Psychological Research
DOI 10.1007/s00426-014-0555-7

ORIGINAL ARTICLE

Bottom-up influences of voice continuity in focusing selective auditory attention

Scott Bressler · Salwa Masud · Hari Bharadwaj · Barbara Shinn-Cunningham

Received: 16 August 2013 / Accepted: 19 February 2014

Abstract Selective auditory attention causes a relative enhancement of the neural representation of important information and suppression of the neural representation of distracting sound, which enables a listener to analyze and interpret information of interest. Some studies suggest that in both vision and in audition, the "unit" on which attention operates is an object: an estimate of the information coming from a particular external source out in the world. In this view, which object ends up in the attentional foreground depends on the interplay of top-down, volitional attention and stimulus-driven, involuntary attention. Here, we test the idea that auditory attention is object based by exploring whether continuity of a non-spatial feature (talker identity, a feature that helps acoustic elements bind into one perceptual object) also influences selective attention performance. In Experiment 1, we show that perceptual continuity of target talker voice helps listeners report a sequence of spoken target digits embedded in competing reversed digits spoken by different talkers. In Experiment 2, we provide evidence that this benefit of voice continuity is obligatory and automatic, as if voice continuity biases listeners by making it easier to focus on a subsequent target digit when it is perceptually linked to what was already in the attentional foreground. Our results support the idea that feature continuity enhances streaming automatically, thereby influencing the dynamic processes that allow listeners to successfully attend to objects through time in the cacophony that assails our ears in many everyday settings.

Introduction

In a complex scene, more information reaches the ears and eyes than can be processed in detail. What information is actually processed is determined by a complex interplay between what is inherently salient in the sounds and sights reaching an observer and what information he or she is trying to attend, a form of "biased competition" between bottom-up stimulus characteristics and top-down task goals (e.g., see Desimone & Duncan, 1995). Top-down attention is often controlled by observers focusing on some stimulus feature, some attribute that differentiates the inputs. Volitional attention to features has been shown to enhance filtering of information to favor processing of stimuli with that feature (Desimone & Duncan, 1995; Kidd, Arbogast, Mason, & Gallun, 2005; Lakatos et al., 2013; Marrone, Mason, & Kidd, 2008; Zion Golumbic et al., 2013). In vision, features like spatial location, color, texture, or orientation can be used to direct attention; likewise, in audition, attention can be directed to features such as spatial location, pitch, and voice quality.

The relationship between attentional processes and auditory scene analysis (or how we organize the content of the acoustic mixture reaching the ears into discrete perceptual objects; Bregman, 1990) has been a matter of great debate. Some previous auditory studies argue that attention is critical for auditory stream segregation; only when a

Electronic supplementary material The online version of this article (doi:10.1007/s00426-014-0555-7) contains supplementary material, which is available to authorized users.

S. Bressler · S. Masud · H. Bharadwaj · B. Shinn-Cunningham (✉)
Center for Computational Neuroscience and Neural Technology, Boston University, 677 Beacon St., Boston, MA 02421, USA
e-mail: shinn@cns.bu.edu

S. Masud · H. Bharadwaj · B. Shinn-Cunningham
Department of Biomedical Engineering, Boston University, Boston, MA, USA

Published online: 16 March 2014

stream is attended is it segregated from a sound mixture (Alain & Woods, 1997; Carlyon, Cusack, Foxton, & Robertson, 2001; Cusack, Deeks, Aikman, & Carlyon, 2004; Jones, 1976). Others provide evidence that the formation of auditory objects (estimates of the content of independent, discrete physical sources of sound present in the acoustic mixture) is automatic and stimulus driven, operating independently from attention (Bregman, 1990; Macken, Tremblay, Houghton, Nicholls, & Jones, 2003; Sussman, Horvath, Winkler, & Orr, 2007). The truth likely lies somewhere in between: stream formation can be driven automatically by stimulus features that define a target object unambiguously; however, in some cases, attention to a particular perceptual feature may help "pull out" an auditory stream from an ambiguous sound mixture (e.g., see Alain, Arnott, & Picton, 2001). For instance, a recent study showed that continuity of a task-irrelevant feature (either spatial location or pitch) biased listeners instructed to attend to the other feature (pitch or location), supporting the idea that bottom-up processes automatically stream sound elements together through time (Maddox & Shinn-Cunningham, 2012). However, how a particular acoustic mixture was parsed into streams depended on which feature was being attended, and the perceptual strength of the task-relevant feature determined how strongly continuity of the task-irrelevant cue biased performance, showing that object formation was influenced by attentional focus. These results support the idea that stream formation (the formation of an auditory object that extends through time) and top-down attention are not independent, but rather interact in determining how a complex auditory scene is parsed and what information in the scene is attended.

This kind of interaction between object formation and attention is consistent with the idea that auditory attention is object based. Theories of object-based attention argue that observers parse the complex field into separate objects, and then focus attention on one object at a time, even when attention is first focused on a particular stimulus feature (Desimone & Duncan, 1995; Duncan, 1984; Fritz, Elhilali, David, & Shamma, 2007; Shinn-Cunningham & Best, 2008; Shinn-Cunningham, 2008). Recent studies argue that when attention is focused on a given auditory feature, like a location, all perceptual features associated with an acoustic source at the desired location are perceptually bound together and enhanced in the neural representation (Shamma, Elhilali, & Micheyl, 2011).

If auditory objects are the focus of attention, it makes sense that the process of auditory scene analysis is intimately linked to how auditory attention operates. For instance, in a complex auditory scene, object formation can take time (e.g., see Cusack et al., 2004) and can be unstable (e.g., see Hupe, Joffo, & Pressnitzer, 2008; Pressnitzer & Hupe, 2006), so there should be an interaction between the dynamics of object formation and those of selective attention. Unfortunately, most past work on attention studied the visual modality, where the dynamics of attention are less critical than in audition; as a result, little effort has gone into studying the dynamics of sustaining attention to an ongoing object, and relatively little is known about the dynamics of attention.

While it has been shown that switching attention requires time (Shomstein & Yantis, 2004), sustaining attention on an ongoing stream in a complex acoustic scene has its own dynamics that directly impact the ability to understand sound sources in a complex acoustic setting (Best, Ozmeral, Kopco, & Shinn-Cunningham, 2008; Best, Shinn-Cunningham, Ozmeral, & Kopco, 2010). The study that motivated the current experiments explored the ability to focus attention on a sequence of target digits in the presence of competing digits spoken by different talkers, which were presented simultaneously from different loudspeakers; lights on top of the loudspeakers were used to identify which loudspeaker was presenting the target in each time position (Best et al., 2008). When the target location jumped from one loudspeaker to the next, performance was worse than when the target location was fixed. This difference was due to a "buildup" of spatial auditory attention when the location was fixed, with performance getting better and better from digit to digit in the sequence. As the inter-digit delay (IDD) increased, the benefit of spatial continuity became smaller. When listeners had to switch attention from one location to another, there was almost no advantage of having the lights on the loudspeakers switch before the digits began compared to having the lights turn on simultaneously with the spoken digits. Together, these results showed that the "cost" of switching attention was not truly a cost of disengaging and reengaging attention, but rather a relative benefit of keeping attention focused at one location in space. The benefit of spatial continuity was enhanced when the target talker identity was the same from one digit to the next and voice continuity worked in concert with spatial continuity. A series of follow-up experiments showed that these effects were due to perceptual continuity of space, not to prior knowledge of the location or other potential confounds in the original study (Best et al., 2010). These studies demonstrate that when targets were defined by their spatial position, perceptual continuity of the target location and of the target talker identity (two features that enhance grouping of the target digits into a unified perceptual stream) enhances the ability of listeners to attend to a sequence of target digits.

