Psychometric functions for sentence recognition in sinusoidally amplitude-modulated noises

Yi Shen,^{1,a)} Nicole K. Manzano,² and Virginia M. Richards²

¹*Department of Speech and Hearing Sciences, Indiana University Bloomington, 200 S Jordan Avenue, Bloomington, Indiana 47405-7000, USA*

²*Department of Cognitive Sciences, University of California, Irvine, 3151 Social Science Plaza, Irvine, California 92687-5100, USA*

(Received 6 March 2015; revised 7 November 2015; accepted 21 November 2015; published online 16 December 2015)

Listeners' speech reception is better when speech is masked by a modulated masker compared to an unmodulated masker with the same long-term root-mean-square level. It has been suggested that listeners take advantage of brief periods of quiescence in a modulated masker to extract speech information. Two experiments examined the contribution of such "dip-listening" models. The first experiment estimated psychometric functions for speech intelligibility using sentences masked by sinusoidally modulated and unmodulated speech-shaped noises and the second experiment estimated detection thresholds for a tone pip added at the central dip in the masker. Modulation rates ranging from 1 to 64 Hz were tested. In experiment 1 the slopes of the psychometric functions were shallower for lower modulation rates and the pattern of speech reception thresholds as a function of modulation rate was nonmonotonic with a minimum near 16 Hz. In contrast, the detection thresholds from experiment 2 increased monotonically with modulation rate. The results suggest that the benefits of listening to speech in temporally fluctuating maskers cannot be solely ascribed to the temporal acuity of the auditory system. © 2015 Acoustical Society of America.

[<http://dx.doi.org/10.1121/1.4937613>]

[FJG]

Pages: 3613–3624

I. INTRODUCTION

A characteristic feature of speech communication is its robustness against background noises. Typical background noises (maskers) in everyday listening environments have fluctuations of sound energy over time and normal-hearing listeners take the advantage of such temporal fluctuations in the maskers to improve speech understanding (e.g., Miller and Licklider, 1950; Takahashi and Bacon, 1992; Howard-Jones and Rosen, 1993; Gustafsson and Arlinger, 1994). Previous studies have also found that this fluctuating masker benefit (FMB) is reduced or absent for listeners with hearing impairment (e.g., Festen and Plomp, 1990; Gustafsson and Arlinger, 1994). Traditional models of speech intelligibility in noise often fail to capture these features (e.g., Speech Intelligibility Index or SII, ANSI, 2007). Because these models depend on the long-term average audibility of speech, often determined by the long-term target-to-masker ratio (TMR), speech intelligibility in modulated maskers for normal-hearing listeners is underestimated. Such models also fail to predict the large performance differences in understanding speech between normal-hearing and hearing-impaired listeners when amplitude-modulated maskers are used.

More recent research has led to the introduction of models that incorporate dynamic components to evaluate the effects of modulated maskers on speech intelligibility. One class of such models assumes that normal-hearing listeners

can effectively "listen-in-the-valley" (Buus, 1985) and recognize speech based on incomplete speech segments, usually available during the momentary high target-to-masker-ratio epochs (e.g., Bronkhorst *et al.*, 1993; Rhebergen and Versfeld, 2005; Rhebergen *et al.*, 2006). An alternative approach is to extend the evaluation of the short-term target-to-masker ratio further into the amplitude-modulation domain (e.g., Jørgensen and Dau, 2011; Jørgensen *et al.*, 2013). A common feature in these recent models is that the predicted FMB depends on temporal acuity of the system. The current study investigates the relationship between the FMB and temporal acuity by contrasting results for sentence recognition and the detection of brief tone pips in amplitude-modulated maskers.

Many previous studies assessed the FMB by the decrease in the speech reception thresholds (SRTs) measured using amplitude-modulated maskers compared to unmodulated maskers (e.g., Peters *et al.*, 1998; Dubno *et al.*, 2002; George *et al.*, 2006; Jin and Nelson, 2006). The SRT is the TMR required for 50% correct identification of keywords in a target sentence, i.e., a single point on the psychometric function. The SRT is often estimated using an adaptive psychophysical procedure (e.g., Levitt, 1970), and once estimated the difference in SRTs for modulated relative to unmodulated maskers forms the FMB. This measure of the FMB, however, does not take into account the fact that introducing masker modulation changes not just the SRT, but also the slope of the psychometric function (Oxenham and Simonson, 2009; Bernstein and Grant, 2009; Gregan *et al.*, 2013; MacPherson and Akeroyd, 2014). As a result,

^{a)}Electronic mail: shen2@indiana.edu

improvement in keyword identification due to masker modulation is usually greater at lower TMRs, or the lower percent-correct regions of the psychometric function. Since the slope of the psychometric function might carry as much information regarding the mechanisms underlying the FMB as the SRT, the current study quantifies the FMB by measuring the psychometric function, which is described by both the SRT and the slope of the psychometric function.

There are a number of difficulties in estimating psychometric functions for sentence recognition across a large number of conditions (e.g., various masker modulation rates). First, traditional procedures for the estimation of the psychometric function can be time consuming. Typically, speech recognition is measured using a set of pre-selected TMRs. Without knowing the domain of the psychometric function, the pre-selected TMRs may include values that lead to 0% or 100% correct, which is inefficient. Second, the speech corpus used in an experiment may include a relatively small number of unique sentences. If estimating a single psychometric function requires a large number of independent sentences, it limits the total number of experimental manipulations that could be made. The current study adopted a Bayesian adaptive procedure (Shen and Richards, 2012) to improve the efficiency in estimating psychometric functions. As a result, experiment 1 provided a comprehensive description of how the psychometric function for sentence recognition varies with the masker modulation rate for normal-hearing listeners, which provided a basis for testing hypotheses with regard to the origin of the FMB.

It has been hypothesized that the FMB, at least in part, arises from “dip-listening” (Buus, 1985). That is, the fluctuations in the temporal envelope of the masker form momentary dips where the short-term TMR is high. It is possible that listeners “glimpse” segments of the target speech during the envelope dips of the masker. Notably, the availability of the glimpses of the target speech is constrained by the temporal resolution of the auditory system. Temporal resolution describes the auditory system’s ability to rapidly encode varying information contained in the temporal envelope of sound stimuli, and is usually modeled by postulating a temporal integrator (e.g., Viemeister, 1979). When the masker’s modulation rate is rapid, the auditory system fails to encode the amplitude modulation, leading to a smoothed internal representation of the masker envelope. This, in turn, predicts that the TMR available for glimpses of the speech sound would depend on listeners’ temporal resolution. Dubno *et al.* (2002, 2003) measured the recovery from forward masking to assess the temporal resolution in young and older adults. Older listeners showed increased amounts of forward masking but not simultaneous masking compared to the younger group, potentially revealing poorer temporal resolution. Correlation analyses suggested that the degraded forward masking thresholds was associated with reduced FMB for single-syllable recognition. Jin and Nelson (2006) replicated this result with normal-hearing and hearing-impaired listeners tested with stimulus levels adjusted to equal audibility between the hearing-impaired listeners and their normal-hearing counterparts. For both groups, forward-masked thresholds were related to the syllable recognition scores.

However, they also found that forward masking results were not strongly correlated to sentence recognition performance.

