Speech-on-speech masking with variable access to the linguistic content of the masker speech

Lauren Calandruccio[a)]
Department of Linguistics and Communication Disorders, Queens College of the City University of New York, Flushing, New York 11367

Sumitrajit Dhar
Roxelyn and Richard Pepper Department of Communication Disorders, Northwestern University, Evanston, Illinois, 60208

Ann R. Bradlow
Department of Linguistics Northwestern University, Evanston, Illinois 60208

(Received 21 October 2009; revised 11 March 2010; accepted 9 June 2010)

It has been reported that listeners can benefit from a release in masking when the masker speech is spoken in a language that differs from the target speech compared to when the target and masker speech are spoken in the same language [Freyman, R. L. *et al.* (1999). J. Acoust. Soc. Am. **106**, 3578–3588; Van Engen, K., and Bradlow, A. (2007), J. Acoust. Soc. Am. **121**, 519–526]. It is unclear whether listeners benefit from this release in masking due to the lack of linguistic interference of the masker speech, from acoustic and phonetic differences between the target and masker languages, or a combination of these differences. In the following series of experiments, listeners' sentence recognition was evaluated using speech and noise maskers that varied in the amount of linguistic content, including native-English, Mandarin-accented English, and Mandarin speech. Results from three experiments indicated that the majority of differences observed between the linguistic maskers could be explained by spectral differences between the masker conditions. However, when the recognition task increased in difficulty, i.e., at a more challenging signal-to-noise ratio, a greater decrease in performance was observed for the maskers with more linguistically relevant information than what could be explained by spectral differences alone.
 [DOI: 10.1121/1.3458857]

PACS number(s): 43.71.Es, 43.71.Hw, 43.72.Dv [MSS] Pages: 860–869

I. INTRODUCTION

Native-English listeners can improve their performance on an English recognition task when the background, or competing, speech is spoken in a language that differs from the target speech compared to a background of English speech (Freyman *et al.*, 1999; Garcia Lecumberri and Cooke, 2006; Van Engen and Bradlow, 2007). It is unclear whether listeners receive this release from masking because 1) the masker speech is spoken in a language that is simply different from the target speech, and therefore has different spectral and temporal properties, providing less energetic masking, 2) the masker speech is spoken in a language unknown to the listener, therefore providing no linguistic meaning, and in turn providing less informational masking, or 3) a combination of these two possibilities. One way to probe this question is to examine listeners' recognition of speech in the presence of competing speech maskers that vary in the amount of linguistic content, or in the amount that listeners are able to "understand" the masking speech. In the current study, we present data for native-English monolingual listeners in the presence of a series of masker conditions that vary in linguistic content.

Understanding speech in the presence of background noise is a complex task. The complexity originates from both peripheral and central phenomena (e.g., Carhart *et al.*, 1969; Gelfand *et al.*, 1986; Jerger, 1992). Peripherally we contend with energetic interference, or scenarios in which it is difficult to understand the target signal because of similar excitation patterns along the auditory periphery caused by the competing signal(s). Some competing auditory signals (e.g., speech), however, may decrease listeners' performance on an auditory task further than what would be predicted due to energetic masking alone (Carhart, *et al.*, 1969; Lutfi *et al.*, 2003; Watson, 2005; Durlach *et al.*, 2003). This additional masking is often referred to as informational and potentially occurs, though perhaps not exclusively, in the central auditory system. When thinking of informational masking as a simple speech-in-speech recognition task, it is easy to understand how listeners might confuse competing speech signals with the speech signal of interest, causing greater difficulty with the recognition task. This has been documented numerous times in the laboratory, with listeners demonstrating difficulty attending to the target talker when competing talkers, especially those with similar voices (e.g., two females), are speaking simultaneously (e.g., Hornsby *et al.*, 2006; Rajan and Cainer, 2008; Helfer and Freyman, 2008; Cooke *et al.*, 2008).

[a)]Author to whom correspondence should be addressed. Electronic mail: lauren.calandruccio@qc.cuny.edu

Informational masker signals can be manipulated in various ways, resulting in interesting and informative patterns of release in masking. For example, switching the gender of the talker of the competing signals to oppose that of the target talker (e.g., Helfer and Freyman, 2008), time reversal of the competing speech (Freyman *et al.*, 2001), spatial separation of the competing speech (Freyman *et al.*, 1999; Aaronson *et al.*, 2008), and interaural time delays between the target and competing speech signals (Carhart *et al.*, 1967) all can improve listeners' recognition. Recently, several researchers have documented a release in masking on an English speech recognition task when the masker is changed from competing English speech to a language unfamiliar to the listener. Such a release in masking has been reported for competing Dutch (Freyman *et al.*, 2001), Spanish (Garcia Lecumberri and Cooke, 2006), Mandarin (Van Engen and Bradlow, 2007), and Croatian (Calandruccio *et al.*, 2008) speech maskers. It has been argued that this release in masking is caused by the lack of "understanding" of the masker speech by the listener group, resulting in the foreign-language masker providing less informational masking in comparison to the English-speech masker. However, it remains unclear whether the release in masking is caused by the masker speech being spoken in a language that 1) is linguistically unfamiliar to the listeners and/or 2) simply differs at the spectro-temporal level from the target language. That is, this release in masking may be energetically driven since the target and masker languages not only vary in linguistic content, but also vary phonetically and acoustically.

We hypothesize that if listeners receive this release in masking due to the inability to understand, or obtain meaning from, the masker speech (causing less confusion or providing less information), then we should see listeners' speech recognition progressively increase as the access to the linguistic content of the speech masker decreases. Data are reported for 2-talker maskers that vary in the degree of linguistic content (English, Mandarin-accented English, and Mandarin speech) and for temporally modulated and steady state spectrally shaped, and temporally modulated white-noise masker conditions. We chose to use 2-talker maskers because 1) this release in masking has been seen with two competing foreign-language talkers (Freyman *et al.*, 2001; Van Engen and Bradlow, 2007) and 2) these maskers provide enough energetic masking to provide us with usable data (no floor or ceiling effects) for both the speech and noise maskers when presented at similar signal-to-noise ratios.

II. EXPERIMENT I: 2-TALKER MASKERS

A. Methods

Twenty-six young-adult normal-hearing listeners (14 females and 12 males) ranging between 19 and 34 years old (M age=22 years, SD=3 years) participated in these experiments. All listeners were native-monolingual speakers of American English with no knowledge of Mandarin. The institutional review board at Northwestern University approved all procedures. Listeners were paid for their participation and provided written-informed consent.