Since both spatial and non-spatial acoustic features affect object formation, both should contribute to the dynamics of selective attention if auditory attention operates on perceptual objects and streams. Imagine ongoing

speech from a particular talker. Each word is an element of a larger speech stream, and a word that is in the focus of attention at one instant shares non-spatial perceptual features with subsequent words (e.g., similarity of pitch, of voice quality, of location, etc.). If attention operates on objects, then those subsequent words should be more likely to be the focus of attention in the future simply because they are more likely to be perceived as part of the currently attended object—due to the continuity of the speech features. Here we ask whether the non-spatial feature of talker identity enhances selective auditory attention through time in a manner analogous to the buildup of spatial selective attention, even in the absence of spatially directed attention. In Experiment 1, we show that when listeners are asked to report back a sequence of spoken target digits embedded in time-reversed digits spoken by other talkers, continuity of target talker identity enhances performance. In Experiment 2, we show that even when subjects cannot predict when the target talker will repeat, continuity of a task-irrelevant feature (talker identity) enhances performance. Together, results of these two experiments support the idea that perceptual continuity of the target voice enhances attention through time in a bottom-up, automatic manner, lending further credence to the idea that auditory selective attention is object based.

Methods

Participants

Eleven subjects (6 males, 5 females; 19–23 years of age) participated in Experiment 1. Seven subjects (4 males, 3 females; 19–26 years of age) participated in Experiment 2. All subjects were screened to confirm that they had pure-tone thresholds within 20 dB of normal hearing limits for frequencies between 250 Hz and 8 kHz. Proper informed consent was obtained, consistent with Boston University Institutional Review Board protocols. Subjects were compensated at an hourly rate for their participation.

Stimuli

Stimuli consisted of the digits 1–9 spoken by two male talkers and two female talkers. Digits were recorded in a sound-protected booth with a large diaphragm condenser microphone (AudioTechnica AT4033, Stow, OH, USA) through a Duet analog-to-digital interface (Apogee Electronics Corp., Santa Monica, CA, USA) at a sampling rate of 44.1 kHz at 16-bit resolution. Sound files were edited on the digital audio workstation, Digital Performer 7 (MOTU, Cambridge, MA, USA). The resulting sound files were then down-sampled in MATLAB (MathWorks, Natick, MA, USA) to 24.414 kHz to accommodate the sampling frequency of the digital-to-analog (D/A) conversion hardware (RP2.1 Enhanced Real-Time Processor, Tucker-Davis Technologies).

For each trial, the target sequence consisted of five digits, chosen randomly, except for constraining the selected sequence of target talkers appropriately (differently for Experiment 1 and Experiment 2; see below). In both experiments, three different masker digits were presented at the same time as each of the target digits. The digits making up the three maskers were time reversed, rendering them unintelligible. The target digit and masker digits that were presented simultaneously each came from a different talker; since there were four talkers and four digits at a time, each talker was present in each of the five Digit Positions making up a trial. The two male and two female talkers were relatively distinct and differed in their average fundamental frequencies (94.71 and 123.19 Hz for the two male talkers and 175.82 and 201.48 Hz for the female talkers). The target-to-masker ratio was −10 dB. Sample stimuli are provided online as supplementary materials.

General procedures

The experiment took place in a sound-insulated booth (Industrial Acoustics Company, Inc.) with interior dimensions of 2.13 m × 2.23 m × 1.98 m. Stimuli were presented diotically through insert-ear ER-1 Earphones (Etymotic Research, Inc.) via the Tucker-Davis Technologies D/A converter and headphone amplifier (HB7 Headphone Driver, Tucker-Davis Technologies). Subjects indicated their response using a number pad graphical user interface designed in MATLAB (MathWorks, Natick, MA, USA). The experiments were self-paced, and subjects were required to enter in a five-digit response to advance to the next trial. In the event the subjects did not hear all five digits, they were instructed to guess, paying attention to the order of presentation. No feedback was provided. MATLAB software was used for stimulus generation, stimulus presentation, data acquisition and analysis.

Statistical tests

Unless otherwise specified, statistical inference was performed by fitting mixed-effects models to the data and adopting a model comparison approach (Baayen, Davidson, & Bates, 2008; Box & Tiao, 1992). Fixed-effects terms were included for the various experimental factors whereas subject-related effects were treated as random. Homoscedasticity of subject-related random effects was not assumed and hence the error terms were allowed to vary and be correlated across the levels of fixed-effects factors. To not over-parameterize the random effects, the

random terms were pruned by comparing models with and without each term using the Akaike information criterion and log-likelihood ratios (Pinheiro & Bates, 2000). The specific random-effects terms included in each case are noted along with the description of the results. All model coefficients and covariance parameters were estimated using restricted maximum likelihood as implemented in the lme4 library in R (Bates, Maechler, Bolker, & Walker, 2013). To make inferences about the experimental fixed effects, the F approximation for the scaled Wald statistic was employed (Kenward and Roger 1997). This approximation is more conservative in estimating Type I error rates than the Chi-squared approximation of the log-likelihood ratios and has been shown to perform well even with fairly complex covariance structures and small sample sizes (Schaalje, Mcbride, & Fellingham, 2002). The p values and F-statistics based on this approximation are reported.

Data analysis

In a similar task when attention is directed to spatial location, performance improves from digit to digit when spatial cues are consistent from one digit to the next (Best et al., 2008). Therefore, we analyzed the probability of reporting a digit correctly as a function of the temporal Digit Position (1–5). In the previous study, the improvement from digit to digit was attributed primarily to improvements of the specificity of spatial auditory attention; however, this interpretation rested on the fact that the competing digits were not time reversed. By analyzing the digits reported on incorrect trials, this previous spatial-attention study showed that attention became more and more spatially precise from one digit to the next when the target digit locations were fixed compared to when they varied (cf. Figure 4 in Best et al., 2008). This kind of analysis of response errors is not possible in the current study, since none of the time-reversed competing digits was a valid response. We therefore did a different kind of analysis to quantify the benefit of perceptual continuity of voice.

