Relative contribution of off- and on-frequency spectral components of background noise to the masking of unprocessed and vocoded speech

Frédéric Apoux[a] and Eric W. Healy
Department of Speech and Hearing Science, The Ohio State University, Columbus, Ohio 43210

(Received 16 December 2009; revised 7 May 2010; accepted 21 July 2010)

The present study examined the relative influence of the off- and on-frequency spectral components of modulated and unmodulated maskers on consonant recognition. Stimuli were divided into 30 contiguous equivalent rectangular bandwidths. The temporal fine structure (TFS) in each "target" band was either left intact or replaced with tones using vocoder processing. Recognition scores for 10, 15 and 20 target bands randomly located in frequency were obtained in quiet and in the presence of all 30 masker bands, only the off-frequency masker bands, or only the on-frequency masker bands. The amount of masking produced by the on-frequency bands was generally comparable to that produced by the broadband masker. However, the difference between these two conditions was often significant, indicating an influence of the off-frequency masker bands, likely through modulation interference or spectral restoration. Although vocoder processing systematically lead to poorer consonant recognition scores, the deficit observed in noise could often be attributed to that observed in quiet. These data indicate that (i) speech recognition is affected by the off-frequency components of the background and (ii) the nature of the *target* TFS does not systematically affect speech recognition in noise, especially when energetic masking and/or the number of target bands is limited. [DOI: 10.1121/1.3478845]

PACS number(s): 43.71.An, 43.71.Gv, 43.66.Ba, 43.66.Lj [MSS] Pages: 2075–2084

I. INTRODUCTION

Speech communication occurs under a variety of adverse conditions including those in which one or more interfering sounds are present simultaneously with the signal of interest, i.e., the target signal. Because most natural sounds, such as speech, are highly modulated both in time and frequency, the relationship between speech and noise intensities, i.e., the signal-to-noise ratio (SNR) is usually nonuniform across frequency. As a consequence, the energy associated with the interfering sounds may not completely overlap with that associated with the target so that a number of frequency regions will be dominated by the noise while others will be dominated by the target. Current models of speech recognition in noise (e.g., the glimpsing model) suggest that in such conditions the normal auditory system recognizes speech by processing primarily the regions of the spectrum that contain a relatively undistorted view of local target signal properties (Celmer and Bienvenue, 1987; Cooke, 2006; Apoux and Healy, 2009). The ability to separate the frequency regions dominated by the target from those dominated by the noise most likely arises from the well established property of the peripheral auditory system to operate as a kind of frequency analyzer (Fletcher, 1940). Accordingly, the frequency extent of each of these regions should correspond to that of an auditory filter and the internal representation of a target signal most likely results from the combination of a limited number of auditory filter outputs.

A central assumption in this view is that the regions of the spectrum considered as noise are essentially ignored. Accordingly, speech recognition should only be affected by the amount of noise present in the auditory filter outputs used to reconstruct the representation of the target. This assumption is partly motivated by the results of psychophysical studies showing that the detection of a pure tone is not substantially affected by the presence of noise so long as the two signals are separated in frequency by more than a certain bandwidth, the so-called critical bandwidth (Fletcher, 1940). Similarly, a small number of studies have since demonstrated that the processing of bands of speech is not substantially affected by the presence of complementary (largely non-overlapping) bands of noise, suggesting that speech recognition in noise also relies on the apparent independence of the auditory channels (Kidd *et al.*, 2005; Apoux and Healy, 2009).

While both psychoacoustic and speech studies indicate that the processing of a target stimulus is not necessarily disturbed by the presence of noise so long as the two signals excite discrete auditory filters, there are many instances in which the target and noise bands are separated in frequency by more than a critical bandwidth and still interact. For instance, it has been reported that the ability to detect amplitude modulation (AM) at a target frequency can be diminished in the presence of modulation at a remote frequency (Yost and Sheft, 1989). This phenomenon has been referred to as modulation detection interference (MDI), and is defined as the difference between threshold modulation depth in the

[a]Author to whom correspondence should be addressed. Electronic mail: fred.apoux@gmail.com

presence of a modulated interferer and that obtained in the presence of an unmodulated interferer. Elevated modulation detection thresholds have also been reported for unmodulated interferers (Bacon and Moore, 1993; Bacon *et al.*, 1995; Bacon, 1999; Gockel *et al.*, 2002). While part of the influence of an unmodulated interferer on modulation detection thresholds can be accounted for by within-channel interactions, other results are more difficult to explain in terms of peripheral interaction. For instance, MDI has been observed in conditions in which the interferer is far outside the passband of the auditory filter centered at the target frequency (Yost and Sheft, 1990; Bacon and Moore, 1993) or in conditions where target and interferer are presented to opposite ears (Bacon and Opie, 1994). More importantly, there is evidence that speech recognition is also susceptible to modulation interference (Apoux and Bacon, 2008a).

Evidence for the influence of an off-frequency masker also includes instances in which the presence of a nonoverlapping noise band significantly *increases* speech recognition. Such beneficial effects of noise were reported by Warren *et al.* (1997). The authors examined the intelligibility of a pair of widely separated narrow speech bands with and without a band of noise in the gap separating the speech bands. Results showed improved sentence recognition in the presence of the noise band. The improvement in intelligibility observed when a spectral hole that would normally contain speech is filled with noise has been referred to as "spectral restoration."

It is apparent from the above that, although the auditory system is able to process the output of each auditory filter with considerable independence, interferences that presumably reflect interactions at a more central level may also occur. Whereas the influence of off-frequency maskers on speech recognition has been evaluated in very specific circumstances (e.g., MDI, spectral restoration), a more general understanding of their role in the recognition of speech is lacking. In particular, it is seemingly important to assess directly the contribution of the off-frequency spectral components of a broadband noise to overall masking. Accordingly, the primary goal of the present study was to evaluate the relative influence of the off- and on-frequency spectral components of a broadband masker on speech recognition and subsequently to determine to what extent the isolated effects of the off- and on-frequency spectral components of this masker combine to produce the effects of these components when simultaneously present.

A potential problem with such investigation is that speech is a broadband signal with no well-defined holes in the spectrum. As a consequence, one cannot present noise bands simultaneously with speech without producing substantial overlap in the spectral domain. In order to circumvent this limitation, previous studies typically restricted speech stimuli to one or two spectral regions. This approach, however, may not accurately reflect the mechanisms involved in the processing of natural, i.e., broadband, signals. The approach used in the present study attempted to better mimic the broadband nature of speech sounds by preserving as much as 2/3 of the speech spectrum (in perceptual units). To achieve this goal, stimuli were divided into 30 contiguous 1-ERB_N (normal equivalent rectangular bandwidth; Glasberg and Moore, 1990) width bands and subjects were presented with as many as 20 speech bands.