Gregan *et al.* (2013) tested the relationship between the FMB for sentence recognition in amplitude-modulated noises and the availability of high-TMR dips in the noise maskers by measuring the detection thresholds for a brief tonal probe as a function of the temporal position of the probe relative to the masker amplitude modulation. They reasoned that a large difference between the probe thresholds at the masker’s envelope peak and dip, i.e., a large amount of masking release, would indicate the availability of a high-TMR in the internal representation of the maskers’ dips. Indeed, they found that the amount of masking release for pure tone was significantly correlated with the FMB for sentence recognition.

The aims of the current study were twofold. First, to provide estimates of the psychometric functions that relate the TMR to the percent correct for keyword recognition from sentences when the masker modulation rate was systematically varied. The second aim was to investigate the contribution of enhanced TMRs in the masker dips to the FMB by comparing the speech recognition results to the audibility of short tone pips within the masker dips. Unlike many previous studies, the current study evaluates the FMB across the entire psychometric function instead of just relying on the SRT or recognition performance at certain TMRs. These psychometric-function data have the potential to provide an improved opportunity to investigate the mechanisms underlying the FMB.

II. EXPERIMENT 1: FLUCTUATING MASKER BENEFIT FOR SENTENCE RECOGNITION

A. Methods

1. Subjects

Ten native speakers of American English participated in this study and four were female. The listeners were between 18 and 31 yr of age and all had audiometric thresholds of 15 dB hearing level (HL) or better at frequencies from 250 to 4000 Hz in both ears. For each listener, the ear with a lower pure-tone average (PTA) threshold (i.e., the mean of the hearing thresholds at 500, 1000, and 2000 Hz) was used. If both ears had equal PTAs, the left ear was tested. Listeners did not receive practice prior to data collection.

2. Stimuli

The target speech stimuli were drawn from the Institute of Electrical and Electronics Engineers (IEEE) sentence corpus (IEEE, 1969). Sentences were spoken by an adult female with an American accent.¹ For each of the 720 sentences, the nouns, verbs, pronouns, adverbs, and adjectives were marked as keywords. Only the 452 sentences with five keywords were retained for testing.² In two conditions, the target speech stimuli were presented at an overall level of 70 dB sound pressure level (SPL). This level was chosen to represent speech levels that often occur in noisy environments.

On each experimental trial, a target sentence was presented in a masker sound, which was a broadband noise

spectrally shaped to match the average long-term spectrum of the 452 test sentences. The masker was turned on 1 s before the onset of the target sentence and turned off 1 s after the offset of the target using 0.5-s raised cosine ramps. In separate conditions, the masker was either unmodulated or sinusoidally amplitude modulated with a 100% modulation depth and modulation rates of 1, 2, 4, 8, 16, 32, or 64 Hz.

The target and masker were generated digitally at a sampling rate of 22 050 Hz and presented monaurally via a 24-bit soundcard (Envy24 PCI audio controller, VIA technologies, Inc.), passed through a programmable attenuator and headphone buffer (PA4 and HB6, Tucker-Davis Technologies, Inc.) and presented to the listener through a Sennheiser HD410 SL headset.

3. Procedures

The experiment was conducted in a single 2-h experimental session for each listener. During the session, the eight conditions (one unmodulated- and seven modulated-masker conditions) were tested in a random order for each listener. On each trial, the target sentence was drawn at random, without replacement, from the 452 potential sentences. Thus, no listener heard the same sentence twice.

Listeners were seated in a double-walled sound-attenuated booth and were asked to follow the on-screen instructions displayed on a PC monitor. A mouse and a keyboard were provided to the listener. On each trial, after the presentation of the stimulus, the listener was asked to type the sentence she/he heard into an editbox on the screen. After the listener confirmed his/her response, the typed text remained on the screen but could not be edited. In addition, five boxes appeared on the screen, each of which was labeled using a true keyword from the target sentence presented on that trial. The listener was instructed to select the keywords that were correctly identified in their typed response by clicking the corresponding buttons. By instruction, errors in tense or obvious typos were counted as correct. The next trial began after the self-scoring process was complete.

The TMR, and consequently the masker level, was manipulated from trial to trial by the updated maximum-likelihood (UML) procedure (Shen and Richards, 2012; Shen, 2013; Shen *et al.*, 2014). The cumulative density function of a normal distribution was used to model the psychometric function (e.g., Smits and Festen, 2011), i.e., the function relates the TMR to the proportion correct in keyword recognition. The shape of the function was described by two parameters: the mean (μ) and the standard deviation (σ) of the normal distribution. The mean parameter (μ) indicated the TMR that corresponded to the center of the psychometric function (the TMR required for 50% correct), therefore, this parameter is equivalent to the SRT (i.e., $\text{SRT} = \mu$ dB). The standard deviation parameter indicated the spread of the psychometric function along the TMR dimension. It is directly related to the slope of the psychometric function (i.e., $\text{Slope} = (100/\sqrt{2\pi}\sigma)\%/ \text{dB}$ when measured at the inflection point, see Strausburger, 2001). The upper asymptote of the psychometric function was fixed at 98% correct keywords, which reduces the computational

burden of the procedure with little if any cost to the accuracy of the estimate thresholds and slopes.

The purpose of the UML procedure was to determine the optimal TMR to be used on the next trial for the estimation of the parameters μ and σ . For the UML procedure there are three TMRs that are optimal sampling levels for the most likely estimate of the psychometric function (one is associated with μ , and two are associated with σ , see Shen and Richards, 2012 for details). For simplicity, these three optimal TMRs will be referred to as the upper, middle, and lower sweet points.

For each experimental condition, the psychometric function was estimated using a single UML adaptive track, which proceeded as follows. First, the prior distributions for the two psychometric-function parameters were set; both were normal distributions with broad spreads. The prior distribution for μ had a mean of 10 dB with a standard deviation of 20 dB and the prior distribution for σ had a mean of 2 dB with a standard deviation of 5 dB. After each trial, the posterior parameter distributions were updated according to the listener's response, which led to an interim estimate of the psychometric function and estimates of the three sweet points. The stimulus was initially sampled at the upper sweet point. If more than half of the keywords were correctly identified on a given trial, the TMR was shifted one sweet point down, unless the current TMR was already at the lower sweet point. On the other hand, if less than half of the keywords were correctly identified, the TMR was shifted one sweet point up, unless the current TMR was already at the upper sweet point. This sweet-point-based 1-down, 1-up procedure was repeated until a total of 16 reversals were reached. On average, each UML track was composed of approximately 25 trials (i.e., 25 target sentences or 125 keywords). Note that the few number of sentences needed to generate an estimate of the psychometric function allowed a relatively large number of modulation rates to be tested.

After all listeners had completed the experiment, two "graders" who were native speakers of American English and naive to the purpose of the experiment were recruited to cross-check the self-reported correctness. For each listener, the graders (independently) read the text typed by the listener during the experiment and indicated which keywords were present. The same grading instructions used for the listeners were given to the graders. That is, errors in tense and obvious typos were counted as correct.

B. Results

Across the ten listeners, the percent agreement between their self-scored keyword correctness and those of the independent graders' ranged from 87.0% to 99.2%. The largest discrepancy was due to a single listener (percent agreement of 87.2% and 87.0% between the listener and graders 1 and 2, respectively). For the remaining listeners and across the two graders, the percent agreement ranged from 98.2% to 99.2%. The data of the aberrant listener were removed from the dataset.

