JARO (2007)
DOI: 10.1007/s10162-007-0073-z

JARO
Journal of the Association for Research in Otolaryngology

Visually-guided Attention Enhances Target Identification in a Complex Auditory Scene

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

Hearing Research Center, Boston University, 677 Beacon St., Boston, MA 02215, USA

Received: 26 October 2006; Accepted: 17 January 2007

ABSTRACT

In auditory scenes containing many similar sound sources, sorting of acoustic information into streams becomes difficult, which can lead to disruptions in the identification of behaviorally relevant targets. This study investigated the benefit of providing simple visual cues for when and/or where a target would occur in a complex acoustic mixture. Importantly, the visual cues provided no information about the target *content*. In separate experiments, human subjects either identified learned birdsongs in the presence of a chorus of unlearned songs or recalled strings of spoken digits in the presence of speech maskers. A visual cue indicating which loudspeaker (from an array of five) would contain the target improved accuracy for both kinds of stimuli. A cue indicating which time segment (out of a possible five) would contain the target also improved accuracy, but much more for birdsong than for speech. These results suggest that in real world situations, information about *where* a target of interest is located can enhance its identification, while information about *when* to listen can also be helpful when targets are unfamiliar or extremely similar to their competitors.

Keywords: spatial attention, temporal attention, cueing, informational masking

INTRODUCTION

Natural acoustic environments are made up of multiple competing sound sources. Often a listener's goal is to pick out and process particular sounds occurring in a mixture of irrelevant sounds. Two kinds of masking make this task difficult. "Energetic" or "peripheral" masking describes interference caused by spectrotemporal overlap of competing sounds, in which portions of the target may become inaudible. "Informational" or "perceptual" masking describes interference at higher stages of processing, whereby maskers disrupt the selective processing of the target (Carhart et al. 1969; Pollack 1975; Watson 1987). Informational masking is related to both the degree of similarity between competing sources and the amount of uncertainty in the task (Wright and Saberi 1999; Lufti et al. 2003; Durlach et al. 2005).

Recent evidence suggests that informational masking can be reduced when listeners attend to a distinguishing feature of the target. For the detection of a pure tone target in the presence of tonal maskers, a priori knowledge about target frequency improves performance (Richards and Neff 2004). For mixtures of natural everyday sounds, attention to the identity of a target source increases the probability that a listener will notice the source disappearing (Eramudugolla et al. 2005). In the spatial domain, Arbogast and Kidd (2000) showed that a priori knowledge about target location in an array of tonal patterns improves a listener's ability to process the target. For speech mixtures, the number of sound sources determines whether spatial cueing enhances target intelligibility. While there is little or no benefit of knowing where to listen for a target in the presence of a single distractor, the benefit increases as the number of

Portions of this work were presented at the 2006 Midwinter meeting of the Association for Research in Otolaryngology.

Correspondence to: Virginia Best • Hearing Research Center • Boston University • 677 Beacon St., Boston, MA 02215, USA. Telephone: +1-617-3538693; fax: +1-617-3537755; email: ginbest@cns.bu.edu

distractors increases (Ericson et al. 2004; Kidd et al. 2005a; Brungart and Simpson 2007). Very few studies have examined the advantage of temporal cueing in auditory tasks. Listeners appear to be relatively unaffected by uncertainty about when the target will occur for simple tone detection tasks (Green and Weber 1980; Wright and Fitzgerald 2004), but they may obtain a benefit from temporal cueing when there is confusion about which of many similar acoustic events correspond to the target.

In the current study, subjects identified a natural target sound embedded in an array of five spatially separated but similar sound sources. Simple visual cues provided information about when and/or where the target would occur in the mixture, but these cues provided no information about target *content*. In experiment 1, subjects identified learned birdsongs in the presence of unlearned songs (a task involving a great deal of informational masking in human listeners, Best et al. 2005). In experiment 2, speech stimuli were used to address the question of whether long-term familiarity with sound sources influences the effectiveness of the different cues. In the visual literature, for example, there is evidence that object-based attention operates more effectively on familiar stimuli (Vecera and Farah 1997; Zemel et al. 2002) and that the speed of visual search depends on target and distractor familiarity (Wang et al. 1994; Shen and Reingold 2001).

METHODS

Subjects

Five subjects (one male, four female, aged 22–27) participated in experiment 1. Subjects S1 and S2 were two of the experimenters and had previously participated in several similar experiments, including many experiments using the birdsong stimuli. Nine subjects (two males, seven females, aged 18–28) participated in experiment 2A. Four subjects (one male, three females, aged 22–28) participated in experiment 2B. Three subjects performed all experiments (S1–S3). Subjects were paid for their participation, and were screened to ensure that they had normal hearing (within 10 dB) for frequencies between 250 and 8 kHz. Experiments were approved by the Boston University Charles River Campus Institutional Review Board.

Stimuli

Each trial consisted of sounds presented simultaneously from five spatially separated loudspeakers. Each loudspeaker presented an ongoing signal that was divided into five contiguous time segments, giving a 5×5 space/time matrix. On any given trial, the target occurred in one of these 25 space/time positions; the other 24 contained maskers. In any time segment, all maskers were different, ensuring that the five channels were spectrotemporally uncorrelated and spatially distinct. Time segments were approximately 600 ms long in experiment 1 and 1,600 ms long in experiment 2, giving approximate total stimulus durations of 3 and 8 s, respectively.

Stimuli for experiment 1 were songs from male zebra finches (*Taeniopygia guttata*). A detailed description of these signals and the recording procedures is given in Best et al. (2005). Subjects were trained to identify five individual birds on the basis of their songs, as described in Procedures. Five separate song tokens from each of the five learned birds were used as targets during the experiment. However, each token from a particular bird was so stereotypical that the token-to-token variability was almost imperceptible. Maskers were unfamiliar songs from ten different birds, recorded in the same fashion as the targets. Within each time segment, the five sounds were set to be all the same length. For time segments containing no target, five maskers were chosen randomly and were set to the length of the shortest masker by windowing the other four maskers with 10-ms cosine squared ramps at onset and offset. For time segments containing a target, four maskers were chosen randomly and were modified to be the same length as the target. Maskers longer than the target were time-windowed, as above. Maskers shorter than the target were looped to make them as long as the target, with a 10-ms cosine squared ramp applied to the end of the looped portion to avoid an abrupt offset.

