Stimulus factors influencing spatial release from speech-on-speech masking

Gerald Kidd, Jr., Christine R. Mason, Virginia Best, and Nicole Marrone[a)]
Department of Speech, Language and Hearing Sciences, and Hearing Research Center, Boston University, Boston, Massachusetts 02215

(Received 24 September 2009; revised 14 July 2010; accepted 17 July 2010)

This study examined spatial release from masking (SRM) when a target talker was masked by competing talkers or by other types of sounds. The focus was on the role of interaural time differences (ITDs) and time-varying interaural level differences (ILDs) under conditions varying in the strength of informational masking (IM). In the first experiment, a target talker was masked by two other talkers that were either colocated with the target or were symmetrically spatially separated from the target with the stimuli presented through loudspeakers. The sounds were filtered into different frequency regions to restrict the available interaural cues. The largest SRM occurred for the broadband condition followed by a low-pass condition. However, even the highest frequency bandpass-filtered condition (3–6 kHz) yielded a significant SRM. In the second experiment the stimuli were presented via earphones. The listeners identified the speech of a target talker masked by one or two other talkers or noises when the maskers were colocated with the target or were perceptually separated by ITDs. The results revealed a complex pattern of masking in which the factors affecting performance in colocated and spatially separated conditions are to a large degree independent. © *2010 Acoustical Society of America.* [DOI: 10.1121/1.3478781]

PACS number(s): 43.66.Lj, 43.66.Dc, 43.66.Pn [MAA] Pages: 1965–1978

I. INTRODUCTION

There is a long history of study of the beneficial effects of spatially separating a target talker from one or more masking sounds. Among the earliest studies were the reports by Licklider (1948) revealing that the interaural phase relations of target and masker played through earphones were crucial to performance, and by Hirsh (1950) who showed that the spatial location of target and masker presented via loudspeakers could significantly affect speech intelligibility. Kock (1950) also demonstrated large detectability advantages when a speech target was varied in horizontal azimuth relative to a noise masker. These early studies were compelling examples of how binaural differences—termed interaural time (ITD) and interaural level (ILD) differences-could be used by the listener to extract one sound source from among other competing sound sources.

Like much of the work of that era examining binaural hearing, those early reports typically used Gaussian noise as the masker (however, see discussion of "subjective" impressions of binaural hearing by Koenig, 1950). Although Gaussian noise, broadband or band limited, is easy to quantify and the interpretation of its masking effects is fairly straightforward, it is rarely representative of the competition among sources in realistic communication situations. Often the difficulty experienced by a listener in selecting one sound source among multiple sound sources is related to aspects of the masking sounds that are not captured by noise. In particular, when the target is speech from one talker, the speech of other talkers may significantly interfere with comprehending the information from the target (although such interference is of course not limited to speech sounds, but occurs for other combinations of similar target(s) and masker(s) cf. Kidd *et al.*, 1998; Best *et al.*, 2005). This difference was recognized in studies by Pollack and Pickett (1958), Shubert and Schultz (1962), and Carhart *et al.* (1969), among others. The effects of noise masking were often explained as limitations due to peripheral overlap of excitation (e.g., Fletcher, 1940). However, the interpretation of listener performance in complex multitalker environments is also thought to involve less well understood processes such as sound source segregation (cf. Bregman, 1990) and central factors such as attention and memory (see classic references by Cherry, 1953, and Broadbent, 1958, Ch. 2; as well as the recent chapter on informational masking by Kidd *et al.*, 2008a).

In a recent study, Marrone *et al.* (2008a) reported a spatial release from masking (SRM) of about 13 dB for a target talker masked by two independent masker talkers. The SRM was computed as the difference between thresholds (target-to-masker ratio, T/M, for 50% correct identification) in dB for conditions in which the sounds were colocated or spatially separated. This large effect occurred despite the reduction in the usefulness of simple acoustic head shadow differences which were minimized by symmetric placement in azimuth of the two maskers. The term "simple" here means that the magnitude of head shadow is computed over a relatively long integration time. Although there was no SRM found in a monaural condition, which presumably provides a control for better ear listening, Marrone *et al.* (2008a) speculated that short-term differences in T/M in one ear or the other (i.e., time-varying ILDs or "moments of better-ear advantage") might still aid in segregating the sound sources, potentially contributing to the overall SRM observed. Be-

[a)]Present address: Department of Communication Sciences and Disorders, Northwestern University, Evanston, Illinois.

cause these "glimpses" of the stimuli are inherently binaural in nature, their putative effects would not be expected to be revealed when listening monaurally.

The magnitude of SRM found in various experiments must be interpreted cautiously. Although binaural differences clearly are the basis for SRM, the underlying mechanisms exploiting ITDs and ILDs may be different in different circumstances. The early binaural studies referenced above were strongly influenced by two empirical results that appeared to be directly related: the demonstration of large advantages in the detectability of a pure-tone signal in noise, and large advantages in speech recognition in noise, when the stimuli were presented dichotically, both referenced to performance when all of the stimuli were diotic or monotic. The dichotic condition causing the largest advantage was usually for the target presented π radians out of phase with an in-phase masker. In fact, the "masking-level difference" (MLD) for detection has been invoked as an explanation for the dichotic advantage in speech recognition for speech targets in noise presented either under headphones (e.g., Levitt and Rabiner, 1967) or via loudspeakers (e.g., Zurek, 1993). According to that interpretation, the MLD may be viewed simply as an improvement in T/M in the various frequency channels without necessarily specifying the underlying mechanism (e.g., an equalization-cancellation mechanism, Durlach, 1963, 1972; Akeroyd, 2004; Culling, 2007; or a binaural decorrelation mechanism, e.g., Culling and Colburn, 2000, Culling *et al.*, 2006).

The relationship between the MLD for detection and the binaural advantage for speech recognition is less clear when the masker(s) are speech rather than noise. This is due in large part to the understanding that energetic masking (EM)—within-channel competition for an adequate neural representation of the target in the presence of the masker(s)—is much less of a factor for speech than for noise maskers (e.g., Carhart *et al.*, 1969; Freyman *et al.*, 2001; Noble and Perrett, 2002; Arbogast *et al.* 2002; Brungart *et al.*, 2006). The SRM observed in speech-on-speech masking, in fact, often appears to be governed more by perceptual and cognitive factors causing release from informational masking (IM) than by lower-level MLD mechanisms (cf. Kidd *et al.*, 2008a). This conclusion has some important implications for the interpretation of SRM in multitalker environments

In the study by Marrone *et al.* (2008a), there were three simultaneous talkers of the same sex uttering similar sentences, a situation thought to be dominated by IM. When all three—the target and two maskers—were colocated, threshold T/Ms were approximately 2–3 dB, a finding that appears to be fairly consistent across various studies using similar stimuli and/or methods (cf., Brungart *et al.*, 2001; Marrone *et al.*, 2008b; Carr, 2010). When the maskers were symmetrically spatially separated from the target (at $\pm 90°$), threshold T/Ms were much lower, about -11 dB, resulting in the large SRM. However, another type of release from masking was observed when the speech of the masker talkers was time-reversed. In that case, threshold T/Ms in the *colocated* condition were nearly 12 dB lower than for forward speech (cf. Freyman *et al.*, 2001; Rhebergen *et al.*, 2005). In the spatially separated condition, the threshold T/Ms were lower for reversed vs. forward speech as well, but only by about 5 dB. If one considers *only* the magnitude of the SRM, the conclusion from Marrone *et al.* would be that SRM is much smaller for reversed speech than for forward speech. However, Marrone *et al.* speculated that once the IM had been reduced (by time-reversal or spatial separation) there was little IM left to reduce via other means (cf. Freyman *et al.*, 2001). Hence, there was relatively little spatial benefit seen for time-reversed maskers and conversely relatively little benefit of time-reversal for spatially separated maskers. Presumably other source segregation manipulations such as using different sex voices for the target and masker talkers would have a similar effect (cf. Freyman *et al.*, 1999; Noble and Perrett, 2002; Allen *et al.*, 2008; Carr, 2010); that is, a release from IM would be seen for the colocated maskers with little additional release observed with spatial separation. Because of these differences in the cues available to segregate sources that are colocated, not all speech-on-speech masking situations will produce large SRM and, in fact, SRM may be reduced under the very conditions producing the best overall performance.

The following experiments were intended to examine further the factors contributing to spatial release from masking in the three-talker listening condition investigated by Marrone *et al.* (2008a). In Experiment 1, the contribution of ITD and ILD cues was investigated by filtering the stimuli into different frequency regions. In Experiment 2, the influence of the number and type of maskers was studied under various presentation conditions in which the stimuli were separated only by ITD cues.

