Effect of number of masking talkers and auditory priming on informational masking in speech recognition

Richard L. Freyman,[a] Uma Balakrishnan, and Karen S. Helfer
Department of Communication Disorders, University of Massachusetts, Amherst, 715 N. Pleasant Street, Room 6 Arnold House, Amherst, Massachusetts 01003

(Received 11 July 2003; revised 18 January 2004; accepted 1 February 2004)

Three experiments investigated factors that influence the creation of and release from informational masking in speech recognition. The target stimuli were nonsense sentences spoken by a female talker. In experiment 1 the masker was a mixture of three, four, six, or ten female talkers, all reciting similar nonsense sentences. Listeners' recognition performance was measured with both target and masker presented from a front loudspeaker (F–F) or with a masker presented from two loudspeakers, with the right leading the front by 4 ms (F–RF). In the latter condition the target and masker appear to be from different locations. This aids recognition performance for one- and two-talker maskers, but not for noise. As the number of masking talkers increased to ten, the improvement in the F–RF condition diminished, but did not disappear. The second experiment investigated whether hearing a preview (prime) of the target sentence before it was presented in masking improved recognition for the last key word, which was not included in the prime. Marked improvements occurred only for the F–F condition with two-talker masking, not for continuous noise or F–RF two-talker masking. The third experiment found that the benefit of priming in the F–F condition was maintained if the prime sentence was spoken by a different talker or even if it was printed and read silently. These results suggest that informational masking can be overcome by factors that improve listeners' auditory attention toward the target. [DOI: 10.1121/1.1689343]

PACS numbers: 43.66.Dc, 43.66.Pn, 43.66.Qp, 43.71.Es [GDK] Pages: 2246–2256

I. INTRODUCTION

Listening to speech in the presence of competing speech is a complex perceptual task that has been the subject of considerable study over many years (e.g., Broadbent, 1952; Cherry, 1953; Duquesnoy, 1983; Yost *et al.*, 1996; Hawley *et al.*, 1999; Arbogast *et al.*, 2002). One of the characteristics of competing speech that should make it less effective than a continuous masker is that it fluctuates over time, both in spectral composition and in amplitude. There are brief pauses in the competing speech at phrase and sentence boundaries, closures during stop consonants, and very weak consonants, such as /f/ and /θ/, that all create instances of reduced masking. In addition, the spectrum of the competing speech fluctuates independently from the spectrum of the target speech. For example, a high-frequency /s/ sound in the interfering speech may be present simultaneously with a lower-frequency vowel sound in the target speech. These spectral and amplitude fluctuations provide the listener with brief but numerous glimpses of the target speech under conditions in which the target-to-masker ratio is favorable. Normal-hearing listeners seem to be able to use these glimpses to help them understand the target speech in the presence of the competition (Peters *et al.*, 1998). In general, research demonstrates that, decibel for decibel, speech maskers are less effective than noise maskers (see Bronkhorst, 2000).

In contrast to this general result, there appear to be conditions in which competing speech produces additional masking processes beyond those existing for noise maskers (Carhart *et al.*, 1969; Freyman *et al.*, 1999, 2001; Brungart, 2001; Hall *et al.*, 2002). Under such conditions, listeners have great difficulty perceptually extracting target speech material from a complex mixture of voices. Carhart *et al.* (1969) used the term ''perceptual masking'' to describe this phenomenon. Borrowing from nonspeech experiments (e.g., Kidd *et al.*, 1994; Oh and Lutfi, 1998) with which this type of speech masking shares several characteristics, we, and others, have more recently used the term ''informational masking.'' The definition of informational masking in speech recognition appears to be quite broad, encompassing features of masking, or release from masking, that cannot be explained in terms of traditional energetic masking. Among these features are unusually shallow slopes of growth of speech recognition performance with increasing S–N ratio (Freyman *et al.*, 1999; Brungart, 2001; Arbogast *et al.*, 2002) and a large release from masking due to horizontal separation of target and masker (Freyman *et al.*, 1999, 2001; Arbogast *et al.*, 2002; Noble and Perrett, 2002). Both of these characteristics were observed in the multitone masking experiments conducted by Kidd *et al.* (1998).

Ultimately, the basis of informational masking in speech recognition may be discovered by identifying factors that overcome informational masking, allowing the listener to perceptually extract a target that is apparently already represented at some levels of the auditory nervous system. Our previous studies (Freyman *et al.*, 1999, 2001) have focused on creating perceived differences in location between target

[a]Electronic mail: rlf@comdis.umass.edu

and masker as a cue for the listener. Specifically, with target and masker produced from a common front (0-deg) location as a reference (the F–F condition), the experimental condition is where a second source of masking from 60 deg to the right is added, with the right loudspeaker leading the front loudspeaker by 4 ms (the F–RF condition). Due to the precedence effect, the masker is heard at a location very close to 60 deg to the right, while the target is heard directly in front. The F–RF condition produces absolutely no advantage in speech recognition when the masker is continuous noise, but can produce a substantial advantage when the masker is one or two additional voices. Perceptually, the problem of finding and following the target speech within a mixture of several voices is resolved when the apparent location of the masker is moved off to the side.

Although it is difficult to specify precisely the conditions under which informational masking of speech occurs, confusability of the target and masker appears to be a critical feature. For example, Brungart and Simpson (2002) found that a great deal more informational masking occurs when target and masking talkers are of the same sex rather than of the opposite sex, presumably because male and female voices are not highly confusable. There also are likely to be variations in the amount of informational masking within talkers of the same sex, depending on as yet unspecified variables, e.g., similarity in fundamental frequency, speaking rate, speech accent, types of speech materials, etc. For example, for target speech produced by a female native speaker of American English, Freyman *et al.* (2001) found differences in the amount of masking produced by different sets of two female talkers. A composite of two Dutch talkers speaking accented English produced less masking than a complex of two native English speakers.

Carhart *et al.* (1975) reported that the amount of perceptual masking is strongly related to the number of masking talkers. They found that perceptual masking grew as the number of masking talkers increased to three, then decreased as the number was increased further. Hall *et al.* (2002) also reported a large amount of masking for two-talker maskers both in adult and child listeners. Brungart *et al.* (2001) found that in diotic listening conditions, two and three masking talkers produced considerably more masking than one masking talker at low S–N ratios, presumably due to increases in both energetic and informational masking. Yost *et al.* (1996) found that a total of three talkers created considerably more difficulty with a divided attention task than a total of two talkers. Moreover, spatial cues were particularly effective in helping to resolve a condition with three voices, as compared to conditions with two voices. Similarly, Freyman *et al.* (1999, 2001) reported much greater masking for a two-talker masker than for either of the individual talkers separately. The increase was substantially larger in the F–F (nonspatial) condition than in the F–RF condition, in which there was a spatial cue.