In experiment 2, targets were taken from a set of 125 sequences of spoken digits taken from the TIDIGIT database. Each sequence comprised five digits from the set 0–9 and was spoken by one of 20 possible male voices. In experiment 2A, the aim was to create maskers that were spectrotemporally similar but unintelligible. This was achieved by concatenating the target sequences together in random order and reversing them in time. Individual maskers were then generated by selecting an arbitrary portion of this string and applying a 10-ms cosine squared ramp to each end. In experiment 2B, intelligible forward speech was used as the interfering stimulus. For each individual masker, a random male-talker sentence was selected from the TIMIT database, cropped to the appropriate duration, and ramped with 10-ms cosine squared ramps. In experiments 2A and 2B, all time segments and thus all maskers were fixed to the length of the target.

Environment

The experiments took place in a single-walled IAC booth with interior dimensions of 12′4″×13′×7′6″ (length, width, height), with perforated metal panels on the ceiling and a carpeted floor. The walls were lined with 8″ wedges of polyurethane Silent Source TF-MAX8® foam (for acoustic analysis of this environment, see Kidd et al. 2005b). The subject was seated on a chair in the center of the room. A headrest attached to the back of the chair cradled the neck and the back of the head to minimize head movements. No instructions were given to subjects regarding eye fixation during stimulus delivery, and eye movements were not measured. Stimuli were presented via five loudspeakers (Acoustic Research 215PS) located on an arc approximately 1 m from the subject at the level of the ears. The loudspeakers were positioned within the visual field of the subject, at lateral angles of −40°, −20°, 0°, 20°, and 40°. Subjects indicated their response using a handheld keypad with an LCD display (QTERM). The booth was kept dark during the experiment, except for a small lamp placed on the floor behind the listener, which helped them to see the keypad.

Digital stimuli were generated and selected via a PC located outside the booth, and fed through five separate channels of Tucker–Davis Technologies hardware. Signals were converted at 40 kHz by a 16-bit D/A converter (DA8), attenuated (PA4), and passed through power amplifiers (Tascam) before presentation to the loudspeakers. Each loudspeaker had an LED affixed on its top surface, which could be turned on and off via the PC using a custom-built switchboard. MATLAB software was used for stimulus generation, stimulus presentation, data acquisition, and analysis.

Procedures

Identification training. For experiment 1, subjects were trained to identify the five target birds on the basis of their unique song motifs. Each target bird was given a name ("Uno", "Junior", "Moe", "Toro", and "Nibbles") that subjects were trained to associate with the specific motifs. First, subjects were presented with example motifs from each of the birds to familiarize them with the songs. After familiarization, training tests were performed, each of which consisted of 25 presentations (five motifs from each bird) from the center loudspeaker in a quiet environment (with no maskers). A numbered list of the five learned birds was posted in front of subjects, and they responded by pressing one of five buttons on the handheld keypad. After a response, they received correct answer feedback from the LCD display. The training test was repeated until a subject achieved 100% accuracy on two consecutive tests. Training was carried out at the beginning of each new testing day to ensure that subjects continued to reliably identify the birds.

For experiment 2, no familiarization was necessary as the targets consisted of spoken numbers, familiar to all subjects.

Testing procedures. In an experimental test, the subject's task was to listen for a familiar target and identify it using the response keypad. For experiment 1, a response was indicated by pressing the number corresponding to the target bird. For experiment 2, a response was a five-digit sequence, which was only scored as correct if all five digits were reported correctly and in the right order (note that other less conservative scoring schemes were explored, but all lead to similar conclusions).

Each subject completed five sessions in an experiment, each on a separate day. A session consisted of four tests, one for each of the four attention conditions (see below). The order of the four conditions was random and different between sessions and subjects. Subjects were informed at the beginning of each test as to the kind of visual cue they would receive during that test.

In experiment 1, a single test consisted of 125 trials. This corresponded to each of the five birds being presented once at each of the 25 unique space/time positions within the stimulus (in random order). Across the five sessions, each subject gave a total of 625 responses in each condition. In experiments 2A and 2B, a single test consisted of 25 trials. These 25 trials used 25 different target sequences at each of the 25 unique space/time positions within the stimulus (in random order). Across the five sessions for each experiment, each subject gave a total of 125 responses in each condition.

Conditions. The four conditions were designed to manipulate the attention of the subject, and were as follows (see Fig. 1):

1. *No cue*: no visual cue was given.
2. *Where*: the LED located on the target loudspeaker lit up synchronously with the onset of the first time segment and remained on for the entire stimulus.
3. *When*: all of the five LEDs lit up at the start of and were turned off at the end of the time segment containing the target.
4. *Where and when*: the LED located on the target loudspeaker lit up only for the duration of the time segment containing the target.

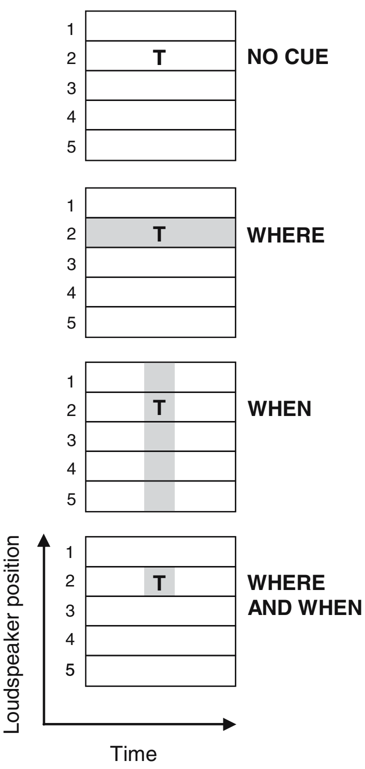

FIG. 1. The four attention conditions. Each panel shows a schematic time course for each of the five loudspeakers. The "T" indicates that the target would occur in a particular loudspeaker (number 2 in this example) and at a point in time corresponding to one of five random-length time segments. The shaded region indicates in which loudspeakers and time segments the LEDs would be active. Note that in the *when* and *where and when* conditions, the LEDs came on synchronously with the onset of the auditory target.