The relative influence of the off- and on-frequency spectral components of a background noise was also evaluated for target stimuli whose temporal fine structure (TFS) had been replaced with tones. Evidence has accumulated to suggest that the TFS is critical for understanding speech in fluctuating backgrounds. In particular, many studies have reported that the ability to take advantage of the spectro-temporal fluctuations in the background is severely reduced if the TFS of the stimulus is disrupted (e.g., Nelson *et al.*, 2003; Qin and Oxenham, 2003, 2006; Stickney *et al.*, 2005; Füllgrabe *et al.*, 2006; Gnansia *et al.*, 2009).

In contrast to previous work (in which TFS cues were simultaneously removed from the target and the masker), the TFS of the masker was always left intact in the present study. The purpose of this manipulation was twofold. First, we were interested in determining the extent to which disruption of TFS information in the target can account for the poorer speech intelligibility in noise reported in the above studies. Indeed, it should be noted that disturbing the speech TFS may affect speech intelligibility in quiet. For instance, Lorenzi *et al.* (2006) showed that replacing the speech fine structure with pure tones can lead to a small decrease in performance in quiet. However, because Lorenzi *et al.* had only 16 spectral channels in the "vocoder" condition, one cannot exclude the possibility that the concomitant reduction in the number of spectral channels contributed to this small decrease. If disruption of the target TFS truly affects performance in quiet, then the effect observed in noise may be attributed, at least partly, to the reduced intelligibility already observed in quiet. Because only the TFS of the target was manipulated and because the number of spectral channels was kept constant across all conditions, the present study provided the opportunity to assess the effect of disrupting the speech fine structure independently from that of reducing the number of channels of spectral information.

Second, we were interested in how TFS cues in the target assist in the segregation of a signal and background into separate auditory objects. At least two alternatives exist. TFS cues could be involved in the separation of the energy related to the target from that related to the masker within each auditory channel (within-channel segregation). This hypothesis, however, is not consistent with the results of studies in which stimuli were presented against a steady background and showing that speech intelligibility is essentially unaffected (Gnansia *et al.*, 2009) or only slightly affected (Hopkins and Moore, 2009) by the disruption of TFS if the number of spectral channels is kept constant across speech processing conditions. Alternatively, TFS cues may primarily assist in identifying which channels are dominated by the target signal so that the output of these channels can be combined at a later stage to reconstruct the internal representation (across-channel integration). The approach used in the present study should help clarify these issues.

II. METHOD

A. Subjects

Thirty-three normal-hearing (NH) listeners participated in the present experiment (22 females). Their ages ranged from 19 to 37 years (average=22 years). All participants had pure-tone air-conduction thresholds of 20 dB HL or better at octave frequencies from 125 to 8000 Hz (ANSI S3.6-2004, 2004). They were paid an hourly wage or received course credit for their participation. This study was approved by the University Institutional Review Board.

B. Speech material and processing

The target stimuli consisted of 16 consonants (/p, t, k, b, d, g, ɵ, f, s, ʃ, ð, v, z, ʒ, m, n/) in /a/-consonant-/a/ environment recorded by four speakers (two for each gender) for a total of 64 vowel-consonant-vowel utterances (VCVs; Shannon *et al.*, 1999). The background noise was a simplified speech spectrum-shaped noise (SSN; constant spectrum level below 800 Hz and 6 dB/oct roll-off above 800 Hz) or a sentence randomly selected from the AzBio test (Spahr *et al.*, 2007). All sentences were played backward to eliminate to some extent linguistic content (see, Rhebergen *et al.*, 2005) and limit confusions with the target while preserving their speech-like acoustic characteristics (time-reversed speech; TRS). While the studies mentioned in the Introduction suggest that speech recognition can be diminished if low-frequency AM is imposed on an off-frequency noise, it is well established that intelligibility is better in fluctuating than in steady-state backgrounds when the speech and noise spectra largely overlap (e.g., Miller and Licklider, 1950; Festen and Plomp, 1990; Takahashi and Bacon, 1992; Gustafsson and Arlinger, 1994). It was therefore important to compare these two types of maskers.

Prior to combination, target and masker stimuli were filtered into 30 contiguous frequency bands ranging from 80 to 7563 Hz using 2 cascaded twelfth-order digital Butterworth filters. Stimuli were filtered in both the forward and reverse directions (i.e., zero-phase digital filtering) so that the filtering process produced zero phase distortion (for more details see Apoux and Healy, 2009). Each band was one ERB_N wide so that the filtering roughly simulated the frequency selectivity of the normal auditory system. Subjects were presented with n target bands (n=10, 15 or 20). These bands were selected randomly from the possible 30 and a new drawing took place at each trial. In one condition (UNP), the target bands were left intact. In the other condition (VOC), a technique similar to vocoder processing was used to replace the speech fine structure with a tone in each band. The envelope was extracted from each band by half-wave rectification and low-pass filtering at cf_m (eighth-order Butterworth, 48dB/oct roll-off). The filtered envelopes were then used to modulate sinusoids with frequencies equal to the center frequencies of the bands on an ERB_N scale. The value for cf_m was independently computed for each band so that it was equal to half the bandwidth of the ERB_N-width filter centered at the sinusoidal carrier frequency (Apoux and Bacon, 2008b). Masker bands were not subjected to vocoder processing. When present, the masker bands were added to the target bands in one of the following ways (see Fig. 1). In one condition (OFF), the n target bands were presented with 30-n complementary masker bands. Therefore, each of the 30 bands was *either* filled with speech *or* noise. In another condition (ON), stimuli were created by adding together the n target bands with the corresponding n masker bands so that target and masker bands completely overlapped in the spectral domain. In the last condition (BB), all 30 masker bands were presented simultaneously with the n target bands. The duration of the masker was always equal to target speech duration.