The effect of perceived spatial separation in the F–RF condition, as well as the effect of number of talkers, may be explained by the auditory attentional processes in which the listener must be engaged to solve the task. Difficulties in focusing and maintaining attention are likely to be greatest for colocated target and maskers (e.g., the F–F condition). Even so, low-level cues for auditory grouping may still allow segregation into multiple speech streams. The listener attempts to selectively attend to the target utterance and ignore the masking utterance(s). However, especially when both target and masking talkers are of the same sex, attention must be paid to the masking utterances to determine whether they are part of the target speech stream. A listener might attend to the beginning of a masking talker's utterance, decide after a short period that it is not the target sentence, shift attention to the target after missing several words, and possibly lose focus again as attention is pulled away by the competing speech. With two masking talkers, there is likely to be even greater competition for attention than with one masking talker. However, as the number of masking talkers increases much further, they may well create mutual masking of one another, appear less like individual speech streams, and compete less with the target for attention. In the F–RF condition, with the masking talkers perceived in a different location from the target talker, it should be easier for the listener to attend to the target. This type of auditory spatial attention has been shown to provide advantages in both response time and accuracy for identification of nonspeech frequency patterns presented within an informational masking background (Arbogast and Kidd, 2001).

As the number of talkers increases, the additional masker waveforms fill in temporal and spectral gaps and increase the amount of energetic masking in both the F–F and F–RF conditions. However, in the F–F condition a substantial informational component may exist which may be nonmonotonically related to number of talkers, as discussed above. Thus, there is a prediction that the effect of number of masking talkers on speech recognition will proceed quite differently in spatial versus nonspatial conditions, and that the difference in performance in the two conditions will narrow considerably for large numbers of talkers. Experiment 1 of the current paper evaluates this prediction. This investigation also will reveal the number of talkers that produces maximum informational masking for the current stimuli, which will be useful in the design of other experiments with these stimuli, including experiments 2 and 3 of the current paper.

The view that the listener's problem in the nonspatial task is one of identifying and maintaining attention on the target suggests that performance will be improved by any manipulation that helps distinguish the target so that sustained attention can be directed toward it. While spatial separation is clearly useful, other cues may also be effective in helping listeners maintain focus on the target. In experiment 2, we evaluated the usefulness of one such cue, namely whether listeners' ability to follow the target within the target–masker complex is improved if the target is presented in quiet just before the masking trial. By hearing a preview of what to listen for, subjects may be better able to focus on the target early in the trial and less likely to have attention drawn away by the maskers. To make any improvement quantifiable, the last of three key words in a nonsense sentence target was omitted from the preview (priming) stimulus and only this last word was scored when the sentence was

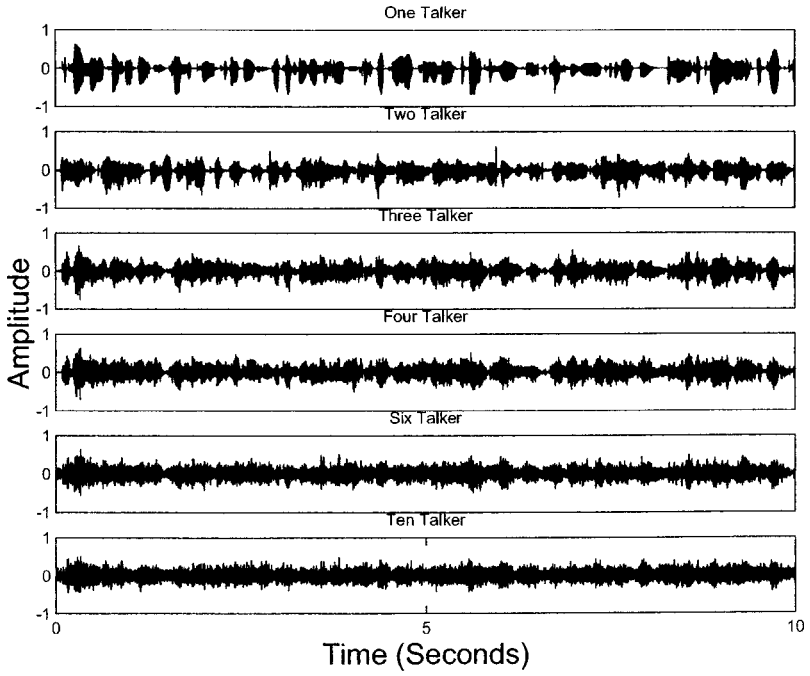

FIG. 1. Ten-second-long segments of single and multitalker maskers.

subsequently presented in masking. Experiment 3 was a follow-up study in which the prime sentence was spoken by a different talker than the target talker or the prime sentence was printed and read by the subject.

II. EXPERIMENT 1: EFFECT OF NUMBER OF TALKERS

A. Method

1. Stimuli

Target stimuli were 320 nonsense sentences spoken by a female talker (Helfer, 1997). These were the same sentences used in experiments described in previous papers (Freyman *et al.*, 1999, 2001). The stimuli had originally been recorded on digital audiotape. The tape recorder's analog output was low-pass filtered at 8.5 kHz and digitally sampled at 20 kHz using a 16-bit A/D converter (TDT AD1). The stimuli were divided into 16 lists of 20 utterances each. Each sentence was semantically incongruous while being syntactically correct, e.g., "The moon could play your love," and contained three key words which were underlined as above for scoring purposes. Percent-correct scores were derived from the number of underlined words correctly identified by the listener.

Four multitalker speech maskers were used: three-talker, four-talker, six-talker, and ten-talker. All maskers were created using the speech of young adult female talkers. Each talker recorded a series of nonsense sentences that was different for each talker and from the 320 target stimuli. The recordings were transferred to a computer (Dell Optiplex GX1p) using a sampling rate of 22.05 kHz. Each talker's recording of discrete nonsense sentences was edited to create an uninterrupted, continuous 35-s-long stream for each talker. The rms outputs of the individual speech streams were equated with one another and then added to build the multitalker maskers as follows: the three-talker masker was created by adding a third talker's speech to the original two talkers (SS and TK) used in Freyman *et al.* (2001), the four-talker masker consisted of the three-talker masker with one more speech stream added to it, and so on. Figure 1 shows 10-s segments of two-, three-, four-, six-, and ten-talker maskers along with the single female talker target for comparison. Note that as the number of talkers in the interference increases, the waveform becomes denser and smoother with a filling in of the peaks and valleys characteristic of the single- and two-talker maskers. Figure 2 displays the long-term one-third-octave spectra of the target and maskers. For ease of viewing, the target was shifted by 20 dB.