RESULTS

Experiment 1

Overall performance. Percent correct scores in experiment 1 are shown in Figure 2a. The first five groups of bars represent individual data from the five subjects, with the mean data shown in the right-most group of bars (error bars represent the across-subject standard error of the mean). The four bars within a group represent the four attention conditions, as labeled.

Individual differences in the percent correct scores are apparent. Subjects S1 and S2 (the two listeners with the most experience with these stimuli) performed consistently better than the other subjects. In general, performance was poorest with no cue (black bars), although it was always better than chance (with five possible targets, chance performance is 20%, but the lowest score was 27% and the average was 43%). Performance was always most accurate in the *where and when* condition (white bars). The *where* condition and the *when* condition produced intermediate results, with some subjects doing better with *where* cues than *when* cues and others doing better with *when* cues than *where* cues. A 3-way repeated measures ANOVA was conducted on the percent correct scores with factors of condition, target loudspeaker, and target time segment. The main effect of condition was significant [$F(3,12) = 18.44$, $p < 0.001$]. Post hoc pairwise comparisons ($p < 0.05$) found a significant difference between the *no cue* and *where and when* cue conditions only.

To examine directly the benefit of the different visual cues, scores in the *no cue* condition were subtracted from scores in the other conditions for each individual subject. These "cue benefits" for each subject, as well as the across-subject means and standard errors of the mean, are shown in Figure 2b. The mean data indicate that the *where*, *when*, and *where and when* cues improved percent correct scores by 12, 13, and 25%, respectively. It is interesting to note that the subjects who performed better overall received the most improvement in percent correct scores from the attention cues (compare panels a and b of Fig. 2). Furthermore, listeners S1 and S2, who had the most experience with the stimuli and the task, received relatively more benefit from the *where* cue than the *when* cue. The other subjects showed the opposite effect (S5) or received an equal benefit from the two cues (S3 and S4).

Effect of target spatial location. To examine whether performance varied as a function of the target

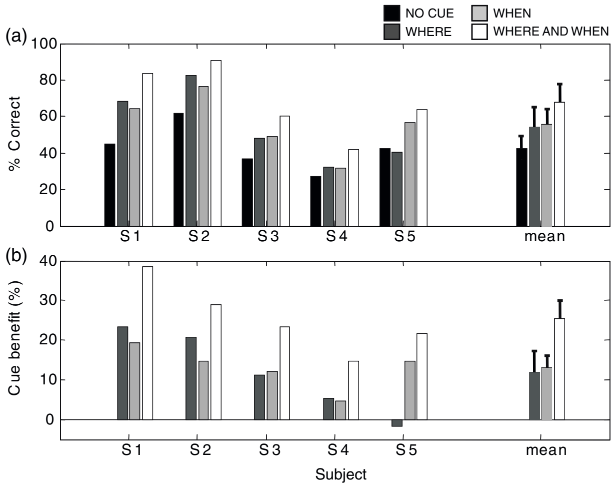

FIG. 2. **(a)** Percent correct scores in experiment 1. Shown are individual data from the five subjects as well as the across-subject mean (error bars represent the standard error of the mean). The four bars within a group represent the four attention conditions, as labeled. **(b)** Cue benefits in experiment 1, calculated by subtracting scores in the *no cue* condition from scores in the other conditions. Shown are the individual benefits as well as the across-subject mean (error bars represent the standard error of the mean).

location in experiment 1, the data were analyzed as a function of which loudspeaker contained the target (regardless of the time segment in which it occurred). Figure 3 shows the across-subject average of performance in the different attention conditions as a function of target location. In all conditions, performance varies in a rough U-shape with target location, meaning that targets to the side tended to be identified more accurately than those towards the center. The ANOVA revealed a significant main effect of target location [$F(4,16) = 11.92$, $p < 0.001$]. There was no significant interaction between target location and condition [$F(12,48) = 0.95$, $p = 0.51$], suggesting that the effect of target location was relatively constant across conditions (consistent with the four curves being approximately parallel). Post hoc pair-wise comparisons ($p < 0.05$) indicated a significant advantage for targets presented from the rightmost loudspeaker compared to those presented from the three central loudspeakers.

Effect of target temporal position. To examine whether performance varied as a function of the temporal position of the target within the trial, the data were pooled across subjects and broken down into five groups based on which time segment contained the target (regardless of the loudspeaker from which it was delivered). Figure 4 shows average performance scores for the different attention conditions as a function of target temporal position. There are not dramatic differences across target time segments; however, there is a slight tendency for scores in the *no cue* and *where* conditions to improve as the target occurs in later time segments. The ANOVA revealed a significant main effect of target time segment [$F(4,16) = 3.86$, $p < 0.05$] and a significant interaction between condition and target time segment [$F(12,48) = 3.06$, $p < 0.005$]. Separate ANOVAs for each condition confirmed that the effect of time segment was only significant in the *no cue* and *where* conditions [$F(4,16) = 6.51$, $p < 0.005$; $F(4,16) = 6.00$, $p < 0.005$], and not in the *when* and *where and when* conditions [$F(4,16) = 1.90$, $p = 0.16$; $F(4,16) = 0.37$, $p = 0.82$]. Post hoc pair-wise comparisons ($p < 0.05$) indicated that scores were higher for targets occurring in the final time segment than those occurring in the first and second time segments (*no cue* condition) or those in the first four time segments (*where* condition).

