Listening to every other word: Examining the strength of linkage variables in forming streams of speech

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

(Received 9 April 2008; revised 12 September 2008; accepted 17 September 2008)

In a variation on a procedure originally developed by Broadbent [(1952). "Failures of attention in selective listening," J. Exp. Psychol. **44**, 428–433] listeners were presented with two sentences spoken in a sequential, interleaved-word format. Sentence one (target) comprised the odd-numbered words in the sequence and sentence two (masker) comprised the even-numbered words in the sequence. The task was to report the words in sentence one. The goal was to determine the effectiveness of cues linking the words of the target (or masker) over time. Three such "linkage variables" were examined: (1) fixed talker, (2) fixed perceived interaural location, and (3) correct syntactic structure. All of the linkage variables provided a significant advantage when applied to the target compared to the baseline condition in which the linkage variables were randomized. However, these linkage variables were not effective when applied to the masker. Word position effects were found such that performance in the baseline condition declined, and the advantages of the linkage variables increased, for the words near the end of the sentence. Overall, this approach appears to be useful for examining interference in speech recognition that has little or no peripheral component. The results suggest that variables that link target words together improve their resiliency to interference and/or their recall. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2998980]

PACS number(s): 43.66.Dc, 43.66.Mk, 43.66.Ba [MW] Pages: 3793–3802

I. INTRODUCTION

In multisource acoustic environments, a listener is faced with the task of determining the various sources of sound and selecting one or more to attend to, while ignoring others that are irrelevant or unimportant. According to Yost (1991, 2008), sound source determination is fundamental to the perceptual organization of the auditory environment and to making sense of the information the sources convey. Individual sounds are sometimes referred to as auditory "objects," which reflects the view that sounds are typically segmented into discrete perceptual elements (e.g., Woods and Colburn, 1992; Griffiths and Warren, 2004; and Shinn-Cunningham, 2008). When a sequence of related elements occurs the listener may perceive an auditory "stream" that can be segregated from other unrelated sounds. Understanding the factors that govern stream formation and maintenance over time is a topic of considerable interest to auditory researchers.

Simple stimuli, such as pure tones, may be arranged in spectrotemporal patterns that clearly lead to the formation of auditory streams that may easily be segregated from other streams (e.g., reviews by Bregman, 1990; Darwin and Carlyon, 1995; and Moore and Gockel, 2002). However, more complex sounds are also organized into and perceived as auditory streams. One important class of such sounds is speech. When multiple talkers are present in an auditory sound field there are typically many cues that may be used to segregate and attend to one particular "target" voice. The fundamental frequency and intonation contour of the talker, as well as a variety of talker-specific acoustic properties, forms one type of cue while another important cue is the spatial location of the talker. Furthermore, because connected speech conveys meaning there are semantic and syntactic continuities that also allow the listener to follow one specific speech stream. Thus, there are normally a variety of acoustic, perceptual, and semantic factors that allow the elements of a speech stream to be linked together and separated from other sounds.

One inherent complication in fully describing the factors governing selective listening in multisource environments is that the sounds produced by the various sources interfere with one another. From the perspective of the listener, irrelevant sounds that interfere with the detection, identification, recognition, or extraction of some other aspect of the target sound cause "masking." From a theoretical perspective, masking is often considered to be composed of two distinctly different components, referred to as "energetic masking" and "informational masking." Energetic masking results from the competition between the target and masker(s) for an adequate representation in the various auditory structures beginning with the basilar membrane and continuing through the ascending auditory neural pathways. In contrast, informational masking occurs despite a presumably adequate neural representation of the target (at some neural level, i.e., the auditory periphery) and is thought to reflect the limitations or natural tendencies of a variety of perceptual and cognitive processes (cf. Kidd *et al.*, 2008).

From an experimental perspective, separating out energetic and informational masking in complex multisource environments is challenging. When two or more talkers are speaking simultaneously, for example, the spectral overlap of

[a]Author to whom correspondence should be addressed. Electronic mail: gkidd@bu.edu

the voices—which leads to energetic masking—varies from moment to moment and it may be difficult to determine how much of the masking that is observed is due to energetic masking and how much is due to informational masking. Informational masking appears to be particularly important to understand when speech is masking other speech because it is often the case that failure in solving the speech recognition task is not due to the inaudibility of an energetically masked target but instead is due to a failure to segregate or focus attention on the target (Brungart *et al.*, 2006).

One experimental approach to separating energetic and informational masking in the multitalker listening situation was proposed by Arbogast *et al.* (2002). They sought to minimize energetic masking by reducing the spectral overlap of the target and masker. This was accomplished by processing the speech into sets of narrow frequency bands such that target bands and masker bands were mutually exclusive and thus had greatly reduced spectral overlap compared to natural speech while maintaining high intelligibility. Furthermore, the speech materials used in their study were from the coordinate response measure test (Bolia *et al.*, 2000) that is known to produce large amounts of informational masking under many conditions. Their results did indeed indicate that large amounts of informational masking were produced concurrently with small amounts of energetic masking, allowing them to test the potency of spatial separation of sources as a cue to reducing informational masking.

However, the approach used by Arbogast *et al.* (2002) has some limitations. The processing of speech into narrow bands reduces, but does not eliminate energetic masking. In certain subject populations—those with sensorineural hearing loss, for example—the energetic/informational ratio may be different than in other populations, a conclusion reached based on the findings of a later study by Arbogast *et al.* (2005). Furthermore, speech processed in this manner lacks strong pitch or intonation cues, and the vocal characteristics that distinguish different talkers are reduced, limiting the investigation of those potentially important cues in the selective listening task.

In the current study, a temporal analog of the approach used by Arbogast *et al.* (2002) is investigated. This approach is based on one originally (to our knowledge) employed by Broadbent (1952). Instead of confining target and masker elements to mutually exclusive frequency regions, here the target and masker elements are confined to mutually exclusive temporal regions. Target and masker words are presented alternately in interleaved sequences with the task of the listener being to report the words contained in the target sequence. Because the words do not overlap in time, they also do not simultaneously overlap in frequency, so, by most definitions of energetic masking, little if any energetic masking would be expected. The possibility of nonsimultaneous energetic masking cannot be excluded and is considered later in this article.

The underlying premise upon which this work is based is that the ability to follow and subsequently report the words in the target sequence depends on the strength of the cues that link the words together. That is, the stronger the linkage between target elements the better the listener will be in solving the speech identification task. This idea may be relevant to natural speech settings in which listeners must often link disconnected "glimpses" of a speech target when it is partially masked or interrupted by other sounds.

