The role of syntax in maintaining the integrity of streams of speech

Gerald Kidd, Jr.[a] and Christine R. Mason
Department of Speech, Language and Hearing Sciences and Hearing Research Center, Boston University, 635 Commonwealth Avenue, Boston, Massachusetts 02215

Virginia Best
National Acoustic Laboratories, Macquarie University, New South Wales 2109, Australia

(Received 19 September 2013; revised 13 December 2013; accepted 23 December 2013)

This study examined the ability of listeners to utilize syntactic structure to extract a target stream of speech from among competing sounds. Target talkers were identified by voice or location, which was held constant throughout a test utterance, and paired with correct or incorrect (random word order) target sentence syntax. Both voice and location provided reliable cues for identifying target speech even when other features varied unpredictably. The target sentences were masked either by predominantly energetic maskers (noise bursts) or by predominantly informational maskers (similar speech in random word order). When the maskers were noise bursts, target sentence syntax had relatively minor effects on identification performance. However, when the maskers were other talkers, correct target sentence syntax resulted in significantly better speech identification performance than incorrect syntax. Furthermore, conformance to correct syntax alone was sufficient to accurately identify the target speech. The results were interpreted as supporting the idea that the predictability of the elements comprising streams of speech, as manifested by syntactic structure, is an important factor in binding words together into coherent streams. Furthermore, these findings suggest that predictability is particularly important for maintaining the coherence of an auditory stream over time under conditions high in informational masking. © *2014 Acoustical Society of America.*

[http://dx.doi.org/10.1121/1.4861354]

PACS number(s): 43.66.Dc, 43.66.Mk [EB] Pages: 766–777

I. INTRODUCTION

In a sound field containing multiple human voices, a listener normally is presented with a variety of cues that can be used to segregate and follow the speech of one specific talker over time. These cues or features often are roughly categorized as either "low-level" or "high-level" depending on the presumed physiological site of origin and the degree of complexity of processing required to exploit the cue. The designation as low-level usually refers to the view that the cue occurs automatically and early on as the neural representations of the sounds propagate from the auditory periphery to the cortex (often characterized as a "bottom-up" process). Furthermore, low-level cues often may be related directly to simple acoustical properties of sounds even to the extent that the important variable is distributed along a single dimension. In contrast, higher-level cues/features are viewed as being extracted from or assigned to the stimulus by processes that originate beyond peripheral levels of the auditory system, such as those causing the "top-down" direction of attention, and are guided or strongly influenced by *a priori* knowledge stored in memory.

Among the important low-level cues used to separate and select speech sources are various talker-specific acoustic properties, such as fundamental frequency (F_0) and the frequencies and bandwidths of the resonances of the vocal tract (e.g., Darwin *et al.*, 2003; Binns and Culling, 2007). Also, the various talkers usually are spatially distributed so that differences in location, as signaled primarily by interaural differences in the waveforms, help the listener to perceptually segregate and attend to specific voices (e.g., reviews in Yost, 1997; Bronkhorst, 2000). These simple acoustical cues are also used to bind together the sequences of sounds into perceptually coherent streams (e.g., reviews in Bregman, 1990; Darwin and Carlyon, 1995; Shamma *et al.*, 2010; Moore and Gockel, 2012).

However, it should be observed that even these putatively low-level features tend not to be stationary over time. The fundamental frequency of one talker varies according to natural patterns of intonation and the distributions of F_0's from different talkers—while perhaps differing in mean values and ranges—often overlap considerably. Likewise, with respect to the perceived locations of actual sound sources, the sources may move and change position. Static differences in F_0 and apparent location, as well as other low-level acoustic features, may not be as important as the values that are *plausible* for those sources at a given moment in time. This suggests that our perceptions of even low-level features are interpreted in the context of past events and our expectation about impending events—interpretations that are relatively high-level in nature (e.g., Denham and Winkler, 2006; Winkler *et al.*, 2012; Rajendran *et al.*, 2013; Kidd *et al.*, 2013). Thus, the dynamic nature of sound sources in realistic multitalker environments suggests that predictability and expectation are key to understanding the formation, segregation,

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

© 2014 Acoustical Society of America

and maintenance of streams of speech (however, see Best *et al.*, 2010, for a counter example).

Speech sounds also convey linguistic information that the listener must process or perhaps sometimes actively ignore. The use of linguistic information—like the exploitation of predictability or sequential dependencies in maintaining auditory streams in general—is a high-level process that also could play an important role in selective listening to speech. Among the fundamental linguistic aspects of normal speech is syntax, which governs the order of words arranged into sentences. Syntactic structure is of particular interest in speech-on-speech masking because it bears directly on the issue of predictability. When the words comprising a sentence follow an established and known syntax, particularly when coupled with semantic content, they become predictable, to a certain extent, so that the listener expects the type of word to be uttered and may even have a sense of the probability of (anticipate) specific words over time on a word-by-word basis. This predictability due to syntax has long been appreciated and forms the basis for certain approaches to automatic speech recognition (e.g., Rabiner, 1989) because it can limit the number of possible words that must be evaluated by an algorithm attempting to identify a given word token. For human listeners in multitalker sound fields, it seems reasonable to assume that exploitation of the predictability of word categories when conforming to a known syntax provides a useful means for maintaining the focus of attention on a specific talker separate from competing talkers. Also, because speech comprehension unfolds as the utterance proceeds, a benefit of syntax—analogous to the automatic speech recognition problem above—may result from a reduction in the number of candidate items in the lexicon that must be evaluated given a specific word token. This possible benefit may be viewed as a reduced processing load or simply as decreased uncertainty for the observer. More broadly, though, the importance of syntactic structure as a means for preserving the integrity of a stream of speech (e.g., preventing it from being confused with competing sounds over time) in multitalker listening situations has not received much attention and currently is not well understood.

The present study examines the role of syntax, and its interaction with the lower-level cues of voice or location, in following the speech of one talker (the "target") among multiple competing talkers (the "maskers"). The underlying hypothesis is that syntactic structure provides an effective means for maintaining the focus of attention on one specific stream of speech in competition with other streams of speech. This putative process is viewed as a manifestation of the more general ability of listeners to exploit the predictability of the elements comprising sequences of sounds generated by a common source (e.g., Kidd *et al.*, 2013). In order to test the specific hypothesis about the benefit of syntax, a series of experiments was conducted using a closed-set speech identification task that allowed talker voice, location and correct vs incorrect syntactic structure to be varied in a controlled manner. A comparison of the benefit of syntax in maintaining the target stream of speech was made for maskers that were predominantly energetic in nature (noise) vs predominantly informational in nature (other talkers).

II. METHODS

The methods common to both experiments 1 and 2 are described here with additional detail about the procedures specific to each individual experiment under those sections.

A. Subjects

A total of seventeen young adults served as subjects in this study. There were 12 subjects total participating in experiment 1 and 9 subjects in experiment 2, with 4 subjects participating in both experiments. The subjects had normal hearing (better than 20 dB HTL at octave frequencies from 250–8000 Hz) as determined by audiometric evaluation. They were paid for their participation.

B. Stimuli

All stimuli were computer-generated and played at a 20-kHz rate through 16-bit DACs (Tucker-Davis Technologies, "TDT"). The stimuli were low-pass filtered at 10 kHz, attenuated, and mixed using TDT modules and presented via a TDT headphone amplifier to the listener wearing Sennheiser HD 280 Pro headphones in a sound-treated IAC booth. The levels were calibrated using a B&K model 2250 sound level meter and artificial ear.