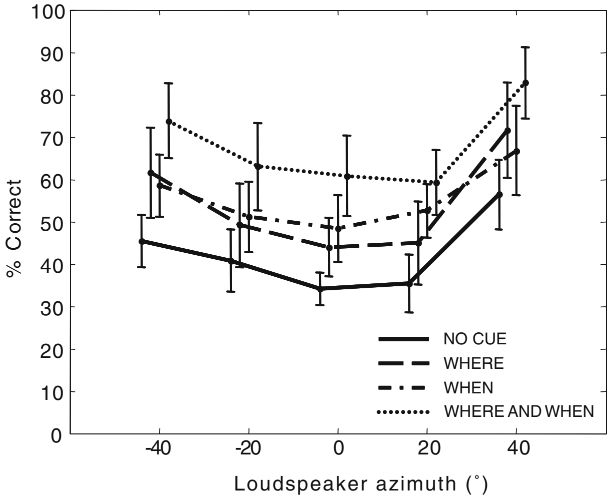

FIG. 3. Mean percent scores in experiment 1 as a function of target loudspeaker (pooled over all subjects and all target time segments). The four lines represent the four attention conditions, as labeled. Error bars indicate standard errors of the across-subject mean, and are staggered horizontally for clarity.

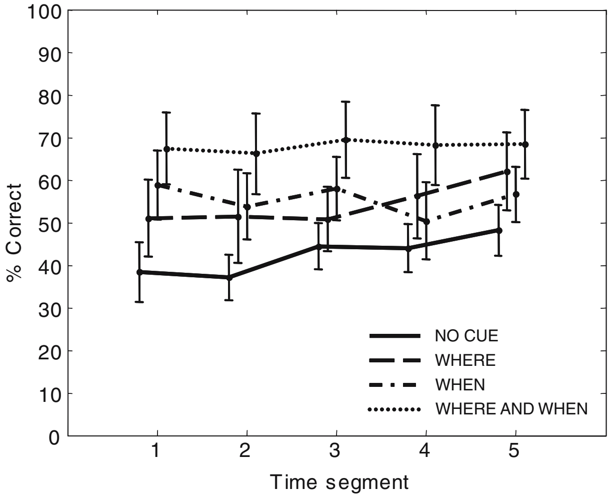

FIG. 4. Mean percent scores in experiment 1 as a function of target time segment (pooled over all subjects and all target loudspeakers). The four lines represent the four attention conditions, as labeled. Error bars indicate standard errors of the across-subject mean, and are staggered horizontally for clarity.

The curves in Figure 4 are roughly parallel for the *no cue* and *where* conditions, reflecting the fact that the *where* cue provided a relatively constant improvement for all target temporal positions. The flat curves for the *when* and *where and when* conditions, however, reflect the fact that a visual cue indicating when to attend afforded a larger advantage when the target occurred in earlier time segments (which produced the poorest scores in the *no cue* condition) compared to later ones.

Experiment 2A

Overall performance. Percent correct scores in experiment 2A are shown in Figure 5a. The first nine groups of bars represent individual data from the nine subjects, with the mean data shown in the right-most group of bars (error bars represent the across-subject standard error of the mean). The four bars within a group represent the four attention conditions, as labeled.

While there are individual differences in overall accuracy, these differences are not as striking as in experiment 1. Performance was always poorest with no cue (except for S2), but all subjects performed well above chance in this condition. Under the strict scoring system, requiring all five digits to be reported in the right order, the probability of guessing the entire sequence is vanishingly small. However, even if subjects heard most of a sequence and guessed just one digit, chance performance would be only 10%. The lowest score in the *no cue* condition was 41% and the average was 55%. A 3-way repeated measures ANOVA was conducted on the percent correct scores with factors of condition, target loudspeaker, and target time segment. The main effect of condition was highly significant [$F(3,24) = 23.25$, $p < 0.0001$]. Post hoc pair-wise comparisons ($p < 0.05$) indicated that (a) scores in the *no cue* and *when* conditions were not significantly different, (b) scores in the *where* and *where and when* conditions were not significantly different, and (c) scores in the *where* and *where and when* conditions were both significantly larger than in the *no cue* and *when* conditions.

To examine directly the benefit of the different visual cues, scores in the *no cue* condition were subtracted from scores in the other conditions for each individual subject. The cue benefits for each subject are shown in Figure 5b, along with the across-subject mean and standard error of the mean. The *when* cue provided very little benefit overall. In fact, every listener obtained a greater benefit from the *where* cue than the *when* cue, and the benefit of the *where and when* cue was essentially the same as for the *where* cue alone. On average, the *where*, *when*, and *where and when* cues improved percent correct scores by 14, 4, and 15%, respectively.

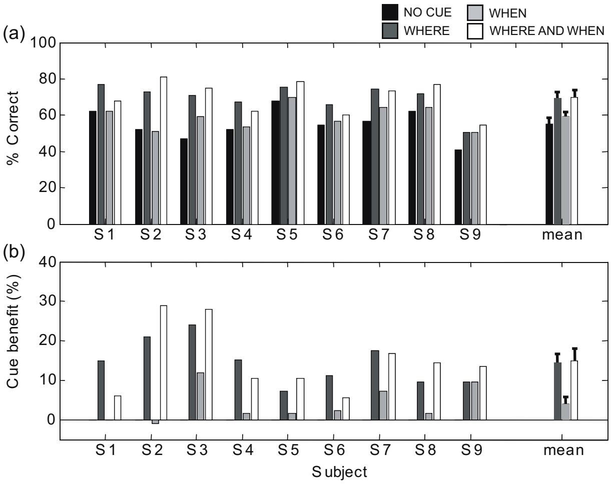

FIG. 5. **(a)** Percent correct scores in experiment 2A. Shown are individual data from the nine subjects as well as the across-subject mean (error bars represent the standard error of the mean). The four bars within a group represent the four attention conditions, as labeled. **(b)** Cue benefits in experiment 2A, calculated by subtracting scores in the *no cue* condition from scores in the other conditions. Shown are the individual benefits as well as the across-subject mean (error bars represent the standard error of the mean).