2. Apparatus

The experiments were conducted in the same anechoic chamber used for previous experiments (Freyman *et al.* 1999, 2001). It measured $4.9\times4.1\times3.12$ m. The walls, floor, and ceiling are lined with 0.72-m foam wedges. The subject was seated in the center of the room in front of a foam-

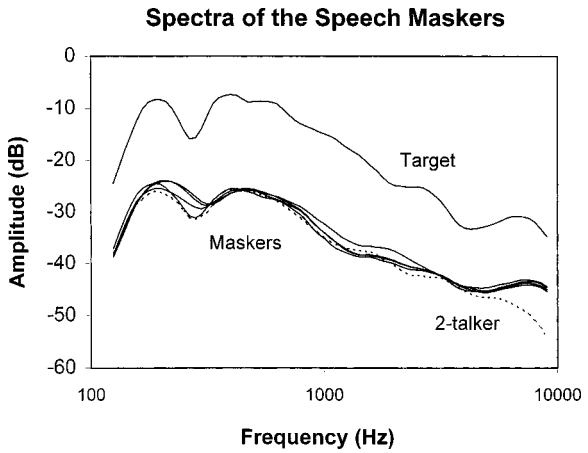

FIG. 2. Long-term one-third-octave spectra of the target and maskers. The target is offset by +20 dB for ease of viewing. The other solid lines represent the spectra of the three-, four-, six-, and ten-talker maskers. The dashed line indicates the spectrum of the two-talker masker.

covered semicircular arc on which two speakers were located. The front loudspeaker was at 0 deg horizontal azimuth and the right speaker was positioned at 60 deg azimuth to the right of the listener. Both were 1.9 m distant from the approximate center of the subjects' heads and at ear height for the typical adult.

The target sentences were delivered via TDT System I instrumentation. The output of the 16-bit D/A converter (TDT DA1) running at 20 kHz was low-pass filtered at 8.5 kHz (TDT), attenuated (TDT PA3), and mixed with the masker before being delivered through a Crown D40 amplifier to a Realistic Minimus 7 loudspeaker. The masker was delivered from the sound board of a Dell Dimension Pentium computer and fed to a delayer (Klark Teknik DN716), one output of which was delayed by 4 ms relative to the other. The delayed output was mixed with the target (TDT SUM3) prior to power amplification and was delivered to the front loudspeaker. The undelayed output was delivered to the right loudspeaker, but was switched off in the conditions in which only the front loudspeaker was to be used. Calibration of presentation level was by means of a 1-in. microphone (B&K 4145) fitted with a random incidence corrector and lowered to the position of the subject's head with the subject absent. A sound-level meter (B&K 2204) located outside the chamber measured the microphone output using the A scale and fast meter response. Small differences measured in the level of the target across the 320 sentences were minimized during the experiment using the TDT PA3 attenuator.

3. Procedures

Two target–masker configurations were used as before: F–F, where the target and masker were presented from the front loudspeaker, and F–RF, where the target was delivered from the front loudspeaker and the masker was delivered from both front and right loudspeakers with a 4-ms time lead to the right. Each of the four maskers was presented at four signal-to-noise (S–N) ratios. The initial data were obtained for the four-talker masker at S–N ratios of −12, −8, −4, and 0 dB. Because of poor performance observed at −12 dB, the S–N ratios for subsequent masker conditions were changed to −8, −4, −2, and 0 dB. The four S–N ratios at two loudspeaker configurations (F–F and F–RF) produced a total of eight conditions per masker.

Specification of target level was based on the median of a large sample of peak needle readings on the sound-level meter. The rms of the maskers (after combining the talkers together) was equated to the rms of a sawtooth wave ($F_0 = 100$ Hz), which was presented daily for calibration. S–N ratios were specified as the difference between the target level of 46 dBA and the measured dBA of the sawtooth wave in the F–F condition. No corrections were made for additional masker energy occurring in the F–RF condition. Across listening blocks of 20 sentences, the desired signal-to-noise ratio was fixed and achieved by varying the level of presentation of the masker while the target level was maintained at 46 dBA.

The subjects were normal-hearing young adults with pure-tone thresholds $\leqslant$20 dB HL in the frequencies 0.5, 1.0, 2.0, 3.0, 4.0, and 6.0 kHz (ANSI, 1996). A different group of eight subjects listened to each masking condition. As in Freyman *et al.* (1999, 2001), a completely within-subjects Latin square design was used for each of the maskers to minimize the potential interaction of subject and sentence list differences. Because there were eight listening conditions for each masker, and 16 lists were available, two consecutive lists were used per condition per listener. Thus, the percentage of key words perceived correctly across subjects for each condition was based on 960 scored items (16 lists $\times$20 sentences$\times$3 key words).

The listener initiated each trial with a button press. The masker was gated on first, with the target sentence following between 0.6 to 1.2 s later, the brief delay in target onset providing a basis for attending to the target. Because the masker was played continuously, its onset during a trial could occur at any point in the continuous speech stream while the target always began with the first word of a nonsense sentence. The target and masker terminated simultaneously. The listener was instructed to repeat the target sentence to the best of his or her ability. While no physical restraints were placed on the listeners, they were advised to maintain a head position facing the front speaker.

Subjects completed the entire listening session in about 1 h, with a break provided halfway through. Prior to listening to the experimental stimuli, subjects listened to five practice sentences to familiarize them with the target speaker's voice. These five sentences were repeated in selected signal-to-noise and speaker conditions to instruct the subject on the task and conditions of the experiment.

B. Results

The basic result of this study is that the improvement in performance in the F–RF condition relative to the F–F condition decreased as the number of masking talkers increased from three to four to six to ten. In the data plotted in panels (b)–(e) of Fig. 3, the narrowing of the difference between the two conditions as the number of talkers is increased is evident. For comparison, the results from the two-talker masker (SS+TK) from experiment 2 of the current paper is displayed in panel (a).[1] There was a considerable narrowing of the F–RF versus F–F difference between the two- and three-talker maskers, and further narrowing as the number of talkers increased to ten.

Signal-to-noise ratios required for a criterion performance of 50% correct were estimated through interpolation of the functions in Fig. 3. Figure 4 [panel (A)] shows the differences in S–N ratio for criterion performance between the F–RF and F–F conditions. In addition to the four maskers studied in the current experiment and the two-talker masker added from experiment 2, the figure also displays single-talker data from Freyman *et al.* (1999, 2001). In Freyman *et al.* (1999), TK was used as a masker, whereas SS was the single-talker masker in Freyman *et al.* (2001). The current figure shows the average of the F–RF versus F–F difference for those two individual talkers. The criterion performance used in these computations for single-talker masking was 60% correct, as subjects never scored as low as 50% at any of the tested S–N ratios. The results show that, among these six conditions, the two-talker masker was associated

