Intelligibility of whispered speech in stationary and modulated noise maskers

Richard L. Freyman[a] and Amanda M. Griffin
University of Massachusetts, Department of Communication Disorders, 358 North Pleasant Street, Amherst, Massachusetts 01003

Andrew J. Oxenham
University of Minnesota, Department of Psychology, 75 East River Parkway, Minneapolis, Minnesota 55455

(Received 25 August 2011; revised 2 August 2012; accepted 7 August 2012)

This study investigated the role of natural periodic temporal fine structure in helping listeners take advantage of temporal valleys in amplitude-modulated masking noise when listening to speech. Young normal-hearing participants listened to natural, whispered, and/or vocoded nonsense sentences in a variety of masking conditions. Whispering alters normal waveform temporal fine structure dramatically but, unlike vocoding, does not degrade spectral details created by vocal tract resonances. The improvement in intelligibility, or masking release, due to introducing 16-Hz square-wave amplitude modulations in an otherwise steady speech-spectrum noise was reduced substantially with vocoded sentences relative to natural speech, but was not reduced for whispered sentences. In contrast to natural speech, masking release for whispered sentences was observed even at positive signal-to-noise ratios. Whispered speech has a different short-term amplitude distribution relative to natural speech, and this appeared to explain the robust masking release for whispered speech at high signal-to-noise ratios. Recognition of whispered speech was not disproportionately affected by unpredictable modulations created by a speech-envelope modulated noise masker. Overall, the presence or absence of periodic temporal fine structure did not have a major influence on the degree of benefit obtained from imposing temporal fluctuations on a noise masker.
 [http://dx.doi.org/10.1121/1.4747614]

PACS number(s): 43.66.Dc, 43.71.Gv, 43.71.Rt [MAS] Pages: 2514–2523

I. INTRODUCTION

Previous research indicates that many hearing-impaired listeners are less able than normal-hearing listeners to take advantage of temporal "valleys" or "dips" in a masker's amplitude fluctuations when listening to speech that is partially masked by modulated noise (e.g., Festen and Plomp, 1990; Bacon *et al*., 1998; Peters *et al*., 1998; Dubno *et al*., 2003; Jin and Nelson, 2006; Bernstein and Grant, 2009). The explanations proposed for this reduced masking release (MR) are varied, and involve a number of factors including, among other things, reduced audibility (Desloge *et al*., 2010), temporal masking (Dubno *et al*., 2003), reduced peripheral compression (Oxenham and Dau, 2004), and higher test signal-to-noise ratios (SNRs) at which masker fluctuations are less useful even in normal-hearing listeners (Bernstein and Grant, 2009; Bernstein and Brungart, 2011).

One further explanation is that the reduced MR is at least partially due to hearing-impaired listeners' relative inability to process temporal fine structure information, especially the periodicity of voiced segments of speech, that would help distinguish the speech during the dips in the modulated masker (e.g., Lorenzi *et al*., 2006; Gnansia *et al*., 2009; Hopkins and Moore, 2009, 2011). The argument has been bolstered by several studies that have also found reduced MR in listeners using cochlear implants, which do not preserve natural temporal fine structure, and normal-hearing people listening to noise- or tone-excited vocoded speech, in which the natural temporal fine structure in each frequency channel is replaced with filtered noise or a pure tone (Qin and Oxenham, 2003; Stickney *et al*., 2004; Hopkins and Moore, 2009; Ihlefeld *et al*., 2010). An alternative explanation for reduced MR is that hearing loss, cochear implants, and vocoding all disrupt and blur the fine spectral variations in speech, and it is more difficult to integrate the degraded spectral information across the relatively audible speech segments. To evaluate this alternative against the explanation based on loss of temporal fine structure, Gnansia *et al*. (2009) compared two forms of spectral smearing in speech stimuli delivered to normal-hearing listeners. The first form involved an overlap-add technique similar to that used by Baer and Moore (1993,1994), which preserves much of the original periodicity in the waveform. The second form was noise-excited envelope vocoding (e.g., Shannon *et al*., 1995), which replaces the original temporal fine structure in each frequency channel with noise. Gnansia *et al*. (2009) found that noise vocoding disrupted MR to a greater extent than the overlap-add technique, suggesting that degradations to temporal fine structure disrupt MR above and beyond what could be explained by spectral smearing. However, the periodicity-preserving manipulation led to better performance, and lower test SNRs, even in unmodulated noise, meaning that the increased MR may have been at least partially due to the lower SNRs that were tested in the periodicity-preserving condition (e.g., Bernstein and Brungart, 2011). The conclusions of Gnansia *et al*. (2009)

[a]Author to whom correspondence should be addressed. Electronic mail: rlf@comdis.umass.edu

 0001-4966/2012/132(4)/2514/10/$30.00

would be strengthened further if the corollary of their experiment produced analogous results; specifically, MR should be substantially disrupted by temporal fine structure degradation even in the absence of spectral degradation.

The purpose of the current study was to investigate the role of natural temporal fine structure in MR by using whispered speech, which has none of the periodic temporal fine structure cues of normal voiced speech, but which preserves the natural spectral variations created by the vocal tract. Whispering is performed by forcing air through a constricted opening between the vocal folds in the larynx. The resulting turbulence noise replaces phonation as a noisy input to be modified by the vocal tract resonances. Considered in terms of the source-filter model of speech production, only the source is affected, with few, if any, effects on the filter. Vestergaard and Patterson (2009) provide a summary of the acoustic and perceptual differences between normally phonated and whispered speech, which include a difference in duration of vowels, spectral tilt, and somewhat reduced intelligibility. Both vocoding and whispering remove the natural temporal fine structure of voiced speech. This may affect the redundancy of speech information and increase the perceptual similarity between a speech stimulus and a masking noise. However, vocoding reduces redundancy further by smearing spectral variations to a degree determined by the number and width of the frequency channels. Whispering also has additional consequences: the loss of phonation decreases the intensity of the portions of speech that are normally voiced, such as vowels, and the consonant-vowel intensity ratio is altered substantially. Modifications to consonant-vowel intensity ratios can be important for intelligibility under conditions where speech is degraded to force the listener to rely on temporal envelope information (e.g., Freyman *et al.*, 1991).

A number of issues arise when attempting to measure MR and compare the size of the effects across populations or types of speech processing. The most important of these, mentioned above, is that the magnitude of MR is dependent on SNR; see Bernstein and Grant (2009) for a quantitative analysis, and also Freyman *et al.* (2008), Oxenham and Simonson (2009), and Bernstein and Brungart (2011) for other discussions and analyses of this issue. Typically the growth of speech recognition performance in modulated noise or other fluctuating maskers (such as a single-talker interferer) is a shallower function of SNR than it is for steady noise. The result of this difference is that MR expressed in percentage points becomes progressively greater as SNR is decreased (A in Fig. 1). However, as it is decreased further eventually floor performance is reached in the steady noise condition, while performance in the modulated condition continues to decrease further with additional decreases in SNR. Because of these floor effects, the size of the measurable MR progressively decreases (B in Fig. 1). It can also be difficult at times to distinguish a true convergence of functions for modulated and steady maskers from ceiling effects (C in Fig. 1). Figure 2 in Stickney *et al.* (2004) is a good example of the type of function shown schematically in Fig. 1, and suggests that the convergence point (C) is around 0 dB SNR in that specific case. The modulated and unmodulated masking results of Qin and Oxenham (2003) converged at slightly below 0 dB SNR (see Freyman *et al.*, 2008). The idea that MR may be dependent on the SNR in unmodulated maskers has led Bernstein and Grant (2009) to propose that the reduced MR found in hearing-impaired listeners and cochlear-implant users may be in large part due to the higher SNRs at which these groups are often tested, rather than to any perceptual deficit that specifically impairs performance in modulated maskers.