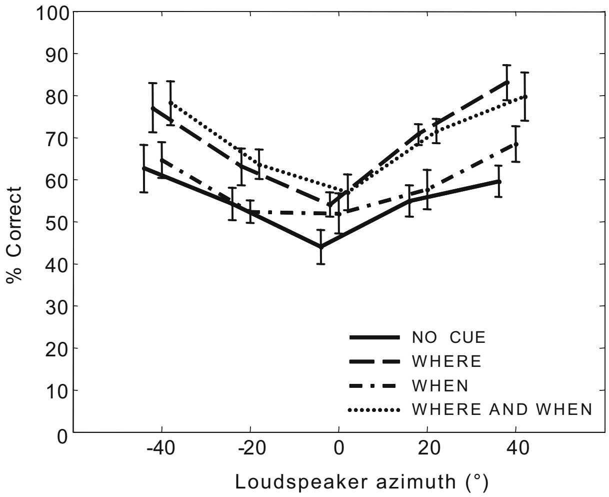

FIG. 6. Mean percent scores in experiment 2A as a function of target loudspeaker (pooled over all subjects and all target time segments). The four lines represent the four attention conditions, as labeled. Error bars indicate standard errors of the across-subject mean, and are staggered horizontally for clarity.

Effect of target spatial location. To examine whether performance varied as a function of the target location in experiment 2A, the data were analyzed in the same way as the data from experiment 1. Figure 6 shows average performance scores for the different attention conditions as a function of target location, in a format identical to that in Figure 3. As in experiment 1, performance in all conditions varies in a rough U-shape with target location, indicating that targets towards the side were identified more accurately than those towards the center. The ANOVA revealed a significant main effect of target location [$F(4,32) = 8.58$, $p = 0.0001$]. There was no significant interaction between target location and condition [$F(12,96) = 1.51$, $p = 0.13$], suggesting that the effect of target location was relatively constant across conditions (consistent with the four curves being approximately parallel). Post hoc pair-wise comparisons ($p < 0.05$) indicated a significant advantage for targets presented from the outermost loudspeakers compared to those presented from the central loudspeaker.

Effect of target temporal position. Figure 7 shows average performance scores for the different attention conditions as a function of target temporal position, in a format identical to that of Figure 4. The ANOVA revealed a significant main effect of target time segment [$F(4,32)=14.53$, $p<0.0001$] and a significant interaction between condition and target time segment [$F(12,96)=2.38$, $p<0.01$]. Separate ANOVAs for each condition showed that the effect of time segment was significant in all conditions [$F(4,32)=7.43$, $p<0.005$; $F(4,32)=15.02$, $p<0.0001$; $F(4,32)=9.39$, $p<0.0001$; $F(4,32)=6.82$, $p<0.0005$], and post hoc pair-wise comparisons ($p<0.05$) indicated that targets in the first time segment were identified more poorly than targets in the third, fourth, and fifth time segments in all conditions. Targets in the final time segment were identified more poorly than in the central time segment in the *no cue* condition only.

The pattern of scores as a function of target time segment is quite different from that seen in experiment 1. Whereas in experiment 1, baseline (*no cue*) performance improved for targets occurring in later time segments, *no cue* performance in experiment 2A was most accurate for a target occurring in the intermediate time segment (segment 3). Furthermore, whereas the *when* cue provided the most benefit for earlier time segments in experiment 1, in experiment 2A the *when* cue only aided the identification of targets occurring in the final time segment. Moreover, *where* and *where and when* cues gave essentially identical benefits, independent of time segment.

Experiment 2B

Percent correct scores are shown in Figure 8a. The first four groups of bars represent individual data from the four subjects, with the mean data shown in the right-most group of bars (error bars represent the across-subject standard error of the mean). The four bars within a group represent the four attention conditions, as labeled. A repeated-measures ANOVA confirmed that the effect of condition was significant [$F(3,9)=11.22$, $p<0.01$] and although post hoc pair-wise comparisons did not reach significance, the difference between the *no cue* and *where and when* cue conditions was on the margin ($p=0.055$).

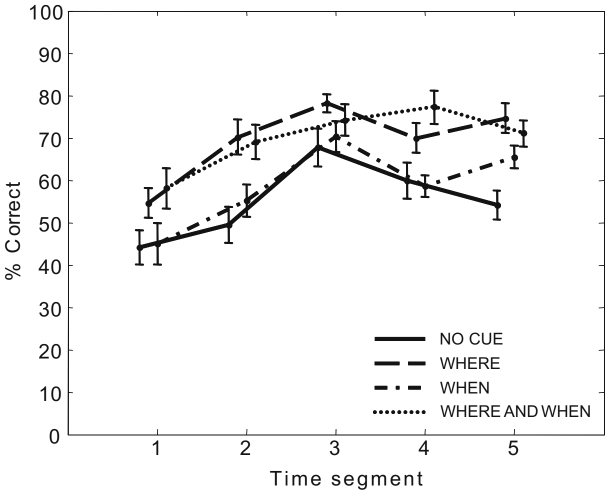

FIG. 7. Mean percent scores in experiment 2A as a function of target time segment (pooled over all subjects and all target loudspeakers). The four lines represent the four attention conditions, as labeled. Error bars indicate standard errors of the across-subject mean, and are staggered horizontally for clarity.

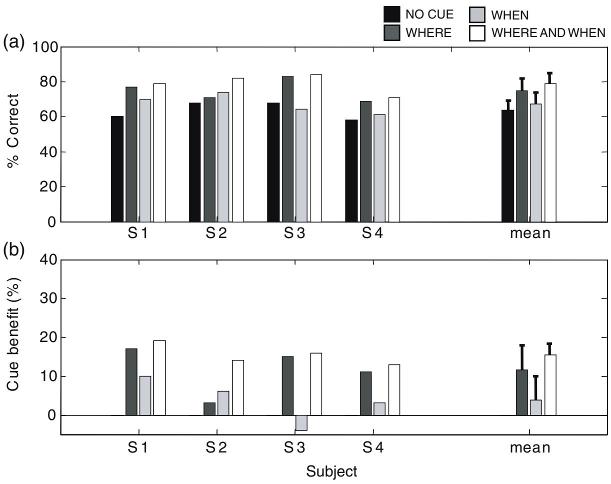

FIG. 8. **(a)** Percent correct scores in experiment 2B. Shown are individual data from the four subjects as well as the across-subject mean (error bars represent the standard error of the mean). The four bars within a group represent the four attention conditions, as labeled. **(b)** Cue benefits in experiment 2B, calculated by subtracting scores in the *no cue* condition from scores in the other conditions. Shown are the individual benefits as well as the across-subject mean (error bars represent the standard error of the mean).

