Spatial release from energetic and informational masking in a selective 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*1

(Received 10 November 2006; revised 12 March 2008; accepted 13 March 2008)

A masker can reduce target intelligibility both by interfering with the target's peripheral representation ("energetic masking") and/or by causing more central interference ("informational masking"). Intelligibility generally improves with increasing spatial separation between two sources, an effect known as spatial release from masking (SRM). Here, SRM was measured using two concurrent sine-vocoded talkers. Target and masker were each composed of eight different narrowbands of speech (with little spectral overlap). The broadband target-to-masker energy ratio (TMR) was varied, and response errors were used to assess the relative importance of energetic and informational masking. Performance improved with increasing TMR. SRM occurred at all TMRs; however, the pattern of errors suggests that spatial separation affected performance differently, depending on the dominant type of masking. Detailed error analysis suggests that informational masking occurred due to failures in either across-time linkage of target segments (streaming) or top-down selection of the target. Specifically, differences in the spatial cues in target and masker improved streaming and target selection. In contrast, level differences helped listeners select the target, but had little influence on streaming. These results demonstrate that at least two mechanisms (differentially affected by spatial and level cues) influence informational masking. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2904826]

PACS number(s): 43.66.Dc, 43.66.Pn, 43.66.Qp [RLF] Pages: 4369–4379

I. INTRODUCTION

When listening selectively for target speech in a background of competing talkers, at least two forms of masking can interfere with performance: energetic and informational masking (see, e.g., Freyman *et al.*, 1999; Brungart *et al.*, 2001; Arbogast *et al.*, 2002; Brungart *et al.*, 2005; Kidd *et al.*, 2005). Spatially separating the target from concurrent masker(s) can improve performance, an effect known as spatial release from masking (SRM; e.g., see Hirsh, 1948; Cherry, 1953; Arbogast *et al.*, 2002). While traditional binaural models can account for spatial release from energetic masking (e.g., Zurek, 1993), the mechanisms underlying spatial release from informational masking are poorly understood.

Two stimulus characteristics are often said to contribute to informational masking: (1) similarity between target and masker with respect to perceptual (e.g., Freyman *et al.*, 1999; Darwin and Hukin, 2000; Brungart, 2001a) or linguistic attributes (Hawley *et al.*, 2004; Van Engen and Bradlow, 2007), and (2) uncertainty about either target or masker (e.g., Lutfi, 1993; Kidd *et al.*, 2005a; Freyman *et al.*, 2007). Past work suggests that at least some of these effects of informational masking are due to failures in segregation and/or attention (e.g., Brungart *et al.*, 2005; Edmonds and Culling, 2006). However, there is no clear consensus on how these two processes contribute to spatial release from informational masking.

Previous studies comparing energetic and informational masking indicate that analysis of response errors can dissociate effects caused by energetic and informational masking. For instance, in selective speech identification tasks with interference from informational masking, subjects often report words from the masker rather than the target message(s) (Brungart, 2001b; Kidd *et al.*, 2005a; Wightman and Kistler, 2005). In contrast, for selective-listening tasks that are dominated by energetic masking, errors are more randomly distributed. The current study attempts to tease apart the mechanisms underlying informational masking through a more detailed analysis of response patterns than has been undertaken in previous studies. The analyses are driven by the hypothesis that similarity between target and masker can interfere with (1) extracting local time-frequency segments from the acoustic mixture, (2) connecting segments from the same source across time (streaming) and/or (3) selecting the correct target segments (or stream) even if they are properly segmented and streamed.

We explored how spatial separation between target and masker influences the pattern of responses, and how these patterns are affected by the level difference between target and masker. Listeners were asked to report a phrase from a variable-level target message that was presented simultaneously with a fixed-level masker message. The locations of target and masker were simulated over headphones to be either co-located or spatially separated by 90°. In addition,

[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.

target level varied over a wide range so that differences in level between target and masker could provide listeners with a cue to select the target and/or better link target keywords across time into a coherent target stream.

Analysis of response errors revealed systematic changes with spatial configuration and target level in the likelihoods of reporting all target keywords, all keywords from the masker, or a mixture of keywords from target and masker. The pattern of results suggests that the relative contributions of energetic and informational masking change systematically with the target-to-masker broadband energy ratio (TMR). At near-zero-dB TMRs, when informational masking occurs, space and intensity cues may help listeners track keywords across time to form a proper stream, as well as enable listeners to select the proper keywords or streams out of the mixture.

II. METHODS

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), which consists of highly predictable sentences of the form "Ready ⟨*call sign*⟩, go to ⟨*color*⟩ ⟨*number*⟩ now." The call sign was one of the set ["Baron," "Eagle," "Tiger," and "Arrow"]; the color was one of the set ["white," "red," "blue," "green"]; and the number was one of the digits between one and eight, excluding the number seven (as it is the only two-syllable digit and is therefore relatively easy to identify). For each session, one of the four call signs was randomly selected as the target call sign.

In each trial, two different sentences were used as sources. The call signs, numbers, and colors in the two utterances were randomly chosen, but constrained to differ from each other in each trial (with one sentence always containing the target call sign). In order to minimize differences between competing messages, talker 0 was used for both sentences.

Each speech signal was 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). All processing was implemented in MATLAB 6.5 [see Fig. 1(a) for a diagram of the processing scheme]. Each target and masker source signal was bandpass filtered into 16 fixed frequency bands of 1/3 octave width, with center frequencies spaced evenly on a logarithmic scale between 175 Hz and 5.6 kHz (every one-third octave). 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. Figure 1(b) shows an example stimulus spectrum with all 16 bands (consecutive bands are shown in alternating shades). In contrast to many previous experiments using amplitude-modulated sine-wave carrier speech (e.g., Arbogast *et al.*, 2002; Gallun *et al.*, 2005; Brungart *et al.*, 2005; Kidd *et al.*, 2005b), the frequency bands of the current stimuli were not equalized to have similar spectral amplitudes, so that the high-frequency bands have less energy than the low-frequency bands.

