2008

Spatial release from masking with noise-vocoded speech

Richard L. Freyman, *University of Massachusetts - Amherst*
Uma Balakrishnan, *University of Massachusetts - Amherst*
Karen S Helfer, *University of Massachusetts - Amherst*

Spatial release from masking with noise-vocoded speech

Richard L. Freyman,[a] Uma Balakrishnan, and Karen S. Helfer
Department of Communication Disorders, University of Massachusetts, 358 North Pleasant Street, Amherst, Massachusetts 01003

(Received 19 June 2007; revised 4 June 2008; accepted 5 June 2008)

This study investigated how confusability between target and masking utterances affects the masking release achieved through spatial separation. Important distinguishing characteristics between competing voices were removed by processing speech with six-channel envelope vocoding, which simulates some aspects of listening with a cochlear implant. In the first experiment, vocoded target nonsense sentences were presented against two-talker vocoded maskers in conditions that provide different spatial impressions but not reliable cues that lead to traditional release from masking. Surprisingly, no benefit of spatial separation was found. The absence of spatial release was hypothesized to be the result of the highly positive target-to-masker ratios necessary to understand vocoded speech, which may have been sufficient to reduce confusability. In experiment 2, words excised from the vocoded nonsense sentences were presented against the same vocoded two-talker masker in a four-alternative forced-choice detection paradigm where threshold performance was achieved at negative target-to-masker ratios. Here, the spatial release from masking was more than 20 dB. The results suggest the importance of signal-to-noise ratio in the observation of "informational" masking and indicate that careful attention should be paid to this type of masking as implant processing improves and listeners begin to achieve success in poorer listening environments. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2951964]

PACS number(s): 43.66.Dc, 43.66.Qp, 43.66.Pn, 43.66.Ts [AJO] Pages: 1627–1637

I. INTRODUCTION

When a target sound is presented together with a masking sound, similarities between the two sounds can create confusion that interferes with the detection and recognition of the target. Evidence for this type of interference caused by target/masker confusability is found where increased thresholds or poor discrimination cannot be explained by traditional conceptualizations of masking. The term most commonly applied in these situations is "informational masking," which in recent years has frequently been considered with regard to the masking of speech by competing speech utterances (e.g., Freyman *et al.*, 1999; Brungart *et al.*, 2001; Arbogast *et al.*, 2002; Hawley *et al.*, 2004; Kidd *et al.*, 2005).

Substantial dissimilarity between target and masker is thought to minimize or eliminate informational masking (Durlach *et al.*, 2003). For example, little evidence of informational masking is found with target and masker speech spoken by talkers of the opposite sex (Brungart *et al.*, 2001; Brungart and Simpson, 2002), presumably because large differences in fundamental frequency between male and female voices minimize confusion. Further, different spatial impressions caused by target and masker are likely to reduce confusion between them and severely reduce or eliminate informational masking (e.g., Freyman *et al.*, 2001; Gallun *et al.*, 2005; Kidd *et al.*, 2005).

The two experiments reported in this paper used noise-excited vocoded speech to investigate informational masking under conditions in which target-masker similarity is expected to be high. This type of speech processing, in which envelopes in different frequency channels are extracted and used to modulate a noise carrier, has been used to model key aspects of processing by cochlear implants (e.g., Shannon *et al.*, 1995; Dorman *et al.*, 1998; Qin and Oxenham, 2003; Stickney *et al.*, 2004; Poissant *et al.*, 2006). This processing can severely reduce or eliminate pitch and other voice difference cues between different talkers, increasing target/masker confusability. Using noise-vocoded speech without target/masker spatial differences, both Qin and Oxenham (2003) and Stickney *et al.* (2004) demonstrated more masking from a single competing talker than predicted from the pattern of results obtained with unprocessed speech. Both reports suggest that vocoded speech may be especially susceptible to informational masking. Stickney *et al.* (2007) offered evidence of improved target recognition under some conditions when temporal fine-structure cues sufficient to provide fundamental frequency information were added to noise-vocoded speech.

Paradoxically, increased susceptibility to informational masking with vocoded speech could be partially mitigated by the difficulty listeners have in understanding speech subjected to this kind of processing. In order to achieve reasonable levels of performance, it is sometimes necessary to present the stimuli at high signal-to-noise (SN) ratios of 10 dB or more (e.g., Qin and Oxenham, 2003; Poissant *et al.*, 2006). At these SN ratios, the target is much louder than the masker. Although partial masking may obscure lower level portions of the waveform and may lead to reduced intelligibility, at these positive SN ratios there should be no overall confusion between target and masker. In fact, Arbogast *et al.* (2005) hypothesized that informational masking may decline

[a]Author to whom correspondence should be addressed. Tel.: 413-545-0298. Electronic mail: rlf@comdis.umass.edu

dramatically above 0 dB SN ratio. However, even though a vocoded target at high SN ratios may stand out because it is louder than interfering speech, the lower level portions of the speech-envelope-modulated noise belonging to the target may be confused with the modulations in the masker (Qin and Oxenham, 2003).

The goal of the present study was to understand the influence of SN ratio on informational masking with vocoded speech. The first experiment examined the recognition of open set nonsense sentences in a background of two-talker masking, where positive SN ratios were required for above-floor performance. The second experiment used the same masking stimuli and a subset of the target stimuli, but there the task was only to *detect* the presence of target stimuli in an adaptive forced-choice task. Negative SN ratios were sufficient for threshold performance in this task. Thus, across the two experiments, identically processed target and masking stimuli were employed at very different SN ratios, giving us insight into how informational masking was influenced by SN ratio. In both experiments, informational masking was quantified by measuring spatial release from masking under two-source speech masking conditions that produce no release from continuous noise masking.

II. EXPERIMENT 1: OPEN-SET NONSENSE SENTENCE RECOGNITION

A. Methods

1. Stimuli

The target stimuli were a set of 320 "nonsense" sentences that were syntactically but not semantically correct, e.g., "A *shop* can *frame* a dog." Each sentence included three key words, as underlined in the example. These sentences, recorded by a female talker, have been used in several earlier studies (e.g., Helfer, 1997; Freyman *et al.*, 1999, 2001, 2007; Li *et al.*, 2004). A full description of the recording methodology can be found in Helfer (1997) or Freyman *et al.* (1999). The maskers were nonsense sentences recorded by ten different female talkers (different sets for different talkers). Details of the recording methodology for these maskers can be found in Freyman *et al.* (2007). Pauses were removed between sentences, creating an approximately 35 s long stream for each talker. The streams were equated in average power (rms) and then combined to form five two-talker maskers, with the selection of pairings roughly according to average fundamental frequency.