II. EXPERIMENT 1

In order to determine the contribution of ITDs and ILDs to the SRM seen in the Marrone *et al.* (2008a) paradigm, filtered stimuli were used. Because ITDs and ILDs are known to be most effective in different frequency regions (cf. Zurek, 1993), filtering the stimuli is a useful approach to determining the effectiveness of these cues. It was expected that low-pass filtering of targets and maskers would reveal the extent to which ITD cues were useful, high-pass filtering would limit the cues to those arising primarily from ILDs, including acoustic head shadow or better ear effects, and stimuli limited to a midrange of frequencies, where neither cue is particularly strong, may or may not show any benefit of spatial separation. In addition, because identical filtering of target and maskers could increase the perceived similarity of the sounds, potentially making the task more difficult and influenced more by IM, the experiment was repeated with stimuli in which only the target was filtered. Because high-frequency better-ear cues were limited by the symmetric masker placement, it was expected that only those conditions in which the critical information was carried by low frequencies (e.g., below about 1500 Hz) would show a significant SRM (cf. Dubno *et al.*, 2002). However, that expectation was tempered by the limited data available for speech masking other speech under conditions in which the stimuli were filtered.

A. Methods

The stimuli and procedures were similar to those employed by Marrone *et al.* (2008a). The sounds were presented via loudspeakers in a large (approximately 12 ft, 4 in. long, 13 ft wide, 7 ft, 6 in. high) custom IAC (Industrial Acoustics Co.) booth with the typical perforated metal ceiling and wall panels and a carpeted floor. The listener was seated in the center of a semicircle of loudspeakers (from left to right in front) positioned 5 ft away. Subjectively, the room has little noticeable reverberation. The direct-to-reverberant ratio (estimated using impulse responses) averaged 6.3 dB for the source positions at 0° and 90° azimuth, and modulation transfer functions (measured according to RASTI procedures) indicated no decrease in modulation depth and hence no predicted decrease in speech intelligibility. The details of these measurements were provided in Kidd *et al.* (2005).

The task was closed set speech identification using the coordinate response measure corpus ("CRM," Bolia *et al.*, 2000) and the target and both maskers were independent sentences from that corpus spoken by different female talkers. The test has the structure "Ready [callsign] go to [color] [number] now." There are eight callsigns, four colors, and eight numbers; and on any given trial the callsigns, colors, and numbers spoken by the three talkers (target and two maskers) were mutually exclusive and chosen randomly. In these experiments the target sentence was denoted by the use of the callsign "Baron" on every trial and the listeners were instructed to report the color and number spoken by the target talker. The response was counted correct only if both the color and the number from the target sentence were reported accurately. The target level was fixed at 50 dB SPL while a 1-up 1-down adaptive procedure was used to estimate the masker level corresponding to 50% correct identification. The two masker sentences were always equal to each other in level. The results are expressed as target-to-masker ratio at threshold (T/M) in dB. A T/M of 0 dB would indicate that the target sentence was at the same level as *each* individual masker (and therefore approximately 3 dB lower than the sum of the level of the two maskers). Both target and masker sentences were filtered identically to create four different conditions that differed in frequency range: "broadband" which was from 0 to 6 kHz; "low-frequency," which was low-pass filtered at 1.5 kHz; "mid-frequency," which was bandpass filtered from 1.5–3 kHz; and "high-frequency," which was bandpass filtered from 3–6 kHz.

In the first part of the experiment, a subset of spatial conditions from Marrone *et al.* (2008a) was tested. In that study, the results for maskers placed symmetrically at ±45° did not differ from the results for the ±90° condition and therefore the ±45° condition was not tested here. That left three spatial conditions: one in which the target and both maskers were colocated at 0° azimuth, and two in which the target was presented at 0° and the maskers were symmetrically placed to the left and right of the listener at either ±15° or ±90°.

The listeners were five young adults with normal hearing as determined by standard audiometry. They completed two or three blocks of target-only identification (at the fixed level used for the target) for each filter condition, two or three adaptive threshold estimates for the target in quiet for each filter condition, and six adaptive threshold estimates for each filter condition and spatial configuration in the masked conditions. In all conditions, speech identification performance was near 100% correct in the absence of the maskers. The listeners participated for three 2-h sessions with each session conducted on a different day.

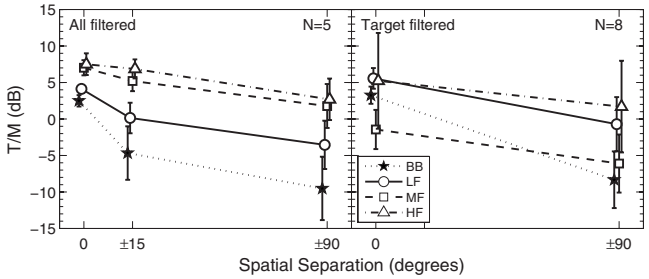

FIG. 1. The results from Experiment 1 expressed as target-to-masker ratio (T/M) in dB. The data are group means and associated standard deviations for the target-masker spatial separations that were tested. The left panel shows the results when target and maskers were filtered equally. The right panel shows the results when the maskers were broadband and only the target was filtered. In both panels, the data for different frequency regions are indicated by the different symbols. The data points are offset slightly along the abscissa and connected by lines for clarity of presentation.

In the second part of this experiment, two of the same listeners and six additional normal-hearing young adults served as subjects. The speech maskers were always broadband while the target was filtered into the same four frequency regions described above. The broadband condition thus represents a replication of the broadband condition from the first part of the experiment. In addition, because the intent of this manipulation was to compare the amount of spatial release from masking, only the colocated and ±90° separation spatial conditions were tested. Otherwise, the procedures were identical to those previously described. This part of the experiment required two listening sessions.

B. Results

Figure 1 shows the group-mean threshold T/Ms in dB for each of the spatial configurations and filter conditions. The left panel displays the results when all three sentences (target and both maskers) were filtered identically ("all filtered") while the right panel displays the results when only the target was filtered ("target filtered"). First, considering the results contained in the left panel, the T/M at threshold in the colocated broadband condition (asterisks; about 2.6 dB) was very similar to that reported previously by Marrone *et al.* (2008a) for nearly identical conditions. Filtering the stimuli raised the thresholds for the colocated condition to about 7 dB T/M for both the mid-(squares) and high-frequency (triangles) conditions. The threshold for the low-pass filtered condition (circles) was intermediate at around 4 dB. Thus, when target and maskers were colocated, threshold T/Ms varied over a range of about 5 dB as a result of filtering.

For the spatially separated condition, the threshold T/Ms were substantially lower in some cases, particularly when the

stimuli were broadband or low-pass filtered. As in the Marrone *et al.* study, there was a substantial benefit from a spatial separation of only ±15° for these two cases, suggesting fairly sharp spatial "tuning" for these stimuli and procedures. For the ±90° separation, threshold T/Ms ranged from about −10 dB for the broadband condition to about 3 dB for the high-frequency condition. Thus, thresholds varied considerably more in the spatially separated condition than in the colocated condition. A repeated-measures analysis of variance indicated that both frequency range [$F(3,12)=211.2$, $p<0.001$] and spatial condition [$F(2,8)=40.1$, $p<0.003$] were significant main factors. Furthermore, as would be expected from inspection of the left panel of Fig. 1, the interaction was also significant [$F(6,24)=12.2$, $p=0.025$].

The difference in performance between individual listeners was in some cases substantial as indicated, for example, by the error bars for the broadband condition (asterisks). The performance across listeners for the colocated case varied by only about 2 dB (a range of thresholds of only 1.4 to 3.3 dB) whereas their performance for the ±90° case varied over almost 10 dB (−14.5 to −4.7). A similar trend was seen for the other filter conditions.

In the right panel, displaying the results from the target-filtered conditions, the pattern of results was very similar to that which occurred when all of the stimuli were filtered, except for the mid-frequency (1.5–3 kHz) condition (squares). For some reason that is not clear to us, threshold T/Ms were lower by several dB for both colocated and spatially separated conditions compared to the same frequency range condition in the left panel. This finding needs further evaluation before any firm conclusions may be drawn. Otherwise, the results from the two conditions (left and right panels of Fig. 1) were nearly identical. It should be pointed out that, as in the first part of the experiment, these T/Ms are defined by the overall level of both target and individual maskers, and on this basis they look quite similar. It appears that these trends due to spatial separation and frequency range may occur when only the target is restricted in frequency content. Thus, with the possible exception of the mid-frequency result, the qualitative differences in the sound of the target and masker due to this manipulation did not lead to substantially different findings. An alternative way to represent the T/Ms for the results plotted in the right panel of Fig. 1 would be to reference the target to the masker level only in the frequency region matching the target band. Based on that computation the thresholds in the broadband condition would not change, the low-frequency thresholds would increase by only about 0.4 dB (because most of the masker energy is already in this region), the mid-frequency thresholds would increase by about 11.8 dB, and the high-frequency thresholds would increase by about 20 dB. These adjustments might be warranted if an attempt was made to factor in the influence of the greater amount of EM that would likely result from the additional masker energy outside of the target frequency region relative to the case of both target and masker filtered identically.

Individual differences were noteworthy here as well. The high-frequency condition produced the largest intersubject variability for both colocated and separated cases of any of the conditions shown in either panel. The thresholds comprising these two points varied across listeners by nearly 20 dB. Two of the eight listeners had threshold T/Ms greater than 10 dB (computed re. the entire masker bandwidth) in the colocated condition (a value at least 6 dB higher than any of the other listeners), while one listener had an unusually low threshold T/M of −11.6 dB for the ±90° spatial separation—a value that was nearly 10 dB better than the next best threshold.