TABLE I. The 40-word corpus arranged in eight rows and five columns where each column is a word category (name, verb, number, adjective, and noun) and a choice of one word from each column in order from left to right would produce a syntactically correct but unpredictable sentence.

Bob	Bought	Two	Big	Bags
Gene	Found	Three	Blue	Cards
Jane	Gave	Four	Cold	Gloves
Jill	Held	Five	Hot	Hats
Lynn	Lost	Six	New	Pens
Mike	Saw	Eight	Old	Shoes
Pat	Sold	Nine	Red	Socks
Sue	Took	Ten	Small	Toys

The general approach of interleaving target and masker sequences is extremely flexible and many types and strengths of linkages are possible. In this study a new version of the interleaved-word paradigm and new stimulus set are described, and the results of an initial examination of a small set of linkage variables are presented. In addition, given that simple tonal stimuli stream more readily at faster rates (van Noorden, 1975), two different rates of presentation are examined. Furthermore, the current work examines the effects of linking together target versus masker words and the implications of that comparison for the ability to selectively attend to one source or actively ignore a second source (see also de Cheveigné *et al.*, 1995 and Brungart and Simpson, 2007).

II. METHODS

A. Subjects

Seven young adult listeners (ages 21–25) participated in the experiments. Four completed experiment 1 and four completed experiment 2. Listener S1 was common to both experiments and completed experiment 2 before experiment 1. The listeners were screened audiometrically to ensure that they had normal hearing (thresholds equal to or less than 20 dB hearing level) for octave frequencies between 250 Hz and 8 kHz. The listeners were paid for their participation.

B. Stimuli

A laboratory-designed corpus of monosyllabic words was recorded by Sensimetrics, Inc. (Somerville, MA) for use in the experiment. The corpus is comprised of 40 words in five categories (eight names, eight verbs, eight numbers, eight adjectives, and eight nouns). The words are contained in Table I. A choice of one word from each category in order (as listed above or in column order from left to right in the table) would yield a syntactically correct yet unpredictable sentence. The entire corpus includes recordings of each word spoken by 16 native speakers of American English (8 male, 8 female). The words were recorded in isolation with neutral inflection rather than in natural sentence form so that all possible combinations of words could be selected in a closed

set test and the effects of coarticulation across word boundaries, and imprecise or "smeared" word boundaries themselves, would be eliminated. Furthermore, the design of these experiments required that individual words be available to be interleaved with other sounds as described in the conditions below.

In the current experiments, the words from a subset of the talkers (three male, three female) were used. The words varied between 385 and 1051 ms in duration with a mean of 624 ms. The main experimental conditions consisted of five-word target strings interleaved with five-word masker strings. The target was defined as the five odd-numbered words in the sequence (i.e., first, third, fifth, etc.) while the masker comprised the even-numbered words in the sequence, the result being a ten-word string. All concatenation was done with no overlap and no silent gaps. Individual words were scaled to the same rms amplitude and no attempt was made to equate word duration across words or talkers. In control conditions, the masker words were replaced with silence, Gaussian noise, or reversed speech tokens. When silent gaps were used, the duration of each silence was set to the length of a randomly selected word from the corpus and therefore varied in the same way that masker words in those positions would vary. The duration of a single noise token varied in the same way as the individual words (i.e., a masker word was chosen as it would have been in the speech masker conditions but instead a noise with equal duration and rms amplitude was created to replace it in the string), and a 50 ms cosine-squared ramp was applied to each onset and offset. When the masker tokens were reversed speech, the masker words were again chosen as in the speech masker conditions but the word was simply reversed in time prior to concatenation.

C. Procedures

Stimuli were presented over headphones (Sennheiser HD265 linear) to listeners seated in a double-walled Industrial Acoustics Corporation booth at approximately 77 dB sound pressure level (as measured on KEMAR). The stimuli were created on a PC located outside the booth and were D/A converted and attenuated using Tucker-Davis Technologies hardware before being routed to the headphones. A mouse and liquid crystal display monitor were located inside the booth to obtain listener responses. After a stimulus was presented, a graphical user interface showing a grid of the 40 words in the set (arranged in five columns and eight rows as per Table I) was displayed on the monitor, and the listener was required to select the five target words in the order they were presented by clicking on the grid with the mouse. The grid was not visible during the stimulus to avoid listeners responding before the entire stimulus was presented, "mapping out" their response pattern visually or leaving the mouse pointing to a response. Of course there are many memory aids in these tasks such as rehearsal that were not explicitly controlled.

D. Experimental conditions

Five different conditions were tested in experiment 1. In all five conditions, ten words (five target and five masker) of the 40 possible were selected at random without replacement. In the baseline condition (referred to as "random"), talkers of each word were chosen such that the voice varied randomly among the six choices (with replacement) across target and masker strings. In addition, each word was given a random interaural location by introducing a delay between the left and right headphones. These interaural time differences were selected from a set of six values (± 150, ± 450, and ± 750 μs). In two "fixed target" conditions, the target words were constrained to have either a fixed voice or a fixed location on each trial. In these cases, the fixed parameter value was prevented from also occurring in the masker sequence for that trial. In two "fixed masker" conditions, the masker words were constrained to have either a fixed voice or a fixed location on each trial, and the fixed parameter value could not occur in the target sequence. The five experimental conditions were tested at two different speech rates. One rate (called "normal rate") was simply the words as they were naturally spoken although the rate after concatenation of words spoken in isolation (of approximately 1.6 syllables/s) is slower than average normal speaking rate. Each condition was also tested using a rate that was twice as fast (called "fast rate"), or approximately 3.2 words/s. This was achieved by shortening each word in the corpus to half its original duration (while maintaining the original pitch) using the PRAAT software package (Boersma and Weenink, 2007).[1]

In experiment 2, the main experimental conditions manipulated whether the target words, masker words, or both were arranged in syntactically correct sentences. A syntactically correct sentence comprised one word from each column of Table I in order from left to right. These choices resulted in three experimental conditions: target-correct masker-random, target-random masker-correct, or both target and masker correct. In addition, a baseline condition consisted of both target and masker sequences in random (not syntactically correct) order for a total of four test conditions. The random order sequences in experiment 2, unlike experiment 1, consisted of one word selection from each column. However, the word order (choice of columns) was randomized rather than left to right so that the syntax was almost always incorrect. The word choices for target and masker were always mutually exclusive so that every condition consisted of two words per column. The stimuli were presented at the normal speaking rate. The voice and location were random selections from among the same set of values described in experiment 1 for each word in the sequence.

