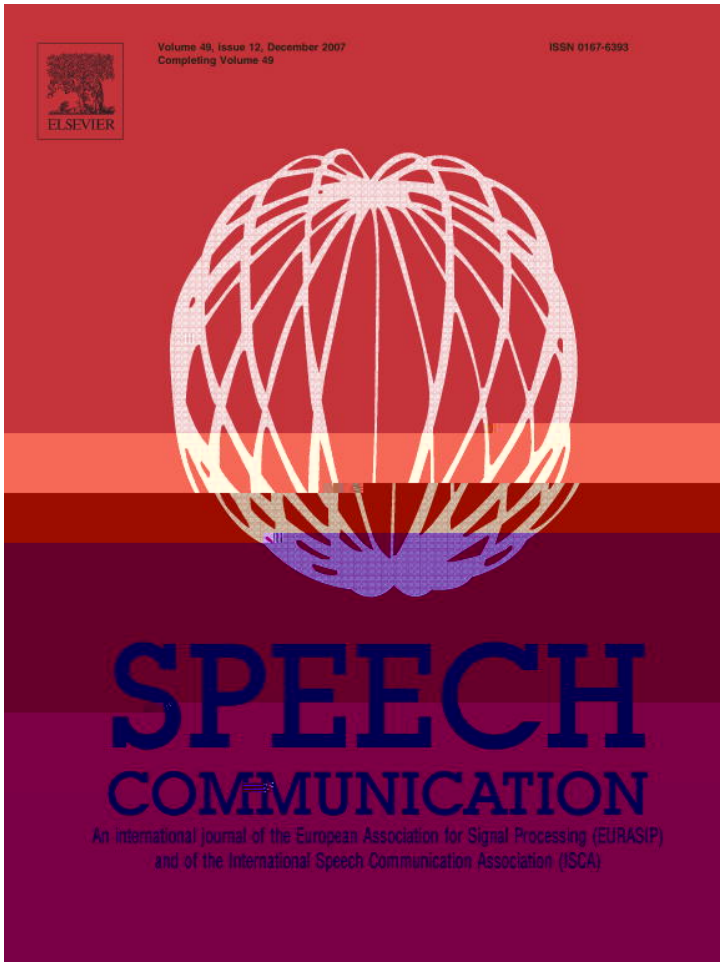
Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article was published in an Elsevier journal. The attached copy is furnished to the author for non-commercial research and education use, including for instruction at the author's institution, sharing with colleagues and providing to institution administration.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

http://www.elsevier.com/copyright

Author's personal copy

The effect of voice cuing on releasing Chinese speech from informational masking ☆

Zhigang Yang [a], Jing Chen [a], Qiang Huang [a], Xihong Wu [a], Yanhong Wu [a], Bruce A. Schneider [b], Liang Li [a,b,*]

[a] *Department of Psychology, National Key Laboratory on Machine Perception, Speech and Hearing Research Center, Peking University, Beijing 100871, China*

[b] *Department of Psychology, Centre for Research on Biological Communication Systems, University of Toronto at Mississauga, Mississauga, Ontario, Canada L5L 1C6*

Received 14 December 2006; received in revised form 16 May 2007; accepted 17 May 2007

Abstract

In a cocktail-party environment, human listeners are able to use perceptual-level and cognitive-level cues to segregate the attended target speech from other background conversations. At the cognitive level, priming the listener with part of the target speech in quiet can markedly improve the recognition of the remaining parts when the target speech and competing speech are presented at the same time. Hence, knowledge of content (content cuing) improves speech recognition when other people are talking. In addition, familiarity or knowledge of the voice characteristics of the target talker could also help the listener attend to the target talker when other talkers are present. The present study investigated the extent to which a cognitive-level cue (content cuing) and a perceptual-level cue (voice cuing) can improve word identification for speech masked by noise or by other speech in Chinese listeners. Specifically, listeners were primed with part of a sentence in quiet before a sentence was repeated in the presence of either noise or speech. The priming sentence was always in the same voice as the target sentence. Two kinds of primes were investigated: same-sentence primes, and different-sentence primes. Under speech-masking conditions, each of the two prime types significantly improved recognition of the last key word in the full-length target sentence. Under noise-masking conditions, same-sentence primes had a weak but significant rer9(re729.6(effec,k)-26542(but)-570.7(different-s

Author's personal copy

1999, 2001, 2004; Kidd et al., 1994, 1998; Li et al., 2004; Lutfi, 1990; Oxenham et al., 2003; Shinn-Cunningham et al., 2005; Summers and Molis, 2004; Wu et al., 2005). Energetic masking occurs when peripheral neural activity elicited by a signal is overwhelmed by that elicited by maskers, leading to a degraded or noisy neural representation of the signal, making it difficult for subsequent cognitive processes to extract the signal.

In addition to energetic or peripheral masking, which is always present especially when there are competing sound sources (Arbogast et al., 2002), competing sound sources can interfere with the processing of the target speech at levels beyond the cochlea. In particular, when the masker is speech, processing of the information in the masker may interfere with processing of the target speech at any one of a number of perceptual (e.g., phonemic identification) or cognitive (e.g., semantic processing) levels. As a result, the listener may find it difficult to segregate the target from the masker because the information in the speech masker is interfering with the processing of information in the target-talker's speech (informational masking). It is important to note that the distracting speech can also mask the target speech at a peripheral (i.e., energetic) level because the speech masker will activate the same or nearby regions on the basilar membrane that are processing the target speech. Hence a speech masker will interfere with speech processing because of energetic masking and informational interference.