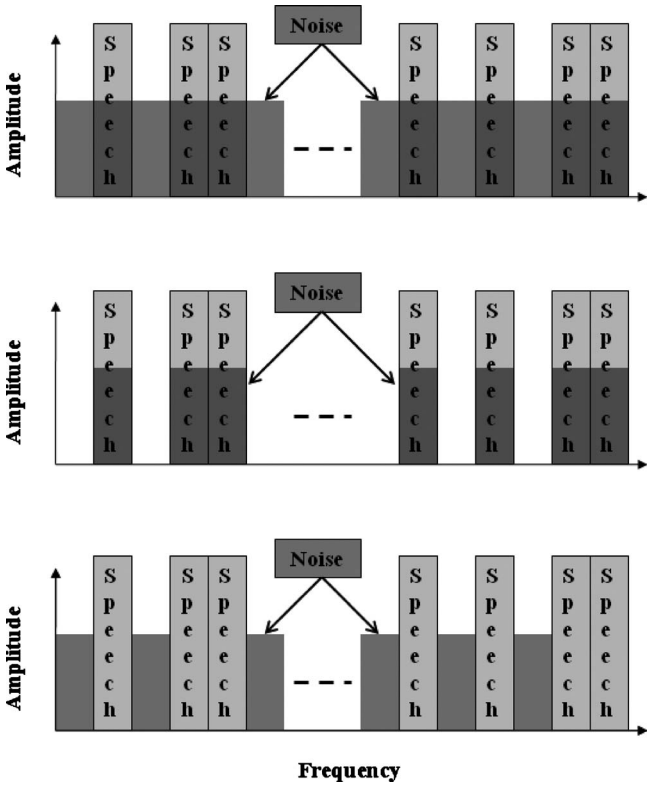

FIG. 1. Schematic of the three target/masker configurations used in the present study. The top, middle and bottom panels show examples for the broadband (BB), on-frequency (ON) and off-frequency (OFF) conditions, respectively.

The overall A-weighted level of the 30 summed target speech bands was normalized and calibrated to produce 65 dB. The overall level of the 30 summed masker bands was adjusted to achieve 6 or 0 dB SNR when compared to the 30 summed target speech bands. The motivation for testing two SNRs partly arose from the findings of Warren *et al.* (1997) indicating that the amount of spectral restoration varies with noise level. Target and masker bands were combined after level adjustment so that their spectrum levels remained identical to those in the broadband condition. Because spectrum levels were held constant, the long-term root mean square level of the stimuli generally increased with increased numbers of bands and overall SNR generally varied across number-of-band conditions. This approach was chosen because it best mimics what occurs in natural listening.

C. Procedure

The 33 participants were divided randomly and equally into three groups. Each group was tested in only one number

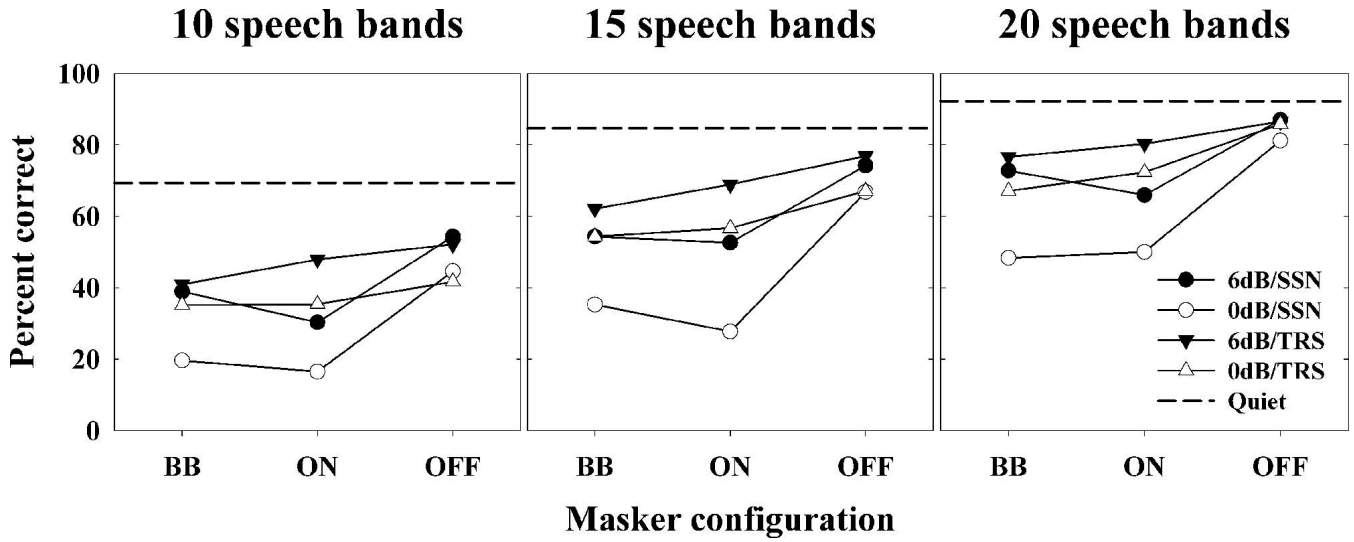

FIG. 2. Percent correct scores for consonant identification as a function of masker configuration [broadband (BB), on-frequency (ON) and off-frequency (OFF)] for the four combinations of SNR (0 and 6 dB) and masker type [speech-shaped noise (SSN) and time-reversed speech (TRS)]. The left, middle, and right panels correspond to the 10-, 15-, and 20-band conditions, respectively. In each panel, performance in quiet is indicated by a dashed line.

of ERB_N speech bands condition: 10, 15 or 20. All combinations of speech processing (UNP and VOC), masker type (SSN and TRS), masker configuration (BB, ON and OFF), and SNR (6 and 0 dB) were tested, resulting in 24 conditions. In addition, subjects were also tested in quiet in both speech processing conditions for a total of 26 conditions in each of the three number-of-band conditions.

Listeners were tested individually in a single-walled, sound-attenuated booth. Stimuli were played to the listeners binaurally through Sennheiser HD 250 Linear II circumaural headphones. The experiments were controlled using custom Matlab routines running on a PC equipped with high-quality D/A converters (Echo Gina24). Percent correct identification was measured using a single-interval 16-alternative forced-choice procedure. Listeners were instructed to report the perceived consonant and responded using the computer mouse to select 1 of 16 buttons on the computer screen. Approximately 30 min of practice was provided prior to data collection. Practice consisted of six blocks with each block corresponding to recognition of all 64 VCVs. In the first two blocks, vocoded VCVs were presented in quiet with all 30 bands present. In the remaining four blocks, vocoded VCVs were presented in quiet in a condition corresponding to the number of speech bands used for the experimental session. Feedback was given during the practice session but not during the experimental session. After practice, each subject completed the 26 experimental blocks in random order. Each experimental block corresponded to recognition of all 64 VCVs, presented in random order. The total duration of testing, including practice, was approximately four hours, which was divided into two sessions.