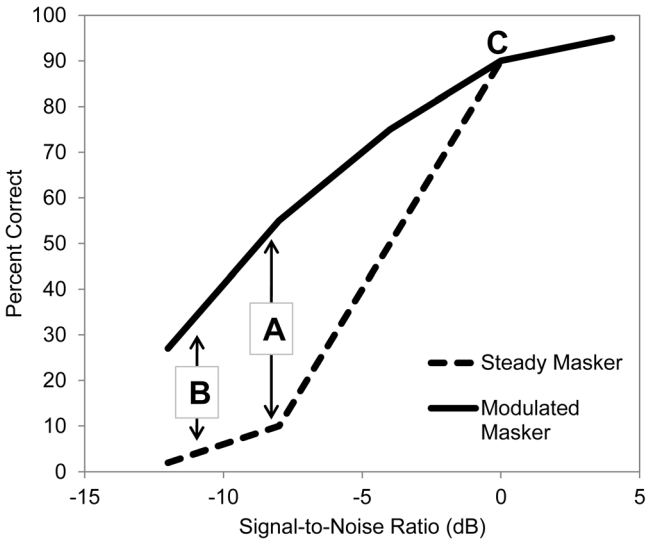

FIG. 1. Idealized psychometric functions for speech in steady and modulated masking. The slope in steady noise is steeper and the difference between recognition performance increases to a maximum (A) before floor effects in the steady masker cause a decrease (B). There is usually a convergence of the functions at higher SNRs (C).

Finally, while some researchers (e.g., Peters *et al.*, 1998; Nelson *et al.*, 2003; Jin and Nelson 2006) have examined the isolated effect of creating dips in the masker without making any other adjustments, other studies, particularly those using more complex modulation patterns such as a speech envelope (e.g., Festen and Plomp, 1990; Qin and Oxenham, 2003), but also those using periodic modulations (e.g., Hopkins and Moore, 2009; Ihlefeld *et al.*, 2010), adjust noise levels to equate root-mean-square (rms) levels in modulated and steady noise. Under these equal rms comparisons, the size of the improvement in modulated noise reveals specifically the net result of two opposing changes, the much improved SNRs in the temporal valleys of the modulated masker, and the slightly poorer SNRs during the increased masker peaks.

The current experiments attempted to deal with these issues in several different ways. Data from both whispered and natural conditions were collected as a function of SNR using the method of constant stimuli, rather than adaptive tracking, and are plotted as a function of SNR for each condition. The SNRs were kept the same for the comparisons of interest. To help determine whether floor and ceiling effects influence the data, different scoring criteria were applied in Experiment 1 to the same data, with the criterion for correctness in sentence recognition ranging from any part of the sentence correct to the entire sentence correct. Finally the SNRs were spaced in 3-dB steps so that with square-wave modulation, the effect of

masker dips alone, without adjustments in amplitude to equate rms, can be isolated by comparing each modulated masker data point with the steady-state masker result shifted horizontally by one data point (i.e., 3 dB).

The first experiment compared MR for whispered speech with that for natural speech. The second experiment compared the effects of vocoding on MR for both natural and whispered speech and compared each of these to the natural and whispered results from Experiment 1. The third experiment used unpredictable masker modulations created by multiplying speech-spectrum noise by the wideband envelope of running speech. The fourth experiment modified the amplitude distributions of whispered speech to be more like natural speech in an attempt to explain some differences found between the data obtained for whispered and natural speech.

II. EXPERIMENT 1: MASKING RELEASE IN WHISPERED VS NATURAL SPEECH

A. Methods

1. Stimuli

The target stimuli were new recordings of 320 nonsense sentences developed by Helfer (1997). These sentences were syntactically but not semantically correct and contained three key words, e.g., "The *ocean* could *shadow* our *peak*." The talker was an adult female, 24 years of age, who produced the sentences normally and by whispering. The speaking rate of the natural and whispered speech was very similar: the durations of the concatenated natural and whispered sentences were 594.4 and 592.8 s, respectively — a difference of less than 0.3%. They were recorded in a double-walled sound-treated booth using a cardioid condenser microphone (Audio-Technica, AT2020, Audio-Technica Corp., Tokyo, Japan) positioned approximately 8″ from the speaker's mouth. Recordings were immediately amplified through a tube microphone preamplifier (PreSonus TubePRE, Baton Rouge, LA) sampled at 22 050 Hz and stored on a personal computer. They were typically recorded five times while watching a volume unit (VU) meter, with one of the five selected after listening. The excess waveform before the beginning and after the end of the target sentences were trimmed off using Cool Edit 2000 (Syntrillium Software Corporation, Scottsdale, AZ). Each of the natural and whispered sentences was then scaled separately to have the same overall rms. Note that although naturally produced whispered speech was used in this study, Vestergaard and Patterson (2009) used the STRAIGHT vocoder with noise excitation to simulate whispered speech, a method that exerts more control. The rationale for the current use of naturally produced unprocessed whispered speech was to avoid any possible variations in spectrum that might result from that processing.

Two 12-s Gaussian noise maskers were generated and spectrally shaped to match the long-term spectrum of the respective target sentences (whispered and natural). A modulated version of each noise was created by multiplying it by a 16-Hz square pulse waveform (100% modulation, 50% duty cycle).

2. Subjects and procedures

Sixteen normal-hearing young adults (mean age 21 years) who passed a hearing screening at 20 dB hearing level (HL) at frequencies of 0.5, 1, 2, 3, 4, and 6 kHz listened monaurally through TDH-39P headphones (Telephonics, Interacoustics, Assens, Denmark) in a double-walled sound-treated booth. The stimuli were presented from a computer's 16-bit sound card, low-pass filtered at 8500 Hz (TDT FILT5), attenuated (TDT PA4), and sent through a headphone amplifier (TDT HBUFF5) using Tucker Davis Technology equipment (Alachua, FL) and a passive headphone attenuator before being routed to the headphones. Each subject listened to the 320 natural and whispered nonsense sentences in a different randomized order in the presence of the steady-state and modulated noise. On each trial, the target was mixed with a randomly selected segment of the 12-s noise by randomizing the offset of the beginning of the segment with a resolution of one sample point, as a consequence also randomizing the starting phase of modulation. The masker selection began and ended simultaneously with the target. Trials were divided into four blocks of conditions: two sentence types (natural and whispered) × two masker types (modulated and steady). SNRs ranged from −9 to +3 dB in 3 dB steps. Each subject listened to all 320 sentences, one 80-trial block for each of the four conditions, and 16 trials for each of the five SNRs within each block. The SNRs varied randomly from trial to trial within a block. The order of presentation of the four blocks was counterbalanced across listeners. Because there were four conditions, there were a total of 24 possible orders in which the blocks could be presented. Sixteen of those 24 were actually used (one unique order for each listener). The orders used met the criterion that within whispered and natural speech the modulated masker condition preceded the steady masker condition for exactly half the listeners (and vice versa), and each of the four conditions was presented first, second, third, and fourth an equal number of times. The subjects' task on each trial was to repeat as much of the sentence as possible.