Effects of informational masking can be alleviated when certain perceptual cues are provided that enable the listener to perceptually segregate the target from the speech masker. For example, either a real or perceived spatial separation between the target talker and the speech masker will result in improved speech recognition. In the case of a real spatial separation between the target and masker, part of this improvement is undoubtedly due to an increase in the signal-to-noise ratio (SNR) at the ear furthest from the masking speech. However, it is interesting to note that even when the spatial separation is only perceived but not real (spatial separation induced by the precedence effect), and consequently there is no improvement in the SNR at either ear, listeners still find it easier to process the target speech, presumably because the perceived spatial separation makes it easier to ignore or inhibit the processing of the informational content in the speech masker (Freyman et al., 1999, 2001; Li et al., 2004; Wu et al., 2005; Schneider et al., in press). Informational masking is also reduced when listeners are forewarned as to the nature of the target speech (e.g., voice characteristics, content, or location), so that they can selectively attend to the target speech (Brungart et al., 2001; Freyman et al., 2004; Kidd et al., 2005a,b).

In addition to perceived spatial separation induced by the precedence effect, another efficient way to reduce the amount of informational masking is to present part of speech signal without the target component just before the full-length speech sentence containing the target component is presented. For example, in a study by Freyman et al. (2004), part of a syntactically correct but semantically-anomalous English speech sentence, whose last key word was replaced by a noise burst, was used as the prime and presented just before the presentation of full-length target speech spoken by a female talker. Following the presentation of the prime using (1) the target-talker's voice (female), (2) a male's voice, or (3) a written presentation of the prime, the SNR corresponding to 50% correct recognition of the last key word in the full target speech sentence was markedly reduced (by about 4 dB) when the masker was two-talker speech. Interestingly, there was no difference between the three different ways of presenting the prime, suggesting a content-cuing effect that is not affected by vocal information.

However, in every-day speech communication situations, the amount of speech masking is highly dependent on the similarity of the target and masker voices. For example, the voice of the same talker produces a larger masking effect on target speech than the voice of a different talker of the same gender or the voice of a talker of a different gender (Brungart, 2001; Brungart et al., 2001; Festen and Plomp, 1990). Thus, it would be important to know, when neither spatial nor content cues are available, whether there is a cuing effect of the voice of the target talker on recognizing target speech in the presence of a speech masker or a noise masker.

1.2. Are there language-based differences in energetic and informational masking?

Most of the studies of informational masking have been conducted with English listeners. Because informational masking of speech by speech involves potential interference at phonemic, semantic, and linguistic levels, we might expect to find differences in the extent and kind of informational masking across language groups. For example, in Chinese, syllables have more voiceless consonants and fewer voiced consonants than English syllables. Therefore Chinese speech may be more vulnerable to noise masking (Kang, 1998). In addition, it has been shown that listeners can benefit from amplitude troughs (or temporal gaps) in the masker when listening to speech (Gustafsson and Arlinger, 1994; Howard-Jones and Rosen, 1993; Nelson et al., 2003; Summers and Molis, 2004). If the amplitude envelope of Chinese speech is different from that of English speech, we might expect differences with respect to the degree to which competing speech in these two language groups is an effective masker. Moreover, because the pitch contour of the vowel in Chinese is phonemic, the extent of informational masking may differ between Chinese and English. Krishnan et al. (2005) have shown that native speakers of Chinese have stronger neural representation of the pitch contours of vowels than those whose native language is English. Hence the salience of pitch contours will be greater in Chinese than in English. In Mandarin Chinese, the pitch contour will typically changes from syllable to syllable, whereas in English the pitch contour within and across

Author's personal copy

syllables is likely to be more uniform. Hence, there is likely to be greater moment-to-moment variation in the fundamental frequency (F_0) in a Mandarin Chinese utterance, where each syllable has its own pitch contour, than in an English utterance, where the pitch contour is more uniform across syllables. It is well known that talkers differ with respect to how F_0 changes during an utterance (the F_0 contour), and that differences in F_0 contours between a target talker and competing talkers can facilitate tracking of the target talker when there are competing talkers (Assmann and Summerfield, 1989; Darwin and Hukin, 2000; Darwin et al., 2003). Hence, because there is greater variability in the F_0 contour in Chinese than in English, the usefulness of this cue may differ across the two languages.

In addition, in contemporary Mandarin Chinese, a large number of words are two-character compound words in which each of the two characters (two syllables) has its own semantic representation. For example, the Chinese word for "Beijing" is a two-syllable (/Bei3/ and /Jing1/) compound word in which each syllable has a specific referent (character 1 means "north", character 2 means "capital city"). There are 31 two-syllable words (collected from *People's Daily* published from 1994 to 2002) that begin with the third-tone "/*Bei*/", and 119 two-syllable words that begin with the syllable "/*Bei*/" whose tone can be any one of the four. On the other hand, there are 45 two-syllable words that end with the first-tone "/*Jing*/", and 189 two-syllable words that end with the syllable "/*Jing*/" whose tone can be any one of the four. Because there are so many possible two-syllable words beginning with "/Bei/", and so many two-syllable words that end in "/Jing/", a Chinese listener may have to access the meanings of both syllables in order to get the whole word correct. In English, with the exception of compound words, the listener generally does not have to access the meaning of the individual syllables to arrive at the meaning of the whole word. In addition, the sequential dependences between the syllables in the two-syllable Chinese words could also affect the degree of informational masking. In particular, access to the meaning of the first syllable might be expected to facilitate access to the meaning of the second

Author's personal copy

in the experiments and were paid a modest stipend for their participation.

