

Comodulation masking release in speech identification with real and simulated cochlear-implant hearing

Antje Ihlefeld^{a)}

MRC Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge CB2 7EF, United Kingdom

Barbara G. Shinn-Cunningham

Hearing Research Center, Boston University, 677 Beacon Street, Boston, Massachusetts 02215

Robert P. Carlyon

MRC Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge CB2 7EF, United Kingdom

(Received 13 April 2011; revised 21 December 2011; accepted 22 December 2011)

For normal-hearing (NH) listeners, masker energy outside the spectral region of a target signal can improve target detection and identification, a phenomenon referred to as comodulation masking release (CMR). This study examined whether, for cochlear implant (CI) listeners and for NH listeners presented with a “noise vocoded” CI simulation, speech identification in modulated noise is improved by a co-modulated flanking band. In Experiment 1, NH listeners identified noise-vocoded speech in a background of on-target noise with or without a flanking narrow band of noise outside the spectral region of the target. The on-target noise and flanker were either 16-Hz square-wave modulated with the same phase or were unmodulated; the speech was taken from a closed-set corpus. Performance was better in modulated than in unmodulated noise, and this difference was slightly greater when the comodulated flanker was present, consistent with a small CMR of about 1.7 dB for noise-vocoded speech. Experiment 2, which tested CI listeners using the same speech materials, found no advantage for modulated versus unmodulated maskers and no CMR. Thus although NH listeners can benefit from CMR even for speech signals with reduced spectro-temporal detail, no CMR was observed for CI users.

© 2012 Acoustical Society of America. [DOI: 10.1121/1.3676701]

PACS number(s): 43.66.Dc, 43.66.Ts, 43.66.Ba [EB]

Pages: 1315–1324

I. INTRODUCTION

Cochlear implant (CI) listeners struggle when listening to speech in an acoustically crowded environment. One aspect of this perceptual disadvantage is a much reduced ability to listen in the dips of a fluctuating masker (Nelson *et al.*, 2003; Nelson and Jin, 2004; Stickney *et al.*, 2004; Fu and Nogaki, 2005). A possible explanation for this impairment in “dip-listening” is that CI listeners cannot access segregation cues, such as differences in fine temporal and/or spectral detail between a target and masker, to determine when a dip in the masker occurs and the instantaneous signal is dominated by target energy. To the extent that difficulties in sound segregation contribute to impairments in dip-listening, CI listeners may benefit more from dips if salient perceptual differences between target and masker are introduced. Indeed additional segregation cues could be useful for CI listeners even when they do not help much in similar acoustic settings for normal hearing (NH) listeners, where such cues may be redundant.

In a previous article, we showed that when speech was “noise vocoded” to simulate the loss of spectral and temporal detail experienced by CI listeners, adding spatial cues to perceptually differentiate target and masker did not, in fact, enable NH listeners to benefit from the temporary improvements

in the signal-to-noise-ratios that occur in the dips of a modulated masker (Ihlefeld *et al.*, 2010). This failure to listen in the dips may have arisen because the spectro-temporal structure of the noise-vocoded speech was impoverished. This may have degraded binaural cues important to spatial perception (including interaural differences in both the temporal fine structure and in the envelopes of the resulting signals) to the point that the target and masker were not sufficiently different, perceptually, to allow dip-listening.

Here we investigate another possible source of perceptual information that could perceptually differentiate target and masker when their spectra do not overlap completely. When a tone is masked by a modulated noise, detection thresholds can be reduced by adding another band of noise that is spectrally remote from the signal but modulated coherently with the on-frequency masker (Hall *et al.*, 1984). At least for NH listeners, this “comodulation masking release (CMR)” has also been demonstrated for speech tasks with noise interference. For NH listeners performing tasks with narrowband filtered speech in noise, CMR can improve speech detection thresholds by 11 dB and speech discrimination thresholds by 6–7 dB in a two-alternative forced choice vowel discrimination task (Grose and Hall, 1992) but is modestly sized or absent for consonant identification (Grose and Hall, 1992; Kwon, 2002) and open set speech identification (Grose and Hall, 1992; Festen, 1993; Buss *et al.*, 2003). Moreover, vocode simulations of CI processing suggest that CMR may improve tone detection in CI listeners (Pierzycki

^{a)}Author to whom correspondence should be addressed. Electronic mail: ai33@nyu.edu

and Seeber, 2010). As far as we know, no study has investigated mechanisms of CMR for speech processed to simulate CI hearing or with CI listeners.

CMR can emerge both through within- and across-channel mechanisms (e.g., Schooneveldt and Moore, 1987). CMR from within-channel cues, mediated by a change in temporal envelope within a filter band, may have its origins in peripheral processing stages such as the cochlear nucleus (e.g., Pressnitzer *et al.*, 2001). Various explanations of across-channel CMR have been proposed, all of which build on the idea that the temporal correlation between the flanker band masker and the original masker makes it easier for listeners to identify local temporal minima in the masker (improving the ability to listen in the dips) and/or to “null out” the masker, e.g., using an equalization and cancellation mechanism (for a review, see e.g., Verhey *et al.*, 2003). Such across-channel mechanisms contribute to CMR specifically in conditions when a target and masker are poorly segregated and a flanking masker enhances perceptual segregation (Dau *et al.*, 2009). Considering that CI listeners typically show little or no release from masking when speech target and noise masker are spectrally matched (Nelson *et al.*, 2003; Nelson and Jin, 2004; Stickney *et al.*, 2004; Fu and Nogaki, 2005), within-channel mechanisms are unlikely to provide CMR benefits to CI listeners. However, across-frequency processing of temporal envelopes may aid in auditory scene analysis, leading to CMR in CI listeners. Across-frequency envelope processing may be particularly useful in CI listeners, given that CI speech perception relies heavily on temporal envelopes.

The aim of the current study was to measure CMR under conditions in which the target speech has reduced spectro-temporal detail as is the case for CI listeners. We examined whether a flanking band of coherently modulated masker energy outside the spectral range of target speech can enhance simulated and real CI speech perception in the presence of modulated background noise. In Experiment 1, we measured CMR when NH listeners identified noise-vocoded speech in noise, simulating some aspects of CI listening. Experiment 2 tested CI listeners with speech stimuli in a background of noise. The noise masker was either modulated or unmodulated.

