# Modulation masking and glimpsing of natural and vocoded speech during single-talker modulated noise: Effect of the modulation spectrum

Daniel Fogerty,[a] Jiaqian Xu,[b] and Bobby E. Gibbs II
*Department of Communication Sciences and Disorders, University of South Carolina, Columbia, South Carolina 29208, USA*

(Received 17 March 2016; revised 17 August 2016; accepted 25 August 2016; published online 19 September 2016)

Compared to notionally steady-state noise, modulated maskers provide a perceptual benefit for speech recognition, in part due to preserved speech information during the amplitude dips of the masker. However, overlap in the modulation spectrum between the target speech and the competing modulated masker may potentially result in modulation masking, and thereby offset the release from energetic masking. The current study investigated masking release provided by single-talker modulated noise. The overlap in the modulation spectra of the target speech and the modulated noise masker was varied through time compression or expansion of the competing masker. Younger normal hearing adults listened to sentences that were unprocessed or noise vocoded to primarily limit speech recognition to the preserved temporal envelope cues. For unprocessed speech, results demonstrated improved performance with masker modulation spectrum shifted up or down compared to the target modulation spectrum, except for the most extreme time expansion. For vocoded speech, significant masking release was observed with the slowest masker rate. Perceptual results combined with acoustic analyses of the preserved glimpses of the target speech suggest contributions of modulation masking and cognitive-linguistic processing as factors contributing to performance. © *2016 Acoustical Society of America.*
[http://dx.doi.org/10.1121/1.4962494]

[CGC] Pages: 1800–1816

## I. INTRODUCTION

Speech is often heard in the presence of background noise or competing talkers. Speech recognition for young normal-hearing (NH) adults is limited at poor signal-to-noise ratios (SNRs) (Christiansen *et al.*, 2013), but is improved when amplitude fluctuations in the competing noise are introduced to provide momentary improvements in the SNR (Festen and Plomp, 1990; Christiansen *et al.*, 2013). This improvement in speech recognition for fluctuating noise relative to recognition with steady-state noise is called masking release (MR). Under such conditions listeners are able to make use of brief time-frequency units of speech where the local, short-time SNR is deemed favorable, that is, exceeds a particular threshold (Cooke, 2006; Stone and Moore, 2014). In other words, NH listeners achieve higher speech recognition when spectro-temporal fragments of the target speech signal are available for perception (i.e., when energetic masking is limited).

In addition to the role of amplitude modulation in determining time varying changes in energetic masking, the speech temporal envelope also conveys significant cues for speech recognition (Shannon *et al.*, 1995), such as used for phoneme identification (Van Tasell *et al.*, 1987; Gallun and Souza, 2008) and sentence processing (e.g., Fogerty, 2011a, 2014). While amplitude modulation can facilitate stream segregation (e.g., Grimault *et al.*, 2002) and dip listening (e.g., Peters *et al.*, 1998), competing or interrupting signals can significantly interfere or alter the processing of these important temporal envelope cues (e.g., Gilbert *et al.*, 2007; Fogerty, 2011b). This is particularly true when the temporal envelopes of the target and competing speech signals are modulated at similar rates, for example, when dynamic range compression imposes co-modulation of these two signals (Stone and Moore, 2004). Such overlaps in the modulation rates of the two competing signals can reduce speech perception due to similar informational content of the temporal envelopes (i.e., informational masking; see Nelson and Jin, 2004) or impose interference in detecting the target speech modulations (i.e., modulation masking; Bacon and Grantham, 1989; Yost and Sheft, 1989).

Speech recognition in real-life environments entails listening in temporally complex modulated-noise backgrounds, such as from competing talkers. Under such conditions, the temporal characteristics of the competing modulation may significantly influence the intelligibility of the target speech. Indeed, the influence of temporal modulation on MR has emerged as an important variable in several studies (cf. Howard-Jones and Rosen, 1993; George *et al.*, 2006; Buss *et al.*, 2009). Furthermore, even the masking that occurs with "steady-state" (i.e., unmodulated) noise can be defined in terms of modulation properties of the random amplitude fluctuations inherent in the noise masker (Stone *et al.*, 2011a; Stone *et al.*, 2012; Stone and Moore, 2014).

[a] Electronic mail: fogerty@sc.edu
[b] Current address: Medical University of South Carolina, Charleston, SC 29425, USA.

The current study was designed to investigate the acoustic and perceptual consequences of the competing talker's temporal modulation spectrum on speech recognition. Time compression/expansion was used to shift the modulation spectrum of the modulated-masker to faster or slower rates to influence the degree of overlap with the modulation spectrum from the target talker. In addition, perceptual contributions of amplitude and frequency modulations in the target speech were investigated through the comparison of unprocessed speech to noise-vocoded speech. This latter condition directly investigated effects of modulation interference based on the competing temporal envelopes. Finally, an acoustic analysis was conducted to define temporal properties of the resulting speech glimpses (i.e., the temporal intervals of speech that occurred at favorable SNRs) that correlated significantly with speech recognition.

## A. Modulation masking

Modulated maskers have traditionally been discussed as providing a release from the energetic masking of "steady-state" maskers. However, modulated maskers may also lead to masking in the modulation domain. Such modulation masking results in poorer processing of target amplitude modulations that occur at the same modulation rate (e.g., Gustafsson and Arlinger, 1994) or second-order beating (Füllgrabe *et al.*, 2005) of the complex modulated masker.

Dau *et al.* (1997) proposed that modulation-rate-specific auditory filters are possibly involved in modulation masking or interference. Indeed, previous psychoacoustic studies of modulation masking and modulation detection interference suggest perceptual interference of modulation processing (e.g., Bacon and Grantham, 1989; Ewert and Dau, 2000; Millman *et al.*, 2002) when a masker is modulated at the same rate, even if the target and masker are separated in spectral frequency (Yost *et al.*, 1989).

Many studies have investigated speech recognition in the presence of square wave or sine wave modulation at a single modulation rate (e.g., Nelson *et al.*, 2003; Gnansia *et al.*, 2008). However, the speech signal is composed of different modulation rates across the modulation spectrum. As such, modulation masking by a competing talker will result in modulation masking of several different speech modulation frequencies. This affords the possibility of observing complex interactions for the modulation masking of speech that are not predicted by psychoacoustic studies (Kwon and Turner, 2001) or even by analysis of the first order modulation spectrum (Füllgrabe *et al.*, 2005). Indeed, multiple mechanisms appear to be involved in determining speech recognition performance in modulated backgrounds (for a discussion, see Füllgrabe *et al.*, 2006). However, determining differences in speech recognition as a result of the modulation properties of the speech and masker is still of significant importance as it may contribute in part to speech understanding in everyday noisy environments. The overlap between the speech modulation spectra of the competing talker with that of the target talker may, in part, determine speech recognition performance. Using a time-compressed speech-modulated masker, the current study offers a method to examine variations in the masker modulation spectrum and corresponding modulation masking of the target talker.

## B. Factors affecting modulation masking release

When listening to speech in the presence of a fluctuating masker, speech information that occurs within the dips of the masker is temporally surrounded by relatively intense portions of the modulated masker. Forward masking, in which an initial noise suppresses the response to subsequent sounds, may potentially diminish the ability of listeners to glimpse speech within the amplitude dips of the masker (e.g., Festen and Plomp, 1990; Dubno *et al.*, 2003). Consistent with this hypothesis, Dubno *et al.* (2003) found that MR was diminished for listeners with poorer forward-masker thresholds. This observed relationship was stronger at poorer SNRs and higher (i.e., faster) noise interruption rates. However, these correlations were conducted for a group of younger and older normal-hearing listeners, and therefore could have been influenced at least in part by additional age-related factors.

In addition to forward masking, MR is also influenced by the global, long-term average SNR. At favorable SNRs, MR is reduced because the noise is a less effective masker compared to poor SNRs. Therefore, less benefit is provided by introducing intermittent periods at an even more favorable SNR (i.e., glimpses) by modulating the masker (Bernstein and Grant, 2009; Christiansen and Dau, 2012). Most recently Christiansen and Dau (2012) found that the correlation between MR and speech reception thresholds in stationary noise (across conditions of low-pass, high-pass, or noise-vocoded processing) varied between processing conditions for competing speech maskers but not for sinusoidally amplitude-modulated (SAM) maskers. On the basis of this evidence, Christiansen and Dau (2012) concluded that acoustic features of the masker signal also influence the degree of MR. Toward this end, the current study also implements an acoustic analysis to define the acoustic consequences of the different noise maskers.

One acoustic property of particular relevance in the current study is the masker modulation spectrum. Speech recognition and masking release are known to vary as a function of the modulation rate (e.g., Miller and Licklider, 1950; Füllgrabe *et al.*, 2006) as well as the temporal distribution of speech cues (Buss *et al.*, 2009). The current study investigated the effect of the range of masker modulation rates in comparison to the modulation range of the target speech. In this context, greater modulation masking is expected when the modulation rates between the two concurrent signals are most similar (Dau *et al.*, 1997; Bacon and Grantham, 1989; Yost and Sheft, 1989; Moore *et al.*, 2009; Stone *et al.*, 2011a; Stone *et al.*, 2012; Stone and Moore, 2014). However, modulation interference for speech may be more governed by stream segregation and perceptual grouping processes (e.g., Hall and Grose, 1991; Yost, 1992; Kwon and Turner, 2001) rather than due to selective interference of modulation channels (as proposed by Dau *et al.*, 1997). Regardless of the exact mechanism, the current study investigates the potential interference resulting from competing temporal envelopes overlapping in the modulation domain to various degrees.

### C. Speech recognition with primarily temporal cues in the presence of modulated maskers

Modulation masking by a modulated masker occurs due to interactions of the temporal amplitude envelope of the target and competing signals. Thus, factors related to modulation masking can best be observed for speech that is limited to primarily temporal envelope cues. This type of listening occurs for vocoded speech and for listeners with cochlear implants (CI). Both of these types of listening conditions demonstrate reduced or no MR when listening to speech in the presence of modulated maskers (e.g., Füllgrabe *et al.*, 2006; Li and Loizou, 2009; Nelson *et al.*, 2003; Nelson and Jin, 2004; Jin and Nelson, 2010; Stickney *et al.*, 2004).

For example, Stickney *et al.* (2004) tested NH listeners on CI simulations (vocoded speech) with both steady-state noise and a single competing talker at SNR levels ranging from 0 to 20 dB. Results demonstrated not only the absence of MR, but speech recognition during a single competing talker was also poorer than during steady-state noise. This is notable given that NH listeners typically perform better in the same task with single-talker maskers than with steady-state noise (Christiansen *et al.*, 2013). These results are consistent with the hypothesis that fluctuating maskers may result in modulation interference (Nelson *et al.*, 2003), potentially due to modulation masking (e.g., Stone *et al.*, 2011a) or errors in stream segregation (e.g., Kwon and Turner, 2001).