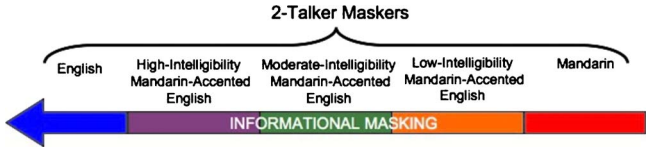

FIG. 1. (Color online) Schematic illustration of the 2-talker maskers used in Experiment I. Ten different male voices were used to create the five masker conditions. All five maskers varied in the amount of linguistic content available to the monolingual-English listeners tested throughout these experiments.

Prior to participation, otoscopic evaluation was performed on all listeners to ensure clear ear canals. All listeners had hearing thresholds <20 dB HL between 250 and 8000 Hz, bilaterally [American National Standards Institute (ANSI), 2004], as tested with standard clinical audiological procedures [American Speech-Language-Hearing Association (ASHA), 2005] using a Maico M26 clinical audiometer.

B. Stimuli

1. Target stimuli

Sentences from Harvard/IEEE sentence lists (IEEE Subcommittee on Subjective Measurements IEEE Recommended Practices for Speech Quality Measurements, 1969) spoken by a native-English speaking male were used for the target stimuli. All target sentences were recorded at Northwestern University in a double-walled sound-treated room at a 44.1 kHz sampling rate with 16-bit resolution. Sentences were digitally edited using custom software developed in MaxMSP (Cycling, 74′ Version 5.0, 2008) to remove silence at the end and at the beginning of each sentence. Once edited, all sentences were root-mean-square (RMS) normalized to the same pressure level using Praat (Boersma and Weenink, 2009).

2. Masker stimuli

Five different 2-talker maskers were created using a total of 10 different male voices. The five distinct maskers were created to assess speech recognition performance when access to linguistic content of the speech masker progressively increased (see Fig. 1 for an illustration of the masker conditions used in Experiment I). Four out of the 10 male talkers spoke (with no detectable accent) in their native language. Two of the male talkers were native-English talkers speaking English and two were native-Mandarin talkers speaking Mandarin. The remaining six male talkers were native Mandarin talkers speaking English. The recordings of these six male talkers were taken from the Northwestern University Foreign Accented Speech Database (NUFASD; Bent and Bradlow, 2003). The NUFASD consists of 32 non-native English talkers producing the same 64 sentences from the BKB-R sentence lists (Bench *et al.*, 1979; ® Cochlear Corporation). The database includes production intelligibility data on all talkers. Intelligibility was assessed based on the perception of native-English speaking, normal-hearing listeners and their ability to understand the non-native English speech in the presence of a white-noise masker presented at 5 dB SNR. The recordings of the six talkers used in this study

were chosen based on their native language (Mandarin), the similarity of the two talkers' production score (for each 2-talker masker) and the overall intelligibility of their English production. The intelligibility scores for the two talkers used for the low-intelligibility, moderate-intelligibility and high-intelligibility Mandarin-accented English maskers were 43 and 45%, 65 and 67%, and 88 and 88%, respectively. The *same* 64 sentences from the BKB sentence lists used in the NUFASD were recorded by the two native-English male talkers. The two native-Mandarin talkers recorded Mandarin-translated versions of the same sentences.

The 64 sentences spoken by all 10 male talkers were normalized to the same root-mean-square (RMS) pressure level. The five 2-talker maskers were created by concatenating the 64 sentences of each talker with no silent intervals between sentences. The order of the concatenation differed between the two talkers in each masker condition. The two strings of 64 sentences were combined together into one single audio file using Audacity©, great care was taken to ensure that the sentences spoken by the two talkers did not end or begin at the same time. Lastly, the five final audio files for each masker condition were RMS normalized to the same overall pressure.

C. Procedure

Listeners were seated in a comfortable chair in a double-walled sound-treated room. Stimuli were presented to the listeners via Etymotic foam insert eartips (13 mm). The target speech and speech masker stimuli were mixed in real time using custom software created using MaxMSP running on an Apple Macintosh computer. Stimuli were passed to a MOTU 828 MkII input/output firewire device for digital-to-analog conversion (44100 Hz, 24 bit), passed through a Behringer Pro XL headphone amplifier and output to MB Quart 13.01HX drivers.

Listeners were presented with a total of 200 target sentences. The level of the target sentences was fixed at 65 dB SPL and the level of the competing speech masker varied around the level of the target speech (e.g., an SNR of −3 indicated that the speech was fixed at 65 dB SPL, while the masker speech was fixed at 68 dB SPL). One target sentence was played on each trial and a random portion of the appropriate babble masker was presented one second longer than the target sentence (500 ms prior to the beginning of the sentence, and 500 ms at the end of the sentence). The first 100 sentences were presented at a fixed SNR of −3 dB, and the second 100 sentences were presented at a fixed SNR of −5 dB. We chose to use two different SNR conditions in order to ensure the collection of usable data on all of our listeners (avoiding ceiling or floor effects, for either the easier or more difficult SNR condition, respectively, or the easier or more difficult language-masker conditions, respectively). That is, we wanted to be able to account for inter-subject variability in performance that has been observed in previous informational masking experiments (e.g., Kidd *et al.*, 1994; Freyman *et al.*, 2007; Van Engen and Bradlow, 2007). This protocol also ensured that any practice effects observed within the experiment would be counterbalanced by the second SNR condition (the more difficult condition) being presented in the second half of the experiments. The presentation of each masker condition was randomly varied across listeners and 20 sentences were presented per masker condition per SNR (5 masker conditions × 2 SNR × 20 sentences = 200 sentences total).

Listeners' responses were scored online by an examiner in the control room adjacent to the test booth based on five keywords/sentence. Listeners' responses were also digitally recorded and later rescored for reliability purposes. Scores that were not in agreement between the two examiners were reassessed and a score was agreed upon. This disagreement occurred in 1.4% of the total trials.

D. Results

Data were transformed into rationalized arcsine units (RAU; Studebaker, 1985) to normalize the error variance of performance scores, and all statistical analyses were calculated based on these transformations. A 5 × 2 repeated-measures analysis of variance (ANOVA) was performed on two within-subject factors (masker condition and SNR). Results indicated a significant interaction of masker condition × SNR [$F(4,100)=4.99$, $p=0.001$] and a significant main effect of masker condition [$F(4,100)=210.39$, $p<0.0001$] and SNR [$F(1,25)=89.54$, $p<0.0001$]. Post-hoc analyses comparing all possible 2-way ANOVAs (using a Bonferroni adjusted critical value) indicated that a significant interaction between masker condition × SNR only existed between the English masker condition and the other four 2-talker maskers [$F(1,25)$ ranging from 7.90–19.61, p ranging from 005–<0.0001]. That is, performance for the English 2-talker masker disproportionately decreased compared to the other four 2-talker maskers at the more difficult SNR. The decrease observed in performance was proportional for the more difficult SNR condition for the other four maskers (English-high intelligibility, English–moderate intelligibility, English–low intelligibility, and Mandarin).