To estimate listeners’ ability to take advantage of the information in the dips of the masker, we calculated the difference in performance between modulated and unmodulated conditions (the modulated-unmodulated difference, or MUD; Carlyon *et al.*, 1989). For unmodulated noise, we expected performance to be limited primarily by energetic masking; moreover, without coherent across-frequency modulation, no CMR can be obtained. Therefore we predicted that speech identification performance would be unaffected by the presence of an additional flanking masker band for unmodulated maskers. For speech in on-target modulated noise, we expected performance to be better than baseline performance in unmodulated noise (i.e., we expected to find a positive MUD). Finally, we were interested in whether the size of the MUD depended on the presence of a flanking masker. An off-band coherently modulated masker might help listeners identify masker dips, improving target-masker segregation and increasing the MUD. This would be a CMR.

II. EXPERIMENT 1

A. Listeners

Eight NH native speakers of British English (ages 18–21) were paid to participate. All had normal-hearing pure-tone thresholds between 250 Hz and 8 kHz as determined by a standard audiometric screening. All listeners gave written informed consent prior to each session, according to the guidelines of the Medical Research Council, Cognition and Brain Sciences Unit.

B. Stimuli

All stimuli consisted of a speech target and noise masker in the right ear, and silence in the left ear. All stimuli were processed using MATLAB R2007a (The Mathworks Inc., Natick, MA).

1. Vocoded speech targets

Speech stimuli were derived from a recording of the Coordinate Response Measure corpus with British talkers (CRM, see Bolia *et al.*, 2000; Kitterick and Summerfield, 2007), with a sampling frequency of 44.1 kHz. Sentences were of the form “ready <call sign>, go to <color> <number> now.” <Color> was one of the set [white, red, blue, and green]. <Number> was one of the digits between one and eight. <Call sign> was one of [arrow, baron, charlie, eagle, hopper, laker, ringo, and tiger]. Only utterances from the four male talkers in the corpus were used. Utterances were time-windowed at the beginning and end of each recording with 2-ms squared cosine windows. Each utterance was then processed by 10 band-pass filters (2nd-order Butterworth, which have a 12-dB-per-octave roll-off), the center frequencies of which are spaced approximately linearly along the cochlea, according to the formula proposed by Greenwood (1990): 3-dB down points were at 200–272, 272–359, 359–464, 464–591, 591–745, 745–931, 931–1155, 1155–1426, 1426–1753, and 1753–2149 Hz (these cut-off frequencies were the same as the lowest 10 bands used in Fu and Nogaki, 2005).¹ The envelope of each narrow band of speech was extracted by half-wave rectification, followed by 50-Hz low-pass filtering. Each envelope was multiplied by a white noise carrier signal and filtered by the band-pass filter corresponding to its spectral band. The resulting amplitude-modulated narrow-band noises were summed, generating noise-vocoded speech. The broadband root-mean square (RMS) was equalized across all processed utterances.

2. Noise maskers

To generate on-target noise, all unprocessed speech utterances were equalized in RMS, summed, and padded with zeros to a length of 8 s. The fast Fourier transform (FFT) of the sum was then computed. For each token of noise, the FFT phase spectrum of the sum was replaced with a draw from a uniform random distribution. The result was inverse-FFT transformed to produce a token of on-target masking noise.

Flanking noise tokens were generated by processing random tokens of 8-s long uniformly distributed white noise

with a steep high-pass filter (30th-order Butterworth filter, 3-dB down point at 6 kHz) followed by a low-pass filter (6th-order Butterworth filter, 3-dB down point at 10 kHz). There were 100 different frozen tokens of on-target noise and 100 different frozen tokens of flanking noise.

The masker always started to play 250 ms before the onset of the target utterance and stopped 4.535 s after the onset of the target. To generate the masker token on each trial, a random on-target noise token was selected as the masker. That noise token was then time-windowed with a rectangular window beginning at a random start position within the token. On the trials with a flanking noise, a token of flanking noise was selected and added to the on-target noise, subject to the same rectangular time windowing as the on-target noise.

These two masker types were presented both in unmodulated and in modulated conditions. To make unmodulated masker tokens, on-target and flanking noise tokens were time-windowed at the beginning and end of each recording (2-ms squared-cosine windows). To generate modulated noise maskers, both on-target and flanking tokens were square-wave modulated with 2-ms cosine-squared windows at a rate of 16 Hz with the same starting phase, 50% duty cycle, and 100% modulation depth. In the co-modulated conditions, the envelope of the modulated flanking noise was the same as that of the modulated on-target noise. The modulation frequency of 16 Hz corresponds to a masker cycle length of 62.5 ms, a duration that is shorter than half of the typical duration of the target keywords in this study. The modulated masker was then scaled such that its RMS equaled that of the corresponding token of the unmodulated masker. As a result, the peak level of the modulated noise was 3 dB higher than that of the unmodulated noise.

Figure 1 shows the magnitudes of the Fourier transforms of an example target and masker combination, with a target-to-masker broadband energy ratio (TMR) of -10 dB. The spectral shape of the long-term averaged on-target noise masker does not perfectly match the spectral shape of the vocoded target speech. This is a consequence of the stimulus processing, as the on-target noise was generated based on

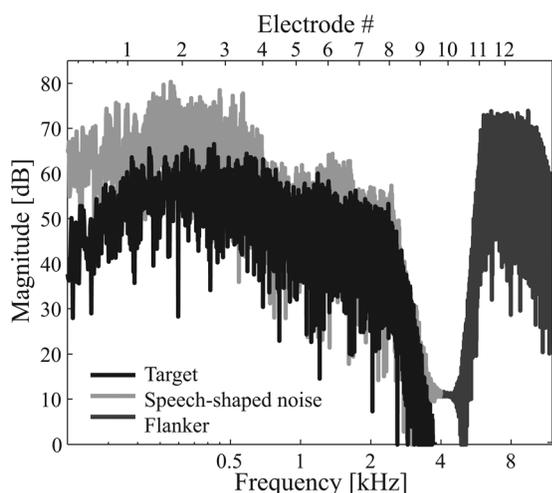


FIG. 1. Magnitudes of the Fourier transforms of an example noise-vocoded target (black), on-target masker (dark gray), and flanking noise band (light gray) at -10 dB TMR.

TABLE I. Actual narrowband TMRs for a nominal broadband TMR of -10 dB.

Center frequency (Hz)	233	312	408	524	664	833	1037	1283	1581	1941
Actual TMR (dB)	-13	-12	-11	-10	-5	-1	-3	-5	-5	-1

unprocessed speech rather than vocoded speech. Nevertheless, the two stimuli have the same overall bandwidth, and, in both cases, the spectrum level drops smoothly with increasing frequency above 0.5 kHz. Table I lists the narrowband TMRs for a nominal TMR of -10 dB, averaged across all stimuli, calculated as the dB difference between the speech target and on-target noise in each of the 10 pass-bands of the target speech. This analysis shows that although the within-band TMRs deviated from the nominal TMR, from band to band these deviations were modest, producing TMRs that were greater than the nominal TMR in higher frequency bands and slightly lower TMRs in the lower-frequency bands.