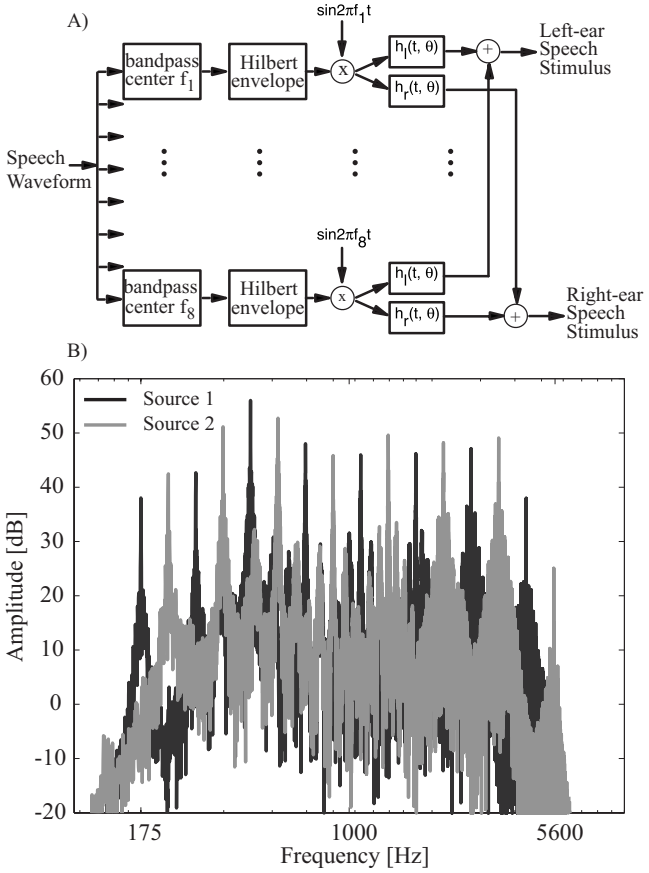

FIG. 1. (a) Flow chart showing how the stimuli were generated. (b) An example of the resulting interleaved spectra of two competing spectrally sparse messages, one in gray and one in black.

On each individual trial, eight of the 16 bands were chosen randomly while ensuring that four of these bands were selected from the lower eight bands (175–882 Hz) and four were selected from the upper eight bands (1.1–5.6 kHz). This resulted in a set of $(8!/(4!4!))^2$ or 4900 different possible spectral combinations. The eight bands were then summed to create the raw waveform for one source. The remaining eight bands were used to construct the other source using otherwise identical processing. As a result, the two raw sources had identical statistics over the course of the experiment, but differed within a trial in their timbre, call sign, and keywords (and, on most trials, level).

The raw source waveforms were scaled to have the same fixed, broadband root-mean-square (RMS) energy prior to spatial processing (described below). When target and masker were set to the same level of broadband RMS energy, the within-band energy ratio of one utterance to another was on the order of 20 dB at all frequencies [cf. Fig. 1(b)]. In fact, a model of the auditory periphery shows that for these

stimuli, energetic masking is likely to play a significant role within each target band only when the selected broadband TMR is −20 dB and less (see Shinn-Cunningham *et al.*, 2005a). Within each trial, the two spatially processed sentences were closely time aligned (this is an inherent property of the CRM corpus; no steps were taken to alter the temporal synchrony present in the original CRM corpus). To determine how synchronized the colors and numbers are when two sentences from Talker 0 are begun simultaneously, the timing of the key words was measured using a routine programmed in MATLAB 7.0. The color and number words were 307 and 290 ms long, on average, with standard deviations of 160 and 148 ms (for colors and numbers, respectively). The onset times of the keywords were almost synchronous across utterances; the standard deviations for the onset times of the color and number words across all stimuli are 29 and 164 ms, respectively.

C. Spatial synthesis

The raw signals were simulated at a distance of 1 m in the horizontal plane containing the ears, either at azimuth 0° (straight ahead) or 90° (to the right of the listener), using head-related transfer functions (HRTFs). HRTFs were measured in a classroom using a Knowles Electronics Manikin for Acoustic Research. The first 10 ms of the HRTFs were time windowed and band limited between 200 Hz and 10 kHz to get pseudo-anechoic HRTFs (see Shinn-Cunningham *et al.*, 2005b for details about the HRTFs used in this spatial processing). The prefiltered utterances were processed with appropriate HRTFs to simulate the desired configuration on a given trial.

D. Procedures

At the start of each session, a random call sign was selected to serve as the target call sign. The target call sign was always in the variable-level talker message. Listeners were instructed to report the color and number of the message with the target call sign, ignoring the message of the fixed-level talker, which will be referred to as the masker. A trial was scored as correct and subjects were given feedback that they were correct if and only if they reported both target keywords.

In each trial, the masker was presented at the same RMS level (which was approximately 70 dB sound pressure level SPL, when presented through the hardware). The target level was varied relative to that of the masker by an amount that was random from trial to trial, chosen from one of six levels (−40, −30, −20, −10, 0, and +10 dB, relative to the level of the masker prior to spatial processing). Subsequently, the binaural signals for the two talkers were summed to produce the two-talker stimulus.

There were four possible spatial configurations, two in which the target and masker were co-located (at either 0 or 90°) and two in which the talkers were spatially separated (target at 0° and masker at 90°, or target at 90° and masker at 0°).

In each run, the spatial configuration of the two talkers was fixed (i.e., the target and masker were played from constant locations throughout the run). In half of the sessions, subjects were told the call sign and location of the target message prior to each run; in the other half of the sessions only the call sign was identified *a priori* (although the location for the target was fixed throughout a block of trials). However, this difference in instructions had no consistent effect on results, so the data were collapsed across the different instructions.[1]

To ensure that subjects could understand the highly processed speech stimuli, subjects went through an initial screening in which they had to report the color and number of one message presented in quiet (processed by 0° HRTF) with at least 90% accuracy over 50 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). Additional data were collected in four additional sessions discussed in the companion paper (Ihlefeld and Shinn-Cunningham, submitted), in which listeners were asked to report both sets of keywords from both of the concurrent messages. Each session consisted of 12 runs (three runs for each of the four spatial configurations). The order of the runs in a session was random, but constrained so that each spatial configuration was performed once before any were repeated. A run consisted of eight repetitions of each of the six TMRs (48 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).

III. RESULTS

A. Percent correct

Given the four possible colors and seven possible numbers, the probability of responding correctly by chance is 1/28 or about 4%; however, if subjects understood only the masker color and number, realized they had heard the masker (not the target), and eliminated these possible responses, the likelihood of responding correctly by chance is 1/18 (6%).

The top panel of Fig. 2 shows the across-subject mean percent correct as a function of TMR for each spatial configuration; error bars show the standard errors of the mean across subjects. Results for all subjects show similar trends, so only across-subject averages are shown. In all configurations, performance improves with increasing TMR. When target and masker were spatially separated, performance was always better than for the co-located cases (dashed lines fall above solid lines). For both co-located and spatially separated conditions, performance at low TMRs was near chance levels. For spatially separated configurations, performance improved to near 100% for near-zero TMRs; however, in the co-located conditions, performance only reached about 80% at the highest TMRs. Moreover, for the co-located configurations, performance at TMR=0 dB, when the target and

