JSLHR

Article

Cross-Language Differences in Informational Masking of Speech by Speech: English Versus Mandarin Chinese

Xihong Wu,[a] Zhigang Yang,[a] Ying Huang,[a] Jing Chen,[a] Liang Li,[a] Meredyth Daneman,[b] and Bruce A. Schneider[b]

Purpose: The purpose of the study was to determine why perceived spatial separation provides a greater release from informational masking in Chinese than English when target sentences in each of the languages are masked by other talkers speaking the same language.
Method: Monolingual speakers of English and Mandarin Chinese listened to semantically anomalous sentences in their own language when 1 of 3 maskers was present (speech-spectrum noise, a 2-talker speech masker in the same language, and a 2-talker speech masker in the other language).
Results: Both groups benefitted equally from spatial separation when the maskers were speech-spectrum noise or cross-language. Chinese listeners benefitted less from spatial separation than did English listeners when a same-language masker was used. Performance was scored in terms of the number of target words correctly identified; because Chinese target words were composed of 2 "stand-alone" morphemes, the authors also scored Chinese target words as correct when either of the morphemes was correctly identified. When this was done, Chinese and English listeners benefitted equally from spatial separation in all conditions.
Conclusion: These results support a model in which release from informational masking in both monolingual English and Chinese listeners occurs because spatial separation facilitates morpheme access in both languages.

Key Words: speech comprehension, speech perception, informational masking, energetic masking, lexical access, English versus Chinese languages, working memory

Listeners often complain that they have more trouble understanding what someone is saying when there are other people talking than when there are competing sound sources of a nonlinguistic nature (such as air-conditioning noise). Why is this the case? Both nonlinguistic sound sources and competing speech can interfere with the processing of the speech target at a peripheral level (by eliciting basilar membrane activity in the same or nearby regions as those elicited by the target speech). This kind of peripheral interference is often referred to as "peripheral" or "energetic" masking (Brungart, 2001; Freyman, Helfer, McCall, & Clifton, 1999). Competing speech, in addition to contributing to energetic masking, could also be interfering with the processing of the target talker's utterances at more central (i.e., cognitive) levels of processing. To comprehend speech, listeners not only have to process and recognize the basic building blocks of speech (the phonemes of the language), they also have to use these units to access the meaning of individual morphemes, words, and phrases, in order to extract meaning from the target talker's utterances. Because competing steady-state noise sources are unlikely to interfere with speech comprehension at phonemic or more central levels, any decrements in perception or comprehension in the presence of such sources can be attributed to energetic masking. Competing speech, on the other hand, in addition to being an energetic masker, is likely to elicit activity in one or more of the processes leading to the extraction of the utterance's meaning. This kind of nonenergetic

[a]National Key Laboratory on Machine Perception, Speech and Hearing Research Centre, Peking University, Beijing, China
[b]University of Toronto Mississauga, Ontario, Canada
Correspondence to Bruce A. Schneider: bruce.schneider@utoronto.ca

Editor: Anne Smith
Associate Editor: Alex Francis
Received October 9, 2010
Accepted April 25, 2011
DOI: 10.1044/1092-4388(2011/10-0282)

interference of competing speech on speech is often referred to as "informational" masking of speech by speech (for recent reviews of informational masking of speech by competing speech, see Schneider, Li, & Daneman, 2007; Schneider, Pichora-Fuller, & Daneman, 2010).

Release From Informational Masking of Speech by Speech in English and Chinese

To effectively process auditory information from a sound source, one must first segregate the target stream from other competing sound sources (Bregman, 1990). Successfully parsing the auditory scene into its component sound sources allows listeners to focus their attention on the target talker and ignore or suppress the processing of information from other sources. A number of auditory and cognitive factors have been shown to facilitate segregation of the target speech from competing speech, including spatial separation (e.g., Freyman et al., 1999; Humes, Lee, & Coughlin, 2006; Li, Daneman, Qi, & Schneider, 2004; Wu et al., 2005), differences in fundamental frequency (Summerfield & Assmann, 1991; Summers & Leek, 1998; Vongpaisal & Pichora-Fuller, 2007), and prior knowledge of part of the target talker's message (Freyman, Balakrishnan, & Helfer, 2004; Yang et al., 2007). In the present study, we employed one of these factors (perceived spatial separation) to investigate why the release from masking is larger in English (e.g., Freyman et al., 1999; Li et al., 2004) than in Chinese (Wu et al., 2005).

Competing speech in one's own native language can interfere with the processing of the target speech because of acoustic similarities between the target and competing speech, and/or because the competing speech interferes with the processing of phonemes, morphemes, or with postmorphemic processes such as the linkage of morphemes in word formation. Hence, it is possible that structural differences between English and Chinese might lead to differential amounts of release from masking at one or more of these levels. In the present study, we present evidence that spatial separation facilitates lexical access to morphemes to the same degree in both languages. However, faciliting morpheme access leads to a greater release from masking in English than in Chinese because of language-based differences in the ways in which morphemes are linked together to form words. To see why this is the case, we will begin with a brief description of the experimental procedure, and then indicate how structural differences in the ways in which words are formed in English and Chinese could result in differential amounts of release from masking in the two languages when the masker is competing speech in one's own native language.

Using Perceived Spatial Separation to Study Release From Informational Masking

A number of studies have shown that spatially separating the target speech from the distracting speech in English listeners can lead to a release from masking on the order of 4–9 dB (e.g., Freyman et al., 1999; Li et al., 2004). The basic paradigm consists of two conditions. In the colocation condition, the target speech (in this case, a semantically anomalous sentence such as "A shop could frame a dog") and the masker are presented over the same loudspeaker with the target sentence starting approximately 1 s after masker onset. The listener is instructed to listen to and repeat the target sentence. This condition is contrasted with one in which the target speech is presented over one loudspeaker, and the masker is presented over a second, spatially separate loudspeaker. The target speech in each condition is presented at several levels of signal-to-noise ratio (SNR) to determine a psychometric function relating percent correct to SNR in each condition. The 50% point on this psychometric function is determined for each of these two conditions, and the degree of release from masking in dB is obtained by subtracting the SNR corresponding to the 50% threshold in the spatially separated condition from the 50% threshold SNR in the spatially colocated condition. Typically two types of maskers are employed: a speech-spectrum noise masker, and a masker consisting of two other people saying the same type of semantically anomalous sentences as the target talker. Because the steady-state speech-spectrum noise masker is unlikely to elicit any activity in the language processing systems, the release from a speech-spectrum noise masker due to spatial separation primarily reflects a release from energetic masking. In contrast, the amount of release from a speech masker due to spatial separation reflects a release from both energetic and informational masking.

In the present study, rather than using actual spatial separation, we used the precedence effect (Zurek, 1980) to achieve a perceived spatial separation between the masker and the target. The c nc $c\acute{}$ is based on the listener's ability to perceptually fuse the direct wave front from a source with its myriad reflections off environmental surfaces. Consider, for the moment, a simplified sound field consisting of a sound source directly to the left of a listener with a single sound reflecting barrier directly to the listener's right. The direct wave from the sound source will be the first wave front to reach the listener's head. This wave front will pass around the head, encounter the sound-reflecting barrier, and be reflected back toward the listener. In effect, the listener receives the direct wave from the sound source located on the listener's left, and then a few milliseconds

later, a second wave front (a filtered and time-delayed version of the direct wave front) coming from the listener's right. Provided that the delay is not too long (under 6–7 ms), the listener will perceptually fuse the information from the two sound waves and perceive a single source located to the left of the listener (Clifton & Freyman, 1989; Li, Qi, Yu, Alain, & Schneider, 2005; Shinn-Cunningham, Zurek, & Durlach, 1993). We used this effect in the laboratory to change the perceived location of the targets and maskers in the following way.