The left panel of Fig. 1 plots the SRTs estimated from the current experiment. The results for nine individual

FIG. 1. The speech reception thresholds (SRTs, left panel) and slopes of the psychometric function (right panel) for sentence recognition in sinusoidally amplitude-modulated noise as functions of the masker modulation rate. Individual listeners' data (gray curves) and the averaged results ($N=9$; filled circles) are shown. Results obtained for unmodulated (UM) noise maskers are plotted to the right (unfilled circles). Error bars indicated standard errors of the means.

listeners (gray lines) and their averages (filled circles) are plotted as a function of the masker's modulation rate and for the unmodulated masker (unfilled circles). The SRTs were lower for the modulated-masker conditions than the unmodulated-masker condition. For the modulated maskers, the SRTs depended on the modulation rate, with a minimum (best performance) at 16 Hz. This non-monotonic dependency of the SRT on masker modulation rate was consistent with reports from past investigations of the FMB using sentence materials (e.g., Gustafsson and Arlinger, 1994; Dubno et al., 2003; Iyer et al., 2007). A repeated-measures analysis of variance (ANOVA) with modulation rate (including the unmodulated condition) as the independent variable confirmed the significant main effect of modulation rate [$F(7,56)=21.5$, $p<0.001$]. *Post hoc* pairwise comparisons indicated that the SRT for the unmodulated masker condition was significantly higher ($p<0.05$, Bonferroni corrected) than for all other conditions except for the 1-Hz modulation-rate condition. Therefore, an FMB was obtained for all but the lowest modulation rate tested. In addition, the SRT at the 16-Hz masker modulation rate was significantly lower than those at masker modulation rates of 1, 2, 4, and 64 Hz. This confirmed that the function relating the masker modulation rate and the SRT was non-monotonic in shape.

Table I lists the correlations among the SRTs estimated at all masker modulation rates tested. The SRT for the unmodulated masker was significantly correlated with the SRTs from all other masking conditions except for modulation rates of 1 and 32 Hz. In particular, at the 16-Hz modulation rate, where the lowest average SRT was obtained, the SRT was strongly correlated with that for unmodulated maskers. These significant correlations indicate that a portion of the individual differences in SRTs were consistent across masking conditions, and are related to the SRT in the unmodulated-masker condition.

The right panel of Fig. 1 depicts the slopes of the psychometric functions estimated at their inflection points in units of %/dB (Strausburger, 2001). On average, the slope of the psychometric functions increased as the modulation rate increased, particularly for the higher modulation rates. A repeated-measures analysis of variance (ANOVA) with

TABLE I. Correlations among the SRTs estimated for various masker modulation rates.

	2 Hz	4 Hz	8 Hz	16 Hz	32 Hz	64 Hz	UM
1 Hz	0.53	0.80 ^a	0.37	0.59	-0.04	0.34	0.50
2 Hz		0.65	0.72 ^a	0.85 ^b	0.42	0.88 ^b	0.83 ^b
4 Hz			0.34	0.72 ^a	0.22	0.54	0.81 ^b
8 Hz				0.87 ^b	-0.25	0.82 ^b	0.69 ^a
16 Hz					0.00	0.88 ^b	0.92 ^b
32 Hz						0.26	0.18
64 Hz							0.79 ^a

^a $p<0.05$.

^b $p<0.01$.

modulation rate (including the unmodulated condition) as the independent variable indicated a significant main effect of modulation rate [$F(7,56)=9.0$, $p<0.001$]. *Post hoc* pairwise comparisons showed that the slope of the psychometric function was significantly steeper ($p<0.05$, Bonferroni corrected) in the unmodulated masker condition than with maskers modulated at 1, 2, and 8 Hz. The slope estimated at the highest masker modulation rate, 64 Hz, was significantly steeper than the slopes at masker modulation rates of 1, 2, 8, and 16 Hz. These results suggest that compared to maskers amplitude-modulated at low modulation rates, unmodulated maskers and maskers modulated at high rates (e.g., 64 Hz) are associated with steeper psychometric function slopes. The similar steep slope observed in the high-modulation-rate condition and the unmodulated masker condition is consistent with the idea that the representation of high-rate envelope fluctuations is limited by temporal acuity, causing an amplitude-modulated masker at a high modulation rate to have a similar effect on speech recognition as an unmodulated masker.

Figure 2 illustrates the effect of amplitude modulation on the shapes of the psychometric functions. These group psychometric functions were generated based on the average of the estimated μ and σ values across the nine listeners. In separate panels, the psychometric functions estimated using the maskers modulated at 2-, 16-, and 64-Hz (solid curves in the top left, top right, and bottom panels, respectively) are compared with the psychometric function for the unmodulated masker (dashed curves). The modulation rates of 2 and 16 Hz were chosen as representative examples because this pair of conditions exhibited the largest contrast in the average SRTs across subjects. The top panels of Fig. 2 suggest that these two modulation rates did not lead to a large difference in the psychometric-function slopes (6.3%/dB and 6.6%/dB for the 2- and 16-Hz modulation rates, respectively). Additionally, although the SRTs in the 2- and 64-Hz conditions (left panels) were both higher than the 16-Hz condition, the slopes of the psychometric functions in these two conditions were quite different, with the psychometric-function slope in the 64-Hz condition (9.0%/dB) being more similar to the slope in the unmodulated-masker condition (10.1%/dB).

Figure 3 plots the group psychometric functions using percent correct (left panel) and TMR (right panel) as parameters. The left panel provides an examination of the FMB for

FIG. 2. Three example psychometric functions from parameters averaged across listeners for masker modulation rates of 2 (top left), 16 (top right), and 64 (bottom) Hz (solid curves). The psychometric function for the unmodulated masker is also plotted in each panel for comparison (dashed curves). The slopes of the psychometric functions are indicated next to the functions.

many fixed percentage correct values, e.g., not just 50%. For example, it is apparent that at low recognition scores, the FMB can be quite large, (e.g., the difference in TMR for the unmodulated masker vs the masker modulated at 16 Hz is quite large, approximately 20 dB). Alternatively, the right panel indicates that the range of percent correct recognition scores vary more extensively for the low-to-mid range TMRs (e.g., -15 and -10 dB) than for high TMRs (e.g., 0 dB), which agrees with previous studies of the FMB using either isolated syllables/words (e.g., [Dubno et al., 2002](#); [Léger et al., 2012](#)) or sentences (e.g., [Nelson et al., 2003](#); [Oxenham and Simonson, 2009](#); [Bernstein and Grant, 2009](#)).

The greater amount of the FMB at 16 Hz than at 2 Hz would seem contradictory to the predictions of the dip-listening hypothesis, which predicts that the FMB should decrease as the masker modulation rate increases because the temporal resolution of the auditory system yields internal representations of envelopes that are less fully modulated as the modulation rates increase. The contribution of temporal

resolution to the FMB is further considered following experiment 2, which provides an independent estimate of temporal resolution.

III. EXPERIMENT 2: FLUCTUATING MASKER BENEFIT FOR THE DETECTION OF SHORT-DURATION TONE PIPS

A. Methods

1. Subjects

Seven (three females) of the listeners from experiment 1 (excluding the listener who demonstrated unreliable self-scoring) participated in this experiment. The same ear was tested as in experiment 1.