Post-hoc pair wise comparisons (also using a Bonferroni adjusted critical alpha level) indicated that listeners' performance was significantly poorer in the presence of the native-English and the native-Mandarin 2-talker masker conditions at the −3 *SNR condition* compared to the other three 2-talker maskers. Performance at the −3 SNR condition in the presence of the English and the Mandarin 2-talker maskers did not significantly differ from each other. Listeners' performance in the English-low intelligibility 2-talker masker condition was significantly better in comparison to the other four 2-talker masker conditions, at −3 SNR.

Similar post-hoc comparisons for performance scores at the −5 SNR condition indicated a significant difference in performance between *all five* 2-talker maskers. That is, listeners performed significantly worse in the presence of the English 2-talker masker compared to all of the other masker conditions. Performance in the presence of the 2-talker Mandarin masker was significantly poorer in comparison to the three degraded Mandarin-accented English speech masker conditions. And performance significantly increased as the degree of intelligibility decreased across the three Mandarin-

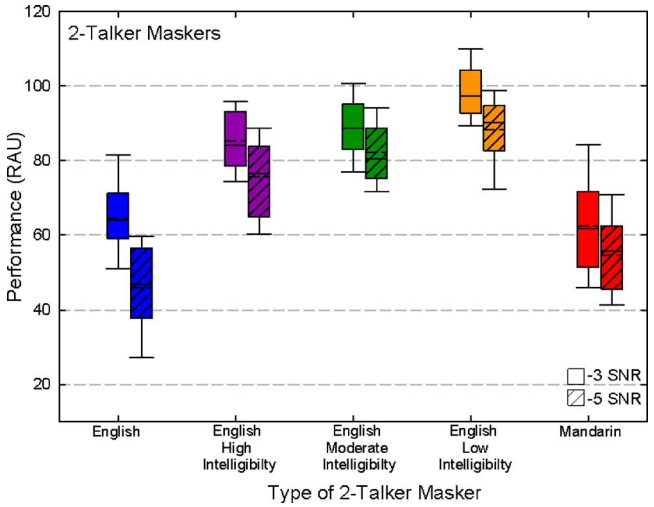

FIG. 2. (Color online) Sentence recognition results for 26 native-English speaking normal-hearing listeners. Performance in RAU is shown for two different SNR conditions for five different 2-talker maskers. The dashed lines within each box plot represent the mean data, while the solid lines represent the median. Significant differences in performance are reported in the text.

accented English-speech masker conditions (see Fig. 2 for all of the results from Experiment I).

In summary, listeners only benefited from a release in masking relative to the English masker condition when the masker speech was spoken in Mandarin during the more difficult (−5 dB) SNR listening condition. For the non-native English speech maskers, performance significantly increased as access to lexical information of the accented-*English* speech maskers decreased. For the easier SNR condition, fewer differences between masker conditions were observed with listeners' showing the poorest performance in the presence of the native-English speech masker. Since all listeners were native monolingual speakers of English the *Mandarin-accented English* speech maskers should have provided greater amounts of lexical information compared to the *Mandarin* speech masker. However, performance in the presence of the Mandarin masker was significantly *poorer* compared to the performance in the presence of the *Mandarin-accented English* speech masker conditions (indicating no release in masking for the Mandarin speech masker compared to Mandarin-accented English speech maskers with varying amounts of intelligibility).

III. EXPERIMENT II: SPECTRALLY MATCHED NOISE MASKERS

To further examine differences between the five 2-talker masker conditions reported in Experiment I, a second experiment was conducted. The purpose of this Experiment was to remove the linguistic differences between the 2-talker maskers used in Experiment I and to examine the contributions from the spectral *and* temporal properties of these maskers that could have influenced the results. Both steady-state and temporally modulated noise maskers were included in Experiment II. These two types of maskers were included to determine how both the spectral properties and the temporal modulations of the 2-talker maskers (used in Experiment I) could have influenced their masking effectiveness. For example, maskers with greater spectral energy in certain frequency regions could have resulted in greater masking. Whereas, those maskers with greater temporal modulations could have allowed listeners to recognize more information from the target speech by "listening in the dips" (Festen and Plomp, 1990) resulting in less masking. All 10 maskers (5 steady state and 5 temporally modulated) were spectrally matched to the original five 2-talker maskers used in Experiment I. Keeping the spectral energy constant across the masker conditions would allow us to examine listener performance when 1) linguistic information was removed from the maskers used in Experiment I (the modulated noise masker conditions) and when 2) both linguistic information and temporal-modulation information were removed (the steady-state masker conditions).

A. Subjects

Twenty-three of the original 26 listeners who participated in Experiment I returned to the laboratory to participate in Experiment II.

B. Stimuli

Target stimuli included 200 additional sentences (not previously used in Experiment I) from the Harvard/IEEE sentence lists spoken by the same male talker, recorded using the same procedures as reported in Experiment I. To remove linguistic content from all of the 2-talker maskers (a) five *temporally modulated spectrally shaped noise maskers and (b) five steady-state spectrally shaped noise maskers both matched to the five 2-talker maskers used in Experiment I were generated. Noise spectrally matched to the average spectrum of each of the five 2-talker maskers used in Experiment I was generated using MATLAB. The noise was created by passing a Gaussian white noise through an FIR filter with a magnitude response equal to the LTASS of the 2-talker masker sentences. These noises were saved into audio files and used for both the steady-state and for the temporally modulated noise maskers. To create the temporally modulated noise maskers the temporal envelopes of the 2-talker maskers were computed in MATLAB. A full-wave rectification Hilbert transform was applied to the stimuli which were then low-pass filtered using a rectangular filter with a cut-off frequency equal to 50 Hz and a sampling frequency of 22.1 kHz. The spectrally matched noise described above was than multiplied by each of the five original respective envelopes to create temporally modulated, spectrally shaped noise maskers (one for each of the five 2-talker masker conditions). All 10 spectrally matched noise maskers (5 steady-state and 5 temporally modulated) were RMS normalized to the same pressure level using Praat. All noise maskers were presented at a fixed SNR of −5 dB. An SNR of −5 dB was chosen to provide a level of difficulty that would not cause listeners' performance to reach either ceiling or floor levels and to have a direct comparison to the −5 dB SNR condition used in Experiment I.*

