Efficiency in glimpsing vowel sequences in fluctuating maskers: Effects of temporal fine structure and temporal regularity

Yi Shen^{a)} and Dylan V. Pearson

Department of Speech and Hearing Sciences, Indiana University Bloomington, Bloomington, Indiana 47405, USA

(Received 30 December 2018; revised 2 April 2019; accepted 3 April 2019; published online 30 April 2019)

Listeners' efficiency in glimpsing the target speech in amplitude-modulated maskers may depend on whether the target is perceptually segregated from the masker and on the temporal predictability of the target. Using synthesized vowel sequences as the target, recognition of vowel sequences in simultaneous amplitude-modulated noise maskers was measured as the signal-to-noise ratio (SNR) and the masker modulation rate were systematically varied. In Experiment I (Exp. I), the temporal fine structure of the target was degraded by synthesizing the vowels using iterated rippled noise as the glottal source. In Experiment II (Exp. II), the vowel sequences were constructed so that they were not isochronous, but instead contained randomized intervals between adjacent vowels. Results were compared to the predictions from a dip-listening model based on short-term SNR. The results show no significant facilitative effect of temporal fine structure cues on vowel recognition (Exp. I). The model predictions significantly overestimated vowel-recognition performance in amplitude-modulated maskers when the temporal regularity of the target was degraded (Exp. II), suggesting the influence of temporal regularity on glimpsing efficiency. Furthermore, the overestimations by the model were greater at lower SNRs and selective to moderate masker modulation rates (between 2 and 6 Hz). © 2019 Acoustical Society of America. <https://doi.org/10.1121/1.5098949>

[GCS]

Pages: 2518–2529

I. INTRODUCTION

When speech is presented in masker sounds with temporally fluctuating envelopes, normal-hearing listeners often demonstrate substantial improvements in speech recognition compared to the situations of steady-state maskers (e.g., Miller and Licklider, 1950; Festen and Plomp, 1990; Takahashi and Bacon, 1992; Gustafsson and Arlinger, 1994). This fluctuating masker benefit (FMB) has been explained by listeners' ability to take advantage of the momentary high signal-to-noise ratios (SNRs) during the dips of the masker (i.e., dip-listening). This short-term SNR based dip-listening hypothesis has been successfully implemented in a number of speech-recognition models for various types of speech materials and masker modulation characteristics (e.g., Bronkhorst *et al.*, 1993; Rhebergen and Versfeld, 2005; Rhebergen *et al.*, 2006; Taal *et al.*, 2011; Schubotz *et al.*, 2016; Biberger and Ewert, 2017; Shen and Pearson, 2017).

However, the benefit from fluctuating maskers predicted by many of these models is based on the assumption that listeners perform dip-listening optimally. That is, the listeners are assumed to have *a priori* knowledge of both the target and masker. Specifically, this implies that they know when the dips are going to occur in the masker, and when the masker dips and the important speech information will coincide. This assumption may be sound in situations where the target and masker envelopes exhibit regular, predictable patterns and when the target and masker sources are

perceptually segregated. The current study investigates the FMB under conditions where the cues for perceptually segregating the target and masker were systematically distorted and the temporal regularity of the target was varied. The experimental results were compared with an ideal dip-listening model. The deviations from the ideal dip-listening model as functions of stimulus manipulations (i.e., manipulations of perceptual segregation and temporal predictability) would provide insight into the non-peripheral processes involved in the FMB. It is hypothesized that reducing segregation cues or temporal regularity would undermine glimpsing efficiency.

It has been repeatedly demonstrated that temporal fine structure (TFS) information plays a crucial role in the FMB (e.g., Qin and Oxenham, 2003; Hopkins and Moore, 2009; Oxenham and Simonson, 2009; Bernstein and Brungart, 2011; Freyman *et al.*, 2012; Steinmetzger and Rosen, 2015). Temporal fine structure may contribute to the perceptual segregation of the target and masker, enabling listeners to selectively attend to the target while preventing interference from the masker. In the current study, the TFS information was systematically degraded in an effort to create target and masker stimuli with either distinct or similar perceptual characteristics. It is expected that when listeners fail to segregate the target from the masker, additional performance degradation may be observed beyond the amount of energetic masking predicted by the short-term-SNR based models.

Temporal regularity in a target stimulus is another cue that could improve speech intelligibility in fluctuating noise. Temporal regularity may give the listener insights into when to glimpse through the masker, as well as allowing the

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

listener to detect the target more quickly and accurately because the signal is temporally expected (Barnes and Jones, 2000). In the current study, we systematically reduced temporal regularity in our stimuli. It is expected that listeners would find it more difficult to identify the target vowels due to this manipulation.

Manipulations of either TFS or temporal envelope of natural speech have been used in many previous investigations of speech recognition in fluctuating maskers. Many previous studies varied the amount of TFS cues using the noise- or sinewave-vocoding techniques that were originally designed for cochlear implant simulation (Shannon *et al.*, 1995; Qin and Oxenham, 2003; Hopkins and Moore, 2009). Common techniques for altering the temporal envelope included temporal interruption and time compression (Gordon-Salant and Fitzgibbons, 2004; Fogerty *et al.*, 2016). Although the importance of TFS and temporal envelope could be investigated through these direct modifications of natural speech, they simultaneously alter speech intelligibility. That is, these modifications affect baseline speech recognition performance in the absence of masking. Experimental paradigms that allow independent manipulations of segregation cues and temporal regularity while controlling for speech intelligibility may be beneficial.

Shen and Pearson (2017) explored the possibility of using synthesized vowel sequences as the identification target to study the FMB. This stimulus paradigm allowed for varying the rate of vowel presentation while maintaining the duration, hence the baseline intelligibility, of individual vowels. When instructed to identify vowel sequences in sinusoidally amplitude-modulated broadband noises, listeners demonstrated substantial FMB. Moreover, the vowel recognition performance depended on both vowel rate and masker modulation rate. This complex pattern of vowel recognition performance was well predicted by a short-term-SNR based dip-listening model, suggesting limited involvements of any non-peripheral processes and near-ideal glimpsing efficiency. This could be partially attributed to two factors. First, since the target vowels were synthesized using harmonic complex carriers, they had much stronger pitch strength than the noise masker. This distinct difference may have facilitated the perceptual segregation of the target and masker sources, leading to more efficient glimpsing. Second, the vowels in the target sequence were isochronous. This, along with the sinusoidally amplitude-modulated masker, resulted in both target and masker envelopes with a high degree of temporal regularity. This may have provided additional cues for the listeners to more efficiently distribute their glimpses over time.

In the current study, modifications to the stimuli used by Shen and Pearson (2017) were made so that segregation between the target vowel sequence and masking noise and the temporal regularity of the target sequence were systematically varied. The same ideal dip-listening model was used to explain the vowel identification performance. It is expected that removing TFS cues or introducing temporal irregularity would undermine the task performance beyond the energetic masking effect predicted by the dip-listening

model. The unpredicted performance degradation is likely due to non-ideal glimpsing.