2. Stimuli

In experiment 2, each experimental trial consisted of three 500-ms intervals, separated by 500-ms inter-stimulus intervals. In one of the intervals, determined at random, a target tone pip was presented at the temporal center of the interval. Listeners were instructed to indicate the interval that contained the tone pip. The target duration was 10 ms, including 5-ms raised-cosine onset and offset ramps. The target level was 40 dB SL (40 dB above the absolute threshold for the same tone pip). Depending on the condition, the target frequency was either 500 or 2000 Hz. All three intervals contained a 500-ms noise masker, gated on and off using 10-ms raised-cosine ramps. The masker was generated as follows. For the modulated maskers, a Gaussian noise was 100% sinusoidally amplitude modulated at 4, 8, 16, 32, or 64 Hz and then band-pass filtered using a fourth-order Butterworth filter with the cutoff frequencies set at one octave below and one octave above the target frequency. For the unmodulated masker, the Gaussian noise was not modulated prior to filtering. Note that the masker used in the current experiment was different from the speech-shaped noise used in experiment 1.³ The unmodulated masker and the modulated maskers were tested in separate conditions, except that the 4-Hz modulation rate was not tested for the 2-kHz target because a masker level potentially harmful to the listeners' hearing would be required to make the 40-dB-SL target inaudible. When amplitude-modulated, the initial phase of the modulator was set so that an envelope dip was

FIG. 3. Group psychometric functions are plotted as percent correct (left panel) or TMR (right panel) contours.

presented at the temporal center of the masker. When present, the target was temporally aligned with that central envelope dip. On each trial a new masker carrier was tested, but across intervals within a trial the maskers were identical in all aspects.

An identical system for the stimulus generation and presentation from experiment 1 was used for the current experiment. On each trial, the listener was instructed to indicate the interval that contained the tone-pip target by pressing the corresponding key on a keyboard. After each response, feedback with regard to correctness was provided.

3. Procedures

Experiment 2 was completed in a single 2-h session, and was run after experiment 1. The experiment was composed of a pre-test and the main experiment.

During the pre-test, listeners' absolute threshold for the 10-ms tone pips were estimated for the 500- and 2000-Hz target frequencies. The UML procedure (Shen and Richards, 2012) was used with the three-interval, three-alternative forced-choice paradigm. The stimuli used for the pre-test were the same as those for the main experiment described above, except that the masker was not presented and the target level on each trial was adaptively controlled by the UML using a two-down, one-up stepping rule. Because no masker was present, the intervals were silent save the target sound. For this reason, the duration of each interval was marked by a displaying the interval number (i.e., "1," "2," or "3") on a computer monitor, which was gated on and off with that interval. Each UML track began with a target level at 55 dB SPL and terminated after ten reversals was reached, corresponding to approximately 25 trials.

In the pre-test, and for each listener, the two target frequencies were tested in random order. For each target frequency, three UML tracks were run sequentially. The target threshold was estimated as the 66.7% point on the psychometric function fitted to the trial-by-trial data across the three UML tracks.

For the main experiment, the detection threshold for a tone pip presented in the noise masker was estimated using the UML procedure. For each listener, the two target frequencies were blocked and tested in a pre-determined random order across listeners. For each target frequency, different various masking conditions were tested in random order, with one UML track being run for each masking condition. For each UML track, the lower plateau of the psychometric function was set to 1/3, as in the pre-test. The prior distribution for μ was a normal distribution with a mean of 10 dB and a standard deviation of 20 dB. The prior distribution for σ was a normal distribution with a mean of 2 and a standard deviation of 5. The TMR, defined in the current experiment as the ratio between the target level and the overall level of the masker, was set at 25 dB at the beginning of each UML track, and was then manipulated by the UML procedure adaptively on subsequent trials. Once all masking conditions for the first target frequency had been tested, the process was repeated two more times using newly randomized orders for the different modulation rates. For these

repetitions, each UML track was began where the earlier track left off. After three UML runs were completed for the first frequency, the same process was used for the second frequency. The final estimated threshold for each masking condition was the 66.7% point on the psychometric function fitted to the trial-by-trial data across the three UML tracks.

B. Results

On average, listeners' absolute thresholds for the 10-ms tone pip were 33.0 and 24.3 dB SPL for 500 and 2000 Hz, respectively. These estimates of the absolute threshold were consistent across the seven individual listeners, leading to small standard deviations of 2.4 and 1.6 dB for 500 and 2000 Hz, respectively.

The top panels of Fig. 4 plot the threshold TMRs as a function of modulation rate for the 500- (left) and 2000-Hz (right) target tone pips. Both individual results (gray curves) and their averages (filled circles) are shown. The average thresholds for the unmodulated masker are indicated using unfilled circles. The results were consistent across listeners. For both 500- and 2000-Hz tone pips, the thresholds increased with increasing modulation rates, and the unmodulated maskers had thresholds approximately equal to those estimated for the highest modulation rate tested, 64 Hz, and thresholds from the 500-Hz condition tended to be higher than those from the 2000-Hz condition.

FIG. 4. (Color online) Top panels: the detection thresholds for a brief tone pip presented in a temporal dip of sinusoidally amplitude-modulated noises as functions of the masker modulation rate. Results for target frequencies of 0.5 and 2 kHz are arranged into the left and right panels, respectively. In each panel, individual and average ($N=7$) thresholds are denoted using gray curves and filled circles, respectively. Results for unmodulated maskers are shown as unfilled circles to the right. Bottom panels: The predicted tone-pip thresholds estimated using a temporal integrator model for individual listeners (gray curves). The averages across the individual predictions are shown as filled and unfilled triangles for modulated and unmodulated maskers, respectively. Error bars indicated standard errors of the means.

A repeated-measures ANOVA treating masker modulation rate (for modulation rates ≥ 8 Hz, and including the unmodulated masker condition) and target frequency as the two independent variables and threshold TMR as the dependent variable indicated significant main effects of modulation rate [$F(4, 24) = 120.3, p < 0.001$] and target frequency [$F(1, 6) = 311.3, p < 0.001$] as well as a significant interaction between the two independent variables [$F(4, 24) = 6.6, p = 0.001$]. Of primary interest is the finding that the threshold TMR increased with increasing modulation rate, which was anticipated based on the low-pass effect of temporal resolution on envelope processing. It suggests that the audibility during the envelope dips of an amplitude-modulated masker improves as the modulation rate decreases. The monotonic effect of masker modulation observed in the current experiment does not resemble the results of experiment 1 where sentence recognition performance was found to depend on the masker modulation rate in a non-monotonic fashion. This discrepancy between the two experiments indicated that the FMB for sentence recognition in noise cannot be explained solely by audibility during the temporal dips of the masker.

To quantify the individual listeners' temporal processing capabilities based on the dependencies of the tone-pip thresholds on the masker modulation rate, a simple temporal integrator model was implemented. The model calculated the TMR within a rectangular temporal integration window (i.e., the local TMR) with the temporal center of the 10-ms tone pip aligned with the center of the integration window. The detection threshold was determined when the local TMR was equal to a criterion K . Therefore, the model had two free parameters: the duration of the temporal integration window τ and the criterion TMR K . Note that the parameter K can be thought of as indicating a listener's efficiency in signal detection. That is, a more efficient detector requires a lower value of K to reach the detection threshold. Because the envelopes of both the tone pip and the masker were deterministic (described by sine functions), the expected target and masker intensities within the temporal window were calculated analytically. The local TMR was the ratio between the expected local intensities of the target and masker, expressed in units of dB. For each frequency and each individual listener, the model was fitted to the threshold TMRs across the masker conditions using least squares criteria, resulting in a pair of τ and K estimates. The lower panels of Fig. 4 plot the predicted tone-pip detection thresholds by the best-fitted models for 500 (left) and 2000 Hz (right) for individual subjects (gray curves) and the average thresholds (triangles). The predicted thresholds represent the experimental data (top panels) well, indicating that the simple temporal integrator model can account for the dependency of the tone-pip detection thresholds on the masker modulation rate.