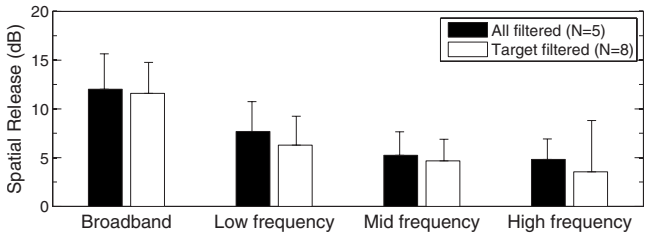

FIG. 2. The amount of "spatial release from masking" (SRM) computed from the values plotted in Fig. 1. The abscissa is the frequency range and the left bar of each pair is when both target and masker were filtered equally while the right bar of each pair indicates SRM when only the target was filtered (see text).

The amount of SRM for each of these conditions is plotted in Fig. 2. These values do not depend on which masker reference level is used. In each case, the filled bar of the pair indicates the SRM when all stimuli are filtered in the same way while the open bar indicates SRM in the target-filtered conditions. In all cases, the target-filtered conditions resulted in slightly less SRM on average than when all stimuli were filtered equally. The SRM in both of the broadband cases (recall that they are identical except for the participating listeners) was about 12 dB—which was quite similar to that which has been reported previously by Marrone *et al.* (2008a). In the other filtered conditions, less of an SRM was observed. To compare the SRM values obtained for filtering both target and masker and only the target, a repeated-measures analysis of variance on the SRM values was conducted with filter condition (all filtered or target filtered) as the between-subjects factor and with frequency region (broadband, low frequency, mid frequency and high frequency) as the within-subjects factor. For this analysis only, the two listeners who completed both experiments were eliminated from the target filtered group. Frequency region was highly significant [$F(3,27)=28.05$, $p<0.001$], filter condition was not significant [$F(1,9)=0.074$, $p=0.79$], and the interaction was not significant [$F(3,27)=0.38$, $p=0.77$]. This is consistent with the impression given from Fig. 2 that the filtering of the masker did not change the SRM.

C. Discussion

In the first experiment, a large SRM was found for two speech maskers symmetrically separated from a speech target that was similar to that reported by Marrone *et al.* (2008a) for broadband targets and maskers. Restricting the frequency region to below 1.5 kHz reduced the SRM somewhat and greater reductions were observed for mid- and

high-frequency bandpass conditions. However, all of the frequency regions tested potentially contributed to the SRM found in the broadband condition. When all of the stimuli were filtered equally the mid- and high-frequency conditions resulted in SRMs of about 5 dB. This is rather surprising given that better ear effects (or ILDs) due to head shadow are greatly reduced by the symmetric placement of the maskers and less than 1 dB of SRM was found by Marrone *et al.* in a broadband "monaural" control condition (with one ear occluded by an earplug and earmuff). In the current study, we re-tested the listener who had the largest SRM in the high-frequency condition when both target and masker were filtered identically in the same "monaural" control used by Marrone *et al.* and also found less than 1 dB of SRM. Therefore, it appears that there is some binaural information in the mid- and high-frequency regions beyond simple head shadow that can produce an SRM.[1]

A reduction in SRM with high-pass or low-pass filtering under conditions of two masker talkers symmetrically placed around the target (in azimuth) was also reported by Noble and Perrett (2002) although the remaining SRM in their study was approximately equal for both the high-and low-pass filtering. In that study, the amount of SRM in the broadband condition (about 5 dB) was much less than found here so the reductions due to filtering were rather modest, leaving SRMs on the order of 2 dB. The results of a separate (unpublished) experiment from our laboratory in which identical filtering was applied with only ITDs separating the high-frequency stimuli suggest that ILDs, not ITDs, give rise to this SRM. Thus it appears that, in the absence of effective ITD cues (i.e., at high frequencies, cf. Schimmel *et al.* 2008), listeners can use brief epochs of ILDs favoring one ear or the other to perceptually segregate a target even when there are no long-term ILDs. Note that the conclusion that this effect is truly binaural is supported by the lack of any SRM in the monaural control condition reported by Marrone *et al.* (2008a) and informally measured on one listener here.

The relative spectral composition of the different stimuli has received less attention in the SRM literature than other factors. It seems clear that the patterns of performance found here in both colocated and separated conditions are primarily mediated by the frequency content of the target. Filtering the target into the various band-limited conditions presented against broadband maskers yielded threshold T/Ms and SRMs that were not appreciably different than when both target and maskers were filtered equally. It seemed possible, for example, that the mismatch in bandwidth between target and masker would provide a qualitative segregation cue much like that which occurs for differences in fundamental frequency between target and masker. However, there is no such evidence in the current findings. Except for the mid-frequency condition (target bandpass filtered from 1.5 to 3 kHz) the threshold T/Ms for the mismatched target-masker spectra were equal to or slightly higher than most of the corresponding threshold T/Ms for the equal-bandwidth case. The picture that emerges here is that all three of the frequency regions that were tested contribute to the overall SRM found in the broadband case or at least contain useable information. This seems likely to be a perceptual mechanism in which SRM- and more specifically performance in the spatially separated condition—depends on the integration of speech information across the spectrum strengthening the target as a separate auditory "object" from the maskers.

III. EXPERIMENT 2

A second experiment was conducted with the stimuli presented via earphones and separated in apparent interaural location only by ITDs. Because of this mode of presentation none of the results could be attributed to or influenced by ILDs or better ear advantages, even those occurring over the brief time frames discussed above. In this experiment several other aspects of the masking were of interest, including the number of maskers (one or two) and their type (speech, noise, reversed speech). The variability in the amount of SRM often found when comparing across studies likely reflects the influence of these specific design choices and includes how much EM and/or IM is present (cf. Kidd *et al.*, 1998; Freyman *et al.*, 1999; Brungart *et al.*, 2001; Noble and Perrett, 2002; Arbogast *et al.* 2002; Best *et al.*, 2005).

In the first part of this experiment the differences in performance for one vs. two maskers were investigated. It seems likely that the cues available for segregating sound sources are quite different for the different numbers of maskers. The influence of head shadow on spatially separated conditions is much greater for single (or asymmetrically placed) maskers than for dual (symmetrically placed) maskers. Removing this factor by separating sources only by ITD permits a more direct estimation of the ability of listeners to segregate the target through binaural processing without the confound of acoustic head shadow differences.

As discussed above, threshold T/M for two speech maskers colocated with the target (broadband) tends to be near 2–3 dB for these methods and is fairly consistent across studies. On the other hand, threshold T/M for a single speech masker masking a target voice—again, in the colocated condition—is often much lower than that value (more than may be accounted for by the decrease in masker energy) and appears to vary more across subjects and studies (cf. Carhart *et al.*, 1969; Brungart, 2001; Arbogast *et al.* 2002; Culling, *et al.*, 2003; Hawley *et al.*, 2004; Rakerd *et al.*, 2006). Because the colocated condition for speech-on-speech masking is thought to be dominated by IM, the perceptual factors forming the basis for segregation may differ substantially depending on whether there are one or two maskers. For instance, when attempting to track a particular voice it could be easier to follow over time when there are more (or longer) glimpses available. Fewer opportunities to glimpse the target during envelope minima would occur when the target is embedded in two independent masker talkers than one. It has also been shown to be possible to listen to the softer talker when there are only two talkers (e.g., Egan *et al.*, 1954; Dirks and Bower, 1969; Brungart, 2001) whereas according to Brungart *et al.* (2001) this cue is not reliable with three talkers. Some listeners may be able to exploit these cues leading to greatly reduced IM in the reference (colocated) condition which consequently affects the amount of SRM that is observed.

TABLE I. A listing of the stimuli and conditions tested in Experiment 2. The conditions that were not tested are indicated by DNT (did not test) and the conditions that were not applicable because only one masker was present are indicated by N/A.

	One masker		Two maskers		
	Spatial condition				
Masker type	Colocated	Separated (+600 μs)	Colocated	Symmetrically separated ($\pm$600 μs)	One colocated and one separated (+600 μs)
Forward speech	✓	✓	✓	✓	✓
Reversed speech	✓	✓	✓	✓	DNT
SSSM noise	✓	✓	✓	✓	DNT
Speech-like noise	✓	✓	✓	✓	DNT
Speech plus SSSM	N/A	N/A	✓	✓	✓[a]✓[b]

[a]Speech colocated and noise separated.
[b]Noise colocated and speech separated.

The masking produced by different types of maskers (i.e., maskers other than speech) was also investigated, with the intent being to assess performance under conditions having different amounts of IM/EM. For the two-masker case we tested conditions in which one masker was speech (predominantly informational, in this context) and the other was noise (energetic). Currently, the interaction of different types of maskers in this situation is not well known. However, in a recent report by Agus *et al.* 2009, a speech-plus-noise masker was used to help quantify amounts of EM and IM. It was assumed that the combined masker had roughly the same amount of EM as the noise-alone masker and therefore differences in performance for the two masker types could be attributed to IM. In the present study, the combination of the two maskers may simply produce a result intermediate to that produced by each separately or one masker may dominate. Furthermore, it may matter which masker is colocated with the target- and presumably at the focus of attention- and which is spatially/perceptually separated and therefore outside of the focus of attention. This stimulus manipulation may thus be informative regarding the viability of a nulling or cancellation type of mechanism applied to the masker sources (cf. Durlach *et al.*, 2003; Gallun *et al.*, 2005; Brungart *et al.*, 2007). To date this issue has been explored only in the Brungart *et al.* (2007) study, to our knowledge, where essentially no advantage in target intelligibility was found for a two-talker masker when only one talker was moved away from the colocated target and other masker talker.