III. RESULTS AND DISCUSSION

A. Relative effects of off- and on-frequency spectral components

1. Results

Mean percent correct identification scores for the unprocessed condition (UNP) are presented in Fig. 2. Each panel corresponds to a given number-of-band condition and shows performance as a function of the masker configuration for the four combinations of SNR and masker type. The standard deviation across listeners ranged from 2 to 17 percentage points (mean=8 points) with no remarkable difference across masker type, masker configuration or number-of-band conditions. For clarity these are not displayed. As can be seen, the pattern of results was very similar across number of bands with the most notable difference being the overall performance level. As expected, consonant identification scores generally increased with increasing number of target speech bands. Performance was systematically better in quiet (dashed lines). It decreased when masker bands were inserted in the non-target bands (OFF condition). This difference was limited in the 20-band condition but increased with decreasing number of target speech bands.

A separate repeated-measures analysis of variance (ANOVA) with factors SNR, masker type and masker configuration was performed for each number-of-band condition.[1] All three analyses indicated a significant effect for each of the three factors ($p<0.0005$). The two-way interactions were all significant ($p<0.05$) except for that between SNR and masker configuration in the 10-band condition ($p=0.28$). Finally, the three-way interaction was significant in the 10- and 15-band conditions ($p<0.05$). Multiple pairwise comparisons (corrected paired t-tests)[2] were also performed separately for each number-of-band condition. The results of the principal comparisons are listed below.

The influence of masker type can be summarized as follows:

i. There was no significant effect of masker type in the OFF condition.
ii. In the ON condition, the presence of the SSN masker was significantly more disruptive to intelligibility than that of the TRS masker in both SNR conditions.

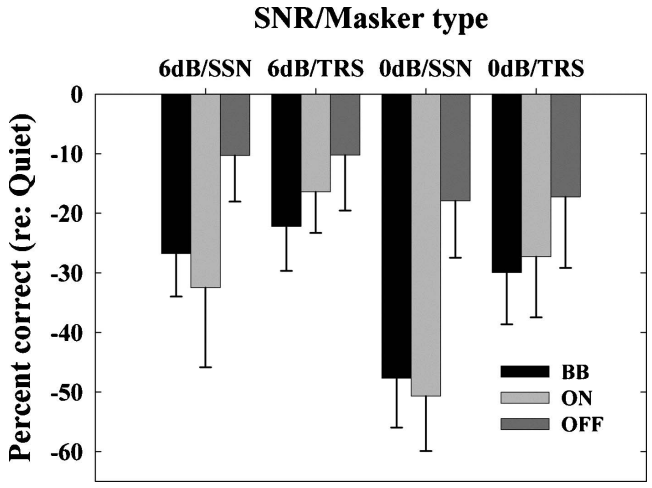

FIG. 3. Percent correct scores relative to those in quiet as a function of the SNR (0 and 6 dB) and masker type [speech-shaped noise (SSN) and time-reversed speech (TRS)] combination for the three masker configurations [broadband (BB), on-frequency (ON) and off-frequency (OFF)]. The error bars show one standard deviation.

iii. In the BB condition, the presence of the SSN masker was significantly more disruptive to intelligibility than that of the TRS masker at 0-dB SNR.

The influence of masker configuration can be summarized as follows:

i. Performance in the OFF condition was always significantly higher than that in the corresponding ON condition with the SSN masker.
ii. Performance in the OFF condition was always significantly higher than that in the corresponding BB condition.
iii. Performance in the BB condition did not differ significantly from that in the corresponding ON condition for seven of the 12 comparisons.

2. Discussion

The effects of masker configuration are summarized in Fig. 3 for each combination of SNR and masker type. To better illustrate the particular effects of the off- and on-frequency components, data have been averaged across number of bands and are now presented in terms of percent identification scores relative to quiet. Therefore, negative values correspond to poorer performance in noise. Figure 3 emphasizes two potentially interesting aspects of off-frequency masking.

A first aspect concerns the effects of an off-frequency masker on consonant recognition and can be deduced directly from the results observed in the OFF condition. As can be seen in Fig. 3, the presence of the off-frequency masker bands negatively affected performance in all conditions. While modulation interference was expected to occur with the TRS maskers, it had been anticipated that the presence of the off-frequency SSN bands would be *beneficial* to intelligibility. It is apparent from Fig. 3 that no such spectral restoration was observed in the SSN condition. Performance actually decreased in this condition and the amount of masking was similar to that observed in the TRS condition. The absence of spectral restoration seems in contradiction with the results of Warren *et al.* (1997) showing that when a spectral hole that would normally contain speech is filled with noise, performance improves. This discrepancy, however, may be attributed to differences between the stimuli used in the present study and those used in Warren *et al.* In their study, Warren *et al.* filtered the stimuli so that "none of the noise bands […] masked spectral components within the speech bands." In other words, the authors attempted to limit energetic masking. In the present study, target and masker bands were contiguous and they overlapped somewhat. As a consequence, the possibility exists that energy from the masker bands spilled out into the target speech bands, interfering with the processing of the latter. This possibility was also suggested by Apoux and Healy (2009). The authors pointed out that several effects (e.g., masking release) observed in their study could only be explained in terms of energetic masking. Accordingly, it may be assumed that the interference observed in the present study with the off-frequency SSN bands resulted from energetic masking (i.e., within-channel interactions).

A second aspect of off-frequency masking concerns the potential influence of the off-frequency components of the background when the on-frequency components are simultaneously present. The changes in performance observed when the off-frequency components are removed from–-or added to—the background should provide a good estimate of this influence. As expected, the amount of masking observed in the ON condition was comparable to that observed in the BB condition, confirming that the overall effect of a broadband masker is primarily governed by the on-frequency components of this masker, especially at relatively low SNRs. One may therefore reasonably conclude from Fig. 3 that the off-frequency components of the masker did not strongly influence performance in the BB condition.

Visual inspection of the data in Fig. 3, however, shows that the amount of masking observed in the ON condition was larger, at least numerically, than that in the BB condition with the SSN masker while it was smaller with the TRS masker. One reason to believe that this may reflect a real influence of the off-frequency components of the masker is that these two opposite effects are consistent with previous reports on MDI and spectral restoration. As mentioned earlier, it has been reported that consonant recognition should improve when a spectral hole that would normally contain speech is filled with noise. Such improvement, however, should not be observed with a temporally fluctuating masker such as speech because these maskers provide multiple opportunities to "listen" in the holes. Moreover, temporally fluctuating maskers have been shown to interfere with speech processing when they do not overlap spectrally with the speech target (e.g., Apoux and Bacon, 2008a). It is apparent in Fig. 3 that the addition of the off-frequency bands followed this expected pattern.