2.2. Apparatus

Listeners were seated in a chair at the center of an anechoic chamber (Beijing CA Acoustics), which was 560 cm in length, 400 cm in width, and 193 cm in height. All acoustic signals were digitized at the sampling rate of 22.05 kHz using the 24-bit Creative Sound Blaster PCI128 (which had a built-in anti-aliasing filter) and audio editing software (Cooledit Pro 2.0), under the control of a computer with a Pentium IV processor. The acoustic signals were delivered to a loudspeaker (Dynaudio Acoustics, BM6 A), which was in the frontal azimuthal plane at 0° position (with respect to the median plane). The loudspeaker height was 106 cm, which was approximately ear level for a seated listener with average body height. The distance between the loudspeaker and the center of the participant's head was 185 cm.

2.3. Stimuli

2.3.1. Chinese nonsense sentences

Speech stimuli were Chinese "nonsense" sentences. Direct English translations of the sentences are similar but not identical to the English nonsense sentences that were developed by Helfer (1997) and also used in studies by Freyman et al. (1999, 2001) and Li et al. (2004). Each of the Chinese nonsense sentences has three key components: subject, predicate, and object, which are also the three key words, with two characters for each (also one syllable for each character). Note that the sentence frame does not provide any contextual support for recognition of the key word.

Based on the database of the Chinese newspaper *People's Daily* published over 9 years (1994–2002), 6000 double-syllable verbs, which were rated as having high frequencies of occurrence, and 12,000 double-syllable nouns, which were also rated as having high frequencies of occurrence, were used. These words were combined randomly into 6000 syntactically correct sentences with the frame of *subject* + *predicate* + *object*. To ensure that sentences used in experiments were not meaningful, the probability of co-occurrence of two nouns with a verb in a normal sentence was determined according to the database of *People's Daily* over 9 years. Only sentences whose probability of co-occurrence of key words in the database was zero were used as the nonsense sentences for the present study. Since Chinese is a tonal language, further selection was made to balance syllable tones across sentences. A double-syllable pronoun was then placed before a noun, and an auxiliary verb was placed before a verb, making a selected sentence more natural. Finally, all sentences were examined by the experimenters to ensure that selected sentences were nonsensical.

Both target speech and different-sentence cuing speech used in this study were spoken by a young female talker (Talker A). Masking speech was a continuous recording of masking Chinese nonsense sentences simultaneously spoken by two other young female talkers (Talkers B and C). Talker B and Talker C spoke different masking sentences. All speech stimuli were recorded digitally onto computer disks, sampled at 22.05 kHz and saved as 16-bit PCM wave files.

Twenty-four lists (18 sentences/list) of nonsense sentences were used as target sentences. To balance information quantity across experimental conditions in this study, the information quantity of a key word in a sentence was calculated as

$$I = -\log\left(\frac{1}{f}\right)$$

where f is word frequency. Information quantity of a sentence was the sum of information quantities of the three key words. All the lists of nonsense sentences were constructed in such a way that the information quantity of each list was about the same. In a target sentence, only the last key word was scored during speech recognition testing. To equate the sentences with respect to audibility, all sentences were rescaled to have the same RMS value, and all sentences (both target and cuing) were presented at the same decibel level (52 dBA).

In the same-sentence cuing condition, the prime, which was spoken by Talker A, was identical to the target sentence except that the last key word was replaced by a white noise burst, whose duration was equal to that of the longest of the last (third) key words in all the target sentences, and whose level was 10 dB lower (both sentence and noise were measured in dBA) than that of the preceding sentence (following Freyman et al., 2004). In the different-sentence cuing condition, a nonsense sentence, whose content was different from that of the target sentence, was also spoken by Talker A, with all other aspects (including the replacement of the last key word with white noise) being identical to the same-sentence cuing condition (Fig. 1). One hundred and forty-four nonsense sentences were used as different-sentence cuing speech materials. Fig. 1 shows the waveform of one of the target sentences, the same-sentence prime, and a different-sentence prime, respectively.

2.3.2. Speech-spectrum noise

Three hundred frequently occurring syllables were chosen from the database of *People's Daily* published for one year. One hundred and thirteen sentences, which appeared in *People's Daily* and contained 317 syllables including all the 300 frequently occurring syllables, were selected as acoustic material for making speech-spectrum noise. The 113 different sentences were assigned to 50 Chinese young female speakers. Fifty-seven sentences were spoken by 25 speakers and 56 other sentences were spoken by another 25 speakers at a medium rate of speech. Recording of the sentences were stored digitally onto computer disks, sampled at 22.05 kHz and saved as 16-bit PCM wave files. All of the 50-voice sentences were mixed using Matlab

Author's personal copy

Author's personal copy

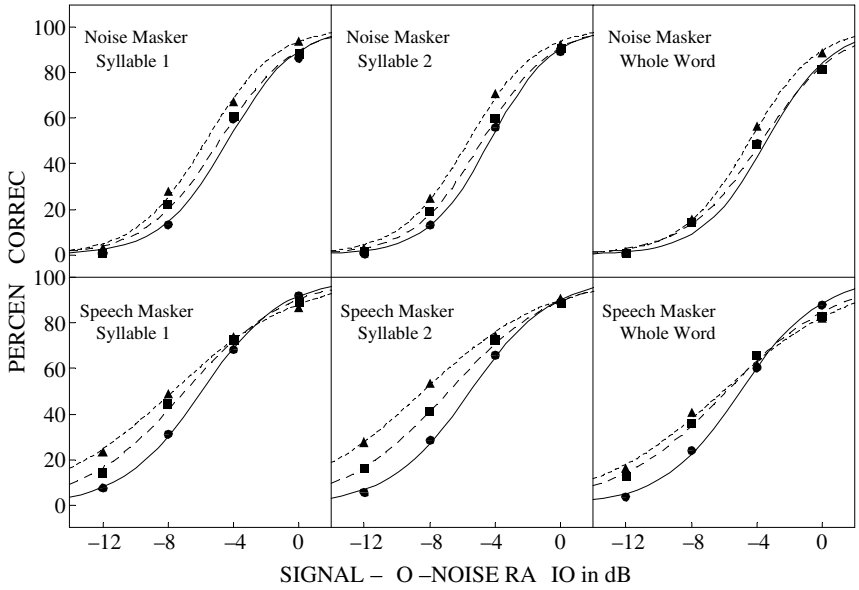
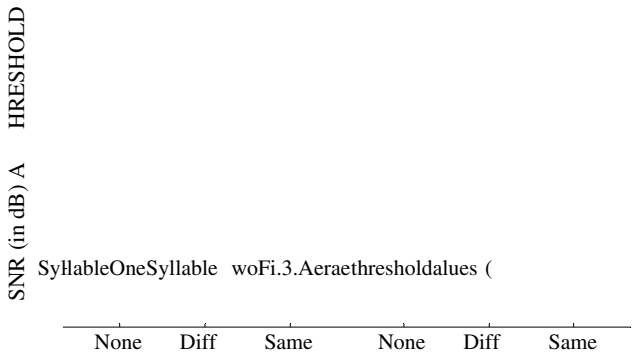