The listener is seated in the center of a soundproof chamber with two loudspeakers, one to the listener's left and the other to the listener's right. To achieve the perception of a single sound source to the listener's right, the same sound is played over both loudspeakers with the sound coming from the left loudspeaker lagging the sound coming from the right loudspeaker by 3 ms. Under these conditions, the listener perceives the sound to be located to his or her right. If, on the other hand, the same sound is played over both loudspeakers with the right loudspeaker sound lagging the left loudspeaker sound by 3 ms, the sound is perceived to be coming from the left. To achieve the perception that both the masking sound and the target sentences are colocated on the right, both are played over the two loudspeakers with the masker and the target played over the left loudspeaker lagging their counterparts played over the right loudspeaker by 3 ms. To achieve the perception that masker and target originate from different sources, both sounds are played over the two loudspeakers with the target played over the left loudspeaker lagging its counterpart by 3 ms, while the masker presented on the right lags the masker presented on the left by 3 ms. Under these conditions, the listener perceives the target sentences as located on the right with the masker perceived as originating from the listener's left. Perceived spatial separation, rather than physical separation, was employed to facilitate comparisons between experiments conducted at Peking University on non-English-speaking participants whose first language was Mandarin Chinese and experiments conducted at the University of Toronto on non-Chinese-speaking participants whose first language was English. An advantage of using perceived rather than physical separation in cross-linguistic research is that previous studies with English listeners have shown that the degree of release from informational masking with perceived spatial separation is relatively independent of the acoustic environments in which testing takes place (Freyman et al., 1999; Li et al., 2004; Marrone, Mason, & Kidd, 2008). This allowed us to test monolingual Chinese and English listeners in their home countries without having to provide identical acoustic environments.

How Competing Sound Sources Might Interfere With Speech Recognition

To aid in determining the processing level or levels responsible for differential amounts of release from masking of speech by speech in Chinese and English, three different maskers were employed with monolingual English-speaking and Chinese-speaking listeners. The first was a steady-state noise whose spectrum matched that of Chinese speech for the Chinese listeners and English speech for the English listeners. The second was a cross-language masker: anomalous English sentences for Chinese listeners, anomalous Chinese sentences for English listeners. The reason for employing a cross-language masker was to assess the degree to which a semantically meaningless speech masker could interfere with target-word recognition. Because maskers are similar in many ways to the listener's native language, we would expect these similarities to give rise to some degree of informational masking. For instance, similarities in fundamental frequency, cadence, and/or phonetic structure could interfere with speech recognition (Calandruccio, Dhar, & Bradlow, 2010; Cooke, Garcia Lecumberri, & Barker, 2008; Rhebergen, Versfeld, & Dreschler, 2005; Van Engen & Bradlow, 2007). Hence, the cross-language masker provides a baseline against which to assess the degree of interference produced when the words are meaningful to the native listener.

A same-language masker was employed to determine the extent to which a semantically meaningful masker interferes with access to the morphemes in a language, and/or with post-morphemic processes such as those that link morphemes together to form words. There is some controversy in the literature as to whether the processes involved in access to the meaning of words or phrases differ between Chinese and English. In all natural languages, including Chinese (see Packard, 2004), phrases and sentences are constructed via syntactic rules that string together words retrieved from a mental lexicon. Chinese differs from English in that most words are multimorphemic, typically compounds consisting of two morphemes (Packard, 1999; Zhang & Peng, 1992). This has led to some discussion in the literature as to whether it is the morpheme or the word or both that is represented in the lexicon. In cohort models of lexical access (e.g., Marslen-Wilson, 1989), it is assumed that the auditory input associated with speech activates most if not all of the cohort of words that are possible given the auditory input up until that point of time. As the speech signal continues to unfold, more and more of the words in the cohort are "ruled out" until the listener recognizes the word and accesses its meaning. Hence, the word is accessed through the process of elimination of alternatives that, up until the point of time in question, were viable. It is here that

differences in the ways in which words are constructed could affect the degree to which English and Chinese are susceptible to informational masking.

In English, many words are monosyllabic, and of those that are multisyllablic, only a small proportion are composed of two or more morphemes that can also stand alone as words (e.g., the individual syllables in words like *an a* are not morphemes, whereas the two morphemes in words like *a a* can stand alone as words). In Chinese, however, most multisyllabic words are also multimorphemic (e.g., the word *ac* in Mandarin is *īn* , a two-morpheme word whose individual morphemes "*īn*" and " ," which mean "ice" and "river," respectively, are also stand-alone words in Mandarin). This raises the possibility that determining the meaning of a multimorphemic word in Chinese could be arrived at in two different ways (see Packard, 1999, 2004, for a thorough discussion of this issue). First, it is possible that the meaning of the word is accessed in the same way in Chinese as it is in English. For example, the cohort model of Marslen-Wilson (1989) could apply in the same way in both languages. According to the cohort model, as the speech signal unfolds, the auditory input begins to limit the number of words that are possible. At the beginning of the utterance the cohort of possible words is quite large. As the utterance unfolds, the auditory input directly narrows down the range of possible words until the word is recognized. It is important to note that if words are recognized through the sequential processing of phonemes,[1] the process would be the same for all words, independent of the number of morphemes in the word, or whether the morphemes were free (stand-alone) or not. On the other hand, there is reason to speculate, especially in difficult listening situations, that the recognition of a multimorphemic word may be mediated through the listener's recognition of one or more of its morphemes. Packard describes two ways such a process might work for the Chinese word for *i a n,* *ˇc ē*, which is composed of two morphemes ("fire" and "vehicle"). One way to access the word *ˇc ē* would be to first access the morpheme *ˇ* ("fire"), then access the morpheme *c ē*, and then to use "a word formation algorithm that combines the two morphemes in real time to form *ˇc ē*." (Packard, 1999, p. 91). Packard argues that such a model is both unwieldy and computationally costly, and he proposes a more limited model in which, after the identification of the morpheme *ˇ* ("fire"), the identified morpheme, rather than its phonemic components, is used to limit the cohort. He argues that this would function in much the same way as does a Chinese dictionary where entries are listed according to the initial morpheme. He notes that "on this view, once the morpheme *ˇ* is identified it should be relatively easy to identify the target word *ˇc ē*, because there are only about 104 words that begin with the morpheme *ˇ*" (Packard, 1999, pp. 91–92). Packard then goes on to dismiss both the first and second model as unlikely because if the listener parses the auditory input string into morphemes rather than words, there must be either a real-time morpheme-combination algorithm to construct words, or, as he puts it, "an improbable lexical retrieval mechanism in which morphemes are identified as the first step in accessing precompiled words" (Packard, 1999, p. 92). In other words, Packard argues that access to the lexicon does not differ between English and Chinese.

