How Visual Cues for when to Listen Aid Selective Auditory Attention

Lenny A. Varghese, Erol J. Ozmeral, Virginia Best & Barbara G. Shinn-Cunningham

Journal of the Association for Research in Otolaryngology

ISSN 1525-3961

JARO
DOI 10.1007/s10162-012-0314-7

JARO (2012)
DOI: 10.1007/s10162-012-0314-7
© 2012 Association for Research in Otolaryngology

JARO
Journal of the Association for Research in Otolaryngology

How Visual Cues for when to Listen Aid Selective Auditory Attention

Lenny A. Varghese,[1] Erol J. Ozmeral,[3] Virginia Best,[2] and Barbara G. Shinn-Cunningham[1,3]

[1]*Department of Biomedical Engineering, Boston University, Boston, MA 02215, USA*

[2]*Department of Speech, Language, and Hearing Sciences, Boston University, Boston, MA 02215, USA*

[3]*Center for Computational Neuroscience and Neural Technology, Boston University, 677 Beacon Street, Boston, MA 02215, USA*

Received: 12 August 2011; Accepted: 10 January 2012

ABSTRACT

Visual cues are known to aid auditory processing when they provide direct information about signal content, as in lip reading. However, some studies hint that visual cues also aid auditory perception by guiding attention to the target in a mixture of similar sounds. The current study directly tests this idea for complex, nonspeech auditory signals, using a visual cue providing only timing information about the target. Listeners were asked to identify a target zebra finch bird song played at a random time within a longer, competing masker. Two different maskers were used: noise and a chorus of competing bird songs. On half of all trials, a visual cue indicated the timing of the target within the masker. For the noise masker, the visual cue did not affect performance when target and masker were from the same location, but improved performance when target and masker were in different locations. In contrast, for the chorus masker, visual cues improved performance only when target and masker were perceived as coming from the same direction. These results suggest that simple visual cues for when to listen improve target identification by enhancing sounds near the threshold of audibility when the target is energetically masked and by enhancing segregation when it is difficult to direct selective attention to the target. Visual cues help little when target and masker already differ in attributes that enable listeners to engage selective auditory attention effectively, including differences in spectrotemporal structure and in perceived location.

Keywords: attention, cross-modal enhancement, energetic masking, informational masking, spatial separation, visual cues

INTRODUCTION

Natural sounds rarely occur without competing or distracting sounds. Distractors may interfere with the perception of a target by causing energetic masking (when the spectral content of target and distractor overlap, yielding portions of the target inaudible) and/or informational masking (where the perceptual interference cannot be explained by energetic overlap; see Kidd et al. 2008). Informational masking occurs when a target and masker are similar in their spectrotemporal structure, which can make it difficult to segregate the competing sources into distinct perceptual objects and/or to identify which object is the target, interfering with selective attention (e.g., Shinn-Cunningham 2008; Shinn-Cunningham and Best 2008).

Spatially separating a target from distractors can help improve the ability to detect and identify the target sound (leading to improved intelligibility, in the case of a speech target; Hirsh 1950; Carhart et al. 1969; Drullman and Bronkhorst 2000; Brungart and Simpson 2002; Marrone et al. 2008). When energetic masking determines performance, the benefit of spatial separation depends on the spectral content of the competing sounds and comes about from both improvements in the

Correspondence to: Barbara G. Shinn-Cunningham • Center for Computational Neuroscience and Neural Technology • Boston University • 677 Beacon Street, Boston, MA 02215, USA. Telephone: +1-617-3535764; fax: +1-617-3537755; email: shinn@cns.bu.edu

Published online: 11 February 2012

target-to-masker ratio reaching the "better ear" at frequencies above about 2 kHz and binaural processing benefits, strongest for frequencies below about 1000 Hz (Zurek 1993). When failures of selective attention limit performance, spatial separation can help both by improving segregation of the target from competing maskers and enabling listeners to direct selective attention to a particular location (e.g., see Kidd et al. 1998; Freyman et al. 1999; Arbogast et al. 2002).

Temporal uncertainty about when a target sound will occur in the presence of ongoing distractors can also interfere with performance, but the influence of temporal uncertainty depends on the experimental conditions. For simple tone-in-noise detection, temporal uncertainty does not appear to reduce sensitivity (Egan et al. 1961; Green and Weber 1980). However, temporal uncertainty seems to have a larger effect on performance when target and masker are more similar, such as when a masker is tonal rather than noise (Bonino and Leibold 2008). Moreover, both speech detection (e.g., Grant and Seitz 2000; Bernstein et al. 2004; Tye-Murray et al. 2011) and speech intelligibility (e.g., Sumby and Pollack 1954; Helfer and Freyman 2005) are enhanced when a visual stimulus depicting appropriate lip movements accompanies speech masked by noise. Bernstein et al. (2004) noted that simpler visual cues, such as the presence of an abstract shape whose size follows the broadband envelope of the target speech or even a static shape whose appearance coincides with the target presentation, can also improve speech-in-noise detection thresholds. These results suggest that visual cues can provide cross-modal enhancement of auditory perception in some energetic masking conditions (Stein and Meredith 1993).

The use of visual cues also helps in more complex listening situations such as when speech is masked by speech (Helfer and Freyman 2005; Best et al. 2007; Gatehouse and Akeroyd 2008) or when birdsongs are masked by other birdsongs (Best et al. 2007). Such benefits can arise not only when listeners are provided with speechreading cues containing potentially useful speech information (Helfer and Freyman 2005) but also when the visual information consists of simple onset/offset information from LED lights (e.g., Best et al. 2007; Gatehouse and Akeroyd 2008). These results suggest that reductions in temporal uncertainty can improve source segregation and target selection in mixtures of similar sounds.