Experiment 1 was run in 30-trial blocks with the condition held constant across trials within a block. In a session, ten blocks (one per condition/rate combination) were completed in random order. Each session took approximately 90 min, and each listener completed three sessions for a total of 90 trials per condition/rate per listener. Experiment 2 was run in 25 trial blocks and a session was comprised of four blocks (one per condition) in random order. Each session took approximately 30 min and each listener completed four sessions for a total of 100 trials per condition per listener.

E. Control conditions

Before commencing the experimental conditions, listeners completed a set of control conditions. These trials were designed both to provide reference data regarding the intelligibility of the target sequences and to familiarize listeners with the basic task and the response grid.

For experiment 1, three of the four subjects completed an initial listening session that consisted of ten blocks of 30 trials. Because listener S1 had completed experiment 2 prior to experiment 1, she was considered sufficiently experienced and did not complete this initial session. The first two blocks (one at each rate) consisted of isolated target sequences ("quiet") with silences in the intervening masker positions. These blocks confirmed that the target words in isolation, even when presented at a faster rate, were highly intelligible (>90% correct). In the second two blocks (one at each rate), broadband noise bursts were presented in the even-numbered temporal positions in the sequence. In the third pair of blocks (one at each rate), masker words were presented as in the random condition of experiment 1, but they were reversed in time such that they were unintelligible. These conditions provided controls for comparison of performance in speech-masked conditions ("speech masking" will refer to the forward-played speech maskers in contrast to the "time-reversed speech" maskers used in one of the control conditions). The remaining four blocks (two at each rate) were identical to the random condition of experiment 1. These blocks allowed listeners to practice the difficult task of ignoring the intervening masker words while remembering and responding to the target sequence (data not reported).

In experiment 2 a shorter control session was conducted on each listener, comprising one block each of the quiet, noise-masked, and reversed speech-masked conditions. The words were spoken at a normal rate and the word order was randomized (not syntactically correct).

III. RESULTS

A. Experiment 1

The group mean results from experiment 1 are presented in Fig. 1. The upper panel shows the results for the normal rate while the lower panel shows the results for the fast rate. These values are group mean percent correct scores with error bars indicating intersubject standard errors. Note that the computation of percent correct for each five-word target sequence is based on the criterion that the words were reported in the correct word position.[2] So, for example, if the sequence of target words was "red, saw, Bob, old, six" and the listener reported "red, Bob, found, old, six…" then the score for these items would be 60% correct (three of five test words reported in the correct word positions). One of the word errors would be scored as a temporal position error ("Bob") and one scored as not present in the target sequence ("found"). The three bars in the left portion of each panel represent the control conditions of no masker ("quiet"), noise masker ("noise"), and reversed speech masker ("rev"). The performance ranged from about 88% to 94% correct when the target sequences were presented in quiet or in Gaussian noise for both speaking rates. Interspersing time-reversed

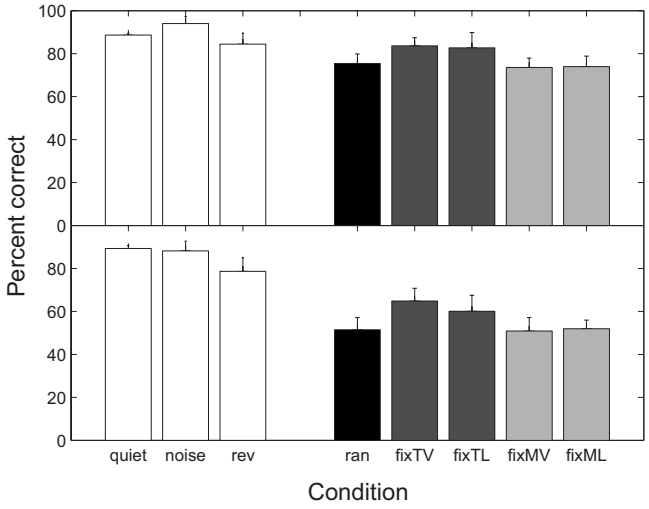

FIG. 1. Group mean percent correct scores from experiment 1. The upper panel displays the results for the normal rate while the lower panel displays the results for the fast rate. The abscissa indicates the experimental conditions. From the left: no masker (quiet), speech-shaped noise (noise), temporally reversed speech (rev), target and masker both random (ran), target voice fixed (fixTV), target location fixed (fixTL), masker voice fixed (fixMV), and masker location fixed (fixML). The error bars represent standard errors of the means.

words degraded performance by about five to ten percentage points for the normal rate and about nine percentage points for the fast rate relative to the quiet or noise-masked scores. For these three conditions, changing the speaking rate did not affect performance.

In the speech masking conditions of experiment 1 (the rightmost five bars in each panel), the linkage variables tested were voice and location. First, averaging across all listeners and word positions, there was a clear decrease in performance in all of these conditions for the fast rate as compared to the normal rate. On average, performance declined by about 19–24 percentage points when the rate was increased across these five masked conditions. In the random baseline condition ("ran"), performance was at about 75% correct for the normal rate. For the two cases in which the target had a fixed parameter value while the masker varied at random, significant improvements in performance relative to the fully random condition were observed. When the target voice was fixed ("fixTV"), group mean performance was about 85% correct, an increase of ten percentage points over random. When target location was fixed ("fixTL"), about the same advantage as fixed voice was observed, with group mean performance near 84% correct. When the same linkage variables were fixed for the masker ("fixMV" and "fixML") while the target varied randomly, no benefit to performance was observed relative to the random condition. A similar pattern of results was observed for the fast rate, although overall performance was significantly poorer. Here the random condition and the two fixed masker conditions were near 50% correct, while the fixed target conditions were somewhat better with average scores of 65% and 60% correct, respectively, for fixed voice and location.