A dismissal on the basis that such models are "unwieldy" and "computationally costly" might be premature if one considers that much of everyday listening occurs in challenging listening environments where word access is typically less than perfect. In such environments, it is conceivable, for instance, that the Chinese listener identifies the morpheme *ˇ* but only part or none of the morpheme *c ē*. Under such circumstances, an algorithm that combines morphemes in real time might be useful in arriving at the meaning of a word even if it were computationally more costly. Hence the masking of speech by speech could reveal the possible role morphemes play in lexical access. In this study, we had monolingual English and Chinese participants listen to and repeat target anomalous sentences presented in (a) speech-spectrum noise, (b) a cross-language masker, and (c) a same-language masker. In all conditions maskers and targets were presented so that they appeared as either colocated or spatially separated. By comparing the relative improvement in performance due to spatial separation and type of masker in the two languages, we are able to show that informational masking does not interfere with processing phonemes or combining morphemes together to form words. However, competing speech does interfere with accessing individual morphemes.

Method

Participants

Twelve English-speaking participants (18–24 years old, mean age = 20 years, four men) were recruited from the student population at the University of Toronto

[1]As Miller and Eimas (1995) note, most models of lexical access assume that there is a stage in which the signal is represented in terms of a sequence of phonological features or phonetic features with the sequential unfolding of each sound leading to a reduction in the cohort of possible words. Marslen-Wilson (1989) argues that these features or units are no larger than a phonetic segment. In the TRACE model (Elman & McClelland, 1986; McClelland & Elman, 1986), there are three levels of representation: phonetic features, phonemes, and words, with the lowest level being phonetic features. Most models assume that the flow of information is bidirectional in order to accommodate a number of different experimental findings. However, in the absence of strongly biasing context, as is the case for the semantically anomalous sentences employed here, it is reasonable to assume that the flow of information is primarily bottom up.

Mississauga (UTM). The first language of all UTM participants was English, and all were raised in English-speaking countries. Audiometric tests indicated that their hearing was well within the normal range and balanced between the two ears (no more than 15 dB difference between the two ears). None of the UTM participants had any familiarity with Mandarin Chinese.

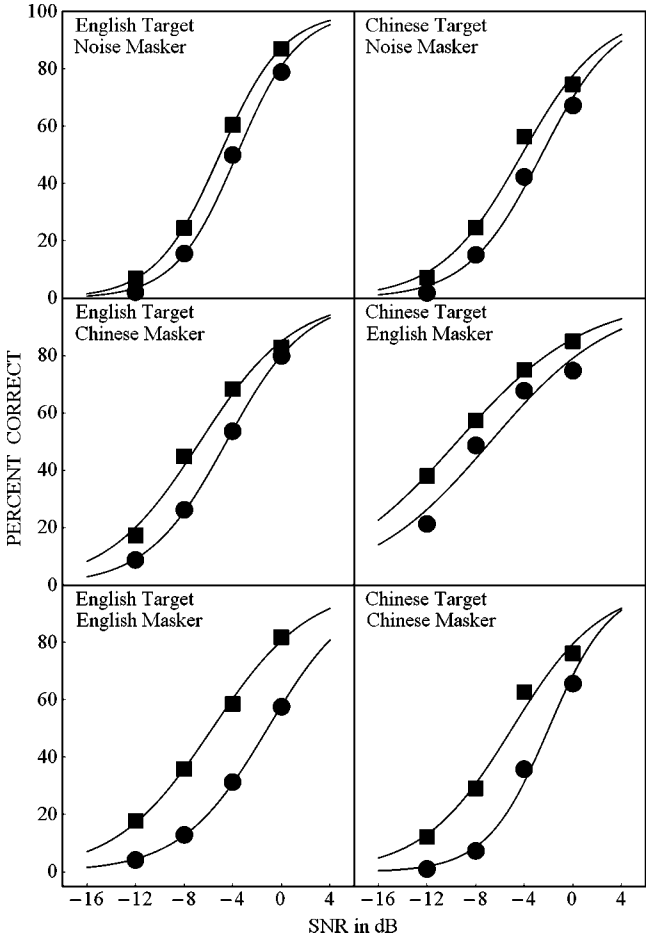
The 12 young Chinese participants (18–25 years old, mean age = 20 years, 12 men) were recruited from the campus police force at Peking University (PU). The first language of all of the PU participants was Mandarin Chinese, and all were raised in Mandarin-speaking areas of China. Like their UTM counterparts, audiometric tests indicated that their hearing was normal and balanced between the two ears. None of them could understand, speak, or read English.

One of the Chinese-speaking participants failed to

of the same type used for the speech target (see Freyman et al., 1999, for a more complete description of the talkers employed for English speech). Hence, the target talker was not the same as either of the two women who provided the masking speech. The Chinese masker also consisted of two women reading aloud a sequence of anomalous Chinese sentences of the same type as used for the speech target (see Wu et al., 2005), neither of whom was the Chinese target talker. For both English and Chinese two-talker maskers, the two talkers were recorded separately and then were mixed digitally.

Twenty-four lists (13 sentences per list) of English and anomalous Chinese sentences were used as targets in each language. Stimulus levels were calibrated by placing a microphone at the position in the booth that would correspond to the center of the average participant's head using Brüel & Kjaer (B & K) sound-level meters. During a session, the target sounds were presented at a level such that each loudspeaker, playing alone, would produce an average sound-pressure level of 57 dBA. This sound-

Figure 1. Percent-correct recognition of English target words by English-speaking listeners (left panels) and Chinese target words by Chinese-speaking listeners (right panels) as a function of signal-to-noise ratio (SNR; in dB) for three types of maskers: matched-language speech-spectrum noise (top panels), cross-language speech maskers (middle panels), and same-language speech maskers (bottom panels). The perceived location of all targets was to the right of the listener. Circles represent the condition in which the masker was perceived on the right; squares represent the condition in which the masker was perceived on the left.



corresponding to 50% correct (the threshold value). Figure 1 suggests that the beneficial effect of shifting the perceived location of either a noise masker (top panels) or cross-language masker (middle panels) away from that of the target is the same for English targets as it is for Chinese targets. However, Figure 1 also indicates that when the speech masker is the same language as that of the target sentences, a shift in the perceived location of the masker provides a greater release from masking for English target words than for Chinese target words.

There are also indications in Figure 1 that the psychometric functions of English listeners for both masker locations are shifted further to the left than those for Chinese listeners when the competing background is noise, whereas the opposite is true when the competing background is a cross-language masker. In other words, the thresholds of English listeners in noise are lower than those of their Chinese counterparts when the masker is masked-language speech-spectrum noise, but not when it is competing speech from a different language. In addition, when the target and masker were perceived as spatially separated, the slopes of the psychometric functions appear to be shallower than when target and masker appear to emanate from the same spatial location for both English and Chinese listeners under all masking conditions. Finally, Figure 1 suggests that slopes of the psychometric functions for both locations of the masker are steeper for English listeners than for Chinese listeners for noise and cross-language maskers, but that the opposite is true when the masker is of the same language as that of the target sentences.