The rationale supporting this experiment is that the magnitude of the spatial release from speech-on-speech masking observed in a given experiment depends crucially on the details of the design of the experiment. It is only by fully understanding the impact of these design details, which, as per discussion above may affect colocated and separated conditions somewhat differently, that SRM may be interpreted.

A. Methods

As with the first experiment, the second experiment also used a closed-set speech identification task. The speech corpus was custom made and is described in more detail in a previous publication (Kidd *et al.*, 2008b). The structure of the speech test is somewhat like the CRM and is also similar to an earlier test by Hagerman (1982; also see Spieth *et al.*, 1954). In this case, there are five words in a sentence with each word chosen from one of eight alternatives for each word position. All of the words were monosyllabic and were recorded when spoken in isolation with neutral inflection so that all combinations of words could be constructed creating a large set (8^5) of unique sentences without coarticulation effects or temporally "smeared" word boundaries. The structure was always ⟨name⟩⟨verb⟩⟨number⟩⟨adjective⟩⟨object⟩. Combinations created in this manner yield syntactically correct but unpredictable sentences. An example is "Bob found three old shoes."

There were 21 experimental conditions made up of combinations of number of maskers, type of masker and spatial conditions (see Table I). The 5-word target source was designated by the name "Bob" and the remaining four words from that talker were scored. There were two numbers of maskers, one or two, and four types of maskers: two kinds of noise, reversed speech and speech. The speech maskers, when present, were also five-word sentences from the same corpus and constructed in the same way as the targets but all talkers and words were different from the target and each other (in the case of more than one masker). All of the talkers were females and the two or three sentences presented on a given trial were always different talkers chosen randomly from a set of eight. The target and masker talkers were always constant across the five words within a sentence but changed from trial to trial. Temporally reversed speech maskers were created by choosing the words exactly as in the forward speech case but applying the time-reversal for each individual word waveform before concatenation into a five-word string. Hence the first reversed word in a masker was always from the "name" choices, the second was a time-reversed verb, and so on.

For the two types of noise, one was speech-shaped speech-modulated (SSSM) noise and the other was called "speech-like" noise. Both noises were based on choosing words in the same manner as for the speech maskers so that each noise masker consisted of 5 bursts of noise that were equal in duration to the masker words for which they were

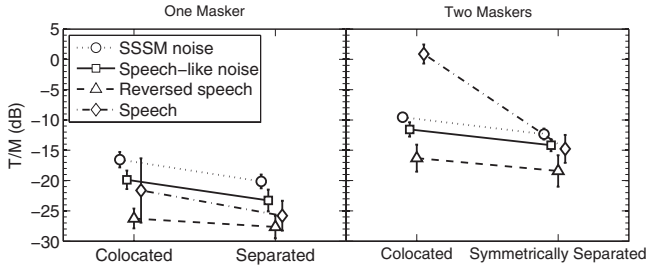

FIG. 3. The results of Experiment 2 plotted as T/M at threshold for one masker (left panel) and for two maskers (right panel) of the same type. The spatial configurations are indicated along the abscissa. The individual functions with different symbols in each panel represent the type of masker (see legend). The data points are offset and connected only for clarity of presentation.

substituted. In the case of SSSM noise, the broadband envelope of the chosen word was used to modulate broadband noise that had the same long-term average spectrum as the entire speech corpus and then each noise burst was concatenated. For the speech-like noise,[2] the only difference was that the spectrum was also taken from that particular choice of masker word and was not an average over the entire corpus. In some two-masker conditions, one masker was speech while the second masker was the SSSM noise. Each word or noise burst was normalized to the same rms value and the T/M was specified by the level of the target sentence relative to the individual level of each masker sequence. Because the words vary naturally in duration there was no imposed time alignment other than the fact that all sequences (target and one or two maskers) began at the same time and had an equal number of words. Sometimes target sentences were longer and sometimes shorter than masker sequences.

In the reference condition (colocated), the target and masker(s) were presented diotically. For the spatial separation conditions the target was diotic and the masker(s) were separated from the target by an ITD of 600 μs. The single masker conditions employed a 600 μs ITD leading toward the right ear. When two maskers were present they either had opposing ITDs of ± 600 μs (symmetrically "spatially" separated) or one was colocated with the target (presented diotically) while the other had a 600 μs ITD leading in the right ear.

In all cases, the level of the individual masker(s) was fixed at 60 dB SPL and the target level was varied adaptively using a one-up one-down technique to estimate T/M at threshold (50% correct sentence identification). In order for a response to be counted correct, all four of the scored words in a target sentence had to be reported correctly.

There were four listeners with normal hearing, each of whom participated for eight 2 h sessions. At the beginning of the experiment, three threshold estimates for the target sentence in quiet were obtained. These trials also served to familiarize the listeners with the corpus and the response interface. In the masked conditions, six threshold estimates per condition were obtained and averaged.

B. Results

Figure 3 shows the group mean results from Experiment 2 for one (left panel) and two (right panel) maskers. In the case of two maskers, only the conditions in which both maskers were of the same type and both were either colocated with the target or were symmetrically separated from the target (i.e., had opposing ITDs) are shown in this figure. For the single masker (left panel), the threshold T/Ms varied over a range of about 10 dB in the colocated condition depending on masker type. The highest threshold was measured for SSSM noise (circles; −16.5 dB T/M) while the lowest threshold was observed for the reversed-speech masker (triangles) at a T/M of −26.3 dB. The individual differences were relatively small except for the speech masker condition (diamonds) where the standard deviation was three to four times larger than the other conditions. For the colocated condition in the two-masker case (right panel), thresholds varied over an even wider range with the threshold for the speech maskers (diamonds) at about 1 dB T/M and the threshold for the reversed-speech maskers (triangles) falling near −16 dB T/M. The threshold for the speech masker was about 10 dB higher than any of the other masker types but, unlike the one-masker case, varied little across listeners.

For the separated conditions the range of threshold T/Ms across masker types was smaller at about 7.5 dB for one masker and about 6 dB for two maskers. In both the one-masker and two-masker cases the lowest thresholds were obtained for the reversed-speech masker and the highest thresholds were for the SSSM noise masker. A repeated-measures analysis of variance indicated that all three main effects: masker type [$F(3,9)=33.1$, $p<0.001$], number of maskers [$F(1,3)=553.5$, $p<0.001$] and spatial condition [$F(1,3)=3523.9$, $p<0.001$] were significant. All three of the two-way interactions and the three-way interaction were also significant. Most obvious in the figure was the three-way interaction in which the highest threshold was for the speech masker when there were two maskers that were colocated with the target.

One general result is that the threshold T/Ms were higher for all of the two-masker conditions than for the corresponding one-masker conditions. As noted above, for these listeners, threshold T/M was about 1 dB when two speech maskers were colocated with the target, which is about the same as past reports using similar closed-set speech tests. By comparison, threshold T/Ms when only one speech masker was colocated with the target averaged about −22 dB. This is an enormous, and somewhat surprising, difference in thresholds that will be revisited in the discussion section. The group mean difference in threshold T/Ms for one versus two maskers (i.e., the additional masking due to the presence of a second masker of the same type) is illustrated in Fig. 4 for the conditions shown in Fig. 3. All of these values were higher than could be accounted for simply by the expected 3 dB increase in total masker power caused by adding a second independent masker (shown by the dashed line in the figure; recall that T/M is computed as the level of the target relative to the level of *each* individual masker). The largest amount of additional masking by far is for speech maskers colocated with the target, which exceeded 20 dB. Otherwise, for the two types of noise as well as reversed speech, increasing the number of maskers from one to two produced an increase in

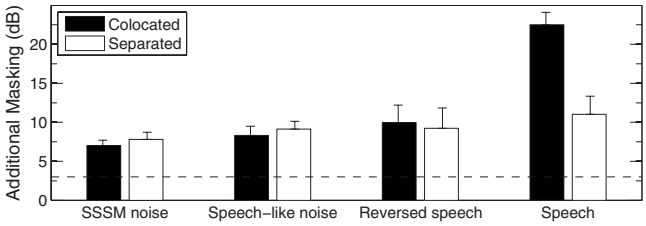

FIG. 4. The increase in masking obtained when two masker sources are present compared to one masker (ordinate). The abscissa is the masker type and for each type the left bar is for the colocated presentation and the right bar is for the separated presentation. The dashed line at 3 dB indicates the increase in threshold T/M expected based simply upon the increase in overall masker level from combining two uncorrelated equal-level maskers.