C. Procedures

1. Stimulus presentation

Stimuli were D/A converted with a sound card (Turtle Beach Sonic Fury; 16 bit resolution, 44.1 kHz sampling frequency) and amplified using a programmable attenuator (TDT PA4), then played through a headphone buffer (TDT HB6). Stimuli were presented over Sennheiser HD 650 headphones to the right ear of a listener seated in a double-walled sound-treated booth. Following each trial, listeners indicated perceived target keywords using a graphical user interface (GUI), after which the GUI indicated the correct response.

The on-target noise maskers were always presented at 67 dB SPL. In the flanking band conditions, an additional flanking band was presented at an RMS that was 10 dB above the RMS of the on-target noise alone, a value selected during pilot listening to create a perceptual difference between speech-shaped and speech-shaped plus flanking band noises. The target was presented at levels of 43, 50, or 57 dB SPL. In the on-target frequency band from 200 to 2149 Hz, this resulted in target to masker energy ratios (TMRs) of -24 , -17 , and -10 dB, respectively.

Listeners completed three sessions lasting 2 h each. Not more than one session was collected for each listener on a single day. At the beginning of each session, listeners completed one block of listening to vocoded speech stimuli in quiet. Ten test blocks followed.

2. Blocking and task

Each test block consisted of 48 trials and contained a fixed masker type: either on-target noise alone or on-target noise plus flanking band noise. Masker types alternated between blocks with an order randomized across listeners. Within a block, modulated and unmodulated maskers and target level were randomly interleaved from trial to trial, such that each of the six combinations of target level and masker modulation condition was presented once before any was repeated; each time the six combinations were repeated,

they were in a new random order. Within a block, each combination was presented eight times. Within each session, listeners always performed an equal number of trials for each experimental condition. Over the course of the experiment, each listener completed 120 trials for each combination of masker type, modulation condition, and target level.

In all experiments in this study, listeners were instructed: "Please report the color and number." The task was a 32-alternative, forced-choice, closed-set speech identification (4 colors \times 8 numbers). Correct-answer feedback was provided after each trial. A trial was scored as correct and listeners were given feedback that they were correct if and only if they correctly reported both target keywords (color and number).

D. Results

Percent correct for each of the eight listeners was calculated separately for each noise condition as a function of target level. They were then converted into logit scores to reduce floor and ceiling effects and to conform better to the heteroscedasticity assumption of analysis of variance (ANOVA). The logit transform for percent correct p is $\text{logit}(p) = \ln[p/(100-p)]$, transforming the percent correct scores into a variable of theoretically infinite range. Chance performance was 3%. To avoid undefined values of the logit transform and to avoid floor and ceiling, $p < 3\%$ was set to 3%, and $p > 99\%$ was set to 99%.

Figure 2(A) shows the logit scores on the left ordinate and the percentage of correct responses on the right ordinate. All listeners showed similar trends, so only the across-listener mean performance is shown. Error bars, where large enough to be visible, show the 95% confidence interval around the mean. The standard error of the mean across listeners was computed and the between-listeners variance was removed by subtracting the grand mean from each listener's mean and using the resulting difference (Loftus and Masson, 1994). Estimates of the 95% confidence intervals were calculated as t_{α} times this corrected standard error, where t_{α}

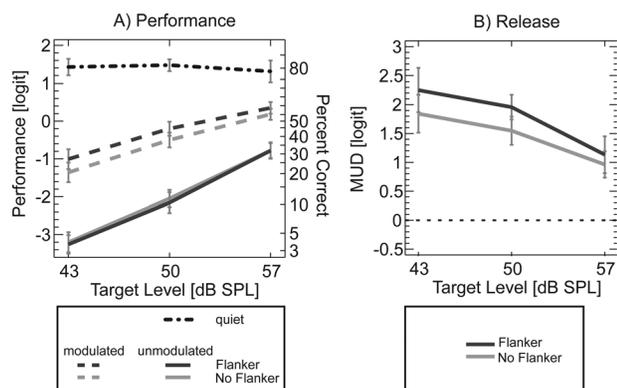


FIG. 2. (A) Mean performance in unmodulated versus modulated noise (solid versus dashed lines, respectively) and for on-target and on-target plus flanking band noise maskers (light versus dark gray lines, respectively) of Experiment 1. The dotted-dashed line shows performance in quiet. Left ordinate shows performance in logit units, right ordinate in percent correct. (B) Mean MUD in logit units. Error bars show 95% confidence intervals of the across-listener mean, after subtracting between-listeners variance. Note that (A) and (B) do not use the same ordinate scaling.

is the t -value corresponding to $\alpha = 5\%$ confidence level (cf., Loftus and Masson, 1994).

Percent correct in quiet did not change appreciably with target level. Averaged across listeners and target levels, quiet performance equaled 78% (black dotted-dashed line). When a masker was present, performance was generally worse than in quiet [all other lines fall below the dotted black line in Fig. 2(A)]. Performance was better in modulated than in unmodulated noise (compare dashed and solid lines). In unmodulated noise, performance was similar whether or not a flanking noise band was present (dark and light gray solid lines fall on top of each other). Performance was better in modulated noise when a flanking noise band was present than when the masker consisted of on-target noise only (dark gray dashed line falls above light gray dashed line). Repeated measures ANOVA with main factors of masker condition, modulation type, and target level found significant main effects of masker condition, modulation type, and target level [$F(1,7) = 7.392, 203.931, 153.056$; $P < 0.0001, P = 0.03, P < 0.0001$, respectively] and, importantly, a significant interaction between masker condition and modulation type [$F(1,7) = 12.221$; $P = 0.01$]. At 43 dB SPL target level, or -24 dB TMR, performance was at chance for both unmodulated masker configurations.

The modulation masking release, or MUD, defined here as the difference in logit units of performance with the modulated masker minus performance with the unmodulated masker, was consistently greater for on-target plus flanking noise maskers than for the on-target masker alone [dark gray line falls above light gray line in Fig. 2(B)]. The average MUD equaled 1.5 logit-units (22.5%) for on-target noise alone, and 1.8 logit-units (28.5%) for modulated noise. The MUD decreased with increasing target level. A repeated measures two-way ANOVA showed a significant difference in MUD between these two conditions [$F(1,7) = 12.21, P = 0.01$], indicating CMR, and a significant effect of target level [$F(2,14) = 13.52, P = 0.001$].