To determine whether the effects found in the average data shown in Figure 1 also characterized the performance of individual participants, logistic psychometric functions were fit to individual data, and analyses of variance (ANOVAs) were conducted on the individually determined μ and σ values. A three-way ANOVA conducted on the 50% threshold values (μ), with target language (Chinese vs. English) as a between-subjects variable, and type of masker (noise, cross-language, same-language) and perceived spatial separation (target and masker colocated vs. target and masker perceptually separated) as within-subject variables revealed significant main effects of (a) perceived spatial separation, $F(1, 21) = 280.710$, $MSE = 0.843$, $< .001$, confirming that thresholds were lower when the maskers and target sentences were perceived as spatially separate as opposed to colocated, and (b) type of masker, $F(2, 42) = 85.691$, $MSE = 1.911$, $< .001$, confirming that thresholds depended on the type of masker. However, there was no main effect due to target language, $F(1, 21) = 1.713$, $MSE = 6.357$, $= .205$. There were also two significant two-way interactions: spatial separation and type of masker, $F(2, 42) = 15.354$, $MSE = 1.039$, $< .001$, with the release from masking due to spatial separation (release from masking = $\mu_{\text{masker,Right}} - \mu_{\text{masker,Left}}$) being larger for same-language maskers than either cross-language or noise maskers, and masker type and target language, $F(2, 42) = 20.086$, $MSE = 1.911$, $< .001$, with threshold being lower for English than for Chinese sentences in noise with the opposite being true when cross-language maskers were used. The third two-way interaction between perceived spatial separation and target language was not significant, $F(1, 21) < 1$. Finally, there was a significant three-way interaction among target language, perceived spatial separation, and type of masker, $F(2, 44) = 4.420$, $MSE = 1.039$, $= .018$.

To clarify the nature of the three-way interaction, we computed the release from masking due to spatial separation for the three kinds of maskers separately for English and Chinese target sentences. Figure 2 shows that for noise and for cross-language maskers, the amount of release from masking was approximately the same for English and Chinese target sentences. However, when the masker is from the same language, there appears to be a greater release from masking for English (4.7 dB) than there is for Chinese (3.0 dB) sentences, indicating the presence of an interaction. A two-way ANOVA conducted on release from masking with target language (Chinese vs. English) as a between-subjects variable and type of masker (noise, cross-language, same-language) as a within-subject variable revealed a significant main effect of masker type, $F(2, 42) = 31.892$, $MSE = 2.077$, $< .001$, and a significant interaction effect, $F(2, 42) = 4.420$, $MSE = 2.077$, $= .018$. The main effect of language was not statistically significant, $F(1, 21) = 0.73$, $MSE = 0.411$, $= .402$. To clarify the nature of the two-way interaction, one-way ANOVAs were conducted separately for Chinese and English targets. The effect of the type of masker was significant for both Chinese targets, $F(2, 20) = 3.723$, $MSE = 1.824$, $= .042$, and English targets, $F(2, 22) = 15.400$, $< .001$. Pairwise Newman–Keuls tests indicated for English targets that the amount of release from masking did not differ between noise and cross-language maskers ($> .05$) but that the amount of release from a same-language masker was significantly larger than for either a noise masker ($< .01$) or a cross-language masker ($< .01$). None of the pairwise comparisons among masker types were significant for Chinese maskers ($> .05$). Two-tailed t tests indicated that there were no significant differences in release from masking between Chinese and English targets when the maskers were either noise, $t(21) = -.483$, $= .634$, or cross-language, $t(21) = -1.176$, $= .252$, but that the amount of release from a same-language masker was greater for English targets than for Chinese targets, $t(21) = 2.34$, $= .029$.

To identify the source of the interaction between target language and type of masker, Figure 3 plots thresholds, averaged across perceived spatial position, as a function of masker type and target language. Figure 3 shows thresholds are lower for English than for Chinese target sentences for noise maskers, higher for English than for Chinese target sentences for cross-language maskers, and approximately the same for English and Chinese target sentences for same-language maskers. Two-tailed t tests indicated that the effect of the noise masker was greater for Chinese than for English targets, $t(21) = 2.398$, $= .026$, whereas the reverse was true for cross-language maskers, $t(21) = -5.246$, $< .001$. Therefore, the interaction between type of masker and the language of the target sentences is due to the fact that the effect of a noise masker was greater for Chinese targets than for English targets, whereas the effect of a

Figure 2. Release from masking in decibels for three types of maskers for English and Chinese target sentences. Standard error bars are shown.

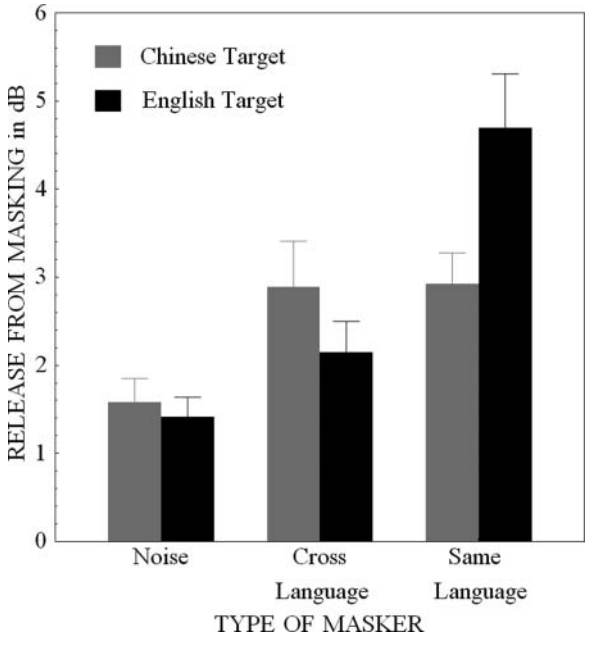
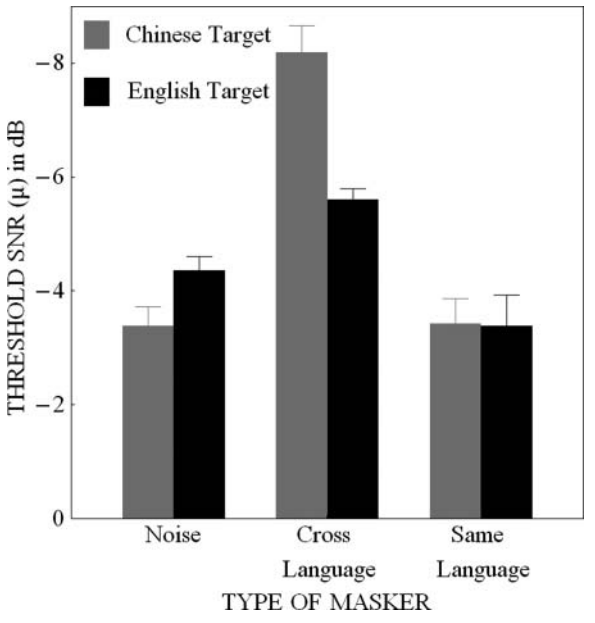