Fig. 2. The top left panel shows, when the masker was noise, mean percent-correct identification of syllable one across 18 listeners as a function of signal-to-noise ratio (SNR) for the three priming conditions: (a) no prime (circles and solid line); (b) different-sentence prime (squares and dashed line); and (c) same-sentence prime (triangles and dotted line). The bottom left panel shows the equivalent priming conditions for syllable one when the masker was speech. The top and bottom middle panels present the same data for syllable two. The two right panels show the results when the whole word was scored as correct. The best-fitting psychometric functions (curves) under all the conditions are also shown.

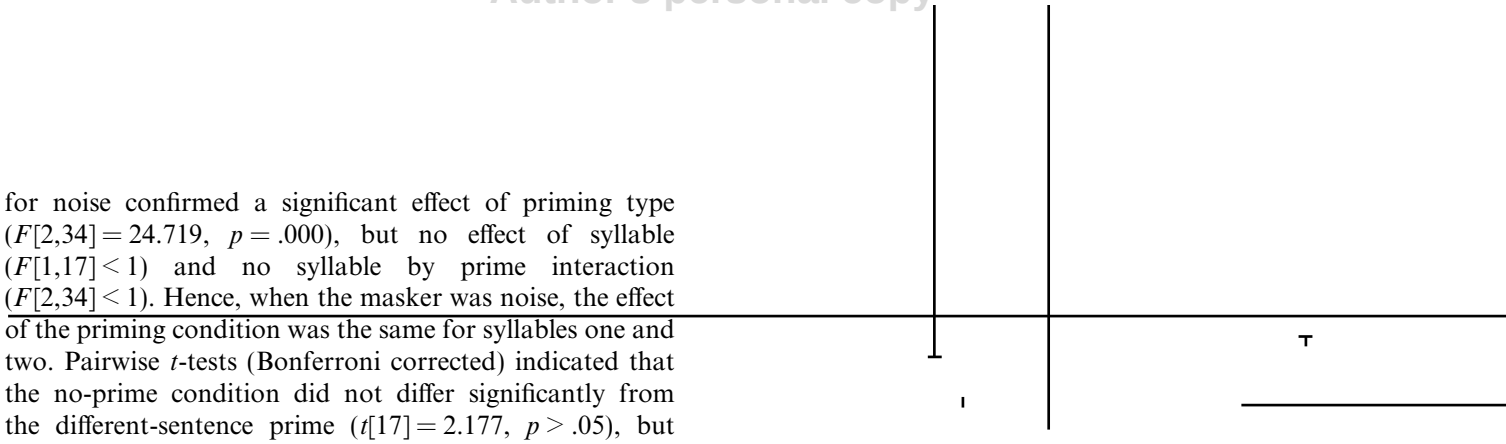
Finally, the two rightmost panels plot the percentage of times the whole word was correctly identified as a function of SNR. Again, both the same- and different-sentence primes appear to provide some release from both noise and speech maskers. However, the amount of release appears to be smaller than that observed when the first and second syllables were considered separately.

To determine whether the psychometric functions shown in Fig. 2 also characterized the individual participants, we fit individual psychometric functions in all of the conditions. Fig. 3 shows how mean threshold values (μ) varied with masker type and priming condition for the first and second syllables considered separately. In all three priming conditions, and for both syllables, lower threshold were observed for speech maskers compared to noise maskers. Also, irrespective of the type of masker, the highest thresholds were observed when no prime was presented, followed by the different-sentence prime, with the lowest thresholds occurring with the same-sentence prime. Fig. 3 also indicates that the extent of release from masking due to priming was larger for the speech masker than for the noise masker (compare the difference between no prime and same-sentence primes between speech masking and noise masking). Finally, the extent of release from masking due to the use of a same-sentence prime appears to be larger for the second syllable than for the first syllable when the masker is speech. A three-factor, within-subject ANOVA confirmed that there was no main effect of syllable ($F[1,17] < 1$). However, the main effects due to masker type ($F[1,17] = 75.438$, $p = .000$) and priming condition ($F[2,34] = 53.236$, $p = .000$) were highly significant. In addition, there was a significant interaction between syllable and type of prime ($F[2,34] = 9.241$, $p = .001$) confirming that the release from masking was larger for the second syllable than for the first. The other two-way interactions between masker and type of prime ($F[2,34] = 2.903$, $p = .069$), and between syllable and the type of masker ($F[1,17] < 1$) were not significant. However, there was a highly significant three-way interaction between syllable, masker type, and priming condition, $F[2,34] = 12.903$, $p = .000$.