In addition to analyzing MUD expressed as the vertical difference in performance, it is instructive to consider the horizontal shift between the modulated and unmodulated performance curves. To this end, logit-transformed percentage correct scores as a function of TMR were fitted with lines using a minimum least squares method (command `polyfit` in MATLAB 7.4.0). Table II lists across-listener averages of the intercepts and slopes of these fits. For all stimulus conditions, the R^2 correlation between the original data and the fit was close to 1. Specifically, across-listener average R equaled 0.98, 0.97, 0.98, and 0.95 for the conditions unmodulated without flanker, unmodulated with flanker, modulated without flanker, and modulated with flanker, respectively. Given these strong correlations, the line fits were deemed appropriate summaries of the results. A repeated measures ANOVA with factors of masker type and modulation condition revealed that slopes were significantly steeper in unmodulated than in modulated conditions but found no statistical difference in slope between the masker types [$F(1,7) = 18.965, 0.248$; $P = 0.03, P = 0.643$ for modulation type, and masker condition, respectively]. Because of the differences in slopes, the horizontal shift in decibels between the

TABLE II. Across listener averages of the intercepts and slopes of the line fits relating the MUD to the TMR; 95% confidence intervals are listed in parentheses.

Experiment 1. Slope (logit/dB)	
Masker: unmodulated noise	0.17 (0.02)
Modulated noise	0.11 (0.02)
Unmodulated noise with flanking band	0.18 (0.02)
Modulated noise with flanking band	0.10 (0.02)
Experiment 1. Intercept (logit)	
Masker: unmodulated noise	0.93 (0.48)
Modulated noise	1.31 (0.36)
Unmodulated noise with flanking band	0.94 (0.43)
Modulated noise with flanking band	1.36 (0.37)
Experiment 2. Slope (logit/dB)	
Masker: unmodulated noise	0.19 (0.04)
Modulated noise	0.29 (0.21)
Unmodulated noise with flanking band	0.18 (0.04)
Modulated noise with flanking band	0.22 (0.14)
Experiment 2. Intercept (logit)	
Masker: unmodulated noise	-0.5 (0.56)
Modulated noise	-0.63 (0.63)
Unmodulated noise with flanking band	-0.22 (0.44)
Modulated noise with flanking band	-0.26 (0.40)

modulated and unmodulated performance curves varied somewhat as a function of target level. When evaluated near 50% correct, the decibel difference between line fits of the logit-transformed modulated and unmodulated performances was, on average, 7.2 dB in the on-target noise conditions and 8.8 dB in the flanking-band conditions.

To summarize, CMR was estimated in two ways: (1) by calculating the grand average “vertical” difference in MUD between the on-target noise alone and the on-target noise with flanking noise conditions; in this case, it was 6% (0.3 logit-units); and (2) by calculating the “horizontal” difference between the line fits in the modulated noise conditions, estimated at 50% correct; in this case, it was 1.7 dB.

III. EXPERIMENT 2

A. Rationale

Experiment 1 showed that a small but significant CMR could be obtained in NH listeners presented with speech stimuli processed to simulate some aspects of CI processing. Experiment 2 explored whether CMR could be obtained with real CI listeners.

B. Listeners

Seven post-lingually deafened users of the MED-EL PULSAR cochlear implant took part. All gave written informed consent prior to each session and were paid for their time. Table III lists details of their etiology. All CI listeners utilized fine structure processing (FSP) speech coding strategies. All hearing loss was post-lingual and of sensorineural origin.

The amount of experimental data that could be collected varied across listeners due to across-listener differences in

TABLE III. Etiology.

CI listener	1	2	3	4	5	6	7
Device	Pulsar	Pulsar	Pulsar	Pulsar	Sonata	Pulsar	Sonata
Age	73	44	70	54	74	60	69
Ear	Right	Left	Left	Left	Right	Right	Left
Duration of CI use (month)	27	21	18	12	9	21	9
TMR _{INP} at C-level	25	15	15	15	15	15	15

speed of response and variations in the amount of time available for each listener. Listeners took part in one to three sessions of about 3 h duration each including breaks, making it desirable to shorten the amount of experimental time. Therefore unlike the method of constant stimuli used in Experiment 1, Experiment 2 measured speech identification thresholds using adaptive tracking.

C. Stimuli

All stimuli consisted of a speech target and noise masker, generated with MATLAB R2007a prior to the experiment. Speech utterances were taken from the same speech corpus as in Experiment 1. Utterances were low-pass filtered with a 7th order Butterworth filter with a cut-off frequency at 2627 Hz,² and equalized in RMS. On-target and flanking band masker noise tokens were identical to those used in Experiment 1. The top axis of Fig. 1 shows the electrode numbers corresponding to the center of each band-pass filter in the listeners’ CI, as determined by the clinical maps for these MedE1 patients.

D. Procedures

1. Stimulus presentation

Using a clinical software tool provided by MedEL, prior to each session and for each listener, the listener’s map with everyday settings was programmed onto a PULSAR speech processor. Spectral input ranges for each of the 12 electrodes, listed in Table IV, were similar across listeners. At the beginning of each psychometric track, the experimenter switched the speech processor between two different programs, depending on the experimental condition. (1) In the on-target noise and no flanker conditions, electrodes 10-12 were silenced by setting the M- and T-levels of these electrodes to zero. Oscilloscope calibration confirmed that this setting resulted in absence of electrical output from these electrodes, while leaving the outputs of all other electrodes intact at their original “everyday” values. (2) In the conditions with the flanking-band masker, only electrode 10 was

TABLE IV. Center frequencies of the acoustic inputs for each of the 12 electrodes.

Electrode Number	1	2	3	4	5	6	7	8	9	10	11	12
Center Frequency (Hz)	149	262	409	602	851	1183	1632	2228	3064	4085	5656	7352

silenced. Except for electrode 10, then, the settings for all electrodes were equivalent to those of the CI listener in everyday life.

Note that we presented the flanker in all conditions and silenced electrodes when the flanker was not desired rather than removing the flanker band from the input in some conditions. This was done so as to avoid complications associated with the addition of flanker energy influencing the response of electrodes responding to the “on frequency” masker and target, via either the automatic gain control (AGC) or the input-output function of the speech processor. In particular, because the AGC operates on the broadband input, energy remote from the cutoff frequencies of a filter assigned to a CI electrode can influence the output of that electrode in several ways. For instance, if the AGC has a knee point, then adding remote energy can put the overall input above that knee point. This would reduce the modulation depth at the output of all electrodes, including those not nominally tuned to the frequency content of the added energy. Our solution was to present the flanker stimulus in all conditions and to disable the basal electrodes in those conditions where we did not want the subject to hear the flanker. As a result, activation of the “on target” electrodes was independent of the presence or absence of the flanker.