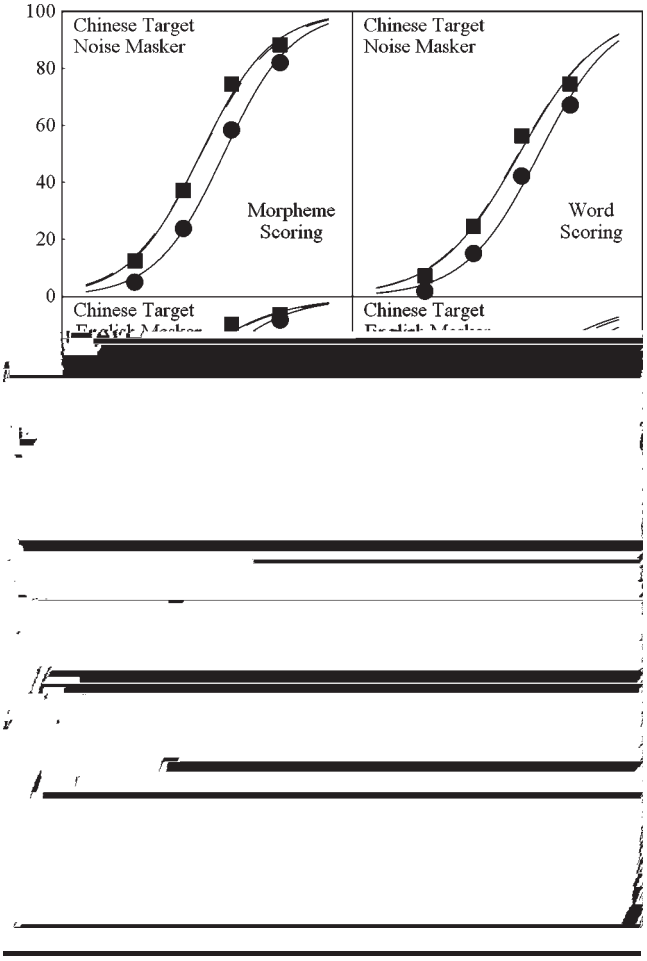
Figure 3. The average threshold (SNR corresponding to 50% correct) for the different types of maskers for the two target languages. Standard error bars are shown.



different-language masker was greater for English targets than for Chinese targets.

A three-way ANOVA on the slope parameter (σ) of the individual psychometric functions, with target language as a between-subjects variable and perceived spatial separation and type of masker as within-subject

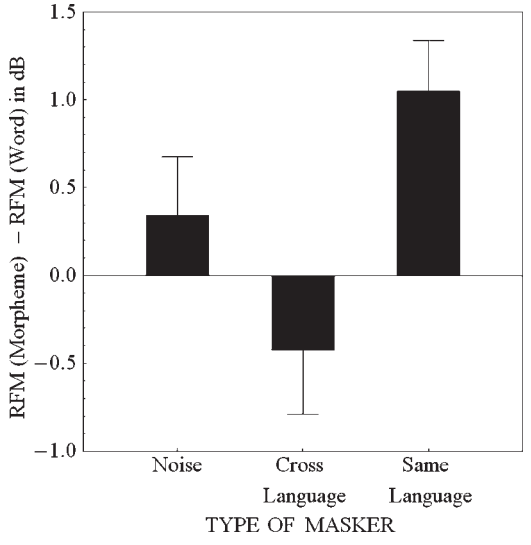
Figure 6. Percent-correct recognition of Chinese targets by Chinese-speaking listeners as a function of SNR (in dB) for three types of maskers: matched-language speech-spectrum noise (top panels), cross-language speech maskers (middle panels), and same-language speech maskers (bottom panels). The perceived location of all targets was to the right of the listener. Circles represent the condition in which the masker was perceived on the right; squares represent the condition in which the masker was perceived on the left. Morpheme-based scoring—that is, scoring a target word as correct if one or more of its morphemes were correctly identified—is presented in the left panels. Whole-word scoring is presented in the right panels. Dashed lines are the predicted psychometric functions based on a model in which spatial separation produced a release from masking on the morpheme level.



ANOVA on the slope difference scores indicated that the change in slope did not differ significantly with the type of masker, $F(2, 20) = 1.233$, $MSE = .003$, $= .313$, or with perceived spatial separation, $F(1, 10) < 1$, nor was there any interaction between type of masker and perceived spatial separation, $F(2, 20) < 1$. This allowed us to average slope differences across conditions to obtain an estimate of the mean slope difference between morpheme-level and whole-word scoring for each of the 11 participants. A two-tailed t test indicated that these slope differences were significantly larger than zero, $t(10) = 5.372$, $< .001$. Thus, a switch from whole-word to morpheme-level scoring resulted in a small but significant increase in slope (from $\sigma = 0.30$ to $\sigma = 0.35$, respectively), that did not vary significantly with either masking condition or perceived spatial separation.

Multiple t tests showed that there was a significant threshold shift to a lower value when morpheme-level scoring was used instead of whole-word scoring in all six conditions ($< .005$, Bonferroni corrected). Hence, in all cases morpheme-level scoring results in a lower threshold (50% point on the psychometric function) than whole-word scoring. Unlike the results for the slope parameter, the extent of the threshold shift did vary across conditions. A two-factor, within-subject ANOVA of the shift in threshold between morpheme-level and whole-word scoring, with type of masker as one factor and perceived spatial separation as the second factor, found a significant main effect for type of masker, $F(2, 20) = 6.198$, $MSE = 0.907$, $= .008$, but not for perceived spatial separation, $F(1, 10) = 3.549$, $MSE = 0.483$, $= .089$. However, there was a significant interaction between masker type and spatial separation, $F(2, 20) = 4.526$, $MSE = .659$, $= .024$. To determine the source of the interaction between masking type and spatial position, Figure 7 plots the amount of release from masking for morpheme-level scoring of Chinese target words minus the amount of release from masking for whole-word scoring of the same target words as a function of the type of masking. Figure 7 shows that the difference in the amount of release from masking between

Figure 7. Decibel difference in the amount of release from masking (RFM) observed for morpheme scoring of Chinese target words and that observed for word scoring of Chinese target words for three types of maskers. Standard error bars are shown.



morpheme-level and whole-word scoring is close to zero for noise masking and cross-language masking of Chinese target words. However, the release from masking is significantly larger for morpheme-level scoring than it is for whole-word scoring when the masker is of the same language. This was confirmed by a t test that found a significant difference only for Chinese masking of a Chinese target word, $t(10) = 3.60$, $p = .005$, two-tailed. Hence the amount of release from a noise masker or cross-language masker was the same, independent of whether the word needed to be correctly identified or only one or both of the two morphemes needed to be correctly identified in order for the target word to be scored as correct.

Figure 6 also suggests that for whole-word scoring there is a greater release from the same-language speech masker than from a noise masker, but that the release from a cross-language masker is approximately the same as that from the same-language masker. Two-tailed t tests confirmed that, when whole-word scoring was used, there was a greater release for same-language masking than for noise masking, $t(10) = 2.89$, $p = .016$, but that there was no significant difference between cross-language and same-language masking, $t(10) = 0.19$, $p = .85$. However when morpheme-level scoring was used, the amount of release from masking was not significantly greater for cross-language masking than for noise masking, $t(10) = 1.91$, $p = .085$, whereas the release from a same-language masker was significantly greater than the release from either a noise masker, $t(10) = 3.91$, $p = .003$, or a cross-language masker, $t(10) = 2.98$, $p = .013$. Note that for both whole-word and morpheme-level scoring, the amount of release from a same-language masker was greater than that observed for a noise masker, but the amount of release from

perceived locations of masker and targets could occur at the level of phonemic recognition, morphemic recognition, or at higher levels of semantic and linguistic processing. Specifically, spatial separation could facilitate the linkage of the two morphemes in the Chinese target words to arrive at the meaning of the target word when listening is difficult. In other words, it might be easier to link the two morphemes when the target and same-language masker were perceived to be spatially separated than when they were perceived to be colocated. To check this, we determined, at each SNR, the probability that Morpheme 2 was correctly identified given that Morpheme 1 was correctly identified for each type of masker and degree of spatial separation.