Although both spatial separation and visual cues can provide large perceptual benefits in masked listening situations, it is possible that they both improve performance in similar ways (helping listeners focus selective attention on the target), making them redundant with one another. If this is the case, it may be that each has the greatest effect in the absence of the other. For example, the largest benefit of visual cues indicating "when to listen" may arise when there are no spatial cues indicating "where to listen" or when the target is not easy to segregate from the background. Consistent with this idea, visual cues (specifically, speechreading cues) interact with spatial cues: while speechreading cues generally provide a benefit regardless of spatial configuration or the dominant form of masking present (energetic or informational), they are most beneficial when target and masker are highly confusable (i.e., similar in content and originating from the same place; see Helfer and Freyman 2005).

Here, we hypothesized that the perceptual benefits of simple visual cues providing information only about the timing of complex, nonspeech auditory targets would be reduced when the target was spatially separated from its competitors or when it was easily differentiated from the background on the basis of its spectrotemporal properties. Listeners heard mixtures consisting of one of five learned targets (zebra finch songs) embedded at a random time within a longer masker. The target was presented either with or without a visual cue that marked its onset and offset. To determine whether the benefit of the visual cue depended on the similarity of the target and masker, we used both steady-state noise maskers and "chorus" maskers made up of other birdsongs. To determine whether the visual cue interacted with spatial cues, we compared performance for various spatial configurations of target and masker. As a control for better-ear effects, we measured all conditions both binaurally [where listeners heard a fully spatialized version of the stimuli as simulated with head-related transfer functions (HRTFs)] and diotically (with both ears receiving an identical, acoustically better-ear signal).

METHODS

Subjects

A total of 11 listeners (four male and seven female, ages 18 to 24) were recruited and paid for their participation in the experiment. All subjects signed an informed consent document approved by the Boston University Charles River Campus Institutional Review Board. The listeners were screened to ensure that they had normal hearing (less than 20 dB hearing loss) for frequencies between 250 Hz and 8 kHz.

Stimuli

Songs from 14 different zebra finches (*Taeniopygia guttata*) were used in this experiment. Songs from five of the birds were used as targets; songs from the remaining nine birds were used to construct maskers.

Songs were recorded as described in Best et al. (2005). All were low-pass filtered at 8 kHz and resampled to 50 kHz in Matlab (Mathworks, Natick, MA) prior to use in this experiment.

For each of the 14 birds, five similar motifs (repeating elements of a longer song) were selected from the bird's repertoire. A zebra finch's motif generally consists of a particular pattern of syllables repeated in a fixed order with nearly identical rhythm, but with a slight jitter in the pitch, loudness, and/or timing of the syllables. These motifs are highly stereotypical for a particular bird but quite distinct from those of the other birds. The chosen motifs varied across birds in the exact syllables making up the motif as well as the number and rhythm of the syllables. The motifs corresponding to the chosen birds were roughly 750–1000 ms in length (Best et al. 2005). Listeners were trained to identify the five motifs from each "target" bird using a specific label (arbitrary names assigned to the bird: "Uno," "Junior," "Nibbles," "Moe," and "Toro").

Two types of masker were used: chorus maskers and song-shaped noise maskers. Both had the same long-term spectral characteristics, but had different short-term statistics. Chorus maskers were constructed such that they were highly confusable with the targets and created temporal uncertainty about when the target occurred. First, 30 motifs were chosen randomly from the set of all possible masker motifs such that each nontarget bird was represented by at least three but not more than four distinct motifs in the masker mixture. The 30 selected songs were each delayed by a random start time and were then summed to produce a sound mixture containing a variable number of songs as a function of time. This mixture (about 7.2 s in duration at this point) was then cropped down to 5.1 s. The process was repeated until 12 "good" maskers were generated with no less than one song present at any point during the masker (Fig. 1). The selected 12 maskers had an average of four to five songs present at any given time. A set of 12 independent noise maskers was created by generating broadband noise that had a spectral profile matching that of the average of the final set of 12 chorus maskers.

Experimental conditions

A total of 16 conditions were tested, made up of all combinations of four factors. (1) The masker was either a chorus masker or song-shaped noise. (2) The stimuli were either colocated (0° separation between target and masker) or separated (90° separation). (3) The stimuli were presented with either realistic spatial cues ("binaural" presentation) or with the acoustically better-ear signal presented to both ears ("diotic" presentation). (4) The acoustic stimuli were presented either alone or in conjunction with a visual cue that was temporally aligned with the target stimulus.

Spatial conditions. Spatial cues were added to the stimuli using pseudo-anechoic HRTFs to simulate different configurations of the target and masker. The HRTFs were measured on a KEMAR manikin at a distance of 1 m in the horizontal plane level with the ears (0° elevation), as described in Shinn-Cunningham et al. (2005). Targets were always simulated from directly in front of the listener (processed with HRTFs corresponding to 0° azimuth). In colocated spatial conditions, the masker was processed using the same HRTF as the target. In the separated conditions, the masker was processed using an HRTF corresponding to 90° azimuth, simulating a masker to the far right of the listener.

Binaural and diotic presentations. When the target and masker were spatially separated (as described above), acoustic shadowing by the head caused the left ear signal to have a better target-to-masker ratio (TMR) than the right ear signal, i.e., the left ear was the acoustically better ear. Diotic stimuli were generated by playing to both ears the left-ear signal from the fully spatialized, binaural stimulus for a given condition. This produced sound mixtures in which the target and masker were both perceived as coming from the auditory midline, but in which the signals at both ears had the TMR of the acoustically better ear for the given spatial configuration. In contrast, in the binaural presentations, interaural differences between the acoustically better, left-ear signal and the right-ear signal (which either had the same or a worse TMR than the left-ear signal) yielded stimuli in which the target and masker either were perceived from distinct locations (in the spatially separated spatial configuration) or were both perceived at the auditory midline (in the colocated configuration).