Moreover, speech recognition in modulated backgrounds may require processing of fast frequency modulations of the acoustic temporal fine structure (TFS) within the dips of the masker (e.g., Lorenzi *et al.*, 2006; Hopkins and Moore, 2009). However, the relative contribution of TFS in dip listening is currently controversial. For example, several studies have found that MR is not associated with the periodic TFS, such as in whispered speech (Freyman *et al.*, 2012) or differentially affected by the presence of resolved low-order harmonics that primarily convey periodic TFS information compared to unresolved high-order harmonics (Oxenham and Simonson, 2009). Indeed, while speech perception for older adults is significantly correlated with TFS sensitivity, MR is not (Füllgrabe *et al.*, 2015). One possibility for this is that TFS information used for speech-in-noise processing may actually be more distributed across a wide range of values (Stone *et al.*, 2011b) or facilitate stream segregation (e.g., Apoux *et al.*, 2013; Fogerty and Xu, 2016). Several of these previous studies focused on TFS cues to pitch periodicity, but listeners appear to use TFS cues across the spectrum to facilitate the perception of masked speech (Hopkins and Moore, 2009). Regardless of how TFS contributes to speech perception, studies demonstrate that listeners who receive vocoder and CI processing that limit acoustic TFS cues are particularly susceptible to modulation masking (e.g., Qin and Oxenham, 2003; Nelson *et al.*, 2003; Stickney *et al.*, 2004; Gnansia *et al.*, 2008; Li and Loizou, 2009; Jin and Nelson, 2010). Of interest here is whether varying the temporal properties of the modulated noise can reduce modulation masking and therefore increase MR for speech signals that either preserve or remove TFS cues.

### D. The current study

The current experiment was designed to investigate MR in the presence of speech-modulated noise. It was hypothesized that an increase in MR would result from altering the modulation spectrum of speech-modulated noise through time compression due to a mismatch between the target modulation spectrum and that of the masker.[1] Thus, we investigated masking by the entire modulation spectrum of a single competing talker. In addition, the role of temporal envelope cues to MR was investigated using unprocessed and vocoded target speech.

Varying the competing modulation spectrum also alters acoustic properties of speech glimpses. A number of acoustic metrics have been developed to define available glimpses (Cooke, 2006; Li and Loizou, 2007). These include the number of available glimpses, duration of individual glimpses, and total proportion of target speech that is glimpsed (where glimpse typically refers to a connected spectro-temporal region of the target speech that exceeds the level of the masker by some threshold, e.g., 0 or 3 dB SNR). It is notable that both Cooke (2006) as well as Li and Loizou (2007) independently found that the total proportion of target speech glimpsed is an important predictor of intelligibility. In this study, we investigated the contribution of different acoustic glimpse properties to speech recognition.

The objectives of this study were:

(1) To analyze how MR is affected by shifting the competing modulation spectrum through time compression/expansion.
(2) To investigate modulation masking by a single-talker modulated noise based on primarily temporal envelope cues through comparison of unprocessed and vocoded speech.
(3) To determine how different acoustic measures of preserved speech glimpses (i.e., glimpse rate, glimpse duration, and sentence proportion) correlate with speech recognition.

## II. PERCEPTUAL MEASURES OF SPEECH INTELLIGIBILITY

### A. Participants

Fifteen NH listeners (4 males and 11 females) between the ages of 19 and 36 years (mean: 22.4 years) participated in this experiment. All listeners had audiometric thresholds of 20 dB hearing level or better at octave frequencies between 250 and 8000 Hz in both ears.

### B. Stimuli and design

Recordings of the Institute of Electrical and Electronics Engineers (IEEE) sentences (IEEE, 1969) by a male talker were used in this study (Loizou, 2007) to create both the target and masker stimuli. Each sentence contained five keywords. For the target talker, both natural and vocoded conditions were created. Each of these conditions was paired with six masker conditions. One masker condition was tested with steady-state speech-shaped noise while the other five conditions tested speech recognition with single-talker modulated noise that was either time compressed or expanded to present the modulation at 25%, 50%, 100%, 200%, or 400% of the original duration. This resulted in a total of 12 conditions (2 target types × 6 masker types). In addition, vocoded speech was also presented in quiet as a baseline condition to determine maximum performance.

## C. Stimulus processing

### *1. Speech processing*

Two speech conditions were used to investigate the effect of spectral resolution on speech recognition in modulated noise: natural and vocoded. Speech vocoding was used to limit spectral resolution, but preserve low frequency amplitude modulations of the speech temporal envelope and faster modulations of speech periodicity that could be used to facilitate perceptual segregation from the noise. Vocoding was implemented in MATLAB using the Hilbert transform. This method retained fast rate modulations associated with periodicity of the talker up to the maximum modulation rate provided by the filter bandwidth. Vocoder analysis used eight channels with equal distance on the basilar membrane (cut-off frequencies of 80, 192, 364, 629, 1037, 1664, 2629, 4115, and 6400 Hz). A bank of finite impulse response filters was designed using a filter order of 572 based on the smallest filter bandwidth. The significant overlap between adjacent filters resulted in 28 Hz wide filter transitions. Extracted Hilbert envelopes were then used to modulate noise that matched the speech spectrum of the target sentence. Modulated noise bands were then summed to produce eight-channel vocoded speech. The natural (i.e., unprocessed) and vocoded sentences were low-pass filtered to 6400 Hz to equate speech bandwidth for these two conditions.

### *2. Masker processing*

The speech-modulated masker was specifically designed to limit modulation interference to the slow modulation rates important for speech (<16 Hz; Drullman *et al.*, 1994; Shannon *et al.*, 1995; Füllgrabe *et al.*, 2009). Thus, faster amplitude modulation rates of the speech associated with prosodic pitch and fundamental frequency (Stone *et al.*, 2008) information were specifically avoided in the design of the modulated masker. However, these cues were present to at least some degree in the natural and vocoded speech stimuli. Thus, an interference of the modulated masker can most likely be assigned to the slow amplitude modulation rates preserved in the speech-modulated noise masker.

To create the masker, a steady-state speech shaped noise (SSN) was generated that matched the long-term average speech spectrum for a concatenation of 40 IEEE sentences that were extracted from sentences not used in the experiment. Silent intervals between sentences were removed. The temporal envelope was then extracted from the concatenated speech sample by half-wave rectification and low-pass filtered using a sixth-order Butterworth filter with a cutoff frequency of 16 Hz. This speech envelope was then used to amplitude modulate the steady-state noise to create the speech modulated noise that matched the modulation spectrum of the target sentences. Thus, the noise masker used here preserved the temporal (see waveforms in Fig. 1) and long-term spectral envelopes (Fig. 2) of the competing speech that it modeled. The modulation spectrum for the masker was then modified using pitch-synchronous overlap-add (PSOLA) time compression/expansion using Praat (Boersma and Weenink, 2014) to run at 25%, 50%, 200%, and 400% of the original duration. This means that the 50% condition was twice as fast as the original (i.e., doubles the modulation rate), while the 200% condition was twice as slow (i.e., halves the modulation rate). Random segments of modulated masker noise at a given time-compression rate were paired with natural and vocoded target sentences to create experimental stimuli and saved as separate channels in a stereo file to preserve information regarding the temporal alignment of the speech and masker. Examples of natural and vocoded waveforms and spectrograms are displayed in Fig. 1 along with displays for the modulated masker noise.

### *3. Speech and noise modulation spectra*

The envelope modulation spectrum was calculated to define the acoustic effect of each time compression condition. First, 40 IEEE sentences were concatenated and time-compressed using the procedures described previously. Next, modulation spectra were calculated for these concatenated files over a 20 s window. The envelope modulation spectrum for speech at each time compression condition was obtained, following the procedures used to create the masking noise. That is, by half-wave rectification, low-pass filtering using a sixth-order Butterworth filter at 16 Hz, and downsampling to 1000 Hz. Next, the fast Fourier transform (FFT) was computed. Partitions of the FFT bins were made to correspond to octave bands with center frequencies ranging from 1 to 32 Hz. The energy in an octave-band FFT partition was divided by the energy in the 0 Hz bin to give the modulation index relative to the DC offset. These values were normalized to sum to one in Fig. 3(A) to display the normalized modulation spectra for the different time compression rates of the masker. As can be observed, the natural speech rate (100%) corresponded to a peak envelope modulation rate of 4-Hz, while time compression or expansion shifted this spectrum to faster or slower rates, respectively.

While the modulation spectrum for the noise was consistent across frequency bands [as can be observed by the spectrogram in Fig. 1(C)], the spectral complexity of the natural and vocoded speech resulted in different temporal envelopes across the frequency spectrum. Thus, the same modulation spectrum analysis was calculated for the target speech for the eight analysis bands used in the vocoder processing. These spectra are displayed in Fig. 3(B). The gray shaded region indicates the wideband modulation spectrum. As can be observed, relative differences across the modulation rates in the low frequency bands are highly consistent

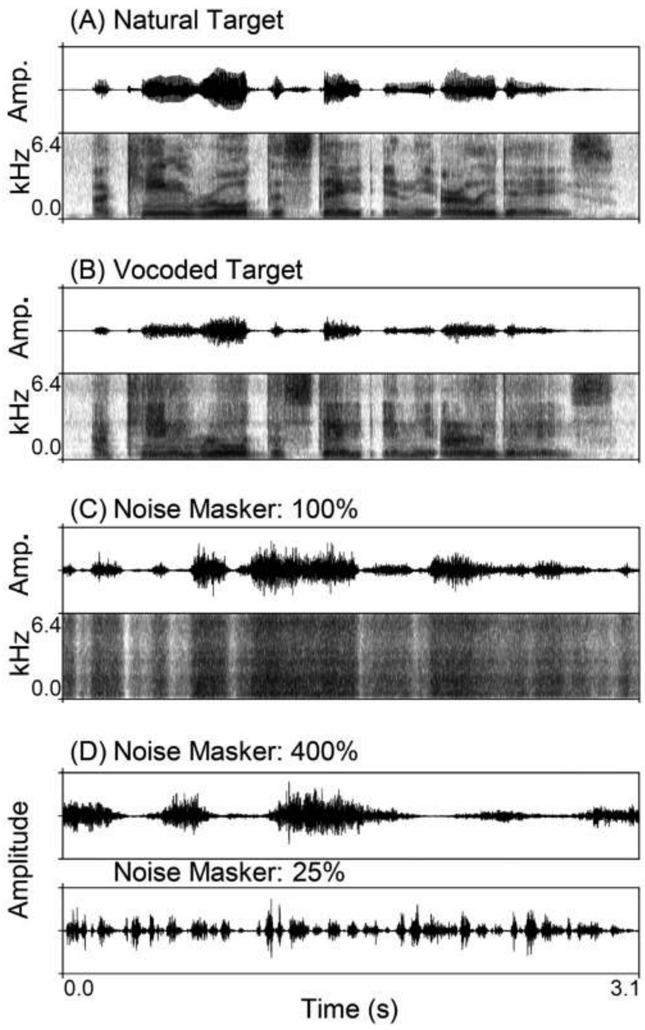

FIG. 1. Waveform (top) and spectrogram (bottom) for (A) natural and (B) vocoded speech and (C) the speech-modulated masker at the original speech rate (100%). (D) The waveforms for speech-modulated noise are displayed following 400% time expansion, which reduces amplitude modulation rates, and 25% time compression, which increases amplitude modulation rates. The example sentence (A-B) is "A king ruled the state in the early days." The modulated noise maskers (C-D) were based on random selections from different sentences.

with the wideband signal. However, some deviation is noted for the high frequency bands, with a shift to more energy at faster modulation rates. This is consistent with earlier analyses of the speech modulation spectrum across frequency (Greenberg *et al.*, 1998).