The target and masking stimuli were processed through six-channel vocoding with a noise carrier using the same algorithm as that in Qin and Oxenham (2003). The frequency range from 80 to 6000 Hz was divided into six channels of equal bandwidth according to the equivalent rectangular bandwidth (ERB) scale (Glasberg and Moore, 1990), using digital sixth-order butterworth bandpass filters. Envelopes were extracted from the filter outputs by digitally low-pass filtering rectified signals with a cutoff frequency of the larger of 300 Hz or half the bandwidth, using a second-order butterworth filter. White noise filtered to have the same bandwidth as the filtered signals was multiplied by the appropriate envelope channel in the time domain to create noises that matched the temporal envelopes in each channel. The six modulated noises were summed to create a broadband six-channel speech-envelope-modulated noise for each of the 320 target sentences and the five two-talker masker speech streams. This type of modulated noise has been shown in several studies to be quite intelligible for sentence stimuli with just four channels (Shannon *et al.*, 1995).

2. Environment and apparatus

The experiment was conducted in a large double-walled sound-treated room (IAC No. 1604) measuring 2.76 $\times$2.55 m. Reverberation times measured in this room ranged from 0.12 s at 6.3 and 8.0 kHz to 0.24 s at 125 Hz (Nerbonne *et al.*, 1983). A previous study conducted in this room using the target front and masker right front (F-RF) condition (Helfer and Freyman, 2005) showed the same kinds of spatial release that have been found in an anechoic chamber (e.g., Freyman *et al.*, 2001). The listener sat on a chair placed with its back against one wall of the room. Two loudspeakers (Realistic Minimus 7), at a distance of 1.3 m from the center of the head when seated in a chair and a height of 1.2 m (the approximate height of the ears of a typical listener), delivered the target and masking stimuli. One loudspeaker was placed at 0 deg azimuth, directly in front of the listener; the second loudspeaker was at 60 deg to the right. The target and masking stimuli were mixed digitally at the appropriate SN ratio on a computer before presentation from two channels of the computer's sound board, attenuated (TDT PA4), amplified (TDT HBUF5), power amplified (TOA P75D), and delivered to the loudspeakers.

In the front-front (F-F) condition, target and masker were presented from the front loudspeaker. In the F-RF condition, the target was presented from the front loudspeaker and the masker from both loudspeakers, with the right leading the front by 4 ms. Due to the precedence effect, the F-RF masking configuration creates the perception that the masker is to the right, well separated from the front target. The 4 ms delay version of this F-RF configuration has been shown to create little or no release from masking for speech targets using continuous or fluctuating noise maskers, indicating no energetic masking release (Freyman *et al.*, 1999; Brungart *et al.*, 2005; Rakerd *et al.*, 2006). Further, Helfer and Freyman (2005) demonstrated no release from continuous noise masking for this configuration in the same sound-treated room used for the current studies. It is assumed, therefore, that when masking release occurs for the F-RF configuration in a competing speech task, the release from masking is due to nonenergetic effects, where perceptual spatial differences make it easier to extract the target from the complex mixture of voices.

3. Subjects

Listeners were ten adult students with hearing thresholds $\leq$20 dB HL in the frequency range of 500–4000 Hz (ANSI S3.6, 1996). The ages ranged from 19 to 21 years. None had extensive experience listening to vocoded speech.

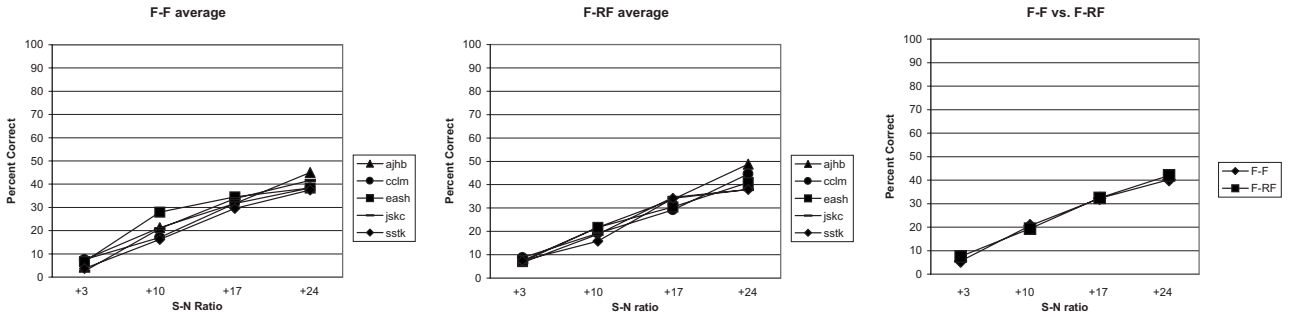

FIG. 1. Percent correct recognition of key words as a function of SN ratio for five two-talker maskers. The left panel shows data for the F-F configuration and the middle panel for the F-RF configuration. The right panel shows F-F and F-RF data combined across all five maskers.

4. Procedures

Subjects were seated in a chair and instructed to face the front loudspeaker but were not physically restrained. At the beginning of each trial, the word "ready" appeared on a computer screen in front of the subject. The presentation of the masker was then initiated, followed by the presentation of the target 0.6–1.2 s later. The masker was terminated simultaneously with the end of the target sentence. The listener repeated the target sentence out loud, and the experimenter, monitoring in a control room, scored the three key words as correct or incorrect.

On each trial a section of two-talker masker waveform was selected randomly from the 35 s stream. The masker onset could occur anywhere in the stream, e.g., at the beginning of a sentence for one of the two talkers and in the middle for the other talker. As noted above, the target sentence began 0.6–1.2 s after the beginning of the masker. Subjects were told that they could use this as a cue for directing attention to the target speech in the presence of the masker.

A total of 40 conditions were presented to each of the listeners: five two-talker maskers, two spatial configurations (F-F and F-RF), and four SN ratios chosen during pilot listening (+3, +10, +17, and +24 dB), presented in a completely crossed design. SN ratio was defined as target rms amplitude relative to masker rms amplitude. SN ratios were manipulated by changing the level of the masker for a fixed-level target, which was always presented at 44 dBA (calibrated at the position of the subject's head using a speech spectrum noise with the same rms as the target sentences). The 320 target sentences were selected at random, without replacement, and with a different random order for each listener. The sentences were presented in 40 blocks of eight sentences each. The SN ratio, masking talkers, and spatial configuration were all fixed within a block. Across sets of four blocks (32 trials), only the SN ratio changed (randomly), while the spatial configuration and masker were fixed. A second set of four blocks followed where the spatial configuration switched from F-F to F-RF or vice versa. Half the subjects received F-F first and the other half F-RF first. After eight blocks were presented (64 trials), the masker changed and the process was repeated until all five maskers had been presented for 64 trials. The order of presentation of the five maskers within a set of 320 trials was random and different for each listener. Subjects completed the experiment in one listening session.

Prior to actual data collection, subjects listened to five practice examples of vocoded nonsense sentences. Each sentence was presented several times in quiet. If the listener did not respond correctly after two or three repetitions, the unprocessed version of the target was presented, and then the vocoded sentence was repeated again until the subject was able to recognize the vocoded sentence.