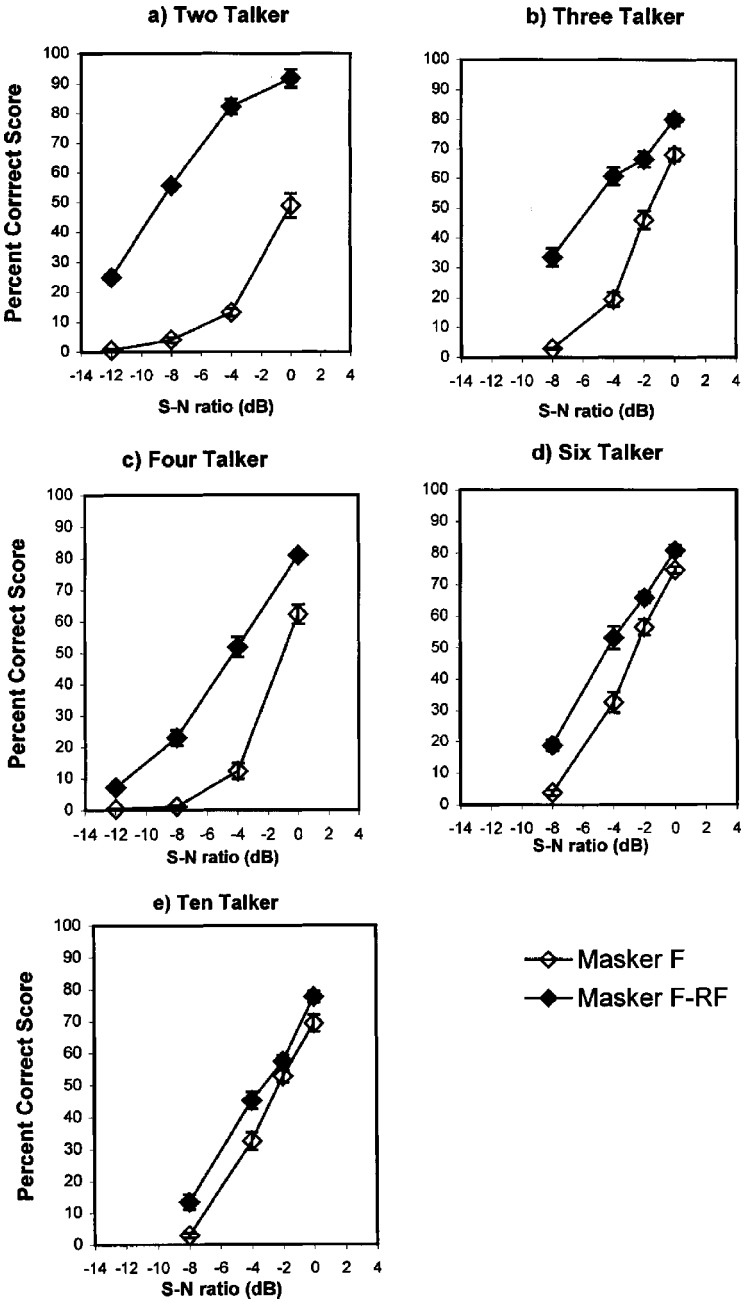

FIG. 3. Group mean-percent correct scores for key words within the target sentences as a function of S–N ratio for the F–F and F–RF conditions. Error bars represent ±1 standard error. Each panel shows the results for a different masker. For comparison, the two-talker masker data obtained in experiment 2 are also included.

with the largest improvement in the F–RF condition relative to the F–F condition. The F–RF benefit was reduced considerably, but did not go to zero, as the number was increased to ten masking talkers. The sources of the narrowing difference are revealed in Fig. 4, panel (B), which shows the actual S–N ratios for criterion performance for the F–RF and F–F conditions individually. The lower line displays the S–N ratio for 50% correct performance for the F–RF condition for the two-, three-, four-, six-, and ten-talker maskers. (The S–N ratio for the single-talker maskers could not be included because specification of S–N ratio was different in the earlier studies and, as described above, performance always was above 50% correct.) The upper line displays the criterion S–N ratios for the F–F condition for these five maskers. This figure shows that the S–N ratio for criterion performance in the F–RF condition, which was presumably largely due to energetic masking, increased as number of masking talkers increased, while the effect was the opposite in the F–F condition. This suggests that the increase in energetic masking over that range (assumed to be the same in F–F and F–RF) was more than offset in the F–F condition by a substantial decrease in informational masking.

III. EXPERIMENT 2: PRIMING BY TARGET TALKER

The first experiment showed that the two-talker masker was most effective in creating informational masking and that perceived differences in spatial location were useful in overcoming this masking, presumably because it facilitated listeners' focused attention on the target. In the current experiment, we explored an alternative means of increasing listeners' ability to identify and focus attention on the target. This experiment investigated the effect of ''priming'' or cuing the listener to the nonsense sentence associated with a

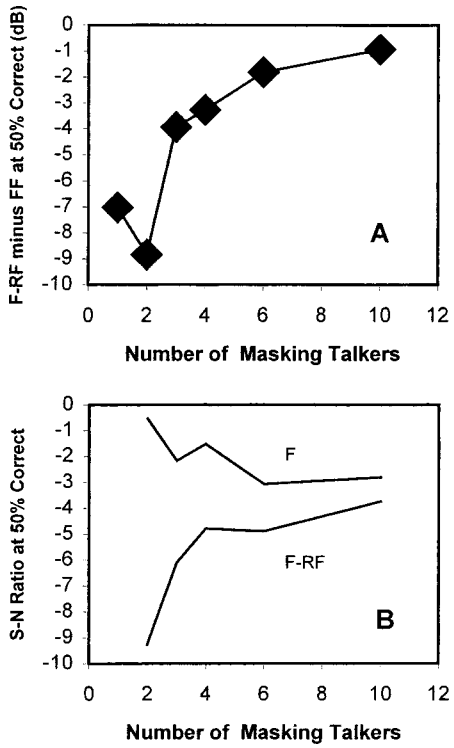

FIG. 4. (A) Difference between the F–F and F–RF conditions for criterion (50%-correct) performance as a function of number of masking talkers. (B) Actual criterion S–N ratios for the F–F and F–RF conditions as a function of number of masking talkers.

target word. In the control condition, the listening task was similar to that described for experiment 1. That is, the listeners heard and repeated back the target nonsense sentences in the presence of masking noise or speech at designated S–N ratios. The comparison, priming condition was identical to the control except that the listener heard the target sentence in quiet just before it was presented in masking. In this preview the last of the three key words in the target sentence was omitted and replaced by noise. Our hypothesis was that hearing the prime would improve listeners' ability to identify and attend to the target utterance when it was presented in a two-talker masker, and therefore improve recognition of the last key word, even though it was not heard during the prime. In continuous noise masking, simple audibility of the target, not attention, is assumed to be the most important factor. Because the last key word could not be predicted from the preceding words in each nonsense sentence, it was hypothesized that the prime would provide no advantage for continuous noise masking.

A. Method

The target sentences were the same set of 320 stimuli used in experiment 1. The priming utterances that preceded these sentences were identical to the target sentences except that the final key word of each utterance was replaced by a noise segment. The noise segment was produced by creating a white-noise token whose duration (700 ms) matched that of the longest third key word segment across all target utterances. The noise was scaled to an rms of approximately 10 dB below the rms of the target speech and appended to the end of each sentence, whose last word had been removed through waveform editing. Figure 5 displays an example target utterance and the corresponding prime utterance.