Except for one single-word condition in experiment 1B (described below), the target was always a string of five words drawn from a corpus of forty monosyllabic words ("BU Corpus," Kidd *et al.*, 2008b). The corpus was subdivided into the five categories with eight items per category of: <*name*> <*verb*> <*number*> <*adjective*> and <*object*>. An example of a syntactically correct target sentence would be "Sue found six red hats." Each word in the corpus was spoken and recorded individually with neutral inflection so that there were no coarticulation effects between words and well-defined word boundaries. Thus, strings of words could be concatenated in any order without affecting the individual tokens. This was an important aspect of the stimulus that allowed manipulating syntactic order in a controlled manner. Likewise, the closed-set corpus and test permitted large numbers of trials without introducing possible learning effects or exhausting the available set of sentence materials. The word set was spoken by 16 voice-trained young-adults (eight male, eight female) under controlled acoustic conditions and recorded by Sensimetrics Corporation (Malden, MA). However, in this study only three female talkers were used for the target and masker voices. The target sentences were presented either in syntactically correct ("*syntactic*," word categories in the order given above) or pseudo-randomly mixed order ("*random*," not syntactically correct but also one word from each category). Nothing prevented the words in *random* strings from comprising partially correct syntax (for example, constructions like "red shoes Bob bought three" were not excluded). The maskers were either speech or speech-shaped noise bursts. On every trial there were two concurrent maskers comprising strings of five words or strings of five independent noise bursts. The speech maskers were strings of words drawn from the same corpus and talkers as the target. The individual words of the two masker strings were presented in pseudo-random order. The three words chosen for the three talkers in each respective word position were always drawn from mutually exclusive word categories (e.g., in one trial for the third word position the words from the three talkers might be target—*six*; masker 1—*red*; masker 2—*Bill*). The noise maskers were generated by shaping Gaussian noise bursts according to the average shape of the power spectrum of the female talkers in the corpus. The durations of the noise bursts corresponded to the durations of random selections of the words from the corpus. In all sequences the words or noise bursts were concatenated with no additional intervening silence. Hence, the three sequences were different total lengths and were only aligned in time at the start of the sequences.

The three sequences were always presented in three different apparent locations in the head: Left, center, and right, achieved by imposing interaural time differences, ITDs, on the stimuli of -800, 0, and 800 μs, respectively.

C. Procedures

The key to the target sequence's voice or location was provided by the word "Ready" presented immediately prior to the three sequences. This cue word directed the attention of the listener to the feature that denoted the target speech on that trial. So, for example, when location was the cue, the word "Ready" was presented simultaneously from all three talkers but with all three voices presented at the ITD designating the target. The other feature (voice) thus was uninformative. That is, when listening to the target location during the sequence, the voice of the target talker varied from word to word. And, conversely, if the target voice was cued, the word "Ready" was presented from all three locations in one voice (the target) with the target location then varying from word to word. Experiment 1 tested only these two cue conditions whereas experiment 2 also tested a "syntax only" condition in which no cue was presented. In that case, both voice and location were uninformative and *only* correct syntax defined the target. The trials were blocked according to condition with the order of blocks randomized. Within each block of trials, the target was fixed at 50 dB sound pressure level (SPL) and a range of masker levels was presented in order to obtain psychometric functions. Each point on the group mean psychometric functions was based on 100 (experiment 1 A), or 105 (experiment 2) scored words *per subject* following 20 practice words. In experiment 1B, the group means were based on 80 scored words per subject per condition (following 20 practice words).

The task of the listener was to identify the sequence of five target words on every trial. The selections were made via a graphical user interface displaying the word choices in columns, one column at a time, in the order in which they were presented. Thus, neither masker errors nor order errors were possible. Chance performance was thus 1 of 8 on a word-by-word basis for all conditions. Response feedback was given after each trial and the percent correct score based on the number of target words identified correctly was given at the end of each block of trials.

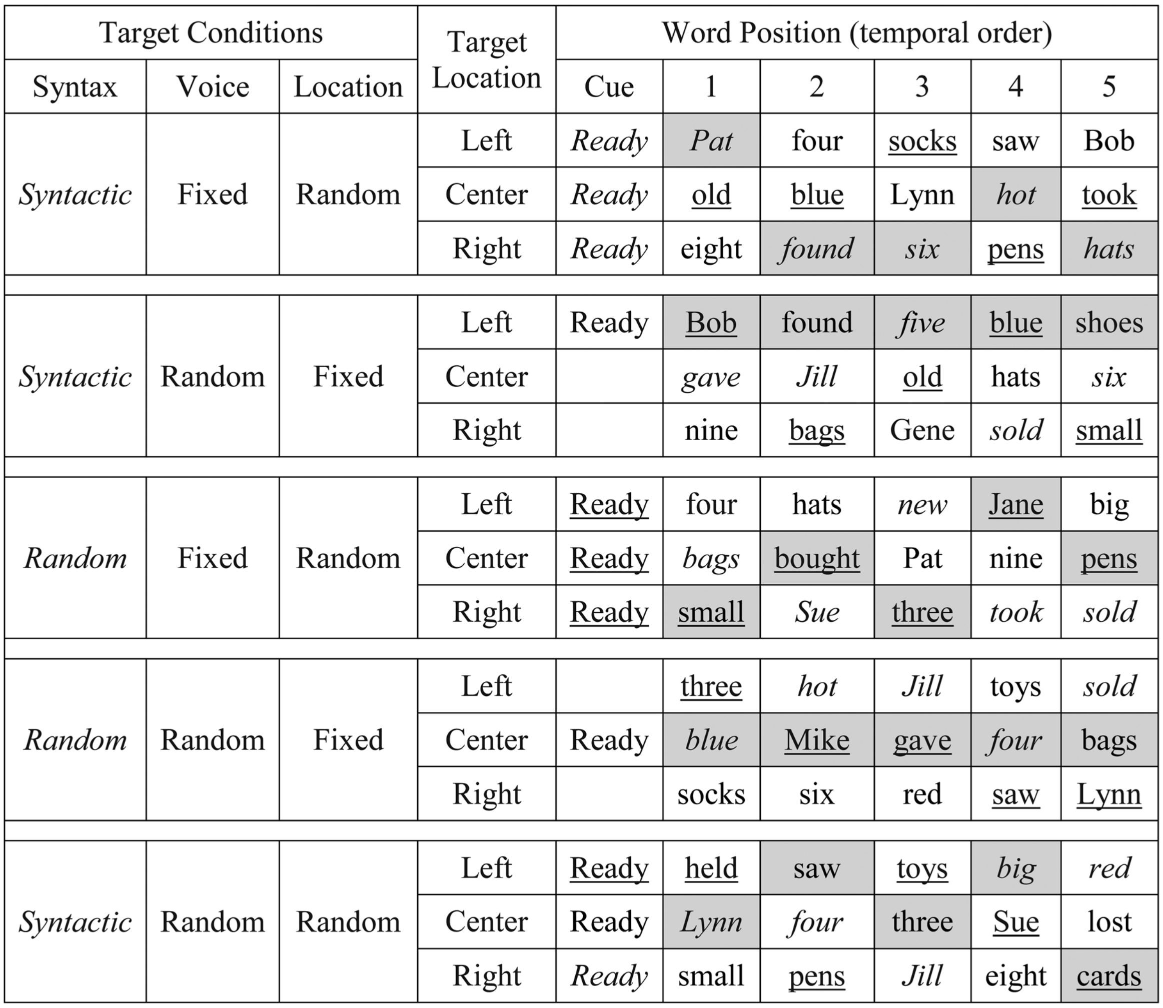
The different target/speech-masker presentation conditions are shown in Fig. 1. The target and masker words are examples of each condition with the row position indicating the apparent interaural location and the font (normal, italic, and underline) indicating the voice identity. The target words are in the shaded boxes. The left three columns designate the five experimental conditions by indicating the syntax (correct "*Syntactic*" or incorrect "*Random*"), voice ("Fixed" or "Random") and location ("Fixed" or "Random") of the target sequence. In experiment 1A, the top four conditions in the figure were also tested using noise maskers (not shown).

III. EXPERIMENT 1: DESIGNATING THE TARGET BY VOICE OR LOCATION

A. Correct vs incorrect target sentence syntax

In the first experiment, the target was identified by a constant interaural location or by a constant voice (but not by both simultaneously) across words within a sequence. When the maskers were noise bursts this designation was unnecessary, but when the maskers were speech it was essential. Furthermore, for these two means of designating the target, the word order of the target string could be presented using either correct (*syntactic*) or incorrect (*random*) syntactic order. All of the selections of values were made independently with replacement on every trial under the constraints of the set of conditions tested; e.g., if location were the target designator, the ITD for the target words was a random selection among the three ITDs for a given trial, but it was held constant for the target words within the trial. For the next trial, the same process was repeated with an independent selection of ITD. The two 5-word masker strings consisted of the non-target voices and locations and were never in the same order (word category type) as the target string.