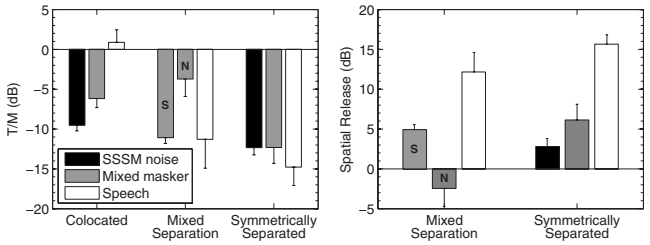

FIG. 6. Results from the mixed masker and mixed separation conditions (see text). The left panel shows the threshold T/Ms while the right panel shows the resulting SRM for each condition. The values for two maskers of the same type (speech or noise) are replotted from Figs. 3 and 5.

threshold T/M of about 7–11 dB in both colocated and spatially separated conditions.

The two-talker speech masker also yielded much larger SRM than any of the other maskers. This is apparent in Fig. 5, which shows the SRM for each condition. The SRM for the two-talker masker was just over 15 dB, a value somewhat larger than the 12.6 dB reported previously by Marrone *et al.* (2008a). For the other maskers, the SRMs were smaller than for the speech maskers due largely to lower T/Ms at threshold in the colocated condition. The SRM observed for the SSSM noise maskers was about 3 dB (2.8 dB for one masker and 3.6 dB for two maskers). The values of SRM for the speech-like noise maskers were nearly identical to those from the SSSM noise, while even smaller SRMs were obtained for the reversed-speech maskers.

The difference between forward and reversed speech—both in terms of threshold T/M and SRM-was substantial, also consistent with the report by Marrone *et al.* (2008a). Although there may be some differences in the EM produced by the two types of masker presentation (cf. Rhebergen *et al.*, 2005), any such effects are likely to be much smaller than the differences found here and, in this case, may also be reduced relative to naturally spoken sentences given that the words were recorded in isolation and reversed individually. So, comparison of the two provides one means of gauging the approximate amount of IM that may be attributed to the meaningfulness of the maskers. The largest reduction in masking due to time reversal was about 17 dB for the two-speech masker colocated condition (Fig. 3).

The second part of Experiment 2 focused on two special situations where there were always two maskers but they were different in some way from the two-masker conditions tested above. These two situations involved the combination

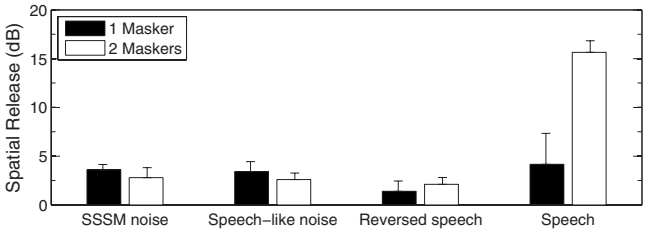

FIG. 5. Spatial release from masking for the conditions displayed in Fig. 3. The ordinate is the amount of SRM in dB and the abscissa is the type of masker that was present. The left bar of each pair is for a single masker while the right bar of each pair is for two maskers.

of a speech masker with a noise masker (referred to as a "mixed masker") and one masker colocated with the target with the other masker separated from the target ("mixed separation"). The exact conditions tested are those listed in the last row and last column of Table I. The only mixed-masker case was for the combination of SSSM noise and speech. The only mixed-separation cases were for two speech maskers or the mixed masker. The results are presented in Fig. 6.

In considering these data, there are a number of comparisons that are of interest. In order to facilitate making these comparisons, some of the conditions from the previous part of Experiment 2 are also shown in Fig. 6 along with the new results. The left panel shows the threshold T/Ms while the right panel gives the SRM. The abscissa in both cases indicates the spatial-separation condition and the masker types are indicated by the shading of the bars: black for noise maskers, gray for mixed maskers and white for speech maskers. For the mixed maskers, the S or N indicates which masker of the pair was *separated.*

For two speech maskers (left panel, white bars), reading left to right, the bars indicate group mean threshold T/Ms for maskers colocated with the target (T/M=0.9 dB), one speech masker colocated with and the other separated from the target by 600 μs ITD (T/M=−11.3 dB), and both maskers symmetrically separated from the target (T/M=−14.8 dB). Relative to the case where both speech maskers were colocated with the target, moving one of the maskers away produced an SRM of about 12 dB on average (right panel, white bar in left group) although there were large differences across listeners for this condition. Moving the second speech masker away from the target so that both were separated reduced thresholds by another 3.5 dB yielding an SRM of about 16 dB (right panel, rightmost white bar).

In contrast to this large effect of separation by ITD for two speech maskers, the values for two SSSM noise maskers were quite different. When the two independent SSSM maskers were colocated with the target, threshold T/M was about −9.5 dB (left panel, leftmost black bar). This value is thus about 10 dB lower than for the case of two colocated speech maskers. When both noise maskers were separated by ±600 μs ITDs, threshold T/Ms were reduced to about −12.3 dB (left panel, black bar in right group of bars). This yielded an SRM of only about 2.8 dB (right panel black bar).

Because of this small SRM when both maskers were separated, the (presumably) intermediate case of mixed separation was not tested.

In the mixed-masker case one masker was speech and the other was SSSM noise. All of the mixed masker conditions produced threshold T/Ms higher than those found for the single speech or single SSSM noise maskers (Fig. 3) in either colocated or separated conditions. The threshold T/M for this combined masker in the colocated case (leftmost gray bar in left panel of Fig. 6) fell in between the corresponding thresholds for either two noise or two speech maskers. When one masker was colocated with the target and the other masker was moved away, there was a large difference in threshold T/M depending on which masker was colocated. When the speech masker was colocated and the noise was separated, the average threshold T/M was −3.7 dB (gray bar marked "N" in middle group of left panel) and a *negative* SRM was observed (right panel, gray bar marked "N;" note that the reference for computing SRM in this case is the mixed-masker colocated threshold). It has been shown in previous studies that noise sometimes can affect the masking caused by speech in rather counterintuitive ways when there is a complex segregation task to perform (cf. Brungart and Simpson, 2002; Gallun *et al.*, 2005; Best *et al.*, 2010). The extent to which this negative SRM, if confirmed, is related to the speech/noise interactions reported in these other studies is not clear at present and requires further study. However, at the very least, there appears to be no positive SRM in this case at all. In contrast, when the speech masker was spatially separated and the noise was colocated with the target (threshold T/M indicated by gray bar marked "S" in left panel of Fig. 6), an SRM of approximately 5 dB was found (gray bar marked "S" in right panel). This is nearly as large an SRM as when both maskers are separated (6.1 dB, rightmost gray bar in right panel for symmetrically separated case). Thus, the difference between moving the speech masker away from the target plus noise masker and moving the noise masker away from the target plus speech masker, was about 7 dB (threshold T/Ms of −3.7 versus −11.1 dB, difference between middle gray bars left panel). This strongly suggests that it was the location of the speech masker relative to the target that tended to determine how much masking occurred in these conditions. However, that generalization must be tempered by some of the other, complex interactions apparent in these data as discussed more fully below.

C. Discussion

In the second experiment, the influence of the type and number of maskers on SRM was investigated using only ITD cues to separate the sounds in perceived interaural location. For all conditions limited to unintelligible maskers (one or two reversed speech, SSSM noise, or speech-like noise maskers) the SRMs were consistently small, ranging from only 1.4 to 3.5 dB. This finding supports the conclusion that SRM is relatively small in the absence of better ear cues under conditions dominated by EM. For all conditions in which a forward speech masker was perceptually separated from the target, regardless of whether it was the only masker or was in combination with another speech or noise masker, the SRMs were larger ranging from 4.9 to 15.7 dB. Within this wide range, the two largest SRMs were found for the situation of two speech maskers when one or both were separated from the target (SRMs of 12.2 and 15.7 dB respectively relative to both colocated). This finding is consistent with the interpretation that large SRMs may be produced under conditions dominated by IM. The determining factor in producing this large SRM appears to be the reference condition comprised of the target colocated with two speech maskers. This produced the highest T/M at threshold of all conditions tested and thus provides the opportunity for the greatest release from masking. When one speech masker was separated and one was colocated, spatial "tuning" (cf. Arbogast and Kidd 2000; Marrone *et al.*, 2008a; Allen *et al.*, 2008) may effectively "attenuate" the separated speech masker allowing the remaining colocated masker to be segregated by monaural/diotic cues much as appears to happen when only one speech masker is present and is at the target location (cf. Figure 3). As suggested by past work on spatial tuning measured perceptually, the phenomenon appears to be complex and sensitive to the specific procedures employed. For example, the threshold T/M for the mixed-masker mixed-separation case when the speech masker is separated from the target (Fig. 6 left panel, gray bar marked "S," −11.1 dB) was essentially the same as the mixed-separation result for two speech maskers (white bar in same group, −11.3 dB) despite the fact that the thresholds for the single speech and SSSM noise maskers were quite different in both colocated and separated conditions (Fig. 3). Under these conditions at least, it did not matter whether the colocated masker was speech or noise when a second speech masker was spatially separated from the target. This does not support the idea that the masker at the point of attentional focus is the sole determinant of the amount of masking especially when one of the maskers is predominantly energetic.