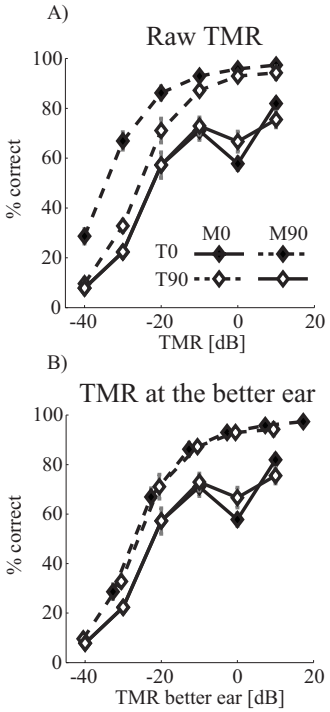

FIG. 2. Overall percent correct as a function of the TMR for the four tested spatial configurations, averaged across subjects. In general, performance improves with TMR, and is better for spatially separated than co-located sources. Error bars show the across-subject standard error of the mean. Filled symbols show results for the target at 0°, open symbols for the target at 90°. Results for spatially separated target and masker 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 masker ratio (TMR). (b) The same results re-plotted as a function of TMR_{be} (correcting for differences in the acoustic TMR at the better ear).

masker were at the same level, was actually worse than at TMR=−10 dB. This is similar to a plateau effect that has been observed in previous studies (Brungart, 2001b; Shinn-Cunningham *et al.*, 2005a).[2]

For co-located configurations, results are essentially identical when the target and masker play from in front of or from the side of the listener. In the separated conditions, results were best when the target originated in front of the listener and the masker from the side. Previous work suggests that the difference in performance for the two spatially separated configurations can be accounted for by considering the differences in the broadband TMR at the acoustically better ear (Shinn-Cunningham *et al.*, 2005a). Prior to the spatial simulation, RMS values of the processed speech waveforms were normalized to have the same broadband RMS, and then the overall level of the target was adjusted downward to produce the desired TMR. However, the spatial processing also altered the target and masker levels, so that the TMR in the signals reaching the ears varied with spatial configuration and could differ at the two ears (depending on the spatial configuration). In the condition where the target was at 0° and the masker at 90°, the TMR at the left ear was, on average, 7 dB greater than the nominal TMR, because the masker energy reaching the left ear was significantly reduced by the acoustic head shadow (note, however, that due to the random selection of frequency bands making up the target and masker on a given trial, the actual TMR at the ears varied somewhat from trial to trial). When the target was at 90° and the masker was in front, the TMR at the right ear was almost equal to the nominal TMR, averaging 1 dB lower than the nominal TMR (of course, for this configuration, the TMR at the left ear is 7 dB lower than the nominal TMR). Note that on average, the TMR for co-located target and masker equaled the nominal TMR.

TABLE I. Possible response types and chance performance for each category. On each trial, subjects responded by reporting a color (C) and a number (N). The subscripts denote whether a keyword was part of the target (T) or masker (M) message; X denotes that the reported word was not present in either the target or the masker.

Response type	Chance	Responses
Masker error	3.6%	$[C_MN_M]$
Mix error	7.1%	$[C_TN_M]$ or $[C_MN_T]$
Drop-1 error	25.0%	$[C_TN_X]$ or $[C_XN_T]$
Drop-2 error	35.7%;	$[C_XN_X]$
Combination error	25.0%.	$[C_MN_X]$ or $[C_XN_M]$
Correct	3.6%	$[C_TN_T]$

To take into account these acoustic effects, data were re-plotted as a function of the TMR at the acoustically better ear (denoted by TMR_{be}) by shifting the raw data in the top panels of Fig. 2 by the appropriate amounts in dB for each spatial configuration (for discussion see Shinn-Cunningham *et al.*, 2005a). As seen in the bottom panels of Fig. 2, this adjustment accounts for differences in performance for the two spatially separated configurations (results for the two separated configurations are indistinguishable following this correction).

B. Analysis of response errors

All incorrect responses were categorized into one of five mutually exclusive error types whose definitions are listed in Table I. Figure 3 plots each kind of error. Errors generally decreased with increasing TMR_{be}. However, the relative likelihoods of the different types of errors depended on TMR_{be} and spatial configuration. At low $TMRs_{be}$ (−20 dB and below), *drop errors* (reporting keywords not in either the target or the masker messages) were the most common errors. For TMR_{be} of −10 dB and greater, *masker* (reporting both masker keywords) and *mix error*s (reporting a mix of target and masker keywords) were the most common errors when sources are co-located, but both types of error were rare for spatially separated sources.

*Drop error*s are shown in panels (A) and (B). The proportion of trials with *drop error*s decreased steeply with increasing TMR_{be} in all spatial configurations, consistent with a decrease in the amount of energetic masking with increasing TMR_{be}. For TMR_{be} of −10 dB and greater very few *drop error*s occurred. For $TMRs_{be}$ between −30 dB and −10 dB (where floor and ceiling effects can be ignored), the number of drop errors was larger for co-located than for separated configurations (dashed lines fall below solid lines), although this difference was small.

Panel (C) displays *mix error*s, where subjects report one target and one masker keyword. At low $TMRs_{be}$ (−40 dB to −20 dB) *mix error*s increased as TMR_{be} in-

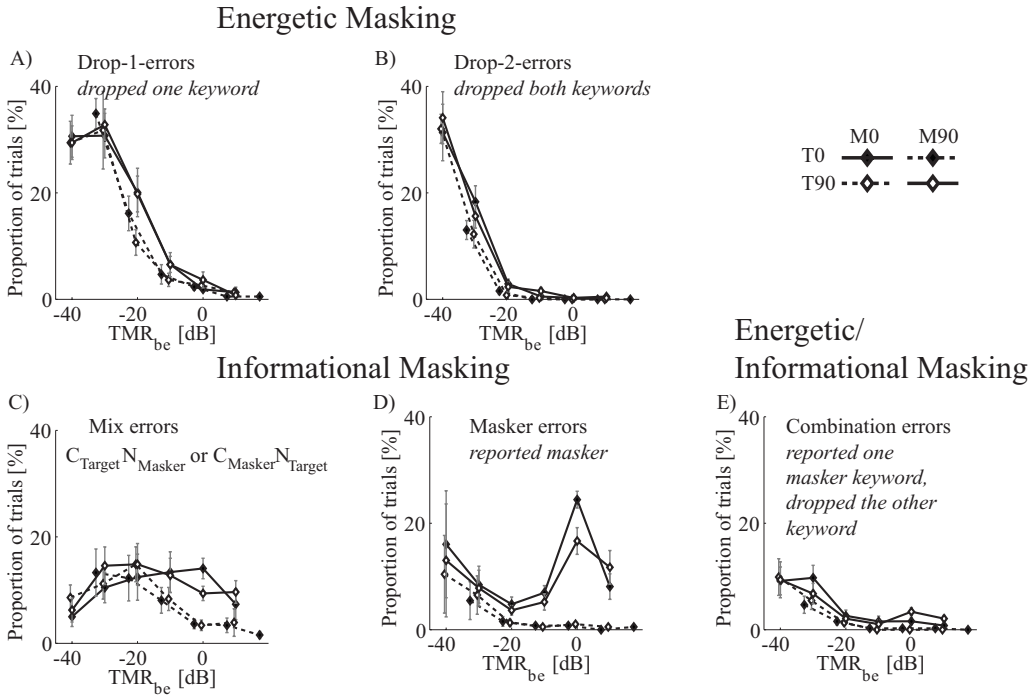

FIG. 3. Analysis of the response errors as a function of TMR_{be} for each kind of error, averaged across subjects. The dependence of each type of response error on TMR_{be} differs, indicating differences in the kind of masking responsible for the errors. Error bars show the across-subject standard error of the mean. Filled symbols show results for the target at 0°, open symbols for the target at 90°. Results for spatially separated target and masker are shown with dashed lines. Results for co-located sources are shown with solid lines. Labels above the panels indicate the posited kind of perceptual interference thought to contribute to the different types of response errors. (a) *Drop-1* errors, where listeners report one target word and one word not present in either the target or masker. (b) *Drop-2* errors, where listeners report two words not present in either the target or masker. (c) *Mix error*s, where listeners report one word from the target and one from the masker. (d) *Masker error*s, where listeners report both words from the masker. (e) *Combination* errors, which encompass all incorrect responses that are not in the above categories (e.g., reporting one masker word and guessing one word).

