Spatial release from energetic and informational masking in a divided speech identification task[a]

Antje Ihlefeld and Barbara Shinn-Cunningham[b]
Auditory Neuroscience Laboratory, Boston University Hearing Research Center, 677 Beacon St., Boston, Massachusetts 02215, USA

(Received 3 January 2007; revised 12 March 2008; accepted 13 March 2008)

When listening selectively to one talker in a two-talker environment, performance generally improves with spatial separation of the sources. The current study explores the role of spatial separation in divided listening, when listeners reported both of two simultaneous messages processed to have little spectral overlap (limiting "energetic masking" between the messages). One message was presented at a fixed level, while the other message level varied from equal to 40 dB less than that of the fixed-level message. Results demonstrate that spatial separation of the competing messages improved divided-listening performance. Most errors occurred because listeners failed to report the content of the less-intense talker. Moreover, performance generally improved as the broadband energy ratio of the variable-level to the fixed-level talker increased. The error patterns suggest that spatial separation improves the intelligibility of the less-intense talker by improving the ability to (1) hear portions of the signal that would otherwise be masked, (2) segregate the two talkers properly into separate perceptual streams, and (3) selectively focus attention on the less-intense talker. Spatial configuration did not noticeably affect the ability to report the more-intense talker, suggesting that it was processed differently than the less-intense talker, which was actively attended. © *2008 Acoustical Society of America.*

[DOI: 10.1121/1.2904825]

PACS number(s): 43.66.Dc, 43.66.Pn, 43.66.Qp [RLF] Pages: 4380–4392

I. INTRODUCTION

Previous studies on selective listening show that listeners are generally good at retrieving information from a source at a location they are attending, but perform poorly when asked to recall messages from unexpected locations (Cherry, 1953; Yost, 1997; Arbogast and Kidd, 2000). Nonetheless, in a recent study, listeners did remarkably well when asked to report two simultaneous messages, and overall performance was only weakly influenced by the amount of spatial separation between two concurrent speech sources (Best *et al.*, 2006). Other studies confirm that while listeners typically can attend to only one auditory message at a time (Cherry, 1953; Broadbent, 1954), they have some capacity to process semantic information from messages outside the immediate focus of attention (e.g., see Moray, 1959; Lawson, 1966; Cowan, 1995; Conway *et al.*, 2001; Rivenez *et al.*, 2006).

Previous work indicates that when trying to understand several sources at the same time, listeners may actively attend to one during the presentation of the stimulus and then selectively read out information about other source(s) from temporary buffers, after the stimulus ended (Conway *et al.*, 2001; Best *et al.*, 2006). This suggests that when asked to report two concurrent sources, listeners may exploit spatial cues to selectively attend to one source during the presentation and then process the other source from memory.

It is not clear whether selective spatial attention operates on a buffered representation of the other source like it does on an ongoing stimulus. If the listener cannot access pre-segregated objects in the buffered representation, competition between sources may play out differently for a recalled message than for a message that is selectively attended during the stimulus presentation. Therefore, it is difficult to predict how the spatial configuration of the sources will influence the ability to extract information about a recalled message. This paper attempts to disentangle how the location of the attended message in a two-talker setting influences the ability to extract information from two simultaneously presented messages.

Studies of selective listening identify numerous forms of interference that can limit performance in speech identification tasks (cf. Brungart, 2001; Brungart *et al.*, 2005; Freyman *et al.*, 2005; Kidd *et al.*, 2005; Ihlefeld and Shinn-Cunningham, 2008). Energetic masking occurs when the masker interferes with the peripheral representation of the target (Cherry, 1953; Spieth *et al.*, 1953; Moray, 1959; Ebata *et al.*, 1968). Informational masking occurs either because listeners (1) cannot segregate the target from the masker, and/or (2) cannot select the target out of a mixture of similar, properly segregated maskers, possibly because they are uncertain as to which sound features constitute the target (Arbogast *et al.*, 2002; Brungart and Simpson, 2002; Durlach *et al.*, 2003; Lutfi *et al.*, 2003; Brungart *et al.*, 2005; Watson, 2005; Best *et al.*, 2005; Shinn-Cunningham *et al.*, 2005).

[a] Portions of this work were presented at the 2005 Mid-Winter meeting of the Association for Research in Otolaryngology.

[b] Author to whom correspondence should be addressed. Electronic mail: shinn@cns.bu.edu.

Both energetic masking and informational masking are likely to affect performance when listeners try to report multiple simultaneous target messages (see also Best *et al.*, 2006). In selective listening, the spatial separation of competing sources influences performance by improving the audibility of the target (reducing energetic masking, e.g., see Zurek, 1993), improving perceptual segregation of the sources (reducing one form of informational masking, e.g., see Freyman *et al.*, 1999), and decreasing confusions between target and masker (reducing the other form of informational masking, e.g., see Brungart, 2001).

The current study has two aims. The first aim is to gain a better understanding of how energetic masking and informational masking interfere with the ability to report the less-intense target when listeners are asked to report two simultaneous targets. The second aim is to examine the roles of spatial factors on performance in a divided listening task, and determine how they change as the relative contributions of energetic masking and informational masking vary.

Listeners were asked to report the content of two concurrent utterances. The spatial separation between the talkers was varied from block to block, and the broadband energy ratio between the talkers was varied within each block to systematically change the relative contributions of energetic masking versus informational masking (see companion paper about a selective attention experiment with identical stimuli; Ihlefeld and Shinn-Cunningham, 2008). To the extent that divided listening consists of first selectively attending to one message and then reporting the other message, spatial separation should improve the ability to report the actively attended message through a combination of acoustic effects at the better ear for the attended message, binaural processing, and spatial release from informational masking through spatially directed attention (e.g., see Best *et al.*, 2006; Ihlefeld and Shinn-Cunningham, 2008). However, spatial separation is also likely to influence the ability to process the other message, either negatively (because listeners attend the location of the initially processed target message, which impairs performance for sources from other locations), or positively (because the two targets are perceptually more distinct). Evidence for both effects was found by Best and colleagues (Best *et al.*, 2006). In the current study, *post-hoc* analysis of response patterns supports the idea that listeners actively attended to the less-intense target and recalled the more-intense target through a different mechanism. We find that spatial separation of the concurrent messages (1) improved the ability to report the actively attended source in ways comparable to improvements in selective listening (Ihlefeld and Shinn-Cunningham, 2008), but (2) neither helped nor hindered the ability to report the other message.

II. METHODS

The methods used in this study are essentially identical to the methods used in the companion paper which tested selective attention (Ihlefeld and Shinn-Cunningham, 2008). The same subjects participated in both studies. Stimuli and procedures were identical, except for the instructions (to report one of the two sources in the selective task described in the companion paper, or to report both messages in the current divided task). Methods are summarized briefly here (see Ihlefeld and Shinn-Cunningham, 2008, for more details).

A. Subjects

Four subjects (ages 21–24) were paid for their participation in the experiments. All subjects were native speakers of American English and had normal hearing, confirmed by an audiometric screening. All subjects gave written informed consent (as approved by the Boston University Charles River Campus Institutional Review Board) before participating in the study.

B. Stimuli

Raw speech stimuli were taken from the Coordinate Response Measure corpus (CRM, see Bolia *et al.*, 2000). Sentences were processed to produce intelligible, spectrally sparse speech signals (e.g., see Shannon *et al.*, 1995; Dorman *et al.*, 1997; Arbogast *et al.*, 2002; Brungart *et al.*, 2005). Each target and masker source signal was bandpass filtered into 16 logarithmically spaced, adjacent frequency bands of 1/3 octave width (center frequencies 175 Hz–5.6 kHz). The envelope of each band was extracted using the Hilbert transform. Subsequently, each envelope was multiplied by a pure-tone carrier at the center frequency of that band.

On each individual trial, eight of the 16 bands were chosen randomly (four from the lower eight frequency bands and four from the upper eight frequency bands) to create the raw waveform for one source. The remaining eight bands were used to construct the other source using otherwise identical processing. The raw source waveforms were scaled to have the same fixed, broadband root mean square (RMS) energy reference level prior to spatial processing (described below).

C. Spatial synthesis

Raw signals were processed with head-related transfer functions (HRTFs) of an acoustic manikin to simulate sources from 0° (in front) or 90° (to the side) azimuth, at a distance of 1 m in the horizontal plane (see Ihlefeld and Shinn-Cunningham, 2008, for details).

D. Procedures

One talker, referred to as the fixed-level talker (target_F) was always presented at the same reference RMS level (set to approximately 70 dB sound pressure level prior to spatial processing). The level of the other, variable-level talker (target_V) was attenuated relative to target_F by an amount that varied randomly from trial to trial, chosen from one of five levels (−40, −30, −20, −10, and 0 dB). Subsequently, the binaural signals for the two talkers were summed to produce the two-talker stimulus. As a result of this manipulation of target_V, the nominal energy ratio between the two talkers varied (without taking into account spatial processing effects). The broadband energy ratio between target$_V$ and target$_F$ will be denoted by T_VT_FR. In this study, T_VT_FR ranged from −40 dB to 0 dB.