Figure 2 (top row) displays the group mean performance in the various speech masking conditions as a function of target word position. The two leftmost panels show results

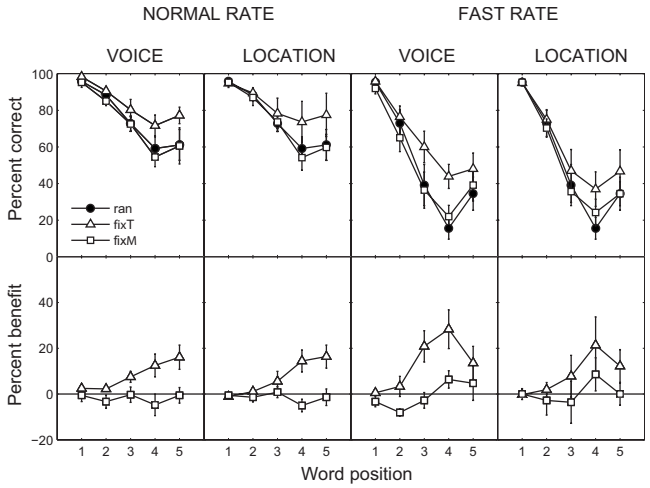

FIG. 2. Top row: Group mean percent correct scored according to target word position in the sequence. Each column shows the results for a different rate/linkage variable. The parameter in each panel is whether the target (open triangles), masker (open squares), or neither (filled circles) was fixed. The error bars represent standard errors of the means. Bottom row: Group mean benefits provided by fixing linkage variables as a function of word position. The benefits are calculated by subtracting the percent correct score in the random condition from the percent correct score in the test condition.

from the normal rate and the two rightmost panels show data for the fast rate. The left and right panels of each pair show results obtained when the linkage variables of voice and location, respectively, were manipulated against the random baseline condition which is replotted for the two panels within each pair. The different functions in each panel show the random ("ran") condition, the fixed target ("fixT") condition, and the fixed masker ("fixM") condition. Plotting the data in this manner reveals a strong word position effect. Performance for the first target word is higher than 90% correct for all conditions, then declines with increasing word position from the second through the fourth word with a tendency for an improvement for word five. There is a clear difference in performance between normal and fast rates which increased with word position; the largest difference occurred for words four and five and was greater than 30 percentage points in some cases.

The benefit provided by the two linkage variables tested is plotted as a function of word position in the bottom row of Fig. 2. Here, the benefit is computed as the difference between fixing a target or masker variable relative to performance in the random condition. The benefit of holding a target parameter constant increased with increasing word position at least through word four where improvements in performance approached 30 percentage points for the fast rate. Benefits were roughly the same for fixing either target voice or target location at a given rate. Fixing the masker did not result in a benefit in performance. A four-way repeated measures analysis of variance was conducted on the benefits with factors of rate (normal or fast), sequence fixed (target or masker), linkage variable (voice or location), and word position (1–5). The analysis indicated significant main effects of sequence $[F(1,3)=19.42,\ p<0.05]$, and a significant interaction between sequence and word position $[F(4,12)=9.86),\ p<0.005]$ which is clearly seen in the figure. The factors of rate, linkage variable, and word position were not significant $(p>0.05)$ and no other interactions were significant $(p>0.05)$.

Figure 3 shows the individual results from experiment 1. The rows are individual subjects while the columns are the four combinations of rate and linkage variable. The parameter of each graph is the fixed sequence (target, masker, or neither). Some of the general trends noted in the group mean data are also apparent for each individual subject. For example, both word position and rate effects are consistent across listeners (although there are large differences in the magnitudes of the effects). However, the benefit of fixing voice or location differed considerably across subjects. For example, S2 derived no benefit from fixing the target location for either rate and only obtained modest benefits from fixing the target voice. The largest benefits for any listener were observed for S4 at the fast rate where both fixed target voice and target location provided benefits of more than 50 percentage points relative to the random case for target word position 4.

An analysis of the different kinds of errors that were made in the experimental conditions is displayed in Fig. 4. The most common error (about one-half of all errors) was reporting a masker word as a target word. The next most frequent error (about 40% of all errors) was reporting a target word in an incorrect word position. Because the task required choosing five words from the grid, the remaining errors consisted of words that were not presented in either target or masker sequences. It appears that a reduction in all error types contributed to the improved performance seen in the fixed target conditions.

B. Experiment 2

In experiment 2, the main variable was the syntactic structure of the sequences. That is, a sequence of words could form a syntactically correct sentence ("correct") or it could simply be a random draw of words from the corpus (random). Group mean performance and standard deviations for random target sentences in the three control conditions (quiet, noise, and time-reversed masker) are shown in the leftmost bars of Fig. 5. On average, performance was about 95% correct for both quiet and noise conditions and declined to about 87% correct when the maskers were time-reversed words.

In the speech masking conditions (right side of Fig. 5) the dominant factor was whether the target sequence formed a syntactically correct sentence or whether the five words forming the target sequence were chosen at random. When syntactically correct, performance was about 90% correct regardless of whether the masker was random ("Tcorr") or syntactically correct ("TMcorr"). When random, performance fell to around 60%–65% correct regardless of the masker condition ("ran" and "Mcorr").

Figure 6 (top panel) shows the group mean results of experiment 2 plotted as a function of word position. Because identification performance was so accurate overall when the