One feature that distinguishes the current study from previous investigations is that the current study systematically reduced the amount of TFS cues by using gradually noisier glottal sources during vowel synthesis. The noise-vocoding approach commonly used by previous studies imposes an uneven effect on the TFS cues across the spectrum (e.g., Shofner and Chaney, 2013; Shofner *et al.*, 2018), while the TFS cues in natural speech are widely distributed across frequencies. It has been shown that listeners can benefit from the TFS information similarly in both high and low frequency regions (Oxenham and Simonson, 2009; Hopkins and Moore, 2010). In order to control for the TFS cues evenly across a wide frequency region, the current study synthesized vowel stimuli using iterated rippled noise (IRN) as the glottal source. Iterated ripple noise is created by taking a copy of Gaussian noise and iteratively adding it to the original after a set delay (Yost, 1996; Shearer *et al.*, 2018). The perceived pitch corresponds to the reciprocal of the delay. As the number of iterations of the delay-and-add process (i.e., the IRN iteration) increases, the periodicity in the signal is strengthened, evoking a stronger perceived pitch strength. In the current study, the amount of TFS cues in the synthesized stimuli was strengthened or weakened by manipulating the IRN iteration.

In the following, two experiments are presented. In Experiment I (Exp. I), we attempted to reduce the perceptual segregation between the masker and target by controlling the strength of the pitch cue using the IRN technique mentioned above. In Experiment II (Exp. II), temporal irregularity was introduced to the target sequence by adding temporal jitter to the onsets of the target vowels in the vowel sequence. For both experiments, the proportion of correct responses in identifying vowel sequences in modulated maskers were compared with the predictions of a dip-listening model that assumes ideal target glimpsing with full *a priori* knowledge of the target and masker.

II. EXPERIMENT I: THE EFFECT OF TFS CUES

A. Methods

1. Listeners

Ten native English speakers (six females) with ages ranging from 18 to 35 years (mean age: 23.3 years) were recruited from the student population at Indiana University. All listeners had audiometric thresholds equal to or better than 15 dB hearing level (HL) between 250 and 4000 Hz in both ears. The test ear for the experiment was chosen by selecting the ear with the lower pure tone average (PTA) threshold (mean of the hearing thresholds at 0.5, 1, and 2 kHz). In the case of both ears having the same PTA threshold the left ear was selected. Listeners were compensated at an hourly rate for their participation. The experimental protocol was approved by the Institutional Review Board at Indiana University. Informed consent was obtained from all listeners before data collection began.

2. Stimuli

In the current experiment, listeners identified sequences of synthesized vowels with five vowels in each sequence. For each position in the sequence, the identity of the vowel stimulus was drawn randomly from /a/, /i/, /u/, /æ/, and /ɜ/ with replacement, leading to a total of 3125 possible vowel sequences. Each vowel was 100 ms in duration, including 1-ms raised-cosine ramps at the onset and offset. A Klatt synthesizer (Klatt, 1980) was used to create each vowel, which involved two steps. First, an IRN was generated by performing the iterative delay-and-add process on a Gaussian noise. The IRN was independently constructed (i.e., based on independent samples of Gaussian noise) across the various positions in a vowel sequence and across different experimental trials. The delay was set to 4 ms corresponding to a fundamental frequency of 125 Hz, and the IRN iteration was set to 1, 4, or 16 iterations in separate conditions. Second, the IRN was filtered to create the formant structure corresponding to the target vowel. For each vowel identity, the formant frequencies and the bandwidths for each formant were the same as those used by Assmann and Summerfield (1994), Lentz and Marsh (2006), and Shen and Pearson (2017). The vowel sequences were generated at two different vowel rates (1 and 4 Hz) by adjusting the duration of the silent intervals between consecutive vowel stimuli. The presentation level for each vowel in the sequence was 65 dB sound pressure level (SPL).

The target vowel sequence was presented with a masker stimulus. The masker was gated on 500 ms before the onset of the target and gated off 500 ms after the offset of the target using 1-ms raised cosine ramps. In Exp. IA, the masker was a steady-state noise whose long-term spectrum matched that of the vowel stimuli. In Exp. IB, the noise masker was sinusoidally amplitude-modulated with a modulation depth of 100% and modulation rates of 0.5, 1, 2, 4, 6, 8, and 16 Hz in separate conditions.

All stimuli were generated digitally at a sampling frequency of 22 050 Hz. During the experiment, the listener was seated in a single-walled sound attenuated booth, and the stimuli were presented monaurally through a 24-bit sound card (MOTU Microbook II, Mark of the Unicorn, Inc.) and a pair of headphones (HD-280 Pro, Sennheiser Electronic GmbH & Co.KG).

3. Procedure

For both Exps. IA and IB, the listener was instructed to identify the vowels in the vowel sequence presented on each trial in the correct order. The five potential vowels were shown as five boxes on a computer monitor in front of the listener (labeled with phonetic symbols and example carrier words). On each trial, the listener identified the vowel sequence by clicking the corresponding boxes using a mouse.

For each listener, Exp. IA was conducted before Exp. IB. In Exp. IA, the psychometric function relating the recognition performance and signal-to-noise ratio (SNR) for the steady-state masker was estimated using the Updated Maximum-likelihood (UML) procedure (Shen and Richards,

2012; Shen *et al.*, 2015a; Shen *et al.*, 2015b). This procedure formulated the psychometric function as a logistic function

$$\text{Proportion Correct} = \gamma + (1 - \gamma - \lambda) [1 + e^{-\beta(\text{SNR} - \alpha)}]^{-1}, \quad (1)$$

where γ represents the chance performance (i.e., $\gamma = 0.2$). The shape of the psychometric function was defined by three parameters: α , β , and λ . The parameter α was the performance threshold, the parameter β was related to the slope of the psychometric function (as β increases, the slope of the function becomes steeper), and the parameter λ described the performance plateau at high SNRs. The prior distribution for α was a uniform distribution between -20 and 10 dB, the prior distribution for β was a uniform distribution between 0.1 and 10 on a logarithmic scale, and the prior distribution for λ was a uniform distribution between 0 and 0.2 . Each UML procedure started with an initial SNR of 5 dB, and the SNRs in subsequent trials were increased or decreased adaptively based on responses collected from earlier trials in the same track. This adaptive stimulus sampling in the UML procedure was used to minimize the expected variability in the estimates of the α , β , and λ parameters. The UML procedure terminated when a total of 20 reversals in the adaptive track was reached.

Five of the listeners were tested with a vowel rate of 1 Hz, while the other five were tested with a vowel rate of 4 Hz. At the beginning of data collection, four familiarization blocks were run using an IRN iteration of 16 iterations. Then, the three IRN iterations (1 , 4 , and 16 iterations) were tested in random order and repeated three times with independent random sequences. Once data collection was complete, the experimental trial under the same IRN iteration was pooled together for each listener and fitted with the logistic psychometric function as specified in Eq. (1) resulting in the final estimates of α , β , and λ .