Visual cues. In half of the experimental sessions, a visual cue provided information about the onset and offset of the target. This visual cue consisted of a large black rectangle displayed on the computer screen in front of the listener, which appeared at the onset of the target in a given trial and stayed on for the duration of the target.

Experimental procedures

Stimulus presentation. Stimulus presentation was handled using Matlab software interfacing with Tucker-Davis Technology (Alachua, FL) hardware operating at a sampling rate of 48.8 kHz. Immediately prior to stimulus presentation, one target and one masker were selected at random, scaled relative to one another to one of seven TMRs (evenly spaced between -36 and +6 dB), and combined according to condition. The TMR was calculated using the broadband RMS levels of the two selected signals prior to spatial processing. The selected TMRs spanned the sloping portion of the psychometric

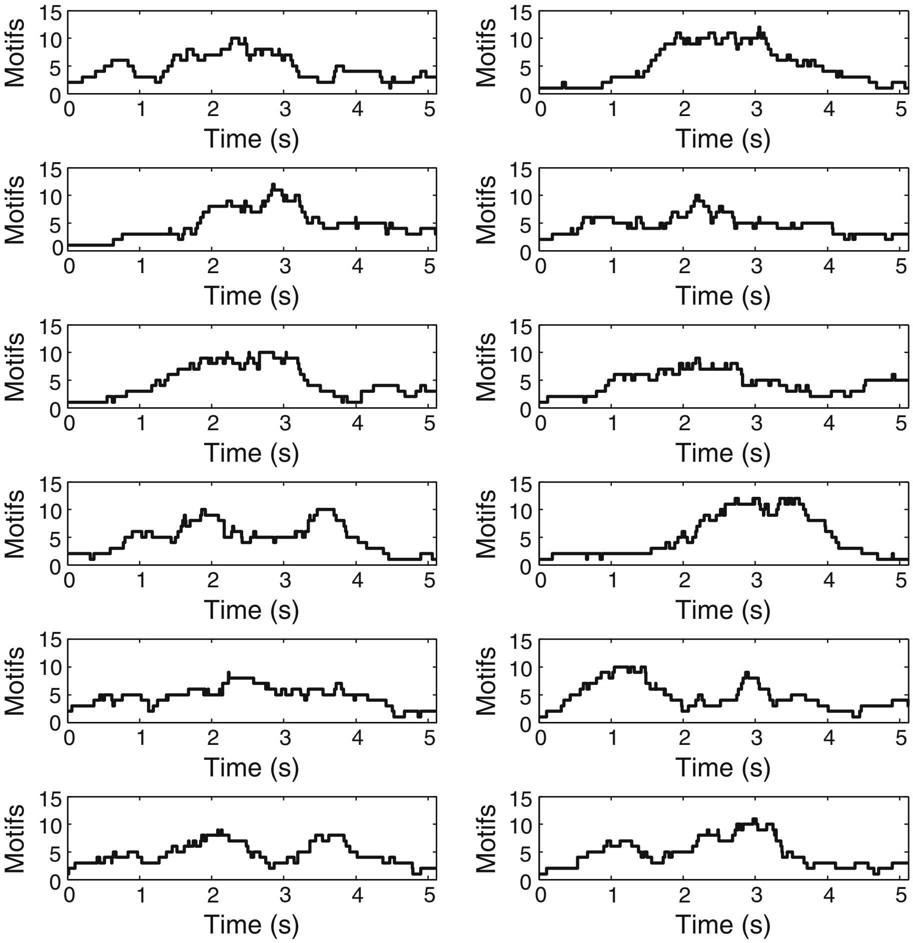

FIG. 1. Chorus masker composition for each of the 12 maskers used in this experiment. Shown are the number of birdsong motifs present at any given time over the duration of the masker. The set of 12 noise maskers was created by generating broadband noise that matched the average spectral profile of these 12 maskers. The target was embedded at a random time in the masker, but in such a way that there was always a buffer of masker sound prior to and after the target (i.e., the target and masker onsets/offsets never coincided).

functions relating identification performance to TMR (based on pilot testing data). Each subject set the overall presentation level of the stimuli to a comfortable level at the start of each session. For a given trial, the target began playback at a uniformly distributed random time between 0.51 and 3.6 s after the start of the masker, so that there was temporal uncertainty as to when the target occurred. These minimum and maximum delay values were chosen such that there was at least a 0.51 s buffer between the start of the target and the start of the masker, as well as between the end of the target and the end of the masker. Finally, the stimuli were passed to the TDT hardware for D/A conversion and attenuation prior to presentation over Sennheiser HD-280 Pro headphones.

Experimental interface. Subjects were seated in a sound-treated booth (Industrial Acoustics, Bronx, NY). Three Matlab-coded graphical user interfaces were used in this experiment, each of which provided different functionality for different parts of the experiment. The "familiarization" interface allowed users to click a button and hear the song corresponding to that target bird. The "testing" interface had a start button that, when clicked, initiated a test in which a random target birdsong was played diotically, in quiet. Listeners clicked one of the five name buttons to identify the song. Correct-answer feedback was provided, and the next song played after a short pause. Finally, the "experimental" interface presented target birdsongs with a masker. While the stimulus was playing, the buttons on the screen were hidden from view. If a visual cue (black rectangle) was required, it appeared in the middle of the interface at the proper time during the acoustic stimulus. The response buttons reappeared after stimulus playback, and users were required to identify the heard song before the next stimulus was presented. No feedback was provided.

Training and testing procedures. Subjects performed seven sessions on seven different days, the first of which was a training session that was not analyzed. In the first, third, and fifth experimental sessions, all trials were presented without a visual cue, while in all other sessions, all trials were presented with a visual cue (post hoc analysis showed no learning effects across the analyzed experimental sessions). Subjects were restricted to performing one session per day, and completed all sessions over the course of 2–4 weeks.