Stimuli were played out from the laptop sound card into the auxiliary input of the research speech processor, implementing the FSP speech coding strategy. Masker conditions and types were similar to those in Experiment 1 with two types of masker noise (on-target noise or on-target plus flanking noise) presented in two conditions (unmodulated or 16-Hz square wave modulated). In addition, two masking controls were presented. The first of these additional controls used an on-target masker that was 16-Hz square wave modulated with a flanking band that was unmodulated. The second of these controls presented a masker consisting of an unmodulated on-target noise masker with a 16-Hz square wave-modulated flanking noise band. These controls allowed us to measure whether adding a flanking band affected performance in ways that did not require an across-frequency combination of masker envelope information.

2. Loudness rating

At the beginning of each session, each listener performed two sets of loudness calibrations. The first set of measurements was intended to both establish a comfortable loudness setting for speech in quiet and to familiarize the listener with the type of sentence material they were going to hear throughout the session. The experimenter repeatedly played the utterance “ready baron go to blue one now,” spoken by the same talker in quiet. With each repetition, the output voltage through the sound card was gradually increased. Monaural thresholds (T_{INP}) were measured when the listener indicated they could detect a soft ongoing sound. Maximally comfortable levels (C_{INP})³ for speech in quiet were then measured by playing the speech at a gradually increasing intensity until the listener rated it very loud but not yet uncomfortable and then gradually decreasing intensity until the listener rated it comfortably loud.

The second set of loudness measurements determined the maximal level at which a mixture of speech and noise could be played. Using the same methodology as in the first set of loudness measurements, T_{INP} and C_{INP} were newly measured for the case where the on-target and flanking band noises were both unmodulated and were mixed with the speech. The broadband target-to-masker RMS ratio at the input to the speech processor (TMR_{INP}) was set to +15 dB. Afterward, the experimenter played the sounds again at C_{INP} with +15 dB TMR_{INP} and asked the listener what they thought the sound was. All listeners except listener CI-1 reported that they could hear a voice and noise. Listener CI-1 reported hearing noise but did not report hearing a voice. Therefore for this listener, T_{INP} and C_{INP} of the mixture were also measured at +25 dB TMR_{INP} . At C_{INP} with +25 dB TMR_{INP} , CI-1 could hear both noise and a voice. T_{INP} and C_{INP} were then also measured for a mixture of speech and 16-Hz modulated on-target plus flanking band noise, at +15 TMR_{INP} for CI-2 through CI-7 and at +25 dB TMR_{INP} for CI-1.

Throughout the remainder of the session, masker levels were conservatively fixed at the smaller of the two C_{INP} obtained during the second set of loudness measurements. Moreover, relative to the masker level, target levels were never played louder than +15 dB TMR_{INP} for listeners CI-2 through CI-7 or +25 dB TMR_{INP} for CI-1.

3. Quiet listening

To further familiarize the listeners with the task, each listener performed between one and four practice runs of 50 trials of identifying target speech in quiet at a fairly soft level until achieving at least 70% accuracy. In these practice runs, speech was played at the target level that, if the target-masker mixture had been present at C_{INP} , would have equaled 0 dB TMR_{INP} (+10 dB TMR_{INP} for CI-1).

4. Speech reception thresholds

Speech reception thresholds were measured using 1-up-1-down adaptive tracking in blocks of six tracks each. Tracks started with the target speech played at +10 dB TMR_{INP} (+20 dB TMR_{INP} for CI-1). The masker was played at its fixed C_{INP} . When the listener reported both target color and number correctly, the target level was decreased by 5 dB unless the resulting target level would have fallen below the T_{INP} , in which case it was left unaltered (note that this never actually happened during the experiments). If the listener reported one or both keywords incorrectly, the target level was increased by 5 dB, subject to the constraint that the resulting TMR_{INP} was set less than or equal to the safe maximum of +15 or +25 TMR_{INP} . After four reversals, the step size was decreased to 2.5 dB. Twenty reversals were collected. The average of the final 12 reversals was used to estimate the speech reception threshold.

Within each block, the six possible maskers were presented one after the other in random order, randomized individually for each listener. After a measurement block of six thresholds was completed, another block was offered to the listener until the experimental time was over. Listeners were

TABLE V. Percent correct performance in quiet.

CI listener	1	2	3	4	5	6	7
Percentage correct	86	72	71	89	97	87	93

encouraged to take breaks often and not to tire themselves. Listeners completed between three and six tracks per condition.

E. Results

Table V lists the percent correct performance for each listener from the last practice run in quiet, confirming that when no masker was present and when the target was relatively soft, CI listeners were clearly and consistently performing better than 50% correct. Average performance in quiet was 85% correct, similar to or, if anything, slightly better than the average performance for NH listeners in Experiment 1, although the difference was not significant [unpaired samples t -test, $df = 12.9$, $t = -1.2$, $p = 0.25$]. For each of the six masker conditions, thresholds for each of the eight CI listeners were calculated separately. Figures 3(A) to 3(G) show TMR_{INP} at threshold for each of the masker conditions and each of the listeners. Error bars show the 95% confidence interval around the mean across tracks (t_{α} times the standard error of the mean across adaptive track thresholds; note that error bars are absent for CI-7 in the control conditions, where only one set of measurements was obtained). Figure 3(H) shows the across-listener average TMR_{INP} (error bars show 95% confidence interval of the mean across listeners after removing the between-listeners variance).