The group-mean results are shown in Fig. 2 plotted as percent correct performance as a function of target-to-masker ratio (T/M) in dB. The error bars indicate ±1 standard error of the intersubject means. The lines shown in Fig. 2 are logistic functions fit to the mean data having parameters of slope, midpoint, and a lapse term plus a constant for chance performance (0.125). The left panel shows the results for the noise maskers while the right panel shows the results for the speech maskers. For both types of maskers, the data were orderly and generally well-described by the fits. However, for the speech maskers small "plateau" regions were apparent in individual or group mean psychometric functions usually in the range near 0 dB T/M (e.g., the small change in performance from -10 to -5 dB T/M). This finding is consistent with others reported in the literature for closed-set speech-on-speech masking experiments (e.g., Brungart, 2001a; Brungart *et al.*, 2001; Wightman *et al.*, 2006). Table I contains the computed slopes and thresholds for the individual subjects and the thresholds and slopes of the fits to the group mean data plotted in Fig. 2. Note that performance was at ceiling for the noise maskers at T/Ms of 0 and +5 dB regardless of the source designation condition and, for the speech maskers, performance approached ceiling in some cases for the highest T/M tested of 15 dB.



Target Conditions			Target Location	Word Position (temporal order)					
Syntax	Voice	Location		Cue	1	2	3	4	5
Syntactic	Fixed	Random	Left	*Ready*	*Pat*	four	socks	saw	Bob
			Center	*Ready*	old	blue	Lynn	*hot*	took
			Right	*Ready*	eight	*found*	*six*	pens	*hats*
Syntactic	Random	Fixed	Left	Ready	Bob	found	*five*	blue	shoes
			Center		*gave*	*Jill*	old	hats	*six*
			Right		nine	bags	Gene	*sold*	small
Random	Fixed	Random	Left	Ready	four	hats	*new*	Jane	big
			Center	Ready	*bags*	bought	Pat	nine	pens
			Right	Ready	small	*Sue*	three	*took*	*sold*
Random	Random	Fixed	Left		three	*hot*	*Jill*	toys	*sold*
			Center	Ready	*blue*	Mike	gave	*four*	bags
			Right		socks	six	red	saw	Lynn
Syntactic	Random	Random	Left	Ready	held	saw	toys	*big*	*red*
			Center	Ready	*Lynn*	*four*	three	Sue	lost
			Right	*Ready*	small	pens	*Jill*	eight	cards

FIG. 1. A schematic illustration of the speech masker test conditions for both experiments. Each row depicts a different condition identified by the left three columns indicating the status of the three target variables: Syntax (*syntactic*/correct or *random*/incorrect for target words in a sentence), voice ("Fixed"—same voice, or "Random"—different voice, for target words in a sentence), and location ("Fixed"—same location, or "Random"—different locations, for target words in a sentence). The fourth column shows the location (in three rows for each condition that are left, center, or right corresponding to ITDs of −800, 0, and 800 μs). The fifth column indicates the cue directing the listener's attention to the target designator (see text). The remaining columns labeled 1–5 show word position within the sentence. The three talkers are indicated by the font type (normal, italic, underlined). The target words are designated by shading.

As expected based on past findings (e.g., Kidd *et al.*, 2002; Arbogast *et al.*, 2002), the slopes of the functions varied depending on the type of masking that dominated. The slopes of the psychometric functions for the noise maskers, which theoretically are governed primarily by energetic masking, were about a factor of 3 steeper than those obtained for speech maskers which caused primarily informational masking. Performance increased about 6%/dB for the noise masker over the linear segment of the functions and about 2%/dB for the speech maskers across those ranges. Note that the intersubject variability around the mean values for the speech maskers was also larger than for the noise maskers, consistent with the shallower slopes.

The group-mean thresholds (T/Ms at 0.5 proportion correct in decibels) extracted from the fitted functions for each condition are plotted in Fig. 3. All of the values were negative indicating that 50% of the target words were intelligible when they were substantially lower in level than the masker. For the noise masker, the differences due to type of target cue—voice or location—were quite small. Furthermore, the differences in thresholds between *syntactic* and *random* word order were also very small. Overall the differences across conditions were less than 2 dB with the group mean thresholds falling in the range from −11.6 to −13.4 dB. The standard errors of the means were about 0.5 dB indicating that the subjects performed very similarly.

In comparison to the results from the noise maskers, the thresholds obtained in the speech maskers varied considerably more across subjects and conditions. When location was the feature defining the target stream the thresholds were lower than when the words were linked by the talker's voice. Furthermore, performance depended more on target syntax for the speech maskers, with the *syntactic* condition producing lower thresholds than the *random* condition. At the extremes, when location was the primary cue for *syntactic* target sentences, the group mean threshold was about −15.7 dB T/M. When the cue was voice and the target words were *random*, the group mean threshold was about −5.2 dB. Importantly for the main question motivating this experiment, the benefit of correct target syntax as determined by

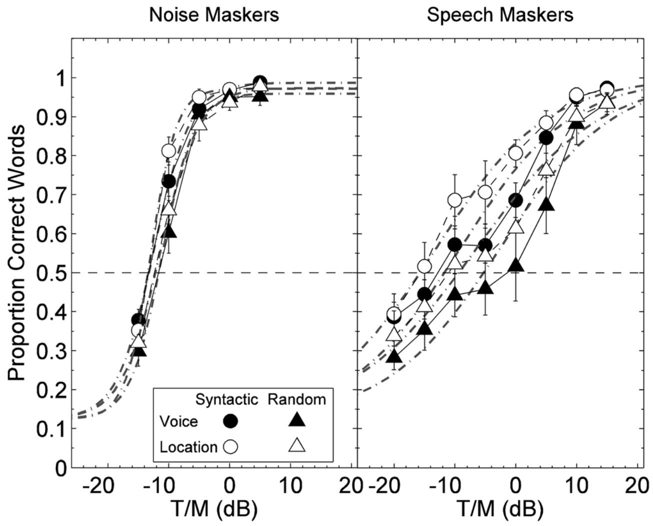


FIG. 2. Group mean results from experiment 1A shown in proportion correct as a function of target-to-masker ratio in dB. The left panel contains the results for the noise maskers while the right panel contains the results for the speech maskers. The data points are group mean proportion correct scores and standard errors of the means. The data points were fit with logistic functions (dot-dashed lines) from which "thresholds" were obtained at the intersection of the fits with 0.5 proportion correct (horizontal dashed line). The filled symbols are for conditions where the target was indicated by constant voice while the open symbols are for conditions where the target was indicated by constant location. Circles indicate that the target sentence was syntactically correct (*syntactic*) while triangles are for syntactically incorrect (*random*) target sentences.

the reduction in T/Ms at threshold was greater under the speech maskers than the noise maskers.

A repeated-measures analysis of variance (ANOVA) on the threshold data indicated that neither masker type (noise vs speech) [F(1,4) = 0.81, p = 0.41] nor cue type (voice vs location) [F(1,4) = 4.02, p = 0.12] were significant main effects. The effect of syntax was significant [F(1,4) = 20.7, p = 0.01] as was the interaction between syntax and masker type [F(1,4) = 10.52, p = 0.032]. This interaction is seen in the result that correct target syntax provided a significantly greater benefit under speech masker conditions than under noise masker conditions, consistent with the pattern of results in Fig. 3.

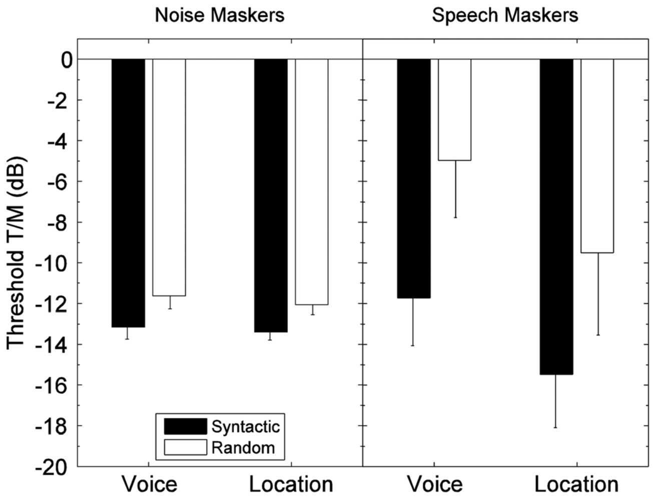


FIG. 3. Group mean "thresholds" (target-to-masker ratios in dB corresponding to the 0.5 proportion correct point on the fitted psychometric functions) and standard errors of the means from experiment 1A. The left panel shows the thresholds from the noise maskers while the right panel shows the thresholds from the speech maskers. In each panel, the thresholds when constant voice designated the target are plotted in the left two bars while the thresholds when constant location designated the target are plotted in the right two bars. The solid bars are when the target sentences were *syntactic* while the open bars are for *random* target sentences.