Two maskers were used: the two-female talker (SSTK) masker used in Freyman *et al.* (2001) and a Gaussian noise whose spectral shape was modeled after filter characteristics described for female speakers of midwestern (Standard) American English (Byrne *et al.*, 1994). Loudspeaker locations and calibration of targets and maskers were as described previously in experiment 1. In the priming condition, an individual trial consisted of the priming utterance (the one with the noise segment at its end) presented in quiet first and followed, after a button press, by the complete target utterance presented against the background masker. Both priming

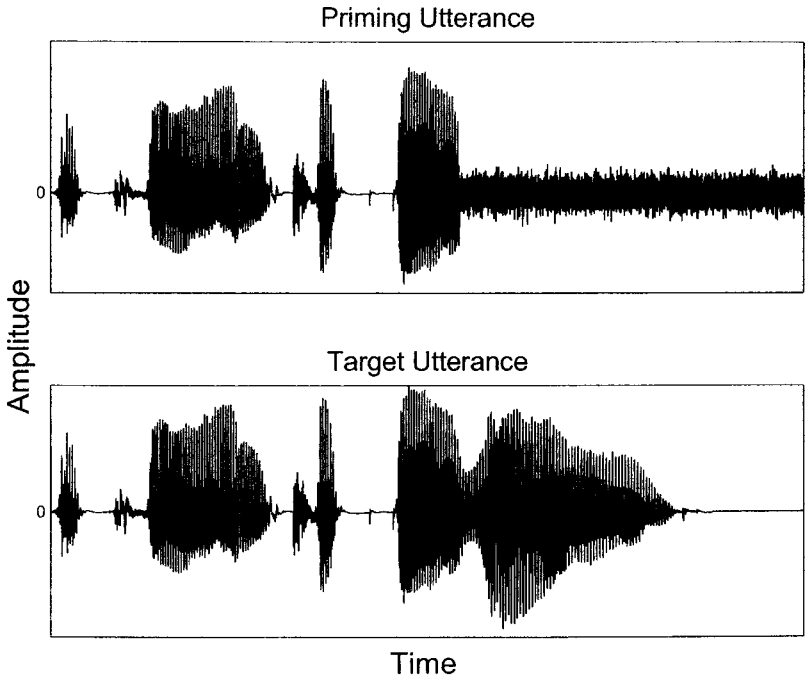

FIG. 5. Time-domain waveforms of the nonsense utterance ''A corn took their wire'' in the priming and target conditions for experiment 2. In the priming condition, the word ''wire'' was replaced by a 700-ms-long segment of white noise.

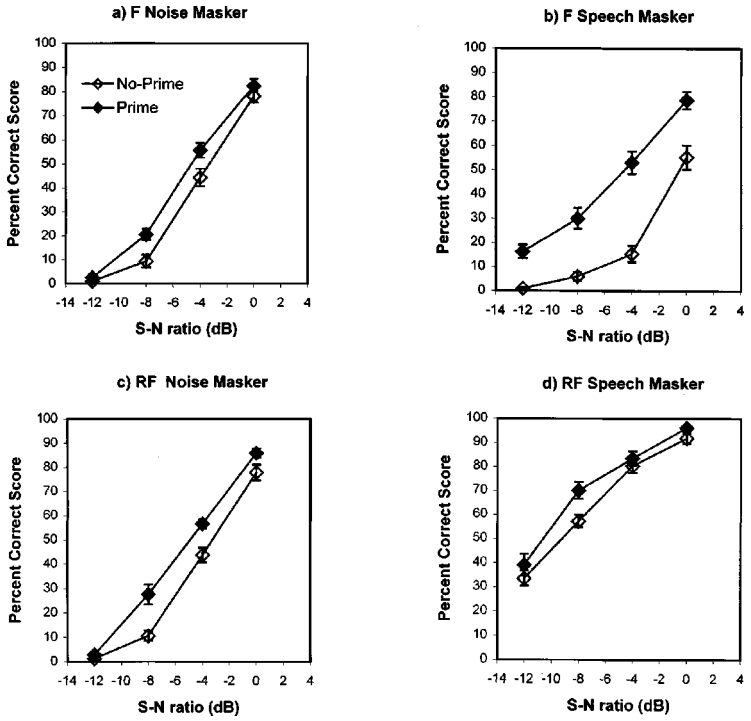

FIG. 6. Effect of priming and spatial separation for a speech-shaped noise masker and two-talker speech masker. Group mean data are shown along with ±1 standard error. Data for the F–F condition were obtained from a different group of subjects than those used for the F–RF condition.

and target utterances were presented at 46 dBA. The listener's task as always was to repeat the entire target sentence, although only the final key word was scored. The no-prime condition was the same except the prime was not presented. Two different groups of 16 subjects participated, one of which listened only to the F–F conditions and the other to the F–RF conditions only. Within the F–F or F–RF configuration, each subject listened to 16 conditions (2 maskers ×2 priming conditions×4 S–N ratios) using a Latin square design in which each condition and subject was assigned all 16 lists, and the condition to list assignment was never repeated across subjects.

B. Results

Figure 6 displays the results for the F–F and F–RF conditions for the noise and speech maskers. Data shown represent the overall percentage-correct score for the last key word in each sentence. Thus, each data point is based on 320 responses (20 items per list×16 subjects/lists). The two functions within each figure, as well as the two figures in the same row, reflect data from one subject group and thus are within-subject comparisons. Comparisons of data across the two rows are between subjects. The results indicate that performance improved with the addition of the priming utterance for every condition. However, the improvement was greatest by far in the F–F two-talker condition where, at the −4-dB S–N ratio, scores improved from 15% correct in the no-prime condition to 53% correct in the prime condition. The improvements displayed in the other three panels at the −4-dB S–N ratio ranged from 3 to 13 percentage points.

As in experiment 1, interpolation of the functions was conducted to derive the S–N ratios at which 50%-correct performance was achieved. Table I displays these derived S–N ratios. Table II displays differences in these S–N ratios for the two maskers to highlight the effect of the priming and F–RF conditions. For the noise masker in both F–F and F–RF speaker configurations and for the two-talker masker in the F–RF speaker configuration, the addition of the prime reduced the criterion S–N ratio by a similar small degree (1.3 to 1.6 dB). The fact that the F–RF speech masker produced priming benefits similar to those obtained with the noise masker suggests that the F–RF two-talker competition, like the noise condition, produced purely energetic masking. That is, it appears that informational masking in the two-talker case was effectively eliminated by spatial separation. It is not clear why performance improved at all in these conditions, as the last key word was never heard in the prime and no semantic context was inherent in the nonsense sentences. It is possible that some phonetic context was provided for the target word by spectral transitions related to coarticulation at the end of the word preceding the (removed) last key word. Another possibility is that the prime decreased the memory load required for the first two words and allowed more resources to be brought to processing of the final word. For the speech masker in the F–F condition, improvement due to the prime was considerably larger, equivalent to an approximately 4-dB reduction in S–N ratio at 50% correct [see also panel (b), Fig. 6]. Here, we assume that informational masking was partially released. Although it gave no direct information about the key word, the priming sentence may have helped the listener to extract the target auditory ''object'' out of the mixture of three talkers. Once the object was extracted and attended to, the last key word was more

TABLE I. S–N ratios (dB) derived by interpolation for 50%-correct response for experiment 2.