Overall, thresholds varied substantially and consistently across listeners, such that listeners who had low thresholds in one condition typically had low thresholds in all other conditions. This across-subject consistency was confirmed by Kendall's coefficient of concordance, which equaled $W = 0.9688$ (Kendall and Babington Smith, 1939). While

there were consistent across-listener differences, there were no consistent effects across conditions. Specifically, in the absence of a flanking band, there was no consistent difference between thresholds obtained with modulated versus unmodulated maskers; in other words, there was no MUD (compare two left-most points for each subject, square symbols). Furthermore, as shown by the points 3rd and 4th from the left for each listener (circles), there was no MUD even when a flanking band was present. Hence our results show no evidence for CMR in CI listeners. One unusual data point can be observed for listener CI-1, whose threshold in the modulated no-flanker condition is greater than in all other conditions. This finding is consistent with the masker modulation interfering with the processing of modulations present in the speech signal (Kwon and Turner, 2001). Although this effect was alleviated by adding a modulated flanking band [compare open square and circle in Fig. 3(A)], this improvement cannot be attributed to CMR; thresholds were also reduced by adding an unmodulated flanker [right-pointing triangle at far right of Fig. 3(A)]. Consistent with the preceding descriptions, repeated-measures ANOVA on the thresholds of listeners CI-2 through CI-7 found no significant effect of modulation type or masker condition [$F(1,5) = 1.26$, 0.01 ; $P = 0.312$, 0.901 , for masker type and modulation condition, respectively], ignoring control conditions, and ignoring results from CI-1.

Figure 4 plots the modulation masking release for the individual listeners, defined here as the difference in TMR at threshold with the modulated masker minus TMR at threshold with the unmodulated masker (MUD_{TMR}). The modulation masking release for each listener is plotted for the no-flanker condition on the horizontal axis and for the flanker condition on the vertical axis. To the extent that CMR improved performance in the conditions with a flanking-band masker, the symbols in Fig. 4 should fall above the diagonal. Although this was true for the NH listeners of experiment 1 (gray "+" symbols), for the CI listeners (black symbols), data cluster around the diagonal. The MUD_{TMR}

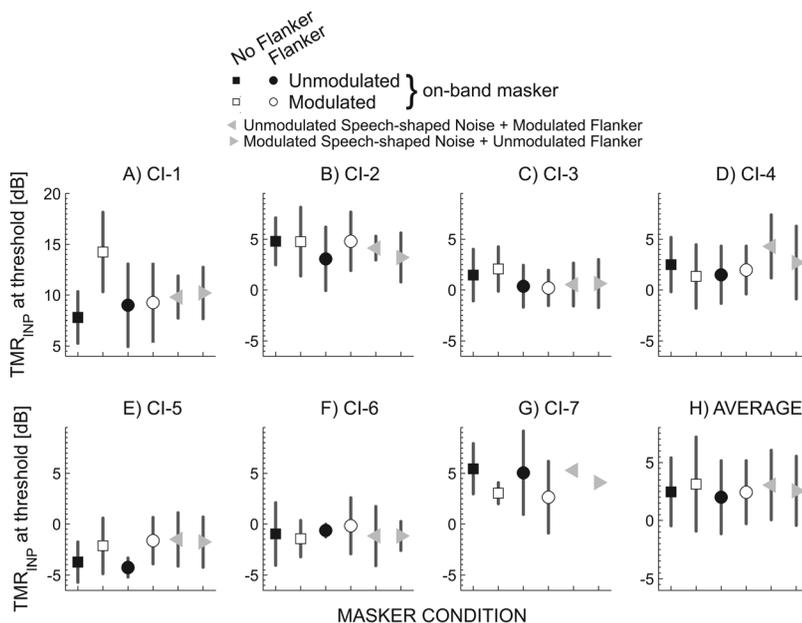


FIG. 3. Thresholds for the six different masker types for individual CI listeners in Experiment 2, except bottom right panel, which shows the mean across CI listeners. Note that the abscissa in the left upper panel ranges from 6 to 20 dB, but is 6 to 8 dB TMR_{INP} in all other panels. Error bars show 95% confidence intervals of the mean across adaptive tracks (A)–(G) or of the mean across CI listeners (H).

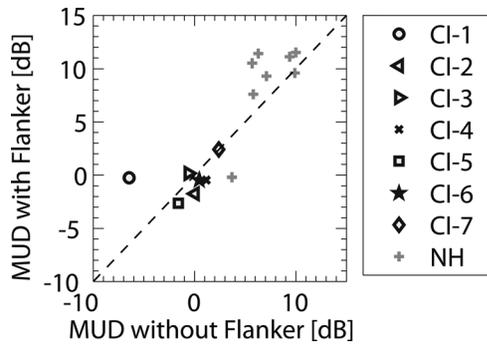


FIG. 4. MUDs for CI listeners (with each black symbol denoting a different CI listener), and for NH listeners for line fits evaluated at 50% correct (gray + symbols).

for CI listeners CI-2 through CI-7 was not statistically different between the on-target and the on-target plus flanking noise conditions, as confirmed by repeated measures ANOVA [$F(1,5) = 3.68$, $P = 0.113$]. Moreover, for both masker types, MUD_{TMR} was not statistically different from zero evaluated with two-tailed t -tests [$t = -0.548$, 0.646 ; $df = 5$; $P = 0.607$ and 0.646 for on-target and on-target plus flanking noise conditions, respectively].

To consider information from the overall shapes of the psychometric functions (which were ignored in the planned comparisons of thresholds), for each listener, all trials from the adaptive tracks were pooled and fitted with cumulative Gaussian functions via a bootstrapping algorithm; these results were logit-transformed. Table II lists across-listener averages of the intercepts and slopes of these psychometric fits. Consistent with the analysis of raw thresholds, this analysis method revealed no consistent differences in slopes between modulated and unmodulated conditions or between the two masker types.

The fact that neither the MUD nor the CMR was significant for CI listeners raises the question of whether no effect exists, or whether any effect was obscured by response variability. Although it is impossible to answer this question with the current results, one can estimate the maximum size that any “true” effect could have been without reaching statistical significance, based on the across-listener variance observed in the data.⁴ The results of such an analysis suggest that a MUD of 2 dB would have reached significance in a one-tailed test. For a CMR to be significant, it would have to be at least 1 dB.

To summarize, in Experiment 2, speech identification thresholds were not statistically different across the six masker conditions. The modulation masking release observed with vocode simulations with NH listeners was absent for CI listeners, both for conditions with and without the flanking band: Provision of a comodulated flanking band did not help CI listeners extract target information from the dips of a modulated masker.

IV. GENERAL DISCUSSION

A. MUD and CMR for NH listeners hearing noise-vocoded speech

Experiment 1 revealed a robust modulation masking release (MUD) for both on-target noise alone and for on-target

plus flanking band noise. It is interesting to compare the current results, which showed a 1.5 logit-unit/7.2-dB MUD in the on-target condition, with results from our own recent study that tested NH listeners in a similar task with either unprocessed or four-band noise-vocoded speech (Ihfeldt *et al.*, 2010). In that study, near 50% correct performance, the MUD equaled approximately 2 logit-units (40%) or 7.4 dB for the unprocessed conditions and 1 logit-unit (12%) or 2.7 dB for the vocoded conditions. For vocoded speech, the size of the MUD increases with the amount of spectral detail preserved by the processing (e.g., see Fu and Nogaki, 2005). The MUD observed in the current study is comparable to the MUD previously obtained with unprocessed speech in a similar task but is substantially larger than what was found for the vocoded conditions in the previous study. Therefore using a 10-channel vocoder here, the ability of our NH listeners to listen in the dips of the modulated masker was not strongly limited by the amount of spectral detail. Instead lack of spectral detail limited the MUD in our previous study, which used only four bands the center frequencies of which ranged from 100 to 6000 Hz.