There were four possible spatial configurations, two in which the two talkers were co-located (at either 0 or 90°) and two in which the talkers were spatially separated (target$_V$ at 0° and target$_F$ at 90°, or target$_V$ at 90° and target$_F$ at 0°). In each run, the spatial configuration of the two talkers was fixed (i.e., the talkers were played from the same location throughout the run).

Stimuli were digital-to-analog converted, amplified using Tucker-Davis System 3 hardware, and presented over Sennheiser HD 580 headphones to subjects seated in a sound-treated booth. Following each trial, subjects indicated perceived target keywords using a graphical user interface (GUI), after which the GUI indicated the correct response.

At the start of each session, a random call sign was selected to serve as the call sign of target$_V$, matching the procedures used in the companion study of selective attention (Ihlefeld and Shinn-Cunningham, 2008). In contrast, the call sign of target$_F$ varied randomly from trial to trial. Target$_V$ and targe$_F$ always had different call signs. Listeners were instructed to report the colors and numbers of both talkers. They were not explicitly instructed to report these keywords in proper pairs corresponding to the two physical sources, nor were they made aware of the fact that target$_V$ had a fixed call sign throughout the session. A trial was scored as correct and subjects were given feedback to that effect only if they reported all of the four keywords in any order.

Prior to testing, all subjects went through an initial screening in which they had to report the color and number of one talker of processed speech presented in quiet (processed by a 0° azimuth HRTF). In order to proceed with the experiment, they had to achieve at least 90% correct over the course of 50 such trials. None of the subjects failed this initial screening. Following the screening, each subject performed a training session consisting of 300 trials (at least one run of 50 trials for each spatial configuration, and an additional run of 50 trials for each of two randomly picked spatial configurations).

Following training, subjects performed four sessions of the experiment (one session per day). In the other four sessions, subjects performed a selective-attention task (reported in Ihlefeld and Shinn-Cunningham, 2008). Each session consisted of 12 runs (three runs for each of the four spatial configurations) of either the selective or the divided task. The order of the sessions and the order of the runs within each session were separately randomized for each subject, but constrained so that each spatial configuration and each of the two tasks was performed once before any were repeated. A run consisted of eight repetitions of each of the five T_VT_FR s (40 trials per run). The orders of the runs within each session were separately randomized for each subject. Given that each subject performed four sessions of this experiment, each subject performed 96 repetitions of each specific configuration (8 repetitions/run × 3 runs/session × 4 sessions).

E. Hypotheses

In the current task, the listener was asked to report the content of both of two simultaneous messages. On each trial, subjects responded by first reporting one color-number pair and then reporting a second color-number pair. The color and number from target$_V$ will be denoted by C_V and N_V, respectively. Similarly, the color and number from target$_F$ will be denoted by C_F and N_F. Color and number responses that are not keywords in either message will be signified by C_X and N_X, respectively. The order and pairing in which keywords were reported was not important for the score that listeners received. Specifically, for stimulus $[C_VN_VC_FN_F]$, the following four responses were scored as correct: $[C_VN_VC_FN_F]$, $[C_FN_FC_VN_V]$, $[C_VN_FC_FN_V]$, and $[C_FN_VC_VN_F]$, where order of report corresponds to the pair order within the brackets.

The ability to correctly report what both talkers were saying depends on whether the listener can hear and segregate the target words. In addition, listeners need to divide their processing resources between the two competing talkers. As in selective listening tasks (Brungart, 2001; Kidd *et al.*, 2005; Ihlefeld and Shinn-Cunningham, 2008), analyzing response errors made in divided listening tasks may illuminate the underlying response strategies that listeners use. Several factors can contribute to a failure to hear a target. Importantly, in the current experiment, listeners may not hear a target message because (1) it was energetically masked by the other source (energetic masking), or because (2) listeners failed to hear out and remember that target, even though it was well above the threshold of audibility (informational masking).

The relative influence of energetic masking compared to informational masking is likely to depend on the energy ratio between the two talkers (Ihlefeld and Shinn-Cunningham, 2008). If listeners truly selectively attend to one target message and then recall the other message, then the pattern of errors should depend on the kind of interference present for the attended target (Brungart *et al.*, 2001), while the ability to report the recalled target will depend on how well it is represented in memory. When target$_V$ is at least 20 dB less intense than target$_F$, target$_V$ may be difficult to hear (energetic masking; Ihlefeld and Shinn-Cunningham, 2008). In such trials, target$_F$ should have a clean representation both in the direct sensory input and in any temporary buffer and should therefore be easy to recall, regardless of the spatial configuration of the talkers. In such conditions, the intelligibility of the less-intense talker should be the main factor limiting divided-attention performance. Thus, performance should improve as the relative level of the less-intense talker increases (much as performance in selective listening improves with increasing target-to-masker ratio; e.g., see Arbogast *et al.*, 2002; Shinn-Cunningham *et al.*, 2005). When the two competing talkers are spatially separated, the overall energy ratio of the less-intense talker relative to the more-intense talker will improve at one ear. Furthermore, binaural cues will increase the audibility of the less-intense talker by a modest amount when it is near detection threshold (Zurek, 1993). Therefore, to the extent that the less-intense target determines performance, divided performance should improve with spatial separation.

If attention can be focused on only one location at a time, increasing spatial separation between the two concurrent messages may also increase the number of *drop* errors for the recalled message (e.g., responding $[C_V N_V C_X N_X]$ if target_V was attended, or $[C_F N_F C_X N_X]$ if target_F was attended, where C_x and N_x denote a color and number not present in either utterance; see Best *et al.*, 2005). Note that while in selective listening a failure to hear the single target can cause listeners to erroneously report the content of the masker message, in the current divided task, listeners will end up guessing the content of the source they tried to attend while still reporting the message of the other target that they heard. Here, in the majority of trials, target_F is relatively intense and salient, whereas target_V is usually much harder to hear than target_F. If listeners therefore attend to target_V at its location, spatial separation may increase the number of drop errors for target_F.

Finally, while listeners were not explicitly instructed to report the keywords of both talkers in proper pairs, they may have a natural tendency to do so. It should be difficult for listeners to report both messages without confusions when the talkers are similar in level and have the same perceived location. Specifically, when both targets are clearly audible but perceptually similar, listeners may have difficulty segregating the talkers; or they may be able to segregate the words and recall keywords from both messages, but may confuse which talker spoke which words (informational masking). Although there was no penalty for responding this way, listeners reporting a mix of target_F and target_V keywords (i.e., $[C_V N_F C_F N_V]$ or $[C_F N_V C_V N_F]$), henceforth *mix responses*, may reflect less complete perceptual segregation and streaming of the two sources compared to trials in which they report the keywords in proper pairs (i.e., $[C_V N_V C_F N_F]$ or $[C_F N_F C_V N_V]$), which will be called *fully correct* responses. Any systematic patterns in the relative likelihood of mix versus fully correct response likely reflect differences in the degree of perceptual segregation of the target and the masker.

III. RESULTS

Section III A analyzes the probability of reporting all four keywords correctly, independent of their pairing and response ordering. More detailed analysis of the kinds of response errors and order of responses are given in subsequent sections.

A. Percent correct

On each trial, subjects responded by reporting two color—number pairs. After each trial, subjects received feedback that they were correct if and only if they reported all four keywords of both utterances, regardless of how they paired keywords from the talkers. Therefore, the likelihood of responding correctly by chance equals $4 \times 1/4 \times 1/7 \times 1/3 \times 1/6$ or 0.8%. However, if subjects heard target_F but could not hear and had to randomly guess the keywords for target_V, the likelihood of being correct by chance would equal $1/3 \times 1/6$ or 6%.

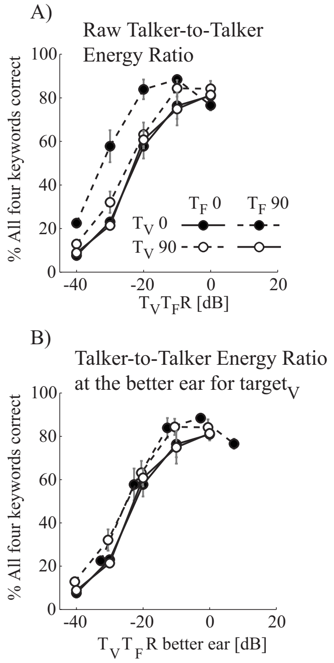

FIG. 1. Percent correct as a function of energy ratio between target_V and target_F ($T_V T_F R$). Error bars show the across-subject standard error of the mean. In general, performance improves with $T_V T_F R$, and is better for spatially separated than co-located sources. Filled symbols show results for target_V at 0° and open symbols show results for target_V at 90°. Results for spatially separated targets are shown with dashed lines. Results for co-located sources are shown with solid lines. (A) Results plotted as a function of the broadband target to target broadband energy ratio ($T_V T_F R$). (B) The same results re-plotted as a function of $T_V T_F R_{\text{be-V}}$ (correcting for differences in the acoustic $T_V T_F R$ at the better ear for target_V).