TABLE I. Thresholds (T/M's corresponding to 0.5 proportion correct performance) and slopes, both from the logistic function fits to the individual and group-mean proportion correct values (the psychometric functions) of experiment 1A. The columns indicate the experimental conditions.

	Noise maskers				Speech maskers			
	Voice		Location		Voice		Location	
Listeners	SYN	RAN	SYN	RAN	SYN	RAN	SYN	RAN
Thresholds								
1	−11.1	−9.6	−11.9	−10.6	−2.9	3.3	−7.9	3.3
2	−14.4	−12.3	−13.7	−11.9	−15.9	−8.4	−11.1	−7.6
3	−14.2	−13.0	−14.1	−12.4	−11.5	−0.0	−18.9	−8.0
4	−12.6	−12.6	−14.0	−13.6	−15.1	−11.3	−22.2	−21.3
5	−13.5	−10.7	−13.2	−11.8	−13.3	−8.5	−17.4	−13.9
MEAN	**−13.1**	**−11.6**	**−13.4**	**−12.1**	**−11.7**	**−5.0**	**−15.5**	**−9.5**
SE	**0.6**	**0.6**	**0.4**	**0.5**	**2.3**	**2.8**	**2.6**	**4.0**
AVG[a]	**−13.3**	**−11.6**	**−13.4**	**−12.2**	**−11.7**	**−5.2**	**−15.7**	**−9.4**
Slopes								
1	0.060	0.088	0.091	0.050	0.031	0.033	0.026	0.034
2	0.081	0.071	0.123	0.081	0.023	0.030	0.033	0.027
3	0.084	0.090	0.095	0.078	0.022	0.022	0.025	0.022
4	0.082	0.075	0.112	0.085	0.025	0.029	0.025	0.019
5	0.070	0.112	0.114	0.061	0.022	0.019	0.023	0.022
MEAN	**0.075**	**0.087**	**0.107**	**0.071**	**0.025**	**0.026**	**0.027**	**0.025**
SE	**0.004**	**0.007**	**0.006**	**0.007**	**0.002**	**0.003**	**0.002**	**0.003**
AVG[a]	**0.070**	**0.079**	**0.101**	**0.066**	**0.024**	**0.023**	**0.024**	**0.021**

[a] Calculated from the psychometric function fit to the group mean results.

Next we examined whether performance varied as a function of word position for noise and speech maskers. The data were pooled from the middle portions of the performance-level functions which were −15, −10, and −5 dB T/M for noise and −10, −5, and 0 dB T/M for speech. The results of this finer-grained analysis are shown in Fig. 4. The group mean proportion correct scores and standard errors of the means are plotted as a function of word position for the noise masker (left panel) and the speech masker (right panel).

First, better performance occurred in all conditions for the last word in the string with relatively constant performance found for other word positions. The better performance for the last word is likely a recency effect—a well-known phenomenon in serial recall—and has been reported previously using similar stimuli but somewhat different methods (Kidd *et al.*, 2008b; see also Ruggles and Shinn-Cunningham, 2011). Also, because the sequences were time-aligned at the beginning and had variable lengths depending on the words chosen, the last target word occasionally extended beyond the two maskers and thus may have been easier to identify. An interesting minor finding was the small bump (better performance) in the noise condition in the third word position (the number category) for the *syntactic* target word order. A re-sorting and analysis of data for the *random* conditions indicated that this performance bump was consistent for the "number" words and not for their specific position in the string. That is, better performance in noise was found when the word was a digit regardless of position. We do not have a definitive explanation for this result. If it were simply that numbers are "overlearned" relative to the other words (and categories of words) in the corpus we might have expected a performance bump in the speech maskers too. So, this small but interesting effect requires further study. It is also noteworthy that there was no primacy effect apparent meaning that the first word in the string—for any of the conditions—was not, on average, reported more accurately than subsequent words.

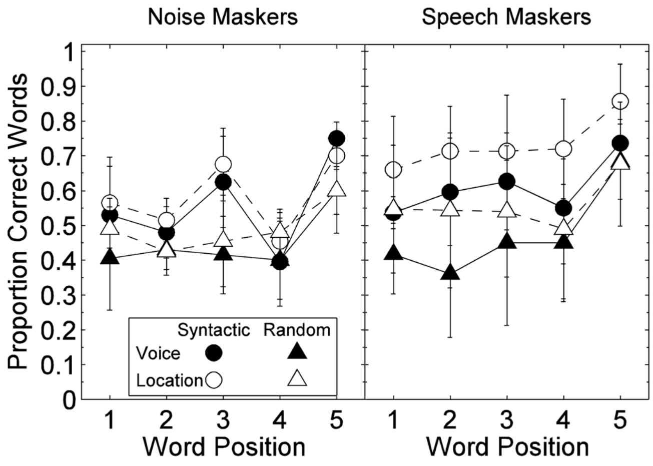


FIG. 4. Group mean proportion correct and standard errors of the means plotted as a function of word position from experiment 1A. The left panel displays the results from the noise maskers and the right panel displays the results from the speech maskers. The values were averaged across T/Ms from the middle portions of the functions in Fig. 2 (see text). The filled symbols are the results when the target designator was voice while the open symbols are the results when the target designator was location. Circles show *syntactic* target results while triangles are for *random* target sentences.

B. Single-word control

In the preceding experiment, imposing syntactic structure on a string of words provided a significant benefit in speech identification performance under conditions of high informational masking when the primary means for designating the target was either a talker's voice or an apparent location. This finding suggested that correct syntax could strengthen the focus of attention on a stream of speech when the words were linked by another feature; i.e., constant voice or location. This led to the question of whether correct syntactic order, in and of itself, would provide a sufficient cue for selecting (from several competing talkers) and linking together target words when other features were unreliable. However, for syntactic structure to have an influence on maintaining the focus of attention on a stream of speech it must—by definition—conform to previously learned rules applied to *sequences* of words whereas features like voice or location may be beneficial in the absence of sequential relations among words or even for isolated words. Preliminary to addressing the question above, we conducted a simple "control" experiment in which only a single test word was presented concurrently with two competing words from different talkers and at different locations. The single-word control was intended to determine how much benefit the cues of voice and location *per se* provided in the absence of any sequential dependencies among the target words. This experiment was identical to the *random* target condition tested above but truncated so that each trial consisted of only one test word and two concurrent masker words, as if only the first of the five words in that experiment were presented. The three concurrent words were drawn from mutually exclusive word categories, were spoken by different talkers, and were presented from different apparent locations. As in experiment 1A, the target was designated either by a voice cue (i.e., report the test word from the same talker as the cue) or by a location cue (i.e., report the test word from the same location as the cue). In addition, performance was measured for a case in which the three words were presented with no explicit *a priori* cue to the target. Following each stimulus within a trial, the set of eight alternatives from which the target was drawn was displayed and the listener selected from among those alternatives for a response. So, the target was essentially designated after the stimulus by limiting the responses to the exemplars from a single word category. For example, the stimuli on one trial might be "red," "Gene," and "six" with each word spoken by a different talker and presented from a different interaural location. No indication was given of which word category, talker or location designated the target until after the stimulus. If the target word was "six," the eight alternatives would include that word as well as other digits but not "red" or "Gene," exactly as in the first experiment for each of the five word categories. For this part of the experiment, the target and masker words were presented only at the T/M ratio of 0 dB.