Further analyses were conducted to investigate the degree to which the values of τ and K , estimated in experiment 2, can provide insights into the patterns of the SRTs and psychometric-function slopes across the masker modulation rates observed in experiment 1. First, consider the duration of the estimated temporal integration window τ . Correlation analyses were run separately for the two tone-pip frequencies tested in experiment 2, using the estimates

of τ as the independent variable. The patterns of SRTs were summarized using the difference between the highest and lowest SRTs across all masker conditions tested ($\text{SRT}_{\text{maxdiff}}$), which formed the dependent variable. Note that because the highest SRT occurred for the unmodulated masker condition for six of the seven listeners, $\text{SRT}_{\text{maxdiff}}$ was highly correlated to the maximum FMB across the various masker modulation rates a subject obtained in experiment 1 [$r = 1.00, p < 0.001$]. The $\text{SRT}_{\text{maxdiff}}$ was chosen as the dependent variable here because a parallel measure was available for the values of the slope of the psychometric function.⁴ Thus, for the psychometric-function slopes, the pattern of results from experiment 1 was summarized using the difference between the steepest and shallowest slopes across masker conditions ($\text{Slope}_{\text{maxdiff}}$). No significant correlation was found between τ and $\text{SRT}_{\text{maxdiff}}$ for either of the tone-pip frequencies tested [$r = 0.53, p = 0.22$ for 500 Hz; $r = -0.72, p = 0.07$ for 2000 Hz], nor was there a significant correlation between τ and $\text{Slope}_{\text{maxdiff}}$ [$r = 0.14, p = 0.77$ for 500 Hz; $r = 0.67, p = 0.10$ for 2000 Hz]. Thus, there is no evidence from this analysis to suggest that the variation in temporal acuity predicts the pattern of SRTs or psychometric function slopes obtained in experiment 1.

When the estimated efficiency parameter K was considered as the independent variable, a marginally significant correlation was found between K and $\text{SRT}_{\text{maxdiff}}$ [$r = 0.74, p = 0.06$] for 2000 Hz but not for 500 Hz [$r = -0.11, p = 0.81$]. Moreover, no significant correlation was found between K and $\text{Slope}_{\text{maxdiff}}$ [$r = 0.06, p = 0.19$ for 500 Hz; $r = -0.15, p = 0.75$ for 2000 Hz]. Therefore, it would seem that K is not a reliable predictor for the patterns of SRTs and psychometric-function slopes observed in experiment 1. Finally, because the parameter K is associated with the efficiency in detecting a tone-pip target in noise, its value might be expected to be related to overall sensitivity in the speech-recognition task (i.e., a positive correlation between K and SRT). Indeed, this result was observed. There was a significant correlation between K at 2000 Hz and the mean SRT across the masker modulation rates [$r = 0.73, p = 0.03$, one-tailed]. This result suggests a certain degree of consistency in individuals' overall sensitivity for detecting 2000-Hz tonal targets and recognizing keywords in sentences.

IV. GENERAL DISCUSSION

A. Comparison with previous studies

One of the goals of the current study was to investigate the extent to which the FMB can be explained by the availability of momentary high-TMR periods in the temporal envelope of the masker. The results from experiment 2 indicate that the audibility of a brief tone pip presented in the masker dip improves with decreases in masker modulation rates, consistent with the prediction of a temporal integration model. However, the observed dependence of the tone-pip threshold on the masker modulation rate does not resemble that for sentence recognition (from experiment 1), providing strong evidence that temporal resolution does not determine the FMB. Moreover, the estimated temporal resolution (τ) from experiment 2 does not predict the dependence of the

psychometric function for sentence recognition on the masker modulation rate in experiment 1, nor is τ related to the amount of FMB for the speech recognition task. This, too, is consistent with the conclusion that the FMB cannot be solely accounted for by temporal acuity estimated from detection thresholds for tone pips masked by amplitude-modulated noises. However, one should be cautious in interpreting the lack of correlations between τ and FMB, because the current data set might not exhibit sufficient variability in either τ or FMB because only young normal-hearing listeners were tested.

The relationship between temporal resolution and FMB has been addressed in a number of previous studies (e.g., Dubno *et al.*, 2002, 2003; Jin and Nelson, 2006; George *et al.*, 2006; Gregan *et al.*, 2013). For example, Dubno *et al.* (2002, 2003) measured the detection thresholds of tone pips forward masked by speech-shaped-noise maskers as a function of the masker-target delay among young and older listeners with relative normal hearing. The forward masked thresholds were used as an indicator of temporal resolution, and they found a significant correlation between forward masked thresholds and consonant recognition when the masker was temporally interrupted (i.e., amplitude-modulated using square waves). Dubno *et al.* (2003) varied the masker modulation rates from 2 to 50 Hz and found stronger correlations between the forward masking and consonant recognition results for higher modulation rates. Using listeners with normal-hearing and sensorineural hearing loss, Jin and Nelson (2006) confirmed this finding in a later study for syllable recognition for masker modulation rates of 8 and 16 Hz. As pointed out by Gregan *et al.* (2013), using forward masked thresholds as indicators of temporal resolution might be problematic because factors other than temporal resolution (e.g., peripheral compression) might affect forward masking, making the interpretation of the results difficult. To address this issue, Gregan *et al.* (2013) measured temporal resolution using the rate at which forward masking recovers with an off-frequency masker. For listeners with sensorineural hearing loss and normal-hearing listeners with simulated hearing loss, this measure of temporal resolution weakly correlated with the FMB for a masker modulation rate of 10 Hz. Results from this and past experiments suggest that temporal resolution may be one among several factors that influence the amount of the FMB, and a significant effect of temporal resolution on the FMB is more likely to be found for higher masker modulation rates.

In the current study, speech recognition and the detection of brief tonal targets were studied as the masker modulation rate is systematically varied. A non-monotonic dependence of speech recognition performance on masker modulation rate was observed, contradicting the temporal integrator model of temporal acuity. Similar non-monotonic dependencies has been reported frequently in the literature (e.g., Miller and Licklider, 1950; Gustafsson and Arlinger, 1994; Füllgrabe *et al.*, 2006; Lorenzi *et al.*, 2006; Phatak and Grant, 2012). Miller and Licklider (1950), who used monosyllabic words as the test material, examined speech recognition performance as functions of the masker's modulation rate. The best performance occurred at about 10 Hz (square-

wave modulation, half duty cycle), and the performance degraded for lower or higher modulation rates. Although the degradation of performance at high masker modulation rates is expected according to the low-pass characteristic of temporal acuity, potential mechanisms underlying the performance at very low masker rates are less obvious. Miller and Licklider suggested that at low modulation rates important speech features might be masked when the masker power is high. According to this hypothesis, when the duration of individual envelope peaks in the masker is longer than the duration of crucial speech features, these speech features might be presented at a far more disadvantageous TMR than the nominal TMR during the masker peaks. Other studies of the recognition of periodically interrupted speech also demonstrated non-monotonicity in performance as the interruption rate increases (e.g., Miller and Licklider, 1950; Powers and Speaks, 1973; Nelson and Jin, 2004; Jin and Nelson, 2010; Shafiro *et al.*, 2011; Bhargava and Başkent, 2012). These studies provided further evidence that important speech features are lost by gating the speech on and off at a low rate (either using a noise masker or interruptions). This phenomenon will be referred to as the *interruption of speech features* in the following discussion.