Another reason to believe in the actual influence of the off-frequency components of a broadband masker is that in several instances the addition of these components did significantly affect performance as revealed by the pairwise

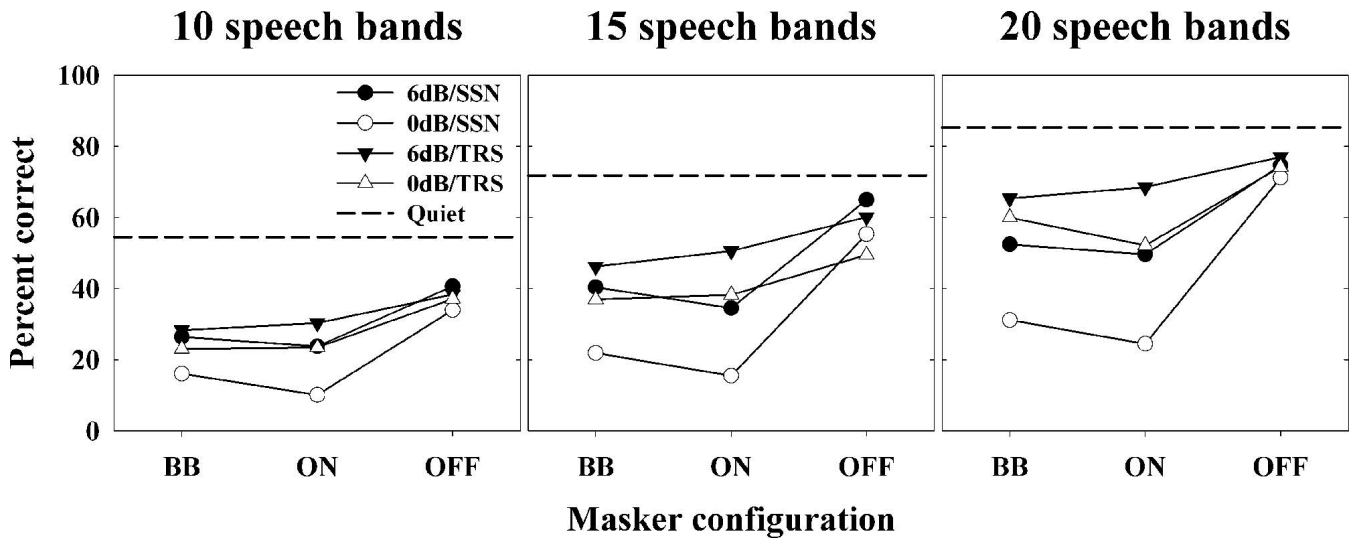

FIG. 4. The same as Fig. 2, but for vocoded target bands.

comparisons reported above between BB and ON conditions. More importantly, every one of the seven significant comparisons fell in the expected direction. Accordingly, one cannot rule out the possibility that the off-frequency components of a broadband masker may affect speech intelligibility. When the masker is a steady-state noise (SSN), these off-frequency components induce spectral restoration, and therefore performance may be better than it would be with the on-frequency components alone. When the masker fluctuates both in time and frequency (TRS), the amplitude fluctuations of the off-frequency components interfere with the processing of the speech target, and accordingly performance may be poorer than it would be with the on-frequency components alone. Taken together, the above findings suggest that while the success of listeners in recognizing speech in noise is primarily governed by the effects of the on-frequency components, an influence of the off-frequency components also exists.

B. Role of the target temporal fine structure in noise

1. Results

Mean percent correct identification scores for the vocoded condition (VOC) are presented in Fig. 4. Each panel corresponds to a given number-of-band condition and shows performance as a function of the masker configuration for each combination of SNR and masker type. A greater variability was observed in the VOC condition with the standard deviation averaging 9.5 percentage points (not displayed). Comparison between Figs. 2 and 4 suggests that all three factors (i.e., SNR, masker configuration, and masker type) had similar effects on unprocessed and vocoded speech. It is also apparent when comparing these two figures that consonant recognition scores were generally poorer in the VOC condition. The finding that performance in quiet was significantly poorer in the VOC condition (corrected paired t-test in each number-of-band condition, $p<0.01$) demonstrates that the speech fine structure conveys indispensable information to identify phonemes in quiet and that NH listeners use this information, irrespective of the total number of speech bands available.

Again, a repeated-measures ANOVA with factors SNR, masker type and masker configuration was performed for each number-of-band condition. The analyses indicated significant main effects of all three factors for the 10-band ($p<0.05$) as well as for the 15- and 20-band conditions ($p<0.0001$). In contrast to the UNP data, all two-way interactions were significant ($p<0.05$). Finally, only one of the analyses indicated a significant three-way interaction (20-band; $p<0.05$). Multiple pairwise comparisons (corrected paired t-tests) confirmed that the patterns of data observed with vocoded speech (i.e., the significant differences) were very similar to those observed with unprocessed speech. There were, however, two exceptions. First, performance was systematically higher in the OFF condition when compared to the ON condition for *both* maskers. Second, one of the significant pairwise comparisons between the BB and the ON conditions in the TRS masker was not consistent with the expected pattern (i.e., modulation interference) by indicating higher (rather than lower) performance in the BB condition.

As mentioned in the Introduction, one motivation for manipulating the speech fine structure was to investigate its role in the unmasking of speech. In an attempt to better illustrate this role or lack thereof, the proportion of responses correct relative to performance in quiet was computed separately for each subject in each condition. We reasoned that a portion of the difference between the recognition scores—in noise—for unprocessed and vocoded speech may be attributed to the difference that already exists in quiet and therefore, the comparison between scores in quiet and in noise should better reflect the specific role of TFS cues in noise. The averaged individual proportions are presented in Fig. 5. Each panel corresponds to a given number-of-band condition and shows proportion as a function of the four SNR and masker type combinations, for each masker configuration and speech processing combination. A series of corrected