In the initial training session, subjects were exposed to the target bird songs repeatedly until they could reliably identify each characteristic song. In all cases, songs were presented without spatial processing and without a masker. Subjects were initially presented with the familiarization interface and were allowed to play the birdsongs as many times as they needed to become comfortable with the songs. When they felt ready to proceed, subjects closed the familiarization interface and were then presented with the testing interface for a 100-trial preliminary testing run. The 100 trials corresponded to the 25 motifs (five birds, five motifs each), each presented four times in random order (chosen without replacement). After completion of the 100-trial preliminary testing run, 25-trial test runs were administered, with each motif appearing once without replacement, until identification performance reached 96%. With this training paradigm, subjects completed a minimum total of 125 identification trials in quiet before starting any experimental sessions. Training was generally completed within approximately 30 min.

All subsequent sessions after the training sessions were experimental sessions whose results were analyzed. At the start of each experimental session, subjects repeated familiarization and 25-trial testing runs to ensure that they could still identify the birdsongs with 96% accuracy at the start of that session. Upon achieving this criterion in the initial testing, the experimental interface was presented to the subjects. For the rest of the experimental session, subjects alternated between performing an experimental run (analyzed to determine how listeners performed for targets presented in different conditions) and a short test run of ten trials each (presented to ensure that listeners were still able to correctly identify target birds in quiet). Subjects had to perform 90% of trials correctly in the interspersed short test runs to be able to advance to the next experimental run; these test runs were repeated until subjects were able to meet the 90% criterion.

The eight experimental runs in a given session were blocked by the eight different conditions comprising the session (all combinations of noise/chorus, colocated/separated configurations, diotic/binaural presentations), presented in a different random order for each subject and each session. There were 35 trials per experimental run (5 birds×7 TMRs per bird; the target motif used for a given bird was randomly selected in each trial). Subjects were typically able to complete all runs in an experimental session in about 1 h.

Data analysis

For each subject, data were averaged over identical TMRs for each unique experimental condition. Psychometric curves were fit to the data for each subject using the psignifit package for Matlab (http://bootstrap-software.org/psignifit/). For each such curve, the threshold was defined as the TMR corresponding to a 60% correct score on the psychometric curve. Within-subject differences in threshold were computed to quantify the effect of the visual cues (threshold with visual cues subtracted from thresholds without visual cues). For each condition, a right-tailed *t*-test was used to test the hypothesis that the unmasking due to the visual cue was greater than 0 dB.

RESULTS

Noise masker

In all conditions, performance improved from chance levels at low TMRs to near perfect at high TMRs (see psychometric functions in Fig. 2A and B, which show mean performance across subjects; note that the variability across subjects was modest, so that these plots are very similar in slope and threshold to the results for any given individual subject). Spatially separating the target and masker generally improved performance both when there was no visual cue and with a visual cue present (in Fig. 2A and B, the red lines are shifted to the left of the corresponding blue lines). In general, when the target and masker were colocated, there was no difference between diotic and binaural presentations, as might be expected (in both Fig. 2A and B, the solid and dashed blue lines fall on top of one another). When there was no visual cue, performance was similar for binaural and diotic presentations in the separated configuration (in Fig. 2A, the solid and dashed red lines fall on top of one another). In contrast, when a visual cue was present, binaural presentation led to slightly better performance than did diotic presentation (in Fig. 2B, the solid red line is shifted leftward compared to the dashed red line). This small improvement of binaural over monaural presentation is likely related to the "binaural masking level difference," in which binaural differences in the target and masker can interact, enabling listeners to detect elements of the target that would be inaudible when the signals are presented diotically (e.g., see Zurek 1993). Figure 2 shows that performance was slightly better when the visual cue was present than when there was no visual cue (results in Fig. 2B are shifted leftward compared to the corresponding results in Fig. 2A).

Threshold values, found by averaging the thresholds of the psychometric function fits to the results for the individual subjects, are shown in Figure 2C to enable a more direct comparison of conditions. Consistent with the observations above, spatial separation of target and masker improved target discrim-