We hypothesized that voice continuity would help listeners direct attention to a subsequent target digit only if attention was already correctly focused on the previous digit spoken by the same talker. To test this idea, we computed the conditional probability of being correct on digit i given that digit i-1 was correctly identified [$P(C_i|C_{i-1})$] and the conditional probability of being correct on digit i given that digit $i-1$ was incorrectly identified [$P(C_i|NC_{i-1})$]. In general, we expected $P(C_i|C_{i-1})$ to be $>P(C_i|NC_{i-1})$ (some correlation in performance for adjacent digits) due to fluctuations in listener state. For instance, if a subject became drowsy at one point in a session, we expected performance to be lower for all Digit Positions within that set of trials. Similarly, if a subject sneezed just as the stimulus was presented on a trial, they were likely to get most digits wrong in that particular trial. Thus, the key comparison of interest was whether the asymmetry between $P(C_i|C_{i-1})$ and $P(C_i|NC_{i-1})$ was greater when the target talker was the same for digit i and for digit $i-1$ compared to when the target talker differed. Such a difference would provide evidence that voice continuity enhanced the likelihood of correctly attending a subsequent digit only if attention was already focused on a digit that shared perceptual attributes with the subsequent target.

In general, the probability of correctly reporting digit i, $P(C_i)$ can be broken down as

$$P(C_i) = P(C_i|C_{i-1})P(C_{i-1}) + P(C_i|NC_{i-1})[1 - P(C_{i-1})] \quad (1)$$

where $P(C_i|C_{i-1})$ and $P(C_i|NC_{i-1})$ are the conditional probabilities of being correct on digit i given that the response to digit $i-1$ was correct or incorrect, respectively. Thus,

$$\begin{aligned} P(C_i) &= [P(C_i|C_{i-1}) - P(C_i|NC_{i-1})] \cdot P(C_{i-1}) \\ &\quad + P(C_i|NC_{i-1}) \text{ or} \\ P(C_i) &= \Delta_i \cdot P(C_{i-1}) + P(C_i|NC_{i-1}), \\ \text{for } \Delta_i &= [P(C_i|C_{i-1}) - P(C_i|NC_{i-1})]. \end{aligned} \quad (2)$$

Ignoring any effects of absolute temporal position, one might assume that $P(C_i|NC_{i-1})$ should equal the probability of getting the first digit in the sequence correct, $P(C_1)$, and that the difference Δ_i is independent of temporal position. Under these circumstances, then Eq. 2 predicts

$$P(C_i) = \Delta \cdot P(C_{i-1}) + P(C_1). \quad (3)$$

Given that $P(C_1)$ is nonnegative, as long as $\Delta > 0$ [i.e., $P(C_i|C_{i-1}) > P(C_i|NC_{i-1})$] one should see an improvement in performance from digit to digit, like that observed in Best et al. (2008). However, these assumptions ignore the fact that when observers are asked to immediately recall a list of items, they often are better at recalling the first and last items in the list, a pattern typically attributed to memory effects (primacy and recency effects; see Jahnke, 1965). In the spatial-attention experiment motivating the current study, such memory order effects may have been weak relative to other temporal order effects, leading to a buildup of performance from digit to digit when target location was the same across digits; however, a failure to see performance improving from digit to digit does not mean that there is not an advantage of stimulus continuity. Here, we asked whether $P(C_i|C_{i-1}) > P(C_i|NC_{i-1})$, and whether this difference was larger when digit i and digit $i+1$ were from the same talker compared to when they were from different talkers.

Finally, to quantify and summarize these conditional probability effects, we calculated the previous digit correct benefit (PDCB) as the log of the ratio of the average probabilities of being correct conditioned on whether or not the previous digit was correct (averaged over digits 2–5):

$$\text{PDCB} = \log\left(\frac{\bar{P}(C_i|C_{i-1})}{\bar{P}(C_i|NC_{i-1})}\right) \quad \text{for } \bar{P}(\cdot) = \frac{1}{4}\sum_{i=2}^{4} P(\cdot). \tag{4}$$

PDCB would be zero if being correct on the digit $i - 1$ had no effect on the probability of getting digit i correct (i.e., the ratio of the conditional probabilities equaled one), and positive if being correct on digit $i - 1$ increased the probability of getting digit i correct. We expected the PDCB to be positive in general, but larger in conditions where the target talker was the same for digits i and $i - 1$ than when the target talker differed.

Experiment 1: blocked by Fixed or Changing Voice

Procedures

In Experiment 1, trials were blocked based on (1) whether the target talker (forward speech) was the same from digit to digit within a trial (Fixed Voice), or switched between every digit in the target sequence so that no two adjacent digits had the same voice (Changing Voice), and (2) in the duration of the silent gap between digits (0 and 500 ms), changing the rate of the isochronous mixture (see Fig. 1). All four combinations of voice condition and IDD were tested. As illustrated in Fig. 1, in the Fixed Voice conditions the target talker varied from trial to trial, but was the same for all Digit Positions within a given trial. Statistically, the distribution of the target digits and reversed masker digits presented in each Digit Position was identical across Digit Positions and was the same in both Fixed Voice and Changing Voice trials; the only difference between the Fixed Voice and Changing Voice blocks was in the continuity of the target talker within a given trial.

Each subject completed 2 days of testing. On each day, they performed four experimental blocks, one of each type of trial. The blocks were randomly ordered, separately for each subject. Each block included 50 trials of the appropriate type. Before each block, subjects were informed as to whether the trials were Fixed or Changing Voice. To mitigate any effects of learning, only results from the second session were analyzed.

Results

Overall, listeners performed better in the Fixed Voice than in the Changing Voice condition for both IDDs (filled symbols are above open symbols in Fig. 2). Performance tended to be better for the first and last digits than the intermediate digits, consistent with primacy/recency effects in recalling lists; however, the way in which performance varied with Digit Position was different for the two IDDs. For the 500 ms IDD, the Fixed Voice performance was better than the Changing Voice performance for all Digit Positions; however, for the 0 ms IDD, performance was equivalent for Digit Position 1 in the Fixed Voice and Changing Voice conditions, and then better in the Fixed Voice condition for all subsequent Digit Positions. These observations are supported by the best-fit model tests using the mixed-effects model comparison approach with fixed-effects factors of Voice, IDD, and Digit Position. Random-

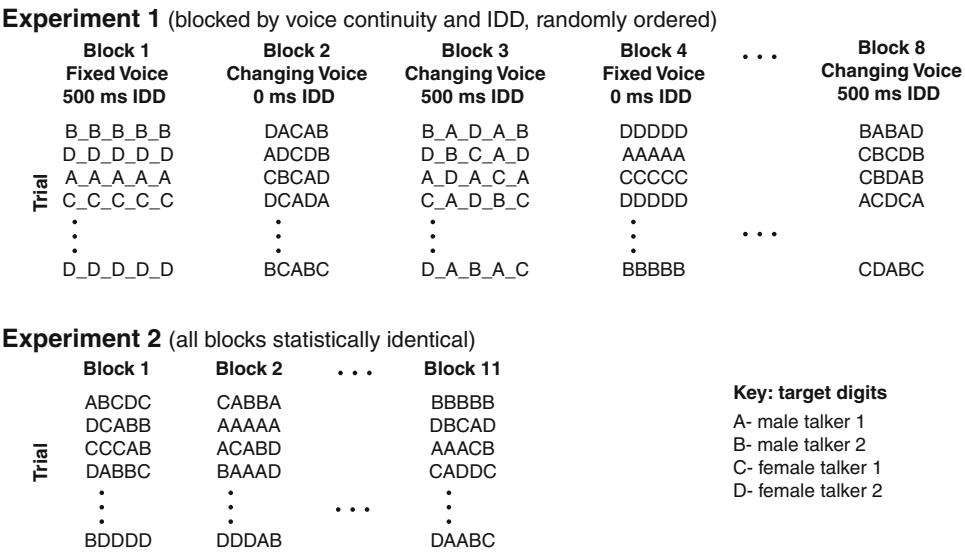

Fig. 1 Schematic illustration showing the target talker sequences used in the different types of blocks in Experiment 1 (*top*) and Experiment 2