creased, with no significant differences between co-located and separated spatial configurations. For TMR_{be} −10 dB and greater, the rate of *mix error*s decreased with increasing TMR_{be} in the spatially separated configurations (dashed lines), but was nearly constant in the co-located configurations (given the across-subject variability; see solid lines). In other words, spatial separation between target and masker reduced the likelihood that listeners reported one target and one masker keyword, but only when the TMR_{be} was larger than −20 dB.

Masker errors are shown in panel (D). When sources were spatially separated (dashed lines), *masker* errors decreased monotonically with increasing TMR_{be}. Essentially no *masker* errors occurred for $TMRs_{be}$ of −10 dB or greater when target and masker were spatially separated. At all TMR_{be}, more *masker* errors occurred when the two sources were at the same location than when they were spatially separated (solid lines are above the dashed lines). *Masker* errors in the co-located configurations were nonmonotonic, decreasing as TMR_{be} grew from −40 dB to −20 dB, increasing as TMR_{be} grew from −20 dB to 0 dB, and then decreasing again for TMR_{be} of 10 dB.

Panels (C) and (D) show different trends in the likelihoods of obtaining *mix* versus *masker* errors. In particular, the ratio of the percentages of *mix* errors and *masker* errors varied nonmonotonically as a function of TMR_{be} (ratio not shown) and depended on whether or not target and masker were spatially separated or co-located. At −40 dB TMR_{be}, there were roughly half as many *mix* errors as *masker* errors. This ratio increased monotonically until around −20 dB TMR_{be}, where *mix* errors occurred at least four times more often than *masker* errors. For −10 dB TMR_{be} and greater, in the spatially separated configurations, *masker* errors were essentially nonexistent and *mix* errors decreased monotonically. In contrast, in the co-located configurations, *masker* errors peaked at 0 dB TMR_{be}, while *mix* errors were roughly constant for TMR_{be} greater than −10 dB. Together, these trends show that the spatial configuration of the sources and TMR_{be} affected *mix* and *masker* errors in different ways. This suggests that the relative importance of different forms of interference depends on the relative levels and locations of target and masker.

Combination errors, shown in panel (E), are relatively uncommon. These errors decreased as a function of TMR_{be}; almost no *combination* errors occurred for $TMRs_{be}$ of −20 dB and greater, and there were no significant differences between the spatially separated and co-located configurations.

C. Spatial gains

For each individual subject, logistic fits for the two co-located configurations were derived and averaged, as were the logistic fits for the two spatially separated configurations (after accounting for the TMR at the better ear; see Appendix for details). Between 30 and −20 dB TMR_{be} the *vertical difference between these averaged spatially separated and averaged co-located logistic fits* (the percent spatial gain) was

approximately 20% for subjects S1 and S4, and 7% for subjects S2 and S3. At the greatest $TMRs_{be}$, the percent spatial gain was between 20% (S1 and S4) and 25% (S2 and S3). The *horizontal shift between the co-located and separated configurations* (the dB spatial gain) was approximately 6 dB for subjects S1 and S4 and 2 dB for subjects S2 and S3 (in the performance range between 40% correct and 60% correct).

IV. DISCUSSION

In this selective attention task, listeners were asked to report the content of the message that contained a particular call sign (Brungart, 2001a). This target message was usually presented at a lower level than the fixed-level masker. To perform well in this selective task, listeners had to be able to segregate the (monosyllabic) target keywords from the acoustic mixture and report the proper keywords. Both energetic masking and informational masking interfered with performance in this task. Two factors emphasize the role of informational masking in the current study, at least at TMRs of −10 dB and above. First, target and masker messages were presented in nonoverlapping spectral bands (for a detailed analysis of simulated auditory nerve firing patterns for these types of stimuli see Shinn-Cunningham *et al.*, 2005a). Second, the target and masker were designed to be perceptually similar (e.g., acoustically, semantically, linguistically, etc.).

A. Cues for distinguishing target and masker

Target and masker keywords were nearly synchronous, and no semantic cues aided in distinguishing the target from the masker message. Because of the way the stimuli were processed, the messages did not have a strong pitch. The timbres of the target and masker signals varied unpredictably from trial to trial and these timbre differences were not very salient, so that it is unlikely that listeners relied on timbre to select the target. Thus, relatively few cues were available to help listeners differentiate the target from the masker. When target and masker were spatially separated, both spatial location and level differences between target and masker could be used to select target segments or the target stream from the mixture. However, when target and masker were co-located, the main cue enabling target selection was the level difference between target and masker (in those trials where target and masker had different levels).

B. Energetic masking and informational masking change systematically with TMR_{be}

Energetic masking is reduced and informational masking further emphasized when target and masker speech are presented in spectrally interleaved narrowbands (Arbogast *et al.*, 2005). In addition, several studies suggest that the amount of interference from informational masking is large in selective listening tasks that use the Coordinate Response Measure (CRM) corpus (Bolia *et al.*, 2000; Brungart, 2001a; Brungart *et al.*, 2005; Kidd *et al.*, 2005b; Wightman *et al.*, 2006). To emphasize the effects of informational masking, the current study employed spectrally interleaved bandpass filtered target and masker speech that were both derived from the CRM corpus.

We expected energetic masking to dominate at low $TMRs_{be}$ and to decrease with increasing TMR_{be}. At low $TMRs_{be}$ (−40 dB to −20 dB), the most common response errors were *drop error*s (nearly 60% of all trials), suggesting that indeed, at these low target levels, energetic masking was the dominant form of masking. In addition, *masker error*s occurred at a rate well above chance. In some ways, it is surprising that *masker error*s existed at all at these low $TMRs_{be}$: listeners were reporting the content of the more intense talker, which they should have realized was the masker and thus excluded from their response. This kind of result, where listeners seem unable to ignore a talker that they should know is the masker, has been observed in other studies of informational masking (e.g., Brungart and Simpson, 2004; Kidd *et al.*, 2005a). Such errors may occur because listeners are not completely certain that the message they heard was from the masker, or because it is too confusing to switch to a strategy of guessing a word not heard in these trials while still reporting the heard words in the other trials. Regardless, these *masker error*s at low TMR_{be} likely reflect a failure to hear the target (energetic masking).