Given this difference for the high frequency bands, it is possible that modulation interference across speech frequencies may be different for the various noise modulation rates. To determine potential differences, we calculated the correlation between the modulation spectrum of the speech band and the modulation spectrum for the five different speech modulated noises. Figure 3(C) displays these correlation coefficients across frequency bands. It can also be observed that across frequency bands, the slower maskers (200% and 400%) result in the highest correlations with the mid frequencies (around 1000–1700 Hz) and small to negative correlations in the high frequency bands (above 2600 Hz). In contrast, the faster maskers (25% and 50%) result in the lowest correlations with the midfrequency region and highest correlations with the high frequency bands. The masker at 100% of the original speech rate in general resulted in the highest correlations and was fairly consistent across frequency bands. The exception to this observation is with the highest frequency bands that also contain the fastest speech modulations. For these frequency bands, correlations with the modulation spectrum for the 50% condition were near or higher than those obtained for the 100% condition. These correlations indicate the expected interference patterns across noise maskers. The most modulation interference was expected for the 100% masker condition, with the least modulation masking occurring for the extreme time compression/expansion conditions (25% and 400%) which resulted in the smallest modulation spectrum overlap with the target speech across frequency bands.

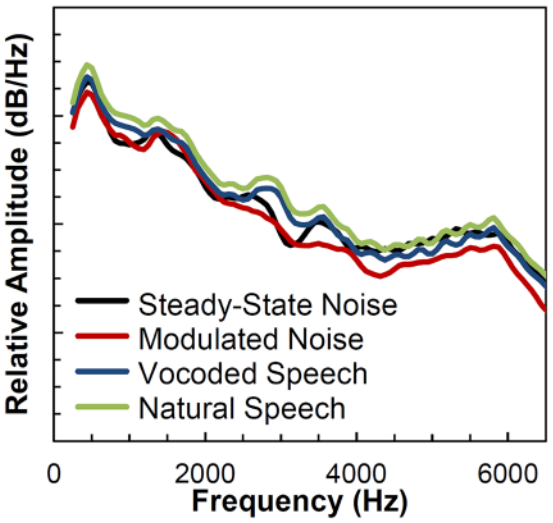

FIG. 2. (Color online) Long-term average spectrum for the natural and vocoded speech and the modulated and steady-state noise. Relative amplitude levels are slightly offset in overall level for clarity in comparing the spectral shape.

### D. Procedure

The 12 conditions (natural and vocoded testing in modulated noise) and vocoded testing in quiet comprised three blocks. The natural target talker block was tested first at an SNR of −7 dB. This block was followed by vocoded testing in quiet, which provided a baseline measure of vocoded intelligibility and familiarization to the vocoder processing. Finally, the vocoded target talker block was tested last at an SNR of 2 dB. These SNR levels were chosen based on preliminary piloting and were selected because they resulted in similar estimated levels of performance in the SSN condition. Before experimental testing began, a short demo of 20 sentences was presented to the listeners to familiarize them with the task and listening condition but the results were not scored. Within each block, the sentences were presented in a different random order for each participant with the different masker time compressions intermixed. Twenty sentences were presented for each condition for a total of 260 experimental sentences.

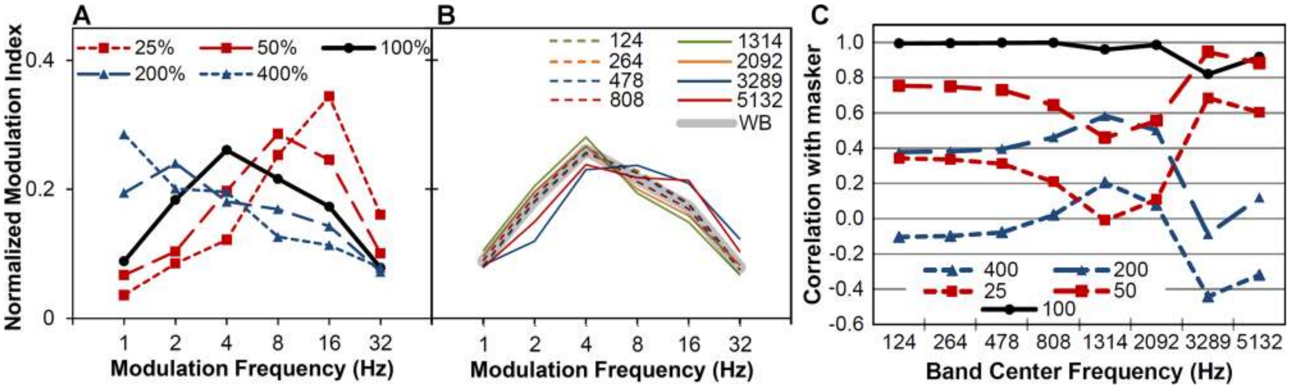

FIG. 3. (Color online) Speech modulation spectra for (A) the different time compression conditions that were used to modulate the masking noise and (B) the eight different filter bands. The modulation spectrum for the original 100% rate (WB) is displayed as the solid black line in A and the thick gray line in B. (C) The correlation coefficient for the five different time compressed/expanded modulation spectra across the eight different filter bands.

The experimental conditions were run using a custom MATLAB stimulus presentation interface in a sound-attenuating booth. Participants wore headphones (Sennheiser HD 280 Pro) that presented the speech through only the right headphone. The presentation level of the target speech was calibrated to 70 dB sound pressure level prior to addition of the noise. Random segments of noise were selected from the masker files and started 100 ms before and ended 100 ms after the target sentence. The unique noise masker used for each trial was saved in a stereo file with the target sentence for later acoustic analyses. Participant responses were recorded for offline scoring and analysis. The recordings were scored manually and the percentage of correct keywords spoken (out of five) was recorded for each sentence. A second rater re-scored data from two participants. The scores between the two raters were highly consistent, Cronbach's $\alpha = 0.98$. MR was calculated by subtracting the steady state noise percent correct score from the modulated noise score.

## E. Results and discussion

Data analysis was first conducted to explain the main effects for overall accuracy and MR. Subsequent contrasts were then conducted to evaluate the significant interactions through comparisons between processing and time compression conditions. Data analysis was conducted using a significance level of 0.05.

### 1. Overall accuracy

Average accuracy for the different experimental conditions is displayed in Fig. 4. Baseline performance for the natural and vocoded conditions was first determined without masker modulation, i.e., SSN. A paired samples $t$-test was used to compare performance between the natural-SSN and vocoded-SSN conditions. Results indicated significantly better performance for the vocoded-SSN condition, $t(14) = -4.7$, $p < 0.001$, $d = 1.2$, by a difference of 6.6 percentage-points. This small but significant difference occurred even after our initial piloting attempts to estimate noise levels that would approximate equal performance levels. Paired $t$-tests were also conducted between performance in SSN and across the five different masker rates. All comparisons were significant ($p < 0.001$) for both natural and vocoded speech conditions, with large effect sizes ($d > 1.2$). This latter result indicates that listeners did obtain a perceptual benefit from modulations in the masker for all natural and vocoded speech conditions. However, there were significant differences in the benefit obtained for the different masker modulations.

To examine the effect of the masker modulation, a 2 (target: natural and vocoded) by 5 (time compression: 25%, 50%, 100%, 200%, and 400%) repeated-measures analysis of variance (ANOVA) test was conducted on the percent correct data. There was a significant main effect of the signal processing condition, $F(1,14) = 185.1$, $p < 0.001$, as well as a main effect of time compression, $F(4,56) = 4.3$, $p = 0.004$. Finally, results also demonstrated a significant interaction between signal processing and time compression factors, $F(4,56) = 76.4$, $p < 0.001$.

Overall, these results indicated better performance for natural speech in speech-modulated noise. This occurred

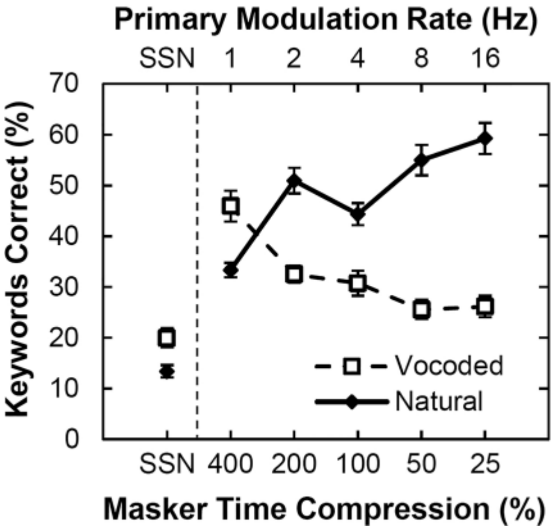

FIG. 4. Mean percent correct is displayed for natural and vocoded talker conditions for the different masker time compression rates and corresponding primary modulation rates, as well as for steady-state noise. The value in parentheses is the global SNR used for testing. Error bars = standard error of the mean.

even though performance in the unmodulated SSN masker was better for the vocoded target. The rate of the modulation also impacted performance.

### 2. Masking release data

Masking release, i.e., the performance difference between the modulated and unmodulated noise backgrounds, was calculated in order to control for the inherent performance differences in energetic masking observed between natural and vocoded conditions. As can be observed from Fig. 5, MR for the natural speech conditions was in the range of 31%–46%, with somewhat poorer performance for the 400% condition (20%). MR was much lower for vocoded speech, generally in the range of 6%–13%. Better performance was observed for vocoded speech in the 400% condition (26%). As reported above in the paired comparisons with SSN, even though some of these values were small, all conditions resulted in significant MR with large effect sizes ($d > 1.2$).

Using this derived data, another 2 (target: natural and vocoded) by 5 (time compression: 25%, 50%, 100%, 200%, and 400%) repeated-measures ANOVA was completed. Consistent with the earlier analysis on the raw percent correct data, results from this analysis demonstrated a significant main effect of signal processing, $F(1,14) = 305.5$, $p < 0.001$, and time compression, $F(4,56) = 4.3$, $p = 0.004$. A significant interaction between signal processing and time compressions was obtained as well, $F(4,56) = 76.4$, $p < 0.001$. These significant differences between conditions were explored in more detail through a set of planned comparisons.

*a. Comparing between natural and vocoded.* A set of two-tailed paired-samples *t*-tests was completed using the MR data to compare the natural and vocoded conditions within each time compression condition. A Bonferroni correction was used to control for multiple comparisons. For all comparisons, participants had significantly more MR for the natural speech than vocoded speech ($p < 0.001$), except for the 400% time compression condition in which listeners had greater MR for the vocoded condition [M = 25.9%, standard deviation (SD) = 6.5%] compared to the natural condition (M = 19.9%; SD = 4.4%), $t(14) = -2.9$, $p = 0.01$, $d = 0.7$. Consistent with the previous literature, these results indicate that listeners did better for natural speech in most modulated backgrounds. However, greater MR was observed for vocoded speech for the background with the slowest modulation rates (i.e., 400%).