Other researchers have hypothesized that the performance is relatively poor at very low masker modulation rates because the benefit of masker dips is counter-balanced by the modulation masking/interference mechanisms (e.g., Kwon and Turner, 2001). That is, the envelope information carried by the target speech could be masked or interfered by the masker envelope in the modulation domain (e.g., Houtgast, 1989; Yost and Sheft, 1989; Bacon and Grantham, 1989; Dau *et al.*, 1999), leading to degraded speech recognition. Kwon and Turner argued that both the dip-listening and modulation masking/interference contributes to the non-monotonic performance patterns observed as a function of masker modulation rate, with temporal resolution being the main performance-limiting factor for high modulation rates. Additionally it is known that different speech materials lead to different patterns of performance as a function of masker modulation rate (e.g., Füllgrabe *et al.*, 2006).

In summary, multiple factors are likely to contribute to the non-monotonic dependence of speech recognition performance on masker modulation rate. Even within the framework of the dip-listening hypothesis, a non-monotonic pattern of results might be expected due to the combined effect of temporal acuity and the interruption of speech features. Therefore, in Sec. IV B, we provide an initial exploration of the contribution of temporal acuity and the interruption of speech features, asking whether a simple dip-listening mechanism can provide predictions that resemble the pattern of the psychometric functions observed in experiment 1.

B. Psychometric functions for speech recognition predicted by a simple dip-listening model

A dip-listening mechanism that includes the effect of temporal acuity, which is expected to dominate sensitivity at high modulation rates, and the effect of the interruption of

speech features, which determines the performance at low masker modulation rates, would reasonably produce a non-monotonic relation between the speech reception threshold and masker modulation rate. To explore whether such a mechanism can capture aspect of the psychometric function estimated in experiment 1, a simple model of dip-listening was implemented. The goal was to learn whether the first-order behavior of such a class of models could account for the current data set rather than to develop a comprehensive model of speech-in-noise perception that would generalize to stimuli other than those tested here. Accordingly, the model was implemented in its simplest form (with the number of free model parameters as low as possible). Of primary interest is the potential that a simple dip-listening model could capture the SRT and slope results from experiment 1 simultaneously.

The subsequent model is a simplified version of ESII model proposed by [Rhebergen and Versfeld \(2005\)](#) and [Rhebergen et al. \(2006\)](#). Briefly, their model first calculated the TMR within a sliding temporal integration windows. In their model, at each epoch the local Speech Intelligibility Index (SII, [ANSI, 2007](#)) is calculated as a function of time, and the prediction of overall performance is based on average of these local SIIs over time. Their model provides satisfactory fits to speech recognition under a variety of masking conditions, including temporally interrupted noises with various duty cycles, and speech-modulated noises. As one might expect, for this model the duration of the temporal integrator effectively governs the amount of the FMB predicted for modulated maskers. Thus, the model predicts monotonic improvement in SRT as the masker modulation rate increases, which contradicts experimental results (e.g., Fig. 9 in [Rhebergen and Versfeld, 2005](#) and Fig. 4 in [Rhebergen et al., 2006](#)).

For the current reduced model, the masker envelope was described using a sine function with an initial phase of φ , relative to the temporal onset of the keyword. Therefore, for a nominal TMR of $\text{TMR}_{\text{nominal}}$, the local TMR within a rectangular integration window can be approximated as

$$\text{TMR}_{\text{local}}(t, \varphi) = \text{TMR}_{\text{nominal}} - 10 \log \left\{ \frac{2}{3\Delta T} \times \int_t^{t+\Delta T} [1 + \sin(2\pi f_m t + \varphi)]^2 dt \right\}, \quad (1)$$

where t indicates time relative to the onset of the keyword, ΔT is the duration of the temporal integration window, and f_m is the masker modulation rate. Defining the psychometric function for recognizing the keyword in steady-state noise as $\text{PF}_{\text{UM}}(\text{TMR}_{\text{nominal}})$, the expected psychometric function for sinusoidally amplitude-modulated maskers is

$$\begin{aligned} & \text{PF}_{\text{MOD}}(\text{TMR}_{\text{nominal}}) \\ &= \frac{1}{2\pi T'} \int_0^{2\pi} \int_{T'} \text{PF}_{\text{UM}}[\text{TMR}'_{\text{local}}(t, \varphi)] dt d\varphi, \end{aligned} \quad (2)$$

where $\text{TMR}'_{\text{local}}$ is the local TMR as in Eq. (1) but bounded to be between -15 and 15 dB, as suggested by ANSI's standard in calculating the SII ([ANSI, 2007](#)) and T' is the

duration of the keyword that contributes to the recognition of the word, which is assumed to be shorter than the total duration of the word T because speech contains redundant information. The redundancy of speech is then reflected in a parameter $r = T'/T$. It is further assumed that the integration duration T' in Eq. (2) is associated with the portions of T with the highest $\text{TMR}_{\text{local}}$.

Compared to the model of [Rhebergen and Versfeld \(2005\)](#), the current model includes significant simplifications for illustration purposes. First, the local TMRs were calculated analytically rather than using the actual temporal envelopes extracted from the target and masker stimuli. Second, no spectral bands were used and hence no spectral weights were implemented. This is because in the current experiment, speech-shaped noises were tested and the long-term TMR was expected to be constant across spectral regions. Last, instead of calculating the SII and then the predicted performance, the current model only predicted the relationship between the psychometric functions for the modulated- and unmodulated-masker conditions (i.e., PF_{MOD} and PF_{UM}), reducing the number of computational steps.

The current model also bears similarity to the model of [Bronkhorst et al. \(1993\)](#), noticeably Sec. III C). In their model, [Bronkhorst et al. \(1993\)](#) considered not only the probability of recognizing a certain phoneme when it is only partially perceived, but also the probably of successfully guessing a missed phoneme based on other recognized phonemes in the same word. That is, for a known word structure (e.g., CVC words), their model allows detailed descriptions of the contextual dependencies among the components of words and how speech information is distributed among various phonemes. In contrast, the current model, assumed a uniform distribution of speech information over the word duration. This simplification in the current model was made so that the application of the model would not be limited to known word structures and in an effort to maintain a very small number of free parameters.

Using Eqs. (1) and (2), the psychometric functions for keyword recognition in amplitude-modulated maskers (PF_{MOD}) were derived for the masker modulation rates tested in experiment 1. The SRT was defined as the value of the dependent variable needed for 50% correct from the predicted psychometric function, and the slope of the psychometric function was defined as the maximum slope within the dynamic range of the psychometric function (i.e., the inflection point). Ultimately, there were three free parameters in the current model: the duration of the temporal integration window ΔT , the redundancy factor r , and the duration of the keyword T . The model simulations were repeated for ΔT 's of 4, 8, and 16 ms, for r 's of 0.5, 0.6, 0.7, 0.8, and 1, and for T 's of 250, 100, and 50 ms.