Two-factor ANOVAs (syllable × priming condition) were conducted separately for noise and for speech to identify the source of this three-way interaction. The ANOVA

Author's personal copy



for noise confirmed a significant effect of priming type ($F[2,34] = 24.719$, $p = .000$), but no effect of syllable ($F[1,17] < 1$) and no syllable by prime interaction ($F[2,34] < 1$). Hence, when the masker was noise, the effect of the priming condition was the same for syllables one and two. Pairwise t-tests (Bonferroni corrected) indicated that the no-prime condition did not differ significantly from the different-sentence prime ($t[17] = 2.177$, $p > .05$), but that it did differ from the same-sentence prime ($t[17] = 7.081$, $p < .001$), and that the different-sentence prime differed significantly from the same-sentence prime ($t[17] = 6.434$, $p < .001$). Hence, when the masker was noise, there was release from masking when a same-sentence prime was used, but not when a different-sentence prime was used.

The equivalent ANOVA for the speech masker found no significant main effect of syllable ($F[1,17] = 1.447$, $p = .246$) but did find significant effects of priming ($F[2,34] = 22.173$, $p = .000$), and a significant syllable × priming interaction ($F[2,34] = 15.570$, $p = .000$), indicating that the effect of priming was stronger for syllable two than it was for syllable one. Multiple t-tests (Bonferroni corrected) confirmed that, for the first syllable, the no-prime condition differed significantly from the two priming conditions (no-prime vs different-sentence prime, $t[17] = 3.078$, $p < .05$; no-prime vs same-sentence prime, $t[17] = 4.610$, $p < .001$), but that the two priming conditions did not differ significantly from one another ($t[17] = 2.470$, $p > .05$). However, t-tests (Bonferroni corrected) showed that all three priming conditions differed from one another for syllable two (no-prime vs different-sentence prime, $t[17] = 3.484$, $p < .01$; no-prime vs same-sentence prime, $t[17] = 6.864$, $p < .001$; different-sentence prime vs same-sentence prime, $t[17] = 4.336$, $p < .005$). Multiple t-tests (Bonferroni corrected) also confirmed that although the difference between the no-prime and different-sentence prime was the same for syllable one as it was for syllable two ($t[1,17] = -2.218$, $p > .05$), the difference between no-prime and same-sentence prime was larger for syllable two than for syllable one ($t[17] = -5.010$, $p < .001$), as was the difference between the different-sentence and same-sentence primes ($t[17] = -3.302$, $p < .05$). Hence, both different-sentence primes and same-sentence primes produce a release from speech masking, with same-sentence primes producing a larger release than different-sentence primes, and with the difference between no prime and same-sentence primes, and the difference between different-sentence and same-sentence primes being larger for syllable two than for syllable one.

Fig. 4 indicates how the slope parameter, σ, varied with masker type and priming condition for syllables one and two of the target word. In general slopes were shallower when the masker was speech than when the masker was noise. It also appears that slopes are steeper when there is no prime than when there is a prime. A three-factor, within-subject ANOVA confirmed that there was a significant main effect of masker ($F[1,17] = 86.348$, $p = .000$), a significant main effect of priming condition ($F[2,34] = 12.989$, $p = .000$), but no main effect of syllable ($F[1,17] = 2.305$, $p = .147$). The only interaction effect that approached significance was the interaction between masker and syllable ($F[1,17] = 4.118$, $p = .058$), which would be consistent with the observation that the slope differences between speech and noise maskers might be slightly larger for syllable two than for syllable one. Multiple t-tests (Bonferroni corrected) showed that slopes in the no priming condition were steeper than those in the different-sentence priming condition ($t[17] = 3.33$, $p < .05$), and those in the same-sentence priming condition ($t[17] = 4.72$, $p < .001$); but that slopes in the different-sentence priming condition did not differ significantly from those in the same-sentence priming condition ($t[17] = 1.65$, $p > .05$).

Figs. 5 and 6 show how thresholds and slopes, respectively, change as a function of masker type and priming condition, when the whole word (both syllables) was considered. Fig. 5 suggests that thresholds are lower for

Author's personal copy

that the primes provided a release from masking. In addition, the amount of release from masking was larger for same-sentence than for different-sentence primes.

Fig. 6 suggests that the slopes in the whole word condition were shallower when the masker was speech than when it was noise, and that they were also shallower when a prime was presented. A two-factor, within-subject ANOVA confirmed that there was a significant main effect of masker type ($F[1,17] = 63.688$, $p = .000$), a significant main effect of priming condition ($F[2,34] = 10.678$, $p = .000$), but no significant masker by priming condition interaction ($F[2,34] = 2.180$, $p = .129$). Multiple t-tests (Bonferroni corrected) indicated that the no-prime condition differed from each of the priming conditions (no-prime vs different-sentence prime, $t[17] = 3.743$, $p < .005$; no-prime vs same-sentence prime, $t[17] = 3.453$, $p < .01$) but the two priming conditions did not differ significantly from each other ($t[17] = -.403$, $p > .05$). Hence slopes were shallower when the masker was speech than when it was noise, and shallower when a prime was presented. Shallower slopes in the presence of a prime, coupled with an asymptote of 100% correct at the higher SNRs under all conditions, means that the amount of release from masking increased as SNR decreased. Hence, the lower the SNR, the greater was the amount of release from masking due to the presentation of a prime.

To determine whether the probability of getting the whole word correct was equal to the product of the probability of getting syllable one correct (p_1) times the probability of getting syllable two correct (p_2), we computed the number of times neither of the syllables were correctly identified, the number of times only syllable one was correctly identified, the number of times only syllable two was correctly identified, and the number of times both syllables

Author's personal copy

probability of correctly identifying the other syllable in all six conditions.