1. Overall percent correct

Figure 1(a) shows percent correct as a function of $T_V T_F R$, averaged across subjects (error bars show the across-subject standard error). All subjects showed relatively similar results, so only the across-subject average results are shown. As the intensity of target_V increased, performance improved. Performance was better in the spatially separated configurations than in the co-located configurations (dashed lines fall above solid lines). At the lowest $T_V T_F R$, performance was near 6% for the co-located configurations, 18% for T_V at 90° and T_F at 0°, and 22% for T_V at 0° and T_F at 90°. For all spatial configurations, performance improved with increasing intensity of target_V until it reached an upper bound of roughly 70%.

Performance was essentially identical for co-located sources whether they played from in front or from the side of the listener. For the spatially separated configurations, performance was better when target_V was playing from in front and target_F from the side of the listener than when their positions were reversed. As shown in the companion paper (Ihlefeld and Shinn-Cunningham, 2008; see also Shinn-Cunningham *et al.*, 2005), differences in the broadband acoustic target-to-masker energy ratio at the better acoustic ear accounted for differences in selective listening performance for different spatially separated spatial configurations. The RMS energy of the two messages was equated prior to spatial processing; the level of target$_V$ was then adjusted to produce the desired T_VT_FR. However, spatial processing also altered the levels of the talkers at each ear. When the two talkers were spatially separated, there was always one ear (the acoustically better ear for target$_V$) that received a higher broadband T_VT_FR than the other ear. When target$_V$ was in front of the listener and target$_F$ was to the right side of the listener, the T_VT_FR at the left ear was on average 7 dB greater than the nominal T_VT_FR prior to spatial processing for the stimuli used in this study (see analysis in Shinn-Cunningham *et al.*, 2005). Conversely, when target$_F$ came from in front and target$_V$ was to the right of the listener, the right ear was the better ear for target$_V$, with a T_VT_FR that was on average 1 dB lower than the nominal T_VT_FR prior to spatial processing. Note that in the co-located configurations, T_VT_FR at both ears equaled the nominal T_VT_FR (on average).

Figure 1(b) shows the data from Fig. 1(a) re-plotted as a function of the T_VT_FR at the better ear for target$_V$ ($T_VT_FR_{\text{be-V}}$) by shifting the raw data horizontally by the appropriate dB amount for each spatial configuration. This adjustment completely accounts for performance differences between the two spatially separated configurations [dashed lines in Fig. 1(b) are virtually identical], just as in the companion study of selective listening (Ihlefeld and Shinn-Cunningham, 2008).

2. Spatial gains

For each subject and spatial configuration, percent correct performance as a function of $T_VT_FR_{\text{be-V}}$ was fitted by logistic functions (see Appendix B). For each individual subject, the psychometric function fits for the two co-located configurations were averaged, as were the psychometric function fits for the two spatially separated configurations (after accounting for the acoustic advantage at the better ear for target$_V$). Between −30 and −20 dB $T_VT_FRs_{\text{be-V}}$, the vertical difference of these averaged spatially separated and co-located psychometric function fits (the percent spatial gain) was 6% for subjects S1 and S3, 10% for subject S2, and 13% for subject S4. At the greatest $T_VT_FRs_{\text{be-V}}$ the percent spatial gain, equal to the difference in upper bounds, was between approximately 5% (subjects S1, S2, and S3) and 11% (S4). The horizontal shift between the logistic fit (the dB spatial gain) was approximately 2–4 dB (for all subjects).

3. Analysis of response pairing

In general, despite the fact that listeners were not instructed to report the messages in correct pairings, they tended to do so. Was there a positive effect of spatial separation on the likelihood of reporting keywords in pairings that correspond to the target messages? To examine this question, all trials where subjects responded correctly were analyzed in more detail. In the majority of the trials in which all four keywords were reported, they were reported in proper pairings (i.e., in 91% of all fully correct trials subjects either reported [$C_VN_VC_FN_F$] or [$C_FN_FC_VN_V$]; fully correct). The pattern of fully correct responses was very similar to the pattern for overall correct responses (shown in Fig. 1) and was not analyzed in more detail. However, analysis of the order in which the proper pairs were reported revealed interesting patterns, considered further in the Appendix A.

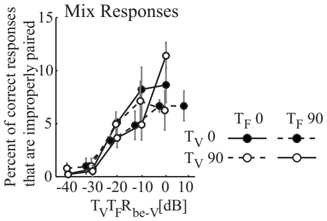

FIG. 2. Mix responses as a function of $T_VT_FR_{\text{be-V}}$ for each spatial configuration, averaged across subjects. Error bars show the across-subject standard error of the mean. Mix responses increase with increasing $T_VT_FR_{\text{be-V}}$. Filled symbols show results for target$_V$ at 0°, while open symbols show results for target$_V$ at 90°. Results for spatially separated sources are denoted with dashed lines and for co-located sources with solid lines.

The less-common trials in which subjects responded with all four keywords correct, but in improper pairings (i.e., in which they reported [$C_VN_FC_FN_V$] or [$C_FN_VC_VN_F$]) are denoted as mix responses (even though listeners were given feedback in that these responses counted as correct) to reflect the fact that subjects mixed the keywords from the two target streams in their responses.

Chance performance for mix responses was $2\times1/4\times1/7\times1/3\times1/6$ or 0.4%. Overall, the rate of mix responses (Fig. 2) was a small subset of the correct responses (shown in Fig. 1). Figure 2 shows the pattern of mix responses as a function of $T_VT_FR_{\text{be-V}}$. Mix responses increased with increasing $T_VT_FR_{\text{be-V}}$. In other words, the more similar the two targets became in level, the more likely listeners were to mix up keywords from the two sources. There are no clear differences in how often listeners made mix responses across different spatial configurations, except near 0 dB $T_VT_FR_{\text{be-V}}$, where slightly more mix responses occurred when sources were co-located compared to when they were separated (dashed lines fall below solid lines near 0 dB $T_VT_FR_{\text{be-V}}$). In other words, subjects were most likely to mix the streams when the sources were both at the same intensity (0 dB $T_VT_FR_{\text{be-V}}$) and at the same location in space. While this effect does not reach statistical significance in this study ($F(1,3)=7.741$, $p=0.069$), it is consistent with results from our companion selective listening study, which showed the greatest number of confusions between target and masker when the sources were co-located and at nearly the same level (Ihlefeld and Shinn-Cunningham, 2008).[1]

In selective listening, differences in both level and location can improve a listener's ability to selectively attend to the target source. To a lesser extent than in the selective listening task, these same factors reduced confusions between the competing talkers in this divided task. This is consistent with the idea that listeners first selectively attended to target$_V$ and then recalled target$_F$.

B. Spatial effects on reporting the second message

In Sec. III A, only those trials in which all four keywords were reported were analyzed. However, this analysis

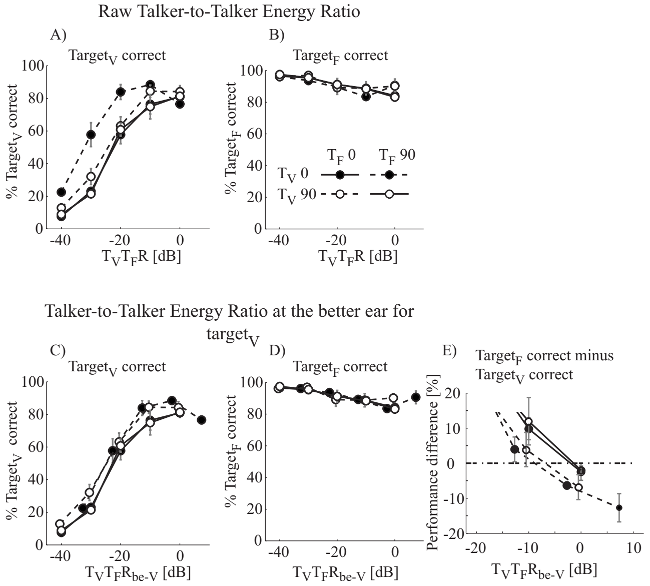

FIG. 3. Probability of reporting each target correctly in a proper pairing as a function of $T_VT_FR_{\text{be-V}}$, averaged across subjects. Error bars show the across-subject standard error of the mean. Spatial separation improves performance for target$_V$ but has no significant effect on target$_F$. Filled symbols show results for target$_V$ at 0° and open symbols for target$_V$ at 90°. Results for spatially separated targets are shown with dashed lines and for co-located sources with solid lines. (A) Results for target$_V$ correct as a function of T_VT_FR). (C) The same target$_V$ results re-plotted as a function of $T_VT_FR_{\text{be-V}}$ (correcting for differences in the acoustic T_VT_FR at the better ear for target$_V$). (B) Results for target$_F$ correct as a function of T_VT_FR. (D) Target$_F$ results re-plotted as a function of $T_VT_FR_{\text{be-V}}$. (E) Target$_F$ correct minus target$_V$ correct (difference between curves in panels D and C) when the two messages are similar in level (i.e., around $T_VT_FR_{\text{be-V}}$ near 0 dB).