Figure 5 plots the SRTs (left panels) and slopes (right panels) of the predicted psychometric functions (curves with no symbols) along with the experimental data (circles) for a T of 250 ms. The top, middle, and bottom panels correspond to ΔT of 4, 8, and 16 ms, respectively. In each panel, different curves are the predictions for different values of r . Regardless of the choices of parameter values (i.e., for ΔT and r), both the SRTs and psychometric-function slopes

FIG. 5. (Color online) The predicted SRTs (left panels) and slopes (right panels) of the psychometric function for sentence recognition in sinusoidally amplitude-modulated are shown as a function of the masker modulation rate (curves without symbols). These predictions were generated using a simple model of speech-in-noise recognition based on the evaluation of local TMR within a temporal integration window. The value of T , the assumed duration of the keyword, 250 ms. The two varying parameters are ΔT , the duration of the temporal integration window, and r , the proportion of the duration of a key word that contributes to the recognition of the word. Predictions for ΔT of 4, 8, and 16 ms are arranged in rows. Different values of r are indicated using different line styles. The average experimental results from experiment 1 replotted from Fig. 1 using circles. The arrow on the abscissa marks the cutoff frequency (f_{cutoff}) associated with the temporal integrator.

increased as the masker modulation rate increased, except that a decrease in the SRT with increases in modulation rates could be observed for the very low modulation rates from 1 to 4 Hz and for values of r less than 1. The cutoff frequency above which the SRT and psychometric-function slope began to increase with increasing modulation rate (f_{cutoff}) depended on the duration of temporal integration window ΔT , with a longer value of ΔT being associated with a lower cutoff frequency [$f_{\text{cutoff}} = 1/(2\pi \Delta T)$, see arrows in Fig. 5].

Below the cutoff frequency, the psychometric-function slope was relatively constant across masker modulation rates. Finally, the value of r affected both the SRT and psychometric-function slope. As r increased from 0.5 to 1, the SRT decreased and the psychometric-function slope became slightly steeper for frequencies below f_{cutoff} .

Figure 6 plots the SRTs (left panels) and slopes (right panels) of the predicted psychometric functions (curves with no symbols) along with the experimental data (circles) for ΔT of 8 ms. The top, middle, and bottom panels correspond to T of 250, 100, and 50 ms, respectively, and different curves are for different values of r . At low masker

FIG. 6. (Color online) The predicted SRTs (left panels) and slopes (right panels) of the psychometric function for sentence recognition in sinusoidally amplitude-modulated are shown as a function of the masker modulation rate (curves without symbols). The value of ΔT is fixed at 8 ms. Results for different values of T are shown in different rows, and results for different values of r are indicated using different curves. The average experimental results from experiment 1 replotted from Fig. 1 using circles.

modulation rates the SRT depends on T . For a T of 250 ms, and as the masker modulation rate increases, the SRT starts to decrease between 1 and 2 Hz. For shorter values of T , the improvement in the SRT with modulation rate shifts to higher modulation rates (e.g., 2–4 Hz for a T of 100 ms, and 4–8 Hz for a T of 50 ms). The effect of T on the SRT is consistent with the interruption of speech features. That is, when the masker modulation rate is sufficiently low, the keyword might be embedded entirely in a masker peak or trough. In this case, the masker modulation would be as likely to facilitate as to undermine the performance, leading to performance similar to the unmodulated masker condition. At low masker modulation rates ($f_m \ll 1/T$) and for $r < 1$, as the masker modulation rate increased, the psychometric-function slope became shallower. In general, for low modulation rates the SRT and the slope of the psychometric function both decrease with increasingly long T 's.

Figure 7 provides another visualization of the measured (left panel, replotted from the right panel of Fig. 3) and predicted psychometric functions. The individual curves in each panel of Fig. 7 indicate the percent correct keyword identification for different values of TMR. Two sets of model predictions are plotted in the middle (for ΔT of 8 ms, an r of 0.7, and a T of 250 ms) and right (for ΔT of 8 ms, an r of 0.7, and a T of 50 ms) panels. Comparing these two sets of predicted psychometric functions, it is apparent that the parameter T mainly governed the performance at low masker

FIG. 7. The group percent correct as a function of masker modulation rate are plotted separately for different TMRs (left panel, reproduced from the right panel of Fig. 3). The center and right panels plot predictions of the simple model for different parameter values, shown at the top of each panel.

modulation rates. Moreover, at these low masker modulation rates, the predicted performance at low TMRs (e.g., -10 and -15 dB) was higher than the empirically estimated levels of performance.

Although the data were not particularly well captured by the simple model considered here, the model did predict shallower slopes of the psychometric functions for lower masker modulation rates. In general, a reduction in the slope of a psychometric function is associated with increased “noise.” For the current model, at low masker rates an increase in variability would be expected due to the fluctuation of the local TMR. As the masker modulation rate increased beyond f_{cutoff} , the degree to which the local TMR varied decreased, leading to steepening of the psychometric function. With regard to the SRTs, the current low-parameter succeeded in producing a non-monotonic relation between masker modulation rate and percent correct recognition, although the qualitative agreement with the experimental data fall short of excellent. With a ΔT of 8 ms, an r of 0.7, and a T of 250 ms, the model did capture the SRTs for masker modulation rates of 16, 32, and 64 Hz, although the predictions for modulations rates between 2 and 8 Hz were quite poor. Although the predicted SRTs could be made closer to the measure SRTs by using a smaller value of T (e.g., 50 ms, see the bottom left panel of Fig. 6), this value is much shorter than the true durations of the keywords in the test material.

Presumably the discrepancy between the measured and predicted psychometric functions reflects the simplicity of the current model. For example, compared to the model of Bronkhorst *et al.* (1993), the current model assumed a uniform distribution of speech information over time and dismissed the fact that the contribution of various phonemes on word recognition might be unevenly distributed. For example, it might be that consonants are more predictive of the word than vowels while they only make up a relatively small proportion of the word’s duration. This might explain why a value of T smaller than the actual word duration confers improved prediction of the experimental results (see the right panel of Fig. 7). In short, the current modeling results do not exclude the possibility that the FMB could be explained by

temporal acuity in combination with the interruption of speech features, however, additional modifications to the current dip-listening model would be required to achieve improved predictions of experimental data.

Notably, the current model did not include stages/mechanisms to account for (1) the potential involvement of modulation masking and interference (e.g., Kwon and Turner, 2001); or (2) the inherent envelope fluctuations in speech-shaped noises (e.g., Stone *et al.*, 2011; Stone *et al.*, 2012). Therefore, one implication of the current modeling result could be that, for speech recognition in temporally fluctuating noise, the involvement of auditory temporal processing might not be limited to temporal resolution. Models with built-in stages for modulation masking and interference (e.g., Jørgensen and Dau, 2011; Jørgensen *et al.*, 2013) might be needed to adequately capture the dependence of the psychometric function on the masker modulation rate.

V. SUMMARY

Using the UML adaptive procedure, psychometric functions for sentence recognition in sinusoidally amplitude-modulated noises were measured for a wide variety of modulation rates (in experiment 1). The speech results are not predicted by audibility of speech sounds during the temporal dips of the masker envelope (evaluated in experiment 2), suggesting temporal resolution is unlikely to be the sole factor contributing to the FMB.

ACKNOWLEDGMENTS

This work was supported by the Undergraduate Research Opportunities Program at the University of California, Irvine awarded to N.K.M., the New Investigator Research Grant from the American Speech-Language-Hearing Foundation (PI: Y.S.), and the NIH Grant No. R21 DC013406 (MPIs: V.M.R and Y.S.).