For Exp. IB, each subject was tested with the same vowel rate as in Exp. IA. The seven masker modulation rates were tested in random order. For each modulation rate, the three IRN iterations were tested in random order. For each IRN iteration, one experimental block of 42 trials was tested. Within each experimental block, seven SNRs (-10 , -7.5 , -5 , -2.5 , 0 , 2.5 , and 5 dB) were presented six times using the method of constant stimuli. Once one block was run for each combination of IRN iteration and masker modulation rate, the above procedure was repeated two more times with independent test sequences. Three of the listeners (two tested using the 1 -Hz vowel rate and one tested using the 4 -Hz vowel rate) only completed two blocks for certain combinations of modulation rate and IRN iteration due to limited availability.

B. Behavioral results

The upper half of Table I shows the estimated psychometric-function parameters (α , β , and λ) for various IRN iteration conditions and for both the 1 and 4 Hz vowel rates from Exp. IA. Mixed-design analyses of variance (ANOVA) were conducted treating IRN iteration as the within-subject factor and

TABLE I. The estimated psychometric-function parameters for vowel recognition in steady-state noise from Exp. IA and Exp. II. The estimated parameters are associated with the SNR at the performance threshold (α), the slope of the psychometric function (β), and the upper performance plateau at high SNRs (λ). The standard errors of the mean are shown in parentheses.

	Vowel rate (Hz)	IRN iterations	α (dB)	β	λ
Exp. IA	1	1	-3.67 (1.14)	0.60 (0.12)	0.09 (0.07)
		4	-4.31 (0.99)	0.62 (0.06)	0.08 (0.07)
		16	-3.76 (1.96)	0.55 (0.12)	0.05 (0.03)
	4	1	-3.50 (0.96)	0.72 (0.15)	0.27 (0.06)
		4	-3.51 (1.23)	0.54 (0.12)	0.23 (0.07)
		16	2.27 (2.26)	0.36 (0.17)	0.11 (0.07)
Exp. II	1	1	-2.85 (0.37)	0.46 (0.03)	0.08 (0.04)
	4	1	-2.95 (0.53)	0.58 (0.04)	0.28 (0.08)

vowel rate as the between-subject factor. In separate analyses, α , β , and λ were treated as the dependent variables. For all three parameters, there was no significant effect of vowel rate [for α : $F(1, 8) = 2.40$, $p = 0.160$, $\eta^2 = 0.23$; for β : $F(1, 8) = 0.40$, $p = 0.546$, $\eta^2 = 0.05$; for λ : $F(1, 8) = 3.54$, $p = 0.097$, $\eta^2 = 0.31$] or IRN iteration [for α : $F(1.12, 8.93) = 3.47$, $p = 0.093$, $\eta^2 = 0.30$, Greenhouse-Geisser corrected; for β : $F(2, 16) = 1.10$, $p = 0.360$, $\eta^2 = 0.12$; for λ : $F(1.06, 8.49) = 1.89$, $p = 0.205$, $\eta^2 = 0.19$, Greenhouse-Geisser corrected]. There was no significant interaction between IRN iteration and vowel rate [for α : $F(1.12, 8.93) = 2.96$, $p = 0.118$,

$\eta^2 = 0.27$, Greenhouse-Geisser corrected; for β : $F(2, 16) = 0.63$, $p = 0.543$, $\eta^2 = 0.07$; for λ : $F(1.06, 8.49) = 0.79$, $p = 0.408$, $\eta^2 = 0.09$, Greenhouse-Geisser corrected]. Therefore, IRN iteration did not affect the psychometric function for vowel recognition in steady-state noise. Although there was a general trend that the upper performance plateau [i.e., lower λ values, see Eq. (1)] was higher for the 1-Hz vowel rate, this trend did not reach statistical significance due to large individual differences.

The top rows of Figs. 1 and 2 show the vowel identification score from Exp. IB as a function of modulation rate for the 1- (Fig. 1) and 4-Hz (Fig. 2) vowel rates. Results plotted in the middle and bottom panels relate to model predictions and will be described later. The average vowel recognition performance across listeners exhibited evident dependencies on SNR and masker modulation rate. An increase in either SNR or modulation rate led to an increase in performance. The effect of modulation rate was the strongest at lower SNRs, and a similar interaction between these two factors has been observed in previous studies (e.g., Shen *et al.*, 2015b). On the other hand, the effect of IRN iteration was not obvious. This indicates that the manipulation of TFS cues in the current experiment did not lead to large changes in the vowel recognition performance. Finally, the vowel rate seems to affect the upper performance plateau. For the 4-Hz vowel rate, the highest performance at high SNRs and high masker modulation rates was barely above 80% on average. Since the lower performance plateau (at low SNRs

FIG. 1. (Color online) Top: results from Exp. IB for the 1-Hz vowel rate. In each panel, the average proportion of correct vowel identification is plotted as a function of modulation rate. The different SNRs are indicated using different symbols. The three IRN iterations are arranged in columns. Middle: vowel recognition performance predicted by a dip-listening model (see text for details). Bottom: the deviation between the behavioral data and model predictions (ΔPC).

FIG. 2. (Color online) Same as Fig. 1, but for the results from Exp. IB for the 4-Hz vowel rate.

and low modulation rates) was determined by the chance performance of 20-% correct, the effect of vowel rate on the upper performance limit also means that the dynamic range of the psychometric function was narrower for the 4 than the 1-Hz vowel rate.

C. Modeling results

The psychophysical data from Exp. IB were compared to the predictions of a short-term SNR based dip-listening model identical to the one described in Shen and Pearson (2017). This model divided the duration of each vowel into brief time windows, in which the short-term SNR was evaluated. The expected vowel recognition performance was predicted based on a subset of the windows with the highest short-term SNRs, assuming ideal glimpsing efficiency. For each of the time windows in the subset, a performance score was predicted based on the psychometric function for vowel recognition in steady-state noise (obtained from Exp. IA). The overall predicted vowel recognition performance was computed as the expected value of the performance scores from all time windows in the subset. Therefore, this model provided a quantitative link between the psychometric functions for vowel recognition in steady-state and fluctuating maskers.

The current model included only two parameters: the duration of the time windows and the minimum duration required for vowel recognition (i.e., the minimum numbers of time windows contributing to vowel recognition). These

two parameters were set to 8 and 32 ms, respectively. These values achieved the best fit to the vowel recognition data reported by Shen and Pearson (2017). These two parameters were fixed instead of being derived from fitting the model to the data from Exp. IB so that the model predictions were independent of the data collected for fluctuating maskers. If the model predicts the performance for vowel identification in fluctuating maskers well, it means that the short-term-SNR based dip-listening model is sufficient to account for the performance difference between the steady-state and fluctuating maskers (i.e., the FMB).