Speech identification performance in NH listeners was statistically indistinguishable between unmodulated on-target and unmodulated on-target plus flanking noise, suggesting that the amount of energetic masking within the critical bands of the target speech did not change when the flanking band was added. Moreover, performance in the modulated conditions was better when a modulated flanking band was present than when no flanking band was added. Together, these results demonstrate that additional modulated masker energy outside the spectral range of the target can improve a listener’s ability to exploit information in the dips of the masker when listening to noise-vocoded speech.

Here we found that modulation in the on-target masker improved performance of NH listeners when there was no flanking band masker. Moreover, MUD increased when a flanking masker was present relative to the conditions without flanking band. These findings support the idea that the flanking band reduces masking caused by a speech shaped masker through an across-frequency CMR mechanism. This is consistent with the idea that CMR can produce a modest improvement in the ability to listen selectively in the dips of the masker and/or stream target information present in the dips.

Previous studies of CMR in speech recognition have used band-pass filtered speech presented to NH listeners (Grose and Hall, 1992; Kwon, 2002; Buss *et al.*, 2003). In one previous study, when stimuli consisted of either band-pass filtered words with six possible forced-choice responses in each trial or open-set sentence material, no CMR was observed (Grose and Hall, 1992). In another study, where listeners were asked to identify intervocalic consonant syllables and performance was measured in a range between 40% and 65% of overall percent correct performance levels, CMR caused less than 4% of improvement in performance (Kwon, 2002). In a third study, when target speech material consisted of spondees with either two or four forced-choice alternatives, CMR was about 3 dB for a two-alternative task or 1 dB for a four-alternative task (Buss *et al.*, 2003). However, when the target speech material in that study was open-set,

modulating the masker *interfered* with performance by 3 dB. Hence the 1.5 dB or 6% CMR we found is in line or even larger than has been observed in past studies.

CMR is often greater for lower-complexity tasks where performance relies on a relatively small set of cues (like detection tasks or forced-choice identification tasks with a small set of possible responses) than for tasks that require finer stimulus analysis, like open-set identification tasks (Buss *et al.*, 2003). Here, although the response set was limited to 32 choices, the number of response choices was greater than in previously published CMR studies using closed-set stimuli. Thus the fact that the NH listeners in our study showed relatively large CMR is not simply due to low task complexity.

B. MUD and CMR in CI listeners

Experiment 2 failed to find evidence for either a statistically significant MUD or CMR in CI listeners. Previous studies suggest that MUD grows non-monotonically with decreasing overall performance level and with decreasing signal level, with maximal release occurring at lower signal to noise ratios where speech is only partially audible and not perfectly intelligible (Gnansia *et al.*, 2008; Oxenham and Simonson, 2009; Bernstein and Grant, 2009). Considering that here thresholds for CI listeners were at positive TMRs, the fact that no MUD was observed in Experiment 2 is consistent with previous findings (cf., Bernstein and Grant, 2009).

It is possible that the 10-channel vocoding simulation in Experiment 1 was too finely resolved to realistically simulate CI listening. For NH listeners in quiet, eight channels can suffice for intelligibility scores that are comparable to CI listeners' performance (Dorman *et al.*, 1997; Shannon *et al.*, 1995). We should point out that Experiment 2 differed from Experiment 1 both in the use of an adaptive procedure rather than the method of constant stimuli and in the age of the listeners and not just in the hearing status of the listeners. However, in general, the adaptive and constant-stimulus methods have similar accuracy (e.g., Dai, 1995). Similarly, at least for tone detection, CMR is not affected by age (Peters and Hall, 1994). Thus these differences are not expected to cause differences in the results. Nevertheless given these methodological differences, a direct comparison between the two experiments should be undertaken with caution. Instead we discuss several factors that may limit the ability of CI listeners both to listen in the dips of a modulated masker and to use across-frequency co-modulation to aid understanding of speech.

One such factor may be the lack of spatial resolution provided by CI current stimulation. It is well known that there is a substantial spread of excitation along the cochlea with CI stimulation, so the neural representation of the input spectrum is likely to be blurred. Excitation spread has been shown to severely limit the ability of CI listeners to process such normally robust segregation cues as a 200-ms difference between the onsets of stimulation on different electrodes (Carlyon *et al.*, 2007). Moreover, previous simulations of CI listening using noise-vocoded speech have revealed

that the size of the MUD decreases as the number of analysis filters is reduced and/or the slopes of the filter skirts are made shallower (e.g., see Fu and Nogaki, 2005). Spread of excitation may also reduce the ability of our CI listeners to use co-modulation cues, as stimulation due to the target may have excited a neural population that substantially overlaps with the population responding to the flanking band masker.

CI listeners typically gain little or no advantage from modulating the masker (Nelson *et al.*, 2003; Nelson and Jin, 2004; Stickney *et al.*, 2004; Fu and Nogaki, 2005). Indeed CI speech identification can even be impaired by modulation of the masker, presumably due to a form of modulation discrimination interference (Kwon and Turner, 2001; Apoux and Bacon, 2008), a factor that may have counteracted potential CMR.

Finally, it is worth noting that one NH listener showed a pattern of results similar to the CI listeners in that he showed only a small MUD both with and without the flanking band (see the leftmost cross in Fig. 4). Hence it may be that when "listening in the dips" is not possible with an on-frequency band, the presence of additional modulation information provided by a flanking band does not help. Perhaps for a very low modulation rate, CI listeners would both show a non-zero MUD for on-target maskers and then would also show CMR. However, CMR was small even for those NH listeners who showed a substantial MUD and was small for all CI listeners with a 16-Hz modulated masker. Together, results indicate that co-modulation is unlikely to provide a substantial benefit for masker modulation rates encountered in everyday speech.

V. CONCLUSIONS

Experiment 1 showed that CMR improved NH listeners' speech identification performance for vocode simulations of speech in modulated noise. However, the small performance improvement seen for NH listeners was not observed for CI listeners. There is potential for quality difference cues to improve dip-listening in cochlear implants, but here we did not see any improvements.