4. Discussion

Under each of the conditions in the present study, percent-correct word identification increased monotonically with the increase of SNR from −12 dB to 0 dB, without displaying plateaus. The absence of nonmonotonicity when both the target and the two-talker speech masker were perceived to be emanating from the same location is in agreement with the results reported by Brungart et al. (2001), Freyman et al. (2001), Li et al. (2004), and Wu et al. (2005).

Also consistent with other previous results (e.g., Brungart, 2001; Freyman et al., 1999; Li et al., 2004; Wu et al., 2005), the results of the present study show that the slopes of the psychometric function for word identification are generally steeper for the noise masker than they are for the speech masker. One explanation is that because there is considerable variation in the energy envelope of the speech masker, the instantaneous SNR is high when there is a pause or unvoiced consonants in the masking speech, and the instantaneous SNR is low when voicing occurs in the masking speech. The effect of these fluctuations in local SNR would be to flatten the psychometric function for a speech masker as compared to a steady-state noise masker, as indicated in the work of Rhebergen and Versfeld (2005) and Rhebergen et al. (2006). (Also see discussion below on differences between Chinese speech and English speech).

4.1. The effects of priming in a noise masker

The results of the present study show that a same-sentence prime produces a larger release from masking than a different-sentence prime when the whole word is considered. In addition, the amount of release due to a same-sentence prime was approximately the same for syllable one (1.34 dB) as it was for syllable two (1.36 dB), and only slightly lower (1.10 dB) when the whole word was considered. Slopes were steeper when no prime was presented than when either a different-sentence or same-sentence prime was given. Hence, the only significant effects observed for noise maskers is that both kinds of primes can lead to a release from masking on the order of 1 dB when the whole word is scored, and that the slopes of the psychometric functions are shallower when a prime is given.

4.2. The effects of priming in a speech masker

When the masker was speech, the introduction of a prime (either same-sentence or different-sentence) produced a reduction in slope for both syllable one and syllable two, and for whole word scoring. Moreover, there were no differences in slope between syllables one and two. Hence, when the masker is speech, the primary effect of priming is to reduce the slope of the psychometric function.

The effect of a prime on thresholds was slightly more complicated. First, same-sentence primes produced a greater release from masking in syllable two than in syllable one (a 1.85 dB release in syllable one and a 3.03 dB release in syllable two). Second, same-sentence primes produced a larger release from masking than different-sentence primes in syllable two and for whole word scoring.

The increased effectiveness of a prime on syllable two in the same-sentence priming condition would be expected if a correct identification of syllable one increased the likelihood of correctly identifying syllable two. Indeed, χ^2 tests indicated that the second syllable was more likely to be correctly identified if the syllable preceding it was correctly identified. Hence, as expected, the second syllable is more easily identified when the first syllable is correctly identified. Presumably, the correct identification of the first syllable reduces the search neighborhood for the second syllable, thereby facilitating the effectiveness of the priming stimulus at a cognitive-level.

The improvement in threshold when the whole word is scored appears to be smaller than the result (4.01 dB) reported in Freyman et al.'s study (2004), if only the 50% point of the psychometric function is examined (1 dB in improvement). Hence, when looking at the 50% threshold, there does not appear to be any significant difference when the same-sentence prime is presented in the presence of a speech masker than when it was presented in the presence of a noise masker. However, because of the flattening of the psychometric function in the presence of a prime when the masker is speech, the separation between the no prime and same-sentence prime in the speech masker condition increases with decreasing SNR. For example, when there is no-priming stimulus, participants in the speech-masking condition correctly identified 20% of the words at a SNR = −8.3 dB, whereas when the target sentence was preceded by the same-sentence prime, participants were able to identify 20% of the words at a SNR of −11.5 dB. Hence, in unfavorable listening conditions (only 20% of the words are correctly identified) the same-sentence prime provides a 3.2 dB advantage, which is closer to the 4.01 dB advantage (for 50% correct identification) reported by Freyman et al. (2004). If only the second syllable is considered, in the same unfavorable listening condition (20% of the second syllables correctly identified), the advantage increases to 4.7 dB. Thus, in spite of the differences between spoken Chinese and English speech (see below), results of the present study indicate that the advantage of same sentence auditory priming in unmasking speech is not limited to English but also extends to tonal Chinese. Since a substantial same-sentence priming effect has been observed in both languages, and presenting the prime does not influence the acoustics at the ears during the presentation of the masker and target, it is unlikely that the release from masking is due to peripheral acoustic features (which differ substantially in these two languages) but rather to the operation of higher-order processes.

Author's personal copy

One possible interpretation of the greater release when the individual syllables in the compound word are scored independently, is that in order to release two-syllable words from informational masking in Chinese, the listener has to access the meaning associated with each syllable, whereas in English, the need to do so does not occur as frequently (the exceptions being compound words). Because each syllable in a compound word must be correctly recognized in order to get the whole word correct, and because each syllable may initiate semantic activity related to its own individual meaning, the processing of two-syllable Chinese words is likely to differ significantly from the processing of most two-syllable words in English where individual syllables often lack meaning (e.g., "rotate"). This difference might help explain why when we score each Chinese syllable independently, the amount of release from informational masking is closer to that observed in English. For almost all of the English words (including multi-syllabic English words) only one meaning needs to be accessed, whereas the meaning of both syllables must be accessed in order to get the Chinese word correct. Hence, when only one meaning needs to be accessed, the amount of release from masking appears to be more comparable in the two languages.