Although substantially smaller, the next largest SRMs for any two-masker situation occurred for the mixed-masker case of speech plus SSSM noise. When both the speech and the noise maskers were separated from the target the SRM was 6.1 dB (compared to both colocated), and it was 4.9 dB when only the speech was separated, leaving the noise masker colocated with the target. The odd case was also for a mixed masker with mixed separation. When the noise masker was separated leaving the speech masker colocated with the target an SRM of −2.5 dB (again relative to both colocated) was found. Given the substantial variation within and across subjects in these conditions it is not known whether this comparatively small effect represents no SRM or if it is genuinely a case of what Brungart and Simpson (2005) call the "reverse cocktail party effect." In their study that effect was observed when a monaural target plus two speech maskers were presented in the same ear with one of the masker talkers at a lower level than the other. The softer talker became more salient when moved to the opposite ear, they argued, and hence produced greater interference than when presented ipsilaterally in that masker context. In the present study when a target talker, another talker and an SSSM noise were all presented diotically, the noise masker

likely simultaneously masked both the target and masker talkers. Moving the noise to the side (via ITD) potentially enhanced the salience of the colocated masker talker producing an effect qualitatively similar to that reported by Brungart and Simpson (2005). Although the interpretations of both of these findings rely on the increased salience of one masker when shifting from diotic (or monaural) to dichotic presentation conditions, note that here the increased masking was due to the colocated masker whereas Brungart and Simpson conclude that the increase in masking was presumably due to the separated masker.

1. Additional masking or "multimasker penalty"

Another interesting finding in Experiment 2 was that greater (and often much greater) masking was produced by two speech maskers than by one. This is an issue that is fundamental to the findings reported here but is not fully understood. There has been other evidence presented for a "multimasker penalty" in which adding a second masker is much more deleterious than might be anticipated based purely on the increase in masker energy (e.g., Yost *et al.*, 1996; Bronkhorst, 2000; Freyman *et al.*, 2004; Durlach, 2006; Iyer *et al.*, 2009; Brungart *et al.*, 2009). In those reports the multimasker penalty was greater when the two maskers were speech—especially highly similar speech—than other sounds (cf. Iyer *et al.*, 2009) thus giving the listener two sources of comprehensible information to ignore. This raises the possibility that the important factor in producing the multimasker penalty may be the extent to which the limitation on performance results from IM where the target and masker are easily confused. It should be noted that further increases in the number of independent speech maskers (beyond two) may decrease the amount of IM (and consequently lessen SRM) as the informational component is reduced and the maskers approach babble or speechlike noise (cf. Freyman *et al.*, 2004).

In the Iyer *et al.* (2009) report they proposed a metric in which an intelligibility score could be obtained that essentially corrects for randomly choosing the key words spoken by one of N talkers in a closed-set forced-choice procedure such as the speech tests used here. The assumption is that when there are multiple talkers the listener's problem is in following the target talker from the initial word (in these cases the callsign indicating the target) to the key words (i.e., maintaining the integrity of the speech "stream"). In their analysis it is assumed that multiple key word combinations are heard and one is chosen at random because it is not known which one came from the target talker. The estimated effect on performance is much larger for three talkers than for two talkers consistent with a multimasker penalty. However, the results for two maskers—one of which is speech—that have been corrected for this random-choice process indicated that performance was similar across various masker types even if the third source is not speech. This suggests that simply having three sources makes it more difficult to follow any one source over time.

In the present study, whatever cues the listeners were using to solve the one speech-masker task were compromised when the number of speech maskers was increased to two, or when a noise was added. Both of these manipulations may have the effect of reducing the number or depth of the dips in the combined masker envelope in which the target is at a favorable T/M. This putative "glimpsing" or "listening in the dips" explanation does not necessarily imply a mechanism that provides a release from EM. The spectro-temporal glimpses provided by such brief epochs may have a strong perceptual effect enhancing segregation, stream continuity over time, and/or source selection (cf. discussion Experiment 1). Thus, this large difference in the effectiveness between one and two speech maskers may reflect differences in the ability to use the various cues available to segregate the target from the maskers and follow it over time under the two conditions. Our group of subjects in Experiment 2 appeared to be particularly successful in segregating a speech target from a single speech masker presented from the same location, based on the very low T/Ms at threshold of about −22 dB measured in that condition, despite the fact that the task required a correct response to all four keywords to decrease the T/M in the tracking procedure. Recalling that our sentences are constructed from words spoken in isolation it is possible that the lack of coarticulation affords better glimpses of target words. It is also possible that there is good fundamental frequency separation among the talkers in this corpus or that the differences in other vocal characteristics provide strong cues. The T/Ms found here are generally similar to the single masker case reported by Carhart *et al.* (1969) using different materials (spondee targets and sentence maskers) but roughly similar spatial conditions (diotic versus opposing ITDs of 800 μs). Despite the similarity for the one-talker masker case though the Carhart *et al.* study found a threshold T/M that was 9 dB better than the current results for the two-talker case, i.e., substantially less additional masking than found here. The smaller threshold difference between one and two masker talkers in the Carhart *et al.* study is not necessarily surprising given that the masker speech was presumably less similar to the target speech than here. Furthermore, Carhart *et al.*'s procedure did not provide the same opportunity for response confusions as with the stimuli/methods used here.

2. Effectiveness of different masker types

When there was only one masker, masking effectiveness, as inferred from the threshold T/Ms in both the colocated case and the spatially separated case, was greatest for SSSM noise, followed by speech-like noise, speech, and reversed speech. In the two-masker case, when both maskers were the same and the two maskers were colocated with the target, the order of masking effectiveness changed such that speech was most effective, followed by SSSM noise, speech-like noise and reversed speech. The mixed masker case predictably fell in between the thresholds for two SSSM noise maskers and two speech maskers. When two maskers were symmetrically separated from the target the threshold T/Ms all fell within a 6 dB range. In the mixed-separation case the most effective masker was the mixed masker when the speech was colocated with the target (and the SSSM noise was separated) with a threshold T/M of −3.7 dB. The other two cases tested, two speech maskers or the mixed masker

with the speech separated, both had threshold T/Ms of about −11 dB. Considering the cases of two maskers in which at least one was speech, when a colocated speech masker was replaced with noise the threshold T/Ms generally improved whereas, if a separated speech masker was replaced by noise there was little change (with the exception of the mixed-masker mixed-separation case when the noise was separated being worse than for mixed separation with two speech maskers).

For speech, noise or speech plus noise, Carhart *et al.* (1969) found similar results. Noise was more effective than speech as a single masker but less effective when there were two maskers of the same type regardless of whether they were colocated or symmetrically separated (via opposing ITDs of 800 μs). They also measured a mixed-masker condition of speech plus noise and found that it was intermediate to two speech or two noise maskers for both colocated and separated conditions. They did not test the mixed separation case tested here.

In an extensive report containing the results from a variety of diotic conditions, Iyer *et al.* (2009) presented data and drew several conclusions that are relevant to this study. In their single masker cases, none of the "contextually-irrelevant" maskers (e.g., reversed speech, foreign language speech, modulated noise or even intelligible speech that was contextually different and hence not "confusable" with the target) decreased the intelligibility of a CRM target sentence unless they were presented substantially higher in level than the target (Fig. 5, p. 20). In the two-masker cases, they concluded that there was no evidence for a "multimasker penalty" when the maskers were these same irrelevant types. However, as they state and is evident from their Fig. 6 (p. 21) this is only the case for positive T/Ms. Regardless, the two-CRM-sentence masker produced poor performance throughout the range of T/Ms tested and the combinations of a CRM sentence masker with an irrelevant masker were intermediate in performance to that and to two irrelevant maskers. Both masker types that contained at least one CRM sentence also resulted in a substantial multimasker penalty. There is also evidence in their findings that the difference between various irrelevant masker types only shows up at negative T/Ms, presumably because the differences in EM effects would be relatively stronger. A remarkable result was that the type of second masker made little difference, even at low T/Ms, once a CRM sentence masker was present and all combinations of a CRM masker and an irrelevant masker were nearly as detrimental as two CRM sentence maskers.

It is interesting that in the current study reversed speech was always less effective than either type of noise, whether as a single or double masker and whether colocated or separated. It is possible that the reversed speech both reduces IM and allows better spectral-temporal glimpses of the target than modulated noise because the noise modulation is like a single-channel vocoder and the noise spectra were long term averages. Better glimpses may also be the reason that the speech-like noise thresholds were always intermediate to the SSSM noise and reversed-speech maskers regardless of whether there were one or two maskers and whether they were colocated or separated. Speech-shaped speech-modulated noise is often used as the EM control for speech but if its spectro-temporal features were made sufficiently similar (e.g., if the modulation were applied in multiple frequency channels) it would soon become intelligible. In that case perhaps the reversed-speech masker, while not perfect, is a better indication of the approximate amount of IM in the forward speech.