The model predictions and the deviation between the data and model predictions (ΔPC), averaged across listeners, are plotted in the middle and bottom panels of Figs. 1 and 2, respectively. For each listener, the model predictions were derived from the psychometric function for vowel recognition in steady-state maskers measured in Exp. IA for the same listener. Since the psychometric function for steady-state maskers did not significantly depend on IRN iteration, data from all three IRN iteration conditions of Exp. IA were pooled together and re-fitted using Eq. (1) in order to improve the reliability of the estimated psychometric function. For this reason, the performance scores predicted by the model for fluctuating maskers were the same for all three IRN iteration conditions (the middle rows of Figs. 1 and 2). The model predicted performance resembled the general trends in the behavioral data (top rows of Figs. 1 and 2). For example, it predicted that the proportion correct monotonically increased with increasing modulation rate from 0.5 to

16 Hz and this predicted effect was stronger at lower SNRs. In the model, this predicted effect of modulation rate for low SNRs was based on the fact that at very low modulation rates the target vowel may completely reside in a single peak or dip of the masker envelope and the overall expected performance was largely governed by the duty cycle of the masker amplitude modulation (i.e., 50% for sinusoidal modulation).

Detailed comparisons between the experimental data and model predictions suggest that the deviation between the two was not governed by random errors, rather it exhibited certain systematic patterns (lower panels in Figs. 1 and 2). In a majority of the conditions, the deviation between the behavioral and modeling results (ΔPC) was negative, indicating an overestimation of the behavioral data by the model. Because the model predictions represent the performance as expected from an ideal dip-listener capable of taking full advantage of the high-SNR periods in the masker envelope, the overestimation by the model observed here reflects the inadequacy of the short-term-SNR based dip-listening model to resemble the glimpsing efficiency of actual listeners.

A mixed-design analysis of variance (ANOVA) was conducted treating IRN iteration, modulation rate, and SNR as the within-subject factors and vowel rate as the between subjects factor. The proportion correct scores from the model and the behavioral experiment were converted to rationalized arcsine units (RAU; [Studebaker, 1985](#)), and the difference between the converted scores (ΔRAU) was treated as the dependent variable. Significant main effects were found for SNR [$F(6, 48) = 21.166, p < 0.001, \eta^2 = 0.726$] and modulation rate [$F(6, 48) = 12.605, p < 0.001, \eta^2 = 0.612$]. Neither the effect of IRN [$F(2, 16) = 1.643, p = 0.224, \eta^2 = 0.170$] or vowel rate [$F(1, 8) = 0.627, p = 0.451, \eta^2 = 0.073$] was significant. Significant interactions were found between modulation rate and SNR [$F(36, 288) = 1.612, p = 0.018, \eta^2 = 0.168$], between IRN and modulation rate [$F(12, 96) = 2.612, p = 0.005, \eta^2 = 0.246$], between SNR and modulation rate [$F(36, 288) = 1.612, p = 0.018, \eta^2 = 0.168$], and between modulation rate and vowel rate [$F(6, 48) = 3.433, p = 0.007, \eta^2 = 0.300$]. None of the other interactions reached statistical significance ($p > 0.05$).

The significant effect of SNR indicated that the difference between the data and the model output (ΔRAU) was sensitive to the SNR. As the SNR decreased, ΔRAU became more negative, indicating a proportion of performance degradation in the experimental data not captured by the model at low SNRs. Moreover, the model was more accurate at the very high and very low modulation rates, while the moderate rates showed the greatest deviation from the behavioral results (i.e., most negative ΔRAU values). *Post hoc* paired *t* tests (Bonferroni corrected) showed that the ΔRAU value at 2 Hz was significantly lower than those at 0.5 Hz [$t(9) = -5.889, p = 0.017$], 8 Hz [$t(9) = -5.936, p = 0.048$], and 16 Hz [$t(9) = -7.76, p = 0.009$]. The ΔRAU value at 16 Hz was also significantly higher than those at 4 Hz [$t(9) = 7.76, p = 0.009$] and 6 Hz [$t(9) = 5.827, p < 0.001$]. Overall, the comparison between the behavioral results and model predictions suggests that the model overestimated human listeners' vowel recognition performance for both slow (e.g., 1 Hz) and fast (e.g., 4 Hz) vowel sequences. The overestimation was

greatest under the conditions with low SNRs (e.g., -10 dB) and moderate masker modulation rates (e.g., between 2 and 6 Hz). The effect of TFS cues was not evident.

III. EXPERIMENT II: THE EFFECT OF TEMPORAL REGULARITY

A. Methods

1. Listeners

Six native English speakers (one female, mean age = 19.83 years) who did not participate in Exp. I were recruited from the student population of Indiana University. All listeners had audiometric thresholds equal to or better than 15 dB HL between 250 and 4000 Hz in both ears. The test ear for the experiment was chosen by selecting the one with the lower PTA threshold, or the left ear in case of identical PTA thresholds across the ears. Listeners were compensated at an hourly rate for their participation. The experimental protocol was approved by the Institutional Review Board at Indiana University. Informed consent was obtained from all listeners before data collection began.

2. Stimuli

Listeners identified vowel sequences in simultaneous maskers. The same procedure was used to generate and present the stimuli as in Exp. I, unless otherwise stated. Each vowel in the target sequence was synthesized using an IRN. The IRN was generated using a delay of 4 ms and an IRN iteration of 1 iteration. Unlike Exp. I, the target sequence was not isochronous. Instead, the onsets of the target vowels were varied by adding a degree of temporal jitter to the temporal gaps between the consecutive vowels. The temporal jitter was implemented using the following procedure. Each sequence contained five vowels, leading to four temporal gaps. For an overall vowel rate of 1 Hz, the cumulative duration of silent gaps was fixed at 3600 ms (total sequence duration: 4100 ms); for an overall vowel rate of 4 Hz, the cumulative duration of silent gaps was 600 ms (total sequence duration: 1100 ms). The cumulative gap duration was divided into four parts according to a ratio of 1:2:3:4 and randomly assigned to the four gaps on every trial. This procedure was used so that the onsets of the vowels in the target sequence became less predictable, while the consecutive vowels were never extremely close or far from each other.¹

The masker in the current experiment was either a steady-state noise as in Exp. IA or a sinusoidally amplitude-modulated noise as in Exp. IB. When modulated, modulation rates were 0.5, 1, 2, 4, 8, and 16 Hz. The masker was gated on 500 ms before the onset of the target sequence and was gated off 500 ms after the offset of the target sequence using 1-ms raised-cosine ramps.

3. Procedure

Experiment II used the same experimental procedure as Exp. IA and IB where listeners were instructed to identify the vowels in the presented sequence in the correct order by

selecting the corresponding boxes on the computer monitor. All listeners were tested using vowel rates of 1 and 4 Hz, in random order. For each vowel rate, the seven different masker modulation rates (0, 0.5, 1, 2, 4, 8, 16 Hz) were tested in random order. Note that a modulation rate of 0 Hz is used to indicate the steady-state masker for simplicity. Within each masker modulation rate, one experimental block was run in which the seven SNRs (−10, −7.5, −5, −2.5, 0, 2.5, and 5 dB) were presented six times using the method of constant stimuli. Once one block was run for each combination of vowel rate and masker modulation rate, the above procedure was repeated two more times with independent test sequences.