As in the preceding experiments, scores in the *no cue* condition were subtracted from scores in the other conditions to show directly the benefit of the different visual cues. The cue benefits for each subject are shown in Figure 8b, along with the across-subject mean and standard error of the mean. Most importantly, the *when* cue provided very little benefit overall and most improvements were due to the *where* cue. On average, the *where*, *when*, and *where and when* cues improved percent correct scores by 12, 4, and 16%, respectively. Overall, the results are very similar to those obtained in experiment 2A (compare to Fig. 5).

DISCUSSION

Evidence for energetic and informational masking

Target identification in both experiments was likely limited both by the audibility of the target ("energetic" masking) as well as confusion caused by the maskers, which were highly salient and similar to the targets ("informational" masking).

The U-shaped pattern of scores as a function of target location (Figs. 3 and 6) that was observed in all conditions may reflect differences in the target-to-masker ratio (TMR) at the ear that is acoustically more advantageous (the better ear; Zurek 1993). To investigate this idea, better-ear TMRs were calculated for each target location. For each target position, the 125 stimuli from session one of experiment 1 were analyzed (corresponding to the five different target birds played from the five loudspeaker positions in the five different time segments). For each stimulus, the target and maskers from the target time segment were filtered using generic head-related transfer functions corresponding to the respective loudspeaker locations. The broadband root-mean-square TMR at the better ear was calculated for every stimulus example. The TMR is shown in Figure 9 (solid line) as a function of target location (averaged across the 25 examples for each location). A similar analysis was performed using the speech stimuli of experiment 2A. For each target position, 25 target segments were randomly generated. For each of these, TMR at the better ear was calculated as described above. This better-ear TMR is also shown in Figure 9 (dashed line) as a function of target location.

These calculations demonstrate that the better-ear TMR is highest for lateral target locations, consistent with the observation that subjects were generally better at identifying a target presented from the more lateral loudspeakers. This advantage is smaller for speech stimuli (~4 dB) than birdsong stimuli (~7 dB), an effect that is likely due to differences in the frequency content of the two kinds of signals. In particular, birdsong contains more high-frequency energy than speech (Best et al. 2005), and thus is more affected by acoustic head shadow effects. Apparently, this difference is not sufficient to give rise to noticeable differences in the dependence of performance on target location in experiments 1 and 2 (compare Figs. 3 and 6). In general, the better-ear TMR at least qualitatively explains the dependence of performance on target location in both experiments.

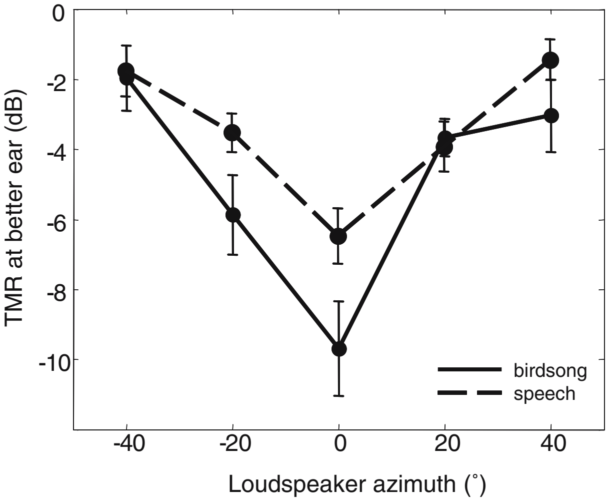

FIG. 9. Target-to-masker ratio (TMR) at the ear closest to the target as a function of target location. Shown are average TMRs for the birdsong stimuli of experiment 1 (solid line) and the speech stimuli of experiment 2A (dashed line). Error bars indicate standard errors of the across-token mean.

As energetic masking is related to the stimulus content and the spatial arrangement of the sources, it can be assumed that its influence is constant across the experimental conditions. In addition to energetic masking, however, target identification in both experiments was also influenced by informational masking. As informational masking is related to uncertainty and an inappropriate allocation of attention, we believe the effect of the visual cues was primarily to modify the influence of informational masking.

Comparison of experiments 1 and 2

In both experiments, the *where* cue provided a robust improvement in target identification performance. In experiment 1, there was also an advantage afforded by the *when* cue, but this advantage was much smaller in experiment 2. The key to this reduced attentional benefit may lie in anecdotal reports that the speech targets "popped out" of the mixture more than the target birdsongs. It seems that the speech maskers used in experiment 2 are less effective maskers for speech targets than unlearned birdsongs are for birdsong targets, giving rise to more *salient* targets in experiment 2.

One contribution to this difference may be differences in target/masker similarity. In experiment 1, the targets and maskers both were composed of very similar birdsong syllables. In experiment 2A, the maskers were time-reversed speech maskers, and there are several reasons why the digit strings may have been relatively distinct in this context. Firstly, reversed speech has a subtly different energy profile from forward speech. Secondly, there is some evidence that the auditory system is preferentially tuned to the spectrotemporal profile of forward speech compared to reversed speech (Asemi et al. 2003). Finally, the maskers were not meaningful in experiment 2A, thus reducing potential confusion from the competitive processing of words (Freyman et al. 2001; Summers and Molis, 2004). These factors suggest that reduced target/masker similarity may have increased the salience of the target in experiment 2A compared to experiment 1. In experiment 2B, these factors were largely eliminated by using forward speech sentence maskers. We assumed that these maskers would be more similar to the targets, reducing the likelihood that the target would emerge automatically from the mixture. However, the results of this experiment were very similar to the results of experiment 2A. The *where* cue provided a benefit but the *when* cue did not (see Fig. 9). It may be that the particular cadence, pitch, and expected linguistic content of the digit string targets was sufficient to make them salient in the midst of sentence maskers. Alternatively, it may be that target/masker dissimilarity is not the only explanation for the salience of the speech targets.