Discussion

The Effects of Language on the Threshold and Slope Parameters of the Psychometric Function

When the masker is matched-language speech-spectrum noise, Chinese listeners find it more difficult to hear the target words than do English listeners, confirming previous observations by Kang (1998) and Wu et al. (2005). There are a number of reasons why Chinese may be more easily masked by noise than English. Unlike English, in which there are relatively frequent occurrences of consonant clusters, Chinese syllables consist of a simple CVC or CV sequences. Hence, mishearing a single consonant is more likely to have a deleterious effect on speech recognition in Chinese than in English. In addition, Chinese has more voiceless consonants than English. Because voiceless consonants have less energy than voiced consonants, it is more likely that Chinese listeners will fail to hear or mishear more consonants than English listeners. Moreover, because Chinese is a tonal language, a failure to correctly apprehend the tonal contour would disrupt phoneme recognition. All of these factors might make Chinese more susceptible to energetic masking than English.

It is interesting to note that when the competing speech is uninterpretable to the listener, it is easier for a Chinese listener to segregate the Chinese talker from an English background than it is for an English listener to segregate an English talker from a Chinese background. Recall that research on English listeners (Summers & Leek, 1998; Vongpaisal & Pichora-Fuller, 2007) indicates that they can use differences in fundamental frequency to segregate out a single talker's voice from a multitalker background. Presumably, the listener does this by "tracking" the fundamental frequency of the target talker. However, given that Chinese is a tonal language, it might be difficult for English listeners to track an English target among the rapidly changing pitch glides provided by the competing Chinese speech. Hence, this cue to stream segregation, which could be a major support to English listeners in an English background, would not be as effective in a Chinese background.

On the other hand, streaming on the basis of fundamental frequency is unlikely to be productive for Chinese listeners. A more likely low-level cue for streaming is the staccato nature of the language. The rate and depth of modulation in the envelopes of Chinese speech is greater than it is in English speech (Yang et al., 2007). It might therefore be very easy for Chinese listeners to track the tempo and rhythm of Chinese speech against the more "monotone" background provided by the competing English speech.

The slope of the psychometric function in either noise or cross-language maskers is somewhat shallower for Chinese than for English listeners, with the reverse being true when the masker is from the same language. At present, we have not been able to generate any reasonable hypotheses as to why this should occur.

Language Effects on Release From Masking

Because a steady-state speech-spectrum noise masker is unlikely to produce any amount of informational masking, the degree of release from a steady-state speech masker can be used to estimate the degree to which spatial separation releases the listener from peripheral or energetic masking. The first result of interest, therefore, is that the amount of release from a speech-spectrum noise masker is the same in both languages, indicating that the degree of release provided by perceived spatial separation does not differ between English and Chinese when the masker is nonlinguistic.

We also noted that acoustic and phonetic similarites between a cross-language masker and the native listener's own language should produce some degree of informational masking. For example, because of phonemic overlap between the two languages, we might expect a cross-language masker to interfere with auditory processing at the phonemic level. (Because both the Chinese and English listeners could not understand each other's language, the interference is unlikely to be occuring at levels beyond the phonemic.) Hence, if phonemic similarity contributes to informational masking in this specific case of cross-language masking, we might have expected to see a greater amount of informational masking and a corresponding larger release from masking for cross-language maskers than for matched-language speech-spectrum noise maskers. The fact that we did not find a significantly higher release from a cross-language masker than from a matched-language speech-spectrum noise masker in either English or Chinese when whole-word scoring was used suggests that the amount of phonemic interference produced by the cross-language masker is minimal in this specific cross-language comparison. (It is possible that we might have observed more phonemic interference by a cross-language masker in situations in which there is a greater degree of phonemic similarity between the two languages as there is, for instance, between German and English.) Hence, it would appear that, in the present case, the majority of informational masking of speech by competing speech occurs at the semantic level in both languages.

When whole-word scoring was used, we also found that the amount of release from masking from a same-language speech masker was greater for English than

for Chinese listeners, confirming a previous report by Wu et al. (2005). There are several possible reasons for this. First, the degree of informational masking produced by a same-language masker might differ between the two languages. Therefore, it might be that there is less release from informational masking in Chinese than in English because there is less informational masking in Chinese than there is in English. Differences in the amount of informational masking in the two languages could arise from several sources. First, the amount of masking produced by a same-language masker could depend upon the acoustical properties of the language. In a previous paper (Yang et al., 2007), we noted that Chinese has more staccato than English, in the sense that the troughs in the speech envelope tend to be deeper and more prolonged in Chinese than in English. Thus, Chinese listeners would have more opportunities to "listen" during these troughs in the Chinese masking sentences than would English listeners during troughs in the English masking sentences. This would have the effect of reducing the amount of energetic masking for Chinese listeners.

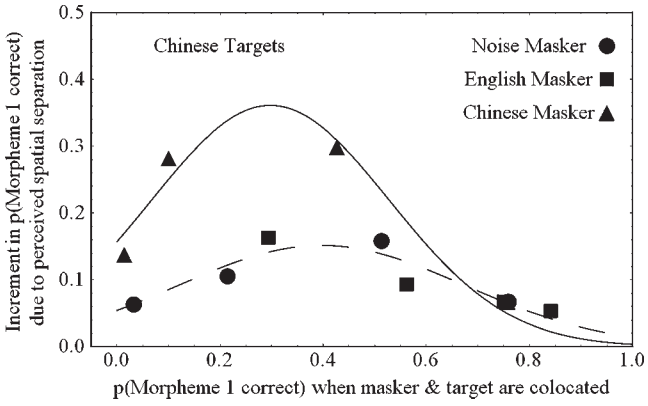
A second potential reason why the release from same-language masking might be less in Chinese than in English may be that auditory scene analysis is easier in English than in Chinese when there are multiple talkers from the same-language group. Presumably, perceived spatial separation results in a release from masking because it facilitates the segregation of the target stream from the other masking streams (see Schneider et al., 2007, for a discussion of this issue). Several factors might make this task more difficult in Chinese than in English. First, Chinese is a tonal language. Hence, pitch glides are both rapid and phonemic in Chinese compared to English. To the extent that segregating the target from the background depends on the listener's ability to "track" the fundamental frequency of the target talker, Chinese listeners might find it more difficult than English listeners to segregate the target talker from same-language competitors because of the rapid and frequent pitch changes that occur in Chinese as opposed to English.

A third potential difference between Chinese and English that could lead to a lesser degree of release from informational masking in Chinese than in English is that access to words might differ substantially in the two languages. Most words in Chinese can be considered as two-morpheme compound words. Hence, two-morpheme words are likely to lead to multiple activations in the semantic systems of Chinese listeners because of their compound nature than are two-syllable words in an English listener's semantic system. Indeed, in difficult listening situations, multiple paths to the access of the meaning of words might be useful. To test this, we also scored the Chinese target words using the less stringent requirement that a word would be scored as correct if either one of its morphemes was correctly identified. Figure 7 indicates that the use of a less stringent criterion for scoring a word as correct did not affect the amount of release from masking when the masker was semantically meaningless (noise or cross-language) but did improve the degree of release from masking when the masker was semantically meaningful (Chinese). Indeed, the amount of release at the morphemic level from a Chinese language masker was the same as the amount of release from an English masker of English. This suggests that the release from an informational masker occurs at the morphemic level.