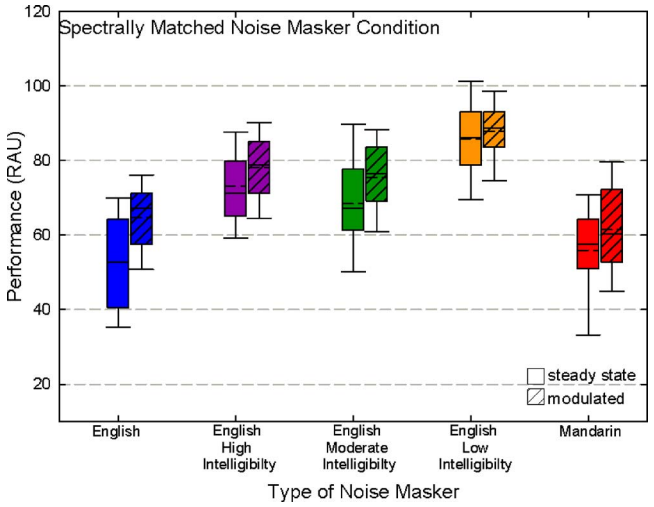

FIG. 3. (Color online) Sentence recognition results for 23 native-English speaking normal-hearing listeners. Performance in RAU is shown for both modulated and steady-state spectrally matched noise maskers presented at an SNR of −5 dB. The types of maskers were generated based on the 2-talker maskers used in Experiment I. Dashed lines within each box plot represent the mean, while solid lines represent the median. Significant differences in performance are reported in the text.

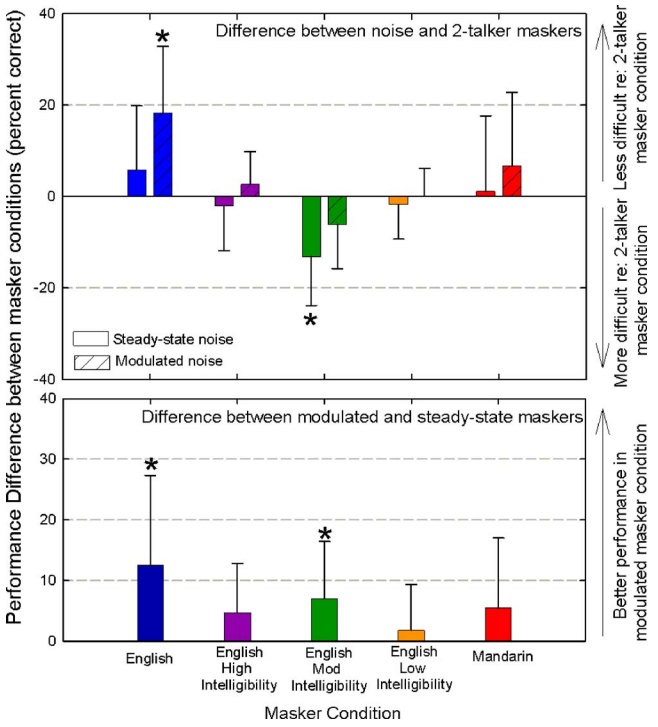

FIG. 4. (Color online) Differences in performance scores between (top) Experiments I and II and −5 dB SNR, and between (bottom) modulated and steady-state noise maskers in Experiment II.

C. Results

All statistical analyses and figures reported for Experiment II are based on percent-correct data transformed into RAUs. Data from Experiment II were analyzed using a 5 ×2 repeated-measure ANOVA with two within-subject factors [type of masker condition (spectral energy based on English, English-High Intelligibility, English-Moderate Intelligibility, English-Low Intelligibility, and Mandarin) and shape of masker condition (temporally modulated vs. steady state)]. Results indicated a significant interaction between type of masker and shape of masker [$F(4,88)=2.86$, $p=0.028$] and a significant main effect of the type of masker [$F(4,88)=86.25$, $p<0.0001$] and shape of masker [$F(1,22)=25.16$, $p<0.0001$; see Fig. 3]. Posthoc 2-way ANOVAs (using a Bonferroni adjusted alpha level) between all combinations of masker pairs indicated that the significant interaction between type of masker and shape of masker was driven by the English matched and the English-moderate intelligibility matched noise masker conditions [$F(1,22)=12.88$, $p=0.002$]. Pairwise posthoc t-tests (using a Bonferroni adjustment) comparing performance scores between modulated and steady-state maskers within each masker type indicated that performance in the modulated noise masker conditions were significantly better for the English and the English-Moderate Intelligibility matched maskers. Though average performance for the modulated noise maskers for the three remaining noise maskers was slightly higher compared to the performance of their respective steady-state maskers it did not reach significance for the English-High Intelligibility, English-Low Intelligibility and the native-Mandarin matched masker conditions (see bottom of Fig. 4).

Additional pairwise comparisons indicated the same pattern of results for the five temporally modulated and five steady-state matched noise maskers. That is, we observed no significant difference in performance between the English and Mandarin masker conditions or the English-High Intelligibility and English-Moderate Intelligibility masker conditions. Average performance scores between all other combinations of masker conditions were significantly different from each other.

In summary, performance for the spectrally matched maskers for the native English and native-Mandarin type maskers was significantly worse in comparison to the other masker conditions. This pattern of results is similar to the pattern of results we observed in Experiment I for the easier (−3 dB) SNR condition only. Recall, however, that all noise masker conditions (Experiment II) were conducted at an SNR of −5 dB, and performance for all masker conditions in Experiment I were significantly different from each other at the −5 dB SNR. These results suggest that the absence of linguistic contributions of the maskers used in Experiment II had an impact on the pattern of performance results observed for these maskers.