Target salience may also be related to target familiarity. Listeners in experiment 1 who were highly familiar with the birdsong stimuli reported that the target songs seemed to pop out of the mixture for them, and these subjects received less benefit from the *when* than naïve listeners. Highly familiar stimuli (birdsongs for these trained listeners, and speech for all listeners) may emerge from a mixture more automatically than unfamiliar stimuli. In the absence of this pop-out, listeners presumably must attend throughout the entire stimulus and are much more likely to be distracted by the maskers and/or miss the target completely. In such conditions, the *when* cue may enable the listener to increase their vigilance at the appropriate time and successfully ignore irrelevant stimuli.

Different modes of attention

Given that the spatial and temporal components of the visual cues in this study were influenced differently by stimulus context, it is likely that they act via different mechanisms. Furthermore, the fact that the *where* cue and the *when* cue gave benefits that were approximately additive (summing to give the *where and when* cue benefit) suggests that the mechanisms are likely to be independent. Most researchers now agree that attention comprises several subsystems that are functionally and anatomically distinct (Posner and Boies 1971; Raz and Buhle 2006), and in this section we consider different modes of attention that may be relevant to the present study.

Spatial cues and orienting. Orienting describes the allocation of attention to a behaviorally relevant stimulus. In the auditory system, there are several examples that the processing of auditory targets can be enhanced by preceding cues that orient the listener to the correct spatial location (Spence and Driver 1994; Mondor and Zatorre 1995; Buchtel et al. 1996). In situations involving multiple competing stimuli, orienting is often described as a mechanism that enables targets to be "selected" from the mixture. Orienting is influenced by the interaction of exogenous (bottom-up) and endogenous (top-down) factors. In the present study, bottom-up factors likely helped listeners orient in some conditions. In particular, spoken digits stand out in a background of reversed speech maskers (experiment 2A) or in a background of spoken sentences (experiment 2B), while recently learned birdsongs do not stand out readily in a background of other birdsongs. In addition, regardless of how much stimulus-driven orienting occurred, top-down orienting in response to the *where* cue provided a consistent enhancement of performance. This is consistent with the idea that exogenous and endogenous processes act separately, but interact, to bias competition in favor of relevant information (Beck and Kastner 2005).

Our conclusion that the *where* cue aided performance via top-down attentional orienting (rather than simple bimodal interactions, for example) is emphasized in the results of experiment 2, where the *when* cue contributed very little to performance. In this experiment, the visual cue came on synchronously with the target in the *where and when* condition, but not in the *where* condition (where it could come on several seconds earlier). Nonetheless, the gains were essentially the same in these two cases, indicating that it was knowledge of the spatial location that was important, rather than a cross-modal enhancement that relies on temporal synchrony of the visual cue and the auditory target.

One common explanation for the benefit of attending to a spatial location involves the modulation of a neural representation of stimuli at that location. While most physiological support for this modulatory effect has come from the visual system, recent data is providing complementary evidence in the auditory system (Winowski and Knudsen 2006). Shiu and Pashler (1994) provided a useful conceptual model of spatial cueing in the visual system, whereby noncued locations are attenuated and the possibility of mistaking a masker for the target is reduced ("decision noise" is reduced). In the auditory system, it seems that spatial attention plays an equivalent role, reducing uncertainty about which acoustic information is relevant.

Temporal cues and alerting. Very few behavioral or physiological studies have investigated the influence of directing attention to a relevant position in *time*, and as a result this mode of attention is poorly defined in the literature. A useful term that has been used is "phasic alerting", which describes a task-specific increase in vigilance or arousal (Parasuraman et al. 1998; Raz and Buhle 2006). Alerting may occur in response to a stimulus or as a result of expectation during rhythmic stimuli (Jones et al. 2002).

The current experiments support the idea that directing listeners' attention to the correct point in time is quite different than directing their attention to the correct spatial location. Compared to knowing *where* to listen, the benefit of knowing *when* to listen depends on stimulus parameters and subject experience with the stimuli. In general, the *when* cue provided the largest benefits when target and masker were most similar and listeners were unfamiliar with the targets. This suggests that temporal attention acts to reduce the likelihood of a listener mistakenly attending to a time segment containing only maskers.

The role of visual cues in speech perception

It is well documented that visual speech information obtained via lip reading can supplement speech intelligibility in noise (Sumby and Pollack 1954), and can even override acoustic cues in some cases of audiovisual conflict (McGurk and MacDonald 1976). Furthermore, correlations between acoustic events and a speaker's facial movements appear to help listeners segregate the message from a particular talker out of a multitalker sound mixture (Helfer and Freyman 2005). This shows that visual speech cues provide information beyond explicit cues to speech content. However, previous results do not tease apart whether visual cues for target *location* improve performance by allowing a listener to focus spatial attention on the target. Experiment 2 shows that spatial knowledge from a simple visual cue can indeed enhance speech intelligibility. Furthermore, the results of experiment 1 suggest that gross temporal information indicating when a target of interest will occur can be of use when a target is not salient enough to emerge automatically from a mixture. This may be relevant for speech situations that are more challenging than that simulated in experiment 2. For instance, temporal cues may benefit hearing-impaired listeners who are trying to understand speech in a complex environment (where salient target features are degraded), or nonnative speakers trying to converse in a new language (where target familiarity is low).

ACKNOWLEDGMENTS

This work was supported by grants from ONR (N00014-04-1-0131) and NIH/NIDCD (DC04663).

REFERENCES

Arbogast TL, Kidd G Jr. Evidence for spatial tuning in informational masking using the probe-signal method. J. Acoust. Soc. Am. 108:1803–1810, 2000.