To check whether the greater release at the morphemic than at the word level in Chinese might simply reflect a relaxation of the criterion for being correct (at least one of the morphemes in the two-morpheme target words being correctly identified), we computed separate psychometric functions for monosyllabic and multisyllabic target words in English (Figure 8). If it were simply the number of syllables that led to a larger release from informational masking, we would expect more release for monosyllabic English words than for multisyllabic English words. The fact that we did not find any differences indicates that the greater release from masking for morpheme scoring than for word scoring in Chinese is not simply due to the fact that only one of the two morphemes needed to be identified in order for the target word to be scored as correct.

The greater release from masking when morpheme scoring was used suggests that release from informational masking might be occurring at the morphemic level. One way of investigating the amount of release from masking that may be occurring at the morpheme level in Chinese due to spatial separation is to measure the increment in the probability of correctly identifying the first morpheme of a target word when the masker and target are spatially separated. In Figure 11, the ordinate is the probability of getting the first morpheme correct when the masker and target are spatially separated minus the probability of getting the first morpheme correct when masker and target are perceived to be colocated. This increment in probability due to spatial separation was computed at each SNR for each of the masking conditions. We then plotted this increment as a function of the probability of getting the first morpheme correct when the masker and target were colocated. Figure 11 shows that when the masker is either noise or English, the amount of release due to spatial separation is approximately the same function of the probability of correctly identifying the first phoneme when there is no spatial separation. However, when the masker is Chinese (same-language masker), the increment in the probability of identifying the first morpheme due to spatial separation is much larger.

Figure 11. Observed increment in percent correct due to spatial separation as a function of the probability of getting the first morpheme in a two-morpheme Chinese target word correct when the masker and target are colocated. The increment in percent correct at any one of the four SNRs is defined as the probability of getting Morpheme 1 correct when the perceived masker is on the left and the target is on the right minus the probability of getting Morpheme 1 correct when both the masker and target are perceived to be on the right. Normal distributions were fit to the data for the linguistic (Chinese; solid line) and nonlinguistic (noise, English; dashed line) maskers separately.



Hence, it appears that the masker must be semantically meaningful to Chinese listeners to obtain a substantial release from informational masking.

We also investigated whether spatial separation might improve the likelihood of linking the two syllables together to correctly access the target words. Figures 9 and 10 indicate that there is no evidence to suggest that spatial separation improves the ability to access the meaning of a word apart from the effect it has on increasing the likelihood of individual morpheme recognition. These results have interesting implications for language processing in noisy and complex listening situations. Figures 6 and 11 clearly indicate that there is a substantial amount of release at the morphemic level (due to spatial separation) from a same-language masker of Chinese. However, this does not translate into a substantial release from masking at the word level because spatial separation does not improve the ability of the listener to link the two morphemes together to access the word (see Figures 9 and 10). Indeed, by using the functions shown in Figures 9, 10, and 11, we can predict the psychometric functions when masker and target are spatially separated from the functions describing how accuracy in identifying the first morpheme of a target word increases as a function of SNR when maskers and targets are colocated. This was done in the following way: First, we determined the psychometric function relating accuracy of identifying the first morpheme to SNR for each of the three maskers (noise, cross-language, same language). We then used these psychometric functions to estimate the probability of getting the first morpheme correct at each of the four SNRs employed in this experiment. This was followed by using the functions in Figure 11 to compute the increment in the probability of correctly identifying the first morpheme at that SNR when the masker and target were perceived to be spatially separated. We then used the functions shown in Figures 9 and 10 to derive the predicted percent correct as a function of SNR for spatial separation of target and masker when morpheme-level scoring was used and when whole-word scoring was used. These predicted functions are shown in Figure 6 as dashed lines. Clearly, a model in which release from masking occurs only at the morpheme level can predict the pattern of results observed for each of the three types of maskers for both whole-word and morpheme level scoring when target and masker are spatially separated. (For details of the model's prediction, see the Appendix.)

It is interesting to note that the processes linking morphemes together to form words were unaffected by spatial separation. There were reasons to expect otherwise. Recall in the introduction we reviewed the notion that there may be algorithmic processes operating to link morphemes together to form words. In quiet situations, it would be difficult to observe such processes in action because there are no errors in identifying the words. Hence, in a quiet background it would be difficult to accumulate experimental evidence to support a model in which there is a pathway to precompiled words via morpheme identification and refute Packard's (1999) dismissal of such pathways as "unwieldly" and "computationally costly." In noisy situations with Chinese listeners, however, sometimes listeners recognize one morpheme of a word and not the other. The existence of a supplemental pathway to precompiled words via morpheme identification would be of some value in such cases. The presence of a significant amount of masking therefore gives us a window to study the processes that bind morphemes together to form words, and to determine whether informational masking might be interfering with these binding processes. There are, indeed, reasons to expect that interference might be occurring at that level, and if it is, spatial separation should mitigate its effects. For example, suppose the binding of morphemes required the resources of working memory. Current information-processing models use the term *n m m* to refer to the limited-capacity system that is responsible for the processing and temporary storage of task-relevant information during the performance of everyday cognitive tasks such as language comprehension (Baddeley, 1986; Daneman & Carpenter, 1980; Daneman & Merikle, 1996; Miyake & Shah, 1999). If the combining of morphemes together to form words required working memory resources, then

we would expect spatial separation to facilitate these processes. The logic behind this argument is as follows: If working memory resources are required to combine morphemes, any factor that reduces the demands on working memory should facilitate these combinatorial processes. A prevalent theory of how spatial separation leads to release from informational masking is that it facilitates stream segregation (for a review of such theories, see Schneider et al., 2010). If so, spatial separation should make it easier to inhibit the processing of the morphemes spoken by the competing talkers (Hasher & Zacks, 1988; Hasher, Zacks, & May, 1999), thereby reducing the burden on working memory by reducing the intrusion of irrelevant information from the competing streams into working memory, or facilitating its purging from working memory. In either case, spatial separation should facilitate the listener's ability to combine morphemes together to form words. The fact that spatial separation does not do this suggests that combining morphemes to form words does not require working memory resources when the spoken sentences are syntactically correct but semantically anomalous. Note that this does not mean that working memory resources are not required for word access in all cases. For instance, in difficult listening situations in which words may be misheard, working memory resources may be required for recovery or repair of the misheard words from the context provided by the sentence (e.g., Pichora-Fuller, Schneider, & Daneman, 1995). Hence, when top-down processing is involved, word access may require working memory resources. But in situations like the present one where the anomalous sentences do not provide any contextual support for word identification, it appears that word access does not require working memory resources. This, in turn, suggests that the automatic retrieval of morphemes and/or words precedes the processing of linguistic material within working memory and that competing anomalous sentences interfere with this automatic retrieval process.

In summary, a model in which spatial separation reduces interference at the morpheme level only is able to account for the amount of release from informational masking in both monolingual English and Chinese listeners.