Prior to test sessions, subjects completed a short practice session to familiarize themselves with the experimental procedures. Sentences used during the practice trials were not employed in the main experiment. There were ten practice trials in total, five whispered and five natural. Subjects were exposed to all five SNRs employed in the main experiment. In the later experiments, the practice was again ten trials with the stimuli and maskers adjusted to the experiment specifics. No feedback was given during the practice or test sessions.

B. Results and discussion

Consistent with previous findings in normal-hearing listeners, speech recognition performance for natural speech was much better in modulated than steady-state noise at the lower SNRs, with the difference diminishing to near zero at +3 and 0 dB SNRs [Fig. 2(a)]. Overall performance (total key words correct) was poorer with whispered speech in all conditions [Fig. 2(b)]. For example, the mean recognition score at −6 dB SNR was 41% correct for natural speech and

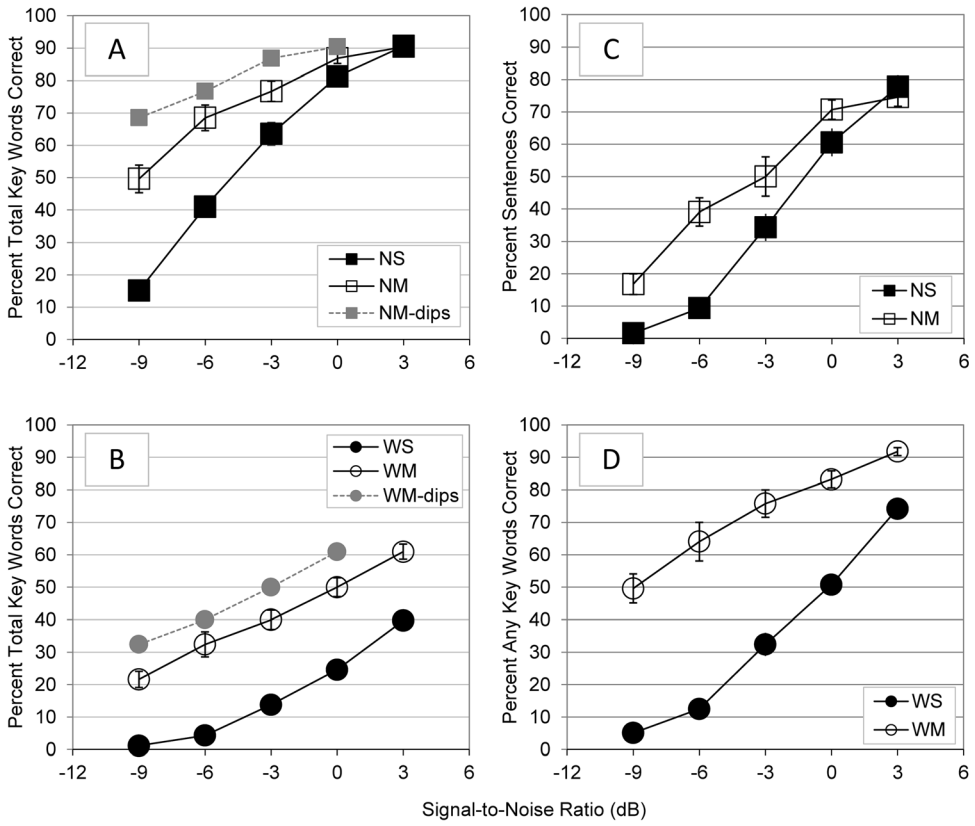

FIG. 2. Average speech recognition performance as a function of SNR in Experiment 1 in both steady (S) and 16-Hz modulated (M) masking. Error bars indicate +/− one standard error of the mean. (a) Percent total key words correct for natural speech (N); (b) Percent total key words correct for whispered speech (W). (c) Percent of sentences completely correct for natural speech. (d) Percent any key word correct in each sentence for whispered speech. In (a) and (b), the small gray symbols show the results in modulated noise shifted 3 dB to left, indicating performance if the modulations were imposed without the compensatory 3 dB increases during the noise on-times that were necessary to equate rms (NM-dips and WM-dips). The difference between the two sets of filled symbols reveals the isolated effect of imposing the masker gaps, and the difference between the open symbols and the small filled symbols is the effect of the compensatory 3 dB increases in the noise. For readability, the additional small symbols are not included in the other figures, but can be constructed by shifting the open symbols 3 dB to the left.

only 4.3% correct for whispered speech. However, the benefits of masker modulation for whispered speech seemed just as large as for natural speech [Fig. 2(a) and 2(b)], and the benefits were more robust across SNRs. The small gray symbols show the data for the modulated condition shifted 3 dB to the left. This negates the 3-dB amplitude compensation imposed on the modulated noise to maintain equal rms with the steady noise. The comparison between these gray symbols and the steady masker conditions allows the measurement of the isolated effect of interrupting the masker, separate from the effect of increasing the noise level in the on-periods by 3 dB to maintain an equal rms level overall. Subsequent figures do not display the shifted data points, but they can be constructed by a reader if desired. The pure effects of interruption are large for both natural and whispered speech. Expressed as the change in SNR required for 40% correct performance, the isolated effect of interruptions for whispered speech was 9 dB, but for the equal rms comparison the effect was only 6 dB [Fig. 2(b)].

Although MR seems similar for natural and whispered speech, a detailed comparison of the size of benefit in terms of percent correct improvement is complicated somewhat by the different levels of performance in the two conditions. For the whispered speech there may have been floor effects at the lowest SNRs, and for the natural speech there may have been ceiling effects for the highest SNRs. The right two panels of Fig. 2 adjust the criterion for what is considered a correct response differently for the whispered and natural speech in order to avoid floor and ceiling effects, respectively. The criterion for the natural speech was that all three keywords must be correct, otherwise, the response was considered to be incorrect. For the whispered speech, the response was considered correct if *any* of the three keywords was answered correctly. These criteria changes created the desired effect of similar baseline performance in steady noise for whispered and natural speech. The results for whispered speech Fig. 2(d) show large improvements in percent correct at SNRs where floor performance was suspected with the total-percent-correct measurement. For natural speech, the absence of MR at +3 dB SNR in Fig. 2(a) is observed not to be a ceiling effect in Fig. 2(c), as there was certainly the possibility of getting 80, 90, or 100% correct even with the "whole sentence correct" criterion.