B. Results and discussion

Results of experiment 1, displayed in Fig. 1, show that even at high SN ratios only moderate levels of performance were obtained. Performance with the same six-channel processing was found to be much better with meaningful sentences (Poissant *et al.*, 2006), so the poorer results probably reflect the use of difficult nonsense sentences where there is no semantic context. The data in the left and middle panels show limited variability in masking efficiency across the five different maskers; in contrast a substantial variation was seen across the same five maskers in their unprocessed version, particularly for the F-F condition (Freyman *et al.*, 2007). For example, at the second highest SN ratio used in that study, performance ranged from 18% correct with one two-talker masker to 71% correct with a different two-talker masker. To be sure, the upper range of performance in the current study may be constrained by the intrinsic difficulty of the stimuli even in the absence of masking. However, the striking homogeneity of performance across masking talkers suggests that some aspect of voice pitch or quality lost in the vocoding process accounted for the variability observed with the unprocessed stimuli in Freyman *et al.* (2007). It suggests little variation in masking produced by other features of the maskers that were preserved in the vocoding process, such as gross temporal patterns. To the extent that vocoding simulates listening with a cochlear implant (e.g., Shannon *et al.*, 1995; Dorman *et al.*, 1998; Qin and Oxenham, 2003; Stickney *et al.*, 2004; Poissant *et al.*, 2006), one might expect to find reduced variability in performance across masking talkers in competing speech situations among cochlear implant users. As an example, Stickney *et al.* (2004) did not find statistically significant differences in implanted listeners' performance in the presence of three different masking talkers,

although the same conditions delivered to normal-hearing listeners showed significant variability across masking talkers.

Perhaps the most important result, shown in the right panel, is that there was no benefit of presenting the target and maskers in the spatial (F-RF) configuration. The same manipulation improved performance by as much as 30 percentage points with the unprocessed stimuli in Freyman *et al.* (2007), representing a 6 dB shift in SN ratio for equivalent (50% correct) performance. The lack of improvement with the current stimuli seems exactly the opposite of what might have been expected. That is, as noted in the Introduction, it would not have been surprising to observe a great deal of informational masking in the F-F configuration because of the absence of quality and pitch differences between the target and maskers in the vocoded tokens. If this type of interference was released by the F-RF presentation of the masker, the improvement could have been quite large. At least three explanations can be considered as to why no advantage was found.

First, it is possible that informational masking was present in the F-F configuration but was not released in the F-RF configuration because the latter condition did not produce different spatial perceptions for the target and masker. Informal observations by the experimenters and four naïve listeners who marked perceived spatial positions on a photocopy of a protractor suggested that this explanation was not correct. The vocoding process did not interfere with localization; i.e., the masker was heard well to the right in the F-RF configuration, and the target appeared from the front loudspeaker.

A second possibility is that a vocoded masker does not, in general, produce informational masking with a vocoded target, despite expectations to the contrary as explained above. When both the masker and target are impoverished in fine structure, leading to decreased intelligibility, the effectiveness of the interfering talkers in producing informational masking could have been theoretically reduced. On the other hand, the results of a number of studies (e.g., Arbogast *et al.*, 2002, 2005; Gallun *et al.*, 2005; Shinn-Cunningham *et al.*, 2005) using vocoded spectrally interleaved targets and maskers indicate that informational masking can be quite strong when the stimuli are spectrally degraded.

The third and, we believe, most likely reason that evidence of informational masking was not seen was the high SN ratios needed in this experiment. The combined use of nonsense sentence materials and envelope vocoding required SN ratios to be very high—so high that the target was clearly distinguishable from the maskers. That is, the difference in loudness between the target and masker probably reduced the confusability between them and limited the interference to purely energetic masking. Indeed, such was the impression of the investigators listening to the materials. If we assume that loudness differences reduce confusability, the greatest chance of seeing a benefit of the F-RF configuration would be where the intensity of the target and masker was most comparable (+3 dB SN ratio). However, even that SN ratio may be high enough to overcome informational masking, as the target is 6 dB above the level of either of the individual maskers in the two-talker complex. Also, the task at that SN ratio may have been too difficult for a benefit to be seen due to a floor effect.

The finding of no indications of informational masking in experiment 1 is an interesting result because it provides a counterexample to the results of Qin and Oxenham (2003) and Stickney *et al.* (2004), both of which seem to indicate increased informational masking with vocoded targets and maskers. The difference could be related to our use of a combination of two masking talkers rather than the one masking talker used in the earlier studies. Also, in Qin and Oxenham (2003) and Stickney *et al.* (2004), the indications of informational masking come from increased masking with speech maskers relative to continuous noise maskers. The method of assessing informational masking in our study is to measure spatial release from masking for conditions where release from energetic masking is not expected. In the second experiment we asked whether spatial release from masking could be observed with the processed targets and maskers using a different task that required lower SN ratios.

III. EXPERIMENT 2: DETECTION THRESHOLDS

When both target and masker are processed with the same vocoding and presented from the same (front) loudspeaker, it is likely that many of the cues necessary for extracting the target from the interference are absent. In experiment 1 the SN ratio was high enough so that the target may have stood out against the less intense masker, minimizing target/masker confusion and eliminating the benefits of providing spatial differences between the target and masker. In order to evaluate the hypothesis that the high SN ratios used in the recognition study were responsible for the absence of spatial release from masking, we sought to create stimulus conditions that were similar but where the task could be performed at lower SN ratios. In this experiment we used words excised from the target sentences, one of the five maskers from experiment 1, and the identical vocoding process. The primary difference was that the task for the subject was only to detect the presence of the target words in a four-interval four-alternative forced-choice (4AFC) adaptive paradigm. Assuming that the identical processing for the target and masker could create confusion that would affect even the detection of the presence of the target, alleviation of this confusion through the introduction of spatial differences could potentially lead to the kind of spatial release from masking expected, but not seen, in experiment 1.

A. Methods

1. Stimuli

Target stimuli were 20 consonant-vowel-consonant (CVC) words excised from the nonsense sentences used in experiment 1. Details described below about word selection and excision are identical to those in Balakrishnan and Freyman (2008). The 20 target words were chosen for clarity of production and ease of excision from the nonsense sentence waveforms. Typically, the target word was the second of the three keywords of each utterance. In those instances where the second word in the sentence was not easy to extract, first

or third keywords were taken. The 20 target words were either processed with the same six-channel vocoding used in experiment 1, or they were left unprocessed. They were scaled to equate their average power (rms) and then postpadded with zeroes to match the duration of longest word in the list (500 ms). The 20 words were concatenated to create a single file from which the experimental software randomly selected a single word and played it on each trial.