If we compare the results from Experiment II to those from Experiment I (at the −5 dB SNR condition) we can see that the only 2-talker masker condition that caused a significant amount of informational masking was the native-English condition. That is, the noise masker that retained the spectral and temporal information [but did not include the linguistic information (the temporally modulated, spectrally shaped noise masker)] was significantly less effective in terms of masking than the 2-talker masker (see Top of Fig. 4). Also, the temporal modulations within the moderate-intelligible English masker condition were enough to improve listeners' sentence recognition (as observed by significantly poorer performance in the steady-state noise masker condition compared to the 2-talker and modulated-noise

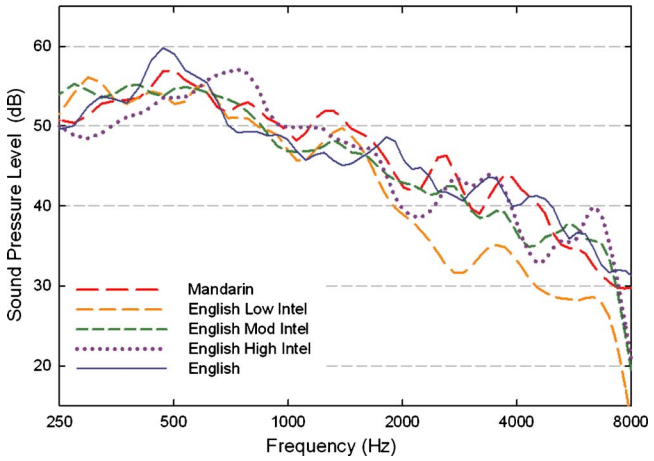

FIG. 5. (Color online) The LTASS of the five two-talker maskers used in Experiment I.

masker conditions). The linguistic contributions from all of the 2-talker Mandarin-accented English (high-to low-intelligibility) and the Mandarin masker conditions did not cause additional (informational) masking (as observed by the modulated noise masker being equally as difficult as the 2-talker masker condition).

Figure 5 shows the LTASS for the original five 2-talker maskers used in Experiments I. The low-intelligibility English masker condition provided significantly less spectral energy in the higher frequency range in comparison to the other maskers. Even though the intelligibility of the low-intelligible English masker was only ~44%, the Mandarin maskers should have provided no intelligibility for our listeners. Therefore, it is likely that the reduced spectral energy played a more significant role than the linguistic content itself in the ineffectiveness of the low-intelligible English masker (recall, that for all conditions tested in both Experiments I and II the low-intelligibility English masker condition was the easiest).

The temporal modulation spectra of each of the original 2-talker maskers were calculated as described in Gallun and Souza (2008) and plotted in Fig. 6. The native English

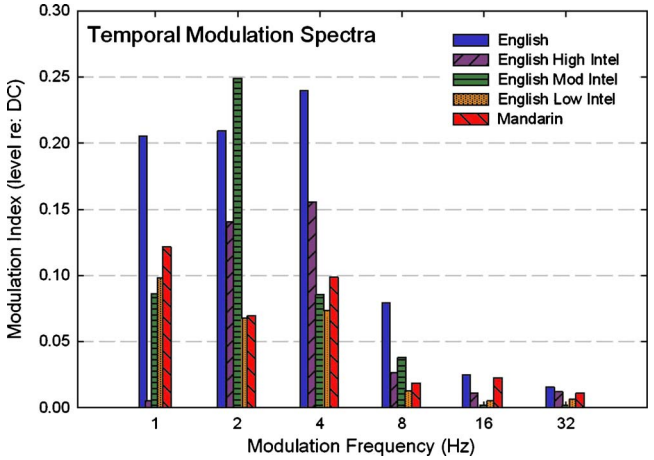

FIG. 6. (Color online) Modulation spectra relative to the energy at 0 Hz (see Gallun and Souza, 2008) analyzed for the five 2-talker masker used in Experiment I.

masker has noticeably greater temporal modulations compared to other masker conditions. These modulations were enough to improve listeners' recognition between the steady-state and temporally modulated noise maskers (see bottom Fig. 4), however, the benefit listeners gained from the temporal modulations were negated by the information provided by the competing native-English talkers (as illustrated by no significant difference in performance between the English 2-talker and steady-state spectrally matched noise maskers; see top Fig. 4).

IV. EXPERIMENT III: TEMPORALLY MODULATED WHITE-NOISE MASKERS

Due to the similar pattern of results between the *temporally modulated*- and the *steady-state*-spectrally matched noise maskers examined in Experiment II, it was difficult to interpret whether the results from the *temporally modulated* noise maskers were driven by the spectral information alone *or* a combination of the spectral and temporal information of these maskers. In addition, it was unclear how the differences between the temporal modulations of the five maskers (shown in Fig. 6) impacted listeners' ability to understand the target. To examine whether the *temporal modulations* of the original 2-talker maskers influenced listeners' performance *without* the influence of the spectral information of each masker, a third experiment was conducted. It was hypothesized that if the performance in Experiments I and II was influenced by temporal differences between the maskers, then a similar pattern of results should emerge when listening in the presence of white-noise temporally modulated to match the original five 2-talker maskers used in Experiment I.

A. Methods

In Experiment III listeners' sentence recognition performance was tested in the presence of white-noise temporally modulated to match the original five 2-talker maskers used in Experiment I. Twenty listeners, who also participated in both Experiments I and II, returned to the laboratory for a third visit and were presented an additional 120 sentences from the Harvard/IEEE sentence (not used in either Experiments I and II) lists spoken by the same male talker, recorded using the same procedures as reported in Experiments I and II. Twenty sentences/condition were presented at a fixed SNR of −5 dB across five masker conditions including five white-noise masker conditions temporally modulated to match the temporal envelopes from the 2-talker masker conditions used in Experiment I. The envelopes of the 2-talker maskers were extracted using the techniques described in Experiment II and white noise generated in MATLAB was than multiplied by each of the five temporal envelopes.

B. Results

Results from Experiment III (see Fig. 7) indicated that average performance for the Mandarin shaped white-noise masker condition was significantly poorer in comparison to the other four modulated white noise maskers (paired t-test results indicating $t(1,19)$ ranging between 5.84–8.02, p

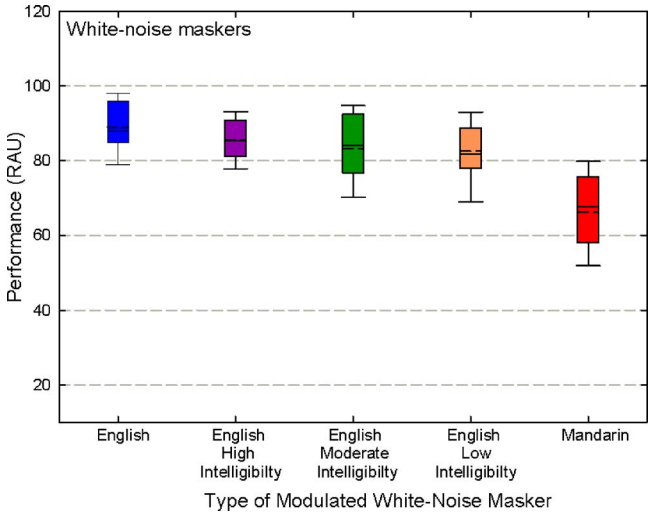

FIG. 7. (Color online) Sentence recognition results for 20 native-English speaking normal-hearing listeners. Performance in RAU is shown for five white-noise masker conditions presented at an SNR of −5 dB. Dashed lines within each box plot represent mean data, while solid lines represent the median.