Acknowledgments

Li, L., Qi, J. G., Yu, H., Alain, C., & Schneider, B. A. (2005). Attribute capture in the precedence effect for long-duration noise sounds. *H a n R a c , 202,* 235–247.

Marrone, N., Mason, C. R., & Kidd, G. (2008). The effects of hearing loss and age on the benefit of spatial separation between multiple talkers in reverberant rooms. *T J na ¿ Ac ¿ ca S c ¿ A n ca, 124,* 3064–3075.

Marslen-Wilson, W. (1989). Access and integration: Projecting sound onto meaning. In W. Marslen-Wilson (Ed.), *L ca n¿a¿ n an c* (pp. 3–24). Cambridge, MA: MIT Press.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *C n ¿ ¿ P c , 18,* 1–86.

Miller, J. L., & Eimas, P. D.

Appendix (p. 1 of 2). Details of the model's prediction.

The first step in fitting the model was to plot mean percent-correct identification of the first morpheme of the target word as a function of SNR for noise, English, and Chinese maskers when targets and maskers were perceived to be coming from the same location. Psychometric functions of the form indicated in Equation 1 were then fit to the mean data. Hence, the following psychometric functions were determined:

$$y_{1,N,R}[x] = \frac{1}{1+e^{-\sigma_{N,R}(x-\mu_{N,R})}}$$
$$y_{1,E,R}[x] = \frac{1}{1+e^{-\sigma_{E,R}(x-\mu_{E,R})}}$$
$$y_{1,C,R}[x] = \frac{1}{1+e^{-\sigma_{C,R}(x-\mu_{C,R})}},$$

where subscript 1 indicates that the y values are for Morpheme 1; N, E, and C stand for noise, English, and Chinese maskers, respectively; R indicates that the masker was perceived to be on the right (masker and target colocated); and x is the SNR in dB. In Figure 11, there are two functions specifying how spatial separation leads to an increment in the probability of correctly identifying the first morpheme of the target. The separate functions fit to semantically empty (SE) and semantically meaningful (SM) maskers were

$$r_{SE}[y] = 0.15173 * e^{-6.50732(y-.396899)^2}$$
$$r_{SM}[y] = 0.361296 * e^{9.39338(y-.29744)^2},$$

where r specifies the increment in the probability of correctly identifying the first morpheme when masker and target are spatially separate, as a function of y, the probability of correctly identifying the first morpheme correctly when masker and target are colocated. Hence, the probability of getting Morpheme 1 correct at SNR = x, when masker and target are separated, is

$$y_{1,N,L}[x] = r_{SE}[y_{1,N,R}[x]] + y_{1,N,R}[x]$$
$$y_{1,E,L}[x] = r_{SE}[y_{1,E,R}[x]] + y_{1,E,R}[x]$$
$$y_{1,C,L}[x] = r_{SM}[y_{1,C,R}[x]] + y_{1,C,R}[x],$$

where L indicates that the masker is perceived on the left, that is, spatially separated from the target.

It follows that the probability of getting the whole word correct is the probability of getting Morpheme 1 correct times the probability of getting Morpheme 2 correct given that Morpheme 1 is correct. In Figure 10, the probability of getting Morpheme 2 correct given that Morpheme 1 is correct is

$$p(y_2|y_1) = .673912 + .289612 * p(y_1)$$

under all six conditions, where y_2 is a correct identification of the second morpheme and y_1 is a correct identification of the first morpheme. Hence, the probability of getting the whole word correct on the left or right sides as a function of SNR is

$$y_{1\&2,N,L}[x] = y_{1,N,L}[x](.673912 + .289612 * y_{1,N,L}[x])$$
$$y_{1\&2,E,L}[x] = y_{1,E,L}[x](.673912 + .289612 * y_{1,E,L}[x])$$
$$y_{1\&2,C,L}[x] = y_{1,C,L}[x](.673912 + .289612 * y_{1,C,L}[x])$$
$$y_{1\&2,N,R}[x] = y_{1,N,R}[x](.673912 + .289612 * y_{1,N,R}[x])$$
$$y_{1\&2,E,R}[x] = y_{1,E,R}[x](.673912 + .289612 * y_{1,E,R}[x])$$
$$y_{1\&2,C,R}[x] = y_{1,C,R}[x](.673912 + .289612 * y_{1,C,R}[x]),$$

where subscripts 1 and 2 stand for getting both morphemes in a word correct—that is, getting the whole word correct.

Finally the probability of getting either Morpheme 1 or Morpheme 2 correct (subscript 1 or 2) is computed by first determining the probability of getting the first morpheme correct. To this is added the probability of getting the first morpheme wrong times the probability of getting the second morpheme correct given that the first morpheme is incorrect. This latter probability of getting the second morpheme correct given that the first is incorrect is

$$p(y_2|\overline{y_1}) = .533084 - 1.1076 * \overline{y_1}[x] + 0.603891 * \overline{y_1}[x]^2,$$

where $\overline{y_1}[x] = 1 - y_1[x]$.

Appendix (p. 2 of 2). Details of the model's prediction.

Hence,

$$
\begin{aligned}
y_{1\ or\ 2,N,L}[x] &= y_{1,N,L}[x] + \overline{y_{1,N,L}}[x] * (.533084 - 1.1076 * \overline{y_{1,N,L}}[x] + .603891 * \overline{y_{1,N,L}}[x]^2) \\
y_{1\ or\ 2,E,L}[x] &= y_{1,E,L}[x] + \overline{y_{1,E,L}}[x] * (.533084 - 1.1076 * \overline{y_{1,E,L}}[x] + .603891 * \overline{y_{1,E,L}}[x]^2) \\
y_{1\ or\ 2,C,L}[x] &= y_{1,C,L}[x] + \overline{y_{1,C,L}}[x] * (.533084 - 1.1076 * \overline{y_{1,C,L}}[x] + .603891 * \overline{y_{1,C,L}}[x]^2) \\
y_{1\ or\ 2,N,R}[x] &= y_{1,N,R}[x] + \overline{y_{1,N,R}}[x] * (.533084 - 1.1076 * \overline{y_{1,N,R}}[x] + .603891 * \overline{y_{1,N,R}}[x]^2) \\
y_{1\ or\ 2,E,R}[x] &= y_{1,E,R}[x] + \overline{y_{1,E,R}}[x] * (.533084 - 1.1076 * \overline{y_{1,E,R}}[x] + .603891 * \overline{y_{1,E,R}}[x]^2) \\
y_{1\ or\ 2,C,R}[x] &= y_{1,C,R}[x] + \overline{y_{1,C,R}}[x] * (.533084 - 1.1076 * \overline{y_{1,C,R}}[x] + .603891 * \overline{y_{1,C,R}}[x]^2).
\end{aligned}
$$

Four points, corresponding to SNRs of −12, −8, −4, and 0, were determined for each of the six functions in Equation A7. Psychometric functions were then fit to these four points. The dashed lines in Figure 6 represent these predicted psychometric functions.

Cross-Language Differences in Informational Masking of Speech by Speech: English Versus Mandarin Chinese

Xihong Wu, Zhigang Yang, Ying Huang, Jing Chen, Liang Li, Meredyth Daneman, and Bruce A. Schneider

J Speech Lang Hear Res 2011;54;1506-1524
DOI: 10.1044/1092-4388(2011/10-0282)