The majority of conditions employed a two-talker speech masker (sstk), one of the five maskers used in experiment 1. The unprocessed version was used with the unprocessed target words, and the six-channel vocoded version was used as a masker for the vocoded targets. An additional masker, speech spectrum noise (Byrne *et al.*, 1994), was used with three listeners to verify earlier findings in recognition experiments that spatial separation produced no release from purely energetic masking.

2. Apparatus

The detection experiments were conducted in an anechoic chamber measuring $4.9 \times 4.1 \times 3.12$ m. The walls, floor, and ceiling are lined with 0.72 m foam wedges. Subjects were seated in the center of the room in front of a foam-covered semicircular arc on which two loudspeakers were positioned. The front loudspeaker was at 0 deg horizontal azimuth; the right loudspeaker was at 60 deg to the right. Both were 1.9 m from the approximate center of the subjects' head and were at ear height for the typical adult.

The target words were delivered via TDT System I instrumentation. The output of the 16 bit digital to analog 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 to a Crown D40 amplifier and a Realistic Minimus 7 loudspeaker. The masker was delivered from a second computer (Dell Dimension XPD 333) via audio software (COOL EDIT PRO). The 35 s long interference segment was played continuously in loop mode over the duration of an adaptive track. The masker was attenuated (TDT PA4) before being mixed with the target. Calibration of the target and maskers was completed by measuring sound levels at the position of the subject's head with the subject absent. The target was calibrated to a sawtooth noise equated for average power to the target words. The maskers were calibrated for each channel using the speech-shaped noise masker.

3. Subjects

Listeners were five college students with hearing thresholds $\leqslant$20 dB HL in the frequency range of 500–4000 Hz (ANSI S3.6, 1996). The ages ranged from 19 to 43 years with a median of 21 years. None of the subjects participated in experiment 1, and none were well practiced in listening to vocoded signals.

4. Conditions and procedures

The target words always originated from the front loudspeaker. The same F-F and F-RF masking configurations from experiment 1 were used in this experiment also. In addition, a target front and masker front right (F-FR) condition was used with the two-talker maskers, where the masker from the front led the masker from the right by 4 ms, the opposite of the F-RF configuration. Thus, for a given run the target-masker configuration could be F-F, F-RF, or F-FR.

Subjects responded to each 4AFC trial using a button box with LEDs that marked the four intervals, one of which, selected randomly with equal probability, contained the target. Listeners were instructed to indicate the interval in which they thought they heard the target. Other than informing them that the target was always from the front, no special instructions were given to participants on how they might solve the task. Feedback was provided via an LED display that illuminated the target interval. For all conditions, the masker level was fixed at 53 dBC in each masker channel while the target level was adapted. A two-down one-up stepping rule was employed to estimate 70.7% criterion performance (Levitt, 1971). An individual adaptive track consisted of ten reversals, with the threshold computed as the arithmetic mean of the last six reversals. The initial step size for the adaptive track was 16 dB, which was halved after each reversal until a final step size of 2 dB was reached. For each condition, three consecutive adaptive tracks were run and final threshold determined as the arithmetic mean of the three individual thresholds.

At the beginning of the first listening session, subjects were verbally instructed, familiarized with the list of 20 target words (via print and audio), and given practice runs in quiet to familiarize them with the task. Before the main experiment was begun, one adaptive threshold estimate was obtained in quiet for both the unprocessed and vocoded targets. These were no higher than 6 dBC for both processing conditions. The sequence of runs for the main experiment was as follows.

First, data were collected for the F-F and F-RF configurations for both the vocoded and unprocessed stimulus conditions. The order of the collection of these four thresholds was randomly determined for each listener.

Second, all five subjects participated in a brief investigation of potential learning effects. A previous study with unprocessed target and masking stimuli (Balakrishnan and Freyman, 2008) had shown that some of the listeners' performance improved (albeit erratically) over successive exposure to the F-F condition. In the current experiment, learning effects were studied by obtaining ten additional adaptive runs in the F-F configuration for both the vocoded and unprocessed stimulus conditions. The order of the processing conditions was counterbalanced as best as possible with five listeners. The ten runs for each processing condition were run consecutively in a block with a break after the first five runs.

Third, thresholds were obtained for the F-FR configuration, with a random ordering of vocoded and unprocessed conditions across subjects. Unlike the F-RF condition, the F-FR configuration does not provide a large angular separation of the target and masker. This is because the lead sound in the two-source masker emanates from the same location (front) as the target. Several speech recognition studies using unprocessed target and maskers (Freyman *et al.*, 1999; Brun-

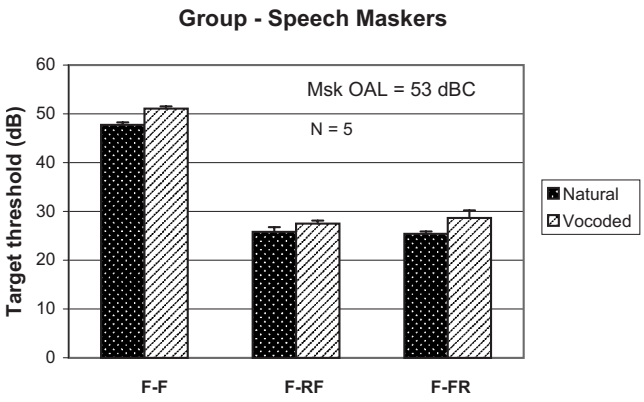

FIG. 2. Across-subject mean masked detection thresholds (in dBC) for CVC words excised from the nonsense sentences used in experiment 1. Thresholds for the F-F, F-RF, and F-FR configurations are shown for both unprocessed and vocoded speech. Error bars represent one standard error of the mean.

gart *et al.*, 2005; Rakerd *et al.*, 2006) and one detection study (Balakrishnan and Freyman, 2008) had shown that wide angular separations were not essential for release from informational masking. In the present study, we examined whether informational masking release would occur in the F-FR condition for the degraded and presumably more confusable vocoded stimuli.

Finally, three of the listeners also completed three adaptive tracks with a speech-shaped noise masker (Byrne *et al.*, 1994) for the F-F and F-RF loudspeaker configurations. Data were collected for only a subset of listeners for this masker because prior data collected for speech targets in noise showed no benefits of perceived spatial separation (Freyman *et al.*, 1999).

B. Results

Figure 2 displays mean detection thresholds and ±1 standard error for the nonspatial (F-F) and spatial (F-FR and F-RF) target/masker configurations. For the F-F configuration, detection thresholds averaged approximately 48 and 51 dBC for the unprocessed and vocoded speech, respectively. These correspond to SN ratios of −5 and −2 dB. Each spatial masking configuration produced reductions in detection thresholds of approximately equivalent sizes for both unprocessed and vocoded speech. The reductions, at 20 dB or greater, were considerably larger than the 5–8 dB spatial release from masking observed with recognition of the full unprocessed sentence stimuli and the same F-RF maskers in previous studies (Freyman *et al.*, 2001, 2007). The amount of release from masking is comparable to that observed for a different group of untrained subjects who listened to the unprocessed targets and maskers in a separate study (Balakrishnan and Freyman, 2008). The large spatial release could be due to the fact that in the F-RF configuration, the listener only had to detect the presence of any sound from the front loudspeaker, which required very low SN ratios in the fluctuating masker. On the other hand, in the F-F condition, merely detecting the presence of sound from the front loudspeaker was obviously not sufficient, and much higher levels were necessary for the target to stand out from the background in one of the intervals.