Masker	F–F no prime	F–F prime	F–RF no prime	F–RF prime
Noise	−3.33	−4.64	−3.27	−4.90
Two talker	−0.53	−4.54	−9.21	−10.59

TABLE II. Advantage of priming and of spatial separation in dB for the noise and speech maskers in experiment 2. These values were derived from the S–N ratios reported in Table I. The benefit of priming can be seen for the different spatial conditions in the left half of the table. The right half of the table shows the benefit of spatial separation for the no-prime and prime conditions.

Benefit of Priming (dB)			Benefit of F–RF versus F–F (dB)		
	F–F	F–RF	No Prime	Prime	
Noise	1.31	1.63	−0.06	0.26	Noise
Speech	4.01	1.38	8.68	6.1	Speech

understandable because it was connected to that object.

A comparison of the benefit of priming versus spatial separation in Table II suggests that perceived spatial separation was more effective in release from informational masking than was priming. The 4-dB improvement due to priming for the speech masker was not as large as the 8.7-dB improvement obtained in the F–RF condition. Further, in the F–RF condition, the effect of priming was small (1.4 dB), while in the priming condition, the effect of spatial separation was substantial (6.1 dB). This implies a considerable additional release from informational masking in a condition where there was already some release due to priming.

IV. EXPERIMENT 3: COMPARISON OF TYPES OF PRIMING STIMULI

The fact that priming resulted in release from informational masking led to the question of what features of the priming utterance were important for cuing the listener. Because both the priming and target sentences were spoken by the same person, it was possible that one salient cue was the voice and delivery characteristics of the speaker. On the other hand, it could also have been the case that listeners were helped by the priming utterance because they were able to attend to the specific words that had just been presented in the prime. In the next experiment, we varied the priming stimulus to try to distinguish between these possibilities. Because the effect of priming was more robust in the F–F condition, only the F–F condition was used for this experiment.

A. Method

Three priming conditions were used. The first was the same as in the previous experiment; that is, the target speaker's utterance was used as the priming and test utterance (the ''target-talker'' condition). The second condition consisted of the same priming sentences recorded by a young adult male talker (''male talker''). The processing of this prime was identical to that of the target-talker prime, except that the noise segment substituted for the last key word was slightly longer, at 715 ms, in order to match the longest last key word within his recordings of the 320 sentences. The third priming condition (''reading'') consisted of the priming sentences presented in print form with the last word omitted. At the start of each reading prime block, the subject was provided with a set of 20 utterances typed out on index cards with blank cards following each utterance card. The subject was instructed to read the priming utterance, turn that card over to reveal a blank card, and then press a button to listen to the complete target sentence presented with the masker. For comparison, a fourth, no-prime condition was also included.

Only the two-talker speech masker from the previous experiment was used. Hence, there was a total of 16 conditions (4 priming conditions$\times$1 masker$\times$4 S–N ratios). A new group of 16 young normal-hearing subjects was presented with the conditions in a Latin square design as described previously. Signals and maskers were calibrated and presented as described earlier.

B. Results

Figure 7 displays the mean percent-correct scores for the three priming conditions and the no-prime condition. It is apparent that the availability of all three priming conditions improved performance relative to the no-prime condition by approximately the same amount. The two dashed lines replot the results for the target prime and no-prime F–F only conditions from experiment 2, which were obtained with a different group of listeners. These conditions were identical to the target-talker condition of the present study. As can be seen, the effects of priming are consistent across subject groups. See Table III.

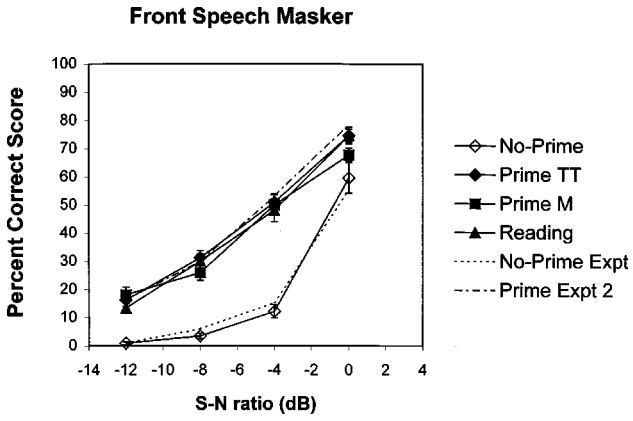

FIG. 7. Comparison of group mean-percent correct scores (with $\pm$1 standard error) for different priming conditions used in experiment 3. The control condition was the ''no-prime'' condition, in which the subjects received no priming utterance (open diamonds). ''Prime TT'' (filled diamonds) refers to the condition where the target talker produced the priming utterance. ''Prime M'' (filled squares) is the condition where the priming utterance was produced by a male talker. ''Reading'' (filled triangles) refers to the priming utterance being presented in print. Dashed lines show the primed and unprimed percent-correct scores obtained in experiment 2.

TABLE III. S–N ratio (dB) derived by interpolation for 50%-correct response for experiment 3. Values obtained for the same conditions in experiment 2 are displayed again to show consistency across subject groups.

	Target talker	Male talker	Reading	No prime
Expt. 3	−4.19	−4.00	−3.72	−0.82
Expt. 2	−4.54	···	···	−0.53

V. DISCUSSION

The first experiment, when considered together with the earlier results presented by Freyman *et al.* (1999, 2001), showed that the release from masking created by the F–RF condition increased when the number of masking talkers was increased from one to two, but decreased as the number was increased further. Performance in the F–F and F–RF conditions followed essentially opposite patterns as the number of talkers increased from two to ten, resulting in a narrowing of the difference between the two conditions. The F–F condition became easier with increasing number of talkers, while the F–RF condition became more difficult. The effect of number of talkers on energetic masking is assumed to be equivalent in the two conditions. Therefore, the different pattern seen in the F–F condition presumably reflects the rise, then fall, of informational masking as the number of masking talkers is increased (see Fig. 4). The overall shape of the function is reminiscent of the nonmonotonic effect of number of masking components on detection of a 1000-Hz signal within a multicomponent informational masker (Oh and Lutfi, 1998). The specific peak observed here for the two-talker masking may be unique to the current stimuli and conditions. However, as noted in the Introduction, the literature includes other examples in which two masking talkers caused significant difficulty for the listener (e.g., Hall *et al.*, 2002; Yost *et al.*, 1996; Carhart *et al.*, 1975). One difference is that Carhart *et al.* (1975) found that three masking talkers produced more perceptual masking than two, whereas with the current stimuli three masking talkers produced less informational masking. These differences can presumably be attributed to the specifics of the stimuli and task.