3. Forward versus reversed speech

Although the reversed-speech thresholds were the lowest at each condition, the largest difference between forward and reversed speech occurred for two maskers of the same type in the colocated condition (Fig. 3). In that case, the T/M at threshold for reversed speech was about 17 dB lower than the corresponding threshold for forward speech. This is an even greater effect than that reported by Marrone *et al.* (2008a) which, on average, was about 12 dB for CRM sentences. There is a long history of study comparing the effects of forward and reversed speech (e.g., Cherry and Taylor 1954) and the theoretical issues involved in that comparison have been discussed in some detail dating at least from the report by Kimura (1967; see also Kimura and Folb, 1968). The most consistent issue appears to be whether the meaningfulness of forward speech creates more masking or "distraction" than the same speech played backwards. Kimura and Folb (1968) were primarily concerned with uncovering differences between ears in the processing of speech vs. nonspeech. Using a dichotic listening paradigm, they concluded that reversed speech was processed "…by neurophysiological systems overlapping those for normal speech sounds…and provides strong support for the suggestion that the critical distinguishing characteristics of speech sounds are not related to meaningfulness, familiarity, or conceptual content" (p. 396). In a chapter reviewing the distraction effects of various "irrelevant" sounds on the serial recall of visually presented information, Jones (1995) stated "…that reversed speech has effects equivalent to those of normal narrative speech (Jones, 1990) (which) further confirms the hypothesis that meaning plays a minor role, but also reinforces the suspicion that some process related to the low-level analysis of speech is at work" (p. 89). In the Jones (1990) work to which he refers in the chapter, forward speech, reversed speech and speech in a foreign language (Welsh) each had an equally disruptive effect on performance and all were more disruptive than noise interference. In a speech-on-speech masking experiment with two talkers Brungart and Simpson (2002) compared forward and reversed speech in the unattended ear when there was both a target and masker talker in the attended ear. In that case, the effectiveness of forward and reversed speech contralateral maskers depended on T/M. When the T/M in the target ear was positive contralateral forward speech was somewhat detrimental (relative to no contralateral masker) while reversed speech was not. However, at T/Ms of 0 dB and lower, forward and reversed speech maskers were equally detrimental. In contrast to these studies suggesting that forward and reversed speech may be processed similarly and might have equal distraction effects, other studies have found substantial differences. Freyman *et al.* (2001) concluded that some—but

not all—of the IM they observed in a speech-on-speech masking experiment could be attributed to the meaningfulness of the maskers. In their colocated condition, they found differences between forward and reversed speech of about 6 dB or more (depending on S/N) as inferred from the performance-level functions they present (see Fig. 8, p. 2119). As with the current study, though, some SRM was observed even for the reversed speech suggesting that time-reversal may not completely overcome IM, or it is possible the remaining SRM is due to binaural processing and therefore on the order of that seen for noise maskers. Hawley *et al.* (2004) also found less masking for colocated reversed speech maskers than for forward speech maskers. However, in the condition most like those tested here (two interfering talkers colocated with the target speech), the advantage appeared to be less than 5 dB. Likewise, relatively small differences between forward and reversed speech maskers have been reported by Noble and Perrett (2002). For conditions in which there is less IM for a colocated speech masker (perhaps because of other segregation cues like different sex talkers) it is possible that reversing the masker speech provides only a small benefit (relative to colocated forward speech) but almost completely eliminates the IM that is present. In that case, the SRM would be based on the remaining EM and would be expected to be much less. Because it is thought that the benefit of temporal reversal of speech is mostly due to the loss of intelligibility of the distracter and hence confusability with the target it is not surprising that a similar release from IM can occur when the speech is in a language that is foreign to the listener (cf. Freyman *et al.*, 2001; Van Engen and Bradlow, 2007; Iyer *et al.*, 2009).

None of these studies, however, have found the large effects reported by Marrone *et al.* (2008a) and here (cf. Figure 3). One possible factor in the different findings is that both the Marrone *et al.* study and the current study used a closed-set speech identification task in which the masker words could be confused with the target words because they both were drawn from the same set and hence were allowable response alternatives. One would think that if that were the basis for the different magnitudes of the forward versus reversed speech effects, it would be reflected in greater masking in the colocated forward-speech condition. However, the T/Ms in that case are roughly similar to those reported by Freyman *et al.* (2001) and Hawley *et al.* (2004); instead the main difference was seen for the colocated reversed speech values. The threshold T/Ms for our single masker talker case are much lower than those reported by Hawley *et al.* (2004), although they are very close to that reported by Carhart *et al.* (1969) at −22.3 dB. It may be that our subjects are unusually proficient at sound source segregation, our stimuli provide better target glimpses, or that some other details of the procedure contributed to the different findings. However, our results and those of Marrone *et al.* (2008a) seem to warrant the observation that forward and reversed speech may produce very different amounts of IM under certain conditions. Thus, broad generalizations about the two being processed equivalently—at least with respect to masking—are not supported by these results.

IV. SUMMARY

In the first experiment, speech-on-speech masking was studied under conditions in which the stimuli were spatially separated in a sound field and subjected to various types of filtering in order to determine the contributions of different frequency regions to spatial release from masking (SRM). The results indicated that the lowest thresholds in spatially separated conditions occurred when the listener had access to the full bandwidth of the stimulus, suggesting that the binaural cues produced in different frequency regions were integrated to maximize performance.

In a second experiment, the binaural cues were limited to interaural time differences (ITDs) with the stimuli presented through earphones. These conditions allowed the examination of the effects of the type of masking that was produced (i.e., varying proportions of energetic and informational masking, EM and IM, respectively) and the sometimes subtle differences owing to the number and relative perceptual locations of the independent maskers that were present. The results revealed that large SRM could be produced when target and masker(s) were separated only by ITDs, but this only occurred when significant IM was present. Furthermore, the factors governing thresholds were to some degree different in colocated and separated conditions, suggesting that SRM alone does not provide a very complete characterization of listener performance.

ACKNOWLEDGMENTS

This work was supported by AFOSR Grant No FA9550-08-1-0424 and by Grant Nos. DC004545, DC00100 and DC004663 from NIH/NIDCD. Virginia Best was also supported by a University of Sydney Postdoctoral Research Fellowship and Nicole Marrone was supported by Grant No. DC004453 from NIH/NIDCD. The authors thank Suzanne Carr Levy for her comments on an earlier version of this manuscript and Nathaniel I. Durlach for many interesting discussions of this work.

[1]The loudspeaker locations, heights and orientations were verified relative to the position of the listener. The listeners were not restrained in their seats and their head movements were not monitored. They were instructed to face the forward loudspeaker (0° azimuth) during testing with their heads supported by a headrest that orients the head in the forward position. The target was always presented from the loudspeaker directly in front. However, because head movements were not physically prohibited it is possible that the head position re. the loudspeakers was slightly asymmetric for some listeners or that movements caused small acoustic asymmetries during testing. Our experience with this system in multiple studies has convinced us that the advantage of the comfort of this arrangement outweighs the greater control that might be achieved by a more rigid head restraint system for experiments requiring extended listening.

[2]The "speech-like noise" masker used in Experiment 2 was called "speech-like" because it had that quality, sometimes sounding like the vowels from the words it had been based on. It was created by choosing an appropriate masker word, obtaining its spectrum and applying that spectrum to Gaussian noise. This masking stimulus varied qualitatively from word to word, was dominated by the energy and spectral shape of the vowels in the word, and was unintelligible. This masker was included because it has a strong speech-like character which varies significantly from word to word but does not preserve the normal amplitude fluctuations of intelligible speech.

Agus, T. R., Akeroyd, M. A., Gatehouse, S., and Warden, D. (**2009**). "Informational masking by young and elderly listeners for speech masked by