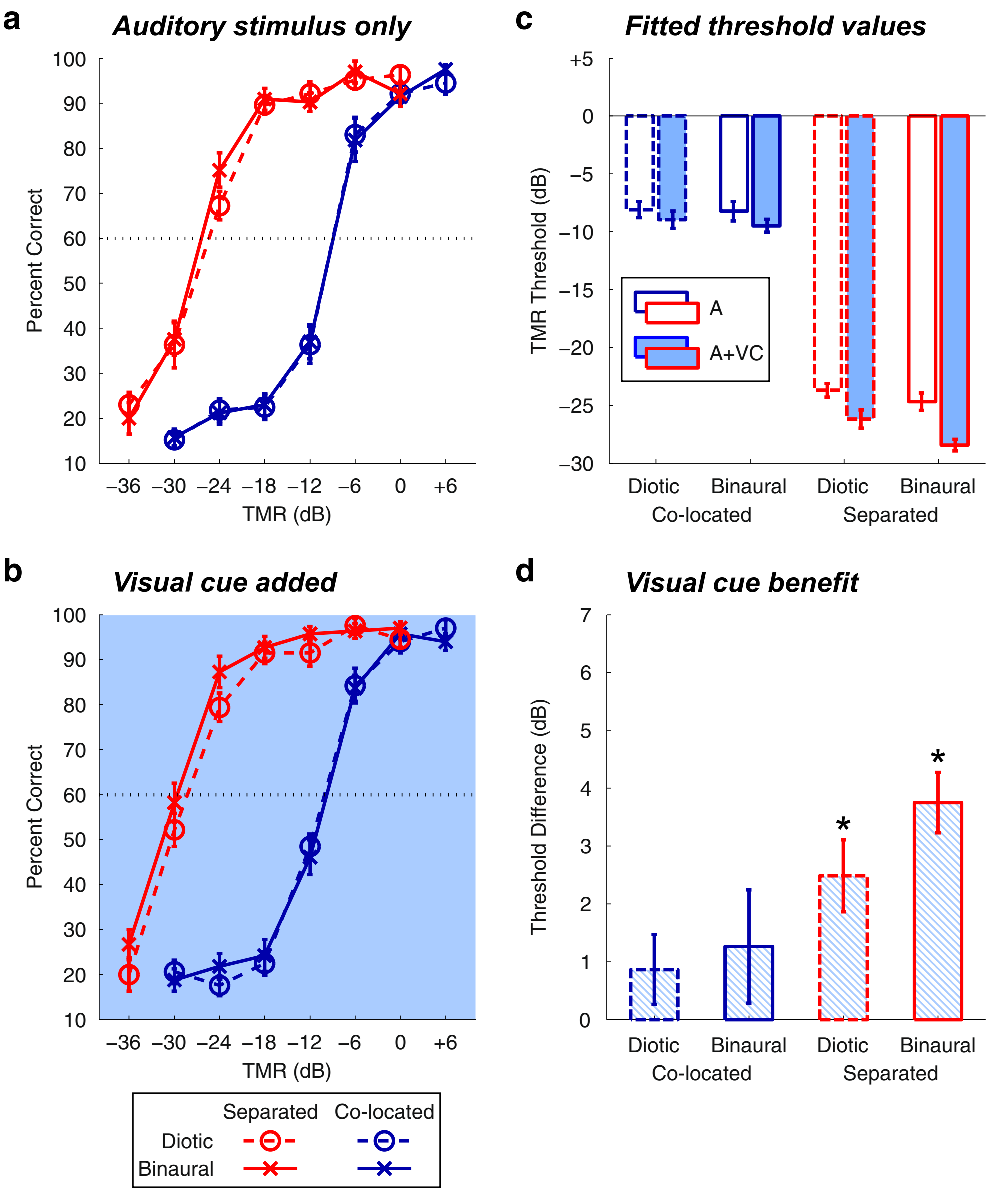

FIG. 2. Subject performance in noise-masked conditions. All values shown are mean±SEM. **A** Raw performance (percent correct vs. TMR) when the auditory stimulus was presented alone. **B** Raw performance (percent correct vs. TMR) when the auditory stimulus was presented in conjunction with a visual cue. **C** 60% threshold values derived from psychometric fits for auditory stimuli only (*A*, *open bars*), and auditory stimulus plus visual cue (*A+VC*, *filled bars*) for the various spatial configurations. **D** Visual cue benefit, calculated by subtracting thresholds for the auditory stimulus alone from the thresholds for the auditory stimulus plus visual cue (i.e., *A+VC-A*). A *star* denotes that the mean is significantly greater than 0 (p<0.05), as determined by a right-tailed t-test.

ination performance (the four bars to the right are lower than the four bars to the left). Binaural presentation had no effect on thresholds for colocated sources, whether or not there are visual cues present (the two leftmost pairs of bars are similar); similarly, presenting stimuli binaurally did not affect performance for separated target and masker when there were no visual cues (in the two rightmost pairs of bars, the white bars are similar). In contrast, when visual cues were present and the sources were spatially separated, performance was slightly better when listeners heard a full, spatial presentation than when they heard the better-ear signal in both ears (the rightmost filled bar is lower than the filled bar second from the right). Finally, adding visual cues tended to improve performance in all conditions (in each pair of bars, the filled bar is lower than the white bar).

To directly assess performance benefits due to the visual cue, the thresholds with the visual cue present were subtracted from the corresponding threshold when no visual cue was present for each subject. The across-subject means of these values are shown in

Figure 2D. Although the means of all these differences were positive, the benefit of the visual cue did not reach significance when target and masker were colocated, regardless of whether presentations were binaural or diotic (Fig. 2D, leftmost two bars). However, the visual cue significantly improved performance when the sources were spatially separated for both the binaural and diotic presentations (Fig. 2D, rightmost two bars; $p<0.05$).

Chorus masker

When a chorus masker was present, performance in all conditions improved smoothly with TMR, just as with the noise masker; moreover, just as with the noise masker, there was a large improvement in performance due to spatial separation of the target and masker (in Fig. 3A and B, the red lines are shifted to the left of the corresponding blue lines). As with the noise masker, when the target and chorus masker were colocated, there was no difference between diotic and binaural presentations (in Fig. 3A and B, the solid and dashed blue lines fall on top of one another). However, performance was affected differently by presentation style (binaural vs. diotic) and the presence of visual cues for the chorus masker compared to the noise masker. Both with and without a visual cue, binaural presentation led to a large improvement in performance over diotic presentation of the target and a competing chorus masker (in Fig. 3A and B, the solid red lines fall to the left of the corresponding dashed red lines). Figure 3A and B shows that performance was generally better when the visual cue was present than when there was no visual cue (results in Fig. 3B are shifted leftward compared to the corresponding results in Fig. 3A). This shift was smallest when spatially separated stimuli were presented binaurally (compare solid red lines in Fig. 3A and B).

Direct comparison of the thresholds derived from the psychometric functions (Fig. 3C) confirms that spatial separation of target and masker improved target discrimination performance (the four bars to the right are lower than the four bars to the left). Binaural presentation improved performance over diotic presentation for the separated conditions both with and without a visual cue present (the rightmost white and filled bars are lower than the corresponding bars in the second pair from the right). However, presenting stimuli binaurally had no effect when the chorus masker was colocated with the target (the two leftmost white bars are similar, and the two leftmost filled bars are similar). When the visual cue was present, thresholds were generally lower than when there was no visual cue (the filled bars are generally lower than the corresponding white bars).