The second experiment showed that hearing a preview of the target before it was presented in a two-talker masker improved speech recognition performance for a tested key word that was not included in the preview. At some S–N ratios, the subjective impression is that the target message stands out much more clearly in the speech background in trials preceded by the priming utterance. Once the target utterance was perceptually pulled out of the mixture of voices, the listener was often able to follow the message well enough to correctly perceive the unprimed last key word. This idea of ''latching on'' to a talker's voice is consistent with Brungart's (2001) finding that, once listeners decide which of two voices is the target (correctly or incorrectly), they persist in responding with that talker's words. The third experiment demonstrated that the priming utterance need not be the exact target waveform. In fact, it is sufficient for the prime to be spoken by an entirely different talker, or even read by the subject from a printed page. The prime gives the subject information about what to listen for, and this knowledge makes it much easier to attend to the target words and ignore the jumbled utterances of the other two talkers.

These results are consistent with the idea that there were sufficient cues in the target plus two-talker waveform for some level of segregation of the target, even in the nonspatial condition. These cues would include those traditionally thought to be important for auditory grouping, such as temporal asynchrony and differences in fundamental frequency between the target and masking messages (Darwin and Carlyon, 1995), in addition to other differences between target and masking speech (e.g., Cherry, 1953; Darwin and Hukin, 2000; Darwin *et al.*, 2003). Despite the fact that the cues are available, the three mixed utterances are difficult to sort out in the brief observation interval. The voices of the masking talkers compete for attention with the target talker. They are in many of the tested conditions at least as loud as the target, come from the same location as the target, and are perceptually similar to the target along some dimensions in that both target and maskers were produced by adult females. This last issue of similarity between target and masker was studied explicitly by Kidd *et al.* (2002) for the identification of nonspeech auditory patterns. They studied the effectiveness of informational maskers that varied in the extent to which they were likely to perceived as auditory streams that could be confused with the masker. Kidd *et al.* (2002) concluded that it was not possible to distinguish between explanations that relied on similarity between target and masker and those that depended on the allocation of attentional resources to maskers that formed their own perceptual streams. Likewise, in the current experiments, the fact that target and maskers are all produced by female talkers could contribute to the difficulty of the task in more than one way. The speech of the masker female talkers may be easily confused with the target. Additionally, and partially because of the similarity, the masking speech attracts the listener's attention.

The nonmonotonic effect of the number of masking talkers on recognition of the target in the F–F condition (experiment 1) also can be considered in terms of similarity and attention. As the number of masker talkers increases and eventually becomes a general babble, the similarity of the masker and target decreases. The target stands out from this background as long as the S–N ratio is sufficient for audibility, i.e., energetic masking is the limiting factor. The attentional demands imposed by the masker might be expected to follow a nonmonotonic course, similar at least qualitatively to the data. Two masking talkers would be expected to necessitate more attentional resources than one masking talker; however, as the number of talkers is increased much further and the individual utterances are less well recognized, the competition for attention is likely to decrease.

Another factor that may influence auditory attention is the relative loudness of competing utterances. It is reasonable to assume that a listener would be more likely to attend to a louder voice. Evidence for this comes from the single-talker masker data of Brungart *et al.* (2001, Fig. 1, top panel). In their conditions in which the masking has been shown to be almost entirely informational, it might be assumed that loudness difference between target and masker in either direction might become a cue for following the target

message. Under this assumption, 0-dB S–N ratio should be most difficult, with improvements occurring when the masker is either louder or softer than the target. However, their data show that performance is relatively unchanged over a region from −12- to 0-dB S–N ratio, then improves sharply as S–N ratio is increased above 0 dB. This suggests that listeners' attention is drawn to the louder of the two messages, which may supersede any benefit that might have resulted from a simple difference in level.

The current data suggest that the relative loudness of the target voice to the individual masking voices might also be important when there is a small number of masking talkers. For example, in Fig. 3(a) (two-talker), performance in the F–F condition is extremely poor until the 0-dB S–N ratio condition is reached. At that S–N ratio, the target is 3 dB higher in level than either of the masking talkers, whereas at all other S–N ratios, the target is about the same as (−1 dB at the −4-dB S–N ratio condition) or below the level of any individual masking voice. As the number of masking talkers increases and the masker is perceived more as a complex babble than individual voices, the loudness difference between the target and any one voice is not likely to be as important.

The current experiments employed conditions that improved the listener's ability to identify and focus attention on the target talker and not on the masking utterances. In the case of the priming conditions, attention to the target stream is made easier because the subject has heard or seen what sentence to listen for. In the case of the F–RF condition, directional information preserved by the precedence effect allows the already-segregated speech streams to be distinctly localized, and this makes it much easier to attend to, and correctly perceive, the target. There is no suggestion in the data that the precedence effect actually creates the initial segregation. This view is consistent with data showing that in general, basic cues for localization, such as interaural time delay (ITD), are not strong cues for auditory segregation. For example, in the ''double-vowel'' experiments reported by Culling and Summerfield (1995), ITD was not sufficient to segregate vowels when it was the only cue. Rather, interaural differences appear to be important for lateralization of signal components that have been segregated by other means (Hill and Darwin, 1996), and may assist in connecting segregated signals across time (Darwin and Hukin, 1999, 2000). The data from experiment 2 show that the F–RF condition produces greater benefit than the prime condition, and creates additional advantages for recognizing sentences that have already received benefit from priming. Thus, although localization cues are considered to be weak cues for sound-source segregation, localization is extremely useful in the task of selectively attending to one message while ignoring others.

In conclusion, the current results suggest that informational masking is most likely to be observed when one must attend to the speech of one person in the presence of one or two nearby conversations. Conditions that allow the listener to better attend to the target will help overcome this type of masking. Knowing most of what the target talker is going to say ahead of time partially releases informational masking, as evidenced by the enhanced recognition of key words that were omitted from a preview of the target sentences. This enhancement is assumed to be due to an improved ability to identify the target message, requiring fewer attentional resources to be devoted to the maskers. As being exposed to a preview of even a subset of the words to be spoken is unrealistic, future work concerned with finding solutions for overcoming informational masking should consider whether simply knowing the topic provides some benefit.

Informational masking appears to be substantially released by conditions that create a perceived difference in horizontal location between target and interfering speech. This type of release from masking may be unavailable to persons who must listen under conditions in which spatial hearing cannot be well exploited. These would include individuals wearing earmuff hearing protection, people who have bilateral hearing losses but are wearing monaural hearing aids, binaurally fitted hearing aid users who have poor ability to localize sound, and most cochlear implant users, who are generally implanted in one ear. For these situations and individuals, alternative methods will be necessary to achieve target/masker distinctions that facilitate focused and sustained attention on the target message.