The large advantage observed with the modulated masker for whispered speech at the highest SNRs, i.e., +3 dB, is unusual and does not match the expected form of the function depicted in Fig. 1. A follow-up experiment obtained the identical data with only whispered speech using a set of eight listeners (including one from the original 16), with a higher range of SNRs that overlapped with the original set (0 dB to +12 dB). Instead of running the natural speech conditions, the second set of listeners (selected with the same criteria as in the main experiment) listened to four whispered speech blocks of 80 trials each, two blocks in steady noise and two blocks in modulated noise. Thus, each of the eight listeners contributed twice the amount of data for the whispered speech condition in comparison to the original sixteen listeners, and so the aggregate data are based on the same number of trials across listeners. The results displayed in Fig. 3 show a diminishing degree of benefit from modulation with increasing SNR, but the effects are still observed at +9 dB SNR. This is unexpected from previous data, where typically little MR is observed at positive SNRs (Oxenham and Simonson, 2009; Bernstein and Grant, 2009).

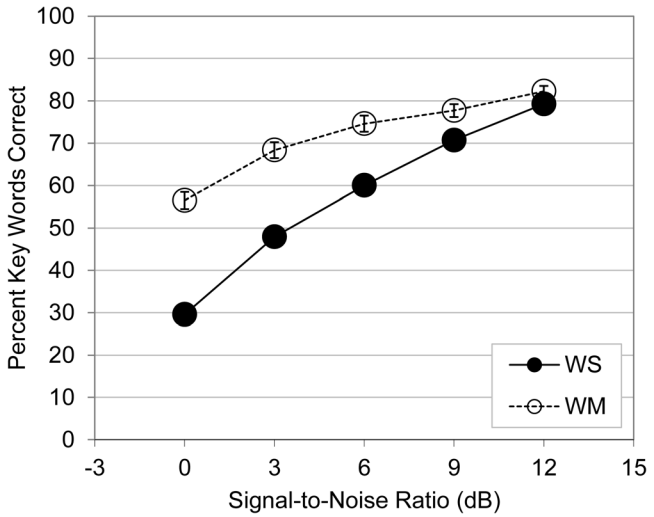

FIG. 3. Average percent key words correct as a function of SNR for whispered speech in steady (WS) and modulated (WM) noise, in a group of eight subjects who listened at higher SNRs. Error bars indicate +/− one standard error of the mean.

One potential reason for this outcome involves differences in the amplitude distributions of natural and whispered speech and is explored further in Experiment 3. Recognition performance for this group of subjects was 5–10 percentage points better than for the first group of listeners for the common SNR conditions of 0 and +3 dB. This could have been due to small subject differences between groups or to the extra exposure that the second group had with the whispered speech (since they ran it twice). When only the first of the two blocks of data obtained in steady and modulated noise from the second group were analyzed, the small between-group difference became even smaller.

The results of this experiment indicate that listeners derived considerable benefit from masker modulations even when temporal fine structure resulting from voicing was removed through the production of whispered speech. This is unlike some results reported in the literature on vocoding, which has been shown to produce diminished benefits from masker modulations (e.g., Qin and Oxenham, 2003). However, previous studies of vocoding differ from the present one in various ways, including the speech stimuli and other experimental details. Therefore, a second experiment was conducted to determine the effect of vocoding on the benefits of masker modulations using the same stimuli and experimental paradigm as used in the first experiment. In addition, it was of interest to observe the result when the whispered speech itself was vocoded.

III. EXPERIMENT 2: MASKING RELEASE IN VOCODED SPEECH

A. Methods

The nonsense sentences were processed off-line using a 16-channel noise-excited vocoder. The processing was largely as described by Qin and Oxenham (2003) using equal filter widths on the ERB_N scale (Glasberg and Moore, 1990) spanning the frequency range from 80 to 6000 Hz. The processing consisted of bandpass filtering the speech signals into 16 channels using sixth-order Butterworth bandpass filters, half-wave rectification, lowpass filtering using a second-order Butterworth filter with a cutoff frequency of half the bandpass-filter bandwidth or 300 Hz, whichever was less, multiplication by white noise, filtering again with the same 16 bandpass filters as used in the input filter bank, and summing the channel outputs. The output signals were concatenated and the long-term spectrum was inspected to ensure that it matched the corresponding masker, as was used for the unprocessed speech in Experiment 1. Sixteen listeners meeting the same criteria described in Experiment 1 were recruited to participate. Four of these listeners had participated in Experiment 1, and three others had some previous experience with the nonsense-sentence stimuli. Four total conditions were run using the same procedures and counterbalanced block orderings as in Experiment 1. The conditions were natural vocoded and whispered vocoded speech in both steady and 16-Hz modulated speech-shaped noise. The range of SNRs was from −3 dB to +9 dB in 3 dB steps, as determined from the results of pilot testing.

B. Results and discussion

The results, displayed in Fig. 4, show no benefits from masker modulations for natural vocoded speech, consistent with Qin and Oxenham's (2003) data for 8 - and 4-channel vocoding and with the data of Nelson and Jin (2003, 2004) for 4-channel vocoding. Both the current and earlier findings could have been at least in part due to the higher range of SNRs that were necessary for listeners to recognize speech in those conditions. The SNR required for 50% correct performance in steady noise was approximately +4 dB in both the current and the Qin and Oxenham studies. In cases where the SNR in steady noise was less than or equal to 0 dB (Qin and Oxenham, 2003; Hopkins and Moore, 2009; Stone *et al.*, 2011), some benefit of masker modulation has been observed for vocoded speech.

For vocoded whispered speech, the benefits of masker modulations were reduced considerably relative to unprocessed whispered speech (Fig. 4 vs Fig. 2), although the difference was not completely eliminated, as it was for natural speech. A comparison between the two figures also shows that vocoding reduced overall performance relative to the unprocessed whispered speech used in Experiment 1, especially in modulated conditions, despite the fact that the temporal fine structure was already noisy prior to vocoding. There was a chance that this difference could have been underestimated by the fact that seven of the listeners in Experiment 2 had previous experience with the nonsense sentences. However, when these seven were excised from the comparison, all of the within-subjects trends previously noted remained, and the reduction in overall performance observed for the vocoded speech increased slightly for both the whispered and natural recordings.