The mean across-subject improvement in performance due to the presence of the visual cues is shown in Figure 3D. Visual cues significantly improved performance in three of the four conditions (the three leftmost bars are all greater than zero; $p<0.05$); only when the target and competing masker were spatially separated and presented binaurally, so that the target was perceived as coming from a different direction than the competing chorus masker, was the effect of the visual cue insignificant.

SUMMARY AND DISCUSSION

This study examined the ability of listeners to identify target stimuli in the presence of an ongoing background masker and explored how a simple visual cue informing the listener of when to listen influences identification. Given past work showing differences in the kinds of perceptual interference that can arise with different kinds of maskers, we compared the ability to identify a birdsong in the presence of a steady-state noise masker (which primarily causes energetic masking) and a chorus masker (which is thought to primarily disrupt sound segregation and selective attention to the target). We found that the introduction of a visual cue supplying timing information about the target aided listeners when the masker was noise and the target and masker were separated in space, regardless of whether the target and masker were perceived as separated in space (binaural) or when there was a diotic presentation of the better-ear signal to both ears. In contrast, when the masker was a chorus of birdsongs, the visual cue aided listeners only when the target and masker were perceived as originating from the same location (i.e., in the colocated conditions and in the separated, diotically presented condition), and not when the target and masker were perceived as originating from two separate spatial locations.

Spatial separation provides different benefits for different kinds of maskers

For both types of maskers and in all listening conditions, spatially separating the target from the masker improved performance. A large part of this benefit comes about not from the neural processing of spatial cues, per se, but rather the improvement in TMR at the acoustically better ear. This can be seen directly by considering the diotic stimulus results. For both a steady-state noise masker and a complex masker and for both presentations with visual cues and without visual cues, performance for diotic

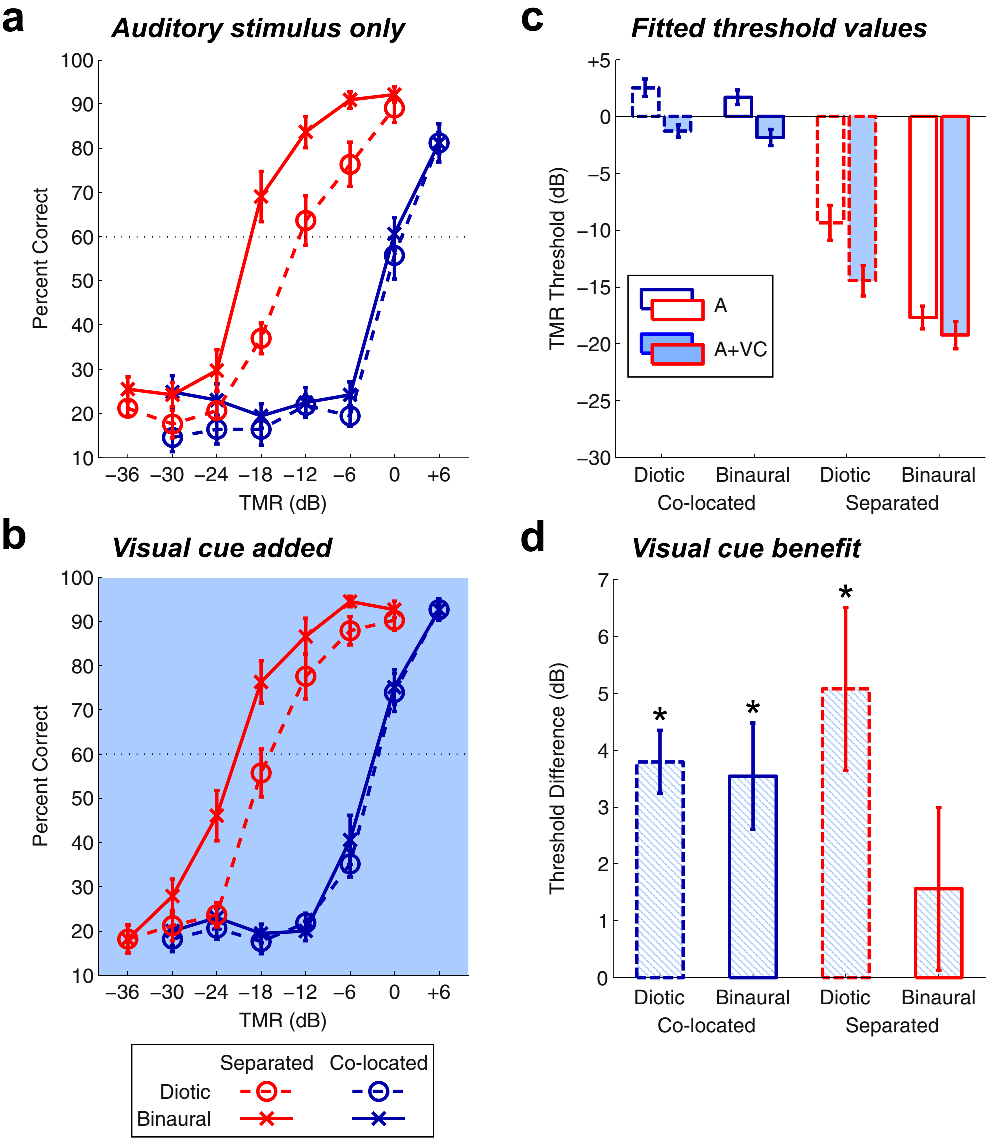

FIG. 3. Subject performance in chorus-masked conditions. All values shown are mean±SEM. **A** Raw performance (percent correct vs. TMR) when the auditory stimulus was presented alone. **B** Raw performance (percent correct vs. TMR) when the auditory stimulus was presented in conjunction with a visual cue. **C** 60% threshold values derived from psychometric fits for auditory stimuli only (*A, open bars*), and auditory stimulus plus visual cue (*A+VC, filled bars*) for the various spatial configurations. **D** Visual cue benefit, calculated by subtracting thresholds for the auditory stimulus alone from the thresholds for the auditory stimulus plus visual cue (i.e., *A+VC-A*). A *star* denotes that the mean is significantly greater than 0 ($p<0.05$), as determined by a right-tailed *t*-test.