ACKNOWLEDGMENTS

This research was supported by a grant from the National Institute for Deafness and other Communicative Disorders (DC01625). The authors would like to thank Cara Caminiti, Hilary Brown, Joni Skinner, and Wendy Levesque for their assistance in data collection for these experiments.

[1]The priming condition in experiment 2 used a two-talker masker but only the last of the three key words was scored and plotted. For comparisons within experiment 2, the control (no-prime) condition was also scored in the same way. However, for the purpose of comparing the two-talker data from experiment 2 with the other data from experiment 1, all three key words in the no-prime two-talker masker condition were scored for each sentence, exactly as in experiment 1. The primary difference between the collection of the two-talker data plotted in Fig. 1(a) and the data plotted in the other panels was that the data in panel (a) were obtained with two different groups of 16 subjects each (one for F–F and one for F–RF). The data for each of the other panels were obtained within different groups of eight subjects. There is considerable confidence in between-group consistency for identical conditions with these stimuli (see Fig. 7 of the current paper). Further, a within-group comparison for this two-talker masker has already been completed [Fig. 3(B) from Freyman *et al.*, 2001] with similar results, although a difference in the specification of S–N ratio makes a direct comparison with the current data difficult.

ANSI (**1996**). ANSI S3.6-1996, ''Specifications for audiometers'' (American National Standards Institute, New York).

Arbogast, T. L., and Kidd, Jr., G. (**2001**). ''Evidence for spatial tuning in informational masking using the probe-signal method,'' J. Acoust. Soc. Am. **108**, 1803–1810.

Arbogast, T. L., Mason, C. R., and Kidd, Jr., G. (**2002**). ''The effect of spatial separation on informational and energetic masking of speech,'' J. Acoust. Soc. Am. **112**, 2086–2098.

Broadbent, D. E. (**1952**). ''Listening to one of two synchronous messages,'' J. Exp. Psychol. **44**, 51–55.

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

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

Brungart, D., and Simpson, B. (**2002**). ''The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal,'' J. Acoust. Soc. Am. **112**, 664–676.
Brungart, D. 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.
Byrne, D., Dillon, H., Tran, K., Arlinger, S., Wilbraham, K., Cox, R., Hagerman, B., Hetu, R., Kei, J., Lui, C., Kiessling, J., Nasser Kotby M., Nasser, N. H. A., El Kholy, W. A. H., Nakanishi, Y., Oyer, H., Powell, R., Stephens, D., Meridith, R., Sirimanna, T., Tavarkiladze, G., Frolenkovgi, G. I., Westerman, S., and Ludvigsen, C. (**1994**). ''An international comparison of long-term average speech spectra,'' J. Acoust. Soc. Am. **96**, 2108–2120.
Carhart, R., Johnson, C., and Goodman, J. (**1975**). ''Perceptual masking of spondees by combinations of talkers,'' J. Acoust. Soc. Am. **58**, S35.
Carhart, R., Tillman, T. W., and Greetis, E. S. (**1969**). ''Perceptual masking in multiple sound backgrounds,'' J. Acoust. Soc. Am. **45**, 694–703.
Cherry, E. C. (**1953**). ''Some experiments on the recognition of speech, with one and two ears,'' J. Acoust. Soc. Am. **25**, 975–979.
Culling, J. F., and Summerfield, Q. (**1995**). ''Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay,'' J. Acoust. Soc. Am. **98**, 785–797.
Darwin, C. J., and Carlyon, R. P. (**1995**). ''Auditory grouping,'' in *The Handbook of Perception and Cognition (Hearing)*, edited by B. C. J. Moore (Academic, London).
Darwin, C. J., and Hukin, R. W. (**1999**). ''Auditory objects of attention: The role of interaural time differences,'' J. Exp. Psychol. **25**, 617–629.
Darwin, C. J., and Hukin, R. W. (**2000**). ''Effectiveness of spatial cues, prosody, and talker characteristics in selective attention,'' J. Acoust. Soc. Am. **107**, 970–977.
Darwin, C. J., Brungart, D. S., and Simpson, B. D. (**2003**). ''Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers,'' J. Acoust. Soc. Am. **114**, 2913–2922.
Duquesnoy, A. J. (**1983**). ''Effect of a single interfering noise and interfering speech upon the binaural sentence intelligibility of aged persons,'' J. Acoust. Soc. Am. **74**, 739–743.
Freyman, R. L., Helfer, K. S., McCall, D. D., and Clifton, R. K. (**1999**). ''The role of perceived spatial separation in the unmasking of speech,'' J. Acoust. Soc. Am. **106**, 3578–3588.
Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2001**). ''Spatial release from informational masking in speech recognition,'' J. Acoust. Soc. Am. **109**, 2112–2122.
Hall, J. W., Grose, J. H., Buss, E., and Dev, M. B. (**2002**). ''Spondee recognition in a two-talker masker and a speech-shaped noise masker in adults and children,'' Ear Hear. **23**, 159–165.
Hawley, M. L., Litovsky, R. Y., and Colburn, H. S. (**1999**). ''Intelligibility and localization of speech signals in a multisource environment,'' J. Acoust. Soc. Am. **105**, 3436–3448.
Helfer, K. S. (**1997**). ''Auditory and auditory-visual perception of clear and conversational speech,'' J. Speech Lang. Hear. Res. **40**, 432–443.
Hill, N. I., and Darwin, C. J. (**1996**). ''Lateralization of a perturbed harmonic: Effects of onset asynchrony and mistuning,'' J. Acoust. Soc. Am. **100**, 2352–2364.
Kidd, Jr., G., Mason, C. R., Deliwala, P. S., Woods, W. S., and Colburn, H. S. (**1994**). ''Reducing informational masking by sound segregation,'' J. Acoust. Soc. Am. **95**, 3475–3480.
Kidd, Jr., G., 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, Jr., G., Mason, C. R., Rohtla, T. L., and Deliwala, P. S. (**1998**). ''Release from masking due to spatial separation of sources in the identification of nonspeech auditory patterns,'' J. Acoust. Soc. Am. **104**, 422–431.
Noble, W., and Perrett, S. (**2002**). ''Hearing speech against spatially separate competing speech versus competing noise,'' Percept. Psychophys. **64**, 1325–1336.
Oh, E. L., and Lutfi, R. A. (**1998**). ''Nonmonotonicity of informational masking,'' J. Acoust. Soc. Am. **104**, 3489–3499.
Peters, R. W., Moore, B. C. J., and Baer, T. (**1998**). ''Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people,'' J. Acoust. Soc. Am. **92**, 3132–3138.
Yost, W. A., Dye, R. H., and Sheft, S. (**1996**). ''A simulated cocktail party with up to three sound sources,'' Percept. Psychophys. **58**, 1026–1036.