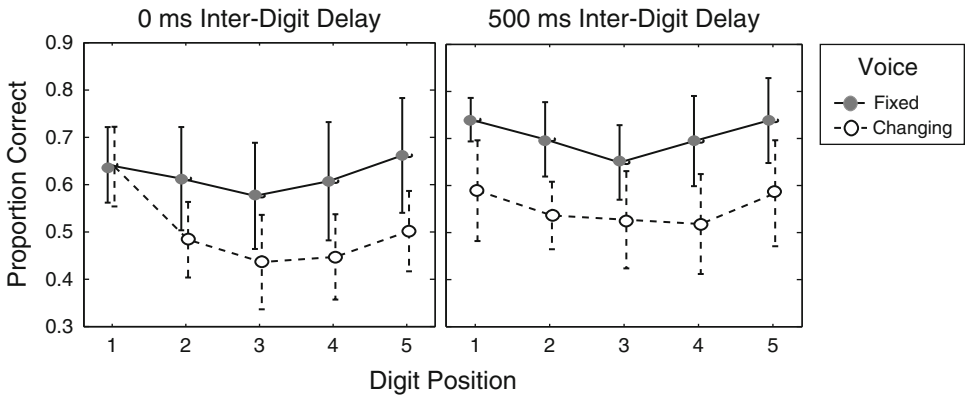

Fig. 2 Across-subject average percent correct in Experiment 1 (±SEM) as a function of Digit Position for the 0 ms IDD (*left panel*) and the 500 ms IDD (*right panel*)

effects terms included the subject-specific intercept, subject-specific slope for Voice, and subject-specific slope for IDD. All three main fixed-effects factors of Voice [$F(1, 10) = 40.37$, $p << 0.0001$], IDD [$F(1, 10) = 15.67$, $p = 0.00271$], and Digit Position [$F(4, 170) = 15.14$, $p << 0.0001$] were significant, as was the second-order interaction of Voice-by-Digit Position [$F(4, 170) = 2.658$, $p = 0.0346$]. No higher order interactions or other two-way interactions were included in the best, most parsimonious model ($p > 0.05$).

Given the significant interactions, we performed separate best-fit model tests for the 0 and the 500 ms results with fixed-effects factors of Voice, and Digit Position and random-effects terms of subject-specific intercept, subject-specific slope for Voice, and subject-specific slope for IDD. For the 0 ms IDD, the main fixed-effects factors of Voice [$F(1, 10) = 10.32$, $p = 0.00928$], and Digit Position [$F(4, 80) = 19.62$, $p << 0.0001$] were significant, as was their interaction, Voice-by-Digit Position [$F(4,80) = 7.703$, $p << 0.0001$]. For the 500 ms IDD, only the two main fixed-effects factors were significant [Voice: $F(1, 10) = 43.29$, $p << 0.0001$; Digit Position: $F(4, 80) = 5.44$, $p = 0.000630$; interaction, $p > 0.05$]. Inspection of Fig. 2 suggests that for the 0 ms IDD, the significant Voice × Digit Position interaction is driven primarily by the fact that performance for the first digit is equal, independent of whether subsequent digits are spoken by the same or a different talker. In contrast, for the 500 ms IDD, whether or not the voice was continuous impacted performance even on the very first digit. As discussed below, the ways in which Digit Position impacted performance are likely due to storage/recall and memory effects. Since our primary interest was in the effects of continuity on performance and our experiment was not designed to tease apart how storage and recall might depend on Digit Position, we did not do additional post hoc statistical analyses on these results.

The probability of reporting a digit correctly was generally higher when the previous digit was correct compared to when the previous digit was incorrect; however, this difference was quite large in Fixed Voice trials and small in Changing Voice trials (see Fig. 3; dark, left bars are higher than the light, right bars in each duplet of bars, but this difference is much larger for the solid bars than the striped bars). To explore whether being correct on the previous digit had a larger effect in the Fixed Voice than in the Changing Voice condition, we used the mixed-effects model comparison approach to test for significant interactions between Voice and previous Digit Response (right or wrong). The model included fixed-effects terms for factors of Voice, IDD, Digit Position, and previous Digit Response. Three random-effects terms were included: subject-specific intercept, subject-specific slope for Voice, and subject-specific slope for previous Digit Response. All four main factors were significant [Voice: $F(1, 10) = 29.33$, $p << 0.0001$; IDD: $F(1, 10) = 14.49$, $p = 0.00345$; Digit Position: $F(3, 293) = 11.84$, $p << 0.0001$; previous Digit Response: $F(1, 10) = 61.60$, $p << 0.001$]. In addition, the interaction Voice × previous Digit Response was statistically significant [$F(1, 293) = 29.33$, $p = 0.000295$], as was the interaction Digit Position × previous Digit Response [$F(3, 293) = 4.003$, $p = 0.00814$], but no other interactions were significant ($p > 0.05$). To better understand the source of this interaction, we performed four post hoc single-tailed paired t tests comparing conditional probabilities for Fixed Voice to those for Changing Voice conditions (Bonferroni-adjusted for multiple comparisons). We found that $P(C_i|C_{i-1})$ was greater for Fixed Voice than for Changing Voice for both IDDs [0 ms: $t(10) = 9.487$, $p << 0.0001$; 500 ms: $t(10) = 8.008$, $p << 0.0001$); however, $P(C_i|NC_{i-1})$ was not statistically different for Fixed Voice and for Changing Voice for either IDD ($p > 0.05$). These results support the hypothesis that continuity of the target talker voice helps listeners once they latch onto the target stream, but not when they miss the previous digit.