ACKNOWLEDGMENTS

This research was funded by the Medical Research Council. We would like to thank all listeners for participating in this research project. Dr. John Deeks helped with cochlear implant testing. Two reviewers provided valuable feedback on an earlier version of this manuscript.

¹Note that at 2149 Hz the upper frequency cut-off of the highest filter removed speech information that is important for recognition. However, enough information was preserved to maintain an intelligible signal. Moreover, this choice was motivated by consideration of CI listeners in Experiment 2, where the relatively low upper cut-off frequency weakened the possibility of current spread masking from the flanking masker bands (see Experiment 2 for details).

²We wanted to give our CI listeners access to speech information from at least eight electrodes. Electrode 8 has a center frequency of 2227 Hz, above the highest frequency of 2149 in the NH listeners' stimuli. The filter bank used to create stimuli for the NH listeners consisted of 10 filters that were spaced evenly along the cochlea. The upper cut-off for what would have been the 11th equidistant filter is 2627 Hz. This is what we chose as

upper cut-off for the CI listeners' speech spectrum. To avoid stimulation of electrode 9, we then chose a much steeper filter roll-off than in Experiment 1.

³We use the terms T_{INP} and C_{INP} to distinguish these levels from the terms T_s and C_s , expressed in terms of electrical current, which are commonly used in the CI literature.

⁴For each CI listener and flanking band condition, the mean of the difference between thresholds in modulated and unmodulated noise was estimated via bootstrap sampling with 1000 draws. We then computed the amount of MUD necessary for it to reach statistical significance, assuming the observed across-listener variance and basing the statistical analysis on a t -test with a 5% confidence interval. Analogously, for each listener, the mean in CMR was estimated via bootstrapping; then the across-listener variance was calculated and the smallest difference that should lead to a statistically significant effect was computed.

- Apoux, F., and Bacon, S. P. (2008). "Selectivity of modulation interference for consonant identification in normal-hearing listeners," *J. Acoust. Soc. Am.* **123**, 1665–1672.
- 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.
- Buss, E., Hall, J. W. III, and Grose, J. H. (2003). "Effect of amplitude modulation coherence for masked speech signals filtered into narrow bands," *J. Acoust. Soc. Am.* **113**, 462–467.
- Bolia, R. S., Nelson, W. T., Ericson, M. A., and Simpson, B. D. (2000). "A speech corpus for multitalker communications research," *J. Acoust. Soc. Am.* **107**, 1065–1066.
- Carlyon, R. P., Buus, S., and Florentine, M. (1989). "Comodulation masking release for three types of modulator as a function of modulation rate," *Hear. Res.* **42**, 37–46.
- Carlyon, R. P., Long, C. J., Deeks, J. M., McKay, C. M. (2007). "Concurrent sound segregation in electric and acoustic hearing," *J. Assoc. Res. Otolaryngol.* **8**, 119–133.
- Dai, H. (1995). "On measuring psychometric functions: A comparison of the constant-stimulus and adaptive up-down methods," *J. Acoust. Soc. Am.* **98**, 3135–3139.
- Dau, T., Ewert, S., and Oxenham, A. J. (2009). "Auditory stream formation affects comodulation masking release retroactively," *J. Acoust. Soc. Am.* **125**, 2182–2188.
- Dorman, M. F., Loizou, P. C., and Rainey, D. (1997). "Speech intelligibility as a function of the number of channels of simulation for signal processors using sine-wave and noise-band outputs," *J. Acoust. Soc. Am.* **102**, 2403–2411.
- Festen, J. M. (1993). "Contributions of comodulation masking release and temporal resolution to the speech-reception threshold masked by an interfering voice," *J. Acoust. Soc. Am.* **94**, 1295–1300.
- Fu, Q. J., and Nogaki, G. (2005). "Noise susceptibility of cochlear implant users: The role of spectral resolution and smearing," *J. Assoc. Res. Otolaryngol.* **6**, 19–27.
- Gnansia, D., Jourdes, V., and Lorenzi, C. (2008). "Effect of masker modulation depth on speech masking release," *Hear. Res.* **239**, 60–68.
- Greenwood, D. D. (1990). "A cochlear frequency-position function of several species—29 years later," *J. Acoust. Soc. Am.* **87**, 2592–2695.
- Grose, J. H., and Hall, J. W., III (1992). "Comodulation masking release for speech stimuli," *J. Acoust. Soc. Am.* **91**, 1042–1050.
- Hall, J. W., Haggard, M. P., and Fernandes, M. A. (1984). "Detection in noise by spectro-temporal pattern analysis," *J. Acoust. Soc. Am.* **76**, 50–56.
- Ihlef, 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.
- Kendall, M. G., and Babington Smith, B. (1939). "The problem of m-rankings," *T. Ann. Math. Stat.* **10**, 275–287.
- Kitterick, P. T., and Summerfield, A. Q. (2007). "The role of attention in the spatial perception of speech," *Assoc. Res. Otolaryngol. Abstr.* **30**, 423.
- Kwon, B. J. (2002). "Comodulation masking release in consonant recognition," *J. Acoust. Soc. Am.* **112**, 634–641.
- 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.
- Loftus, G. R., and Masson, M. E. J. (1994). "Using confidence-intervals in within-subject designs," *Psychon. Bull. Rev.* **1**, 476–490.
- 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.
- 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.
- Peters, R. W., and Hall, J. W. (1994). "Comodulation masking release for elderly listeners with relatively normal audiograms," *J. Acoust. Soc. Am.* **96**, 2674–2682.
- Pierzycki, R. H., and Seeber, B. U. (2010). "Indications for temporal fine structure contribution to co-modulation masking release," *J. Acoust. Soc. Am.* **128**, 3614–3624.
- Pressnitzer, D., Meddis, R., Delahaye, R., and Winter, I. M. (2001). "Physiological correlates of comodulation masking release in the mammalian ventral cochlear nucleus," *J. Neurosci.* **21**, 6377–6386.
- Schooneveldt, G. P., and Moore, B. C. (1987). "Comodulation masking release (CMR): Effects of signal frequency, flanking-band frequency, masker bandwidth, flanking-band level, and monotic versus dichotic presentation of the flanking band," *J. Acoust. Soc. Am.* **82**, 1944–1956.
- 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.
- Verhey, J. L., Pressnitzer, D., and Winter, I. M. (2003). "The psychophysics and physiology of comodulation masking release," *Exp. Brain Res.* **153**, 405–417.