stimuli was much better when sources were spatially separated than when they were colocated. Given that in the diotic presentations the target and masker both sound as if they are at midline, this large improvement can be attributed to purely acoustic effects on the TMR at the acoustically better ear; there are no binaural or spatial perceptual effects in diotic presentations (e.g., see Best et al. 2005).

When target and masker were colocated, diotic and binaural presentations yielded similar performance, as expected. When sources were spatially separated, providing binaural cues that allowed target and masker to be perceived at different locations, performance improved beyond the acoustic better-ear improvements seen for diotic presentation. We found that improvements due to binaural presentation depended on the kind of masker. When the masker was steady-state noise and energetic masking was the primary form of interference, the benefit of binaural presentation was very small, on the order of 1 dB at threshold. In contrast, when the target song was played along with a chorus, the improvements obtained by presenting the sources binaurally rather than diotically were larger, on the order of 4–8 dB. Statistical tests confirm that for spatially separated target and masker in absence of a visual cue, the benefit of binaural presentation over diotic presentation is larger for the chorus masker than for the noise masker (paired *t*-test, $p<0.05$). These results replicate those of a previous study that also directly compared the contributions of better-ear acoustics and spatial perception to spatial unmasking for different maskers (Best et al. 2005). In that study, the target and masker turned on and off simultaneously, which both removes temporal uncertainty about when to listen and might promote grouping target and masker into one auditory object. However, for a steady-state noise masker, the target tends to be easily discriminable from the rest of the sound mixture, no matter what the temporal structure of the stimuli. Perceptually, the dissimilarity between a complex target and a noise masker reduces temporal uncertainty about when the target occurs (in the current study) and counteracts the tendency for target and masker to be heard as a single object (in the previous study). Thus, it makes sense that the effects of spatial unmasking in the presence of a noise masker are similar across the two studies. For the chorus masker, segregating and attending to the target song should be difficult whether the onsets and offsets of target and masker are simultaneous (as in the previous study) or the target comes on at a random, unpredictable time (as in the current study). In both cases, spatial perception is likely to help listeners selectively attend to the target, above any benefits of the acoustically better ear. Again, it makes sense that spatial unmasking plays out similarly here and in this previous study. Our results are also consistent with past studies showing that spatial unmasking is greater in the presence of a masker that is similar to the target than in the presence of an energetic masker (e.g., see Kidd et al. 1998; Freyman et al. 1999; Arbogast et al. 2002). Together, these studies show that in everyday situations, spatial separation of competing sources improves performance through improvements in the TMR at the better ear, no matter what kind of competing sounds are present; however, when masking sounds are similar to the target, perceiving sources from different directions yields additional improvements in performance by helping segregate target from masker and providing a cue that can allow listeners to direct selective attention.

Visual cues provide different benefits for different kinds of maskers

We hypothesized that a simple visual cue that only provided timing (onset and offset) information about the target would be beneficial in situations where it is hard to select the target from an ongoing masker (i.e., when the target and masker are not perceived as spatially separated or when target and masker are difficult to differentiate based on their spectrotemporal properties).

When the target and noise masker were colocated, we saw no significant benefit of visual cues; however, visual cues improved performance when sources were spatially separated, regardless of whether the stimulus presentation was the better-ear signal presented diotically or was the full binaural signal. These benefits may be the result of cross-modal enhancement (Stein and Meredith 1993), in which near-threshold auditory signals are perceptually enhanced by synchronous visual inputs. It is possible that when there is only energetic masking and target and masker are not otherwise confusable, cross-modal enhancement is maximally beneficial when the auditory input is at a very low absolute level: for the separated configurations, the target was presented about 16 dB lower at threshold on average than in the colocated conditions. Consistent with this explanation, the visual cue provided, if anything, a greater benefit for spatially separated sources when presented binaurally, which had the lowest threshold overall, than when presented diotically. The magnitude of the benefit we observed, which was on the order of a few decibels, is in the range found by other studies examining the benefits of visual cues for the detection and identification of speech (e.g., Sumby and Pollack 1954; Grant and Seitz 2000; Helfer and Freyman 2005).

When the masker was a chorus of birdsongs and thus easily confusable with the target, the visual cue provided a significant benefit when the target and masker were perceived as coming from the same location (for both binaural and diotic presentations of the colocated configuration and for diotic presentation of the spatially separated configuration). Only when the masker was perceived as coming from a different direction than the target did the visual cue fail to provide any significant benefit to target identification. Perceived spatial separation can yield large benefits for listeners in complex settings; it improves the ability to understand and recognize a target in the presence of an otherwise confusable masker (Kidd et al. 1994; Freyman et al. 1999; Best et al. 2005), presumably by helping to segregate the target from the masker and allowing listeners to focus selective attention on the target (Shinn-Cunningham 2008). The current results might suggest that visual cues for when to listen provide benefits that are similar to spatial cues, helping listeners to focus on the song syllables that correspond to the target song and segregate target from masker. In this view, visual cues for when to listen are helpful when target and masker are not already perceptually segregated; however, if other cues are present that promote target segregation from the chorus, such as perceived differences in location, visual cues are redundant and provide no significant benefit.