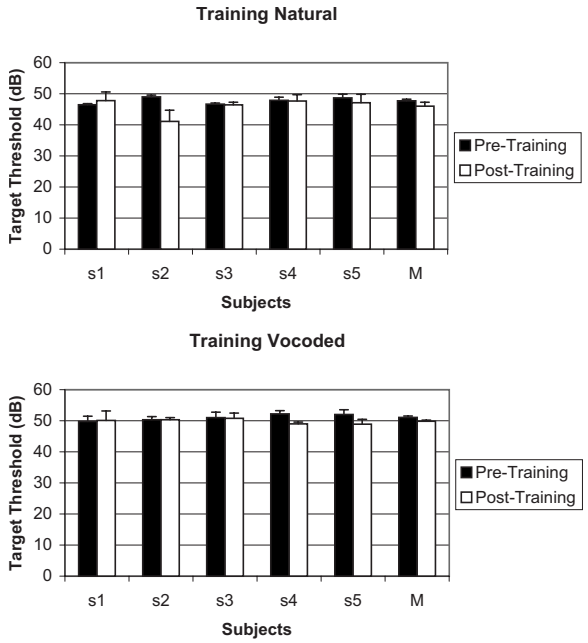

FIG. 3. Individual-subject and mean (*M*) masked detection thresholds (in dBC) for the words in the F-F configuration before and after extended listening experience. The solid bars are the average of the initial three runs, with the across-subject mean reproduced from Fig. 2. The open bars are the average of the last three of ten consecutive additional runs on the same condition. The top panel displays the data for the unprocessed conditions, and the bottom panel those for the vocoded conditions. Error bars for the individual subject data represent the standard deviation across three runs, while the error bars for the mean thresholds are one standard error.

Individual detection thresholds for the average of the last three of training runs completed for the F-F condition are shown in Fig. 3. They are compared with the average of the initial three runs for the same condition (the across-subject mean is a replot of data from Fig. 2). From the unprocessed data (top panel) it is clear that four of the subjects show pre- and post-training thresholds that are within error of measurement; a fifth subject (S2) showed an 8 dB improvement post-raining but with high variability across runs. The mean difference across subjects between pre-and post-training thresholds was not statistically significant ($t=1.064$, $p=0.347$). A few of the listeners in Balakrishnan and Freyman (2008) also had shown improvements for unprocessed targets and maskers in the post versus pretraining runs, but once again with high variability across runs. As shown in the bottom panel of Fig. 3, for all five listeners performance with the vocoded stimuli was relatively unaffected by the amount of experience acquired in the present study, and the mean difference between pre- and post-training thresholds was not statistically significant ($t=1.576$, $p=0.190$).

The results of the control condition when the masker was steady-state noise are shown in Fig. 4. In the recognition task (e.g., Freyman *et al.*, 2001, Helfer and Freyman, 2005; Brungart *et al.*, 2005; Rakerd *et al.*, 2006), the F-RF configuration with noise masking and a 4 ms delay provides no reliable improvement relative to the F-F condition. The data

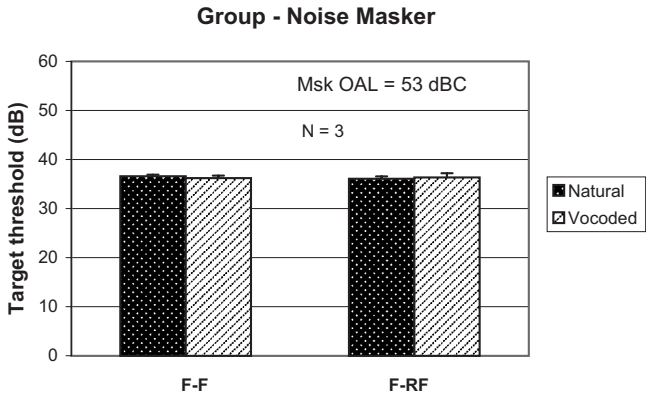

FIG. 4. Across-subject mean detection thresholds for target words (in dBC) in the presence of the noise maskers. Thresholds are shown for both vocoded and unprocessed conditions in both the F-F and F-RF configurations. Error bars represent one standard error of the mean.

in Fig. 4 confirm this lack of improvement in the current detection task for both unprocessed and vocoded speech.

IV. DISCUSSION

A. Summary of results

(1) *A priori* expectations to the contrary, no release from masking was found for recognition of six-channel noise-vocoded nonsense sentences using a spatial configuration that has been shown to produce a large release from informational masking with unprocessed stimuli.
(2) Variability in masking effectiveness across five different pairs of masking talkers was extremely small with the vocoded stimuli, in contrast to the observation of considerably larger variability with unprocessed versions of the same targets and maskers (Freyman *et al.*, 2007).
(3) A 4AFC adaptive detection task using words excised from the sentences revealed large amounts of spatial release from presumably informational masking, although not larger than what was observed for unprocessed stimuli. The spatial release was nearly equally effective whether or not there was a large difference in the target/masker perceived angle (F-RF versus F-FR).
(4) Short-term experience in the nonspatial condition through extended repetitions of adaptive runs did not reveal substantial evidence that listeners could learn to overcome masking in a condition where the informational component was presumably large.

B. Interpretation of the recognition data

In experiment 1, no release from masking was found when the vocoded target speech was presented from the front and the vocoded masking speech was presented from both the front and right loudspeakers. This lack of improvement contrasts strongly with the sizeable release from masking found for the same target and masker without vocoding (Freyman *et al.*, 2007). In one sense, the absence of masking release for the vocoded stimuli in the F-RF condition is counterintuitive because in this configuration masking release has been assumed to be related to a reduction in target/masker confusion caused by the spatial differences between the target and masker (Freyman *et al.*, 1999). With noise-excited vocoded speech, the confusion between target and masker could be expected to increase sharply, creating more informational masking and therefore increasing the release of masking that could be realized with spatial differences.