Using Praat software (Boersma and Weenink, 2011), an analysis was conducted on the natural, whispered, and vocoded speech to determine the extent to which whispering and vocoding affected the periodicity of the speech

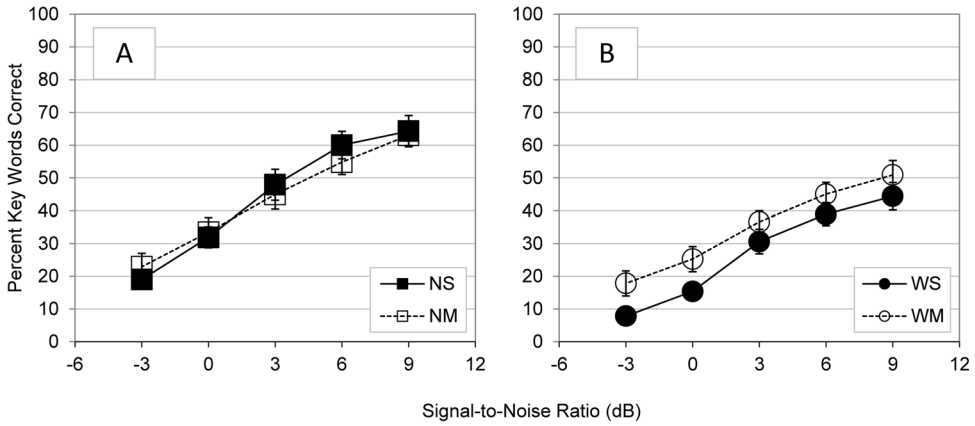

FIG. 4. Average percent key words correct as a function of SNR for vocoded speech in Experiment 2 in steady and 16-Hz modulated noise. Error bars indicate +/− one standard error of the mean. (a) natural speech, (b) whispered speech. The abbreviations in the legend are as in Fig. 2.

waveforms. The expectation was that a substantial portion of the natural utterances would show evidence of periodicity, given the prevalence of vowels and voiced consonants in normal sentence production. The statistic used was a broadband measure of harmonics-to-noise ratio, expressed in dB, computed every 10 ms, using the function "Harmonicity (ac)." The algorithm computes the harmonics-to-noise ratio using the height of the maximum peak of the normalized autocorrelation function. A typical harmonics-to-noise ratio for normal voiced speech is between 10 and 20 dB (e.g., Qi and Hillman, 1997). Figure 5 displays histograms of the harmonics-to-noise ratios in the 10-ms segments, based on the first 32 000 such samples (320 seconds) of the concatenated waveforms, which represented well over half of the corpus. The data points on the left show the number of samples in which no significant periodicity was found, due to either non-voiced speech sounds (such as fricatives) or silence. For the natural speech, a substantial proportion of the 10-ms segments had harmonic-to-noise ratios of between 10 and 20 dB, as expected from the voiced segments. Neither the whispered nor the vocoded speech had a similar distribution. Instead, very few of the whispered-speech segments had measurable harmonics-to-noise ratios. The apparent periodicity in the vocoded stimuli, as reflected by the small proportion of segments with harmonics-to-noise ratios of between 0 and 5 dB seems to reflect segments where a single one of the 16 bands dominates, producing a narrow-band stimulus that is categorized by the algorithm as having some periodic structure. In summary, it is doubtful that the reduced MR in vocoded speech, relative to whispered speech, is due to any significant preservation of periodic temporal fine-structure in whispered speech.

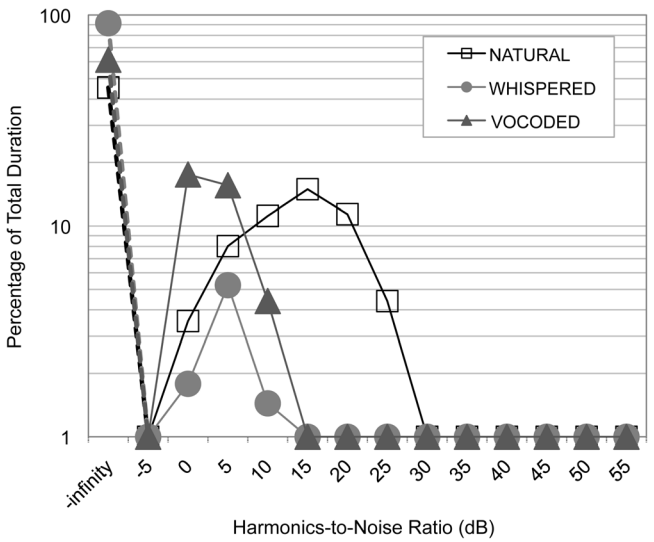

FIG. 5. Distributions of harmonics-to-noise ratios computed using Praat (see text for details). Minus infinity indicates that no periodicity was detected in a sample of waveform. In order to accommodate the logarithmic ordinate, percentages less than 1.0 were set equal to 1.0. These instances were greatest for whispered speech. The percentage of observations demonstrating strong periodicity was greatest for natural speech, followed by vocoded speech, and then by whispered speech.

IV. EXPERIMENT 3: SPEECH-ENVELOPE MODULATED NOISE

The idea that natural temporal fine structure is necessary for listeners to take advantage of gaps in a modulated noise does not appear to be supported by the current data. However, the modulations used in this and many other studies are highly regular and predictable. It may be that temporal fine structure would be more important in distinguishing the target from maskers with less predictable modulations. In this experiment, the speech-shaped noise masker was modulated by the broadband envelope extracted from the natural speech of the target talker. The comparison condition was square-wave modulated noise with an 8-Hz modulation rate. The 8-Hz rate was used rather than the 16-Hz rate from the first two experiments for two reasons. First, the 16-Hz data were already available for a (between-subjects) comparison, if necessary. Second, the slower 8-Hz frequency of modulation is closer to the peak of the modulation spectrum of speech (e.g., Houtgast and Steeneken, 1985). The purpose of the experiment was to determine if performance was worse in noise with irregular modulations in comparison to the square-wave modulations, and whether any such effects were more pronounced with whispered than with natural speech.

A. Methods

The speech-envelope-modulated (SEM) maskers were created by concatenating a number of target sentences together to build an extended (45-s) waveform and multiplying the wideband amplitude envelope extracted from that waveform by the whispered and natural steady-state speech masking noises (after having repeated those 12-s noises four times),

creating two separate SEM maskers of 45 s in duration. The envelope extraction was conducted by full-wave rectification and low-pass filtering at 20 Hz (third-order Butterworth), thereby incorporating the majority of envelope modulation energy from the speech. The 8-Hz square wave modulation was created as described earlier for the 16-Hz modulation. As before, a random starting point within these longer maskers was used to randomize the sample of masking and starting phase of modulation on every trial.

Sixteen new listeners meeting the same criteria as described in Experiment 1 were recruited to participate. Four conditions were run using the same procedures and counterbalanced block orderings as in Experiment 1. The conditions were SEM and 8-Hz modulated noise with both whispered and natural target sentences. The range of SNRs, determined from pilot testing, was different for natural and whispered speech, although there was considerable overlap.

B. Results

The results are shown in Fig. 6. It is clear that the SEM noise created slightly more difficulty than 8-Hz modulated noise for equivalent SNRs, especially at the lower SNRs for each masker. The differences in performance could have been related to any differences in modulation frequency, depth of modulation, other unspecified aspects of the maskers' modulation spectrum, as well as the regularity or irregularity of the modulations. The effects, however, are quite similar for whispered and natural speech, and so suggest that the unpredictability of the masker modulation (or the fact that the modulator had values other than 0 and 1) does not differentially affect the whispered speech that lacked normal periodic temporal fine structure.