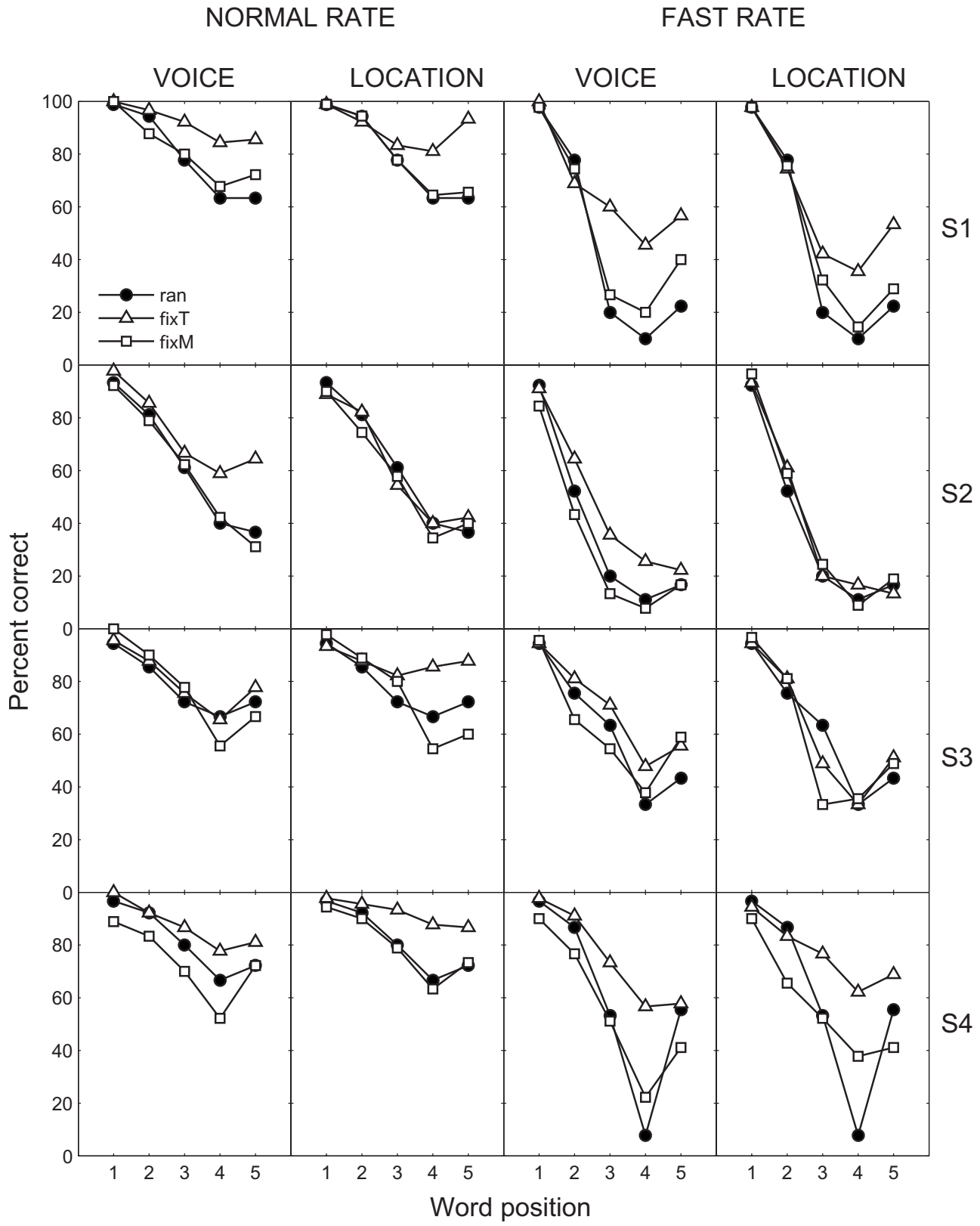

FIG. 3. Percent correct performance as a function of word position for the four individual subjects in experiment 1. Each row shows the results from a single subject, and each column shows the results for a different rate/linkage variable. The parameter in each panel is whether the target (open triangles), masker (open squares), or neither (filled circles) was fixed.

target was syntactically correct, the effect of word position appears relatively small. The general pattern of results noted in experiment 1 with respect to word position was also apparent here when the target was randomized, with word position 4 yielding the poorest overall performance.

The amount of benefit provided by correct syntax is illustrated in the bottom panel of Fig. 6. Again, the reference for computing benefit was performance in the random baseline condition. When the masker alone was correct, no consistent benefit was obtained. However, large and nearly equal

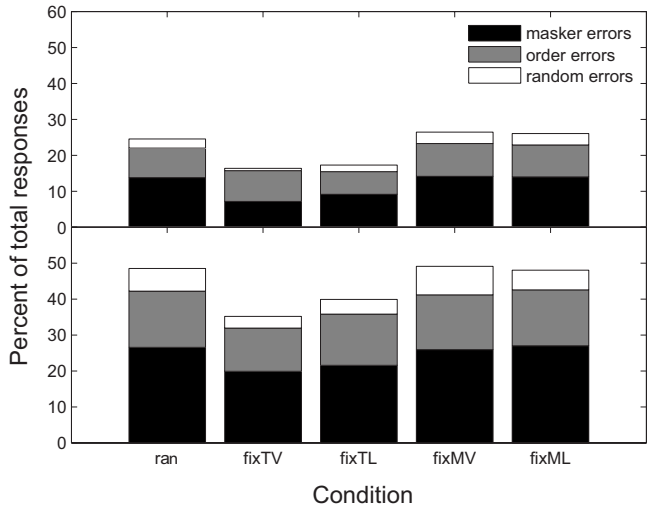

FIG. 4. Group mean error patterns for experiment 1. The upper panel displays the results for the normal rate while the lower panel displays the results for the fast rate. The abscissa indicates the experimental conditions. Errors are classified as one of three types: masker errors (black), target order errors (gray), or random errors (white) where subjects reported words that were not presented.

benefits that are word position dependent are apparent in both conditions when the target was syntactically correct (regardless of masker sequence syntax). For word positions 3–5, benefits of 30–40 percentage points are seen. A two-way repeated measures analysis of variance conducted on the benefits, with factors of sequence correct (target, masker, or both) and word position (1–5), revealed significant main effects of sequence [$F(2,6)=23.13$, $p<0.005$] and word position [$F(4,12)=12.02$), $p<0.001$], and a significant two-way interaction [$F(8,24)=11.10$, $p<0.001$].

Figure 7 shows the results from individual subjects in experiment 2. In general, the trends seen in the group data are apparent for each subject, with the syntactic structure of the target determining performance. However, the different listeners differ considerably in their overall performance, particularly in the random condition, and those that perform more poorly receive a greater benefit from correct target syntax.

Figure 8 shows an analysis of the different kinds of errors that were made in experiment 2. When the target was not syntactically correct, there was an approximately equal

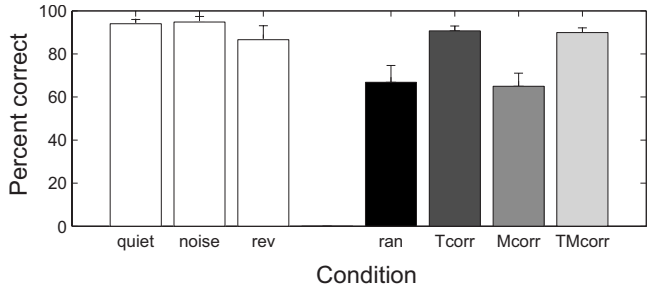

FIG. 5. Group mean percent correct scores from experiment 2. The abscissa indicates the experimental conditions. From the left: no masker (quiet), speech-shaped noise (noise) and temporally reversed speech (rev), target and masker both with random word order (ran), correct target syntax (Tcorr), correct masker syntax (Mcorr), and correct target and masker syntax (TMcorr). The error bars represent standard errors of the means.