Both *drop* and *masker error*s decreased as TMR_{be} increased from −40 dB to −20 dB, consistent with a decrease in energetic masking. Moreover, spatial separation yielded a slightly lower rate of both of these energetic-masking-caused errors, supporting the idea that binaural decorrelation processing provides a small release from energetic masking for the low-frequency portions of the target within 10–15 dB of masked threshold (Zurek, 1993; Durlach, 1972; Shinn-Cunningham *et al.*, 2005a). This release from energetic masking due to binaural decorrelation may not have required *perceived* spatial differences between target and masker (Colburn and Durlach 1965; Edmonds and Culling, 2005a; Edmonds and Culling, 2005b; Culling *et al.*, 2006), but simply a change in interaural correlation caused by the addition of the near-threshold signal (when the masker energy falling within the target bands can rival the target energy).

The pattern of errors was very different for $TMRs_{be}$ of −10 dB and greater. As the TMR_{be} increased from −20 dB, *drop* and *combination* errors disappeared, suggesting that energetic masking became negligible at the mid- to high-range $TMRs_{be}$. In this range, target-masker similarity of level and location (see Secs. IV D and IV E, below) determined how well listeners could extract the target from the mixture. *Masker* and *mix error*s were more likely when target and masker were co-located. This shows that once the target was audible and properly segmented from the acoustic mixture, informational masking dominated. Moreover, this pattern suggests that when the listener had trouble selecting the target from the properly segmented mixture, perceived spatial location was a salient, robust cue for identifying the target segments and/or target stream.[3]

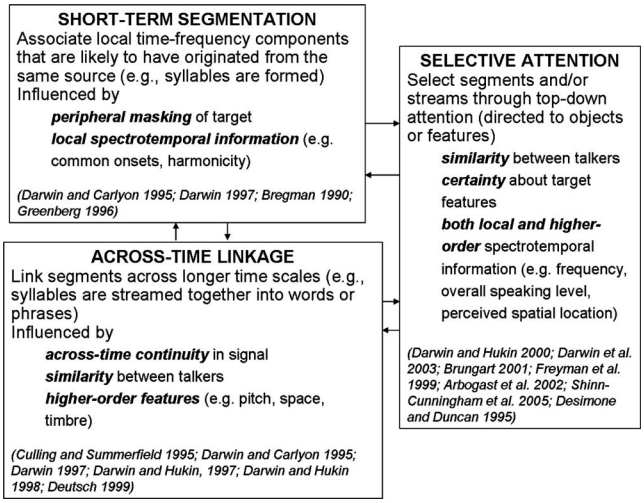

FIG. 4. Flow chart of the proposed conceptual framework of masking.

C. Conceptual framework explaining different forms of masking

At least three intricately linked mechanisms should affect masking in this kind of speech identification task: (1) short-term segmentation, (2) across-time linkage, and (3) selective attention (see definitions below and Fig. 4).

(1) Short-term segmentation is defined as the process by which all or part of the acoustic mixture is automatically segregated into local time-frequency components that are likely to have originated from the same source (e.g., see Bregman, 1990). Although segmentation may be influenced by attention, it is primarily based on the primitive spectro-temporal structure of the sound sources (e.g., see Darwin and Carlyon, 1995; Darwin, 1997). We assume that in the current task, when energetic masking is low, short-term segmentation can robustly extract speech segments on the time scale of syllables (Greenberg, 1996). Conversely, when energetic masking is high enough, target segmentation should fail, forcing listeners to guess the target keywords (or to adopt a different response strategy such as reporting the masker message).

(2) Temporal discontinuities (e.g., stop consonants, silent gaps) limit the duration of segments. Proper stream formation depends on "across-time linkage," the process of binding short-term segments across such discontinuities. Previous studies suggest that continuity of higher-order features such as timbre, perceived location, and overall intensity are important for across-time linkage (e.g., see Culling and Summerfield, 1995; Darwin and Carlyon, 1995; Darwin, 1997; Darwin and Hukin, 1997; Darwin and Hukin, 1998; Deutsch, 1999).

(3) Even if short-term segmentation and across-time linkage are performed flawlessly, listeners still must choose the correct stream from a sound mixture using "selective attention," a mechanism that may be directed both to local and higher-order spectro-temporal information (Freyman *et al.*, 1999; Darwin and Hukin 2000; Arbogast *et al.*, 2002; Darwin *et al.*, 2003; Shinn-Cunningham *et al.*, 2005a). Depending on the degree of similarity between target and interferers and the amount of certainty the listener has about the target features, selective attention can enhance sounds with desirable features and suppress others by biasing the sensory representation (e.g., Desimone and Duncan, 1995), thereby bringing a selected object into the perceptual foreground.

In performing the current selective listening task, listeners may use one of two response strategies. First, listeners may select all segments or the stream with a desired feature. In that case, the listener either must know these target features ahead of time, or they must estimate features of the target call sign when it occurs and selectively attend to the streams or segments with the estimated features of the target call sign. Alternatively, if listeners link segments into proper streams using across-time continuities inherent in each message, they may solve the task by attending to the stream that contains the target call sign. In fact, in conditions where higher-order acoustic features do not disambiguate which segments or which stream to attend to (e.g., when two perceptually similar messages are presented, and only a call sign within the target utterance defines which message is the target), listeners may have to use this strategy to solve the task. Specifically, given that timbre differences are relatively weak and unreliable in this task, listening for the target call sign and then properly linking it to the subsequent target words may be the only way to perform the task in the co-located configuration when the TMR_{be} is 0 dB. How do the current results support the conceptual framework? The following two sections shed light on this question.

D. Different mechanisms underlie masker errors and mix errors

Even though listeners were asked to report only the target message, listeners could almost always understand both the target and the masker messages when the TMR was −10 dB or higher and the target could be segmented properly.[4] Based on the conceptual framework (Sec. IV C), this observation suggests that *masker* errors occur either because the listener independently selects the wrong keywords for both the color and the number, or because the listener selects the wrong one of two properly formed streams. In contrast, *mix* errors may occur when the listener independently selects one correct keyword and one masker keyword, or when the listener selects a perceptual stream that is a mix of target and masker keywords. Both of these possibilities occur only when the listener fails to link keywords properly across time.[5]

In the current study, TMR affects performance in two ways. In general, listeners perform better with increasing target level. However, because *differences* in the levels of target and masker provide a cue to select the target segments or streams, performance does not improve monotonically with TMR_{be}. When target and masker are equally intense and co-located, subjects perform noticeably worse than when the masker is 10 dB more intense than the target. Thus, level must provide a cue that aids the target selection and/or improves the proper across-time linkage of the target keywords.