V. EXPERIMENT 4: ALTERED AMPLITUDE DISTRIBUTIONS

The final experiment attempted to uncover the basis of the unexpectedly robust MR found for whispered speech at higher SNRs. As noted in the Introduction, the constriction in the area of the vocal folds during whispered speech production reduces the amplitude of what are normally higher-amplitude phonated speech segments. Measurements were made for a few sample utterances to demonstrate this using Adobe Audition (San Jose, CA) waveform analysis software. For example, in the sentence "A shop can frame a dog," the /ʃ/ in "shop" was 1 dB lower in amplitude than the following /ɑ/ sound in natural speech, but 9 dB above the /ɑ/ in whispered speech. It is conceivable that the relationship between the rms amplitude of the utterances used to compute the SNR and the information bearing elements of speech would be altered by these differences. This could have potentially affected the performance in steady-state noise at a given SNR as well as the size of MR. This study adjusted the amplitude distribution of the whispered sentences to match the natural sentences more closely and examined speech recognition in both steady and modulated noise, using within-subject comparisons to "uncorrected" whispered speech.

A. Method

1. Stimuli

The stimuli were created by first analyzing the amplitude distributions of both the natural and whispered corpuses. All the individual sentences within each corpus were concatenated to form extended waveforms. The wideband rms amplitude in each adjacent 30-ms segment of these long waveforms was then calculated. The forms of the distributions of amplitudes were clearly different for the natural and whispered speech, as indicated in Fig. 7(a). After transforming these to cumulative distributions, we calculated the differences in dB between whispered and natural speech for each percentile from 1 to 99 and adjusted the amplitude of each 30-ms segment in each whispered sentence to compensate for this difference. For example, if the rms of a particular segment of whispered speech placed it at the 70th percentile, but the same percentile was 5 dB higher for the natural speech, the amplitude of that whispered segment was amplified by 5 dB. Some adjustments, particularly in the lower percentiles, involved attenuating the whispered segment. Cumulative distributions for the 10th to 90th percentile data are shown in Fig. 7(b). The result of the adjustments is shown in both panels by the open circles. For whispered speech, approximately 84% of the segments had rms values below the total rms that was used to define the SNR during experiments presented thus far, whereas only 64% of the segments were below the overall rms level in the natural speech. Thus, even though the overall SNRs were equated in different conditions, there was a large difference in the proportion of time that the SNR was actually

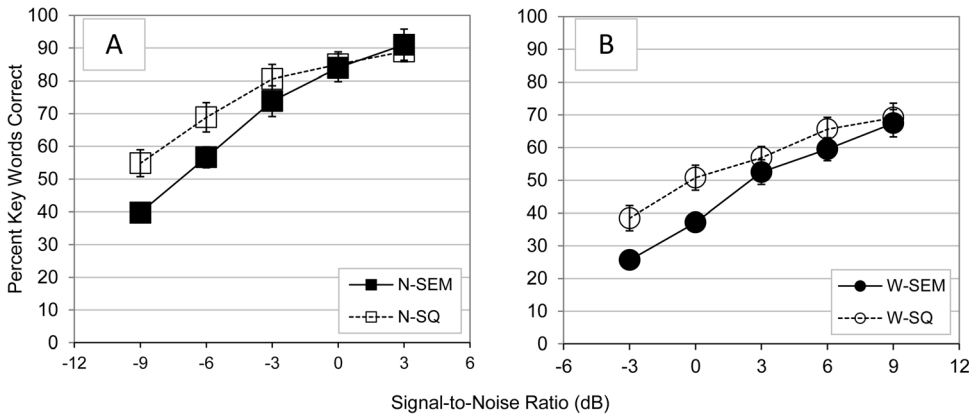

FIG. 6. Average percent key words correct as a function of SNR for natural (N) and whispered (W) speech for both 8-Hz square-wave modulated masking (SQ) and speech envelope modulated (SEM) masking. Error bars indicate +/− one standard error of the mean. (a) natural speech, (b) whispered speech.

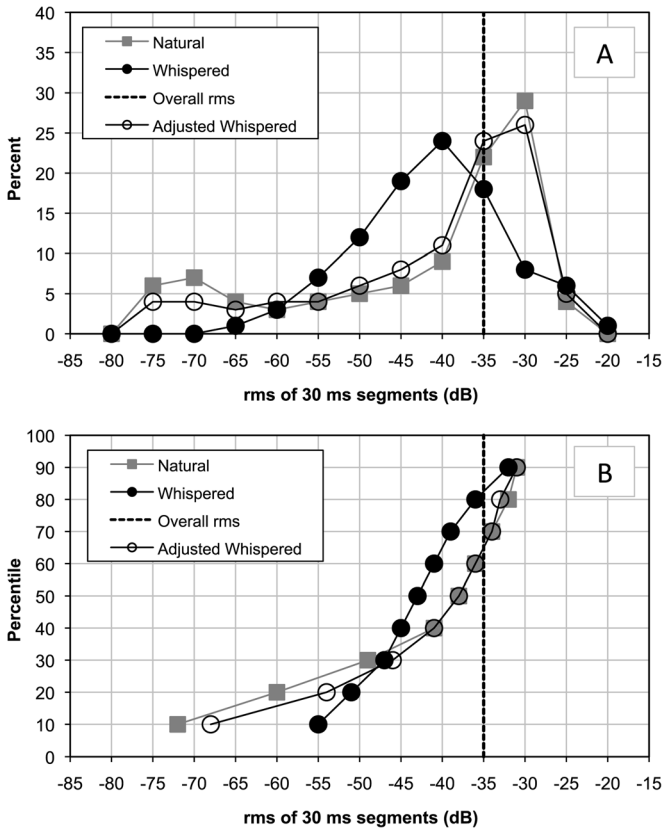

FIG. 7. Distribution (a) and cumulative distribution (b) of the rms amplitudes of 30-ms segments of waveforms in both whispered and natural speech. The open circles show the distribution after the whispered speech segments had been adjusted in amplitude in an attempt to correct for the difference between whispered and natural speech.

lower than that value, and this could tend to cause whispered speech to be more difficult to understand at equivalent SNRs. The result of the rms corrections is shown as the unfilled circles in Fig. 7. Applying the adjustment to the broadband stimulus may have introduced additional within-source modulation correlation to the speech (Stone and Moore, 2007), although this effect has not been shown to correlate strongly to speech intelligibility. Also, although each adjacent segment was analyzed and adjusted independently, without any smoothing of the transitions between segments, the resultant adjusted waveforms sounded like natural whispered speech to the experimenters, without any obvious distracting discontinuities.

2. Subjects and procedures

A new group of 16 young normal-hearing listeners participated. To facilitate a within-subjects comparison of the effects of the amplitude adjustments, the subjects listened to the unmodified stimuli from Experiment 1 also. Thus, there were again four conditions: uncorrected and corrected whispered speech in both steady and 16-Hz square-wave modulated speech-shaped noise. Because the primary interest was the robust MR for whispered speech at the higher SNRs, the range was shifted slightly and ranged from −6 dB to +6 dB in 3 dB steps. The randomizations of SNRs within a block and the counterbalancing of blocks were as before.