The differences in how visual cues improve performance for a steady-state masker and for a chorus masker highlight the different ways in which competing sounds can interfere with the ability to understand complex signals. When a masker is energetic, the primary limitation is in detecting target energy. In this case, visual cues may benefit listeners when the target is low in intensity and detecting the target energy is difficult. Visual cues can also help improve detection by helping a listener focus on the target elements, an effect that is strong even when the target elements are suprathreshold and clearly audible; however, such benefits are not observed if some other feature, such as location, is available to help guide selective auditory attention.

ACKNOWLEDGMENTS

This work was supported by a grant from the Office of Naval Research to BGSC. Three anonymous reviewers provided helpful feedback on this paper. The authors would like to thank Lorraine Delhorne for assisting with data collection.

REFERENCES

Arbogast TL, Mason CR, Kidd G (2002) The effect of spatial separation on informational and energetic masking of speech. J Acoust Soc Am 112:2086–2098. doi:10.1121/1.1510141

Bernstein LE, Auer ET, Takayanagi S (2004) Auditory speech detection in noise enhanced by lipreading. Speech Commun 44:5–18. doi:10.1016/j.specom.2004.10.011

Best V, Ozmeral E, Gallun FJ, Sen K, Shinn-Cunningham BG (2005) Spatial unmasking of birdsong in human listeners: energetic and informational factors. J Acoust Soc Am 118:3766–3773. doi:10.1121/1.2130949

Best V, Ozmeral EJ, Shinn-Cunningham BG (2007) Visually-guided attention enhances target identification in a complex auditory scene. J Assoc Res Otolaryngol 8:294–304. doi:10.1007/s10162-007-0073-z

Bonino AY, Leibold LJ (2008) The effect of signal-temporal uncertainty on detection in bursts of noise or a random-frequency complex. J Acoust Soc Am 124:321–327. doi:10.1121/1.2993745

Brungart DS, Simpson BD (2002) The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal. J Acoust Soc Am 112:664–676. doi:10.1121/1.1490592

Carhart R, Tillman TW, Greetis ES (1969) Perceptual masking in multiple sound backgrounds. J Acoust Soc Am 45:694–703. doi:10.1121/1.1911445

Drullman R, Bronkhorst AW (2000) Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation. J Acoust Soc Am 107:2224–2235. doi:10.1121/1.428503

Egan JP, Greenberg GZ, Schulman AI (1961) Interval of time uncertainty in tone detection. J Acoust Soc Am 33:771–778. doi:10.1121/1.1908795

Freyman RL, Helfer KS, McCall DD, Clifton RK (1999) The role of perceived spatial separation in the unmasking of speech. J Acoust Soc Am 106:3578–3588. doi:10.1121/1.428211

Gatehouse S, Akeroyd MA (2008) The effects of cueing temporal and spatial attention on word recognition in a complex listening task in hearing-impaired listeners. Trends Amplif 12:145–161. doi:10.1177/1084713808317395

Grant KW, Seitz PF (2000) The use of visible speech cues for improving auditory detection of spoken sentences. J Acoust Soc Am 108:1197–1208. doi:10.1121/1.1288668

Green DM, Weber DL (1980) Detection of temporally uncertain signals. J Acoust Soc Am 67:1304–1311. doi:10.1121/1.384183

Helfer KS, Freyman RL (2005) The role of visual speech cues in reducing energetic and informational masking. J Acoust Soc Am 117:842–849. doi:10.1121/1.1836832

Hirsh IJ (1950) The relation between localization and intelligibility. J Acoust Soc Am 22:196–200. doi:10.1121/1.1906588

Kidd G, Mason CR, Deliwala PS, Woods WS, Colburn HS (1994) Reducing informational masking by sound segregation. J Acoust Soc Am 95:3475–3480. doi:10.1121/1.410023

Kidd G, Mason CR, Rohtla TL, Deliwala PS (1998) Release from masking due to spatial separation of sources in the identification of nonspeech auditory patterns. J Acoust Soc Am 104:422–431. doi:10.1121/1.423246

Kidd G, Mason CR, Richards VM, Gallun FJ, Durlach NI (2008) Informational masking. In: Yost WA, Fay RR (eds) Auditory perception of sound sources, Springer handbook of auditory research. Springer Science+Business Media, New York, pp 143–189

Marrone N, Mason CR, Kidd G (2008) Tuning in the spatial dimension: evidence from a masked speech identification task. J Acoust Soc Am 124:1146–1158. doi:10.1121/1.2945710

Shinn-Cunningham BG (2008) Object-based auditory and visual attention. Trends Cogn Sci 12:182–186. doi:10.1016/j.tics.2008.02.003

Shinn-Cunningham BG, Best V (2008) Selective attention in normal and impaired hearing. Trends Amplif 12:283–299. doi:10.1177/1084713808325306

Shinn-Cunningham BG, Kopco N, Martin TJ (2005) Localizing nearby sound sources in a classroom: binaural room impulse responses. J Acoust Soc Am 117:3100–3115. doi:10.1121/1.1872572

Stein BE, Meredith MA (1993) The merging of the senses. MIT Press, Cambridge

Sumby WH, Pollack I (1954) Visual contribution to speech intelligibility in noise. J Acoust Soc Am 26:212–215. doi:10.1121/1.1907309

Tye-Murray N, Spehar B, Myerson J, Sommers MS, Hale S (2011) Cross-modal enhancement of speech detection in young and older adults: does signal content matter? Ear Hear 32:650–655. doi:10.1097/AUD.0b013e31821a4578

Zurek PM (1993) Binaural advantages and directional effects in speech intelligibility. In: Studebaker GA, Hochberg I (eds) Acoustical factors affecting hearing aid performance II. Allyn and Bacon, Boston