<0.0001). There were no significant differences in performance between the other modulated white-noise maskers using an adjusted critical value based on a Bonferroni correction for multiple pairwise comparisons.

These data indicate that some of the difficulty listeners had when listening in the presence of the Mandarin speech masker were driven by the temporal properties of the 2-talker Mandarin babble. Also, these data suggest that the results observed in Experiment II for the temporally modulated, spectrally matched English and Mandarin-accented English noise maskers where driven by the spectral information of the maskers, not the temporal differences. These results are in agreement with the temporal modulation analysis (reported in Fig. 6) and the performance results from Experiments I and II (reported in the top of Fig. 4) that indicated that listeners were able to take advantage of the temporal modulations within the native English masker condition.

V. DISCUSSION

A. General Discussion

The results from these three Experiments portray the complexity of interpreting performance scores for speech-in-speech recognition tasks. Based on the results of these Experiments, the release in masking that has been observed for monolingual listeners when the competing speech is spoken in an unfamiliar language compared to speech matched to the language of the target speech is not due solely to the presence, or absence, of linguistic interference. In fact, linguistic, spectral and temporal differences between the target and masker speech, and the overall difficulty of the listening task, all influenced the effectiveness of the speech maskers used in Experiment I.

It is often assumed that we stress a system (in our case the auditory system) more when forcing the system to perform a more difficult task. In Experiment I listeners were asked to complete the speech-in-speech recognition task at two different SNRs. The significant interaction between SNR and masker language indicated that as the task became more difficult for the listeners, and the auditory system was therefore more taxed, the effect of linguistic interference became more influential. That is, it was only at the more difficult SNR that a significant difference in performance between the five maskers was observed. At the more difficult SNR for the native-English and the non-native English speech maskers, performance significantly decreased as the listeners' access to lexical information of the masker speech increased. However, the results for the performance scores of the Mandarin masker condition question whether the significant differences in performance between the native-English and non-native English maskers are entirely linguistically driven. Out of all of the masker conditions, native English listeners should have the least lexical access to the Mandarin speech, even when compared to the least intelligible non-native English masker. Therefore, if we conclude that as lexical access to the speech masker increased (e.g., from low-intelligibility accented English to native-English speech), speech recognition performance decreased, then the Mandarin masker condition should have resulted in the highest performance compared to all of the English (both native and non-native accented) maskers. However, this was not the case, and indeed the competing Mandarin speech proved to be more difficult for listeners compared to any of the non-native English maskers.

A closer look at the spectral and temporal properties of the speech maskers examined in Experiments II and III reveals that the spectral properties of the speech maskers may be driving the results observed in the 2-talker data (especially for the easier SNR condition). Interestingly, it appears that when the task is less difficult for listeners, the pattern of results for the 2-talker masker is identical to the pattern of results for the spectrally matched noise maskers (those used in Experiment II and stripped of linguistic content). That is, there are no significant differences between the English and the Mandarin masker conditions, nor are there differences in performance between the high and moderate intelligibility masker conditions. However, when we make the task more difficult (during the −5 dB SNR condition), the performance scores for the 2-talker masker conditions begin to separate, showing significantly decreased performance for those masker conditions to which the listeners would have greater lexical access (e.g., English vs. Mandarin and High vs. Moderate intelligibility). These results indicate that the lexical information of the masker appears to cause greater detriment when the entire auditory system is more stressed, or is forced to perform a more difficult task. This statement is most evident when we observe the significant interaction between language of the masker and SNR in Experiment I, which was only found to be significant for the masker condition with the greatest amount of lexical information (native English). In summary, it appears that spectral differences between masker signals for a speech-in-speech recognition task using two-talker maskers drive differences in performance when the auditory task at hand is not too difficult. However, as the task is increased in difficulty (e.g., with increasing difficulty of

SNR) "informational" factors of the masker signal begin to become more problematic, causing an additional layer of complexity for the auditory system. In addition, for some competing speech (in these experiments the Mandarin maskers) temporal modulations cannot be discounted when determining masker effectiveness.

Freyman *et al.* (2001) reported a similar release in masking when listeners attended to English speech in the presence of two competing Dutch talkers compared to two English talkers. The spectra of the two maskers used in their study were very closely matched (the Dutch and English maskers were spoken by the same two female talkers). Though details about the specific statistically significant differences across their SNR conditions were not reported, the visual representation of their data shows a significant release in masking for the −4 dB SNR condition they tested, but not for the easier SNR condition of 0 dB. Thus, their data appear to be in agreement with our finding that performance in the presence of an English masker decreases disproportionately compared to the foreign language maskers as SNR increases. One main difference to note between our data and those reported by Freyman *et al.* (2001) is the typological difference between the competing masker languages used. Dutch is very close typologically to English; both are descendents of the Germanic language and have a number of similar phonetic features. Mandarin, on the other hand, is typologically very distant from English, most notably with respect to it having a system of lexical tones. Future experiments are warranted to assess the effect of typological distance between target and masker on this reported disproportionate decrease in performance as SNR increases.

B. "Informational" Masking

An interesting finding from the data reported in Experiments I and II is that we only observed significant "informational" masking for the native-English masker condition. That is, performance only significantly decreased between the 2-talker and the modulated spectrally shaped noise maskers for the native-English condition (both tested at the −5 dB SNR). The mean difference between the paired comparison was 17.6 RAU, indicating that it was, on average, 17.6 RAU more difficult for listeners when the information from the linguistic content of the speech was included in the masker. None of the other comparisons of 2-talker and their respective *modulated* spectrally matched noise maskers reached significance.

The following findings make the lack of "informational" masking between the other four maskers interesting. First, the high-intelligibility Mandarin-accented English masker provided ~80% intelligibility. Yet, performance results in the presence of this masker compared to its temporally modulated spectrally shaped counterpart did not provide significant "informational" masking. Therefore, it is plausible to conclude that the competing speech signal needs to be greater than 80% intelligible to cause significant "informational" masking. This being said, a second observation from the more difficult SNR condition in Experiment I also indicated that the *high*-intelligibility masker caused performance to significantly decrease compared to the *moderately* intelligible English masker condition. Yet, when linguistic content was removed from these two maskers they did not result in significant differences in performance based on either their spectral or temporal properties (as reported in Experiments II and III). Therefore, this result suggests that even when significant "informational" additions of masking cannot be obviously measured [i.e., as depicted in (top) Fig. 4, no measurable differences in performance caused by a given speech masker compared to its respective temporally modulated spectrally shaped noise masker], linguistic contributions can still affect (decrease) performance, perhaps causing greater confusion for the listener.