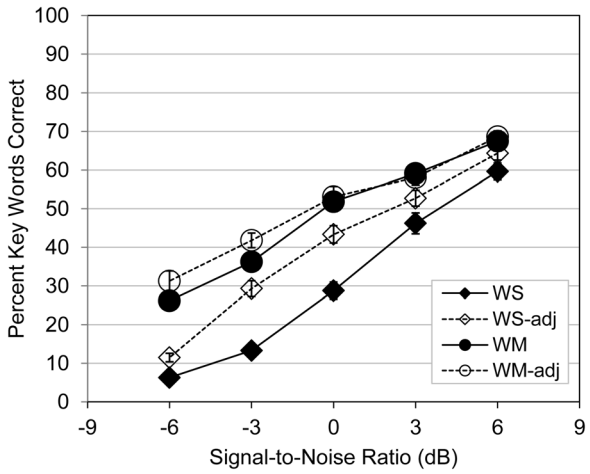

FIG. 8. Average percent key words correct as a function of SNR for unprocessed whispered speech (filled symbols) in steady (WS) and 16-Hz modulated noise (WM; similar to Experiment 1, but with new subjects), as well as the same data for whispered speech with adjusted amplitude distributions (WS-adj and WM-adj; open symbols) as shown in Fig. 7. Error bars indicate +/− one standard error of the mean.

B. Results

The results displayed in Fig. 8 indicate that the adjustments made to the amplitude distribution of whispered speech improved recognition performance in the presence of steady noise, especially at the intermediate SNRs, but had little effect on performance in the modulated masker. This had the effect of reducing the MR substantially at SNRs of −3 dB and above, similar to what was observed for natural speech in Experiment 1 (see Fig. 2). It appears that the maintenance of strong MR at the higher SNRs in unaltered whispered speech, as seen in Fig. 2, was indeed the result of the unusual amplitude distribution of whispered speech.

VI. DISCUSSION

The results of these experiments suggest that the periodic temporal fine structure of natural speech is not essential in order for listeners to take advantage of the amplitude valleys in modulated noise maskers. Whispered speech did not reduce MR relative to natural speech (Fig. 2), whereas vocoding with 16 channels of spectral resolution reduced MR considerably (Fig. 4). Speech recognition was slightly poorer when the masker modulations were derived from the wideband amplitude envelope of speech utterances, relative to 8-Hz square wave modulation. However, the reduction in performance relative to that for predictable masker modulations was approximately the same for whispered and natural speech (Fig. 6), suggesting that irregular modulations were not especially problematic for speech without periodic temporal fine structure. The unexpectedly large MR for whispered speech at higher SNRs appears to be due to differences in the amplitude distributions between whispered and natural speech (Figs. 7 and 8). This result is consistent with the analysis of Bernstein and Grant (2009), which indicated that the intensity importance functions are a critical feature in explaining and predicting MR.

The results do not mean that the processing of temporal fine structure is unimportant for speech recognition, as whispered speech was more difficult to understand than natural speech at any given SNR, even when amplitude distributions were adjusted. However, with these stimuli and conditions, it did not appear that temporal fine structure provided a disproportionate benefit for the perception of speech in modulated noise. This conclusion is different from that of Hopkins and Moore (2009), who used vocoded speech. Beyond differences in stimuli and methods between the two studies that might explain some of the differences, one additional explanation is that the changes in spectral resolution in vocoded speech, which occur in conjunction with a loss of temporal fine structure, also have an impact on MR. This explanation is supported by the results of Experiment 2 of the current paper, where vocoding reduced MR for whispered sentences that did not have periodic temporal fine structure to begin with. Thus, some other effect of our 16-channel vocoding, such as poorer spectral resolution and somewhat degraded temporal envelope representation (produced in part by the additional modulations of the vocoder noise carriers; see Whitmal *et al*., 2007; Stone *et al*., 2011,2012), must also impair speech intelligibility, particularly in modulated maskers.

The results of Gnansia *et al*. (2009) suggest a more important role for temporal fine structure in MR than does the current study, because spectral smearing produced by vocoding created much greater reductions in MR than did a different method of spectral degradation (Baer and Moore, 1993,1994) that does not include significant degradation of temporal fine structure. The vocoded conditions were tested at considerably higher SNRs than the alternative method of spectral smearing, because of the extra difficulty of the vocoded speech and because the experiment was designed so that the baseline performance was approximately 50% in all steady-noise masking conditions. It has become clear that SNR is a significant factor affecting the size of MR (Bernstein and Grant, 2009), and may have been partially responsible for differences in MR observed between the two types of spectral smearing. The importance of SNR was recently supported by Bernstein and Brungart (2011), who also measured MR with unprocessed speech, spectrally smeared speech, and speech with reduced temporal fine structure cues. They found roughly equal MR in all three conditions once SNR in steady noise was equated by changing the size of the response set, again suggesting that temporal fine structure is not critical in producing MR.

The conclusions from the present study are also consistent with those of Oxenham and Simonson (2009), who compared MR using lowpass- and highpass-filtered stimuli. The cut-off frequencies were selected to produce equivalent speech intelligibility in the two spectral regions, and to ensure that temporal fine structure was not readily available to listeners in the high-frequency region (as confirmed by the poor pitch discrimination thresholds for stimuli in the high-frequency region). Oxenham and Simonson (2009) found roughly equal MR for the lowpass- and highpass-filtered stimuli for a variety of maskers, whereas a specific benefit for temporal fine structure would have predicted greater MR in the lowpass-filtered conditions.

Accepting that the preservation of both spectro-temporal fine structure and envelope are important for recognizing speech in both steady and modulated noise, the mechanisms that pertain specifically to listeners' ability to take advantage of temporal valleys in modulated maskers are nevertheless not fully understood. The current data suggest that a loss of normal temporal fine structure does not prevent listeners from perceptually extracting target speech during the more audible time segments from within either predictable or unpredictable masker fluctuations. Rather, present and past findings on the topic appear to indicate that any type of degradation that reduces the number of redundant cues to speech recognition will make it more difficult to perceptually reconstruct disconnected audible speech segments into an understandable stream (e.g., Kwon and Turner, 2001; Nelson and Jin, 2004; Chatterjee *et al*., 2010; Gilbert and Lorenzi, 2010; Gnansia *et al*., 2010). These same degradations affect listening in steady-state noise as well, so the precise effect of a particular speech modification on MR is likely to be a complex interplay between various factors.

Whispering, like vocoding, does not produce a pure degradation of periodic temporal fine structure in the absence of other consequences. In the case of vocoding, the major additional consequence is spectral smearing; in the case of whispering, the primary side effect appears to be the altered amplitude relationships between sound segments. The results of Experiment 4 showed large changes in performance in steady-state noise when the amplitude distribution of whispered speech was manipulated, and support the conclusion of Bernstein and Grant (2009) that the intensity importance function (Studebaker and Sherbecoe, 2002) is a significant factor in determining MR. Whispering altered the normal distribution of amplitudes and presumably the intensity importance function as well. It will be important in the future to determine the effects of restoring within whispered speech the natural amplitude relationships of voiced speech in narrow frequency channels and even finer time segments, to augment the results of the broadband amplitude compensation in 30-ms segments used here. Other types of processing, such as dynamic range compression, that alter such amplitude distributions would be therefore expected to affect MR as well.