simultaneous speech and noise," J. Acoust. Soc. Am. **126**, 1926–1940.
Akeroyd, M. A. (**2004**). "The across frequency independence of equalization of interaural time delay in the equalization-cancellation model of binaural unmasking," J. Acoust. Soc. Am. **116**, 1135–1148.
Allen, K., Carlile, S., and Alais, D. (**2008**). "Contributions of talker characteristics and spatial location to auditory streaming," J. Acoust. Soc. Am. **123**, 1562–1570.
Arbogast, T. L., and Kidd, G., Jr. (**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 masking and energetic masking of speech," J. Acoust. Soc. Am. **112**, 2086–2098.
Best, V., Gallun, F. J., Mason, C. R., Kidd, G., Jr., and Shinn-Cunningham, B. G. (**2010**). "The impact of noise and hearing loss on the processing of simultaneous sentences," Ear Hear. **31**, 213–220.
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**, 3766–3773.
Bolia, R. S., Nelson, W. T., Ericson, M. A., and Simpson, B. D. (**2000**). "A speech corpus for multitalker communications research," J. Acoust. Soc. Am. **107**, 1065–1066.
Bregman, A. S. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT, Cambridge, MA).
Broadbent, D. E. (**1958**). *Perception and Communication* (Pergamon, New York).
Bronkhorst, A. (**2000**). "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," Acust. Acta Acust. **86**, 117–128.
Brungart, D. S. (**2001**). "Evaluation of speech intelligibility with the coordinate response measure," J. Acoust. Soc. Am. **109**, 2276–2279.
Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. (**2006**). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. **120**, 4007–4018.
Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. (**2009**). "Multitalker speech perception with ideal time-frequency segregation: Effects of voice characteristics and number of talkers," J. Acoust. Soc. Am. **125**, 4006–22.
Brungart, D. S., Iyer, N., and Simpson, B. D. (**2007**). "Selective spatial attention in a dynamic cocktail party task: Evidence for a strategy based on masker minimization," J. Acoust. Soc. Am. **121**, 3119.
Brungart, D. S., and Simpson, B. D. (**2002**). "Within-ear and across-ear interference in a cocktail-party listening task," J. Acoust. Soc. Am. **112**, 2985–95.
Brungart, D. S., and Simpson, B. D. (**2005**). "Evidence for a reverse cocktail party effect in a three-talker dichotic listening task," Acta. Acust. Acust. **91**, 564–566.
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.
Carhart, R., Tillman, T. W., and Greetis, E. S. (**1969**). "Release from multiple maskers: Effects of interaural time disparities," J. Acoust. Soc. Am. **45**, 411–418.
Carr, S. P. (**2010**). "The effects of pitch, reverberation, and spatial separation on the intelligibility of speech masked by speech in normal-hearing and hearing-impaired listeners," Ph.D. thesis, Boston University, Boston, MA.
Cherry, E. C. (**1953**). "Some experiments on the recognition of speech, with one and two ears," J. Acoust. Soc. Am. **25**, 975–979.
Cherry, E. C. and Taylor, M. K.(**1954**). "Some further experiments upon the recognition of speech, with one and two ears," J. Acoust. Soc. Am. **26**, 554–559.
Culling, J. F. (**2007**). "Evidence specifically favoring the equalization-cancellation theory of binaural unmasking," J. Acoust. Soc. Am. **122**, 2803–2813.
Culling, J. F., and Colburn, H. S. (**2000**). "Binaural sluggishness in the perception of tone sequences and speech in noise," J. Acoust. Soc. Am. **107**, 517–527.
Culling, J. F., Edmonds, B. A., and Hodder, K. I. (**2006**). "Speech perception from monaural and binaural information," J. Acoust. Soc. Am. **119**, 559–565.
Culling, J. F., Hodder, K. I., and Toh, C. Y. (**2003**). "The effect of reverberation on perceptual segregation of competing voices," J. Acoust. Soc. Am. **114**, 2871–2876.
Dirks, D. D., and Bower, D. R. (**1969**). "Masking effects of speech competing," J. Speech Hear. Res. **12**, 229–45.
Dubno, J. R., Ahlstrom, J. B., and Horwitz, A. R. (**2002**). "Spectral contributions to the benefit from spatial separation of speech and noise," J. Speech Lang. Hear. Res. **45**, 1297–1310.
Durlach, N. I. (**1963**). "Equalization and cancellation theory of binaural masking level differences," J. Acoust. Soc. Am. **35**, 1206–1218.
Durlach, N. I. (**1972**) "Binaural signal detection: Equalization and cancellation theory," *Foundations of Modern Auditory Theory*, edited by J. V. Tobias (Academic, New York), Vol. **2**, Chap. 10.
Durlach, N. I. (**2006**). "Auditory masking: Need for improved conceptual structure," J. Acoust. Soc. Am. **120**, 1787–1790.
Durlach, N. I., Mason, C. R., Kidd, G., Jr., Arbogast, T., Colburn, H. S., and Shinn-Cunningham, B. G. (**2003**). "Note on informational masking," J. Acoust. Soc. Am. **113**, 2984–2987 (L).
Egan, J. P., Carterette, E. C., and Thwing, E. J. (**1954**). "Some factors affecting multi-channel listening," J. Acoust. Soc. Am. **26**, 774–782.
Fletcher, H. (**1940**). "Auditory patterns," Rev. Mod. Phys. **12**, 47–65.
Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2001**). "Spatial release from informational masking in speech recognition," J. Acoust. Soc. Am. **109**, 2112–2122.
Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2004**). "Effect of number of masker talkers and auditory priming on informational masking in speech recognition," J. Acoust. Soc. Am. **115**, 2246–2256.
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.
Gallun, F. J., Mason, C. R., and Kidd, G., Jr. (**2005**). "Binaural release from informational masking in a speech identification task," J. Acoust. Soc. Am. **118**, 1614–1625.
Hagerman, B. (**1982**). "Sentences for testing speech intelligibility in noise," Scand. Audiol. **11**, 79–87.
Hawley, M. L., Litovsky, R. Y., and Culling, J. F. (**2004**). "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," J. Acoust. Soc. Am. **115**, 833–843.
Hirsh, I. J. (**1950**). "The relation between localization and intelligibility," J. Acoust. Soc. Am. **22**, 196–200.
Iyer, N., Brungart, D. S., and Simpson, B. D. (**2009**). "Intelligibility of target signals in sequential and simultaneous segregation tasks," Final Report No. AFRL-RH-WP-TR-2009-0033, Air Force Office of Scientific Research (AFOSR), Arlington, VA.
Jones, D. (**1995**). "Objects, streams and threads of auditory attention," in *Attention: Selection, Awareness, and Control: A Tribute to Donald Broadbent*, edited by A. Baddeley and L. Weiskrantz (Oxford University Press, New York).
Jones, D. M. (**1990**). "Recent advances in the study of performance in noise," Environ. Int. **16**, 447–58.
Kidd, G., Jr., Best, V., and Mason, C. R. (**2008b**). "Listening to every other word: Examining the strength of linkage variables in forming streams of speech," J. Acoust. Soc. Am. **124**, 3793–802.
Kidd, G., Jr., Mason, C. R., Brughera, A., and Hartmann, W. M. (**2005**). "The role of reverberation in release from masking due to spatial separation of sources for speech identification," Acta. Acust. Acust. **91**, 526–536.
Kidd, G., Jr., Mason, C. R., Richards, V. M., Gallun, F. J., and Durlach, N. I. (**2008a**). "Informational masking," in *Auditory Perception of Sound Sources*, edited by W. A. Yost, A. N. Popper, and R. R. Fay (Springer Science+Business Media, LLC, New York), pp. 143–190.
Kidd, G., Jr., Mason, C. R., Rohtla, T. L., and Deliwala, P. S. (**1998**). "Release from masking due to spatial separation of sources in the identification of nonspeech auditory patterns," J. Acoust. Soc. Am. **104**, 422–431.
Kimura, D. (**1967**). "Functional asymmetry of the brain in dichotic listening," Cortex **3**, 163–178.
Kimura, D., and Folb, S. (**1968**). "Neural processing of backwards-speech sounds," Science **161**, 395–396.
Kock, W. E. (**1950**). "Binaural localization and masking," J. Acoust. Soc. Am. **22**, 801–804.
Koenig, W. J. (**1950**). "Subjective effects in binaural hearing," J. Acoust. Soc. Am. **22**, 61–62.
Levitt, H., and Rabiner, L. R. (**1967**). "Binaural release from masking and gain in intelligibility," J. Acoust. Soc. Am. **42**, 601–608.
Licklider, J. C. R. (**1948**). "The influence of interaural phase relations on the masking of speech by white noise," J. Acoust. Soc. Am. **20**, 150–159.
Marrone, N., Mason, C. R., and Kidd, G., Jr. (**2008a**). "Tuning in the spatial

dimension: Evidence from a masked speech identification task," J. Acoust. Soc. Am. **124**, 1146–1158.
Marrone, N., Mason, C. R., and Kidd, G., Jr. (**2008b**). "Effects of hearing loss and age on the benefit of spatial separation between multiple talkers in reverberant rooms," J. Acoust. Soc. Am. **124**, 3064–3075.
Noble, W., and Perrett, S. (**2002**). "Hearing speech against spatially separate competing speech versus competing noise," Percept. Psychophys. **64**, 1325–1336.
Pollack, I., and Pickett, J. M. (**1958**). "Stereophonic listening and speech intelligibility against voice babble," J. Acoust. Soc. Am. **30**, 131–133.
Rakerd, B., Aaronson, N. L., and Hartmann, W. M. (**2006**). "Release from speech-on-speech masking by adding a delayed masker at a different location," J. Acoust. Soc. Am. **119**, 1597–1605.
Rhebergen, K. S., Versfield, N. J., and Dreschler, W. A. (**2005**). "Release from informational masking by time reversal of native and non-native speech," J. Acoust. Soc. Am. **118**, 1274–1277.
Schimmel, O., van de Par, S., Breebaart, J., and Kohlrausch, A. (**2008**). "Sound segregation based on temporal envelope structure and binaural cues," J. Acoust. Soc. Am. **124**, 1130–1145.
Schubert, E. D., and Schultz, M. C. (**1962**). "Some aspects of binaural signal selection," J. Acoust. Soc. Am. **34**, 844–849.
Spieth, W., Curtis, J. F., and Webster, J. C. (**1954**). "Responding to one of two simultaneous messages," J. Acoust. Soc. Am. **26**, 391–396.
Van Engen, K., and Bradlow, A. R. (**2007**). "Sentence recognition in native- and foreign-language multi-talker background noise," J. Acoust. Soc. Am. **121**, 519–526.
Yost, W. A., Dye, R. H., Jr., and Sheft, S. (**1996**). "A simulated 'cocktail party' with up to three sound sources," Percept. Psychophys. **58**, 1026–1036.
Zurek, P. M. (**1993**). "Binaural advantages and directional effects in speech intelligibility," in *Acoustical Factors Affecting Hearing Aid Performance*, 2nd ed., edited by G. A. Studebaker and I. Hochberg (Allyn and Bacon, Needham Heights, MA), Chap. 15.