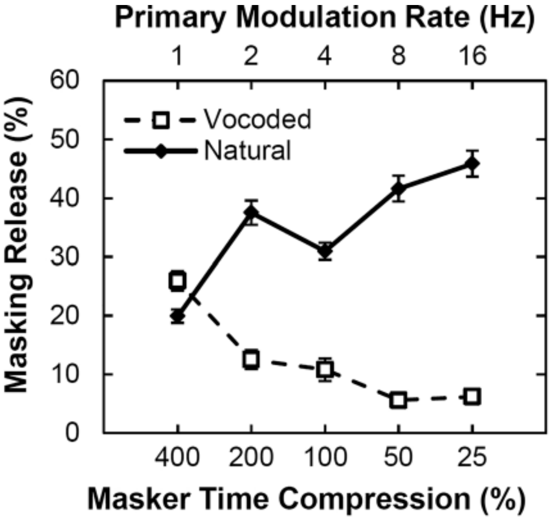

FIG. 5. Masking release (i.e., the difference from SSN) is displayed for natural and vocoded talker conditions for the different masker time compression rates and corresponding primary masker modulation rates. The value in parentheses is the global SNR used for testing. Error bars = standard error of the mean.

*b. Comparing among time compressions.* Another set of two-tailed paired-sample *t*-tests was completed using MR data to compare among the time compressions within the natural and vocoded conditions. Bonferroni correction was again used to control for multiple comparisons. Planned comparisons were conducted between the original 100% rate and the different time compression conditions. For the natural condition, MR in the 100% condition was significantly poorer than for the 25%, 50%, and 200% time compressions (Bonferroni-adjusted $p < 0.05$); while for the vocoded condition, the 100% condition was not significantly different from the 25% and 200% conditions ($p > 0.05$) but was significantly better than the 50% condition (Bonferroni-adjusted $p < 0.05$, $d = 0.82$). However, the meaningfulness of this latter comparison is considered tentative considering the similar performance across the other rate comparisons.

Performance for natural and vocoded speech was most different for the 400% condition. For the natural condition, the 100% condition resulted in significantly more MR than the 400% condition [$t(14) = 7.9$, $p < 0.001$], while for the vocoded condition, 100% condition resulted in significantly less MR than 400% condition [$t(14) = -6.7$, $p < 0.001$].

For the natural condition, an improvement in performance was noted for faster and slower modulations, compared to the natural 100% speech rate, with the exception of the slowest 400% condition. This performance profile across rates is different from the vocoded condition, which shows no appreciable difference in performance except for the improvement noted at the slowest 400% condition. Overall, results demonstrate that speech recognition in noise is affected, at least in part, by the interaction between the modulation spectra of the target speech and competing masker.

*c. Keyword analysis for the 400% condition.* At faster modulation rates, glimpses would have occurred more frequently across the sentence, thus providing some acoustic sampling of each keyword. However, at the slowest modulation rate (400%), it is likely that only some of the keywords were glimpsed during the sentence. In the absence of bottom-up perceptual cues, listeners would have had to cognitively fill-in the missing words that were not glimpsed. To examine the use of contextual cues for this “filling-in” process, we analyzed the accuracy for keywords that were glimpsed versus those that were not glimpsed. For this application we defined a glimpsed keyword as any keyword that had at least one 16-ms window that was at or above a local SNR of 0 dB. Note that this was a very lose criterion that counted keywords as having been "glimpsed" even if only minimal energy and duration of the speech was above the level of the noise. This definition resulted in 71% and 83% of the keywords defined as having been glimpsed for natural and vocoded conditions, respectively. Of course, for those keywords that were "glimpsed," only a portion was actually presented above 0 dB SNR. On average, only 39% of the natural keyword duration and 46% of the vocoded keyword duration occurred above this local SNR criterion.

Figure 6 displays a sub-analysis of glimpsed and not-glimpsed keyword accuracy for the 400% condition. The left panel displays recognition accuracy for the five keywords in the sentence, partitioned by the glimpse status of the keyword, as defined above. Note that the sum of correct responses across glimpsed and not glimpsed keywords is equal to the total accuracy reported in Fig. 4. As expected, correct keywords were most frequently those keywords that contained some glimpsed acoustic information. Of interest in this analysis is the ability of listeners to fill in the missing (i.e., not glimpsed) keywords. The right panel of Fig. 6 displays the proportion of responses that were correct for the subset of words that were glimpsed, and separately, for the subset of keywords that were not glimpsed. Interestingly, very similar glimpsed keyword recognition accuracy is seen for natural and vocoded conditions [41% versus 44%, respectively, $t(14) = 1.3$, $p = 0.22$]. Note that these proportions are relative to the total keywords glimpsed, and so account for the greater number of glimpsed keywords for vocoded (83%) compared to natural (71%) speech conditions. In contrast, listeners were much less successful at filling in missing keywords in the natural condition compared to the vocoded condition; the latter of which had a fewer number of not-glimpsed keywords per sentence. On average, listeners were only able to accurately report 10% of keywords that were not glimpsed in the natural condition, compared to 59% in the vocoded condition [$t(14) = 12.8$, $p < 0.001$]. This greater success at filling in missing keywords is combined with a significantly higher recognition [$t(14) = 3.6$, $p < 0.01$, $d = 0.92$] in the vocoded condition when only considering the glimpsed keywords (see the black bars in the left panel of Fig. 6). Thus, performance in the vocoded 400% condition was better than the natural condition due to better recognition of glimpsed keywords and a greater ability to use context to fill in the fewer missing keywords.

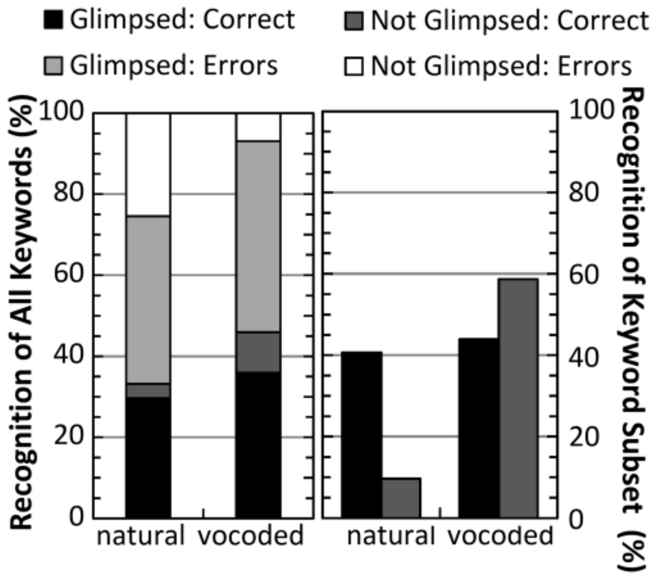

FIG. 6. Recognition of keywords that were glimpsed (i.e., had at least 16-ms above 0 dB SNR) or not glimpsed in the natural and vocoded 400% time-expansion condition. The left panel displays the proportion of responses that were correct or incorrect for keywords that were glimpsed and for keywords that were not glimpsed. Responses are displayed as a proportion of all keywords presented. The right panel displays recognition accuracy for the glimpsed and not glimpsed keyword subsets. Accuracy is displayed as a proportion of the number of keywords presented in each subset.

In the vocoded condition listeners had a greater number of keywords with some acoustic information preserved, a greater duration of those keywords preserved, and better recognition for those keywords. All of these factors likely contributed to the successful use of context in the vocoded condition. Furthermore, the results indicate that listeners were successfully able to perceptually resolve speech information during the glimpsed keyword intervals. In contrast, performance in the natural condition appears limited as glimpsed speech intervals were less distributed across the sentence, resulting in fewer glimpsed keywords, and therefore limiting listeners' abilities to contextually fill-in missing (i.e., not-glimpsed) keywords. This was combined with poorer recognition accuracy for glimpsed keywords compared to the vocoded condition, potentially due to shorter glimpsed durations of the keywords. These observations suggest that, compared to other time-compression conditions that provided more regular glimpsing opportunities across the sentence, performance in the natural 400% condition was likely limited by sparseness in perceptual glimpsing opportunities and difficulty using contextual cues to cognitively fill in the missing speech information.

## III. ACOUSTIC ANALYSIS OF SPEECH GLIMPSES

As demonstrated earlier in Fig. 3, time compression of the speech-modulated masker resulted in substantial differences in the masker modulation spectrum across conditions (i.e., changes in peak modulation from 4 Hz in the target to the extremes of 1 Hz or 16 Hz for 400% and 25%, respectively). As proposed by Stone *et al.* (2011a), Stone *et al.* (2012), and Stone and Moore (2014), performance in these conditions could be determined, in part, by modulation masking due to the overlap in the modulation domain between the target speech and competing noise waveforms. This overlap was systematically varied across conditions. To demonstrate this effect, modulation spectra were calculated (following the same procedures as used for Fig. 3) for the target sentence and corresponding noise for each sentence trial. The absolute difference in the modulation index was calculated for each octave modulation band and averaged to provide a summary statistic of the modulation overlap between the speech and noise. Figure 7 summarizes these data for the natural and vocoded conditions. As predicted, the greatest overlap between the speech and noise

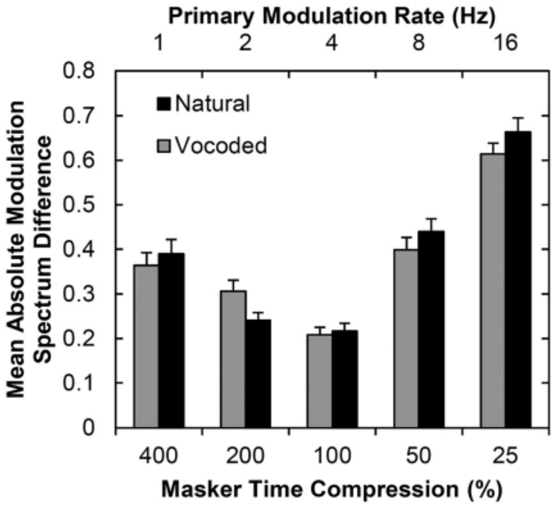

FIG. 7. Absolute difference between the speech and noise modulation spectra, averaged across octave modulation bands, for each experimental trial. Greater difference values indicate greater dissimilarity in modulation rates present for the speech and noise waveforms.

modulation spectra occurred for the original 100% rate, with greater differences occurring as the modulated noise was either time compressed or time expanded. Interestingly, the greatest masking release occurred for time compressions where greater separation was achieved between the competing modulation spectra. However, such a summary statistic based on the modulation spectrum cannot fully account for the results observed in this study. For example, observations of reduced MR for natural conditions at 400% and for vocoded conditions at 25% and 50% suggest a more complicated picture than that explained solely by differences in the modulation spectrum. This may in part be due to complex interactions between modulation spectra that could have significant implications for how preserved portions of the speech signal are available across time. Toward this end, an acoustic analysis was conducted to define the acoustic preservation of the speech signal in each of the experimental conditions.