ignored those trials in which part but not all of the response was correct. As discussed in the Introduction, we hypothesized that spatial separation negatively affects the ability to process both messages. In particular, if listeners attend to the location of the initially processed target message, it may impair performance for the second target when the second target comes from a different location than the initially processed target. To examine the effect of spatial separation on reporting the second message, here, all trials when listeners succeeded in reporting one of the two messages are analyzed separately for target$_V$ and target$_F$. When subjects reported both keywords of target$_V$ as a pair (i.e., either responded either $[C_VN_VC\_N\_]$ or $[C\_N\_C_VN_V]$, where "_" denotes a target$_F$ keyword or a keyword not present in either message), a trial was scored as target$_V$ correct. Analogously, a trial was scored as target$_F$ correct when the response contained the color and number of target$_F$ in one pair (i.e., subjects responded either $[C_FN_FC\_N\_]$ or $[C\_N\_C_FN_F]$. Note that although it was not explicitly pointed out to them, in principle, listeners could differentiate between target$_V$ and target$_F$, because target$_V$ (the target that was usually softer and therefore harder to hear) had a call sign that was fixed throughout the course of the session. In contrast, the call sign of target$_F$ varied randomly from trial to trial (but never equaled the call sign of target$_V$; see also Sec. II D).

Figure 3 plots the percentage of trials in which a message was reported in correct pairing for target$_V$ [Fig. 3(a)] and target$_F$ [Fig. 3(b)] as a function of T_VT_FR and as a function of $T_VT_FR_{\text{be-V}}$ [target$_V$ and target$_F$ in Figs. 3(c) and 3(d), respectively]. As a function of T_VT_FR, performance for target$_V$ [Fig. 3(a)] was better when the two talkers are spatially separated than when they are co-located, and was best when target$_V$ is at 0° and target$_F$ is at 90°. When plotted as a function of $T_VT_FR_{\text{be-V}}$, results for target$_V$ were similar for the two spatially separated configurations [Fig. 3(c)]. Plotted this way, there was a small spatial gain of about 2–3 dB for both target$_V$ in front and target$_F$ to the side (dashed lines fall above solid lines). In contrast, performance for target$_F$ [Fig. 3(b)] did not depend strongly on spatial configuration (the four lines are overlapping). Normalization to take into account the better-ear ratio for target$_F$ [plotting as a function of $T_VT_FR_{\text{be-V}}$; Fig. 3(d)] had little effect on the psychometric curve describing the probability of correctly reporting target$_F$ and did not reveal any systematic effect of spatial configuration on performance.[2] However, ceiling effects may account for the lack of spatial effects on performance for target$_F$.

Finally, while for the majority of trials performance for the more-intense target$_F$ was better than for target$_V$, one might expect that the two talkers were equally hard to understand near 0 dB $T_VT_FR_{\text{be-V}}$, when they were near the same intensity. However, subjects performed better for target$_V$ than for target$_F$. Given the scales of panels 3(C) and 3(D), it is difficult to make direct comparisons and see this difference. To make this pattern clearer, for each spatial configuration, the differences between the percentages correct for target$_V$ correct and target$_F$ are shown in Fig. 3(e) for points near 0 dB $T_VT_FR_{\text{be-V}}$ (error bars show across-subject standard errors). For trials with $T_VT_FR_{\text{be}}$ near 0 dB, the differences in performance between target$_V$ and target$_F$ are consistently negative for all spatial configurations [i.e., performance is better for target$_V$ than for target$_F$; Fig. 3(e); coding of line style and symbols is the same as in the other panels].

Further analysis of report order in Appendix A shows that there are systematic differences in how listeners prioritize the two messages. These results suggest that both (1) when listeners selectively attended to target$_V$ and (2) when sources were spatially separated in this divided task, the source from in front of the listener was inherently more salient than the source to the side (see Appendix A).

In summary, spatial separation improved the ability to hear out target$_V$, but had no significant effect on performance for target$_F$. Moreover, subjects appeared to devote more processing resources to target$_V$, as evidenced by the fact that they were better at reporting target$_V$ than target$_F$ when both were equally intense (or even when target$_V$ is slightly less intense than target$_F$), even when the two messages were co-located. In other words, listeners appear to have attended to the location of the initially processed target message, but this did not impair performance for the second target when the second target came from a different location than the initially processed target.

C. Incorrect responses

Performance for all trials in which responses were not counted correct was analyzed in more detail to see if there

TABLE I. Definitions of and chance probabilities of different error types

Response type	Chance	Responses
Target$_V$ drop error	6.7%	[C_FN_F C_XN_V], [C_FN_F C_VN_X], [C_FN_F C_XN_X], [C_XN_V C_FN_F], [C_VN_X C_FN_F], [C_XN_X C_FN_F]
Target$_F$ drop error	6.7%	[C_VN_V C_XN_F], [C_VN_V C_FN_X], [C_VN_V C_XN_X], [C_XN_F C_VN_V], [C_FN_X C_VN_V], or [C_XN_X C_VN_V]
Target$_V$ combination error	6.7%	[C_VN_X C_XN_V], [C_VN_X C_FN_V], [C_VN_F C_XN_V], [C_XN_V C_VN_X], [C_FN_V C_VN_X], or [C_XN_V C_VN_F]
Target$_F$ combination error	6.7%	[C_FN_X C_XN_F], [C_FN_X C_VN_F], [C_FN_V C_XN_F], [C_XN_F C_FN_X], [C_VN_F C_FN_X], or [C_XN_F C_FN_V]
Mix response	0.4%	[C_VN_F C_FN_V] or [C_FN_V C_VN_F]
Fully correct	0.4%	[C_VN_V C_FN_F] or [C_FN_F C_VN_V]
Other	72.2%	All trials that were not fully correct, mix responses, drop errors, or combination errors were scored as other.

was any evidence for how listeners prioritized the messages. Table I shows the definitions of the different response types and the distributions of the different incorrect responses [Tables I and II, respectively; Note that the total sum of the percentages of incorrect responses, fully correct responses, and mix responses equals 100%].

Across all $T_VT_FR_{\text{be-V}}$, drop errors (where subjects reported both keywords of one target but failed to report both

TABLE II. Distribution of types of incorrect responses in percent as a function of $T_VT_FR_{\text{be-V}}$ averaged across subjects (standard error in parentheses). The top half of the table shows results when the targets are coming from the same location; the bottom half shows results when the targets are spatially separated.

T_VT_FR	−40 dB		−30 dB		−20 dB		−10 dB		0 dB	
	Co-located configurations									
Incorrect Response [%]	T_V0° T_F0°	T_V90° T_F90°	T_V0° T_F0°	T_V90° T_F90°	T_V0° T_F0°	T_V90° T_F90°	T_V0° T_F0°	T_V90° T_F90°	T_V0° T_F0°	T_V90° T_F90°
Target$_V$ drop error	88 (3)	87 (2)	72 (3)	74 (3)	35 (9)	34 (8)	16 (8)	19 (13)	10 (6)	12 (4)
Target$_F$ drop error	0 (1)	0 (1)	2 (1)	2 (1)	4 (2)	5 (4)	7 (2)	7 (2)	13 (2)	14 (3)
Target$_V$ combination error	1 (1)	2 (1)	3 (1)	3 (2)	4 (2)	3 (2)	5 (3)	3 (4)	3 (2)	3 (2)
Target$_F$ combination error	0 (0)	0 (0)	1 (1)	1 (2)	1 (1)	1 (1)	2 (1)	2 (1)	4 (2)	3 (2)
Other	3 (1)	2 (1)	2 (3)	2 (1)	3 (1)	3 (3)	2 (2)	3 (1)	3 (2)	2 (2)
	Spatially separated configurations									
	T_V0° T_F90°	T_V90° T_F0°	T_V0° T_F90°	T_V90° T_F0°	T_V0° T_F90°	T_V90° T_F0°	T_V0° T_F90°	T_V90° T_F0°	T_V0° T_F90°	T_V90° T_F0°
Target$_V$ drop error	74 (2)	83 (4)	36 (15)	63 (11)	11 (7)	30 (9)	8 (1)	12 (5)	7 (7)	7 (3)
Target$_F$ drop error	1 (0)	0 (1)	2 (1)	1 (0)	7 (2)	6 (6)	14 (2)	8 (8)	19 (4)	14 (6)
Target$_V$ combination error	2 (2)	2 (2)	3 (2)	3 (2)	4 (3)	4 (3)	2 (1)	2 (1)	2 (1)	2 (2)
Target$_F$ combination error	1 (1)	0 (0)	1 (1)	1 (1)	2 (2)	2 (2)	1 (1)	2 (2)	3 (2)	1 (1)
Other	2 (1)	2 (4)	3 (4)	1 (2)	1 (2)	3 (2)	2 (2)	2 (2)	1 (1)	1 (1)