Asemi N, Sugita Y, Suzuki Y. Auditory search asymmetry between normal Japanese speech sounds and time-reversed speech sounds distributed on the frontal-horizontal plane. Acoust. Sci. Technol. 24:145–147, 2003.

Beck DM, Kastner S. Stimulus context modulates competition in human extrastriate cortex. Nat. Neurosci. 8:1110–1116, 2005.

Best V, Ozmeral E, Gallun FJ, Sen K, Shinn-Cunningham BG. Spatial unmasking of birdsong in human listeners: energetic and informational factors. J. Acoust. Soc. Am. 118:3766–3773, 2005.

Brungart DS, Simpson BD. Cocktail party listening in a dynamic multitalker environment. Percept. Psychophys., 2007, in press.

Buchtel HA, Butter CM, Ayvasik B. Effects of stimulus source and intensity on covert orientation to auditory stimuli. Neuropsychology 34:979–985, 1996.

Carhart R, Tillman TW, Greetis ES. Perceptual masking in multiple sound backgrounds. J. Acoust. Soc. Am. 45:694–703, 1969.

Durlach NI, Mason CR, Gallun FJ, Shinn-Cunningham BG, Colburn HS, Kidd G Jr. Psychometric functions for fixed and randomly mixed maskers. J. Acoust. Soc. Am. 118:2482–2497, 2005.

Eramudugolla R, Irvine DRF, McAnally KI, Martin RL, Mattingley JB. Directed attention eliminates "change deafness" in complex auditory scenes. Curr. Biol. 15:1108–1113, 2005.

Ericson MA, Brungart DS, Simpson BD. Factors that influence intelligibility in multitalker speech displays. Int. J. Aviation Psychol. 14:311–332, 2004.

Freyman RL, Balakrishnan U, Helfer KS. Spatial release from informational masking in speech recognition. J. Acoust. Soc. Am. 109:2112–2122, 2001.

Green DM, Weber DL. Detection of temporally uncertain signals. J. Acoust. Soc. Am. 67:1304–1311, 1980.

Helfer KS, Freyman RL. The role of visual speech cues in reducing energetic and informational masking. J. Acoust. Soc. Am. 117:842–849, 2005.

Jones MR, Moynihan H, MacKenzie N, Puente J. Temporal aspects of stimulus-driven attending in dynamic arrays. Psychol. Sci. 13:313–319, 2002.

Kidd G Jr, Arbogast TL, Mason CR, Gallun FJ. The advantage of knowing where to listen. J. Acoust. Soc. Am. 118:3804–3815, 2005.

Kidd G Jr, Mason CR, Brughera A, Hartmann WM. The role of reverberation in release from masking due to spatial separation of sources for speech identification. Acustica united with Acta Acustica 114:526–536, 2005.

Lufti RA, Kistler DJ, Callahan MR, Wightman FL. Psychometric functions for informational masking. J. Acoust. Soc. Am. 114:3273–3282, 2003.

McGurk H, MacDonald J. Hearing lips and seeing voices. Nature 264:746–748, 1976.

Mondor TA, Zatorre RJ. Shifting and focusing auditory spatial attention. J. Exp. Psychol. Hum. Percept. Perform. 21:387–409, 1995.

Parasuraman R, Warm JS, See JE. Brain systems of vigilance. In: Parasuraman R (ed) The Attentive Brain. Cambridge, Massachusetts, MIT Press, 1998.

Pollack I. Auditory informational masking. J. Acoust. Soc. Am. 57:S5, 1975.

Posner MI, Boies SJ. Components of attention. Psychol. Rev. 78:391–408, 1971.

Raz A, Buhle J. Typologies of attentional networks. Nat. Rev., Neurosci. 7:367–379, 2006.

Richards VM, Neff DL. Cuing effects for informational masking. J. Acoust. Soc. Am. 115:289–300, 2004.

SHEN J, REINGOLD EM. Visual search asymmetry: the influence of stimulus familiarity and low-level features. Percept. Psychophys. 63:464–475, 2001.

SHIU L, PASHLER H. Negligible effects of spatial precuing on identification of single digits. J. Exp. Psychol. Hum. Percept. Perform. 20:1037–1054, 1994.

SPENCE CJ, DRIVER J. Covert spatial orienting in audition: exogenous and endogenous mechanisms. J. Exp. Psychol. Hum. Percept. Perform. 20:555–574, 1994.

SUMBY WH, POLLACK I. Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. 26:212–215, 1954.

SUMMERS V, MOLIS MR. Speech recognition in fluctuating and continuous maskers: effects of hearing loss and presentation level. J. Speech Lang. Hear. Res. 47:245–256, 2004.

VECERA SP, FARAH MJ. Is visual image segmentation a bottom-up or an interactive process? Percept. Psychophys. 59:1280–1296, 1997.

WANG Q, CAVANAGH P, GREEN M. Familiarity and pop-out in visual search. Percept. Psychophys. 56:495–500, 1994.

WATSON CS. Uncertainty, informational masking and the capacity of immediate auditory memory. In: Yost WA and Watson CS (eds) Auditory Processing of Complex Sounds. Hillsdale, NJ, Lawrence Erlbaum, 1987.

WINOWSKI DE, KNUDSEN EI. Top-down gain control of the auditory space map by gaze control circuitry in the barn owl. Nature 439:336–339, 2006.

WRIGHT BA, FITZGERALD MB. The time course of auditory attention in a simple auditory detection task. Percept. Psychophys. 66:508–516, 2004.

WRIGHT BA, SABERI K. Strategies used to detect auditory signals in small sets of random maskers. J. Acoust. Soc. Am. 105:1765–1775, 1999.

ZEMEL RS, BEHRMANN M, MOZER MC, BAVELIER D. Experience-dependent perceptual grouping and object-based attention. J. Exp. Psychol. Hum. Percept. Perform. 28:202–217, 2002.

ZUREK PM. Binaural advantages and directional effects in speech intelligibility. In: Studebaker GA and Hochberg I (eds) Acoustical Factors Affecting Hearing Aid Performance. Boston, Allyn and Bacon, 1993.