A corollary of this hypothesis is that, when the language is English, the degree of release observed for two-syllable words should not differ substantially from the degree of release observed for one-syllable words. To check this, we examined the data from Li et al. (2004), in which there were 936 key words (312×3), of which there were 541 single-syllable key words (57.8%) and 395 multi-syllable key words (42.2%) (including 391 double-syllable key words and four triple-syllable key words). Fig. 7 plots the degree of release from masking as a function of perceived spatial position (target and masker perceived to originate from the same location vs target and masker perceived to originate from different locations). Fig. 7 shows that when the masker was noise, the psychometric functions for both same and different positions were virtually identical for one- and two-syllable target words. When the masker was speech, listeners were slightly better at identifying two-syllable words than one-syllable words, particularly when the target and masker were perceived as originating from the same position. Also, the degree of release from masking due to a shift in the perceived location of the speech masker was only slightly higher for one than for two-syllable words. Thus the results show that the degree of release observed for two-syllable words in English is comparable to that observed for one-syllable words. Presumably, this would not be the case for single-syllable as opposed to two-syllable Chinese words (see Fig. 8).

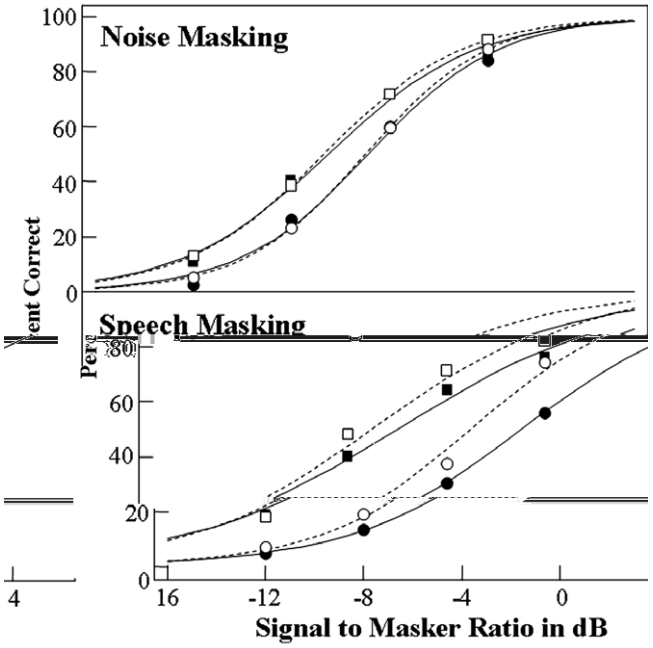


Fig. 7. Mean percent-correct identification of single-syllable key words (filled symbols and solid curves) and that of double-syllable key words function of SNR when the speech target and the masker were perceived as co-located (circles) or as spatially separated (squares). Top panel, noise (2004).

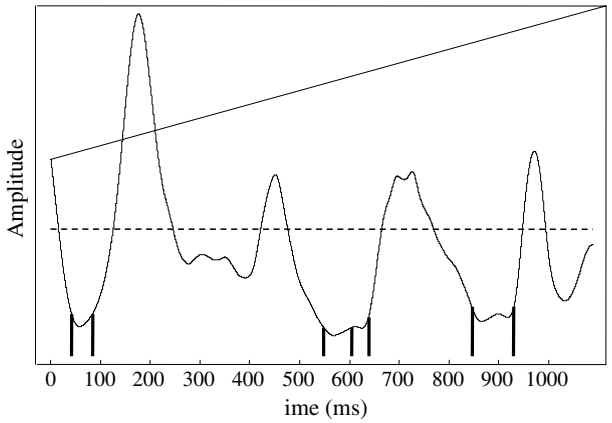


Fig. 8. The amplitude envelope for a segment of the two-talker Chinese speech masker. The locations and widths of deep troughs are identified by thick vertical bars. The horizontal dashed line indicates the mean amplitude of the envelope.

4.3. The priming effects of the voice

When we consider syllables one and two separately, there is a small but significant release from masking when a different-sentence prime precedes a speech masker. How can the voice alone (irrespective of the content of the utterance) provide a release from speech masking? One possibility is that knowledge of the vocal characteristics of the target talker is useful in parsing an auditory scene because it helps the listener identify and track the target talker when there are other competing talkers. However, when the masker is noise and the target is speech, knowledge of the characteristics of the speaker's voice is unlikely to further improve performance since the target voice is already easy to track in a background of stationary noise. On the other

Author's personal copy

hand, knowledge of the content of the sentence can provide certain cues at a cognitive level (such as knowing when the target word will occur) when the key word is not completely masked by stationary noise (see Fig. 2).

Hence, based on the present results we propose that a same-sentence prime exerts its effect because knowledge of the content of the first part of the sentence aids word recognition at a higher-order cognitive level. Specifically, knowledge of the first part of the semantically-anomalous sentences allows the individual to track the voice that is producing these sentences. As reported by Freyman et al. (2004), using the target-talker's voice (female), a male's voice, or a print form to present the same-sentence prime causes the same amount of improvement in recognizing the last key word in the full target speech sentence (about 4 dB) when the masker was two-talker speech, indicating that the effect is just a content-cuing effect. Because this cuing effect appears to be independent of the voice of the speaker and the mode of presentation (auditory and visual primes both lead to unmasking), the cuing stimulus is clearly exerting its influence at a central (cognitive) level rather than at an auditory level. However, the current study shows that priming the listener with the voice of the target sentence (by presenting a different-sentence in the same voice) leads to an equivalent amount of unmasking for Chinese listeners. Hence, knowledge of the characteristics of a speaker's voice facilitates word recognition because it leads to better segregation of the target talker's voice from competing talkers' voices at a perceptual level.

Hence, to fully understand the nature of masking in Chinese as opposed to English, we need to examine how structural differences between Chinese and English can affect the degree to which listeners in the two languages may benefit from factors which should produce a release from informational masking. Clearly, more work is needed with respect to this issue.