of the keywords of the other target) were the dominant kind of response error. In all spatial configurations, the relative likelihood of target$_V$ drop errors decreased with increasing $T_VT_FR_{be\text{-}V}$, while target$_F$ drop errors increased. Target$_V$ drop errors were less common when the targets were spatially separated than when they were co-located ($F(1,3)=92.549$, $p=0.002$). In contrast, the percentage of target$_F$ drop errors did not vary significantly with the spatial configuration of the talkers ($F(1,3)=2.131$, $p=0.24$). Combination errors, which occur when listeners succeed in segregating one of the targets out of the acoustic mixture but fail to properly stream it across time, tended to increase with increasing $T_VT_FR_{be\text{-}V}$. However, while the relative number of target$_F$ combination errors increased monotonically with increasing $T_VT_FR_{be\text{-}V}$ ($F(4,12)=4.871$, $p=0.014$), the percentage of target$_V$ combination errors increased between −40 and −20 dB $T_VT_FR_{be\text{-}V}$, and then either decreased or remained constant as the two targets became more similar in level (no significant effect of $T_VT_FR_{be\text{-}V}$, $F(4,12)=7.921$, $p=0.341$). Although this is not a strong trend, it was consistent across all spatial configurations. This result hints that level cues influence the segregation and streaming of target$_V$ more than target$_F$. Other errors are very uncommon and do not depend consistently on $T_VT_FR_{be\text{-}V}$ or spatial configuration.

IV. DISCUSSION

A previous divided-listening study found that spatial separation between concurrent messages improves performance slightly, but that the dominant benefit of spatial separation was from purely acoustic effects (Best *et al.*, 2006). However, that study presented two messages of equal intensity, making it difficult to assess the full impact of other spatial factors. That study also found evidence for two opposing effects of spatial separation in divided listening: spatial separation leads to an improvement in perceptual segregation of the concurrent sources, but a degradation in the ability to process both of the two simultaneous sources. In one experiment in that study, the two competing sources were equated such that they were equally intelligible in a selective listening task, but listeners were instructed to report the target message that was relatively more to the left before the target message that was relatively more to the right. These instructions caused listeners to devote more attentional resources to the left source that they had to report first. As a result, listeners made more errors for the lower-priority, right source. Moreover, the effects of spatial separation on the two sources differed. Best *et al.* concluded that listeners actively attended the higher-priority, left source and then recalled the lower-priority, right source, and that spatial separation had very different effects on the ability to report the two sources.

The current results support and extend these findings. Here, intelligibility of two spectrally degraded competing targets was investigated as a function of their broadband energy ratio for different spatial configurations. In this divided task, the ability to understand the less-intense talker dominated the pattern of performance, and performance improved as the energy ratio of the less-intense talker to the more-intense talker increased at the ear that had the more favorable energy ratio for the less-intense talker. Although listeners in the current study were not explicitly instructed as to which source to give higher priority, results suggest that listeners actively attended to the less-intense talker, which was usually harder to hear. As in the study by Best *et al.* (2006), this prioritization caused different effects of spatial separation on the lower- and higher-priority messages. Specifically, we found that spatial separation of the messages improved the ability to report the higher-priority message, but had little effect on the lower-priority message.

In the visual literature, three main models of spatial attention have been proposed. When extended to auditory tasks, these models predict that spatial separation will impair performance in a divided-listening task (cf. McMains and Somers, 2005). In the "zoom lens" model, the tuning of a single, spatial attentional filter widens in order to encompass spatially dispersed targets of interest, causing a trade-off between response accuracy and the size of the attentional field. The "multiple spotlights" model proposes simultaneous sampling of the auditory scene at several target locations, predicting a trade-off between processing efficiency and the total spatial extent of the attended regions. The "rapidly moving spotlight" model assumes that a single spotlight switches between spatially separated talkers, predicting that performance should degrade with increasing spatial separation of the targets.

The current results show that performance was better when the targets are spatially separated compared to when they are co-located, suggesting that these models of visual spatial attention cannot readily be applied to the current auditory divided-attention task. Of course, there are a number of differences in the demands of our auditory task and those of the visual tasks that typically are used to test models of dividing visual attention. For instance, by their very nature, auditory messages evolve over time, requiring listeners to sustain attention on a target message in order to analyze it and extract its meaning. The need to sustain attention on a message over time may make a strategy in which listeners switch attention between targets ineffective. Instead, the current results are consistent with listeners prioritizing the two targets differently, and processing them through different mechanisms.

While performance for the keywords of target$_F$ was essentially unaffected by the spatial configuration of the concurrent sources, performance for the actively attended target$_V$ was better in the spatially separated than in the co-located configurations. Most of the effects of spatial separation on performance for target$_V$ are consistent with the effects of spatial separation in selective listening (Arbogast *et al.* 2002; Ihlefeld and Shinn-Cunningham, submitted). However, no such effects occurred for target$_F$ (e.g., there was no reduction of target$_F$ drop errors when sources were spatially separated). Moreover, even at 0 dB $T_VT_FR_{be\text{-}V}$ where the two talkers should have been equally salient (albeit somewhat difficult to keep segregated), performance was slightly better for target$_V$ than for target$_F$ [cf. Fig. 3(e)]. We infer that listeners actively attended to target$_V$, and did so in part based on its fixed call sign (which was the only cue distinguishing target$_V$ from target$_F$ when sources were co-located and at the same level).

Current results show that as in selective listening, spatially separating the two targets improved the intelligibility of the actively attended message (target$_V$), presumably through some combination of acoustic improvements at the better ear for target$_V$, binaural processing benefits that improved the audibility of target$_V$ (e.g., see Zurek, 1993) and spatial attention benefits that allowed listeners to selectively attend to target$_V$ by directing attention to its location. In this task, where the ability to report target$_V$ determined overall performance, a strategy of actively attending to target$_V$ may have been near optimal, at least if listeners could not actively attend to both messages simultaneously. After the better-ear advantage for target$_V$ was taken into account, the dominant remaining spatial effect (ignoring report order; see Appendix) was that target$_V$ drop errors were less common for spatially separated than for co-located sources.

In contrast, with the exception of performance at 0 dB $T_VT_FR_{\text{be-V}}$, mix responses and combination errors did not vary with spatial separation for either target$_V$ or target$_F$. When both talkers were relatively easy to hear, spatial separation did not influence the ability to segregate the competing messages, except when spatial cues were the sole reliable feature for differentiating the two talkers. Informal listening suggested that for −20 dB T_VT_FR and greater, two distinct sources could be heard. However, we did not measure whether listeners heard the two target messages from two distinct locations. Therefore it is difficult to assess the extent to which listeners used spatial attention to perform the current task.

In order to perform this task, listeners needed to properly identify the two messages; it was not necessary to link each keyword to the proper source in order to have a trial scored as correct. However, percent correct performance in this divided listening task was nearly as good as performance in the companion selective listening task in which listeners were asked to report only one of the two messages (Ihlefeld and Shinn-Cunningham, 2008). This suggests that listeners were indeed able to link the keywords to distinct sources, but further studies are needed to gain a better understanding of how the ability to identify keywords and the ability to correctly pair a message with its source influence divided listening.

The relatively high incidence of drop errors at high $T_VT_FR_{\text{be-V}}$ suggests that the ability to track two simultaneous talkers was limited. However, overall performance in the divided task was surprisingly high compared to performance in many previous studies. Many researchers (Cherry, 1953; Broadbent, 1954; Moray, 1959; Treisman and Geffen, 1967; for a review see Stifelman, 1994) suggest that listeners are limited in their ability to report two or more simultaneous messages. For instance, although listeners can recall basic properties of a channel that is not actively attended (such as the sex of the talker), most of the target words from that channel cannot be reported correctly (e.g., Cherry, 1953; Treisman and Geffen, 1967). However, these previous studies investigated identification tasks with a relatively high processing load, such as asking listeners to shadow sustained messages (i.e., "Repeat what you hear in the right ear"). In a study that examined a detection task with a lighter processing load (using tones instead of word targets), listeners could detect targets equally well in attended and rejected channels (Lawson, 1966). The processing and memory load required for the highly predictable, relatively short CRM messages used in the current task may have been low enough that listeners could process and/or temporarily store the contents of both of the two simultaneous utterances.