These results may also indicate that once sufficient differences between the target and masker speech exist (e.g., the competing speech becoming more and more accented), listeners can benefit from a release in masking. Therefore, it may be feasible to suggest that if there was a way to add distortion (as we did in these experiments by increasing the accentedness of the speech masker) to the competing signals we might be able to improve speech recognition. That is, it may be possible to think of improving listeners' recognition in noise, not simply by making the target signal more "clear" or enhanced with better signal-processing technology, but perhaps to address this problem from the opposite view point; leaving the target signal undisturbed and perturbing the competing signals so they sufficiently differ from the target potentially resulting in a release in masking.

C. Listening in the dips

Though there was a trend for all of the temporally modulated maskers tested in Experiment II to be less effective in terms of masking compared to their steady-state counterparts, only the English and moderate-intelligible English matched temporally modulated maskers provided a significant release in masking. These performance results are in agreement with the greater modulation depth observed for those two maskers at low-modulation frequencies (as seen in Fig. 6). Also, the results from Experiment III indicate a trend that the modulations within the native English masker condition provide the greatest release in masking (as observed in Fig. 7).

Previously it has been reported that normal-hearing listeners are able to benefit from the temporal modulations within competing speech signals until four or more talkers are competing in the background (see Miller, 1947; Simpson and Cooke, 2005). That is, we would have predicted based on previous literature that the temporally modulated spectrally matched noise maskers used in Experiment II would have provided a significant release in masking (due to the temporal modulations) compared to their respective steady-state maskers. This release in masking only occurred for the English and moderate-intelligible English matched masker conditions, but not for the high- or the low-intelligible English, or the Mandarin matched maskers. There is a potential that the talkers used in this study had less low-frequency modulations compared to those typically reported in the lit-

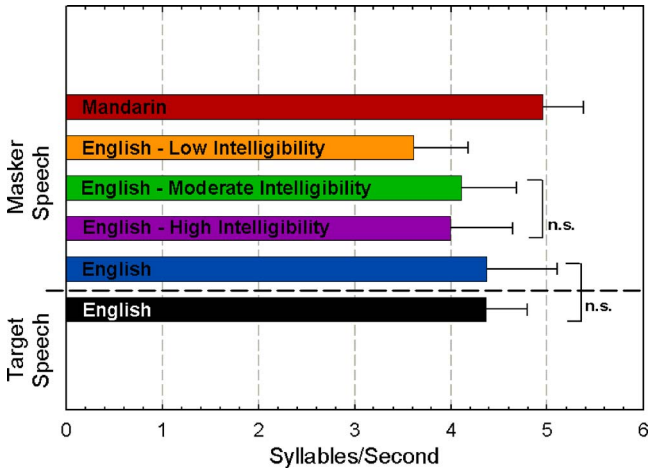

FIG. 8. (Color online) Average syllables/second for each of the five two-talker maskers used in Experiment I. The native-English masker speech was spoken using statistically the same speech rate as the target speech. All other two-talker maskers significantly differed from the rate of the target speech.

erature. However, it is also possible that accented English speech has less low-frequency modulation depth compared to native English speech.

D. Speaking rate

To assess how the rate of the masking speech might have affected listeners' performance, we conducted an analysis based on the expected syllables/second of all the speech stimuli used in these experiments. Recall that all speech maskers were created by concatenating individual sentences. The expected syllables/second were calculated for every sentence spoken by each talker by counting the number of expected syllables in each sentence and dividing that number by the duration of the sentence (in seconds). The two talkers used in each masker condition were included in the syllable/second masker calculations. Figure 8 demonstrates that the average rate of the English masker was not significantly different compared to average rate of the target speech. Also, the rate of the high-intelligibility and moderate-intelligibility English maskers were not significantly different from each other. All other rate comparisons between masker conditions reached significant differences. Therefore, it is plausible that having a similar speaking rate in the English masker condition added to the difficulty for listeners in this masking condition. That being said, Calandruccio *et al.* (2008) recently reported that sentence recognition performance was not significantly affected by various speaking rates in the masker speech (by comparing recognition performance in the presence of clear and conversational speech masker conditions). A future experiment in which speech recognition is assessed in the presence of speech maskers all spoken at different rates could help to probe this question further.

E. Learning effects

In a recent paper, Helfer and Freyman (2009) investigated performance differences between the first and last 25 trials of several speech-recognition experiments that varied in the amount of energetic and informational masking contributions. They hypothesized that even though listeners should indicate slight performance increases across trials once becoming familiar with the target talker's voice, listeners should indicate greater learning (or improvement across trials) for those conditions that provided more informational masking (or greater confusion between the target and masker signals). An analysis of their data indicated, however, the opposite. Specifically, those masker conditions less likely to cause informational masking resulted in greater learning across trials. It was hard to determine based on their data whether this result was truly due to differences between energetic and informational masking contributions, or simply differences in the difficulty of the task. To probe this idea further we analyzed our data based on the first half and the last half of keywords/condition [50 (non-repeating) keywords per half]. Repeated measure ANOVAs indicated that for both steady-state and temporally modulated spectrally shaped noise maskers [Experiment II; examining two within subject factors (Language of Masker × First and second half of trials)] no significant main effect of first and second half of trails was found [$F_{(1,22)}=.379$, $p=0.544$ and $F_{(1,22)}=0.665$, $p=0.445$, respectively], indicating no significant learning for our energetic masker conditions.

The same lack of a significant effect for the first and second half of the trials was found for a repeated measure ANOVA examining the 2-Talker data from Experiment I for the easier (−3 dB) SNR condition [$F_{(1,25)}=3.09$, $p=0.091$]. However, a significant main effect of the first and second half of trials was found for a repeated measure ANOVA examining the 2-Talker data from Experiment I at the more difficult (−5 dB) SNR condition [$F_{(1,25)}=16.49$, $p<0.0001$]. Therefore, this result suggests that for those masker conditions that have information contributions, the more difficult the task, the greater potential for improvement. This being said, no significant interaction was found between language of the masker and the first and second half of the trials [$F_{(4,100)}=0.606$, $p=0.659$] for either SNR. Though no significant interaction was found, there was a clear trend indicating greater learning for the English condition compared to the other four maskers. Our results imply that task difficulty impacts the degree of learning that is observed over the course of the experiment, and at least for our data, significant learning was only observed for those maskers that included informational contributions. The most difficult SNR tested in the Helfer and Freyman (2009) study was −4 dB (and their data also included SNRs of −1 and 2 dB), therefore, the discrepancy between the learning effects observed in our data and theirs may be due to task difficulty.