Speech recognition in noise is to some extent dependent on the portions of the speech signal that occur at favorable SNRs, i.e., the speech glimpses (Cooke, 2006). Many studies have now demonstrated that speech recognition is influenced by how speech information is distributed across these glimpses, for example, in terms of the rate of glimpsing, duration of individual glimpses, or proportion of the speech preserved (e.g., Miller and Licklider, 1950; Li and Loizou, 2007; Wang and Humes, 2010; Shafiro *et al.*, 2011). The present acoustic analysis was conducted to determine how the different competing modulation rates influenced these different glimpse properties and in turn, how those properties explained performance across the different experimental trials.

## A. Signal analysis

Stimuli for the previous experiment were saved with the speech and noise in separate channels of a stereo file. This analysis calculated, for each sentence, the running root mean square (RMS) level for the speech channel and noise channel independently using a 16 ms non-overlapping window. From these measurements, the short-time SNR was obtained across the sentence, accounting for the appropriate presentation level of the speech signal relative to the noise (i.e., −7 dB or 2 dB for the natural and vocoded conditions, respectively). Figure 8 displays these measurements for an example sentence with natural and vocoded processing at different time compression values for the speech-modulated noise. Visual inspection of these displays demonstrates very different glimpse profiles across the different time compression conditions.

Four glimpse metrics were calculated using a threshold of 0 dB SNR to define the relatively preserved portions of the speech signal. Thus, glimpses were defined as temporal intervals of speech that occurred at positive SNRs for durations of at least 16 ms. Our initial analysis uses a 0 dB SNR threshold as it is a commonly accepted and theoretically motivated criterion (Li and Wang, 2009). Subsequent analyses reported in a final analysis also investigated results across a range of threshold values for the following:

(1) Sentence proportion: The proportion of the total sentence duration that occurred at positive SNRs.
(2) Glimpse duration: The average duration of time that the short-time SNR was at or exceeded 0 dB SNR.
(3) Glimpse proportion: The proportion of the glimpsed sentence provided by each glimpse. This was calculated by dividing the number of glimpses by the sentence proportion.
(4) Glimpse rate: The average number of glimpses that occurred per second. This was calculated by dividing the number of glimpses by the total sentence duration.

These four metrics were calculated for each sentence trial based on the masking noise that was presented on that trial. This was done in order to assess the association between average speech intelligibility (i.e., keyword scores averaged across participants) and the acoustic measures for each sentence. Glimpse metrics were calculated over the predominant speech portion of each file by removing the initial and final 250 ms of the file from the glimpse analysis. This procedure avoided having the initial and final silence padding influence the reported results.

## B. Results and discussion

### 1. Main effect of each glimpse metric

A set of statistical tests were first run to determine if the four proposed glimpse metrics captured significant acoustic differences between the experimental conditions. This consisted of a separate by-item ANOVA for each of the glimpse metrics using a 2 (signal processing: natural vs vocoded) × 5 (time compression: 25%, 50%, 100%, 200%, 400%) design across a total of 200 sentences (20 per condition). Significant main effects of signal processing and time compression were found and are summarized in Table I. In general, the vocoded condition demonstrated more frequent glimpses that were longer in duration and accounted for a greater proportion of the total sentence compared to the natural condition. This was

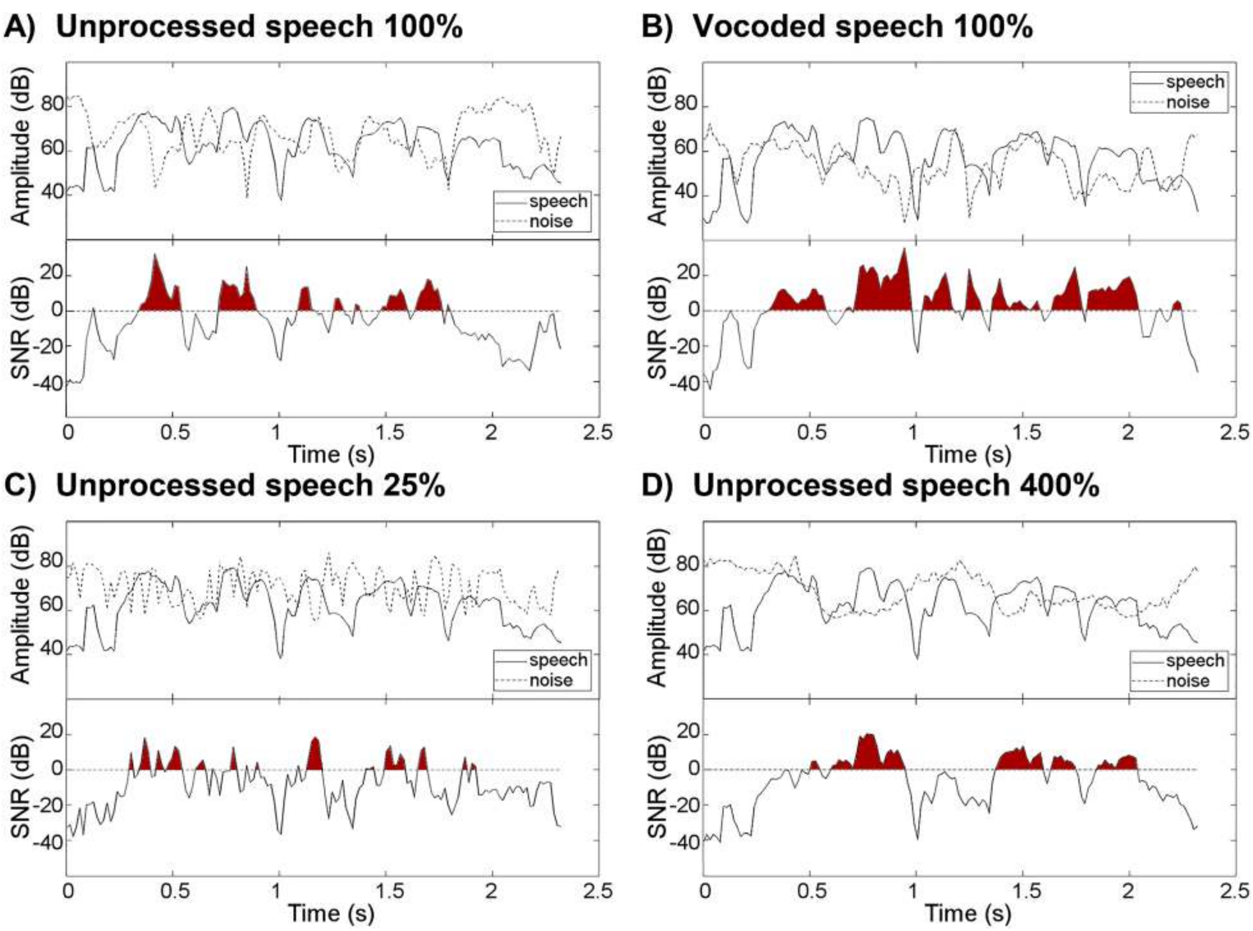

FIG. 8. (Color online) The running RMS level was calculated within 16-ms windows for the speech and noise (top display in each panel). The resulting short-term running SNR (bottom display in each panel) is displayed at the original speaking rate for (A) unprocessed and (B) vocoded speech. Highlighted portions above 0 dB SNR were defined as speech glimpses. For comparison, displays are also provided for unprocessed speech during (C) time-compression and (D) time-expansion.

due to the better long-term average SNR that was used during vocoded testing. As expected, the main effect of time compression demonstrates longer glimpses at the slower rates and more frequent glimpses at the faster rates.

TABLE I. Mean values and ANOVA results for the four different glimpse metrics.

| | Time Compression (%) | Sentence proportion | Glimpse duration (ms) | Glimpse proportion | Glimpse Rate (Hz) |
|---|---|---|---|---|---|
| Natural | 25 | 0.28 | 46.72 | 0.02 | 6.10 |
| | 50 | 0.29 | 63.59 | 0.03 | 4.69 |
| | 100 | 0.31 | 85.08 | 0.03 | 3.68 |
| | 200 | 0.30 | 105.45 | 0.04 | 2.99 |
| | 400 | 0.27 | 108.95 | 0.06 | 2.71 |
| | Mean | 0.29 | 81.96 | 0.04 | 4.03 |
| Vocoded | 25 | 0.51 | 83.69 | 0.03 | 6.16 |
| | 50 | 0.55 | 111.39 | 0.05 | 5.1 |
| | 100 | 0.53 | 133.53 | 0.06 | 4.23 |
| | 200 | 0.54 | 152.96 | 0.07 | 3.79 |
| | 400 | 0.52 | 156.43 | 0.07 | 3.62 |
| | Mean | 0.53 | 127.6 | 0.05 | 4.58 |
| Processing | $F(1,190)$ | 588.6[a] | 72.9[a] | 37.1[a] | 16.8[a] |
| Time Compression | $F(4,190)$ | 1.4 | 22.8[a] | 15.3[a] | 67.0[a] |
| Interaction | $F(4,190)$ | 0.2 | 0.2 | 0.8 | 1.3 |

[a] $p < 0.001$.

### 2. Independence and associations between acoustic glimpse measures

The four glimpse metrics reflect constructs that partially overlapped in their definitions. Therefore, correlations were conducted between each of the measures to determine the degree to which each glimpse metric reflected independent acoustic properties of glimpsing. Correlations were conducted across 100 sentences (20 sentences × 5 rates) separately for natural and vocoded processing. The resulting correlation coefficients are displayed in Table II. Results indicate significant correlations among the acoustic measures for both natural and vocoded processing. High correlations were obtained between glimpse proportion and glimpse duration ($r > 0.9$). Moderate correlations were obtained among the other comparisons with one exception: sentence proportion and glimpse rate appear to be independent measures. Similar patterns were observed for analyses based on the natural and vocoded conditions.

### 3. Association between acoustic glimpse measures and intelligibility

Of primary interest in this study is whether a particular glimpse metric was associated with intelligibility. This was investigated by determining the trial-by-trial correlation between each glimpse metric and sentence intelligibility in

TABLE II. Correlation coefficients for comparisons between the acoustic glimpse metrics calculated for each stimulus trial. Italics above the diagonal reflect correlations during vocoded processing.

| | Sentence proportion | Glimpse duration | Glimpse proportion | Glimpse Rate |
|---|---|---|---|---|
| Sentence proportion | | *0.48*[a] | *0.50*[a] | *−0.13* |
| Glimpse duration | 0.39[a] | | *0.93*[a] | *−0.87*[a] |
| Glimpse proportion | 0.37[a] | 0.93[a] | | *−0.74*[a] |
| Glimpsing Rate | 0.14 | −0.67[a] | −0.56[a] | |

[a] $p < 0.01$.

rationalized arcsine units[2] (RAU; Studebaker, 1985) on that trial, averaged across all participants. Correlation coefficients for each metric with mean intelligibility across time compression conditions is displayed in Table III.