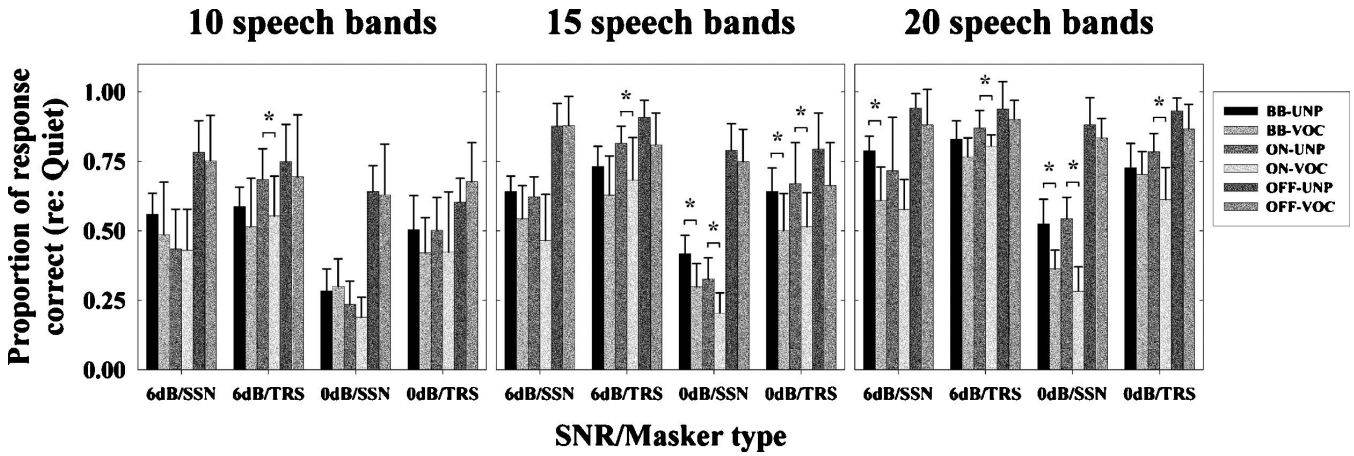

FIG. 5. Proportion of responses correct relative to performance in quiet as a function of the SNR (0 and 6 dB) and masker type [speech-shaped noise (SSN) and time-reversed speech (TRS)] combination. The left, middle, and right panels correspond to the 10-, 15-, and 20-band conditions, respectively. In each panel, the bars correspond to combinations of masker configuration [broadband (BB), on-frequency (ON) and off-frequency (OFF)] and speech processing [unprocessed (UNP) and vocoded (VOC)]. Asterisks indicate the UNP proportions that were significantly different from the corresponding VOC proportion. The error bars show one standard deviation.

t-tests indicated that only 11 out of the 36 proportions observed in the VOC condition were significantly different from those observed in the corresponding UNP condition (marked with asterisks).

2. Discussion

The modest number of significant comparisons in Fig. 5 suggests that the nature of the speech fine structure (i.e., speech or tones) did not play a critical role in most conditions and one may reasonably conclude a correspondingly modest role of the target TFS in the unmasking of speech.

Similarly, the number of significant differences between UNP and VOC proportions was fairly equally distributed across masker types, suggesting a limited interaction between the TFS in the target and that in the masker. In other words, the present data suggest that replacing the TFS of the target signal with tones has the same effect on performance, irrespective of the nature of the masker. This last finding is not consistent with previous studies indicating that speech recognition in fluctuating maskers is usually more affected than speech recognition in steady maskers by degradation of the TFS (e.g., Gnansia *et al.*, 2009). This differential effect is often attributed to a reduction in the ability to take advantage of momentary improvements in SNR, i.e., reduced masking release (see Füllgrabe *et al.*, 2006). The present study shows that masking release is not systematically reduced when only the TFS in the target is degraded. The results of a complementary experiment, however, confirmed that masking release can be reduced provided that the TFS in the target and masker are simultaneously degraded (see Appendix). Consequently, the results of the present study cannot be generalized to situations in which the TFS of the entire sound mixture is degraded.

Considering the number of factors manipulated in the present study, it may be worthwhile to look at the above proportions in more detail to identify specific situations in which the auditory system does in fact use the TFS of the target to extract speech from noise. A closer inspection of the data revealed that the significant differences between UNP and VOC proportions were not randomly distributed across conditions and two clear patterns can be described. First, the number of significant comparisons increased with increasing number of bands. In the 10-band condition, there was only one VOC proportion that differed significantly from the corresponding UNP proportion. There were five in the 15- and 20-band conditions.[3] This apparent relationship between number of speech bands and importance of TFS cues suggests that the latter may in fact play a role in segregating speech from noise in situations where the target speech is not restricted in frequency.

This finding is consistent with an earlier study by Oxenham and Simonson (2009) investigating the role of TFS cues in masking release. Based on previous work suggesting that pitch is an important grouping cue and that pitch perception depends on the accurate reception of TFS cues, the authors hypothesized that less masking release (i.e., less effective grouping) should be observed when stimuli are restricted to the high-frequency region, relative to the low-frequency region, because pitch information is less readily available from high frequencies. Indeed, it is well established that low-numbered harmonics produce a far stronger pitch than high-numbered harmonics. To test their hypothesis, Oxenham and Simonson simply compared the amount of masking release measured for unfiltered (i.e., broadband), low- and high-pass filtered sentences. The results showed that masking release was roughly equal in the two filtered conditions but systematically larger in the unfiltered condition. This result can also be interpreted as evidence that the auditory system is more efficient at using TFS cues in broadband conditions.

A second pattern that emerges from closer inspection of the current data suggests a second situation in which the auditory system uses TFS cues to extract speech from noise. This involves masker configuration. As can be seen in Fig. 5, removing TFS from the target did not significantly affect performance in the OFF condition. All 11 significant comparisons occurred either in the BB or in the ON condition. An obvious feature common to these two conditions is that the target and the masker overlapped in the spectral domain.

The fact that listeners only benefited from the presence of TFS information in the overlapping conditions may be interpreted as evidence that TFS information *in the target* was used by the auditory system to separate speech from noise *within* a given auditory channel.

It is difficult, however, based on the present data to exclude the alternative hypothesis that TFS in the target assisted in the separation of auditory channels dominated by the target signal from those dominated by the background noise (across-channel integration). The possibility exists that the auditory system simply did not need TFS cues to identify the channels containing the target in the OFF condition. Indeed, the target channels were not substantially corrupted in this condition. In these circumstances, cues other than TFS may have been available and sufficient to determine which channels contained speech. These potential cues include common onset/offset and amplitude modulation, which are well known to support the integration of information across channels (e.g., Darwin and Carlyon, 1995). In contrast, noise was present in the target bands in the BB and ON conditions. It is possible that this noise interfered with the cues used in the OFF condition, forcing the auditory system to rely more heavily on TFS cues. While there is no direct evidence suggesting that TFS cues are more robust to corruption than other cues, it has been suggested that amplitude modulation cues, i.e., envelope cues, are most affected by the addition of noise. For example, Dubbelboer and Houtgast (2007) evaluated the effects of noise on speech recognition. The authors divided the overall effect of noise into three types of interference: (i) a systematic lift of the envelope equal to the mean noise intensity, (ii) the introduction of stochastic envelope fluctuations, and (iii) the corruption of fine structure. Using a wavelet transformation based approach, Dubbelboer and Houtgast were able to evaluate the effect of each of these types of interference individually. Their results indicated that a systematic lift of the envelope was the most detrimental type of interference, suggesting that envelope cues are more affected by the introduction of noise.