Looking in even more detail at the errors in the co-located configurations, the percentage of *masker* errors shows a pronounced peak at 0 dB TMR_{be}, while, in stark contrast, the percentage of *mix* errors does not. This differ-

ence in the patterns for *masker* and *mix* errors suggests that at least two different mechanistic failures contribute to informational masking: failures in across-time linkage of segments and failures in the selection of target segments and/or the target stream. In particular, mix errors are no more common at 0 dB TMR_{be} than at −10 or 10 dB TMR_{be}. Therefore, the across-time linkage of the target keywords is unaffected by TMR_{be}, suggesting that level cues do not provide a strong cue for streaming the keywords together. In contrast, masker errors are more common at 0 dB TMR_{be} than at −10 dB and 10 dB TMR_{be}. Thus, attention can be directed to a source based on a level difference between target and masker (explaining the jump in masker errors at 0 dB TMR_{be}).

Previous studies found that the usefulness of level cues depends on the types of stimuli used. The utility of intensity differences between the target and masking talkers decreases in importance as the number of maskers increases (Brungart *et al.*, 2001a; Freyman *et al.*, 2004), and is reduced when the masker is very different from the target in perceptual quality (Brungart 2001c). These results are consistent with the idea that level helps in selecting the target from the sound mixture (a process that should get more and more challenging the greater the number of competing talkers), but is redundant when other cues differentiate target and masker. None of these results suggest a role of level in automatic streaming of utterances.

E. Spatial release from masking

At the lowest $TMRs_{be}$, there is no difference in the rate of *mix error*s for co-located and spatially separated sources. This is consistent with previous studies that suggest that spatial release from informational masking depends on perceiving competing streams from different locations (Freyman *et al.*, 2001; Arbogast *et al.*, 2002; Shinn-Cunningham *et al.*, 2005a; Gallun *et al.*, 2005). Evidence suggests that objects start to be perceptually separated before they are heard at different locations (Woods and Colburn, 1992; Litovsky and Shinn-Cunningham, 2001; Best *et al.*, 2007). The current results are consistent with the idea that spatial release from informational masking only occurs when syllables are properly segmented and syllables are perceived at different locations, and that perception of the segmented target at the correct location only occurs at high $TMRs_{be}$, above the $TMRs_{be}$ that first allow the target to be segmented from the mixture.

In principle, spatial similarity could also cause difficulty in segmenting the two messages, by increasing uncertainty about which time-frequency components belong to which message. However, past studies show that spatial cues have little influence on short-term segmentation (Darwin, 1997).

At the mid- to high-range $TMRs_{be}$, when the target is intense enough to be segmented from the masker, the spatial gains in performance are markedly greater than at the low $TMRs_{be}$. The spatial release from masking at these $TMRs_{be}$ is primarily caused by a reduction in *masker* errors and *mix* errors in the spatially separated configurations compared to the co-located configurations. Together with the interpretation of how *masker* errors and *mix* errors are affected by across-time linkage and selective attention (see Sec. IV C and IV D), the spatial differences in the pattern of *masker* errors suggest that spatial cues improve the ability to select either independent target segments or a properly formed target stream.

The rate of *mix* errors decreases as TMR_{be} increases from −20 dB for spatially separated sources, but is essentially constant for co-located sources. At first glance, this reduction in *mix* errors appears to show that listeners can use spatial location as a cue for linking segments across time. However, because mix errors can also occur when listeners independently select one correct and one wrong segment, this spatial improvement in mix errors may merely reflect an improvement in the likelihood that listeners independently select the correct target keywords in the spatially separated configurations compared to the co-located configurations. Overall, these results are consistent with previous studies suggesting that spatial location can help listeners selectively attend to already-formed syllables (e.g., see Darwin and Hukin, 1999).

V. SUMMARY AND CONCLUSIONS

The results of this selective listening task give strong evidence that the relative influences of energetic masking and informational masking change systematically as a function of TMR_{be}. The pattern of results is consistent with the idea that different attributes of two competing signals can be used to select a target out of the mixture and to link short-term segments across time, including level differences between target and masker and the spatial cues that were the main focus of this study.

The pattern of errors as a function of the level difference between target and masker suggests that distinct mechanisms contribute to the types of errors in this selective speech identification task. In particular, *drop* errors appear to be caused predominantly by energetic masking; *masker* errors are most likely caused by energetic masking at low $TMRs_{be}$ and failures in selective attention at higher $TMRs_{be}$; and *mix* errors are most likely to occur when both across-time linkage and selective attention fail.

Spatial separation improves performance at all $TMRs_{be}$; however, the improvements come from different mechanisms at different $TMRs_{be}$. At the lowest TMR_{be}, binaural processing reduces energetic masking of the target, which, in turn, makes the target easier to segment from the mixture. There is no evidence that spatial cues improve selection or across-time linkage at these low $TMRs_{be}$. At higher $TMRs_{be}$, spatial release occurs by increasing the likelihood that the listener selects the correct keywords or the correct stream out of the mixture. The data hint that spatial differences between target and masker may also improve across-time linkage of syllables, but this conclusion is confounded by the possibility that selective attention alone may reduce the probability of selecting the color and number (independently), which would in turn reduced *masker* as well as *mix* errors. Finally, level differences between target and masker allow a listener to select the proper keywords from a mixture, but do not improve the perceptual linkage of the adjacent keywords into a single stream.

TABLE II. 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 upper asymptote of performance is higher for spatially separated sources than for co-located sources, but 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.

	T0M0	T90M90	T0M90	T90M90
(A) Midpoint of dynamic range α [dB]	27.4 (1.4)	24.5 (1.6)	23.5 (1.8)	24.2 (1.0)
(B) Slope at the midpoint of dynamic range $1/\beta$ [% correct / dB]	27.8 (11.0)	24.8 (16.3)	16.7 (4.5)	19.0 (3.4)
(C) Upper asymptote of performance $1-\lambda$ [% correct]	71.3 (4.6)[a]	72.4 (5.2)[a]	96.5 (2.0)	93.9 (4.2)

[a]Statistically significant difference between co-located and spatially separated configurations.

ACKNOWLEDGMENTS

This work was supported in part by grants from AFOSR and NIDCD. The authors are grateful to Gerald Kidd, Virginia Best, Christine Mason, Frederick Gallun, Steve Colburn, Richard Freyman, and three anonymous reviewers for their helpful comments.

APPENDIX

For each of the four spatial configurations, and separately for each subject, percent correct performance as a function of TMR_{be} was fitted by a logistic function using a maximum-likelihood method with bootstrapping, *psignifit* version 2.5.6 (http://www.bootstrap-software.com/psignifit/, see Wichmann and Hill, 2001a). The probability of responding correctly at a given TMR, $\hat{P}(x)$, equals

$$\hat{P}(x) = \gamma + (1-\lambda-\gamma)\frac{1}{1+e^{\alpha-x/\beta}}, \tag{A1}$$

where γ is the lower bound on performance, $1-\lambda$ is the upper bound on performance at the largest TMR, α is the energy ratio at which percent correct performance is exactly halfway between chance and the best observed performance, and $1/\beta$ is the slope of the psychometric function evaluated at $x=\alpha$. Note that this fitting algorithm places a higher emphasis on the steep portion of the psychometric function than a minimum least square fitting constraint would have.