B. Behavioral results

The bottom half of Table I shows the estimated psychometric-function parameters (α , β , and λ) for the steady-state maskers from Exp. II. Paired t tests were performed for α , β , and λ to investigate the effect of vowel rate. Vowel rate did not have a significant effect on α [$t(5) = 0.14$, $p = 0.894$] or β [$t(5) = -1.76$, $p = 0.139$], and it only had a marginally significant effect on λ [$t(5) = 2.32$, $p = 0.068$]. Similar to Exp. IA, the higher vowel rate was associated with a lower upper performance limit for some listeners.

The behavioral results from the modulated-masker conditions of Exp. II are shown in the top row of Fig. 3. The middle and bottom rows show results related to the model predictions which will be discussed later. Similar to the results of Exp. IB, the vowel recognition performance improved as either the SNR or masker modulation rate increased. These two factors also interacted with each other, so that the effect of modulation rate was most evident at lower SNRs. When the vowel rate was 4 Hz, the upper performance limit at high SNRs and high modulation rates became lower than 80 % correct, leading to a narrow dynamic range for the psychometric function under this condition.

To visualize the effect of temporal regularity, data from the current experiment tested using temporally irregular vowel sequences are replotted in Fig. 4, together with those from Exp. IB tested using isochronous sequences under the corresponding conditions. Although temporal regularity did not affect vowel recognition in steady-state maskers (unconnected symbols to the left of each panel), irregular sequences led to degradation in vowel recognition performance when the masker was amplitude modulated. This effect of temporal regularity was observed across a wide range of SNRs and a wide range of masker modulation rates. For isochronous vowel sequences (circles), the vowel recognition scores for modulated maskers were almost all above those for steady-state maskers, exhibiting benefits from fluctuation maskers (i.e., the FMB). On the other hand, for irregular vowel sequences (triangles), the recognition scores for modulated maskers were not always higher than those for steady-state maskers. The FMB was observed only at low SNRs and high masker modulation rates. At high SNRs and certain moderate modulation rates (e.g., see the bottom right panel near the 2-Hz masker modulation rate), amplitude-modulating the masker led to no benefit but additional performance degradation compared to the unmodulated maskers.

FIG. 3. (Color online) Average proportion of correct vowel identification as a function of modulation rate from Exp. II (top row) and as predicted by the dip-listening model (middle row). The deviation between the experimental results and model predictions (ΔPC) are shown in the bottom row. The 1-Hz (left) and 4-Hz (right) vowel rates are arranged in the two columns. Within each panel, various SNRs are indicated using different symbols.

To further investigate the effect of temporal regularity, two mixed-design ANOVA were conducted separately for the 1- and 4-Hz vowel rates, treating modulation rate (from 0.5 to 16 Hz) and SNR (from −10 to 5 dB) as the within-subject factors and temporal regularity as the between-subject factor. Proportion correct was converted into RAU units and treated as the dependent variable. For the 1-Hz vowel rate, significant main effects were found for modulation rate [$F(5, 45) = 63.591$, $p < 0.001$, $\eta^2 = 0.876$], SNR [$F(6, 54) = 123.173$, $p < 0.001$, $\eta^2 = 0.932$], and temporal regularity [$F(1, 9) = 5.87$, $p = 0.038$, $\eta^2 = 0.395$]. Significant interactions were found between modulation rate and SNR [$F(30, 270) = 8.312$, $p < 0.001$, $\eta^2 = 0.48$] as well as a three-way interaction among temporal regularity, modulation rate, and SNR [$F(30, 270) = 1.732$, $p = 0.013$, $\eta^2 = 0.161$]. No significant interaction between temporal regularity and SNR [$F(6, 54) = 1.721$, $p = 0.134$, $\eta^2 = 0.161$] or between temporal regularity and modulation rate [$F(5, 45) = 0.7614$, $p = 0.69$, $\eta^2 = 0.064$] was found. For the 4-Hz conditions, significant effects were found for modulation rate [$F(5, 45) = 39.091$, $p < 0.001$, $\eta^2 = 0.813$] and SNR [$F(6, 54) = 63.641$, $p < 0.001$, $\eta^2 = 0.876$]. The effect of temporal regularity was

FIG. 4. Average proportion of correct vowel identifications for an IRN iteration of 1 for temporally regular (from Exp. IB, squares) and temporally irregular (from Exp. II, triangles) vowel sequences. Data shown are for SNRs of -10 , -5 , 0 , and 5 dB (from top to bottom). The 1-Hz (left) and 4-Hz (right) vowel rates are arranged in the two columns. In each panel, the results for the steady-state masker (SS) are plotted to the left. Error bars indicate $\pm$ one standard error of the mean.

not significant [$F(1, 9) = 2.536$, $p = 0.146$, $\eta^2 = 0.22$]. Significant interactions were found between modulation rate and SNR [$F(30, 270) = 4.351$, $p < 0.001$, $\eta^2 = 0.326$] as well as between modulation rate and temporal regularity [$F(5, 45) = 3.703$, $p = 0.007$, $\eta^2 = 0.292$]. No significant effect was found for the interaction between temporal regularity and SNR [$F(6, 54) = 0.684$, $p = 0.684$, $\eta^2 = 0.071$] or for the three-way interaction between temporal regularity, SNR, and modulation rate [$F(30, 270) = 1.175$, $p = 0.250$, $\eta^2 = 0.115$].

For the 1-Hz vowel rate, the significant effect of temporal regularity shows that listeners' vowel identification performance was sensitive to the temporal jitter added in Exp. II. For the 4-Hz rate, temporal regularity was not a significant factor by itself, but it significantly interacted with modulation rate. As the modulation rate increased, the performance improved more for regular targets than irregular targets with the largest effect of temporal regularity found at higher modulation rates (e.g., 8 and 16 Hz). The observed difference between the 1- and 4-Hz vowel rates may be

related to the decreased dynamic range of the psychometric function seen at the faster vowel rate. It is also worth noting that the data for the temporally regular target (i.e., Exp. IB) at vowel rates of 1 and 4 Hz were from two different groups of listeners, therefore it is possible that the difference between the two vowel rates reflected individual differences.

C. Modeling results

The middle row of Fig. 3 shows the model predictions from the same short-term SNR based dip-listening model as in Exp. I. These predictions were based on the psychometric functions estimated from the behavioral results from the unmodulated masker. The deviation of the behavioral results from the model predictions (ΔPC) is shown in the bottom row of Fig. 3. The ΔPC values were negative for both vowel rates and for all modulation rates and SNRs tested, indicating an overestimation of behavioral data by the model. A repeated-measures ANOVA was conducted treating modulation rate, vowel rate, and SNR as the within-subject factors. The model and behavioral proportion correct scores were converted into RAU units, and the difference between the behavioral and model RAU scores (ΔRAU) was treated as the dependent variable. Significant effects were found for modulation rate [$F(5, 25) = 7.618$, $p < 0.001$, $\eta^2 = 0.604$] and SNR [$F(6, 30) = 3.246$, $p = 0.014$, $\eta^2 = 0.394$], while the effect of vowel rate was not significant [$F(1, 5) = 0.739$, $p = 0.429$, $\eta^2 = 0.129$]. There was a significant interaction between SNR and modulation rate [$F(30, 150) = 2.15$, $p = 0.001$, $\eta^2 = 0.301$]. Other interactions did not reach statistical significance ($p > 0.05$).