Figure 4 shows that the PDCB is large for the two Fixed Voice cases, but is near zero for the two Changing

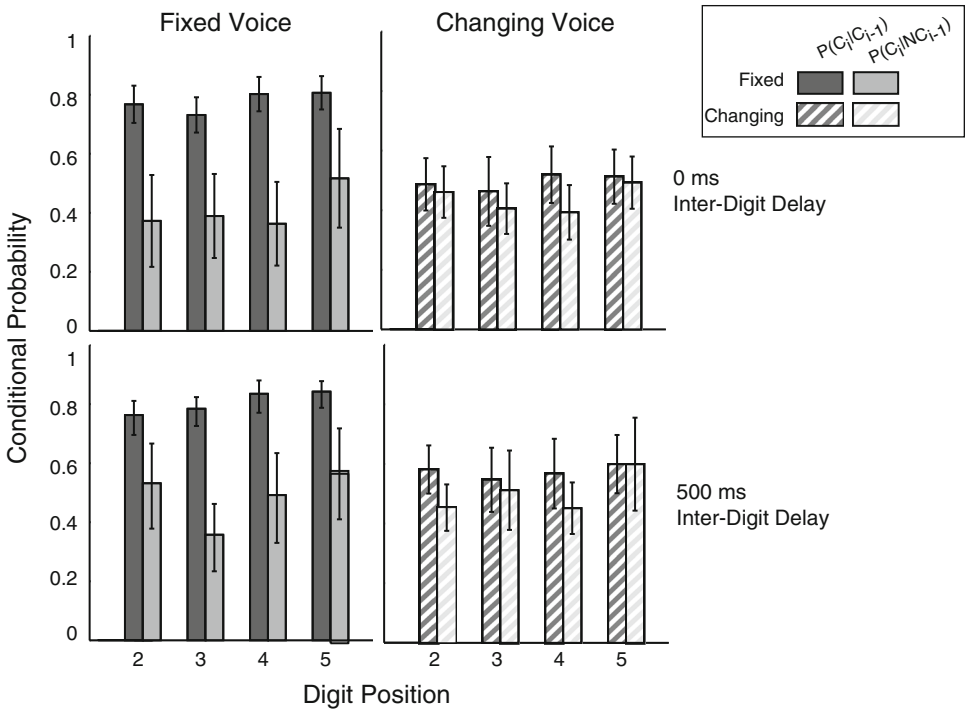

Fig. 3 Across-subject average percent correct in Experiment 1 (±SEM), conditioned on whether or not the previous digit was correctly reported, all as a function of Digit Position. Fixed Voice results are on the *left* and Changing Voice on the *right*. *Top row* shows results for 0 ms IDD, while *bottom row* shows results for 500 ms IDD

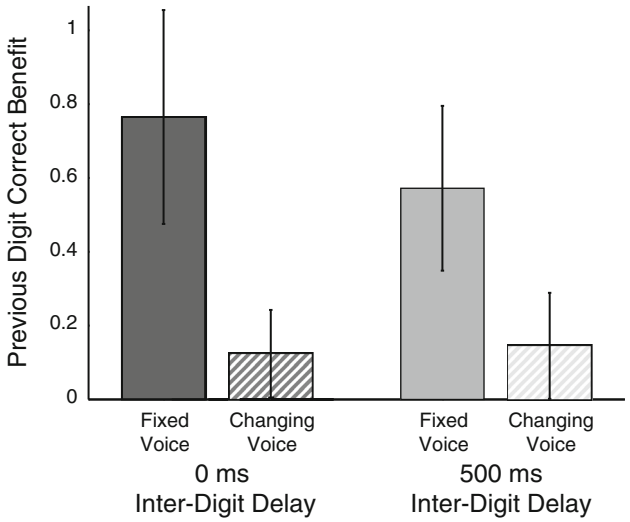

Fig. 4 Across-subject average previous digit correct benefit in Experiment 1 (±SEM) for Repeating Voice and Switching Voice

Voice cases. Three paired one-tailed t tests (Bonferroni-adjusted for multiple comparisons) confirmed that the PDCB was significantly greater in the Fixed Voice than in the corresponding Changing Voice case for both IDDs [0 ms IDD: $t(10) = 3.9762$, $p = 0.0039$, corrected; 500 ms IDD: $t(10) = 4.0276$, $p = 0.0036$, corrected], and that for the Fixed Voice conditions, the PDCB was larger for the faster presentation rates (0-ms IDD) than for the slower 500-ms IDD trials [$t(10) = 2.6673$, $p = 0.0354$, corrected].

Discussion

Results from Experiment 1 show that voice continuity enhances the ability of listeners to report a sequence of digits: overall, Fixed Voice performance is better than Changing Voice performance. Critically, however, this benefit comes about because the PDCB is large when the target talker is fixed compared to when the target talker changes from digit to digit. Indeed, performance in the Fixed Voice cases conditioned on *missing* the previous digit is worse than performance in the Changing Voice cases. This may come about because of how we constructed our stimuli. It may be that, as in the previous study (Best et al., 2008), listeners fail to report the target because they focused attention on a competing (here unintelligible) utterance. In the current experiment, that would correspond to attending to a reversed digit from one of the three other talkers. In the Fixed Voice condition, that other talker is guaranteed to be a distractor voice in the current Digit Position; if the reversed speech in the foreground of attention automatically enhances sound like it, the enhancement favors a distractor. This makes it even more likely that the listener will fail on the current digit than if the talkers were changing randomly.

Performance is better overall when there is a temporal gap between the digits (IDD = 500 ms), and the benefit of voice continuity is greater when the digits are temporally abutting compared to when they are separated (the PDCB in the Fixed Voice condition is significantly larger for 0 ms than for 500 ms IDD). This result is consistent with there

being strong automatic processes of perceptual continuity at play; when the digits are close together in time, streaming effects are stronger.

While we see very clear effects of Voice condition on performance, we did not see an explicit buildup of attention in this experiment. Instead, there is a significant main effect of Digit Position for both 0 and 500 ms IDDs. For the 0 ms IDD, there is a significant interaction between Digit Position and Voice, but not for the 500 ms IDD. The main effect of Digit Position appears to be due to primacy and recency effects; that is, performance is generally better for the first and last digits than for the intermediate digits. This result suggests that memory factors play a large role in the current results. As seen from Eq. 2, if there are differences in the overall likelihood of being correct for digits in different temporal positions, then performance does not necessarily have to improve from one digit to the next even if there is a benefit of perceptual continuity.

The idea that memory plays a significant factor in the current results is further supported by how performance varies across conditions for the very first digit in the sequence. In all cases, before the first digit plays, the knowledge a listener has about what the first target digit will sound like is the same, independent of whether the Voice is Fixed or Changing and whether the IDD is 0 or 500 ms. When the IDD is 0 ms, performance for the first digit is similar for Fixed and Changing Voices; moreover, this level of performance is roughly equal to the performance when the Voice is Changing and the IDD is 500 ms. However, when the Voice is Fixed and there is a large gap between digits, performance is better, even for the very first digit presented. It is possible that this influence of continuity may be due to improved retrospective recall, similar to that observed in reflective attention studies utilizing retro-cueing in delayed match to sample paradigms (Backer and Alain, 2013). Regardless, the fact that the sequence of digits coming later has an effect on performance for the first digit strongly implicates storage/recall processes as playing a key role in these results. Specifically, this finding shows that in the Fixed Voice condition when there is plenty of time to process and store each digit, later-arriving digits in the sequence cause less interference for the first digit compared to when the target talker changes from digit to digit, and compared to when the digits are presented close together in time.