The across-channel interpretation also has the advantage of falling more in line with the models of speech recognition in noise described in the Introduction, as these models do not emphasize the separation of the energy related to the target from that related to the background within a given channel.

IV. SUMMARY AND CONCLUSION

The present study evaluated the relative influence of the off- and on-frequency spectral components of a masker on consonant recognition. It showed that the off-frequency components of a masker have limited effects on consonant recognition. The amount of masking produced by the on-frequency components alone was similar to that produced by the broadband masker. A similar pattern of results was obtained with vocoded stimuli. The relative drop in performance associated with the loss of TFS information in the target signal was not systematically larger in noise than in quiet. The following conclusions can be drawn from this study:

i. As one may reasonably expect, manipulation of the spectral components present in the masker confirms that speech recognition in noise is primarily affected by the on-frequency components.

ii. Although limited, the influence of the off-frequency components may be observed in broadband conditions. The nature of this influence depends on the nature of the masker and occurs in conjunction with energetic masking. As a consequence, broadband maskers may produce opposing effects on speech recognition. While the on-frequency components of a steady-state masker can produce energetic masking, the off-frequency components have the potential to produce spectral restoration. Similarly, imposing modulation on a steady-state masker will decrease the amount of masking from the on-frequency components while it may lead the off-frequency components to induce modulation interference instead of spectral restoration.

iii. While eliminating fine structure information from the target signal resulted in a systematic decrease in performance, the decrease observed in noise was more often than not proportional to that observed in quiet. In other words, the drop in performance was not *relatively* larger in noise than in quiet in most of the conditions tested in the present study. It should be noted that this observation may not hold in situations where the TFS of the masker is also disrupted.

iv. In specific instances, however, TFS cues may contribute to the unmasking of speech. In particular, when the speech signal is not severely restricted in frequency, the auditory system may be able to use TFS cues *in the target signal* to segregate speech from noise. Whether these cues are used for within-channel segregation or across-channel integration remains unclear.

ACKNOWLEDGMENTS

This research was supported by grants from the National Institute on Deafness and other Communication Disorders (NIDCD Grant No. DC009892 awarded to author FA and DC008594 awarded to author EWH).

APPENDIX: INTELLIGIBILITY OF VOCODED SPEECH IN THE PRESENCE OF UNPROCESSED AND VOCODED MASKERS

1. Methods

Unless noted otherwise, all methodological and procedural details were identical to those used in the main experiment.

a. Subjects

Four NH listeners participated in the present experiment (3 females). One was the first author. All participants had pure-tone air-conduction thresholds better than 20 dB HL at octave frequencies from 125 to 8000 Hz (ANSI S3.6-2004, 2004).

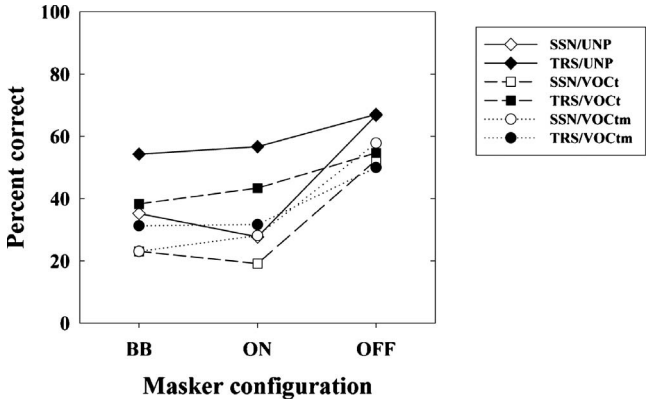

FIG. 6. Percent correct scores for consonant identification as a function of masker configuration [broadband (BB), on-frequency (ON) and off-frequency (OFF)] for the four combinations of masker type [speech-shaped noise (SSN) and time-reversed speech (TRS)] and masker processing [unprocessed (VOC_t) and vocoded (VOC_{tm})]. Data were collected in the 15-band condition only. For reference, the UNP data from Fig. 2 are also shown (diamonds).

b. Procedure

The four participants were presented with 15 ERB_N target speech bands, which were vocoded as described earlier. Two masker types (SSN and TRS) and three masker configurations (BB, ON and OFF) were tested with the overall SNR set to 0 dB. In contrast to the main experiment, it is the TFS of the masker bands that was either left intact (VOC_t) or replaced with tones (VOC_{tm}). When the TFS of the masker was replaced with tones, the processing took place prior to combination with the target, but the same tonal carriers were used for both target and masker. This is equivalent to modulating a single carrier with two envelopes and was done to avoid introducing potential extraneous TFS cues. The combination of masker processing, masker types, and masker configurations resulted in 12 conditions. In addition, subjects were also tested in quiet for a total of 13 conditions.

2. Results

Mean percent correct identification scores as a function of masker configuration for the four combinations of masker processing and masker type are presented in Fig. 6. For reference, the UNP data from Fig. 2 are also shown (diamonds). The standard deviations across listeners ranged from 6 to 13 percentage points (mean=9 points), but for clarity these are not displayed. In the OFF condition, recognition scores were generally comparable across masker processing. In the ON condition, two opposite effects were observed. When the masker was TRS, replacing the TFS of that masker with tones (VOC_{tm}) resulted in a 12 percentage point drop in performance relative to the VOC_t condition. When the masker was SSN, replacing the TFS of that masker with tones (VOC_{tm}) resulted in a 9 point improvement. In the BB condition, vocoding the masker lead to only a small decrease in performance in the TRS condition (7 percentage points) while no effect was observed in the SSN condition.

The above data indicate that a reduced improvement in scores in a modulated relative to a steady noise is observed primarily when the target and masker are both vocoded, as has been reported in previous studies. The so-called masking release was approximately 29 percentage points when fine structure was intact (see Fig. 6, ON masker configuration, diamonds). Masking release was reduced slightly, to approximately 24 percentage points when only the target lacked fine structure (see Fig. 6, ON masker configuration, squares). However, masking release was reduced to less than 4 percentage points when both target and masker lacked fine structure (see Fig. 6, ON masker configuration, circles). Therefore, the results observed in the main experiment are not in contradiction with previous works. Instead, they suggest a complex interaction between the nature of the TFS in the target and that in the masker.

[1]Consonant percent correct data were subjected to the following arcsine transform before all analyses $2\times\sin^{-1}(\sqrt{x/100})/\pi$ where x is the score in percent. The same transform was applied to the VOC data.