The significant effects and interactions indicate that the data-model deviation was not random, but systematically depended on SNR and modulation rate. The overestimation by the model tended to be larger (i.e., more negative ΔRAU) as the SNR decreased, and the magnitude of the deviation was greatest at moderate masker modulation rates. The lack of significant effect of vowel rate indicates that the mechanism underlying the data-model deviation was not very sensitive to the overall rate at which the vowels were presented.

Compared to the data-model deviations shown in Figs. 1 and 2, the deviations from the current experiment were larger in magnitude. In other words, when the target sequence was made less predictable, the vowel-recognition performance deviated further from the expected performance from the ideal dip-listening model. For both isochronous (Exp. IB) and irregular (Exp. II) vowel sequences, the model predictions were based on the psychometric function for the corresponding type of vowel sequences in steady-state maskers. Therefore, the model predictions were corrected for the baseline performance difference caused by the introduction of temporal irregularity. The difference in ΔRAU between the temporally irregular and isochronous target sequences reflected an effect of target temporal regularity that was specific to amplitude-modulated maskers.

IV. GENERAL DISCUSSION

A. The effect of TFS cues

The main purpose of the current study was to investigate the effects of TFS cues and temporal regularity on the

efficiency of glimpsing target speech in fluctuating maskers. The amount of TFS cues in the target was manipulated through the IRN processing with increasing IRN iterations corresponding to greater TFS information. In Exp. IB, it was found that the vowel recognition performance was not sensitive to manipulation of the IRN iteration.

The lack of effect of IRN iteration found for amplitude-modulated maskers in Exp. IB contradicts the importance of TFS cues in the FMB demonstrated in previous studies (Qin and Oxenham, 2003; Hopkins and Moore, 2009). However, it was not completely unexpected as previous research with degraded target TFS have shown mixed results. Apoux and Healy (2011) measured consonant recognition using vowel-consonant-vowel utterances words in four types of maskers, including steady-state and amplitude-modulated noises. When the target TFS was disrupted, the recognition performance became poorer for both the steady-state and amplitude-modulated maskers, by a similar amount. That is, the target TFS did not lead to a differentially larger impact on the fluctuating maskers. Freyman *et al.* (2012) studied the effect of the target TFS cues by comparing recognition performance using either naturally-produced or whispered speech. Although the whispered speech exhibited degraded TFS cues, it did not lead to reduced benefits from fluctuating maskers. Steinmetzger and Rosen (2015) investigated the importance of the target TFS cues using vocoded speech with varied amount of periodicity presented in steady-state or amplitude-modulated maskers. No difference in speech intelligibility was found for aperiodic and periodic target speech in either steady-state or amplitude-modulated maskers. On the other hand, it was shown that the periodicity in the masker significantly aided speech recognition, suggesting a potentially stronger effect of the masker TFS cues than target TFS cues.

In summation, this experiment showed no compelling evidence for the hypothesis that TFS-based perceptual segregation between the target and masker plays an important role in speech glimpsing in fluctuating maskers. It may be necessary to include independent manipulations of target and masker TFS cues in future studies.

B. Temporal regularity

Experiment II examined the impact of temporal regularity in the target vowel sequence on the ability to recognize the vowels in a modulated noise masker. Results showed a significant effect of temporal regularity, where temporally unpredictable stimuli led to a significant decrease in vowel recognition performance at all modulation rates for the 1-Hz stimuli and led to a significant decrease of performance at the higher modulation rates for the 4-Hz stimuli. Listeners' inefficiency in glimpsing, shown as the deviation between the behavioral data and the predicted performance from the ideal dip-listening model, is greater for the temporally irregular stimuli. This indicates that the irregular temporal pattern in the target may have undermined the listeners' benefits from the high-SNR periods in the modulated maskers.

Temporal regularity of the target sequence may have influenced glimpsing efficiency in several ways, and these potential mechanisms may not be mutually exclusive. One

hypothesis is that to glimpse the target speech optimally, listeners need to predict when the high-SNR periods in the target-masker mixture will occur. Freyman *et al.* (2012) compared sentence recognition in noises that were amplitude-modulated by either a temporally regular, 8-Hz square wave or the envelope of running speech. Speech recognition was poorer for speech-envelope than square-wave modulated maskers, indicating that temporal regularity facilitated glimpsing the target speech in fluctuating maskers. This result was replicated by Schubotz *et al.* (2016). In their study, sentence recognition in noises amplitude-modulated by an 8-Hz sinusoid or speech envelope was measured. The speech recognition threshold was consistently lower for the sinusoidally modulated masker. More interestingly, models based on short-term SNR, for example the Extended Speech Intelligibility Index (ESII) model (Rhebergen and Versfeld, 2005; Rhebergen *et al.*, 2006), cannot predict the effect of temporal regularity. For the speech-envelope masker, the ESII model predicted better performance than the experimental data. The short-term SNR based models, including the one used in the current study, assumes full *a priori* knowledge about the target and masker envelopes. Because the instances in the target-masker mixture with the highest SNRs are known to the models, they can be effectively utilized to improve speech intelligibility. For human listeners, on the other hand, the high-SNR periods are estimated online. When either the target or masker envelope is unpredictable, the timing for glimpsing the target speech would be poorly estimated, leading to poorer performance than expected from an ideal dip-listener.

An alternative hypothesis is that the effect of temporal regularity is related to stream segregation. For example, Andreou *et al.* (2011) studied the effect of temporal regularity on stream segregation using sequences of alternating A- and B-tones of different frequencies. The inter-tone-intervals for each of the two sequences were randomly varied between 100 and 400 ms. Under certain condition (i.e., when A- and B-tones were two semitones apart in frequency), as one of the two sequences were made temporally isochronous, the perceived segregation between the two sequences was enhanced. Rimmele *et al.* (2012) presented a high-frequency tone sequence and a low-frequency tone sequence to their listeners, who were asked to detect a high-intensity deviant in one of the two sequences (i.e., the target sequence). Introducing temporal regularity to either the target or the non-target sequence led to improved task performance and enhanced segregation between the two sequences.

These observations could be understood using the predictive processing model (e.g., Bendixen *et al.*, 2012; Bendixen, 2014), which assumes that the auditory system keeps looking for predictable sequential patterns. Auditory inputs that fall within a single predictive pattern would be combined into a perceptual stream. Presence of deviations from the predictive pattern would promote segregation. When two uncoupled predictive patterns are found, two highly segregated streams are formed. In Exp. IB, both the target vowel sequence and the sinusoidally amplitude-modulated masker exhibited highly predictable temporal patterns, but these two patterns were incoherent with each other.

According to the predictive coding model, this would lead to very strong segregation of the target and masker. On the other hand, in Exp. II, the target was made unpredictable, which would lead to weakened segregation of the target and masker compared to Exp. IB.