The fact that absolutely no spatial release from masking was found is due, we believe, to the fact that high SN ratios were required for reasonable performance with the vocoded nonsense sentences. At these SN ratios (ranging from +3 to +24 dB), the target speech was louder than the masking speech and should have stood out from the background. It is likely that there was no overall confusability between the target and masking utterances. Still, although the target was clearly audible there were nevertheless steady improvements with increasing SN ratio. This could occur because low-level portions of the target envelope-modulated noise could have been masked by or confused with the higher level portions of the masker envelope-modulated noise, to a decreasing degree as the SN ratio was increased. If confusion, as opposed to only energetic masking, was involved, then a release from masking would have been expected in the spatial conditions. We interpret the absence of spatial release from masking as an indication that there was no informational masking with the vocoded sentences in this experiment. This explanation is in agreement with the data and interpretations of Arbogast *et al.* (2005), who proposed that 0 dB may be near the upper limit of SN ratios where informational masking is likely to be observed.

It may be useful to consider how well the idea of an SN ratio ceiling fits with data from nonspeech informational masking tasks. A most compelling example is found in the data of Oxenham *et al.* (2003). That study looked at the effect of musical training on informational masking. Tones were detected in the presence of two types of multitone complexes that were either likely or not likely to produce informational masking. Nonmusicians were significantly and substantially more susceptible to informational masking with these signals. Most striking from the perspective of the current study is that for listeners showing significant informational masking (nonmusicians), over half of the informationally masked thresholds and, in particular, thresholds for the five listeners with the most informational masking clustered within a few decibels of 0 dB SN ratio; yet virtually no data point was reliably above 0 dB (Oxenham *et al.*, 2003; Fig. 2). The notion of a ceiling in the vicinity of 0 dB appears to be well supported by their data.

The speech recognition results of Qin and Oxenham (2003) may provide a counterexample to the idea that informational masking cannot be observed at positive SN ratios. Target utterances were lists of HINT sentences (Nilsson *et al.*, 1994). Maskers were male and female talkers, modulated noise, and steady speech spectrum noise. For unprocessed sentences, speech reception thresholds (SRTs) were lowest (best) with the single male and female talker maskers. However, when the targets and maskers were vocoded with the same type of processing used in the current study, the single-talker maskers actually produced the *highest* (poorest) thresholds. Most notably, this tendency continued and even increased slightly when the number of envelope channels

was reduced from 24 to 4 and the SRTs were between +10 and +20 dB SN ratio. To the extent that the increased masking efficiency of the speech maskers relative to the noise maskers demonstrates informational masking, these data must be taken as evidence of informational masking at positive SN ratios. Similar results and interpretations were offered by Stickney *et al.* (2004), and the results were also extended to actual cochlear implant listeners.

Although the results of Qin and Oxenham (2003) and Stickney *et al.* (2004) suggest that informational masking may exist in actual and simulated implant listening, even at positive SN ratios, there are some differences from the current study that could be important. First, both studies used single-talker maskers, while the current experiments used two-talker maskers. Poissant *et al.* (2006), also using six-channel vocoded speech, did not find that a two-talker masker was substantially more effective than a speech spectrum noise masker. There could be something fundamentally different about the confusions that occur with single-talker maskers, especially when the target and maskers are essentially modulated noise. Second, the measure of informational masking is different. Both of the studies referred to above discussed their data in terms of the difference in the masking efficiency of speech maskers versus speech spectrum noise, whereas in the current study we measured the spatial release from masking caused by spatial masker presentations that have produced no evidence of release from energetic masking.

One issue that should be considered when interpreting the speech masker versus noise masker comparison is that the net effect of the valleys and peaks of a fluctuating masker may not remain constant with changes in SN ratio. The fluctuations in a speech masker cause spectrotemporal peaks and valleys relative to an unmodulated noise masker of equal rms amplitude. The traditional data indicate that the effect of the additional masker energy in the peaks is more than offset by the reduced energy in the valleys, which allows relatively unmasked glimpses of the target speech (e.g., Festen and Plomp, 1990). However, to our knowledge, this demonstration has been made mostly, and perhaps exclusively, at low SN ratios [see examples in Festen and Plomp (1990) and George *et al.*, (2006)]. At low SN ratios, spectrotemporal peaks may be relatively inefficient, adding energy but not masking if portions of the target are already inaudible. Conversely, at strongly positive SN ratios where much of the speech is already audible, the benefits of listening in masker valleys may be more than offset by the additional masking resulting from the masker peaks.

At least for unprocessed speech, the differences in slope between the psychometric functions obtained with speech versus noise maskers lend support to the idea that improvements with speech maskers are most obvious at low SN ratios and may disappear at higher SN ratios. For example, although the unprocessed data of Qin and Oxenham (2003) show a better SRT (50% correct) for speech maskers than noise maskers, the mean sigmoidal parameters used to fit the psychometric functions suggest that the efficiency of their male speech masker is reduced relative to a speech spectrum noise masker only below −4 dB SN ratio, at which both functions reach 80% correct. Above that, the fits predict slightly *better* performance in the noise masker. The data for the female masker, which should produce little informational masking of the male HINT sentence talker, show a similar trend of converging with the noise masker data (at −2 dB SN ratio). Results from Stickney *et al.* (2004) (Fig. 2) show no difference between noise and speech maskers at 0 dB SN ratio at approximately 80% correct, but a much slower decline in performance in the presence of the speech masker as SN ratio is reduced. Thus, the relative efficiency of speech spectrum noise maskers and single-talker speech maskers is highly dependent on the SN ratio, which should be taken into consideration when interpreting the difference between them in terms of informational masking.

C. Interpretation of the detection data

Our own measure of informational masking, spatial release from masking with a two-source relative to a single-source masker, must also be considered carefully, as there potentially could be other factors besides positive SN ratios to explain why no masking release was measured. We felt it was important to support our interpretation by demonstrating masking release for the processed stimuli at lower SN ratios. In experiment 2, words excised from the sentence stimuli were used in a detection study. The SN ratios required for detection in the spatial masking conditions were all below −20 dB. If informational masking contributed to thresholds in the nonspatial (F-F) condition there would be ample room to observe it before any truncation could occur at 0 dB SN ratio. Indeed, the nonspatial thresholds were much higher than the spatial thresholds for both vocoded and unprocessed speech. They averaged −2 and −5 dB SN ratio respectively, and suggest an exceptionally large amount of informational masking (20–25 dB).

The fact that the spatial/nonspatial difference is larger than what has been observed previously for the recognition of similar unprocessed stimuli (Freyman *et al.*, 2007) may be partially explained by the number of decibels below 0 dB SN ratio at which the spatial thresholds are obtained. *Recognition* of the unprocessed nonsense sentences in the presence of the same spatial two-talker masker became nearly impossible as the SN ratio was reduced to −12 dB (Freyman *et al.*, 2001). Assuming a ceiling in the vicinity of 0 dB SN ratio in the nonspatial condition, it would not be possible to observe more than about 12 dB of informational masking in that study. With threshold SN ratios in the range of −20 to −25 dB obtained here in the spatial conditions, there was much more room below the ceiling. It is important to recognize, however, that the threshold SN ratio is not the only relevant difference between the word detection and sentence recognition studies. The brevity of words relative to sentences may make them more difficult to perceptually extract from the background speech, thereby increasing informational masking in the nonspatial case. A more modest amount of informational masking was seen for the same two-talker masker using nonsense sentence detection (Helfer and Frey-

man, 2005), even though threshold SN ratios were also low in their spatial condition.