Several observations can be made from these correlations. For the natural condition, accuracy was most strongly associated with sentence proportion and glimpse rate, two factors that were uncorrelated. Further analysis of these correlations demonstrate that these two glimpse measures were most associated with accuracy for the time expanded masker rates that were slower than the speech rate (i.e., 200% and 400%) (sentence proportion: $r = 0.61$, $p < 0.001$; glimpse rate: $r = 0.51$, $p < 0.001$). At these slow masker modulation rates, long durations of the speech signal are significantly masked by the noise. In these cases, the listener may have limited information to perceptually and cognitively fill in the missing speech information. In these conditions, greater preservation of the sentence, as indexed by sentence proportion, and greater sampling of the sentence, indexed by glimpse rate, may facilitate these "filling-in" processes.

On the other hand, a different effect was observed for time-compressed masker rates that are faster than the target speech signal (i.e., 25% and 50%). In these cases glimpse proportion was more associated with performance ($r = 0.37$, $p = 0.02$). This may reflect a potentially different process determined by the amount of information transfer per glimpse relative to the total available information.

Different effects were observed for the vocoded speech conditions, which were determined most by glimpse duration, proportion, and rate measures. The association of performance with the average glimpse duration and glimpse proportion is unique to vocoded processing. These two metrics were highly correlated and reflect the amount of time that is available during each glimpse. Correlations were largely influenced by the slower modulation rates. In contrast to the natural condition, there was also a negative correlation observed between intelligibility and glimpse rate. This is consistent with the earlier reported correlations with glimpse duration and proportion, as these metrics are also negatively correlated with glimpse rate. This suggests that while performance for natural sentences is best with frequent glimpsing of the sentence, the intelligibility of vocoded sentences is best with a less frequent sampling of the sentence. This benefit of reduced glimpses is likely due to the associated increase in glimpse duration, which may afford the required time to process faster speech amplitude modulations within the glimpse.

TABLE III. Correlation coefficients for comparisons between the acoustic glimpse metrics calculated for each stimulus trial and mean intelligibility for that stimulus.

| | Sentence proportion | Glimpse duration | Glimpse proportion | Glimpse Rate |
|---|---|---|---|---|
| Natural | | | | |
| All rates | 0.39[a] | −0.10 | −0.09 | 0.35[a] |
| Slow rates | 0.61[a] | 0.10 | 0.02 | 0.51[a] |
| Fast rates | 0.14 | 0.17 | 0.37[b] | −0.13 |
| Vocoded | | | | |
| All rates | 0.10 | 0.23[b] | 0.23[b] | −0.20[b] |
| Slow rates | 0.09 | 0.20 | 0.23 | −0.14 |
| Fast rates | 0.11 | −0.06 | −0.06 | 0.12 |

[a] $p < 0.01$.
[b] $p < 0.05$.

It is interesting that while glimpse opportunities were more favorable for vocoded conditions according to all glimpse metrics due to testing at a better SNR, listeners performed more poorly overall both in terms of overall accuracy and MR.[3] This is consistent with earlier studies that demonstrate poorer MR when speech is primarily limited to temporal envelope cues (e.g., Nelson *et al.*, 2003). However, the results of this study indicate that there are selective cases where the temporal properties of the modulated masker, relative to the temporal properties of the speech signal, afford a perceptual benefit from glimpse portions of the vocoded speech signal. This occurred when glimpse durations were relatively long due to the slow peak modulation of the masker signal. In these conditions, faster amplitude modulations of the speech would have remained preserved during these long glimpses.

At slow noise modulation rates listeners obtained some benefit from longer glimpses that accounted for a greater proportion of the available speech (i.e., longer glimpse proportions). However, this occurred at different noise modulation rates for natural and vocoded speech. Perception of vocoded speech was best at the slowest rate provided by 400% time expansion, while performance peaked for natural speech for the faster 200% time expansion condition. However, when TFS cues were available in the natural condition, listeners were also able to utilize glimpses particularly well at faster noise modulation rates. This later observation could be due to the ability to resolve spectral segregation cues within relatively brief glimpses and, importantly, to track changes in these spectral cues across the sentence when there is a fast glimpsing rate. Therefore, more glimpses per second may facilitate speech segregation using the TFS. This is consistent with the established benefit of TFS cues for listening to speech-in-noise (e.g., Lorenzi *et al.*, 2006; Gnansia *et al.*, 2008; Stone *et al.*, 2011b). However, results from the vocoded conditions tested here suggest that temporal envelope cues may also contribute to speech glimpsing when glimpses are of long enough duration to track faster speech temporal envelope modulations within the glimpse.

Combined, this acoustic analysis suggests two primary factors that contribute to speech glimpsing that operate to different degrees depending on the acoustic glimpse conditions. First, when listeners have access to very short glimpses that occur frequently across the sentence, listeners are able to access TFS cues within the glimpse and track those cues distributed across the sentence. Thus, in fast-rate modulated noise conditions, access to TFS cues is essential. This is demonstrated by the little to no MR with vocoded processing in these conditions. Alternatively, it may be the reduced spectral resolution of vocoded processing that is responsible for this effect, and not TFS cues directly. Second, when glimpses are relatively long in duration, listeners are able to resolve important speech temporal envelope modulations within these glimpses that can facilitate MR. This was observed by increased MR during vocoded processing at slower modulation rates. However, as observed with these stimuli, longer glimpse durations were associated with a reduced frequency of glimpse occurrences. This results in large portions of the sentence being significantly masked by the modulated noise. Thus, while listeners may obtain benefit from the temporal envelope during glimpses of speech, they may have little information available to cognitively fill in the missing sentence information for which they have no acoustic cues. This may have been responsible for the reduced MR that was observed for the natural condition during these very slow modulated maskers (i.e., 400%). Moreover, while listeners did have access to TFS cues within each glimpse in natural processing (as observed by better performance in the natural compared to the vocoded condition), they were less able to track spectral cues across the sentence to facilitate segregation for these slow glimpse rates. Indeed, lower glimpse rates were significantly correlated with poorer sentence intelligibility in natural processing. Observations suggest that when noise modulations match the modulation spectrum of the speech (100% condition), significant modulation masking is observed. These results now provide a heuristic to explain how TFS cues and modulation masking work together, based on the rate and duration of speech glimpses, to determine the degree of masking release; and therefore, speech intelligibility in the presence of modulated noise maskers.

### 4. Generalization of results to other SNRs

The above analysis is dependent on both the SNR threshold used for defining glimpsed speech intervals and the global SNR of the speech-in-noise mixture. To facilitate comparison to other conditions, we conducted a subsequent analysis that investigated correlations between performance and each of the four glimpse metrics across a range of threshold values (Fig. 9). The local criterion (LC) was defined as the glimpse threshold (i.e., the 0 dB value used previously), and is plotted as the left column in Fig. 8 across the LC range of −10 to +10 dB. The relative criterion (Kjems *et al.*, 2009) was defined as the difference between the LC and the global SNR used to present the mixture. For defining the temporal glimpses used here, there is a direct correlation between the LC used in the analysis and the mixture SNR, such that increasing the LC by 1 dB is akin to decreasing the global SNR by 1 dB (Brungart *et al.*, 2006).

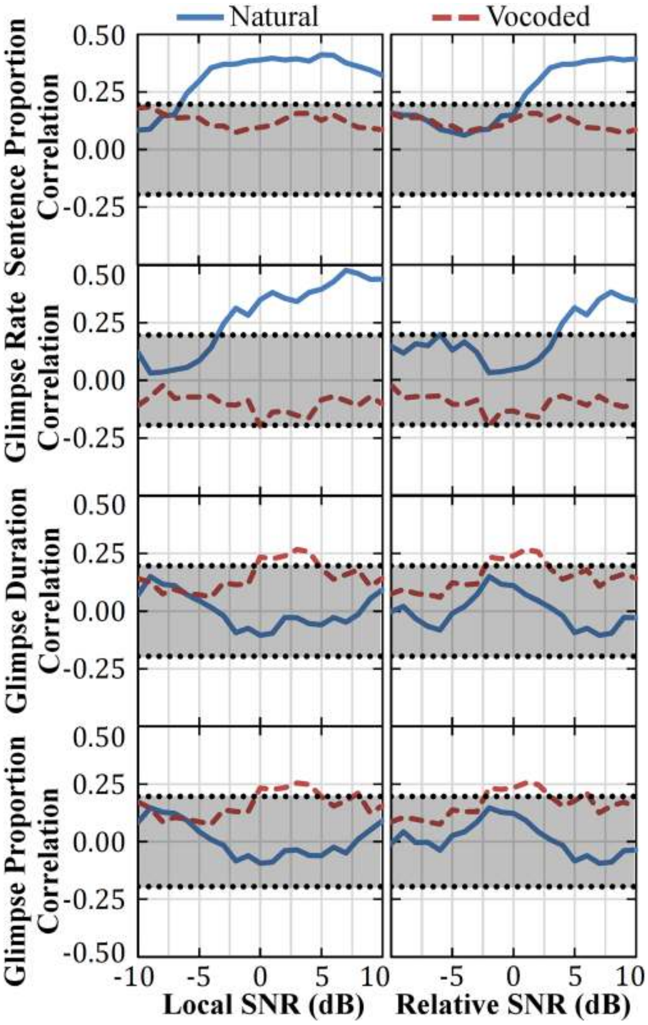

FIG. 9. (Color online) Analysis of correlations between the four glimpse metrics and performance across different glimpse threshold values. (Left column) The local criterion defines the SNR cutoff. (Right column) The relative criterion is the difference, in dB, between the local criterion and the global SNR of the mixture. The relative criterion accounts for intrinsic acoustic glimpse differences between natural and vocoded processing as a result of the different test SNRs. For comparison, the plotted relative criterion range of −10 to 10 dB corresponds to a local criterion range of −17 to 3 dB SNR for natural and −8 to 12 dB SNR for vocoded conditions. Lines that extend beyond the shaded region indicate significant correlations, $p < 0.05$.

The relative criterion is defined as the relative difference between the LC and the mixture SNR. By adjusting the relative criterion, we are able to compare performance between natural and vocoded conditions while controlling for intrinsic differences in the glimpse metrics due to the SNR used during testing. The effect of this relative comparison can be viewed in Fig. 9. The right column plots these relative values across the range of −10 to +10 dB, which, due to the different global SNRs tested, corresponds to a LC range of −17 to +3 dB SNR for the natural condition and −8 to +12 dB SNR for the vocoded condition.

Results of this analysis indicate a large consistency with the results previously reported for 0 dB SNR. Performance with natural speech is most associated with sentence proportion and glimpse rate, as observed by the lines which exceed the shaded zone. In contrast, performance with vocoded speech was more associated with glimpse duration and glimpse proportion. In general, there appears to be a range of threshold values that significantly relate to performance, particularly for the natural speech condition. However, the relative contribution of the glimpse metrics to natural versus vocoded speech conditions appears to remain consistent across threshold values. In addition, the similarity in correlation functions for glimpse duration and glimpse proportion metrics, in combination with the high correlation between these two measures reported previously, further demonstrate that they are not independent.