C. Masker modulation rates

In the current study, vowel recognition was measured as a function of masker modulation rate. In both Exps. IB and II, it was found that at low SNRs (e.g., -10 dB) the performance improved as the masker modulation rate increased from 0.5 to 16 Hz. Near 0.5 Hz, the performance is dominated by whole-word masking (see the top row in Fig. 5 for a visual illustration of whole-word masking). That is, at very low masker modulation rates, the target vowel may coincide with a masker peak, resulting in a very low effective SNR, or a masker dip, resulting in a very high effective SNR. These two situations counterbalance each other, and the net effect on vowel recognition is largely determined by the duty-cycle of the masker envelope (i.e., 50% for sinusoidal modulation). Near 16 Hz, the duration of the target vowel contains multiple cycles of masker modulation, and the vowel-recognition performance reflects the listener's efficiency in glimpsing the target vowel in the dips of the masker (see the bottom row in Fig. 5 for a visual illustration). The ideal dip-listening model used in the current study is able to capture the transition from whole-word masking to glimpsing as the masker modulation rate increases from 0.5 to 16 Hz, as well as the interaction between the masker modulation rate and SNR.

One noticeable feature of the model predictions is that the model systematically overestimated the vowel-recognition performance. In Exp. IB, this was observed at low SNRs (see the bottom rows of Figs. 1 and 2), while in Exp. II, the overestimation of performance was found in all conditions (see the bottom row of Fig. 3). The data-model deviation indicates failures of the ideal dip-listening model and the inability of listeners to estimate the best instances to glimpse the target speech in fluctuating maskers. Furthermore, the data-model

FIG. 5. Illustrations of the temporal relationship between the target vowels (rectangles) and masker envelope (shaped areas) for masker modulation rates of 1 Hz (top), 4 Hz (middle), and 16 Hz (bottom). Two examples of masker envelope with opposite modulation phases are shown in the left and right columns.

deviation depended on masker modulation rate with the greatest amount of deviation (i.e., most negative ΔPC) found at moderate modulation rates (2–6 Hz).

The masker modulation may have an interfering effect on the recognition of vowel sequences when the target and masker fluctuate at similar rates. Many researchers have described the rate-based interference as “modulation masking.” It has been argued that modulation masking may play an important role in speech recognition in noise maskers (Stone *et al.*, 2012). A number of speech intelligibility models implemented modulation masking using a band of band-pass modulation filters within each of the peripheral frequency filters (e.g., Jørgensen and Dau, 2011). The modulation filters describe the auditory system's selectivity in the modulation domain, the target and masker modulation interfere with each other when they fall within the same modulation filter. Fogerty *et al.* (2016) demonstrated the modulation-rate based interference using speech stimuli. In one condition of their experiment, the target was naturally produced sentences and the masker was broadband noise amplitude-modulated using speech envelope. Further, time compression or expansion was applied to the masker to manipulate the masker modulation spectrum. It was shown that the speech recognition performance was poorer when the modulation spectra of the target and masker overlapped more (i.e., exhibiting similar mean modulation rates).

Shen and Pearson (2017) tried to replicate this finding using synthesized vowel sequences as the target and amplitude-modulated noises as the masker. Both the vowel rate and the masker modulation rate were systematically varied, but signs of modulation selectivity (i.e., poorer performance when the target and masker rates were similar) were not found. On the other hand, the current study found that the amount of masking effect not accounted for by the ideal dip-listening model depended on masker modulation rate. The regions with the most data-model deviation (typically between 2 and 6 Hz) generally matched to the rates at which the target vowels were presented (1 or 4 Hz). The main difference between the current study and that of Shen and Pearson (2017) is that the vowel sequences in the current study were processed to remove either TFS information or temporal regularity. Therefore, the effect of masker modulation rate is likely to depend on (1) the perceptual segregation between the target and masker and (2) predictability of the high-SNR periods in the target-masker mixture.

Contrary to the modulation masking hypothesis, the greatest interfering effect from the masker (i.e., the most negative ΔPC) was not observed at 1 Hz for the 1-Hz vowel rate or at 4 Hz for the 4-Hz vowel rate. This may be due to the fact that the tuning in the modulation domain is quite broad (Dau *et al.*, 1997). On the other hand, it is also possible that additional processes other than modulation masking was involved. For example, the observed effect of masker modulation rate may be understood according to the theory of dynamic attention (Jones and Boltz, 1989; Barnes and Jones, 2000). Between 2 and 6 Hz (see the middle row of Fig. 5), the recognition of each target vowel relies on only one glimpse, and the vowel-recognition performance crucially depends on whether the listener's attention is

accurately aligned with the high-SNR period during the target vowel. For modulation rates higher than 6 Hz (see the bottom row of Fig. 5), since the high-SNR periods occur multiple times over the duration of each target vowel at a regular rate, it may be easier for the auditory system to calibrate the attentional rhythm to follow the instances of high-SNR periods. For modulation rates lower than 2 Hz (see the top row of Fig. 5), the duration of each high-SNR period is relatively long and errors in identifying the best glimpsing time matters less. Therefore, when errors in estimating the best glimpsing time occur—for example caused by failures in segregating the target and masker or unpredictable target rhythm—vowel recognition with moderate masker modulation rates is most affected.

In Exp. IB, for the 4-Hz vowel rate, the vowel recognition performance for a masker modulation rate of 6 Hz stood out from surrounding modulation rates (e.g., 4 and 8 Hz). For this modulation rate, the data-model deviation (the bottom row of Fig. 2) was much larger. A similar dip in ΔPC was not found for the 1-Hz vowel rate. The reason for the aberrant results for the 4-Hz vowel rate and 6-Hz masker modulation rate is not clear. One potential explanation is that the chances for glimpsing the target vowel may depend on the beating between the vowel sequence and masker envelope. As illustrated in Fig. 6, when the vowel rate is 4 Hz and the masker modulation rate is either 4 Hz (top) or 8 Hz (bottom), the rate at which glimpses of the target vowels are present (indicated by the arrows) is at the vowel rate. On the other hand, when the masker modulation rate is 6 Hz (middle), the short-term SNR fluctuates at 2 Hz. If the listener's attention entrains to the rhythm of the short-term SNR, then approximately half of the vowels in the sequence would be overlooked.

V. SUMMARY

The current study investigated the roles of target TFS cues and temporal regularity in vowel identification in modulation masking. This was done through both the analysis of behavioral data as well as comparing it to an ideal dip-listening model based on the evaluation of short-term SNRs. The findings showed evident effect of TFS cues in improving

FIG. 6. Illustrations of the available glimpses of the target vowels, indicated by the arrows, in Exp. I for a vowel rate of 4 Hz and masker modulation rates of 4 Hz (top), 6 Hz (middle), and 8 Hz (bottom).

vowel recognition in modulated maskers. Additionally, compared to temporally regular vowel sequences, listeners were less efficient in glimpsing the target vowels with temporally irregular vowel sequences. When temporal regularity was degraded, the ideal dip-listening model overestimated the behavioral results. This data-model deviation is greatest at moderate modulation rates, which may be related to modulation masking or attention dynamics.