Results of Experiment 1 are consistent with the idea that the benefit of voice continuity arises because in the Fixed Voice condition, the sequence of target digits sound more like a continuous stream: the PDCB is greatest when target digits are from the same talker, close together in time. These results may arise because once a listener is attending to one item in an ongoing stream, the subsequent item in that stream is automatically more likely to win the competition for attention. However, given the blocked design of Experiment 1, it is possible that in the Fixed Voice trials subjects may have volitionally directed their attention to the qualities of the target voice within a given trial. Specifically, because trials were blocked according to whether the target talker was Fixed or Changing, listeners may have picked up on the fact that the target voice repeated from one digit to the next during the Fixed Voice blocks and used this information to direct top-down attention to a subsequent digit once they heard out a preceding target digit. Experiment 2 was designed to more directly test the question of whether the benefit of voice continuity arose because listeners directed top-down volitional attention to the target talker once it was identified.

Experiment 2: within-trial randomization of Fixed vs. Random Voice transitions

Procedures

In Experiment 2, pairs of adjacent digits in each trial were either spoken by the same talker (Repeating Voice), or spoken by different talkers (Switching Voice). Within any given trial, there could be both Repeating Voice and Switching Voice transitions, making it impossible to predict whether the next target digit in a trial would come from the same talker as the previous target digit or from a different talker (see Fig. 1). The number of Repeated Voice transitions in a single trial varied from zero (like a Random Voice trial in Experiment 1) all the way to five (like a Fixed Voice trial in Experiment 1). The trials were all randomly intermingled, so that the number of Repeating Voice transitions was unpredictable throughout a block. Repeated transitions were overall slightly less likely than random transitions (45 vs. 55 % of transitions), and were slightly more likely to occur in the middle of a sequence than between the first two or last two digits (probability of a transition being fixed: $1 \rightarrow 2$, 36 %; $2 \rightarrow 3$, 55 %; $3 \rightarrow 4$, 55 %; $4 \rightarrow 5$, 36 %). Importantly, listeners were not informed that the target voice ever repeated from one digit to the next. Listeners' post hoc reports suggest that they were not consciously aware that the target talker sometimes was the same in two consecutive digits in a trial. Because we were primarily interested in the effects of perceptual continuity, and because in Experiment 1, the PDCB was greater when the digits abutted than when there was a gap between them, we only tested stimuli with a 0 ms IDD.

Before starting, subjects participated in a 25-trial practice session and then performed the 550-trial-long main experiment, which was broken down into 11 statistically identical blocks of 50 trials (containing different patterns of

Repeating Voice and Switching Voice transitions within each trial). Each subject completed 2 days of testing, performing roughly half of the blocks each day. The digit 7 was not included in the digits used to generate stimuli for Experiment 2, as it was a little easier to identify than all other digits, based solely on the fact that it was the only two-syllable-long digit. Other than this minor difference, the mixtures of digits presented at any given Digit Position in Experiment 2 were statistically identical to those used in Experiment 1. The only difference in the stimuli in the two experiments was in the transitions between digits. Specifically, what differed was whether all trials within a block consisted of random mixtures of same-target-talker and different-target-talker transitions (Experiment 2), or whether all trials within a block were made up of either only same-target-talker transitions (Fixed Voice blocks of Experiment 1) or only different-target-talker transitions (Changing Voice blocks of Experiment 1).

Results

Figure 5 plots performance as a function of Digit Position. Because the definition of Repeating Voice and Switching Voice depends on the previous digit and all transitions are randomly intermingled, Digit Position 1 cannot be labeled as either Repeating or Switching, so it is plotted disconnected from the other points. For all other Digit Positions, results are plotted separately, broken down by whether the previous target digit in the trial was from the same talker (Repeating Voice) or from a different talker (Switching Voice). In general, results from Experiment 2 were very similar to the results of Experiment 1 for the 0 ms IDD. Specifically, results for digits after a Repeating Voice transition looked very much like results for Fixed Voice trials from Experiment 1; results for digits after a Switching Voice transitions looked very much like Changing Voice trials from Experiment 1. Listeners performed better in the Repeating Voice than in the Switching Voice condition (filled symbols are above open symbols in Fig. 5). Recency effects were present and of similar magnitude for both Repeating and Switching Voice transitions. A mixed-effects model that included fixed-effects terms for factors of Voice and Digit Position and random-effects terms for subject-specific intercept and subject-specific slope for Voice supports these observations, finding statistically significant effects of Voice [$F(1, 42) = 73.25$, $p << 0.0001$] and Digit Position [$F(3, 42) = 5.608$, $p = 0.00251$], but not their interaction ($p > 0.05$).

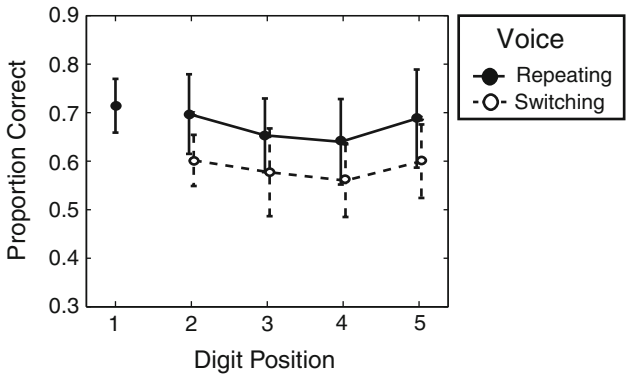

Fig. 5 Across-subject average percent correct in Experiment 2 (±SEM) as a function of Digit Position

As in Experiment 1, the probability of reporting a digit correctly was higher when the previous digit was correct compared to when the previous digit was incorrect, and this difference was much larger in Repeating Voice trials than in Switching Voice transitions (see Fig. 6, where dark, left bars are higher than light, right bars of each duple, and where this difference is much larger for the solid than the striped bar pairs). These observations were supported by a mixed-effects model comparison that included fixed-effects terms for factors of Voice, Digit Position, previous Digit Response, and their interactions. Random-effects terms included subject-specific intercept, subject-specific slope for Voice, and subject-specific slope for previous Digit Response. Results found that all three main effects were significant [Voice: $F(1, 90) = 9.102$, $p = 0.00332$;

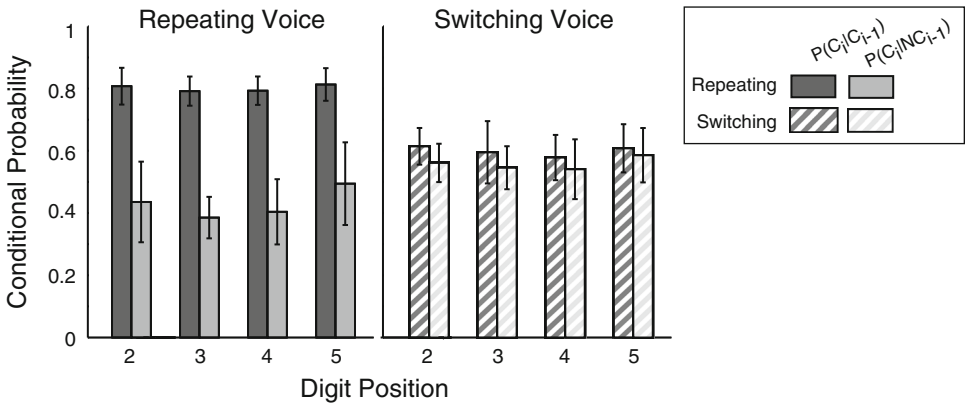

Fig. 6 Across-subject average percent correct in Experiment 2 (±SEM), conditioned on whether or not the previous digit was correctly reported, all as a function of Digit Position. Fixed Voice results are on the *left* and Changing Voice on the *right*