The lower bound on performance γ was set to 6%, chance level assuming that listeners hear and rule out the masker keywords. Although there is ample evidence that listeners do not always do this, the fits were quantitatively better when setting chance performance to 6% rather than the 4% that would occur if listeners chose randomly among all keywords.

The goodness of fit of the psychometric functions was evaluated using Efron's bootstrap technique (Wichmann and Hill, 2001a, Wichmann and Hill, 2001b). Residual differences between the predictions from the fits and data were compared to the error residuals between the predictions and 10 000 runs of Monte Carlo simulated data sets (whose statistics equaled the estimated distribution of the data). A deviance measure was calculated as described in Eq. (A2) (for detailed discussion see Wichmann and Hill, 2001a):

$$D = \sum_{i=1}^{K}\left[n_i y_i \log\left(\frac{y_i}{\hat{p}_i}\right) + n_i(1-y_i)\log\left(\frac{1-y_i}{1-\hat{p}_i}\right)\right], \tag{A2}$$

where y_i denotes performance (either measured or simulated) at the TMR denoted by i, $\hat{p}_i$ is the percent correct predicted by Eq. (A1) for the corresponding TMR, n_i is the number of trials at each TMR ($n_i=96$ for all i), and K equals the number of TMRs tested ($K=6$). The deviance D_{measured} between predictions and measured data was calculated. Similarly, the deviances were calculated for each of the 10 000 sets of simulated data, yielding a set of D^*. Based on the Monte Carlo generated distribution of these 10 000 values of D^*, 95% confidence intervals for D_{measured} were then estimated.

For each measured data set for which D_{measured} falls within the 95% confidence interval, the fitting function was considered to provide a good description of the underlying data. Twelve of the 16 fits (four functions for each of four listeners) fell within the 95% confidence intervals of the distribution fits. In the four cases that did not meet this criterion, the largest errors in the predictions were consistently due to performance dips in the measured data at 0 dB TMR_{be} for co-located configurations. However, those data points cannot be fit by any monotonically increasing function. Other than missing these dips, the fits were deemed adequate descriptions of the data.

The resulting parameters, averaged across subjects, are listed in Table II. Unlike the patterns of the raw percent correct performance, the normalized midpoint parameters α of the psychometric functions (left panel) and the normalized slopes at the midpoints of the psychometric functions, $1/\beta$ (center panel) do not vary significantly with spatial configuration [T-test; $p>0.01$]. The upper bounds $1-\lambda$ are lower in the co-located than in the separated configurations, reflecting the lower level of performance for co-located configurations at the greatest TMRs (T-test; $p<0.01$).

At first glance, the fact that the upper bounds $1-\lambda$ are the only parameters that differ significantly across co-located and spatially separated configurations may appear counterintuitive. While the raw performance data differ for spatially

separated and co-located configurations, when normalized so that they range between 0% and 100%, the logistic fits have similar midpoints and slopes.

The parameters α and $1/\beta$ capture differences in midpoint and slope *relative* to the lower and upper limits of performance. Specifically, the midpoint parameter α is the TMR at which performance is halfway between the lower and upper bounds on performance for a given psychometric function, which will be different absolute levels of performance if the upper bound (parameter $1-\lambda$) varies with condition. Similarly, the slope parameter quantifies the percent of change in performance between lower and upper bounds of the psychometric function per dB, not the change in percent correct per dB, and so will translate to different absolute %/dB slopes if the upper bound varies with condition.

To the extent that logistic fits are adequate descriptions of the underlying data, this may suggest that a difference in the upper limits of the logistic fits between spatially separated and co-located configurations is sufficient to account for performance differences. Indeed, this is consistent with the idea that for the co-located configurations, the listener's attention may have been misdirected more often than in the spatially separated configurations, causing a decrease in asymptotic performance (cf. Lutfi *et al.*, 2003). However, fitting the nonmonotic performance function of our raw data with a monotonic logistic function conceals systematic differences in the midpoints of the raw performance data. In other words, this way of analyzing the data hides the fact that at a given TMR, spatial cues lead to absolute improvements in the ability to select the target keywords from the mixture.

[1]The lack of differences between these two sets of instructions suggests that, across all spatial configurations, listeners did not benefit from *a priori* knowledge of the target location. However, listeners could have computed the spatial location of the target call sign in the first few trials of a run, directed attention to that estimated location on subsequent trials, and then selected the keywords based on their perceived locations. Thus, this lack of any effect of instructions may simply reflect the fact that listeners may have adopted a strategy in which they directed attention to the target location, independent of the instructions.

[2]In each block, the target is softer than the masker in 67% of trials, allowing listeners to perform relatively well simply by focusing on the less-intense talker. The dip in performance at 0 dB TMR_{be} has been attributed to the loss of this relative level cue for selecting target words from the mixture. Not all studies show a drop in performance at 0 dB TMR (e.g., see Arbogast *et al.*, 2002). Similarly, not all of them show an upper bound of 80% (rather than 100%) correct. However, those studies that do not show a drop in performance for equal intensity target and masker (and that had higher high-TMR performance) generally used more speech bands for the target than for the masker, or they used full speech. This may have made the target more salient, even when it was the same broadband level as the masker, and thus easier to understand than the masker. This is in line with findings by Brungart and colleagues, who show that the amount of across-ear interference in a dichotic masking paradigm increases with the number of masker bands for amplitude-modulated sine-wave carrier speech as well as modulated-noise-band speech (Brungart *et al.*, 2005), a result that suggests that performance decreases as the intelligibility of an informational masker increases.

[3]Unlike the processed speech used in the current study, in ordinary conversational speech within-stream continuity cues are much stronger, and pitch, semantic, and linguistic information help listeners to link syllables and words across time. The stimuli used here are likely to make it more difficult to properly track keywords from a target message across time compared to normal everyday discourse. This may cause listeners to rely more heavily on other cues in the stimuli (level, location) important for auditory scene analysis. These stimuli allow us to tease apart whether level and location contribute to across-time linkage and/or selection of keywords and/or streams.

[4]The listeners who participated in this experiment also participated in a companion study with identical stimuli in which they were asked to report keywords from both utterances (Ihlefeld and Shinn-Cunningham, submitted). Listeners could report all four keywords (of both target and masker) nearly as well as they could report the keywords of the target message. Within each of the spatial configurations, the percent correct performance in this selective task (reporting both target keywords correctly) is nearly equal to percent correct performance in the divided task in the companion study (reporting all four keywords correctly), never differing by more than 10%.