If one considers the detection threshold to be the lower bound of the intelligibility function and that detection of sentences would not be expected to be very different from detection of words taken from the sentences, then it must be noted that the intelligibility function takes quite an unusual form in the F-RF condition, requiring a sensation level of almost 30 dB before even 10% intelligibility could be achieved. This must be because, at threshold, the detection of speech in the F-RF condition is based on the awareness of any sound coming from the front loudspeaker; it does not need to be recognized as speech. Because of the impoverished spectral information in the six-channel vocoded nonsense sentences, the stimuli must apparently be well above threshold before any reasonable proportion of them can be understood. An analogous result was reported by Micheyl *et al.* (2006), who showed that the detection of a target complex against a competing complex occurred at a level nearly 20 dB below that required for accurate fundamental frequency discrimination. With the F-F condition, the detection threshold was much higher, so the span of levels between detection threshold and the beginnings of recognition was not so great.

One of the suppositions made in the introduction to these studies was that informational masking, as revealed by the amount of spatial release, might be increased with vocoded stimuli because the target-masker similarity and confusability would be increased. In experiment 2, spatial release from informational masking for vocoded stimuli was sizeable, but it was not larger than that observed for unprocessed speech (e.g., Fig. 2). This could be due to the fact that even the unprocessed two-talker female masker produced a great deal of informational masking, as revealed by the F-F versus F-RF difference of more than 20 dB. It is certainly possible that other speech maskers consisting of, for example, the speech of male talkers would produce considerably less informational masking in the unprocessed condition. Although not yet tested, it is hypothesized that vocoding would increase informational masking substantially in that situation.

D. Alternative interpretations

We believe that the above analysis explaining differences in results between the two experiments in terms of SN ratio offers the simplest account for the data and the one that is most consistent with the subjective impression of the authors when listening to the stimuli themselves. However, it must be recognized that achieving the desired difference of lower SN ratios in experiment 2 involved several other potentially important differences from experiment 1. In experiment 1, the target was a large set of sentences, none of which were repeated during a session, whereas experiment 2 used a smaller set of words excised from the sentences. In experiment 1 the task was recognition, whereas in experiment 2 it was detection. Finally, while the two-talker masker was identical across the experiments, in experiment 1 it was gated on and off with the target, whereas in experiment 2 it was presented continuously. These differences between experiments must be evaluated as potential alternative explanations for the differences in the results.

There is a substantial literature employing nonspeech stimuli that has demonstrated significant amounts of informational masking in a wide variety of tasks, including discrimination, identification, and detection (e.g., Kidd *et al.*, 1995, 1998; Oh and Lutfi, 1999; Oxenham *et al.*, 2003). Thus, there is nothing inherent to recognition tasks that should lead to the expectation of an absence of informational masking. The smaller stimulus set size of 20 excised words relative to the 320 sentences might have predicted less informational masking in the word detection experiment because of reduced target uncertainty. However, as noted above, our data from Balakrishnan and Freyman (2008) with unprocessed speech show that the 4AFC detection experiment with words does indeed give a larger spatial release from masking in the spatial masking conditions than we have observed with sentences in other studies (e.g., Freyman *et al.*, 2007). The main difference is a higher threshold SN ratio required in nonspatial detection of words relative to sentences. Most importantly, however, substantial and consistent improvements in the F-RF condition relative to the F-F condition have been observed with the very same target sentences, identical maskers, and masker gating characteristics, and the same recognition task used in the current study (Freyman *et al.*, 2007). Thus, when considered against all the literature that has preceded it, the finding in experiment 1 of no F-RF advantage must be due to some consequence of the vocoding process itself, as opposed to procedural differences between the two experiments reported in the current study.

The consequence of vocoding that has been emphasized in this paper was the higher SN ratio required for reasonable levels of speech recognition performance. However, there was also the possibility that the F-RF configuration did not produce the desired spatial perceptions with the vocoded stimuli that would be helpful in releasing informational masking. Informal listening described in Sec. II for experiment 1, as well as the results of experiment 2 showing large F-RF advantages with the identically processed vocoded stimuli, suggests that the spatial perceptions were available and could be useful. Finally, it is possible that the vocoding process eliminated informational masking in the recognition task, even though it apparently did not in the detection task, for reasons that have nothing to do with the SN ratios at which the stimuli were delivered. One possibility is that informational masking is reduced or eliminated because vocoding destroys the intelligibility of the two-talker masker. It is quite reasonable to consider that the intelligibility of the masker (and certainly that of the target) might be more important in recognition than in detection. Balakrishnan and Freyman (2008) showed that time-reversing a two-talker masker had almost no effect on detection in unprocessed speech conditions, whereas Freyman *et al.* (2001) showed that time-reversing a masker did improve recognition performance in the F-F condition. On the other hand, with both time-reversed maskers and foreign language maskers, a substantial spatial advantage in the F-RF condition remained. Thus, there is evidence of significant informational masking

occurring in recognition tasks in which the maskers were not intelligible to the listeners. Finally, it is also clear that the vocoding process does not by itself eliminate informational masking (e.g., Arbogast *et al.*, 2002, 2005).

In summary, while alternative explanations cannot be ruled out, to a large extent they require suppositions of interactions between variables that the literature provides little foundation for. On the other hand, explanations based on the notion of a ceiling for informational masking in the vicinity of 0 dB SN ratio are consistent with what appear to be truncations around 0 dB SN ratio in studies of both speech and nonspeech stimuli [e.g., Oxenham *et al.*, (2003) and Arbogast *et al.*, (2005), both discussed earlier in this paper]. The biggest challenge to our hypothesis is that there was no F-RF advantage even at +3 dB SN ratio, which is not very much above that hypothetical ceiling. However, it could simply be the case that +3 dB SN ratio (+6 dB above each individual talker in the two-talker complex) is sufficient to allow the target to stand out of the background. Further, because performance for the F-RF condition at +3 dB was less than 10% correct (Fig. 1), there was little room for the F-F condition to show worse performance. Finally, we would not want to argue with the notion that depending on the relevance of the information reaching the listener from a competing talker and message, softer interfering speech could sometimes be distracting, drawing attention away from a louder target talker (e.g., with a highly familiar competing talker or when one's name is spoken). However, this distraction may not be strongly related to similarity or uncertainty assumed to be involved in informational masking. Our interpretation is rather that louder target speech stands out from a background in a way that limits confusion between the target and a softer masker to a degree that makes it difficult to demonstrate further gains from the perception of target/masker spatial differences.