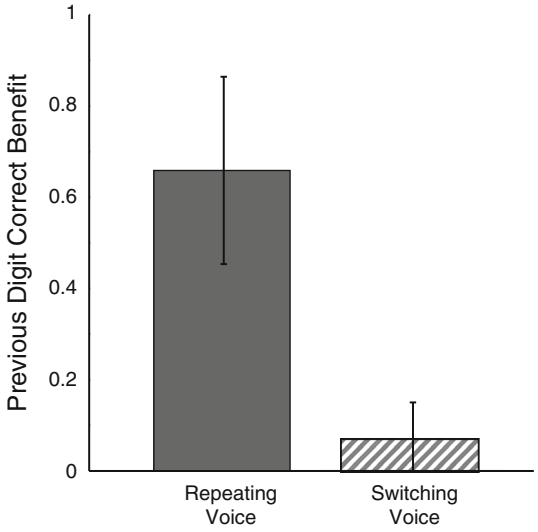

Fig. 7 Across-subject average previous digit correct benefit in Experiment 2 (±SEM) for Repeating Voice and Switching Voice

previous Digit Response: $F(3, 90) = 3.079$, $p = 0.0315$; Digit Position: $F(1, 90) = 283.2$, $p << 0.0001$]. In addition, the Voice × previous Digit Response interaction was significant [$F(1, 90) = 182.3$, $p << 0.0001$], but no other interactions were significant ($p > 0.05$). These results support the conclusion that the benefit of the target voice repeating was driven by cases in which the target voice was both repeated and heard correctly in the preceding Digit Position.

Also consistent with results of Experiment 1, the PDCB was large for the Repeating Voice cases and near zero for the Switching Voice case (see Fig. 7). A one-tailed t test confirmed that the PDCB was significantly greater for Repeating Voice than for Switching Voice transitions [$t(6) = 4.859$, $p = 0.0014$].

Discussion

As noted above, listeners did not report being aware that the target talker sometimes repeated; they were also not told that the target talker ever repeated. Despite this, we cannot completely rule out the possibility that listeners adopted some specific top-down attentional strategy (e.g., attention to the feature of voice pitch or to voice quality). Yet it is difficult to imagine that such a strategy would lead to the pattern of results found here. In the absence of any other clear approach, the most likely top-down strategy would be for listeners to direct attention to whatever pitch or talker they last perceived in a target digit (a strategy that would have been beneficial on 45 % or nearly half of the transitions). Such a strategy could explain why performance was better for Repeating Voice transitions than for Switching Voice transitions. However, it should also lead to differences in the conditional probabilities shown in Fig. 6. Specifically, one would expect performance to be poor for Switching Voice transitions when listeners correctly heard the preceding target digit. On such trials, listeners would always be listening for a pitch or a talker who was uttering a reversed masker digit, since the preceding target (that they just heard) is guaranteed not to be the target talker in the next Digit Position. Then one would expect performance in this condition to be lower than performance in the Repeating Voice condition, conditioned on getting the previous digit incorrect. To address this possibility, we performed a post hoc two-tailed paired t test, comparing $P(C_i|C_{i-1})$ in the Switching Voice condition to $P(C_i|NC_{i-1})$ in the Repeating Voice condition. We found that $P(C_i|NC_{i-1})$ in the Repeating Voice condition was significantly lower than $P(C_i|C_{i-1})$ in the Switching Voice condition [$t(6) = -9.9644$, $p << 0.0001$, corrected], rather than the reverse. In other words, a listener strategy of directing top-down attention to whatever voice was most recently heard predicts that performance should be poor in the Switching Voice condition when listeners heard the previous digit correctly compared to performance on the Repeating Voice when listeners failed to report the previous digit correctly; instead, the reverse is true. To address whether this difference might be solely due to the small temporal correlations that come about from changes in listener state, we then compared $P(C_i|NC_{i-1})$ in the Repeating Voice and Switching Voice conditions (two-tailed paired t test). We found that $P(C_i|NC_{i-1})$ was significantly lower for the Repeating Voice condition than for the Switching Voice condition [$t(6) = -4.1316$, $p = 0.0123$, corrected] suggesting that fluctuations in listener state do not account for the difference between $P(C_i|NC_{i-1})$ in the Repeating Voice condition and $P(C_i|C_{i-1})$ in the Switching Voice condition. Thus, although further experiments are necessary to confirm our conclusion that the effects we see are not the result of some top-down, volitional strategy, the overall pattern suggests that voice continuity provides an automatic, rather than top-down, benefit to performance.

General discussion

Our first experiment found that overall, listeners were better at focusing attention on a sequence of target digits embedded in time-reversed speech when the target talker was the same across all digits than when it changed from each digit to the next. This benefit arose because listeners were better at reporting a digit after correctly hearing the preceding digit spoken by the same talker; there was no benefit of the target voice repeating unless the listener correctly heard the preceding, same-talker digit (see

Fig. 3). However, from results of Experiment 1, there is no way to tell if the PDCB is from top-down attention to the target voice, or from automatic streaming processes.

In Experiment 2, listeners were never explicitly told that voices ever repeated, and the pattern of repeating voices was unpredictable, both from trial to trial and, even more importantly, within each trial. Despite this, the pattern of results from Experiment 2 was similar to the results obtained in Experiment 1. In particular, performance was better overall for target digits preceded by a target digit spoken by the same talker. Just as in Experiment 1, the benefit of the repeated target voice derives from a very large benefit on trials when the preceding digit was not only spoken by the same talker, but also was correctly identified by the subject. Assuming that this bias toward talker continuity, a task-irrelevant feature, is automatic and robust, one might also expect to observe similar benefits in cases where the repeating voices are extremely unlikely (i.e., talker voice repeats once on 10 % of all trials). Plans for a follow-up experiment are currently under development.

In both experiments, the mean PDCB is positive for all conditions. But the PDCD is significantly larger when the target talker repeated compared to when it switched (where the benefit was very small). Overall, our results show that continuity of talker voice enhances the ability to focus attention on a sequence of intelligible target digits embedded in competing unintelligible, time-reversed (but otherwise similar) speech sounds. The similarity of the results of the two experiments suggests that this benefit is not driven by a volitional decision to listen for the same talker from one moment to the next. Instead, our results support the hypothesis that the benefit is obligatory and automatic, adding to the body of evidence that at least some of the processes governing object formation and streaming are hard wired and operate even without focused attention (Alain et al., 2001; Maddox & Shinn-Cunningham, 2012; Pressnitzer, Sayles, Micheyl, & Winter, 2008; Shamma & Micheyl, 2010; Sussman et al. 2007).