Eight subjects participated in this control condition (one subject also participated in experiment 1A). The results are shown in Fig. 5. This figure contains group mean proportion correct scores and standard errors of the means. Performance

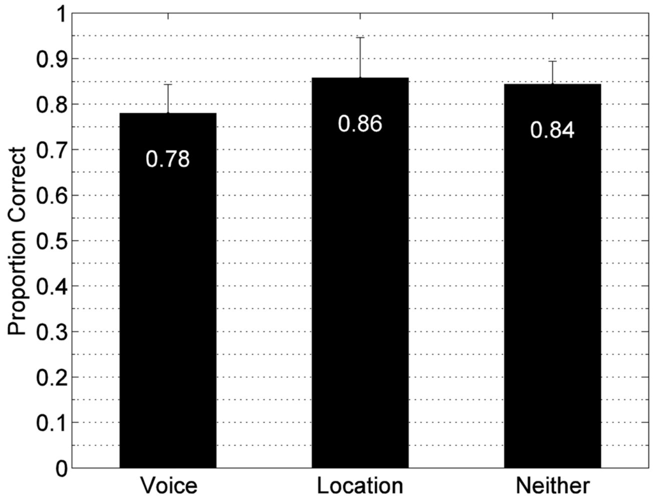


FIG. 5. Group mean proportion correct scores and standard errors of the means for the single-word condition in experiment 1B. The left bar shows the results for the voice cue condition, the middle bar is for the location cue condition, and the right bar is for neither cue with the target word category presented following the stimulus.

was similar across the three conditions with the highest proportion correct score found for the location cue (0.86) and the lowest score found for the voice cue (0.78). Surprisingly, the no-cue condition ("neither") was intermediate at 0.84 proportion correct. A repeated measures analysis of variance revealed a significant effect of cue condition [F(2,14) = 6.14, p = 0.012]. Pairwise comparisons with Bonferroni corrections were not significant except for the voice cue versus no cue. The mean difference between these conditions was 0.064 which was slightly smaller than the difference between location and voice cues (0.078); however, the variance was also smaller (standard errors of 0.017 and 0.029, respectively). These results suggest that the three concurrent words were well-segregated, and that listeners were able to successfully select among them from memory following presentation.

Although this result was expected for the voice and location cues based on the results of experiment 1A, it was not anticipated that listener performance would also be so accurate for the no-cue condition given that, presumably, attention either was unfocused, arbitrarily directed to one source, or perhaps distributed among the three alternatives. Consistent with the performance scores some listeners anecdotally reported that attempting to listen for a specific voice was more difficult than simply listening in an unfocused or global manner perhaps because of the similarity of the three young female talkers used.

IV. EXPERIMENT 2: SYNTAX AS THE ONLY MEANS OF DESIGNATING THE TARGET

In the next experiment, the purpose was to test the benefit of syntactic structure when other cues were not reliable. The primary condition of interest was one in which the *only* means for identifying the target talker was that the selections for the target sentence were made in correct syntactic order with both voice and location randomized across words. As indicated in the last row of Fig. 1, the "cue" was simply the word "Ready" spoken concurrently by the three talkers one from each location. So, the cue did not designate the target per se and only alerted the listener to the beginning of the trial. Two conditions from experiment 1 were also repeated so that within-subjects comparisons were available. These are the first two conditions described in Fig. 1, i.e., the target had correct syntax and in one case the defining feature was talker voice and in the other it was apparent location.

The results from this experiment are shown in the left panel of Fig. 6 plotted as group mean psychometric functions. The slopes and midpoints from the individual subjects are presented in Table II. As expected, the slopes and midpoints of the functions were similar to those reported in experiment 1A for the voice and location designations with correct target syntax. Performance differed most among the three conditions in the region near, or just below, 0 dB T/M. The syntax-only condition exhibited a more pronounced plateau than the other conditions increasing only about 5 percentage points between −15 and 0 dB T/M.

The group mean T/Ms at threshold extracted from these functions are plotted in the right panel of Fig. 6. For the voice cue, the group mean threshold was −9.7 dB T/M and the slope was 0.03 while the group mean threshold for the location cue condition was −15.6 dB T/M also with a slope of 0.03. For the condition of greatest interest here in which correct syntax formed the only cue to the target the group mean threshold T/M was −6.1 dB with slope of 0.02 with individual thresholds ranging from −1.2 to −12.3 dB. Thus, the syntax-only condition yielded a threshold T/M that was about 3.6 dB poorer than those obtained when the additional linking by voice was available and about 9.5 dB poorer than the case of constant location.

A repeated-measures ANOVA indicated that the single factor of cue type (voice, location or syntax-only) was significant [F(2,16) = 27.2, p < 0.001]. Further analysis revealed

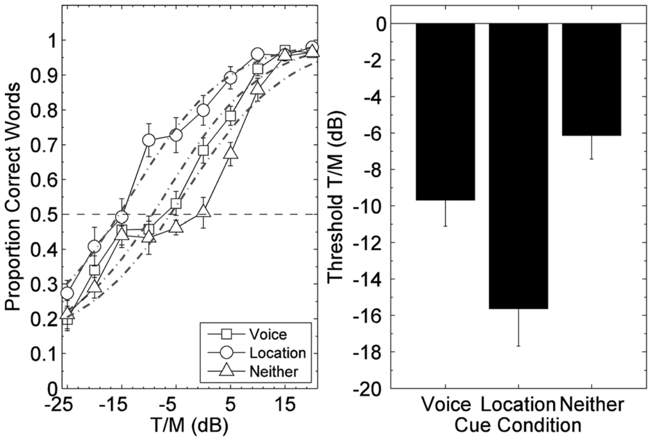


FIG. 6. The left panel shows group mean proportion correct performance and standard errors of the means plotted as a function of T/M from experiment 2. The data points are plotted along with best-fitting logistic functions (dot-dashed lines). The three types of designators for the target sentence are indicated in the symbol key. Group mean "thresholds" (target-to-masker ratios in dB corresponding to the 0.5 proportion correct points on the fitted psychometric functions) and standard errors of the means from experiment 2 are shown in the right panel. The three conditions shown are for when the target designator was voice (left bar), location (middle bar) or neither (right bar, correct syntax only).

TABLE II. Thresholds and slopes from the fitted psychometric functions for the individual and group−mean results from experiment 2. The three cue conditions are voice, location and syntax.

Listener	Voice	Location	Syntax
Thresholds			
1	−9.3	−14.0	−6.3
2	−1.1	−10.1	−1.4
3	−11.5	−13.2	−2.5
4	−6.6	−18.1	−6.6
5	−6.6	−3.7	−1.2
6	−14.0	−17.4	−10.2
MEAN	**−8.2**	**−12.8**	**−4.7**
SE	**1.8**	**2.2**	**1.5**
AVG[a]	**−7.9**	**−12.7**	**−4.7**
Slopes			
1	0.027	0.027	0.026
2	0.041	0.051	0.032
3	0.019	0.021	0.015
4	0.026	0.021	0.021
5	0.033	0.032	0.031
6	0.031	0.044	0.028
MEAN	**0.030**	**0.033**	**0.025**
SE	**0.003**	**0.005**	**0.003**
AVG[a]	**0.027**	**0.026**	**0.024**

[a] Calculated from the psychometric function fit to the group mean results.

all three pairwise comparisons with Bonferroni corrections were also significant; voice versus location ($p = 0.018$), voice versus syntax ($p = 0.04$), and location versus syntax ($p < 0.001$).

Comparing the voice and location conditions across the two experiments (for syntactically correct targets in experiment 1A), this second group of subjects had mean thresholds about 2 dB higher than the first group in the voice condition and nearly identical to the first group for the location condition. Comparison of Tables I and II highlights the large individual differences found between subjects on this task. Across the two experiments, the range of individual values when the target was designated by voice was about 14 dB while it was about 20 dB when designated by location. Because location as an *a priori* cue consistently yielded lower thresholds than voice as a cue in both experiments 1A and 2 (as well as in the single-word control condition in experiment 1B) a further analysis was conducted pooling the results across the two experiments for the conditions in common. This analysis included the first two conditions represented in Fig. 1 (a syntactically correct target designated by voice or location) and the 13 unique listeners (only the results from the first experiment were included for the subject who participated in both). The group-mean thresholds pooled in this manner were thus −10.5 dB (standard error of 1.3 dB) for voice and −15.7 dB for location (standard error of 1.7 dB) cues. A one-way repeated measures ANOVA found that this difference was significant [$F(1,12) = 14.05$, $p = 0.003$]. Thus, for the specific conditions tested here (e.g., three arbitrary female talkers, ITDs of 800, 0, and −800 μs) performance was better when cued by location than by voice.