VI. CONCLUSIONS

Based on the data reported in these three Experiments, we can conclude that spectral differences between target and masker speech play a large role in determining the potential for a release-in-masking when comparing matched and non-matched target and masker language experiments. This being said, linguistic differences between target and masker speech cannot be discounted since they too appear to play a role in this release-in-masking especially when the auditory and

cognitive systems are required to perform a more difficult task. Therefore, the difficulty of the task (or listeners' performance level), and spectral and temporal properties of the speech should be accounted for when interpreting results using linguistically similar and linguistically different speech maskers.

ACKNOWLEDGMENTS

Supported by the Hugh Knowles Center for Hearing Research at Northwestern University and the New Investigator Grant from the American Academy of Audiology Foundation awarded to the first author. The authors are thankful to Nah Eun (NahNah) Kim and Christina Yuen for help with data collection, to Page Puccinini with help in reliability measurements, and especially Rebekah Abel and Chun-Liang Chan for providing fruitful discussion and assistance with software development. We are thankful to Dr. Pamela Souza for the assistance she provided with the temporal modulation analyses. Portions of these data were reported at the 157th Meeting of the Acoustical Society of America Meeting in Portland, OR. We also acknowledge grant support from the NIH (Grant No. R01-DC005794 from NIH-NIDCD).

Aaronson, N. L., Rakerd, B., and Hartmann, W. M. (**2008**). "Release from speech-on-speech masking in a front-and-back geometry," J. Acoust. Soc. Am. **125**, 1636–1648.

American National Standards Institute (ANSI) (**2004**). American national standard specifications for audiometers (ANSI S3.6-2004), ANSI, New York.

American Speech-Language-Hearing Association (ASHA) (**2005**). Guidelines for Manual Pure-Tone Threshold Audiometry, ASHA, Rockville, MD.

Bench, J., Kowal, A., and Bamford, J. (**1979**). "The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children," Br. J. Audiol. **13**, 108–112.

Bent, T., and Bradlow, A. R. (**2003**). "The interlanguage speech intelligibility benefit," J. Acoust. Soc. Am. **114**, 1600–1610.

Boersma, P., and Weenink, D. (**2009**). "Praat: Doing phonetics by computer (Version 5.1.07) [Computer program]," http://www.praat.org/ (Last viewed 5/12/2009).

Calandruccio, L., Yuen, C., Van Engen, K., Dhar, S., and Bradlow, A. (**2008**). "Assessing the clear speech benefit with competing speech maskers," Am Sp-Lang-Hear Ass National Meeting, Chicago, IL.

Carhart, R., Tillman, T. W., and Greetis, E. S. (**1969**). "Perceptual masking in multiple sound backgrounds," J. Acoust. Soc. Am. **45**, 694–703.

Carhart, R., Tillman, T. W., and Johnson, K. R. (**1967**). "Release of masking for speech through interaural time delay," J. Acoust. Soc. Am. **42**, 124–138.

Cooke, M., Garcia Lecumberri, M. L., and Barker, J. (**2008**). "The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception," J. Acoust. Soc. Am. **123**, 414–427.

Durlach, N. I., Mason, C. R., Kidd, G., Jr., Arbogast, T. L., Colburn, H. S., and Shinn-Cunningham, B. G. (**2003**). "Note on informational masking," J. Acoust. Soc. Am. **113**, 2984–2987.

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

Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2001**). "Spatial release from informational masking in speech recognition," J. Acoust. Soc. Am. **109**, 2112–2122.

Freyman, R. L., Helfer, K. S., and Balakrishnan, U. (**2007**). "Variability and uncertainty in masking by competing speech," J. Acoust. Soc. Am. **121**, 1040–1046.

Freyman, R. L., Helfer, K. S., McCall, D. D., and Clifton, R. K. (**1999**). "The role of perceived spatial separation in the unmasking of speech," J. Acoust. Soc. Am. **106**, 3578–3588.

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

Garcia Lecumberri, M. L., and Cooke, M. (**2006**). "Effect of masker type on native and non-native consonant perception in noise," J. Acoust. Soc. Am. **119**, 2445–2454.

Gelfand, S. A., Piper, N., and Silman, S. (**1986**). "Consonant recognition in quiet and in noise with aging among normal hearing listeners," J. Acoust. Soc. Am. **80**, 1589–1598.

Helfer, K. S., and Freyman, R. L. (**2008**). "Aging and speech-on-speech masking," Ear Hear. **29**, 87–98.

Helfer, K. S., and Freyman, R. L. (**2009**). "Lexical and indexical cues in masking by competing speech," J. Acoust. Soc. Am. **125**, 447–456.

Hornsby, B. W., Ricketts, T. A., and Johnson, E. E. (**2006**). "The effects of speech and speechlike maskers on unaided and aided speech recognition in persons with hearing loss," J. Am. Acad. Audiol **17**, 432–447.

IEEE Subcommittee on Subjective Measurements IEEE Recommended Practices for Speech Quality Measurements (**1969**). IEEE Trans. Audio Electroacoust. **17**, 227–46.

Jerger, J. (**1992**). "Can age-related decline in speech understanding be explained by peripheral hearing loss?," J. Am. Acad. Audiol **3**, 33–38.

Kidd, G., Jr., Mason, C. R., Deliwala, P. S., Woods, W. S., and Colburn, H. S. (**1994**). "Reducing informational masking by sound segregation," J. Acoust. Soc. Am. **95**, 3475–3480.

Lutfi, R. A., Kistler, D. J., Oh, E. L., Wightman, F. L., and Callahan, M. R. (**2003**). "One factor underlies individual differences in auditory informational masking within and across age groups," Percept. Psychophys. **65**, 396–406.

Miller, G. A. (**1947**). "The masking of speech," Psychol. Bull. **44**, 105–129.

Rajan, R., and Cainer, K. E. (**2008**). "Ageing without hearing loss or cognitive impairment causes a decrease in speech intelligibility only in informational maskers," Neuroscience **154**, 784–795.

Simpson, S., and Cooke, M. (**2005**). "Consonant identification in *N*-talker babble is a nonmonotonic function of *N* (L)," J. Acoust. Soc. Am. **118**, 2775–2778.

Studebaker, G. A. (**1985**). "A "rationalized" arcsine transform," J. Speech Hear. Res. **28**, 455–462.

Van Engen, K. J., and Bradlow, A. R. (**2007**). "Sentence recognition in native- and foreign-language multi-talker background noise," J. Acoust. Soc. Am. **121**, 519–526.

Watson, C. S. (**2005**). "Some comments on informational masking," Acta. Acust. Acust. **91**, 502–512.