ACKNOWLEDGMENTS

The authors would like to thank Joshua Bernstein for his helpful discussions on the role of intensity importance functions in masking release, and to two anonymous reviewers for their comments on an earlier version of this manuscript. We are grateful to the National Institute on Deafness and other Communication Disorders for supporting this research (Grant No. R01 DC 01625 awarded to R.F. and Grant No. R01 DC 05216 awarded to A.J.O.).

Bacon, S. P., Opie, J. M., and Montoya, D. Y. (**1998**). "The effects of hearing loss and noise masking on the masking release for speech in temporally complex backgrounds," J. Speech Lang. Hear. Res. **41**, 549–563.

Baer, T., and Moore, B. C. J. (**1993**). "Effects of spectral smearing on the intelligibility of sentences in the presence of noise," J. Acoust. Soc. Am. **94**, 1229–1241.

Baer, T., and Moore, B. C. J. (**1994**). "Effects of spectral smearing on the intelligibility of sentences in the presence of interfering speech," J. Acoust. Soc. Am. **95**, 2277–2280

Bernstein, J. G., and Brungart, D. S. (**2011**). "Effects of spectral smearing and temporal fine-structure distortion on the fluctuating-masker benefit for speech at a fixed signal-to-noise ratio," J. Acoust. Soc. Am. **130**, 473–488.

Bernstein, J. G. W., and Grant, K. W. (**2009**). "Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **125**, 3358–3372.

Boersma, P., and Weenink, D. (**2011**). "Praat: Doing phonetics by computer," http://www.praat.org/ (Last viewed August 16, 2011).

Chatterjee, M., Peredo, F., Nelson, D., and Baskent, D. (**2010**). "Recognition of interrupted sentences under conditions of spectral degradation," J. Acoust. Soc. Am. **127**, EL37–EL41.

Desloge, J. G., Reed, C. M., Braida, L. D., Perez, Z. D., and Delhorne, L. A. (**2010**). "Speech reception by listeners with real and simulated hearing impairment: Effects of continuous and interrupted noise," J. Acoust. Soc. Am. **128**, 342–359.

Dubno, J. R., Horwitz, A. R., and Ahlstrom, J. B. (**2003**). "Recovery from prior stimulation: Masking of speech by interrupted noise for younger and older adults with impaired hearing," J. Acoust. Soc. Am. **113**, 2084–2094.

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.

Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2008**) "Spatial release from masking with noise-vocoded speech," J. Acoust. Soc. Am. **124**, 1627–1637.

Freyman, R. L., Nerbonne, G. P., and Cote, H. A. (**1991**). "Effect of consonant-vowel ratio modification on amplitude envelope cues for consonant recognition," J. Speech Hear. Res. **34**, 415–426.

Gilbert, G., and Lorenzi, C. (**2010**). "Role of spectral and temporal cures in restoring missing speech information," J. Acoust. Soc. Am. **128**, EL294–EL299.

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., Pean, V., Meyer, B., and Lorenzi, C. (**2009**). "Effects of spectral smearing and temporal fine structure degradation on speech masking release," J. Acoust. Soc. Am. **125**, 4023–4033.

Gnansia, D., Pressnitzer, D., Pean, V., Meyer, B., and Lorenzi, C. (**2010**). "Intelligibility of interrupted and interleaved speech in normal-hearing listeners and cochlear implantees," Hear. Res. **265**, 46–53.

Helfer, K. S. (**1997**). "Auditory and auditory-visual perception of clear and conversational speech," J. Speech. Lang. Hear. Res. **40**, 432–443.

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.

Hopkins, K., and Moore, B. C. J. (**2011**). "The effects of age and cochlear hearing loss on temporal fine structure sensitivity, frequency selectivity, and speech reception in noise," J. Acoust. Soc. Am. **130**, 334–349.

Houtgast, T., and Steeneken, H. J. M. (**1985**). "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," J. Acoust. Soc. Am. **77**, 1069–1077.

Ihlefeld, A., Deeks, J. M., Axon, P. R., and Carlyon, R. P. (**2010**). "Simulations of cochlear-implant speech perception in modulated and unmodulated noise," J. Acoust. Soc. Am. **128**, 870–880.

Jin, S. H., and Nelson, P. B. (**2006**). "Speech perception in gated noise: The effects of temporal resolution," J. Acoust. Soc. Am. **119**, 3097–3108.

Kwon, B. J., and Turner, C. W. (**2001**). "Consonant identification under maskers with sinusoidal modulation: Masking release or modulation interference?" J. Acoust. Soc. Am. **110**, 1130–1140.

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.

Nelson, P. B., and Jin, S. H. (**2003**). "Understanding speech in modulated interference: Cochlear implant users and normal-hearing listeners," J. Acoust. Soc. Am. **113**, 961–968.

Nelson, P. B., and Jin, S. H. (**2004**). "Factors affecting speech understanding in gated interference: Cochlear implant users and normal-hearing listeners," J. Acoust. Soc. Am. **115**, 2286–2294.

Oxenham, A. J., and Dau, T. (**2004**). "Masker phase effects in normal-hearing and hearing-impaired listeners: Evidence for peripheral compression at low signal frequencies," J. Acoust. Soc. Am. **116**, 2248–2257.

Oxenham, A. J., and Simonson, A. M. (**2009**). "Masking release for low- and high-passed-filtered speech in the presence of noise and single-talker interference," J. Acoust. Soc. Am. **125**, 457–468.

Peters, R. W., Moore, B. C. J., and Baer, T. (**1998**). "Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people," J. Acoust. Soc. Am. **103**, 577–587.

Qi, Y., and Hillman, R. E. (**1997**). "Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals," J. Acoust. Soc. Am. **102**, 537–543.

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.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.

Stickney, G. S., Zeng, F.-G., Litovsky, R., and Assmann, P. (**2004**). "Cochlear implant speech recognition with speech maskers," J. Acoust. Soc. Am. **116**, 1081–1091.

Stone, M. A., and Moore, B. C. J. (**2007**). "Quantifying the effects of fast-acting compression on the envelope of speech," J. Acoust. Soc. Am. **121**, 1654–1664.

Stone, M. A., Füllgrabe, C., Mackinnon, R. C., and Moore, B. C. J. (**2011**). "The importance for speech intelligibility of random fluctuations in 'steady' background noise," J. Acoust. Soc. Am. **130**, 2874–2881

Stone, M. A., Füllgrabe, C., and Moore, B. C. J. (**2012**). "Notionally steady background noise acts primarily as a modulation masker of speech," J. Acoust. Soc. Am. **132**, 317–326.

Studebaker, G. A., and Sherbecoe, R. L. (**2002**). "Intensity-importance functions for bandlimited monosyllabic words," J. Acoust. Soc. Am. **111**, 1422–1436.

Vestergaard, M. D., and Patterson, R. D. (**2009**). "Effects of voicing in the recognition of concurrent syllables (L)," J. Acoust. Soc. Am. **126**, 2860–2863.

Whitmal, N. A., Poissant, S. F., Freyman, R. L., Helfer, K. S. (**2007**). "Speech intelligibility in cochlear implant simulations: Effects of carrier type, interfering noise, and subject experience," J. Acoust. Soc. Am. **122**, 2376–2388.