V. DISCUSSION

For the noise-masker conditions tested in experiment 1A there was no ambiguity about which source formed the target source: Any audible words were necessarily target words. Despite the fact that there were three concurrent sources from three distinct locations, the noise-masker conditions posed no challenge for the listener in selecting the target sound source and presumably only audibility (due to energetic masking) and memory capacity would limit performance. Because performance at high T/Ms for the noise masker for either the *syntactic* or the *random* five-word strings was at ceiling, the capacity for serial recall per se was not a factor in limiting performance. Furthermore, the main variables of this study: Voice versus location cues and correct versus incorrect syntax, were not significant factors affecting the outcomes of the noise-masker conditions. It should be noted that the relatively steep performance-level functions and low intersubject variability (standard errors of the means around 0.5 dB), found here are typical for energetic masking of speech.

For the speech masking conditions, the performance-level functions were shallower and less orderly than those found for noise masking conditions and the intersubject variability in performance was greater (standard errors of the means ranging from 1.5 to 4 dB). The shallower performance-level functions and high intersubject variability are typical of informational masking (e.g., Freyman *et al.*, 1999; Brungart, 2001a; Brungart *et al.*, 2001; Arbogast *et al.*, 2002). Moreover, "plateaus" were apparent in the performance-level functions for group-mean and individual results (not shown). These plateau regions may be indicative of the influence of segregation by level, and perhaps by glimpses of target speech in masker envelope minima. This result has been reported in the past under similar speech-on-speech masking conditions and tends to occur when the targets and maskers are very similar, as with same talker or same-sex talkers (e.g., Brungart, 2001a; Brungart *et al.*, 2001; Wightman *et al.*, 2006). Because performance in speech-on-speech masking is not due simply to acoustic overlap/energetic masking (cf. Brungart *et al.*, 2006), perceptual factors involved in the segregation and selection of sources can strongly influence the results. These perceptual effects causing nonmonotonicities in the psychometric functions lead to some ambiguity in defining "threshold" T/Ms. However, in the absence of any compelling rationale for using a different approach, we computed the best-fitting logistic functions as represented in Figs. 2 and 6 for deriving the points on the functions corresponding to a proportion correct of 0.5 for making comparisons across conditions. Other means for extracting thresholds or comparing performance at different levels (e.g., defining thresholds at higher or lower performance levels on the fitted functions) did not yield values that altered the conclusions reached here.

Designating the target speech source according to voice or location was intended to provide a marker for the focusing of attention, albeit along very different perceptual dimensions. The results of the current study indicating that either of these two cues was highly effective in assisting the listener in segregating and selecting a target talker among competing talkers are not surprising based on a considerable body of past work (e.g., reviews in Yost, 1997; Bronkhorst, 2000; Kidd *et al.*, 2008a; Mattys *et al.*, 2012). The conclusion about the efficacy of voice and location cues from the current study is based on the findings indicating that speech identification performance was much above chance when such cues were the *only* basis for selecting the target words (e.g., by randomizing the complementary variable in the non-syntactic word order conditions; cf. Best *et al.*, 2008; Maddox and Shinn-Cunningham, 2012).

The finding that there was a significant difference in performance observed between the two *a priori* cues under speech masking conditions, as indicated by the pooled analysis reported in the preceding section, suggests that they differed in the degree to which they aided speech stream selection. As discussed more fully below, this difference in effectiveness was also apparent to varying degrees in the single-word control results (Fig. 5) and in the roughly constant performance as a function of word position (Fig. 3) in experiment 1, suggesting that location was the superior cue in both the initial segregation/selection of sources and in the ongoing maintenance of the target streams in competition with the speech maskers. Apparent interaural location was mapped according to a single stimulus variable: ITD. The ITDs were chosen so that the locations were clearly distinct perceptually and were based on the approximate values of the largest ITDs normally encountered. Also, the unidimensional nature of the mapping between the physical variable and its perceptual correlate may have fostered a strong attentional focus. It seems likely that voice was less distinct perhaps because of its representation along multiple physical dimensions and, for this set of three young-adult same-sex talkers, there was a high degree of similarity along some of these dimensions. A different set of talkers, especially if it included both males and females, or if the talkers differed significantly in age, might have increased the effectiveness of voice as a cue. Because the words were spoken individually with neutral inflection when recorded (cf. Kidd *et al.*, 2008b), it also is possible that less of a benefit of voice was found here than would occur for natural speech where coarticulation and intonation patterns may provide cues linking words together. Furthermore, the familiarity of the talkers was not assessed or explicitly manipulated and that factor has been shown to exert an influence on the ability to selectively attend to one talker among competing talkers (e.g., Newman and Evers, 2007; Johnsrude *et al.*, 2013).

One way of viewing the ability of listeners to exploit these acoustic cues is that they served to reduce listener uncertainty. In that sense, there was a top-down component inherent to solving the task. For the listener, *a priori* knowledge fostered *expectation* about the occurrence of a key feature that would reduce ambiguity about which words should be selected from the three concurrent sources and linked with other words stored in memory. In thinking about how syntactic structure may affect performance under conditions of high source uncertainty, consideration of the single-word control condition provides some insight. Listeners were able to choose one of three concurrent words with a high degree of accuracy when cued either by voice or by location. This suggests that energetic masking had little influence on performance consistent with quantitative estimates of energetic masking under reasonably comparable speechmasking conditions (Brungart *et al.*, 2006). Significantly, though, equivalent performance was found when *neither* cue was provided prior to the stimulus. Instead the target was only designated by the allowable list of response alternatives—each of which was an exemplar from the target word category and did not contain either of the masker words—*after* the stimulus was presented. Because the target word category was randomized among the five categories (exemplars from three mutually exclusive categories presented on each trial), each of the three words presented on a given trial was equally likely to be the target until the word category was designated following presentation. Identification performance depended on the listener storing all three of the words in memory and then retrieving the correct word once the word category was given. Thus, the prior focus of selective attention was not necessary to achieve a high level of identification performance when the memory demands were relatively low. On a word-by-word basis then, it seems that all three words were initially available to the listener (i.e., not energetically masked and perceptually segregated) and were robustly maintained in memory (i.e., represented sufficiently well to support identification in a 1 of 8 forced-choice format).

For strings of connected speech, however, the presentation of sequences of concurrent words would likely overwhelm memory capacity in the absence of a means for rapidly implementing selection by prior cuing. For example, Kidd *et al.* (2005) examined performance using the Coordinate Response Measure test (e.g., Brungart, 2001b) when *a posteriori* cues to target location were given for three concurrent sentences from different talkers located at three different source azimuths. They found that performance was no better than chance (i.e., randomly choosing 1 of the 3 sources to attend) under those conditions that required recalling only two test words (1 of 4 colors and 1 of 8 numbers) near the end of the target sentences. When location was cued before the stimulus identification, a high proportion correct score (above 0.9) was observed as was the case in the most comparable conditions tested here. Despite the differences in design, the findings from these two studies suggest that *a posteriori* selection and serial recall of the items from three concurrent sources quickly deteriorates when more than a single item per source must be identified.

Even when the acoustic features of voice and location were cued beforehand, there remained a significant benefit of correct target sentence syntax. It has long been known that recall of word strings is superior when the words are arranged in meaningful sentences. For example, Campoy and Baddeley (2008) state that "…immediate recall of verbal material can benefit from the short-term maintenance of other kinds of information, especially semantic. Studies of the immediate recall of sentences, for example, have shown that participants can recall many more words from sentences than from lists of unrelated words (Brener, 1940). There are a number of factors that could contribute to this phenomenon, including syntactic constraints and lexically based sequential redundancy. However, it seems likely that sentence superiority is in part a consequence of participants being able to maintain the overall meaning of the sentence and use this information at recall" (p. 330). The design of the current study, in which the random within-category selections of words limited the predictive value of semantic content, presumably constrained the listeners to rely principally on syntax rather than semantics[1] for obtaining a benefit in recall.