Qualitative comparison at relative criterion thresholds indicates a potential similarity between natural and vocoded glimpse processing, particularly for the glimpse duration/proportion analysis, with differences related to the magnitude of the correlation. This may indicate the involvement of similar mechanisms contributing to natural and vocoded speech processing but not to the same degree. We would expect that perceptual processes that are involved in vocoded speech processing would also apply in part to natural speech processing which also preserves these temporal envelope cues. However, the better spectral resolution in natural speech appears to enable listeners to also make greater use of other glimpse properties (i.e., rate and sentence proportion). While glimpsing opportunities were available for the vocoded condition at the global 2 dB SNR tested here, future work will have to assess how these functions change as glimpsing mechanisms become more pronounced during vocoded processing, potentially through testing at additional, and poorer, global SNRs or by speech interruption.

## IV. GENERAL DISCUSSION

Overall, the results from this study demonstrate that separation of the modulation spectrum between target and masker speech improves performance due to increasing MR for both natural and vocoded conditions. However, patterns of performance were different between the natural and vocoded conditions, suggesting that performance under natural listening conditions cannot be fully explained by interactions between the temporal envelopes of the speech and masker. The results for the natural and vocoded conditions are discussed separately below.

### A. Natural condition

Lower performance was obtained with natural speech when the modulation spectrum of the competing noise matched that of the target sentence, compared to when the competing modulated noise was time compressed or expanded within a minimal limit (200%, but not 400%). This observation is consistent with early studies of speech interrupted at specific rates. Miller and Licklider (1950) concluded that continuous speech in the presence of an interrupted noise masker at 8–10 Hz results in the best speech intelligibility, compared to different interruption rates, e.g., 4 Hz. As the primary modulation rate for speech is near 4–6 Hz, as measured here and previously by Greenberg *et al.* (1998), the 8 Hz masker rate is most consistent with our 50% time compression condition. Our 50% condition, due to a doubling of the primary modulation rate, presents the peak masker rate at about 8 Hz, which is a less effective masker than at the original 4 Hz rate. This conclusion is also supported by Christiansen and Dau (2012) who found that SAM noise at 8 Hz has less of an effect on MR (in vocoded processing) than a competing-talker masker. This would be expected as a competing talker has most masker energy at the dominant 4–6 Hz region and presumably residual modulation energy in other frequency bins that could also contribute to modulation masking.

This interaction between the speech and masker modulation rate is further supported by Grose *et al.* (2015) who time-compressed the target talker, rather than the modulated noise. In that study they found reduced MR with 50% time compression which resulted in the primary speech fluctuations approximating the 10-Hz fluctuation rate of the masker. Therefore, their result may have been due to a matched modulation rate between speech and masker, which resulted in greater similarity in the modulation domain between the speech and masker. Taken together, these results suggest that MR is reduced when the target and masker have similar modulation rates and improved when they are different. This could potentially be due to either theoretical explanations related to masking of specific modulation rate channels (e.g., Dau *et al.*, 1997) or difficulties with source segregation due to temporal envelope similarities (e.g., Hall and Grose, 1991). Future studies will be required to disambiguate these two possibilities.

These observations indicate that the degree of masker similarity to the modulation spectrum of the target speech is a primary determiner of masking release. However, factors separate from potential modulation masking also play a significant role in overall performance. One example of this in the current study occurred for the 400% time expansion condition. In this condition the masker was time expanded to produce a very slow rate of modulation near 1 Hz. This result may be comparable to some studies using interrupted speech which model the temporal glimpses available during modulated maskers. While the signal for interrupted speech compared to speech in modulated noise is arguably quite different, interruption studies are motivated by overlapping perceptual processes (e.g., glimpsing) that are involved in processing speech in modulated backgrounds. This is supported by strong correlations in performance between the two conditions (Jin and Nelson, 2010). In a seminal speech interruption study, Miller and Licklider (1950) also found a dip in the performance function for 1-Hz interruption rates. The performance difference observed for this condition may be related to different perceptual processes that are involved for processing sentences that are interrupted by slow compared to fast interruption rates. Shafiro *et al.* (2011) and Shafiro *et al.* (2015) have investigated the relative contribution of these processes using sentences interrupted by a primary gating rate (0.5–24 Hz) or by both a primary and faster secondary rate (dual gating). Dual gating sentences resulted in greater intelligibility than single rate sentences for equal proportions of signal preservation due to glimpsing of sentences and then using a cognitive top-down processing to fill in the blanks based on contextual cues. With faster interruption rates (e.g., 8–16 Hz, or 25% and 50% compression), listener performance is determined by perceptual bottom-up processing that depends on the available acoustic cues distributed across the sentence. Once that perceptual processing occurs, the listener can then use top-down cognitive-linguistic processes to fill in additional information about the missing portions of speech. With the slower interruption rates (e.g., 1–2 Hz, or 200% and 400% compression), the listener can use this cognitive top-down processing to fill in the blanks by using the contextual information to infer the missing sections of speech. However, listeners have no or significantly reduced perceptual information during the missing or noise masked portions of the signal during the peaks of the modulated masker. At the slowest rate tested (i.e., 1 Hz, or the 400% condition as defined here), large portions of the speech signal were masked by the competing noise. Thus, listeners have little perceptual information available during those intervals to fill in the missing speech information, which may have resulted in the poorer MR scores. Indeed, the keyword subanalysis of the 400% condition supports this hypothesis that listeners had a limited "filling-in" ability to correctly report keywords that were not glimpsed. Instead, performance appears to be determined mostly by perceptual access to temporal envelope cues that become available during the long glimpse durations. This observation is supported by superior MR for vocoded compared to natural speech in this 400% condition.[4]

## B. Vocoded condition

As supported by previous research (Nelson *et al.*, 2003; Füllgrabe *et al.*, 2006; Li and Loizou, 2009), the vocoded conditions, which simulated the speech heard through a CI, resulted in significantly less MR than the natural condition. This could be explained by the lack of acoustic TFS cues in the vocoded condition, which could result in difficulty separating out the talker from the masker (Apoux *et al.*, 2013; Stone *et al.*, 2011b). Indeed, the minimal MR that was observed (6%–13%) could potentially be attributed to the preservation of periodicity cues in the vocoder processing that may have facilitated source segregation. The poorer performance is not explained by poor intelligibility of the vocoded speech, as listeners performed at high levels for vocoded speech in quiet (mean word recognition was 90.7%). Results also showed that time compression had little effect on MR. That is, listeners received little benefit of separating the target and masker modulation spectrum. This may have been due to the better SNR that was used compared to the natural speech conditions (see Bernstein and Grant, 2009).

The one exception to this observation was the 400% condition that presented masker modulations at significantly slower rates (i.e., by a factor of 4). This condition reflects a critical difference compared to previous investigations by demonstrating significant MR for vocoded speech in the presence of modulated maskers that exceeded MR obtained for the natural speech condition. Here we propose that this effect is due to the long glimpse durations that occur with slow modulated maskers that preserved a relatively long duration of speech at a favorable SNR. These long periods of preserved, continuous portions of speech resulted in listeners' ability to resolve faster amplitude modulations of the vocoded target talker. Accurate recognition during these long glimpse intervals also facilitated listeners' ability to contextually fill-in keywords that were not glimpsed. This further amplified recognition for vocoded speech in the 400% condition.

## C. Further research and limitations

An important consideration in this study is the different long-term average (global) SNRs used in natural compared to vocoded testing. While the findings here depart from most previous work in terms of finding a fluctuating masker benefit in vocoded listening, at least for very slow masker modulation rates, it is possible that greater benefits would surface given comparable SNRs. However, our acoustic analysis across a number of relative SNR criterion values provides some potential insight into possible effects at different global SNR conditions. These results suggest that listeners do make use of different glimpse properties in the natural speech conditions (associated with glimpse rate and sentence proportion) compared to vocoded speech. In the latter case, the duration of usable glimpses provides the best explanation of the listener's ability to process the target amplitude modulation cues, potentially due to energetic masking of the target modulation during the peaks of the masker.

In addition, the ability to generalize these results to additional speech materials needs to be investigated. The current study used short IEEE sentences. However, different effects could be observed for longer passages of speech that may contain more energy at slower modulation rates. However, the interaction between the modulation spectrum of the target and competing messages is expected to be relative and not tied directly to the specific modulation rates present in these IEEE materials. Therefore, we would still expect performance to improve as background modulation becomes faster than the rate of the target speech materials and, potentially, to be limited by significantly slower background modulations that produce long durations of the target speech that are masked. However, future investigations will have to detail this relationship for target speech with different modulation spectra.

As noted in the Introduction, complicating the comparison between overlapping modulation spectra is the finding of interactions between the complex temporal envelopes of the target and masker that result in perceptual beating due to 2nd order amplitude modulation effects (see Füllgrabe and Lorenzi, 2003, 2005). This may be due to modulation distortion products that are introduced by auditory system nonlinearities in response to the complex envelopes (e.g., Shofner *et al.*, 1996). These complex, nonlinear interactions will have to be assessed in order to fully describe modulation spectra interactions of two competing talkers. However, even in the context of these complex interactions, Füllgrabe *et al.* (2006) concluded that consonant identification is still strongly determined by the ability of the listener to glimpse speech and the number and duration of available speech glimpses. Consistent with this conclusion, the glimpse analysis of our study demonstrated significant correlations among the glimpse metrics and intelligibility, even with the complex temporal envelopes that were used.

Finally, the processing used in the current study preserved periodicity cues for both natural and vocoded speech—but not for the modulated masker. As such, we attempted to limit the interference of the masker to the slow modulation rates. We speculate that if faster modulation rates associated with periodicity were included in the masker, this would result in greater interference for natural and vocoded speech. We suspect that periodicity in the masker could result in increased difficulty with source segregation, and therefore reduce MR. From studies of concurrent talkers (e.g., Darwin *et al.*, 2003; Lee and Humes, 2012), we would expect better source segregation with greater differences between target and masker periodicity. Future studies will have to assess the independent contribution of masker periodicity and if it interacts with listeners' ability to capitalize on differences in the modulation spectra and/or in making use of the available speech glimpses.

## V. SUMMARY AND CONCLUSIONS

(1) In the natural condition, time compression or expansion of the single-talker modulated noise resulted in an increase in MR for most conditions. The increase in MR can be attributed to an increased separation of the modulation spectra of the target and masker speech that reduces the degree of modulation masking.

(2) The decrease in MR with the 400% natural condition can be attributed to reduced perceptual cues distributed across the sentence due to the large intervals of noise.

(3) In the vocoded condition, time compression or expansion of the modulated noise had little to no impact on MR, with the exception of the 400% condition. Overall, the vocoded condition resulted in significantly less MR than the natural condition, which can be explained by reduced spectral cues that may facilitate segregation of the target and masker speech and/or the SNR used in testing.

(4) The significant MR with the 400% vocoded condition can be attributed to increased access to amplitude modulations of the target speech due to longer dips in the modulated noise masker.