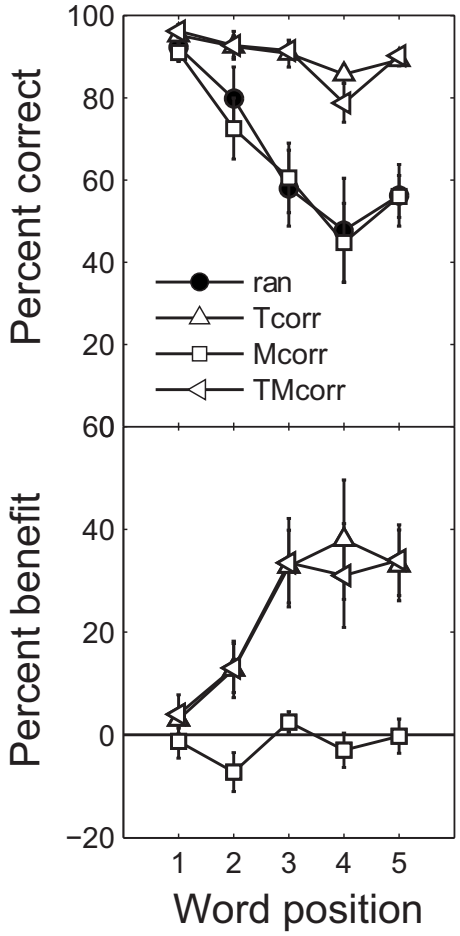

FIG. 6. Top panel: Group mean percent correct scored according to target word position in the sequence. The parameter in each panel is whether the target (open upward-pointing triangles), masker (open squares), target and masker (open left-pointing triangles), or neither (filled circles) had plausible syntax. The error bars represent standard errors of the means. Bottom panel: Group mean benefits provided by giving the target and/or masker-orrect syntax as a function of word position. The benefits are calculated by subtracting the percent correct score in the random condition from the percent correct score in the test condition.

number of each of the three error types made. When the target was given plausible syntax, target order errors were effectively eliminated (because the subject knew to respond in correct syntactic order left to right), but there was also a reduction in the number of masker errors and random errors.

IV. DISCUSSION

In his original article describing the speech intelligibility assessment procedure in which target and masker words alternated, Broadbent (1952) found that irrelevant words interspersed among test words greatly increased the difficulty in reporting the test words. Furthermore, he reported that factors such as familiarity with the talker's voice could improve recognition performance. This occurred even though the test words were perfectly audible and ideally could be selected through an appropriate use of attentional focus over time. He thus concluded that the interference caused by the masker words was evidence for a failure of selective attention. The current results support and significantly extend those early findings. These effects are clearly not attributable to

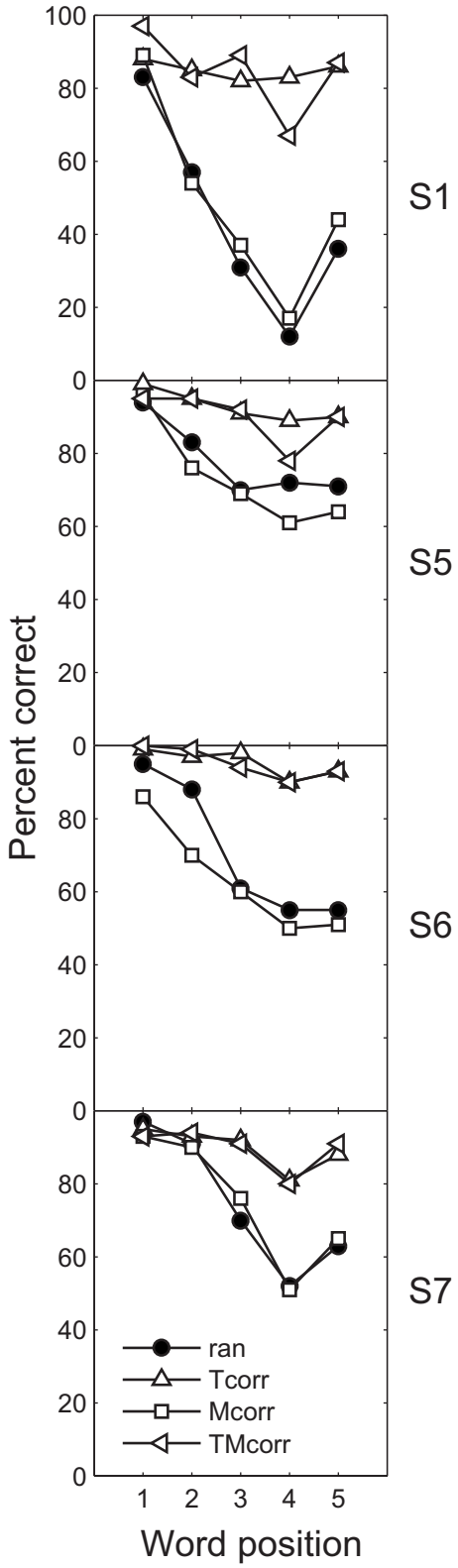

FIG. 7. Percent correct performance as a function of word position for the four individual subjects in experiment 2. Each panel shows the results from a single subject, and the parameter in each panel is whether the target (open upward-pointing triangles), masker (open squares), target and masker (open left-pointing triangles), or neither (filled circles) had plausible syntax.

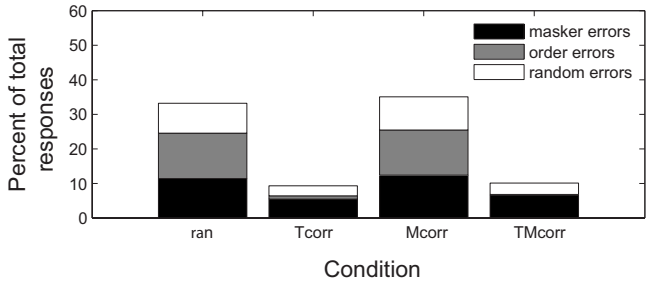

FIG. 8. Group mean error patterns for experiment 2. The abscissa indicates the experimental conditions. Errors are classified as one of three types: masker errors (black), target order errors (gray), or random errors (white) where subjects reported words that were not presented.

nonsimultaneous energetic masking because performance in the noise-masker control condition was equivalent to performance when no masker was present. Moreover, in both of the noise-masker cases, for speech presented at a normal rate and at twice the normal rate, recall of random sequences was generally better than 90% correct.