In the present study, when either voice or location was cued, the listener could prepare to attend to the designated feature and link in memory only those words possessing that feature. Thus, the listener could select or discard competing words upon discrimination of that feature on a word by word basis. *A priori* cuing of voice or location would provide a means for rapidly selecting the target words maintaining the focus of attention on the correct stream of speech (e.g., Freyman *et al.*, 2004; Best *et al.*, 2007; Maddox and Shinn-Cunningham, 2012). This view is consistent with the idea that these are low-level factors that do not require extensive higher-level processing in order to be useful in discrimination. For example, choosing among concurrent stimuli based on the discrimination of a simple acoustic feature would not require the processing necessary to identify the words; that could occur *after* selection. However, *a priori* knowledge about *syntax* would not allow the listener to prepare to attend and sort the stimuli in quite the same manner. Rather than listening for a specific acoustic feature or property, the listener must identify each word first so that its category could be determined. At that point, the word could be selected and linked to the other words stored in memory. However, the extent to which all three concurrent words must be identified before one is selected in preference to the others is not entirely clear from the current findings. It is possible that this concurrent syntactically based identification/selection happens rapidly without fully comprehending the meaning of each word. Regardless of how fully the competing sounds must be recognized, the processing required to accomplish this feat would certainly be more extensive than discrimination based on a simple acoustic feature.

One factor potentially contributing to the beneficial role of syntax found here is the differing degree of uncertainty experienced by the listener for *syntactic* and *random* targets. Although chance performance was the same in all cases (1 of 8 for each word selection), the number of possible target word alternatives is greater for the *random* condition than for the *syntactic* condition viewed from the perspective of the listener prior to, and during, the stimulus.[2] In the *random* case, the number of possible alternatives ranged from 40 for the first word in the string to 8 for the final word because word categories were not repeated during the sequence. The category of each word in sequence was excluded from subsequent presentations so that the number of alternatives for the first word was 40, for the second word was 32, etc. For the *syntactic* word strings, the number of alternatives was known to be eight for each word position because the word category order was predefined. In a sense, this may be thought of as a general benefit of syntax even in naturally occurring speech in that it promotes a degree of predictability of the words in sequence. However, this interpretation does not seem to be sufficient as an explanation for the current results. The uncertainty in the *random* conditions is reduced as the sequence progresses so that, as noted earlier, the number of possible alternatives is equal to the *syntactic* case by the final word. If the degree of uncertainty governed identification, it would be expected that performance would have improved throughout the sequence as uncertainty declined. This prediction would apply equally for both speech and noise maskers. However, there was no evidence of a trend toward improved identification scores for the *random* target—except a recency effect for the final word position that was apparent under all conditions. So, although listener uncertainty may have declined as the target string progressed, the conclusion that the difference in performance between *syntactic* and *random* conditions was due simply to a reduction in the size of the set of possible word alternatives is not supported by these results.

Another plausible explanation for the difference in performance between *syntactic* and *random* target conditions is based on the degree of structural similarity between target and masker speech. Brouwer *et al.* (2012) have proposed a "linguistic similarity hypothesis" that postulates that the greater the degree of similarity—on a linguistic level—between speech sources, the greater the informational masking that results. According to Brouwer *et al.* (2012) formulation of linguistic similarity, syntax and semantic content are considered to be "stimulus-related factors" specific to speech. They varied similarity according to the language of the target and masker talkers (i.e., target in one language, masker in the same or different language while also manipulating whether the languages were primary, secondary or not understood by the listener; cf. Ezzatian *et al.*, 2010) and whether target and masker speech were at the same or different level of semantic content (i.e., semantically meaningful or "anomalous"). For both manipulations—language and semantic content—the observed amount of informational masking increased when the target and masker speech was similar as compared to dissimilar according to their criteria. These differences in performance due to linguistic factors occurred even in the absence of reliable differences in "general auditory distance" between stimuli (which presumably would describe the location and voice manipulations in the present study).

In the current study, the advantage observed for the *syntactic* conditions could be attributed to the dissimilarity between target and masker strings with respect to the linguistic variable of syntax. Viewed from that perspective, the present results are consistent with the linguistic similarity hypothesis described above. Presumably a further test of this hypothesis would be to pair random order target word strings with syntactically correct masker strings, with the prediction being that less masking would occur due to this dissimilarity. We did not attempt to test this condition. Earlier work from our laboratory (Kidd *et al.*, 2008b) using this corpus and similarly constructed five-word target strings of speech masked by equal-length strings of maskers drawn from the same corpus, found that variations in syntactic similarity between target and masker did not significantly affect performance. In the Kidd *et al.* (2008b) study, the target and masker words were interleaved in time so that the target occupied the odd-numbered words in the sequence while the masker occupied the even-numbered words (cf. Broadbent, 1952). Unlike the current study, though, both target and masker syntax varied in the five-word strings. Significant differences were found for identification performance when the target word syntax varied (as here, better performance for correct syntax) but masker syntax was not a significant factor. There were many differences between the present study, the Kidd *et al.* (2008b) study, and the Brouwer *et al.* (2012) study with respect to the methods that were used. Perhaps most importantly, though, Brouwer *et al.* employed semantically more meaningful target utterances than the other two studies.

Another example of informational masking of speech due to linguistic factors is the difference between the effectiveness of a speech masker presented normally versus the same masker presented time reversed. Historically, the observation that time-reversed speech retains many of the low-level features of normal speech while eliminating linguistic content has led to comparisons of performance under the two conditions as a means of gauging linguistic components of masking (e.g., Dirks and Bower, 1969). Although Dirks and Bower (1969) did not find any significant differences between speech maskers presented forward versus reversed, a number of more recent investigations have. For example, in the Kidd *et al.* (2008b) study mentioned previously that used the same five-word strings as in the current study, large differences were found in performance for time-reversed maskers compared to time-forward maskers with both interfering more than noise. Other studies reporting large differences due to time-reversal of speech include Freyman *et al.* (1999), Marrone *et al.* (2008), Kidd *et al.* (2010) and Best *et al.* (2012). And, at an even lower level of target-masker complexity, Uslar *et al.* (2013) found small effects of the linguistic complexity of target sentence construction on speech reception thresholds measured in unmasked or steady-state noise masking conditions but significantly larger effects when the noise was envelope modulated. They concluded that the fluctuating noise interacted with the linguistic complexity of the target sentence identification task to produce a greater cognitive load on the listener. These findings are consistent with the proposition that a hierarchy of complex linguistic factors affects the masking one or more speech source(s) exerts on a target source.

Theories of perceptual and/or cognitive load (e.g., Lavie *et al.*, 2004) offer useful insights into the general problem of speech-on-speech masking (e.g., Francis, 2010). The primary assumption upon which load theory is based, as it is applied here, is that there is a limited "pool of resources" available to the observer for performing the tasks of sound source segregation and subsequent processing (e.g., identification of semantic content of the target/attended source). If the perceptual segregation task is difficult—perhaps because of insufficient or unreliable low-level cues—the interference caused by concurrent maskers is relatively low because all of the available resources are engaged by the segregation task. In contrast, if perceptual segregation is easily accomplished, as in the current study, additional resources are available to engage the nontarget sources essentially passing the representations of those sounds along to higher processing levels. Selection among the segregated sources at this later stage invokes the various cognitive functions required to solve the task. At that point, factors such as the similarity of these segregated source representations, and the linguistic and memory demands they place on the observer, determine the interference they create. Because the sources used in the present study were well-segregated (as inferred from the single-word control case), it would appear that the benefit of syntax, unsurprisingly, is broadly consistent with a decrease in cognitive load. In the case of target designation by syntax only, selection of the well-segregated sources likely occurs at a higher level of processing than when voice or location is available. In that case, some degree of identification of the competing words must occur so that selection according to word category may be accomplished. It is possible that the current approach could be adapted for use in examining the issue of limitations on processing resources at the levels of segregation vs identification by manipulating other variables such as, for example, changing the rate of presentation of the words or perhaps by using a dual task.

A final point concerns the role of timing in these word recall tasks. In the single-word control condition, listeners were highly successful in selecting one of three concurrently presented words. In quiet or high T/M conditions, near perfect recall was observed for both *syntactic* and *random* word order presentation for five-word strings. Taken together, these findings suggest that neither segregation nor recall limited performance, but instead, it was the dynamic aspect of segregation, selection, storage, and retrieval that occurred for the word sequences. Thus it seems likely that the magnitude of the effects found here—the benefit of word predictability under speech masking conditions—depends on factors such as the length of time between words and the number of words comprising the sequences and may be particularly important during the normally rapid flow of conversation.