At 0 dB $T_VT_FR_{\text{be-V}}$, subjects performed better for the keywords from target$_V$ than for the keywords from target$_F$, even though both talkers were equally intense and should have been equally intelligible. This suggests that listeners assigned higher processing priority to target$_V$. At least one previous study shows that the order of responses in a divided attention task reflects the priority that listeners give each target (Bonnel and Hafter, 1998). Examination of response order in Appendix A shows that on those trials where subjects reported all four keywords correctly, as $T_VT_FR_{\text{be-V}}$ increased subjects were increasingly likely to report keywords from target$_V$ first. In contrast, the percentage of responses in which listeners reported one target keyword from the variable-level talker and guessed at least one other word did not change systematically with $T_VT_FR_{\text{be-V}}$. In other words, response order did not just depend on $T_VT_FR_{\text{be-V}}$, but depended on whether listeners got all keywords correct, i.e., how well they extracted each of the two messages on a particular trial. In addition, when talkers were spatially separated, the report order was biased towards reporting the message from in front of the listener before the message from the side.

Overall, these results support the idea that response order depended on the relative certainty that the listener had about the two messages, with the listeners first reporting the message about which they were most sure. The relative certainty of the messages appears to depend on both the relative saliency of the two targets as well as the amount of attention that the listener devoted to a target. In turn, the inherent salience of the messages depended on (1) the audibility of the messages, (2) the relative intensities of the messages, and (3) the spatial locations of the messages (where messages from in front were inherently more salient). In summary, the results support the idea that subjects gave higher priority (and selectively attended) to target$_V$. However, when listeners tried but failed to understand target$_V$, they resorted to reporting target$_F$ first, and then reporting their best-guess response for target$_V$.

Together, these results suggest that listeners used two different processing strategies in monitoring the two concurrent targets. Spatial separation improved the ability to understand keywords from target$_V$, presumably because listeners actively tried to attend to target$_V$ and were more successful in performing this selective attention task when target$_V$ came from a different location than target$_F$. In contrast, performance for target$_F$ showed little effect of spatial separation, consistent with the idea that target$_F$ was recalled from a temporary storage that was at best weakly affected by the spatial configuration of the sources or by spatially directed attention.

V. CONCLUSIONS

In this divided listening task with two concurrent target messages, performance improved as the ratio of the broadband energy of a less-intense talker to the energy of a simultaneous fixed-level talker increased. Overall, listeners were relatively good at reporting the fixed-level talker, which was generally easy to hear.

Results are consistent with listeners actively attending to the harder-to-hear source (target$_V$), and then recalling target$_F$.

Overall performance (the probability of reporting all four keywords) improved with increasing spatial separation.

- After taking into account better ear effects for the high-priority target$_V$, overall performance depended primarily on whether the sources were co-located or separated.
- Improvements with spatial separation of the competing messages came about primarily through spatial gains in performance for the less-intense, high-priority target$_V$. Effects of spatial configuration on the low-priority target$_F$ were negligible.

Listeners naturally tended to report messages in proper pairings, even though they were not instructed to do so. Spatial separation of sources reduced the likelihood of confusing the two messages and reporting the keywords in inconsistent pairings. However, this benefit was very small and was only observed near 0 dB $T_VT_FR_{\text{be-V}}$, where listeners had few other cues to segregate the mixture.

ACKNOWLEDGMENTS

This work was supported in part by ONR and AFOSR. We thank Virginia Best, Gerald Kidd, Steve Colburn, and Chris Mason, Richard Freyman and two anonymous reviewers for helpful comments on earlier versions of this manuscript.

APPENDIX A: REPORT ORDER

In a companion study, when asked to report only the keywords from target$_V$ (ignoring the message from target$_F$) subjects performed nearly as well as they did here, when asked to report both messages. Together with the current results, this finding suggests that listeners had little difficulty reporting the usually more-intense target$_F$ in addition to target$_V$, and that the ability to report target$_V$ was the main factor limiting performance. Therefore, both saliency (i.e., the inherent, bottom-up strength of target$_V$ relative to target$_F$) and attention (i.e., the listener's ability to select target$_V$ from the mixture) should have influenced how well listeners perform in this task. The order in which subjects naturally choose to report the target keywords can reflect how they prioritize each target (Bonnel and Hafter, 1998). Therefore, results were analyzed *post-hoc* to examine whether there was a consistent pattern in the order in which listeners chose to report the color-number pairs.

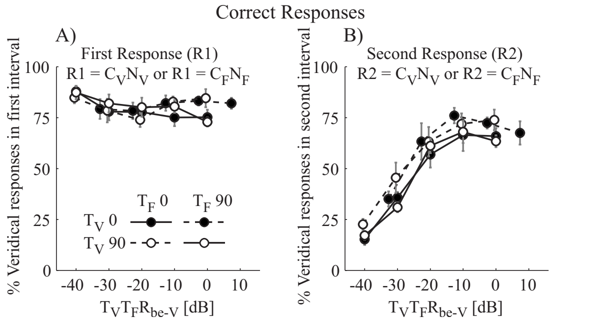

FIG. 4. Probability of reporting the first and second response pairs correctly in a proper pairing as a function of $T_VT_FR_{\text{be-V}}$, averaged across subjects. Error bars show the across-subject standard error of the mean. The probability of correct first responses is always greater than that of correct second responses. Subjects are more likely to respond without error in the second interval when the targets are spatially separated than when they are co-located. Filled symbols show results for target$_V$ at 0° and open symbols show results for target$_V$ at 90°. Results for spatially separated sources are shown with dashed lines and co-located sources are shown with solid lines. (A) First response interval. (B) Second response interval.

Figure 4(a) shows the percentage of trials in which the first color-number pair corresponded to the keywords from either one of the two messages (i.e., where subjects correctly reported either $[C_VN_V]$ or $[C_FN_F]$ as the first pair, ignoring the second response, which could be correct or wrong), as a function of $T_VT_FR_{\text{be-V}}$. Figure 4(b) shows the corresponding probability of a correct color-number pair being reported in the second pair (i.e., either $[C_VN_V]$ or $[C_FN_F]$), ignoring responses in the first color-number pair).

Figure 4(a) shows that for all spatial configurations, subjects responded without error in the first interval in 80% or more of the trials. The likelihood that the first pair was correct was very similar for all spatial configurations. However, in the spatially co-located configurations, the percentage of those first-pair responses that were correct decreased slightly with increasing $T_VT_FR_{\text{be-V}}$ (consistent with subjects confusing the two target messages when they were both at the same level and from the same location), whereas this probability did not change with $T_VT_FR_{\text{be-V}}$ in the two spatially separated configurations.

Figure 4(b) shows that for all spatial configurations, the percentage of correct second-pair responses increased with increasing $T_VT_FR_{\text{be-V}}$. Moreover, subjects were more likely to respond without error in the second interval when the two talkers were spatially separated than when they are co-located [dashed lines are above solid lines in Fig. 4(b)].

Comparing results of Figs. 4(a) and 4(b), the probability of a correct first-pair response was much greater than the probability of a correct second-pair response for all conditions, indicating that subjects tended to respond first with a color-number pair that they were more sure was correct (though this was not the only criterion, as target$_V$ influenced the report order too; see analysis below).

All of the first-pair responses (both correct and incorrect) were broken down into six possible response types, depending on whether subjects reported both the color and number of target$_V$ ($[C_VN_V]$, a correct response on the first pair), both color and number of target$_F$ ($[C_FN_F]$; another form of correct response on the first pair), a mix of keywords from both targets ($[C_VN_F]$ or $[C_FN_V]$; a mix response), one keyword from target$_V$ and one word that was not from either talker ($[C_VN_X]$ or $[C_XN_V]$; a form of drop error), or two

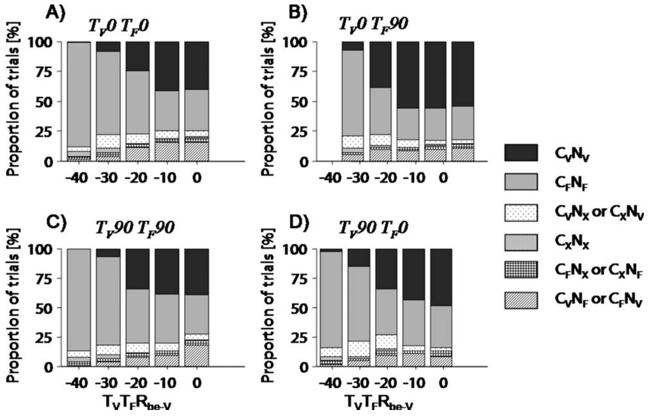

FIG. 5. Analysis of the first-pair responses as a function of $T_VT_FR_{\text{be-V}}$ for each spatial configuration, averaged across subjects. As target$_V$ becomes increasingly more audible, subjects become more likely to report it first. Different fill patterns denote different errors. $[C_VN_V]$ responses are solid black, $[C_FN_F]$ solid gray. target$_V$ guesses ($[C_VN_X]$ and $[C_XN_V]$) are represented by sparsely dotted fill and completely random guesses ($[C_XN_X]$) by densely dotted fill. $[C_FN_X]$ and $[C_XN_F]$ are denoted by square-grid fill. $[C_VN_F]$ and $[C_FN_V]$ are represented by rightward diagonal hatches. Each panel shows one spatial configuration. The left two panels (A, C) show results when the targets are coming from the same location. The right two panels (B, D) show results when the targets are spatially separated. The top row (A, B) shows results for the two configurations with target$_V$ in front of the listener. The bottom row (C, D) shows results when target$_V$ is to the side of the listener.