One conclusion based on the current results is that this procedure is useful for examining informational masking under conditions in which the results are not confounded by a peripheral masking component. This is a stronger conclusion than may be made about speech processed into very narrow frequency bands presented simultaneously (cf. Arbogast *et al.*, 2002; Kidd *et al.*, 2005; and Gallun *et al.*, 2007). Conversely, it also appears to be a useful means for exploring how various factors bind sequences of words together to provide a release from informational masking. As compared to the noise masker, a small amount of interference was obtained by using a temporally reversed speech masker. This finding suggests that it was not necessary for the masker elements to be meaningful but instead that some degree of informational masking may occur if the target and masker elements are qualitatively similar even when no spectrotemporal overlap is present.

All of the linkage variables tested provided significant improvements in performance when applied to the target. A constant apparent location provided by a fixed interaural time difference (ITD) that was different from that of the masker and a constant voice uttering all target words in a sequence—again, different from the masker—provided significant benefit in solving the speech identification task. Likewise, imposing a sensible syntactic structure on the target sequence provided a significant performance benefit. Increasing the rate significantly degraded performance in the speech masking conditions but did not affect performance in quiet, in noise, or in reversed speech. In those three control conditions, there apparently was not a sufficient processing load placed on the listener for rate to matter. This finding is similar to that reported by Brungart and Iyer (2007) who also noted an increase in the potency of speech maskers at faster-than-normal speech rates. In the present study, the effect of rate on the benefit provided by each linkage variable was not significant. This was somewhat surprising given that faster rates are known to promote stronger perceptual streaming for simple stimuli (e.g., van Noorden, 1975), but may perhaps be explained by the absence of silent gaps between words at either rate in the current conditions (e.g., Bregman *et al.*, 2000).

Characterizing these performance benefits as a consequence of strengthening the linkage between target words

seems warranted for several reasons. Simultaneous segregation (or the formation of auditory objects) is not a factor here given that each individual word of both the target and the masker is already perfectly segregated. Thus, unlike the case of multiple simultaneous talkers, the issue of whether each is perceptually segregated, or the strength of the segregation, does not apply. It seems likely that the intervening masker words interrupt the process of linking, storing, rehearsing, and retrieving the stream of target words in memory. The beneficial effect of the linkage variables applied here is manifested in an improved ability to recall the target items. It also makes sense that the benefit of a given linkage variable would increase as the number of items that share a common property increases, and an increase in the benefit of the linkage variables across time/word position that is broadly consistent with this notion was found. It remains to be seen, however, whether or not the early items would benefit more when they are more difficult in the baseline condition.

The idea that strengthening the cues that perceptually bind words together can influence word recall has been raised previously (e.g., Martin *et al.*, 1989; Nicholls and Jones, 2002a; and Nicholls and Jones 2002b). Nicholls and Jones (2002a) examined how "sandwiching" irrelevant words around a sequence of digits in a serial recall task reduced the detrimental "suffix" effect. The suffix effect, which is a classical finding in memory research, refers to the interference in the recall of a series of items when an irrelevant item (not to be remembered) is inserted following the list of target words. The largest effect of the suffix is on the terminal word in the test sequence, which is usually easier to recall than the preceding test word due to recency in memory. Nicholls and Jones (2002a) demonstrated that the suffix effect could be reduced when the sandwiched items (irrelevant words inserted between target words, very much like the current procedure) had acoustic properties known to promote perceptual streaming. The most effective "capture" (e.g., Bregman and Rudnicky, 1975) of the suffix (pulling it into the masker stream) occurred when the masker words and suffix had the same fundamental frequency. In contrast, no capture was observed when the fundamental frequencies of the masker words varied randomly. In a separate, but related article Nicholls and Jones (2002b) examined the sandwich experiment (cf. Hitch, 1975 and Baddeley *et al.*, 1993) more fully. They reported that the irrelevant words had the greatest effect on target recall when they had the same pitch as the target words. Furthermore, they found that more interference in target word recall was obtained when the masker words were randomly selected than when a single masker word was simply repeated throughout the sequence. The findings of Nicholls and Jones (2002a, 2002b) seem inconsistent with our finding that strengthening the linkage variables of the masker words provided no benefit in reporting the target words. The only benefit, as discussed more fully below, was obtained when the linkage variables were applied to the target words. There are a number of procedural differences that may account for the different findings. Most importantly, in Nicholls and Jones' work the target was always spoken by the same voice. It is possible that masker continuity can play a role only if there is target continuity. This condition was not tested with the voice and location linkage variables in experiment 1, and although it was tested for the parameter of syntactic structure in experiment 2, ceiling effects would not allow us to see any additional benefit of masker linkage. In addition, syntactic structure and acoustic linkage variables likely involve very different processing mechanisms and thus may be expected to differ in their effects.

Viewed in another way, the fact that uncertainty in the target was more disruptive in the present experiments than uncertainty in the masker is relevant to discussions of the "listener max vs listener min" observer models (e.g., Durlach *et al.*, 2003). As defined by Durlach *et al.* (2003), this contrast in listener strategies reflects whether the available processing resources are devoted to emphasizing the representation of the target (e.g., applying gain at a particular point along the relevant stimulus dimension, called listener max) or to de-emphasizing or ignoring the masker (attenuating a point or points along the relevant stimulus dimension, called listener min). However, on a practical level it can be very difficult to separate the actions of these two hypothetical observer strategies in actual experiments (however, cf. de Cheveigné *et al.*, 1995). This is partly because, in masked conditions, both produce similar effects and both may result in selective/tuned responses along a given perceptual dimension. The current results appear to be more consistent with a listener max strategy because the only consistent benefit of the various linkage variables (i.e., constant physical value or syntactic structure) occurred when they were applied to the target. If the listener was attempting to "null out" the masker words then we might have expected some benefit from application of the linkage variables relative to random conditions. In contrast, Brungart *et al.* (2007) recently reported evidence from a multitalker speech identification task that appeared to be more consistent with a listener min model. A primary variable in their study was a difference in spatial location, a manipulation that is very similar to the ITD manipulation used here. They employed a procedure in which the predictability of the locations of the target and the maskers was varied systematically. As in experiment 1 here, either target or masker location(s) was/were fixed while the other changed location unpredictably, although on a trial-to-trial basis rather than within a trial. They found that listeners were more sensitive to variability in the masker location than in the target location. They interpreted this result as consistent with a listener min (or "masker min" in their terms) approach in which knowledge about masker location would lead to a nulling of the masker which was more useful than knowledge about the target. In their study, the difference in performance between the two contrasting conditions was small but was statistically significant. Moreover, they provided converging evidence in support of their position in a set of conditions under which the relative positions of the target and two masker locations were varied. It seems plausible that both listener max and listener min strategies are available to listeners to some degree and that they may adopt one or the other—or a combination of both—as appropriate to the demands of the task and the specific listening situation.