¹The speech stimuli were downloaded from <http://www.tigerspeech.com/>.

²Note that the sentences in the original IEEE corpus were arranged into 72 lists of 10 sentences each. These lists were carefully balanced in terms of the expected recognition performance. By using a subset of the sentences

and randomizing the sequence in which the sentences were presented, results from the current experiment could potentially be affected by the unequal intelligibility of individual sentences.

³Pilot results suggested that if the speech-shaped noise was used as the masker, the masker level at threshold for a 40-dB SL target tone pip at 2 kHz would be extremely high.

⁴The analysis presented here using SRT_{maxdiff} as the dependent variable was repeated using two different measures of FMB as alternative dependent variables, i.e., the maximum and mean difference between the SRTs in the modulated and unmodulated conditions. Both measures lead to similar conclusions as when SRT_{maxdiff} was used. No significant correlation was found between either τ or K and the amount of FMB, except a marginal correlation between K for 2000 Hz and the maximum FMB [$r=0.75$, $p=0.05$].

ANSI (2007). ANSI S3.5-2007, *American National Standard Methods for Calculation of the Speech Intelligibility Index* (American National Standards Institute, New York).

Bacon, S. P., and Grantham, D. W. (1989). "Modulation masking: Effects of modulation frequency, depth, and phase," *J. Acoust. Soc. Am.* **85**, 2575–2580.

Bernstein, J. G., 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.

Bhargava, P., and Başkent, D. (2012). "Effects of low-pass filtering on intelligibility of periodically interrupted speech," *J. Acoust. Soc. Am.* **131**, EL87–EL92.

Bronkhorst, A. W., Bosman, A. J., and Smoorenburg, G. F. (1993). "A model for context effects in speech recognition," *J. Acoust. Soc. Am.* **93**, 499–509.

Buus, S. (1985). "Release from masking caused by envelope fluctuations," *J. Acoust. Soc. Am.* **78**, 1958–1965.

Dau, T., Verhey, J., and Kohlrausch, A. (1999). "Intrinsic envelope fluctuations and modulation-detection thresholds for narrow-band noise carriers," *J. Acoust. Soc. Am.* **106**, 2752–2760.

Dubno, J. R., Horwitz, A. R., and Ahlstrom, J. B. (2002). "Benefit of modulated maskers for speech recognition by younger and older adults with normal hearing," *J. Acoust. Soc. Am.* **111**, 2897–2907.

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 normal 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.

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

George, E. L., Festen, J. M., and Houtgast, T. (2006). "Factors affecting masking release for speech in modulated noise for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.* **120**, 2295–2311.

Gregan, M. J., Nelson, P. B., and Oxenham, A. J. (2013). "Behavioral measures of cochlear compression and temporal resolution as predictors of speech masking release in hearing-impaired listeners," *J. Acoust. Soc. Am.* **134**, 2895–2912.

Gustafsson, H. Å., and Arlinger, S. D. (1994). "Masking of speech by amplitude-modulated noise," *J. Acoust. Soc. Am.* **95**, 518–529.

Houtgast, T. (1989). "Frequency selectivity in amplitude-modulation detection," *J. Acoust. Soc. Am.* **85**, 1676–1680.

Howard-Jones, P. A., and Rosen, S. (1993). "The perception of speech in fluctuating noise," *Acustica* **78**, 258–272.

IEEE (1969). "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.* **17**, 225–246.

Iyer, N., Brungart, D. S., and Simpson, B. D. (2007). "Effects of periodic masker interruption on the intelligibility of interrupted speech," *J. Acoust. Soc. Am.* **122**, 1693–1701.

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

Jin, S. H., and Nelson, P. B. (2010). "Interrupted speech perception: The effects of hearing sensitivity and frequency resolution," *J. Acoust. Soc. Am.* **128**, 881–889.

Jørgensen, S., and Dau, T. (2011). "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," *J. Acoust. Soc. Am.* **130**, 1475–1487.

Jørgensen, S., Ewert, S. D., and Dau, T. (2013). "A multi-resolution envelope-power based model for speech intelligibility," *J. Acoust. Soc. Am.* **134**, 436–446.

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.

Léger, A. C., Moore, B. C., and Lorenzi, C. (2012). "Temporal and spectral masking release in low- and mid-frequency regions for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.* **131**, 1502–1514.

Levitt, H. (1970). "Transformed up-down methods in psychoacoustics," *J. Acoust. Soc. Am.* **49**, 467–477.

Lorenzi, C., Husson, M., Ardoint, M., and Debrulle, X. (2006). "Speech masking release in listeners with flat hearing loss: Effects of masker fluctuation rate on identification scores and phonetic feature reception," *Int. J. Audiol.* **45**, 487–495.

MacPherson, A., and Akeroyd, M. A. (2014). "Variations in the slope of the psychometric functions for speech intelligibility: A systematic survey," *Trends Hear.* **18**, 2331216514537722.

Miller, G. A., and Licklider, J. C. R. (1950). "The intelligibility of interrupted speech," *J. Acoust. Soc. Am.* **22**, 167–173.

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., Moore, B. C., 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.

Phatak, S. A., and Grant, K. W. (2012). "Phoneme recognition in modulated maskers by normal-hearing and aided hearing-impaired listeners," *J. Acoust. Soc. Am.* **132**, 1646–1654.

Powers, G. L., and Speaks, C. (1973). "Intelligibility of temporally interrupted speech," *J. Acoust. Soc. Am.* **54**, 661–667.

Rhebergen, K. S., and Versfeld, N. J. (2005). "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Am.* **117**, 2181–2192.

Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (2006). "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," *J. Acoust. Soc. Am.* **120**, 3988–3997.

Shafiro, V., Sheft, S., and Risley, R. (2011). "Perception of interrupted speech: Effects of dual-rate gating on the intelligibility of words and sentences," *J. Acoust. Soc. Am.* **130**, 2076–2087.

Shen, Y. (2013). "Comparing adaptive procedures for estimating the psychometric function for an auditory gap detection task," *Atten. Percept. Psychophys.* **75**, 771–780.

Shen, Y., Dai, W., and Richards, V. M. (2014). "A MATLAB toolbox for the efficient estimation of the psychometric function using the updated maximum-likelihood adaptive procedure," *Behav. Res. Methods* **47**(1), 13–26.

Shen, Y., and Richards, V. M. (2012). "A maximum-likelihood procedure for estimating psychometric functions: Thresholds, slopes, and lapses of attention," *J. Acoust. Soc. Am.* **132**, 957–967.

Smits, C., and Festen, J. M. (2011). "The interpretation of speech reception threshold data in normal-hearing and hearing-impaired listeners: Steady-state noise," *J. Acoust. Soc. Am.* **130**, 2987–2998.

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

Strasburger, H. (2001). "Converting between measures of slope of the psychometric function," *Percept. Psychophys.* **63**, 1348–1355.

Takahashi, G. A., and Bacon, S. P. (1992). "Modulation detection, modulation masking, and speech understanding in noise in the elderly," *J. Speech Hear. Res.* **35**, 1410–1421.

Viemeister, N. F. (1979). "Temporal modulation transfer-functions based upon modulation thresholds," *J. Acoust. Soc. Am.* **66**, 1364–1380.

Yost, W. A., and Sheft, S. (1989). "Across-critical-band processing of amplitude-modulated tones," *J. Acoust. Soc. Am.* **85**, 848–857.