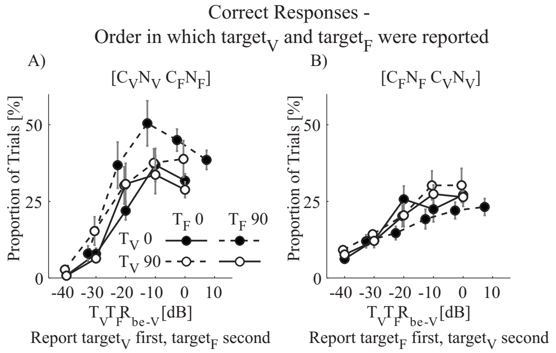

FIG. 6. Order in which listeners reported properly streamed keywords of target$_V$ and target$_F$ in the fully correct trials as a function of $T_VT_FR_{\text{be-V}}$. Error bars show the across-subject standard error of the mean. Results suggest that report order is a measure of the relative certainty the listener has about the content of the two messages. The absolute spatial configuration affected report order, suggesting that the source in front of the listeners was inherently more salient than source to the side of the listener. Filled symbols show results for the target at 0° and open symbols for the target at 90°. Results for spatially separated sources are shown with dashed lines and for co-located sources with solid lines. (A) First response is target$_V$ and second response is target$_F$. (B) First response is target$_F$, and second response is target$_V$.

words that were not part of either target ($[C_XN_X]$; another form of drop error). All first responses that did not fit any of these criteria were scored as other responses ($[C_FN_X]$ or $[C_XN_F]$), but such responses were rare. Note that the probabilities of responding ($[C_VN_V]$), ($[C_FN_F]$), ($[C_VN_X]$ or $[C_XN_V]$), $[C_XN_X]$, ($[C_VN_F]$ or $[C_FN_V]$), and ($[C_FN_X]$ or $[C_XN_F]$) sum to 1.0.

Figure 5 shows the distribution of the first responses as a function of $T_VT_FR_{\text{be-V}}$ for each spatial configuration. For $T_VT_FR_{\text{be-V}}$ of −20 dB and below, $[C_FN_F]$ was the dominant response type (gray solid fill). For $T_VT_FR_{\text{be-V}}$ greater than −20 dB, the most common first-pair response was $[C_VN_V]$ (black solid fill). This shows that as target$_V$ became louder and easier to hear, subjects became more and more likely to report it first. The proportion of trials in which subjects heard only part of target$_V$ (i.e., reported one keyword from target$_V$ and guessed the other word, $[C_VN_X]$ or $[C_XN_V]$, shown as sparsely dotted fill) was small and did not change systematically with $T_VT_FR_{\text{be-V}}$ (compare size of sparsely dotted-fill areas from left to right in each panel). This suggests that when listeners were not sure of the content of target$_V$, they tended to report it second, rather than first.

The percentage of trials in which listeners intermingled keywords from both talkers (reported $[C_VN_F]$ or $[C_FN_V]$) increased with increasing $T_VT_FR_{\text{be-V}}$ (see rightward diagonal hatch areas in Fig. 5), especially for the co-located spatial configurations (panels A and C). This increase in mix responses in the first responses was consistent with the overall pattern of mix responses (cf. Sec. III A 3). Completely random drop errors in the first response (reporting $[C_XN_X]$) only occurred at the lowest $T_VT_FR_{\text{be-V}}$ (densely dotted fill), and were very unlikely compared to the other responses. Similarly, other errors did not occur often and did not change consistently with either $T_VT_FR_{\text{be-V}}$ or spatial configuration (square-grid fill).

The ways in which subjects ordered and paired responses on the subset of trials when they were fully correct was analyzed to see how listeners naturally grouped the keywords, conditioned on them being fully correct. Figure 6(a) shows the percentage of correct trials in which subjects first reported target$_V$ and then target$_F$ ($[C_VN_VC_FN_F]$). Figure 6(b) shows the percentage of correct trials in which subjects first reported target$_F$ and then target$_V$ ($[C_FN_FC_VN_V]$). In both panels, performance is plotted as a function of $T_VT_FR_{\text{be-V}}$ for the four different spatial configurations.

For both report orders, the percentage of fully correct trials increased with increasing $T_VT_FR_{\text{be-V}}$. For $T_VT_FR_{\text{be-V}}$ less than −20 dB, subjects were more likely to report target$_F$ before target$_V$ [plotted percentages are higher in Fig. 6(b) than in Fig. 6(a)]. For $T_VT_FR_{\text{be-V}}$ of −20 dB and greater, subjects were most likely to report target$_V$ first [plotted percentages are higher in Fig. 6(a) than in Fig. 6(b)]. Interestingly, there were differences in these likelihoods that depended on the absolute locations of the talkers: Subjects were more likely to report keywords from target$_V$ first when target$_V$ was in front and target$_F$ was to the side than when target$_V$ was to the side and target$_F$ was in front [in Fig. 6(a), the dashed line with filled symbols is above the other lines], even after taking into account the talker energy ratios at the better ear for target$_V$ (or target$_F$, see[2]). This trend reverses for trials in which listeners first reported target$_F$ and then reported target$_V$. In those trials, the percentage of correct reports was greater when target$_F$ was from in front of the listener and target$_V$ was to the side than in the reverse configuration [in Fig. 6(b), the dashed line with filled symbols is below the other lines].

Overall, these results suggest that the listeners were actively attending to target$_V$, but tended to report the message

TABLE III. Mean parameters of the psychometric function fits for the different spatial configurations, averaged across subjects (across-subject standard error of the mean is shown in round brackets). The midpoint parameters are greater for co-located than for spatially separated sources; the upper bound of performance is higher for spatially separated sources than for co-located sources; no other differences are significant. (A) Estimates of α, the TMR at the midpoint of the dynamic range in the psychometric function. (B) Estimates of $1/\beta$, the slope of the psychometric function at the midpoint of the dynamic range. (C) Estimates of $1-\lambda$, the upper asymptote of the functions.

	T_v0T_f0	T_v90T_f90	T_v0T_f90	T_v90T_f90
(A) Midpoint of dynamic range α [dB]	27.3 (2.2)	27.2 (2.9)	25.2 (3.1)	25.1 (3.2)
(B) Slope at the midpoint of dynamic range $1/\beta$ [% correct/dB]	19.2 (5.4)	20.4 (7.6)	19.5 (5.0)	16.0 (1.6)
(C) Upper asymptote of performance $1-\lambda$ [% correct]	85.6 (6.2)	85.4 (5.8)	91.6 (6.0)	92.5 (4.3)

that they were most sure of first. The effect of the absolute locations of the talkers on report order suggests that a message from in front of the listener was more salient (and that listeners were therefore more sure of its content) than a message from the side of the listener. Note that this was the only aspect of performance for which the absolute locations of the talkers mattered (after accounting for the acoustic effects of the better ear for target$_V$); all other effects of spatial configuration depended only on whether the talkers were spatially separated or co-located.

We conclude that at least three factors affected the relative certainty listeners had about the content of the competing messages: listeners were actively trying to attend to target$_V$, which enhanced the neural representation of target$_V$ (when listeners were successful at hearing target$_V$). However, the ability to hear target$_V$ depended directly on $T_VT_FR_{be\text{-}V}$. On top of both of these effects, the source from in front of the listener appeared to be inherently more salient than the other source, which caused an asymmetry in report orders for the two spatially separated configurations.

In this task, listeners were not instructed to report the two messages in any particular order, and were not penalized if they incorrectly paired keywords from the two competing messages. Despite this, report order depended systematically on $T_VT_FR_{be\text{-}V}$, on whether the messages were spatially separated or co-located, and on the absolute spatial configuration of the sources. The consistency of these effects, even without any explicit instruction to the subjects, suggests that listeners naturally adopted a strategy in this divided attention task in which they gave top priority to the usually harder-to-hear variable-level target over the fixed-level target.