There has been relatively little work on this issue in the auditory domain and the question of whether and to what degree listeners adopt these hypothetical strategies remains open.

V. CONCLUSIONS

This study demonstrated that large amounts of informational masking may be obtained by temporally interleaving a sequence of target and masker words when the task is to recall the target sequence. Presenting the words in a temporally nonoverlapping manner greatly reduces the chance of any peripherally based energetic masking contributing to the results. It was found that all three linkage variables examined were effective in reducing informational masking and improving accuracy of report when applied to the target words. Generally, the beneficial effects of the linkage variables were greatest for the later-occurring items where performance was relatively poor. None of the linkage variables were effective when applied to the masker words. These findings are compatible with a listener max model (Durlach *et al.*, 2003) in which the available processing resources are devoted to emphasizing the representation of the target.

ACKNOWLEDGMENTS

This work was supported by Grant Nos. DCO4545 and DC0100 from the National Institute on Deafness and Other Communication Disorders (NIH/NIDCD) and by Grant No. FA9950-05-1-2005 from the Air Force Office of Scientific Research (AFOSR), United States Air Force. Virginia Best was also supported in part by a University of Sydney Postdoctoral Research Fellowship.

[1]Despite the large range of individual word durations, we found no evidence that the identification of the target words was related to duration in either the normal-rate or fast-rate conditions.

[2]It is often the case that serial recall is scored without regard to errors of word position. We analyzed our results in that manner and found that the pattern of word position effects was not greatly affected by scoring method.

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.

Baddeley, A. D., Papagano, C., and Andrade, J. (**1993**). "The sandwich effect: The role of attentional factors in serial recall," J. Exp. Psychol. Learn. Mem. Cogn. **19**, 862–870.

Boersma, P., and Weenink, D. (**2007**). Praat: Doing phonetics by computer (Version 4.6.40) (Computer program). Last viewed December, 2007), from http://www.praat.org/

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* (MIT Press, Cambridge, MA).

Bregman, A. S., Ahad, P. A., Crum, P. A., and O'Reilly, J. (**2000**). "Effects of time intervals and tone durations on auditory stream segregation," Percept. Psychophys. **62**, 626–636.

Bregman, A. S., and Rudnicky, A. J. (**1975**). "Auditory segregation: Stream or streams?," J. Exp. Psychol. Hum. Percept. Perform. **1**, 263–267.

Broadbent, D. E. (**1952**). "Failures of attention in selective listening," J. Exp. Psychol. **44**, 428–433.

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., and Iyer, N. (**2007**). "Time-compressed speech perception with speech and noise maskers," Proceedings of Interspeech 2007, Antwerp, Belgium.

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.

Darwin, C. J., and Carlyon, R. P. (**1995**). "Auditory grouping," in *Handbook of Perception and Cognition*, edited by B. C. J. Moore (Academic, New York), Vol. **6**, pp. 387–424.

de Cheveigné, A., McAdams, S., Laroche, J., and Rosenberg, M. (**1995**). "Identification of concurrent harmonic and inharmonic vowels: A test of the theory of harmonic cancellation and enhancement," J. Acoust. Soc. Am. **97**, 3736–3748.

Durlach, N. I., Mason, C. R., Shinn-Cunningham, B. G., Arbogast, T. L., Colburn, H. S., and Kidd, G., Jr. (**2003**) "Informational masking: Counteracting the effects of stimulus uncertainty by decreasing target-masker similarity," J. Acoust. Soc. Am. **114**, 368–379.

Gallun, F. J., Mason, C. R., and Kidd, G., Jr. (**2007**). "The ability to listen with independent ears," J. Acoust. Soc. Am. **122**, 2814–2825.

Griffiths, T. D., and Warren, J. D. (**2004**). "What is an auditory object?," Nat. Rev. Neurosci. **5**, 887–892.

Hitch, G. J. (**1975**) "The role of attention in visual and auditory suffix effects," Mem. Cognit. **3**, 501–505.

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," Acust. Acta Acust. **114**, 526–536.

Kidd, G., Jr., Mason, C. R., Richards, V. M., Gallun, F. J., and Durlach, N. I. (**2008**). "Informational masking," in *Auditory Perception of Sound Sources*, edited by W. A. Yost, A. N. Popper, and R. R. Fay (Springer Handbook of Auditory Research, New York), pp. 143–190.

Martin, C. S., Mullennix, J. W., Pisoni, D. B., and Summers, W. V. (**1989**). "Effects of talker variability on recall of spoken word lists," J. Exp. Psychol. Learn. Mem. Cogn. **15**, 676–684.

Moore, B. C. J., and Gockel, H. (**2002**). "Factors influencing sequential stream segregation," Acust. Acta Acust. **88**, 320–333.

Nicholls, A. P., and Jones, D. M. (**2002a**). "Capturing the suffix: Cognitive streaming in immediate serial recall," J. Exp. Psychol. Learn. Mem. Cogn. **28**, 12–28.

Nicholls, A. P., and Jones, D. M. (**2002b**). "The sandwich effect reassessed: Effects of streaming, distraction and modality," Mem. Cognit. **30**, 81–88.

Shinn-Cunningham, B. G. (**2008**). "Object-based auditory and visual attention," Trends Cogn. Sci. **12**, 182–186.

van Noorden, L. (**1975**). "Temporal coherence in the perception of tone sequences," Ph.D. thesis, Technical University Eindhoven, Eindhoven, The Netherlands.

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.

Yost, W. A. (**1991**). "Auditory image perception and analysis: The basis for hearing," Hear. Res. **56**, 8–18.

Yost, W. A. (**2008**). "Perceiving sound sources," in *Auditory Perception of Sound Sources*, edited by W. A. Yost, A. N. Popper, and R. R. Fay, (Springer Handbook of Auditory Research, New York), pp. 11–13.