[5]Failures in short-term segmentation, across-time linkage, and/or selection can occur at any time during the presentation of the target and masker utterances. However, we only measured performance for color and number keywords. Therefore, based solely on our results, we cannot determine how across-time linkage between (for instance) the call sign and color depends on stimulus manipulations.

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.

Arbogast, T. L., Mason, C. R., and Kidd, G., Jr. (**2005**). "The effect of spatial separation on informational masking of speech in normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **117**, 2169–2180.

Best, V., Gallun, F. J., Carlile, S., and Shinn-Cunningham, B. G. (**2007**). "Binaural interference and auditory grouping," J. Acoust. Soc. Am. **121**, 1070–1076.

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.

Bregman, A. S. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA).

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

Brungart, D. S. (**2001b**). "Evaluation of speech intelligibility with the coordinate response measure," J. Acoust. Soc. Am. **109**, 2276–2279.

Brungart, D. S., Simpson, B. D., Ericson, M. A., and Scott, K. R. (**2001a**). "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. (**2004**) "Within-ear and across-ear interference in a dichotic cocktail party listening task: Effects of masker uncertainty," J. Acoust. Soc. Am. **115**, 301–310.

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.

Colburn, H. S., and Durlach, N. I. (**1965**) "Time-intensity relations in binaural unmasking," J. Acoust. Soc. Am. **38**, 93–103.

Culling, J. F., and Summerfield, Q. (**1995**). "Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay," J. Acoust. Soc. Am. **98**, 785–797.

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.

Darwin, C. J. (**1997**). "Auditory grouping," Trends Cogn. Sci. **1**, 327–333.

Darwin, C. J., and Carlyon, R. P. (**1995**). "Auditory grouping," in *Hearing*, edited by B. C. J. Moore (Academic, San Diego), pp. 387–424.

Darwin, C. J., and Hukin, R. W. (**1997**). "Perceptual segregation of a harmonic from a vowel by interaural time difference and frequency proximity," J. Acoust. Soc. Am. **102**, 2316–2324.

Darwin, C. J., and Hukin, R. W. (**1998**). "Perceptual segregation of a harmonic from a vowel by interaural time difference in conjunction with mistuning and onset asynchrony," J. Acoust. Soc. Am. **103**, 1080–1084.

Darwin, C. J., and Hukin, R. W. (**1999**). "Auditory objects of attention: The role of interaural time differences," J. Exp. Psychol. Hum. Percept. Perform. **25**, 617–629.

Darwin, C. J., and Hukin, R. W. (**2000**). "Effectiveness of spatial cues, prosody, and talker characteristics in selective attention," J. Acoust. Soc. Am. **107**, 970–977.

Darwin, C. J., Brungart, D. S., and Simpson, B. D. (**2003**). "Effects of

fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers," J. Acoust. Soc. Am. **114**, 2913–2922.
Desimone, R., and Duncan, J. (**1995**). "Neural mechanisms of selective visual attention," Annu. Rev. Neurosci. **18**, 193–222.
Deutsch, D. (**1999**). "Grouping mechanisms in music," in *The Psychology of Music*, 2nd ed., edited by D. Deutsch (Academic, San Diego).
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. (**1972**). "Binaural signal detection: Equalization and cancellation theory," in *Foundations of Modern Auditory Theory*, edited by J. Tobias (Academic, New York), pp. 369–463.
Edmonds, B., and Culling, J. (**2005a**). "The role of head-related time and level cues in the unmasking of speech in noise and competing speech," Acta Acust. **91**, 546–553.
Edmonds, B. A., and Culling, J. F. (**2005b**). "The spatial unmasking of speech: Evidence for within-channel processing of interaural time delay," J. Acoust. Soc. Am. **117**, 3069–3078.
Edmonds, B. A., and Culling, J. F. (**2006**). "The spatial unmasking of speech: Evidence for better-ear listening," J. Acoust. Soc. Am. **120**, 1539–1545.
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.
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 masking talkers and auditory priming on informational masking in speech recognition," J. Acoust. Soc. Am. **115**, 2246–2256.
Freyman, R. L., Helfer, K., and Balakrishnan, U. (**2007**). "Variability and uncertainty in masking by competing speech," J. Acoust. Soc. Am. **121**, 1040–1046.
Gallun, F., Mason, C., and Kidd, G. J. (**2005**). "Binaural release from informational masking in a speech identification task," J. Acoust. Soc. Am. **118**, 1614–1625.
Greenberg, S. (**1996**). "Understanding speech understanding: Towards a unified theory of speech perception," in *Proceedings of the ESCA Workshop on the "Auditory basis of speech perception,"* Keele University, Keele, England, pp. 1–8.
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. (**1948**). "The influence of interaural phase on interaural summation and inhibition," J. Acoust. Soc. Am. **20**, 536–544.
Ihlefeld, A., and Shinn-Cunningham, B. G.. "Spatial release from energetic and informational masking in a divided speech identification task." J. Acoust. Soc. Am. **123**, 4380–4392.
Kidd, G. J., Arbogast, T., Mason, C., and Gallun, F. (**2005a**). "The advantage of knowing where to listen," J. Acoust. Soc. Am. **118**, 3804–3815.
Kidd, G. J., Mason, C., and Gallun, F. (**2005b**). "Combining energetic and informational masking for speech identification," J. Acoust. Soc. Am. **118**, 982–992.
Litovsky, R. Y., and Shinn-Cunningham, B. G. (**2001**). "Investigation of the relationship among three common measures of precedence: Fusion, localization dominance, and discrimination suppression," J. Acoust. Soc. Am. **109**, 346–358.
Lutfi, R. A. (**1993**). "A model of auditory pattern analysis based on component-relative entropy," J. Acoust. Soc. Am. **94**, 748–758.
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.
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. (**2005a**). "Bottom-up and top-down influences on spatial unmasking," Acta Acust. **91**, 967–979.
Shinn-Cunningham, B. G., Kopco, N., and Martin, T. J. (**2005b**). "Localizing nearby sound sources in a classroom: Binaural room impulse resonses," J. Acoust. Soc. Am. **117**, 3100–3115.
Van Engen, K. J., and Bradlow, A. R. (**2007**). "Sentence recognition in native- and foreign-language multi-talker background noise," J. Acoust. Soc. Am. **121**, 519–526.
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.
Wightman, F. L., and Kistler, D. J. (**2005**). "Informational masking of speech in children: Effects of ipsilateral and contralateral distractors," J. Acoust. Soc. Am. **118**, 3164–3176.
Wightman, F. L., Kistler, D. J., and Brungart, D. (**2006**). "Informational masking of speech in children: Auditory-visual integration," J. Acoust. Soc. Am. **119**, 3940–3949.
Woods, W. S., and Colburn, H. S. (**1992**). "Test of a model of auditory object formation using intensity and interaural time difference discrimination," J. Acoust. Soc. Am. **91**, 2894–2902.
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).