VI. CONCLUSIONS

The current results suggest that voice and location may be effective cues for segregating and selecting a target talker masked by competing talkers. When the target word strings conformed to a known syntactic structure, speech identification performance was better than when the target word strings were presented in random order, regardless of whether the primary source selection cue was voice or location. The benefit of syntax was greater when the maskers were competing speech, producing high amounts of informational masking, than when the maskers were independent noises, producing high amounts of energetic masking.

When neither voice nor location provided reliable cues for selecting and maintaining the target speech stream, syntax alone was sufficient to support better than chance identification performance. This suggests that both low-level and high-level cues serve to select and maintain the focus of attention on one specific talker in competition with other talkers and to extract the information contained in the stream of speech uttered by that talker.

ACKNOWLEDGMENTS

Work supported by NIH/NIDCD grants DC04545 and DC04663 and by AFOSR award FA955012-1-0171. The authors are grateful to Sudha Arunachalam for comments on an earlier version of this manuscript.

[1] Although the random selection of words within categories mitigated meaning as a reliable basis for predicting the occurrence of specific words, it could be argued that the structure imposed by the syntactic form used here provided some level of semantic knowledge; i.e., an actor (name) engaged in an action (verb) on an object (noun), etc. that provided some basis for predictability/expectation.

[2] The number of response alternatives—and thus chance performance—is the same for syntactically correct and random-order target word strings. However, during the test sequence, the listener knows the order of the target word categories—i.e., by definition it conforms to the syntactic structure used—and therefore their uncertainty on a word-by-word basis is expected to be equal to one of eight possibilities. This is true for the random word order case when the response word categories are identified at the end of the stimulus, but during the sequence listener "uncertainty" is presumably greater; i.e., they do not know which word category to expect in the next word position.

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

Best, V., Marrone, N., Mason, C. R., and Kidd, G., Jr. (**2012**). "The influence of non-spatial factors on measures of spatial release from masking," J. Acoust. Soc. Am. **131**, 3103–3110.

Best, V. Ozmeral, E., Kopčo, N., and Shinn-Cunningham, B. G. (**2008**). "Object continuity enhances selective auditory attention," Proc. Natl. Acad. Sci. **105**, 13173–13177.

Best, V., Ozmeral, E. J., and Shinn-Cunningham, B. G. (**2007**). "Visually-guided attention enhances target identification in a complex auditory scene," J. Assoc. Res. Otolaryng. **8**, 294–304.

Best, V., Shinn-Cunningham, B. G., Ozmeral, E. J., and Kopčo, N. (**2010**). "Exploring the benefit of auditory spatial continuity," J. Acoust. Soc. Am. **127**, EL258–EL264.

Binns, C., and Culling, J. F. (**2007**). "The role of fundamental frequency contours in the perception of speech against competing speech," J. Acoust. Soc. Am. **122**, 1765–1776.

Bregman, A. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sounds* (Bradford Press, MIT Press, Cambridge, MA), pp 1–792.

Brener, R. (**1940**). "An experimental investigation of memory span," J. Exp. Psych. **26**, 467–482.

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

Bronkhorst, A. W. (**2000**). "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," Acta Acust. Acust. **86**, 117–128.

Brouwer, S., Van Engen, K., Calandruccio, L., and Bradlow, A. R. (**2012**). "Linguistic contributions to speech-on-speech masking for native and non-native listeners: Language familiarity and semantic content," J. Acoust. Soc. Am. **131**, 1449–1464.

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

Campoy, G., and Baddeley, A. (**2008**). "Phonological and semantic strategies in immediate serial recall," Memory **16**, 329–340.

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.

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

Denham, S. L., and Winkler, I. (**2006**). "The role of predictive models in the formation of auditory streams," J. Physiol. Paris **100**, 154–170.

Dirks, D. D., and Bower, D. R. (**1969**). "Masking effects of speech competing messages," J. Speech Hear. Res. **12**, 229–245.

Ezzatian, P., Avivi, M., and Schneider, B. A. (**2010**). "Do nonnative listeners benefit as much as native listeners from spatial cues that release speech from masking?," Speech Commun. **52**, 919–929.

Francis, A. L. (**2010**). "Improved segregation of simultaneous talkers differentially affects perceptual and cognitive capacity demands for recognizing speech in competing speech," Atten. Percept. Psychophys. **72**, 501–516.

Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2004**). "Effect of number of masker talkers and auditory priming on informational masking in speech recognition," J. Acoust. Soc. Am. **115**, 2246–2256.

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

Johnsrude, I. G., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., and Carlyon, R. P. (**2013**). "Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice," Psychol. Sci. **24**, 1995–2004.

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

Kidd, G., Jr., Best, V., and Mason, C. R. (**2008b**). "Listening to every other word: Examining the strength of linkage variables in forming streams of speech," J. Acoust. Soc. Am. **124**, 3793–3802.

Kidd, G., Jr., Mason, C. R., and Arbogast, T. L. (**2002**). "Similarity, uncertainty and masking in the identification of nonspeech auditory patterns," J. Acoust. Soc. Am. **111**, 1367–1376.

Kidd, G., Jr., Mason, C. R., Best, V., and Marrone, N. L. (**2010**). "Stimulus factors influencing spatial release from speech on speech masking," J. Acoust. Soc. Am. **128**, 1965–1978.

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

Kidd, G., Jr., Mason, C. R., Streeter, T., Thompson, E. R., Best, V., and Wakefield, G. P. (**2013**). "Perceiving sequential dependencies in auditory streams," J. Acoust. Soc. Am. **134**, 1215–1231.

Lavie, N., Hirst, A., de Fockert, J. W., and Viding, E. (**2004**). "Load theory of selective attention and cognitive control," J. Exp. Psychol. **133**, 339–354.

Maddox, R. K., and Shinn-Cunningham, B. G. (**2012**). "Influence of task-relevant and task-irrelevant feature continuity on selective auditory attention," J. Assoc. Res. Otolaryn. **13**, 119–129.

Marrone, N., Mason, C. R., and Kidd, G., Jr. (**2008**). "Tuning in the spatial dimension: Evidence from a masked speech identification task," J. Acoust. Soc. Am. **124**, 1146–1158.

Mattys, S. L., Davis, M. H., Bradlow, A. R., and Scott, S. K. (**2012**). "Speech recognition in adverse conditions: A review," Lang. Cognit. Processes **27**, 953–978.

Moore, B. C. J., and Gockel, H. (**2012**). "Properties of auditory stream formation," Philos. Trans. R. Soc. B. **367**, 919–931.

Newman, R. S., and Evers, S. (**2007**). "The effect of talker familiarity on stream segregation," J. Phonetics **35**, 85–103.

Rabiner, L. R. (**1989**). "A tutorial on Hidden Markov Models and selected applications in speech recognition," Proc. IEEE **77**, 257–286.

Rajendran, V. G., Harper, N. S., Willmore, B. D., Hartmann, W. M., and Schnupp, J. W. H. (**2013**). "Temporal predictability as a grouping cue in the perception of auditory streams," J. Acoust. Soc. Am. **134**, EL98–EL104.

Ruggles, D., and Shinn-Cunningham, B. G. (**2011**). "Spatial selective auditory attention in the presence of reverberant energy: Individual differences in normal-hearing listeners," J. Assoc. Res. Otolaryn. **12**, 395–405.

Shamma, S. A., Elhilali, M., and Micheyl, C. (**2010**). "Temporal coherence and attention in auditory scene analysis," Trends Neurosci. **34**, 114–123.

Uslar, V. N., Carroll, R., Hanke, N., Hamann, C., Ruigendijk, E., Brand, T., and Kollmeier, B. (**2013**). "Development and evaluation of a linguistically and audiologically controlled sentence test," J. Acoust. Soc. Am. **134**, 3039–3949.

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.

Winkler, I., Denham, S., Mill, R., Bohm, T. M., and Bendixen, A. (**2012**). "Multistability in auditory stream segregation: A predictive coding view," Proc. R. Soc. B. Biol. Sci. **367**(1591), 1001–1012.

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