E. Implications

As noted in the Introduction, the type of noise-excited vocoded speech employed in the current studies has been used in the past to simulate important features of cochlear implant processing. Potential extensions of the results to wearers of those devices must be made with extreme caution, and there was no attempt to simulate the kind of spatial cues that implant users might receive. Nevertheless, the current experiments were interpreted to reveal a great deal of spatial release from informational masking with competing vocoded speech presented at poor SN ratios. Therefore, to the extent that spatial hearing could be partially restored in implant listeners through bilateral implantation, the present results suggest that there is the potential to provide this additional advantage beyond other benefits realized from bilateral implantation. At positive SN ratios, our interpretation of the absence of spatial release from masking is that there was no informational masking, even in nonspatial conditions. This, when considered in the context of other speech recognition studies that can be interpreted to show evidence of informational masking at positive SN ratios (Sec. IV B), presents an equivocal picture of informational masking and the role of spatial hearing at such SN ratios. It is likely that the role of informational masking is highly stimulus dependent. However, if our explanations for the current data are correct, this study adds evidence of target-to-masker intensity ratio dependence in addition to stimulus dependence. As implant processing improves and implant users have the expectation of succeeding in more difficult SN ratio conditions, the cautious prediction from the current study is that the challenge of informational masking will remain unless improvements in implant design include features that allow better talker recognition and segregation of voices.

ACKNOWLEDGMENTS

The authors would like to thank the Associate Editor, Andrew Oxenham, and two anonymous reviewers for their helpful comments on an earlier version of this manuscript. We also acknowledge J. Ackland Jones for his assistance with data collection and the support of the National Institute on Deafness and other Communication Disorders (DC 01625).

ANSI. (**1996**). *ANSI S3.6–1996; Specifications for Audiometers* (American National Standards Institute, New York).

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

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

Balakrishnan, U., and Freyman, R. L. (**2008**). "Speech detection in spatial and non-spatial speech maskers," J. Acoust. Soc. Am. **123**, 2680–2691.

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

Brungart, D. S., Simpson, B. D., and Freyman, R. L. (**2005**). "Precedence-based speech segregation in a virtual auditory environment," J. Acoust. Soc. Am. **118**, 3241–3251.

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.

Dorman, M. F., Loizou, P. C., Fitzke, J., and Tu, Z. (**1998**). "The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6–20 channels," J. Acoust. Soc. Am. **104**, 3583–3585.

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

Festen, M. J., and Plomp, R. (**1990**). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. **88**, 1725–1736.

Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2001**). "Spatial release from informational masking in speech recognition," J. Acoust. Soc. Am. **109**, 2112–2122.

Freyman, R. L., Helfer, K. S., and Balakrishnan, U. (**2007**). "Variability and uncertainty in masking by competing speech," J. Acoust. Soc. Am. **121**, 1040–1046.

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

Gallun, F. J., Mason, C. R., and Kidd, G., Jr. (**2005**). "Binaural release from

informational masking in a speech identification task," J. Acoust. Soc. Am. **118**, 1614–1625.
George, E. L. J., Festen, J. M., and Houtgast, T. (**2006**). "Factors affecting masking release for speech in modulated noise for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **120**, 2295–2311.
Glasberg, B. R., and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**, 103–138.
Hawley, M. L., Litovsky, R. Y., and Culling, J. F. (**2004**). "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," J. Acoust. Soc. Am. **115**, 833–843.
Helfer, K. S. (**1997**). "Auditory and auditory-visual perception of clear and conversational speech," J. Speech Lang. Hear. Res. **40**, 432–443.
Helfer, K. S., and Freyman, R. L. (**2005**). "The role of visual speech cues in reducing energetic and informational masking," J. Acoust. Soc. Am. **117**, 842–849.
Kidd, G., Jr., Mason, C. R., and Brughera, A., and Hartmann, W. M. (**2005**). "The role of reverberation in release from masking due to spatial separation of sources for speech identification," Acta. Acust. Acust. **91**, 526–536.
Kidd, G., Jr. Mason, C. R., and Rohtla, T. L. (**1995**). "Binaural advantage for sound pattern identification," J. Acoust. Soc. Am. **98**, 1977–1986.
Kidd, G., Jr., Mason, C. R., Rohtla, T. L., and Deliwala, P. S. (**1998**). "Release from masking due to spatial separation of sources in the identification of nonspeech auditory patterns," J. Acoust. Soc. Am. **104**, 422–431.
Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477.
Li, L., Daneman, M., Qi, J. G., and Schneider, B. A. (**2004**). "Does the information content of an irrelevant source diffeentially affect spoken word recognition in younger and older adults?," J. Exp. Psychol. Hum. Percept. Perform. **30**, 1077–1091.
Micheyl, C., Bernstein, J. G. W., and Oxenham, A. J. (**2006**). "Detection and F0 discrimination of harmonic complex tones in the presence of competing tones or noise," J. Acoust. Soc. Am. **120**, 1493–1505.
Nerbonne, G. P., Ivey, E. S., and Tolhurst, G. C. (**1983**). "Hearing protector evaluation in an audiometric testing room," Sound Vib. **17**, 20–22.
Nilsson, M., Soli, S. D., and Sullivan, J. A. (**1994**). "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**, 1085–1099.
Oh, E. L., and Lutfi, R. A. (**1999**). "Informational masking by everyday sounds," J. Acoust. Soc. Am. **106**, 3521–3528.
Oxenham, A. J., Fligor, B. J., Mason, C. R., and Kidd, G., Jr. (**2003**). "Informational masking and musical training," J. Acoust. Soc. Am. **114**, 1453–1459.
Poissant, S. F., Whitmal, N. A., III., and Freyman, R. L. (**2006**). "Effects of reverberation and masking on speech intelligibility in cochlear implant simulations," J. Acoust. Soc. Am. **119**, 1606–1615.
Qin, M. K., and Oxenham, A., J. (**2003**). "Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers," J. Acoust. Soc. Am. **114**, 446–454.
Rakerd, B., Aaronson, N. L., and Hartmann, W. M. (**2006**). "Release from speech-on-speech masking by adding a delayed masker at a different location," J. Acoust. Soc. Am. **119**, 1597–1605.
Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Shinn-Cunningham, B. G., Ihlefeld, A., Satyavarta, L., E. (**2005**). "Bottom-up and top-down influences on spatial unmasking," Acta. Acust. Acust. **91**, 967–979.
Stickney, G. S., Assman, P. F., Chang, J., and Zeng, F. G. (**2007**). "Effects of cochlear implant processing and fundamental frequency on the intelligibility of competing sentences," J. Acoust. Soc. Am. **122**, 1069–1078.
Stickney, G. S., Zeng, F. G., Litovsky, R. Y., and Assman, P. (**2004**). "Cochlear implant speech recognition with speech maskers," J. Acoust. Soc. Am. **116**, 1081–1091.