APPENDIX B: FITS TO PSYCHOMETRIC FUNCTIONS

Psychometric functions were fit to the percent correct scores as a function of T_VT_FR for each subject and condition (Wichmann and Hill, 2001a; see also Ihlefeld and Shinn-Cunningham, 2008). The estimated probability of responding correctly at a given T_VT_FR, $\hat{P}(x)$ was fit as

$$\hat{P}(x) = \gamma + (1-\lambda-\gamma)\frac{1}{1+e^{\alpha-x/\beta}}, \tag{B1}$$

where γ is the lower bound on performance (chance performance, set to 6%), $1-\lambda$ is the upper bound on performance at the largest T_VT_FR, α is the energy ratio at which percent correct performance is halfway between chance and asymptotic performance, and $1/\beta$ is the slope of the psychometric function evaluated at $x=\alpha$.

The goodness of fit of the psychometric functions was evaluated with a deviance criterion that was derived using Efron's bootstrap technique (Wichmann and Hill 2001a, Wichmann and Hill, 2001b). Fourteen of the 16 fits (four functions for each of four listeners) meet the 95% confidence interval deviance criterion. The relatively poor data fit in the other two cases was not caused by outliers (subjectively, even these fits summarized the results adequately).

The upper bound parameter $1-\lambda$ and the slope parameter $1/\beta$ were fitted to maximize the likelihood of observing the actual data given the psychometric function, using the psignifit toolbox in MATLAB 6.5. The resulting parameters, averaged across subjects, are shown in Table III. *T*-tests were employed to test for differences between the within-subject averages of the two spatially co-located configurations and the two spatially separated configurations. The midpoint parameter α of the psychometric function was significantly larger in the spatially co-located than in the spatially separated configurations (*t*-test; $p<0.01$). The slopes at the midpoints of the psychometric functions, $1/\beta$, did not vary significantly with spatial configuration (*t*-test; $p<0.01$). The upper bounds $1-\lambda$ were significantly lower in the co-located than in the separated configurations (*t*-test; $p<0.01$), reflecting the lower level of performance for co-located configurations at the greatest $T_VT_FR_{be\text{-}V}$.

[1] Considering that (1) listeners were not instructed to report keywords in proper pairing and that (2) listeners also received correct feedback for mix responses, mix responses could have resulted from a response strategy whereby listeners did not attempt to report keywords in proper pairings. However, given that listeners had a strong natural tendency to report the keywords in proper pairings, this is not a very likely explanation for the occurrence of mix responses (see also Appendix A).

[2] Responses were also analyzed as a function of the better ear for target$_F$ ($T_VT_FR_{be\text{-}F}$; results not shown here). However, this analysis does not reveal any consistent pattern in the data that could help in explaining any of the effects of spatial separation in the data. In fact, when plotted as a function of $T_VT_FR_{be\text{-}F}$, the performance curves for the spatially separated configurations end up being shifted away from each other, seemingly increasing the difference between spatially separated configurations.

Arbogast, T. L., and Kidd, Jr., G. (**2000**). "Evidence for spatial tuning in informational masking using the probe-signal method," J. Acoust. Soc. Am. **108**, 1803–1810.

Arbogast, T. L., Mason, C. R., and Kidd, G., Jr. (**2002**). "The effect of spatial separation on informational and energetic masking of speech," J. Acoust. Soc. Am. **112**, 2086–2098.

Best, V., Gallun, F. J., Ihlefeld, A., and Shinn-Cunningham, B. G. (**2006**). "The influence of spatial separation on divided listening," J. Acoust. Soc. Am. **120**, 1506–1516.

Best, V., Ozmeral, E., Gallun, F. J., Sen, K., and Shinn-Cunningham, B. G. (**2005**). "Spatial unmasking of birdsong in human listeners: Energetic and informational factors," J. Acoust. Soc. Am. **118**, 3766–3733.

Bolia, R. S., Nelson, W. T., and Ericson, M. A. (**2000**). "A speech corpus for multitalker communications research," J. Acoust. Soc. Am. **107**, 1065–1066.

Bonnel, A., and Hafter, E. (**1998**). "Divided attention between simultaneous auditory and visual signals," Percept. Psychophys. **60**, 179–190.

Broadbent, D. (**1954**). "The role of auditory localization in attention and memory span," J. Exp. Psychol. **47**, 191–196.

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

Brungart, D. S., Simpson, B. D., Ericson, M. A., and Scott, K. R. (**2001**). "Informational and energetic masking effects in the perception of multiple simultaneous talkers," J. Acoust. Soc. Am. **110**, 2527–2538.

Brungart, D. S., and Simpson, B. D. (**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.

Brungart, D., Simpson, B., Darwin, C., Arbogast, T., and Kidd, G. J. (**2005**). "Across-ear interference from parametrically degraded synthetic speech signals in a dichotic cocktail-party listening task," J. Acoust. Soc. Am. **117**, 292–304.

Cherry, E. C. (**1953**). "Some experiments on the recognition of speech, with one and with two ears," J. Acoust. Soc. Am. **25**, 975–979.

Conway, A. R., Cowan, N., and Bunting, M. F. (**2001**). "The cocktail party phenomenon revisited: The importance of working memory capacity," Psychon. Bull. Rev. **8**, 331–335.

Cowan, N. (**1995**). *Attention and Memory: An Integrated Framework* (Oxford University Press).

Dorman, M., Loizou, P., and Rainey, D. (**1997**). "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs," J. Acoust. Soc. Am. **102**, 2403–2411.

Durlach, N. I., Mason, C. R., Kidd, G., Jr., Arbogast, T. L., Colburn, H. S., and Shinn-Cunningham, B. G. (**2003**). "Note on informational masking," J. Acoust. Soc. Am. **113**, 2984–2987.

Ebata, M., Sone, T., and Nimura, T. (**1968**). "Improvement of hearing ability by directional information," J. Acoust. Soc. Am. **43**, 289–297.

Freyman, R., Helfer, K., and Balakrishnan, U. (**2005**). "Spatial and spectral factors in release from informational masking in speech recognition," Acta Acust. **91**, 537–545.

Freyman, R. L., Helfer, K. S., McCall, D. D., and Clifton, R. K. (**1999**). "The role of perceived spatial separation in the unmasking of speech," J. Acoust. Soc. Am. **106**, 3578–3588.

Ihlefeld, A., and Shinn-Cunningham, B. G. (**2008**). "Spatial release from energetic and informational masking in a selective speech identification task." J. Acoust. Soc. Am. **123**, 4369–4379.

Kidd, G. J., Arbogast, T., Mason, C., and Gallun, F. (**2005**). "The advantage of knowing where to listen," J. Acoust. Soc. Am. **118**, 3804–3815.

Lawson, E. A. (**1966**). "Decisions concerning the rejected channel," Q. J. Exp. Psychol. **18**, 260–265.

Lutfi, R. A., Kistler, D. J., Callahan, M. R., and Wightman, F. L. (**2003**). "Psychometric functions for informational masking," J. Acoust. Soc. Am. **114**, 3273–3282.

McMains, S., and Somers, C. (**2005**). "Processing efficiency of divided spatial attention mechanisms in human visual cortex," J. Neurosci. **25**, 9444–9448.

Moray, N. (**1959**). "Attention in Dichotic Listening: Affective cues and the influence of instructions," Q. J. Exp. Psychol. **11**, 56–60.

Rivenez, M., Darwin, C. J., and Guillaume, A. (**2006**). "Processing unattended speech," J. Acoust. Soc. Am. **119**, 4027–4040.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.

Shinn-Cunningham, B. G., Ihlefeld, A., Satyavarta, and Larson, E. (**2005**). "Bottom-up and top-down influences on spatial unmasking," Acta Acust. **91**, 967–979.

Spieth, W., Curtis, J., and Webster, J. (**1953**). "Responding to one of two simultaneous messages," J. Acoust. Soc. Am. **26**, 391–396.

Stifelman, L. (**1994**). "The cocktail party effect in auditory interfaces: A study of simultaneous presentation," MIT Media Laboratory Technical Report.

Treisman, A., and Geffen, G. (**1967**). "Selective attention: Perception or response?" Q. J. Exp. Psychol. **19**, 1–17.

Watson, C. (**2005**). "Some comments on informational masking," Acta Acust. **91**, 502–512.

Wichmann, F., and Hill, N. (**2001a**). "The psychometric function: I. Fitting, sampling and goodness-of-fit," Percept. Psychophys. **63**, 1293–1313.

Wichmann, F., and Hill, N. (**2001b**). "The psychometric function: II. Bootstrap-based confidence intervals and sampling," Percept. Psychophys. **63**, 1314–1329.

Yost, W. A. (**1997**). "The cocktail party problem: Forty years later," in *Binaural and Spatial Hearing in Real and Virtual Environments*, edited by R. Gilkey and T. anderson (Erlbaum, New York), pp. 329–348.

Zurek, P. M. (**1993**). "Binaural advantages and directional effects in speech intelligibility," in *Acoustical Factors Affecting Hearing Aid Performance*, edited by G. Studebaker and I. Hochberg (College-Hill Press, Boston, MA).