[2]Repeated-measures t-tests were performed and the results were corrected using the incremental application of Bonferroni correction described in Benjamini and Hochberg (1995).

[3]The lack of significance across the 20-band-6dB/SSN-ON conditions is presumably due to one subject whose performance in the unprocessed condition was especially low (17.19% correct). The performance of this subject in the remaining 25 conditions was well within the normal range.

ANSI S3.6-2004 (**2004**). *Specifications for Audiometers* (American National Standards Institute, New York).

Apoux, F., and Bacon, S. P. (**2008a**). "Selectivity of modulation interference for consonant identification in normal-hearing listeners," J. Acoust. Soc. Am. **123**, 1665–1672.

Apoux, F., and Bacon, S. P. (**2008b**). "Differential contribution of envelope fluctuations across frequency to consonant identification in quiet," J. Acoust. Soc. Am. **123**, 2792–2800.

Apoux, F., and Healy, E. W. (**2009**). "On the number of auditory filter ouputs needed to understand speech: Further evidence for auditory channel independence," Hear. Res. **255**, 99–108.

Bacon, S. P. (**1999**). "Some effects of background noise on modulation detection interference," Hear. Res. **129**, 20–26.

Bacon, S. P., and Moore, B. C. J. (**1993**). "Modulation detection interference: Some spectral effects," J. Acoust. Soc. Am. **93**, 3442–3453.

Bacon, S. P., Moore, B. C. J., Shailer, M. J., and Jorasz, U. (**1995**). "Effects of combining maskers in modulation detection interference," J. Acoust. Soc. Am. **97**, 1847–1853.

Bacon, S. P., and Opie, J. M. (**1994**). "Monotic and dichotic modulation detection interference in practiced and unpracticed subjects," J. Acoust. Soc. Am. **95**, 2637–2641.

Benjamini, Y., and Hochberg, Y. (**1995**). "Controlling the false discovery rate: A practical and powerful approach to multiple testing," J. R. Stat. Soc. Ser. B (Methodol.) **57**, 289–300.

Celmer, R. D., and Bienvenue, G. R. (**1987**). "Critical bands in the perception of speech signals by normal and sensorineural hearing loss listeners," in *The Psychophysics of Speech Perception*, edited by M. E. H. Schouten (Nijhoff, Dordrecht).

Cooke, M. P. (**2006**). "A glimpsing model of speech perception in noise," J. Acoust. Soc. Am. **119**, 1562–1573.

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

Dubbelboer, F., and Houtgast, T. (**2007**). "A detailed study on the effects of noise on speech intelligibility," J. Acoust. Soc. Am. **122**, 2865–2871.

Festen, J. M., 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.

Fletcher, H. (**1940**). "Auditory patterns," Rev. Mod. Phys. **12**, 47–65.

Füllgrabe, C., Berthommier, F., and Lorenzi, C. (**2006**). "Masking release for consonant features in temporally fluctuating background noise," Hear. Res. **211**, 74–84.

Glasberg, B. R., and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**, 103–138.

Gnansia, D., Péan, V., Meyer, B., and Lorenzi, C. (**2009**). "Effects of spec-

tral smearing and temporal fine structure degradation on speech masking release," J. Acoust. Soc. Am. **125**, 4023–4033.
Gockel, H., Carlyon, R. P., and Deeks, J. M. (**2002**). "Effects of modulator asynchrony of sinusoidal and noise modulator on frequency and amplitude modulation detection interference," J. Acoust. Soc. Am. **112**, 2975–2984.
Gustafsson, H. A., and Arlinger, S. D. (**1994**). "Masking of speech by amplitude modulated noise," J. Acoust. Soc. Am. **95**, 518–529.
Hopkins, K., and Moore, B. C. J. (**2009**). "The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise," J. Acoust. Soc. Am. **125**, 442–446.
Kidd, G., Jr., Mason, C. R., and Gallun, F. J. (**2005**). "Combining energetic and informational masking for speech identification," J. Acoust. Soc. Am. **118**, 982–992.
Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., and Moore, B. C. J. (**2006**). "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure," Proc. Natl. Acad. Sci. U.S.A. **103**, 18866–18869.
Miller, G. A., and Licklider, J. C. R. (**1950**). "The intelligibility of interrupted speech," J. Acoust. Soc. Am. **22**, 167–173.
Nelson, P. B., Jin, S.-H., Carney, A. E., and Nelson, D. A. (**2003**). "Understanding speech in modulated interference: Cochlear implant users and normal-hearing listeners," J. Acoust. Soc. Am. **113**, 961–968.
Oxenham, A. J., and Simonson, A. M. (**2009**). "Masking release for low- and high-pass-filtered speech in the presence of noise and single-talker interference," J. Acoust. Soc. Am. **125**, 457–468.
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.
Qin, M. K., and Oxenham, A. J. (**2006**). "Effects of introducing unprocessed low-frequency information on the reception of envelope-vocoder processed speech," J. Acoust. Soc. Am. **119**, 2417–2426.
Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (**2005**). "Release from informational masking by time reversal of native and non-native interfering speech," J. Acoust. Soc. Am. **118**, 1274–1277.
Shannon, R. V., Jensvold, A., Padilla, M., Robert, M. E., and Wang, X. (**1999**). "Consonant recordings for speech testing," J. Acoust. Soc. Am. **106**, L71–L74.
Spahr, A. J., Dorman, M. F., and Loiselle, L. H. (**2007**). "Performance of patients using different cochlear implant systems: Effects of input dynamic range," Ear Hear. **28**, 260–275.
Stickney, G. S., Nie, K., and Zeng, F.-G. (**2005**). "Contribution of frequency modulation to speech recognition in noise," J. Acoust. Soc. Am. **118**, 2412–2420.
Takahashi, G. A., and Bacon, S. P. (**1992**). "Modulation detection, modulation masking, and speech understanding in noise in the elderly," J. Speech Hear. Res. **35**, 1410–1421.
Warren, R. M., Hainsworth, K. R., Brubaker, B. S., Bashford, J. A., and Healy, E. W. (**1997**). "Spectral restoration of speech: Intelligibility is increased by inserting noise in spectral gaps," Percept. Psychophys. **59**, 275–283.
Yost, W. A., and Sheft, S. (**1989**). "Across-critical-band processing of amplitude-modulated tones," J. Acoust. Soc. Am. **85**, 848–857.
Yost, W. A., and Sheft, S. (**1990**). "A comparison among three measures of cross-spectral processing of amplitude modulation with tonal signals," J. Acoust. Soc. Am. **87**, 897–900.