(5) Acoustic analyses of glimpsed portions of the speech signal were conducted. In the natural condition, sentence proportion and glimpse rate were most correlated with sentence intelligibility. For vocoded processing, glimpse duration and glimpse proportion showed the strongest intelligibility correlations, particularly at noise modulation rates that were slower than the target speech.

(6) Investigation of the natural and vocoded speech conditions suggests the importance of spectral cues to glimpsing, particularly when the noise masker was modulated at rates that were faster than the target speech. However, at very slow masker modulation rates, MR appears to be more explained by the relative preservation of speech amplitude modulations during long glimpses.

## ACKNOWLEDGMENTS

This work was supported, in part, by a Magellan Scholars grant (J.X.) and an ASPIRE-I grant (D.F.) from the University of South Carolina. Support was also provided by the National Institutes of Health, National Institute on Deafness and Other Communication Disorders, Grant Nos. R03-DC012506 and R01-DC015465 (D.F.).

[1] However, reductions in MR could still be observed even with separated modulation spectra due to second-order beating at target modulation rates (Füllgrabe *et al.*, 2005).

[2] Masking release data were not available for this analysis over individual items because sentences were not presented in both modulated and unmodulated background noise. Therefore, accuracy data were used. In order to stabilize the error variance across the different trials, percent correct scores for each sentence were transformed into RAU scores.

[3] Of course, this comparison comes with the caveat that natural and vocoded speech were tested at different SNRs.

[4] Note that, contrary to our observations, less MR would be expected for the vocoded speech because of the comparably better test SNR (Bernstein and Grant, 2009). This underscores access to significant information provided by the vocoded speech modulation in this condition.

Apoux, F., Yoho, S. E., Youngdahl, C. L., and Healy, E. W. (**2013**). "Role and relative contribution of temporal envelope and fine structure cues in sentence recognition by normal-hearing listeners," J. Acoust. Soc. Am. **134**, 2205–2212.

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

Boersma, P., and Weenink, D. (**2014**). "Praat: Doing phonetics by computer [computer program] (version 5.3.80)," http://www.praat.org/ (Last viewed June 29, 2014).

Brungart, D., Chang, P. S., Simpson, B. D., and Wang, D. L. (**2006**). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. **120**, 4007–4018.

Buss, E., Whittle, L. N., Grose, J. H., and Hall, J. W. (**2009**). "Masking release for words in amplitude-modulated noise as a function of modulation rate and task," J. Acoust. Soc. Am. **126**, 269–280.

Christiansen, C., and Dau, T. (**2012**). "Relationship between masking release in fluctuating maskers and speech reception thresholds in stationary noise," J. Acoust. Soc. Am. **132**, 1655–1666.

Christiansen, C., MacDonald, E. N., and Dau, T. (**2013**). "Contribution of envelope periodicity to release from speech-on-speech masking," J. Acoust. Soc. Am. **134**, 2197–2204.

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

Darwin, C. J., Brungart, D. S., and Simpson, B. D. (**2003**). "Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers," J. Acoust. Soc. Am. **114**, 2913–2922.

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.

Drullman, R., Festen, J. M., and Plomp, R. (**1994**). "Effect of reducing slow temporal modulations on speech reception," J. Acoust. Soc. Am. **95**, 2670–2680.

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.

Ewert, S. D., and Dau, T. (**2000**). "Characterizing frequency selectivity for envelope fluctuations," J. Acoust. Soc. Am. **108**, 1181–1196.

Festen, J. M., and Plomp, R. (**1990**). "Effects of fluctuating noise and interfering speech on speech-reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. **88**, 1725–1736.

Fogerty, D. (**2011a**). "Perceptual weighting of individual and concurrent cues for sentence intelligibility: Frequency, envelope, and fine structure," J. Acoust. Soc. Am. **129**, 977–988.

Fogerty, D. (**2011b**). "Perceptual weighting of the envelope and fine structure across frequency for sentence intelligibility: Effect of interruption at the syllabic-rate and periodic-rate of speech," J. Acoust. Soc. Am. **130**, 489–500.

Fogerty, D. (**2014**). "Importance of envelope modulations during consonants and vowels in segmentally interrupted sentences," J. Acoust. Soc. Am. **135**, 1568–1576.

Fogerty, D., and Xu, J. (**2016**). "Speech recognition interference by the temporal and spectral properties of a single competing talker," J. Acoust. Soc. Am. **140**, EL197–EL203.

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.

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

Füllgrabe, C., and Lorenzi, C. (**2003**). "The role of envelope beat cues in the detection and discrimination of second-order amplitude modulation," J. Acoust. Soc. Am. **113**, 49–52.

Füllgrabe, C., and Lorenzi, C. (**2005**). "Perception of the envelope-beat frequency of inharmonic complex temporal envelopes," J. Acoust. Soc. Am. **118**, 3757–3765.

Füllgrabe, C., Moore, B. C., Demany, L., Ewert, S. D., Sheft, S., and Lorenzi, C. (**2005**). "Modulation masking produced by second-order modulators," J. Acoust. Soc. Am. **117**(4), 2158–2168.

Füllgrabe, C., Moore, B. C. J., and Stone, M. A. (**2015**). "Age-group differences in speech identification despite matched audiometrically normal hearing: Contributions from auditory temporal processing and cognition," Front. Aging Neurosci. **6**, 347.

Füllgrabe, C., Stone, M. A., and Moore, B. C. J. (**2009**). "Contribution of very low amplitude-modulation rates to intelligibility in a competing-speech task," J. Acoust. Soc. Am. **125**, 1277–1280.

Gallun, F., and Souza, P. (**2008**). "Exploring the role of the modulation spectrum in phoneme recognition," Ear Hear. **29**, 800–813.

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.

Gilbert, G., Bergeras, I., Voillery, D., and Lorenzi, C. (**2007**). "Effects of periodic interruptions on the intelligibility of speech based on temporal fine-structure or envelope cues," J. Acoust. Soc. Am. **122**, 1336–1339.

Gnansia, D., Jourdes, V., and Lorenzi, C. (**2008**). "Effect of masker modulation depth on speech masking release," Hear. Res. **239**, 60–68.

Greenberg, S., Arai, T., and Silipo, R. (**1998**). "Speech intelligibility derived from exceedingly sparse spectral information," in *Proceedings of the International Conference on Spoken Language Processing*, Sydney, Australia, pp. 2803–2806.

Grimault, N., Bacon, S. P., and Micheyl, C. (**2002**). "Auditory stream segregation on the basis of amplitude-modulation rate," J. Acoust. Soc. Am. **111**, 1340–1348.

Grose, J. H., Griz, S., Pacifico, F. A., Advincula, K. P., and Menezes, D. C. (**2015**). "Modulation masking release using the Brazilian-Portuguese HINT: Psychometric functions and the effect of speech time compression," Int. J. Audiol. **54**, 274–281.

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

Hall, J. W., and Grose, J. H. (**1991**). "Notched-noise measures of frequency selectivity in adults and children using fixed-masker-level and fixed-signal-level presentation," J. Speech Lang. Hear. Res. **34**, 651–660.

Hopkins, K., and Moore, B. C. J. (**2009**). "The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise," J. Acoust. Soc. Am. **125**, 442–446.

Howard-Jones, P. A., and Rosen, S. (**1993**). "Uncomodulated glimpsing in 'checkerboard' noise," J. Acoust. Soc. Am. **93**, 2915–2922.

Institute of Electrical and Electronics Engineers (IEEE) (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 225–246.

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.

Kjems, U., Boldt, J. B., Pedersen, M. S., Lunner, T., and Wang, D. L. (**2009**). "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," J. Acoust. Soc. Am. **126**, 1415–1426.

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.

Lee, J. H., and Humes, L. E. (**2012**). "Effect of fundamental-frequency and sentence-onset differences on speech-identification performance of young and older adults in a competing-talker background," J. Acoust. Soc. Am. **132**, 1700–1717.

Li, N., and Loizou, P. C. (**2007**). "Factors influencing glimpsing of speech in noise," J. Acoust. Soc. Am. **122**, 1165–1172.

Li, N., and Loizou, P. C. (**2009**). "Factors affecting masking release in cochlear-implant vocoded speech," J. Acoust. Soc. Am. **126**, 338–346.

Loizou, P. C. (**2007**). *Speech Enhancement: Theory and Practice* (CRC Press, Taylor and Francis, Boca Raton, FL), 608 pp.

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., and Moore, B. C. J. (**2006**). "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure," Proc. Natl. Acad. Sci. U.S.A. **103**, 18866–18869.

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

Millman, R., Lorenzi, C., Apoux, F., Fullgrabe, C., Green, G., and Bacon, S. (**2002**). "Effect of duration on amplitude-modulation masking," J. Acoust. Soc. Am. **111**, 2551–2554.

Moore, B. C. J., Füllgrabe, C., and Sek, A. (**2009**). "Estimation of the center frequency of the highest modulation filter," J. Acoust. Soc. Am. **125**, 1075–1081.

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

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

Oxenham, A. J., and 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.

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.

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.

Shafiro, V., Sheft, S., Risley, R., and Gygi, B. (**2015**). "Effects of age and hearing loss on the intelligibility of interrupted speech," J. Acoust. Soc. Am. **137**, 745–756.

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

Shofner, W. D., Sheft, S., and Guzman, S. J. (**1996**). "Responses of ventral cochlear nucleus units in the chinchilla to amplitude modulation by low-frequency, two-tone complexes," J. Acoust. Soc. Am. **99**, 3592–3605.

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

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

Stone, M. A., Füllgrabe, C., and Moore, B. C. J. (**2008**). "Benefit of high-rate envelope cues in vocoder processing: Effect of number of channels and spectral region," J. Acoust. Soc. Am. **124**, 2272–2282.

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

Stone, M. A., and Moore, B. C. (**2004**). "Side effects of fast-acting dynamic range compression that affect intelligibility in a competing speech task," J. Acoust. Soc. Am. **116**(4), 2311–2323.

Stone, M. A., and Moore, B. C. J. (**2014**). "On the near non-existence of 'pure' energetic masking release for speech," J. Acoust. Soc. Am. **135**, 1967–1977.

Stone, M. A., Moore, B. C. J., and Füllgrabe, C. (**2011b**). "The dynamic range of useful temporal fine structure cues for speech in the presence of a competing talker," J. Acoust. Soc. Am. **130**, 2162–2172.

Studebaker, G. A. (**1985**). “A rationalized arcsine transform,” J. Speech Language Hearing Res. **28**(3), 455–462.

Van Tasell, D. J., Soli, S. D., Kirby, V. M., and Widin, G. P. (**1987**). “Speech waveform envelope cues for consonant recognition,” J. Acoust. Soc. Am. **82**, 1152–1161.

Wang, X., and Humes, L. E. (**2010**). “Factors influencing recognition of interrupted speech,” J. Acoust. Soc. Am. **128**, 2100–2111.

Yost, W. A. (**1992**). “Auditory perception and sound source determination,” Curr. Dir. Psychol. Sci. **1**, 179–184.

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

Yost, W. A., Sheft, S., and Opie, J. (**1989**). “Modulation interference in detection and discrimination of amplitude modulation,” J. Acoust. Soc. Am. **86**, 2138–2147.