ACKNOWLEDGMENTS

This work was supported by the NIH Grant Nos. R21 DC013406 (Co-PIs: Virginia M. Richards and Yi Shen) and R01DC013538 (Co-PIs: Gary R. Kidd and J. Devin McAuley). Anna M. Hopkins provided editorial assistance.

¹This method of introducing temporal regularity only varies the onsets of the 2nd, 3rd, and 4th vowels relative to the onset of the masker. This means that any effect of temporal regularity observed in the current experiment may differ depending on the vowel position in the sequence.

- Andreou, L. V., Kashino, M., and Chait, M. (2011). "The role of temporal regularity in auditory segregation," *Hear. Res.* **280**, 228–235.
- Apoux, F., and Healy, E. W. (2011). "Relative contribution of target and masker temporal fine structure to the unmasking of consonants in noise," *J. Acoust. Soc. Am.* **130**, 4044–4052.
- Assmann, P. F., and Summerfield, Q. (1994). "The contribution of waveform interactions to the perception of concurrent vowels," *J. Acoust. Soc. Am.* **95**, 471–484.
- Barnes, R., and Jones, M. R. (2000). "Expectancy, attention, and time," *Cogn. Psychol.* **41**, 254–311.
- Bendixen, A. (2014). "Predictability effects in auditory scene analysis: A review," *Front. Neurosci.* **8**, 60.
- Bendixen, A., Schröger, E., Ritter, W., and Winkler, I. (2012). "Regularity extraction from non-adjacent sounds," *Front. Psychol.* **3**, 143.
- Bernstein, J. G., and Brungart, D. S. (2011). "Effects of spectral smearing and temporal fine-structure distortion on the fluctuating-masker benefit for speech at a fixed signal-to-noise ratio," *J. Acoust. Soc. Am.* **130**, 473–488.
- Biberger, T., and Ewert, S. D. (2017). "The role of short-time intensity and envelope power for speech intelligibility and psychoacoustic masking," *J. Acoust. Soc. Am.* **142**, 1098–1111.
- 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.
- Dau, T., Kollmeier, B., and Kohlrausch, A. (1997). "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers," *J. Acoust. Soc. Am.* **102**, 2892–2905.
- 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.
- Fogerty, D., Xu, J., and Gibbs, B. E. (2016). "Modulation masking and glimpsing of natural and vocoded speech during single-talker modulated noise: Effect of the modulation spectrum," *J. Acoust. Soc. Am.* **140**, 1800–1816.
- Freyman, R. L., Griffin, A. M., and Oxenham, A. J. (2012). "Intelligibility of whispered speech in stationary and modulated noise maskers," *J. Acoust. Soc. Am.* **132**, 2514–2523.
- Gordon-Salant, S., and Fitzgibbons, P. J. (2004). "Effects of stimulus and noise rate variability on speech perception by younger and older adults," *J. Acoust. Soc. Am.* **115**, 1808–1817.
- Gustafsson, H. Å., and Arlinger, S. D. (1994). "Masking of speech by amplitude-modulated noise," *J. Acoust. Soc. Am.* **95**, 518–529.
- Hopkins, K., and Moore, B. C. (2009). "The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise," *J. Acoust. Soc. Am.* **125**, 442–446.
- Hopkins, K., and Moore, B. C. (2010). "The importance of temporal fine structure information in speech at different spectral regions for normal-hearing and hearing-impaired subjects," *J. Acoust. Soc. Am.* **127**, 1595–1608.

- Jones, M. R., and Boltz, M. (1989). "Dynamic attending and responses to time," *Psychol. Rev.* **96**, 459–491.
- 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.
- Klatt, D. H. (1980). "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.* **67**, 971–995.
- Lentz, J. J., and Marsh, S. L. (2006). "The effect of hearing loss on identification of asynchronous double vowels," *J. Speech Hear. Res.* **49**, 1354–1367.
- Miller, G. A., and Licklider, J. C. (1950). "The intelligibility of interrupted speech," *J. Acoust. Soc. Am.* **22**, 167–173.
- Oxenham, A. J., and Simonson, A. M. (2009). "Masking release for low- and high-pass-filtered speech in the presence of noise and single-talker interference," *J. Acoust. Soc. Am.* **125**, 457–468.
- Qin, M. K., and Oxenham, A. J. (2003). "Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers," *J. Acoust. Soc. Am.* **114**, 446–454.
- 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.
- Rimmele, J., Schröger, E., and Bendixen, A. (2012). "Age-related changes in the use of regular patterns for auditory scene analysis," *Hear. Res.* **289**, 98–107.
- Schubotz, W., Brand, T., Kollmeier, B., and Ewert, S. D. (2016). "Monaural speech intelligibility and detection in maskers with varying amounts of spectro-temporal speech features," *J. Acoust. Soc. Am.* **140**, 524–540.
- Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). "Speech recognition with primarily temporal cues," *Science* **270**, 303–304.
- Shearer, D. E., Molis, M. R., Bennett, K. O., and Leek, M. R. (2018). "Auditory stream segregation of iterated rippled noises by normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.* **143**, 378–387.
- Shen, Y., Dai, W., and Richards, V. M. (2015a). "A MATLAB toolbox for the efficient estimation of the psychometric function using the updated maximum-likelihood adaptive procedure," *Behav. Res. Methods* **47**, 13–26.
- Shen, Y., Manzano, N. K., and Richards, V. M. (2015b). "Psychometric functions for sentence recognition in sinusoidally amplitude-modulated noises," *J. Acoust. Soc. Am.* **138**, 3613–3624.
- Shen, Y., and Pearson, D. V. (2017). "Recognition of synthesized vowel sequences in steady-state and sinusoidally amplitude-modulated noises," *J. Acoust. Soc. Am.* **141**, 1835–1843.
- 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.
- Shofner, W. P., and Chaney, M. (2013). "Processing pitch in a nonhuman mammal (*Chinchilla laniger*)," *J. Comp. Psychol.* **127**, 142–153.
- Shofner, W. P., Morris, H., and Mills, M. (2018). "Perception of noise-vocoded tone complexes: A time domain analysis based on an auditory filterbank model," *Hear. Res.* **367**, 1–16.
- Steinmetzger, K., and Rosen, S. (2015). "The role of periodicity in perceiving speech in quiet and in background noise," *J. Acoust. Soc. Am.* **138**, 3586–3599.
- 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.
- Studebaker, G. A. (1985). "A 'rationalized' arcsine transform," *J. Speech Lang. Hear. Res.* **28**(3), 455–462.
- Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2011). "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE/ACM Trans. Audio Speech Lang. Process.* **19**, 2125–2136.
- Takahashi, G. A., and Bacon, S. P. (1992). "Modulation detection, modulation masking, and speech understanding in noise in the elderly," *J. Speech Lang. Hear. Res.* **35**, 1410–1421.
- Yost, W. A. (1996). "Pitch strength of iterated rippled noise," *J. Acoust. Soc. Am.* **100**, 3329–3335.