We hypothesize that once a listener has focused attention on a particular sound element, perception automatically enhances the salience of a subsequent element that is perceived to come from the same external sound source. In this view, perceptual continuity biases attentional competition to favor future elements that fall within the same perceptual stream as the element already in the attentional foreground, automatically enhancing the ability to *maintain* attention on an ongoing perceptual stream. However, the cost of switching attention from one talker after Changing Voice/Switching Voice transitions in the two experiments could also contribute to the differences we see. In the study of spatially focused auditory attention that motivated the current experiments (Best et al., 2008), knowing ahead of time where the next target digit would come from did not aid subjects very much compared to when the visual cue for where to listen was presented simultaneously with the target digits. This finding, along with the pattern of response errors, was taken as evidence that the performance benefit obtained when location was fixed was due to benefits of spatial continuity rather than a reduction in the cost of switching attention. Given the qualitative similarity of the current results to these previous results, we therefore favor the interpretation that voice continuity enhances selective attention, rather than reduces the cost of attention switching; however, additional studies are necessary to fully tease apart these two possibilities. Regardless of whether the obligatory effects we found are due to enhanced maintenance of attention to an ongoing stream or a cost of switching attention between streams, both explanations support the idea that auditory attention is object based.

Acknowledgments This project was supported in part by CELEST, a National Science Foundation Science of Learning Center (NSF SMA-0835976), and by the Office of Naval Research.

References

Alain, C., Arnott, S. R., & Picton, T. W. (2001). Bottom-up and top-down influences on auditory scene analysis: evidence from event-related brain potentials. *Journal of Experimental Psychology: Human Perception and Performance, 27*(5), 1072–1089.

Alain, C., & Woods, D. L. (1997). Attention modulates auditory pattern memory as indexed by event-related brain potentials. *Psychophysiology, 34*(5), 534–546.

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

Backer, K. C., & Alain, C. (2013). Attention to memory: orienting attention to sound object representations. *Psychological Research,*. doi:10.1007/s00426-013-0531-7.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2013). lme4: linear mixed-effects models using Eigen and S4. R package version 1.0-5. http://CRAN.R-project.org/package=lme4.

Best, V., Ozmeral, E. J., Kopco, N., & Shinn-Cunningham, B. G. (2008). Object continuity enhances selective auditory attention. *Proceedings of the National Academy of Science, 105*(35), 13174–13178.

Best, V., Shinn-Cunningham, B. G., Ozmeral, E. J., & Kopco, N. (2010). Exploring the benefit of auditory spatial continuity. *Journal of the Acoustical Society of America, 127*(6), EL258–264.

Box, G., & Tiao, G. (1992). *Bayesian inference in statistical analysis*. New York: Wiley.

Bregman, A. S. (1990). *Auditory scene analysis: the perceptual organization of sound*. Cambridge: MIT Press.

Carlyon, R. P., Cusack, R., Foxton, J. M., & Robertson, I. H. (2001). Effects of attention and unilateral neglect on auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance, 27*(1), 115–127.

Cusack, R., Deeks, J., Aikman, G., & Carlyon, R. P. (2004). Effects of location, frequency region, and time course of selective attention

on auditory scene analysis. *Journal of Experimental Psychology: Human Perception and Performance, 30*(4), 643–656.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review Neuroscience, 18*, 193–222.

Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology: General, 113*(4), 501–517.

Fritz, J. B., Elhilali, M., David, S. V., & Shamma, S. A. (2007). Auditory attention—focusing the searchlight on sound. *Current Opinion in Neurobiology, 17*(4), 437–455.

Hupe, J. M., Joffo, L. M., & Pressnitzer, D. (2008). Bistability for audiovisual stimuli: perceptual decision is modality specific. *Journal of Vision*, *8*(7), 11–15.

Jahnke, J. C. (1965). Primacy and recency effects in serial-position curves of immediate recall. *Journal of Experimental Psychology, 70*, 130–132.

Jones, M. R. (1976). Time, our lost dimension: toward a new theory of perception, attention, and memory. [Research Support, US Gov't, Non-PHS Review]. *Psychological Review*, *83*(5), 323–355.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics, 53*(3), 983–997.

Kidd, G, Jr, Arbogast, T. L., Mason, C. R., & Gallun, F. J. (2005). The advantage of knowing where to listen. *Journal of the Acoustical Society of America, 118*(6), 3804–3815.

Lakatos, P., Musacchia, G., O'Connel, M. N., Falchier, A. Y., Javitt, D. C., & Schroeder, C. E. (2013). The spectrotemporal filter mechanism of auditory selective attention. *Neuron, 77*(4), 750–761.

Macken, W. J., Tremblay, S., Houghton, R. J., Nicholls, A. P., & Jones, D. M. (2003). Does auditory streaming require attention? Evidence from attentional selectivity in short-term memory. *Journal of Experimental Psychology: Human Perception and Performance, 29*(1), 43–51.

Maddox, R. K., & Shinn-Cunningham, B. G. (2012). Influence of task-relevant and task-irrelevant feature continuity on selective auditory attention. *Journal of the Association for Research in Otolaryngology, 13*(1), 119–129.

Marrone, N., Mason, C. R., & Kidd, G. (2008). Tuning in the spatial dimension: evidence from a masked speech identification task. *Journal of the Acoustical Society of America, 124*(2), 1146–1158.

Pinheiro, J., & Bates, D. (2000). *Mixed-effects models in S and S-PLUS* (Vol. Statistics and Computing Series). New York: Springer.

Pressnitzer, D., & Hupe, J. M. (2006). Temporal dynamics of auditory and visual bistability reveal common principles of perceptual organization. *Current Biology, 16*(13), 1351–1357.

Pressnitzer, D., Sayles, M., Micheyl, C., & Winter, I. M. (2008). Perceptual organization of sound begins in the auditory periphery. *Current Biology, 18*(15), 1124–1128.

Schaalje, G., Mcbride, J., & Fellingham, G. (2002). Adequacy of approximations to distributions of test statistics in complex mixed linear models. *Journal of Agricultural, Biological, and Environmental Statistics, 7*, 512–524.

Shamma, S. A., Elhilali, M., & Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. *Trends in Neurosciences, 34*(3), 114–123.

Shamma, S. A., & Micheyl, C. (2010). Behind the scenes of auditory perception. *Current Opinion in Neurobiology, 20*(3), 361–366.

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences, 12*(5), 182–186.

Shinn-Cunningham, B. G., & Best, V. (2008). Selective attention in normal and impaired hearing. *Trends in Amplification, 12*(4), 283–299.

Shomstein, S., & Yantis, S. (2004). Control of attention shifts between vision and audition in human cortex. *Journal of Neuroscience, 24*(47), 10702–10706.

Sussman, E. S., Horvath, J., Winkler, I., & Orr, M. (2007). The role of attention in the formation of auditory streams. *Perception and Psychophysics, 69*(1), 136–152.

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., et al. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron, 77*(5), 980–991.