The present study also shows that when no priming was given, the threshold for recognizing the last key word was over 1 dB lower for a speech masker than for a noise masker. One might have expected a greater degree of masking by a speech masker than by a noise masker (Freyman et al., 1999, 2004; Li et al., 2004), or at least equivalent degree of masking by these two maskers (Wu et al., 2005), since the speech masker has both energetic and informational masking effects and the noise masker has energetic masking only. However, a greater degree of fluctuation in the envelope of the Chinese speech masker than in the English speech masker may have made it easier for the Chinese participants to extract target information (see below). The reason for this is that it has been shown that listeners can benefit from troughs (temporal gaps) in the masker when listening to speech (Gustafsson and Arlinger, 1994; Howard-Jones and Rosen, 1993; Nelson et al., 2003; Summers and Molis, 2004). If the Chinese speech masker used here has deeper and wider troughs than the English speech masker, Chinese listeners will have a greater opportunity to benefit from listening in the troughs than do English listeners. Indeed, a comparison of deep troughs of envelopes between the Chinese two-talker speech masker used in the present study and the English two-talker speech masker (Freyman et al., 2001, 2004; Li et al., 2004) indicates that there appears to be a greater degree of amplitude modulation in the Chinese envelope than in the English envelope, and the duration of the Chinese troughs appear to be longer than those of the English troughs.[1] Hence, Chinese listeners might find it easier to hear the target speech in the presence of competing speech than in an equivalent level of stationary noise because of the greater depth and duration of the troughs in the Chinese speech maskers employed here than in the English speech maskers used in previous studies (Freyman et al., 2001, 2004; Li et al., 2004). It is important to note, however, that a number of factors, such as speech rate, will affect the frequency and depth of troughs in a language. Hence, all we can say is that the Chinese speech masker employed here had deeper troughs than the English speech masker (see Rhebergen and Versfeld, 2005 and Rhebergen et al., 2006, for a discussion of the role of troughs in the masking of speech by speech).

Any manipulation, as long as it helps distinguish the target and direct selective attention toward the target, will help segregate target speech from competing speech (Brungart, 2001; Freyman et al., 2004; Kidd et al., 2005a,b). Brungart and colleagues (Brungart, 2001; Brungart et al., 2001) reported that when a target phrase was masked by one or more competing phrase maskers, informational

[1] To look for peaks and troughs, we full-wave rectified 47 second samples from both the two-talker English speech masker, which were used in the study by Freyman et al. (2001, 2004) and that by Li et al. (2004), and the two-talker Chinese speech masker used in the present study, before passing them through a 20 Hz filter to extract the amplitude envelopes of both English and Chinese speech maskers. These amplitude envelopes were then smoothed using an r-term moving average filter provided by Mathematica (Wolfram Research, $r = 500$ samples). The smoothed samples were then fit using a quadratic interpolation function (Mathematica, Wolfram Research). This interpolation function was then differentiated to find the locations where the derivative of the interpolated function was zero, i.e., the locations of the peaks and troughs in the amplitude envelope. In the fashion troughs in the amplitude envelope were identified.

Troughs in the speech masker are more likely to be useful in listening to the target speech, the deeper, wider, and more frequent they are. We first searched for troughs that were more than 6 dB below the mean amplitude of the envelope. To define the width of these deep troughs, we started at the bottom of the trough and looked at the samples before it until we encountered the closest sample that was more than 3 dB above the floor of the trough. The time at which this sample was taken was defined as the lower boundary of the trough. The upper boundary of the through was obtained by examining successive samples following the bottom of the trough until we encountered a sample that was more than 3 dB above the floor of the trough. The time at which this sample was taken defined the upper boundary of the trough. The difference between the upper and lower boundaries was taken as the width of a trough. In the case that two troughs overlapped, the upper boundary of the first trough became the lower boundary of the second trough to avoid double counting of time spent in a deep trough. Fig. 8 shows the amplitude envelope for a segment of the Chinese speech masker, and identifies the location and widths of troughs. The total amount of time in a deep trough was 19% for the Chinese masker but only 10% for the English masker.

Author's personal copy

rather than energetic masking dominated performance, and the amount of masking was highly dependent on the similarity of the target and masker voices. The results suggest that even the voice of the target talker can have a cuing effect on recognizing the target speech sentence in the presence of speech masker. Specifically, the present study shows that presenting a different-sentence prime using the target-talker's voice can significantly improve recognition of the last key word in the full-length sentence only when the masker is two-talker speech. Therefore, in addition to perceived spatial separation (Freyman et al., 1999, 2001; Li et al., 2004; Wu et al., 2005), a priori knowledge about target location (Kidd et al., 2005b), and the informational content of the prime (Freyman et al., 2004; the present study), knowledge of the target-talker's voice can assist listeners' speech communication in the presence of masking speech when the language was Chinese. It would be interesting to see whether there is an equivalent effect of voice for English listeners.

It is important to note that the effect of a different-sentence or same-sentence prime did not depend on the order in which conditions were experienced. We would have expected such order effects if the prime exerted its effect primarily by familiarizing the listener with the target-talker's vocal characteristics. For if that were the case, we would expect priming to produce a larger release from masking when the no-priming condition preceded the two priming conditions than when the no-priming condition followed the two priming conditions. In the former case, the listener would have no exposure prior to experiencing the no-priming condition and therefore might be expected to show a larger release from masking than in the latter case where the amount of exposure to the target-talker's voice would be extensive before the no-priming condition was experienced. However, because there were no order effects,[2] it is unlikely that it is the total duration of exposure to the talker's voice influenced the amount of release from masking.

5. Summary and conclusions

Presenting a different Chinese sentence spoken by the target talker before the target speech was presented facilitated listeners' recognition of each of the last key syllables when the masker was speech but not when the masker was noise. Moreover, presenting Chinese target speech without the last key word before presenting the full target sentence also facilitated listeners' recognition of the last two syllables and the whole word, but this facilitation effect was smaller when the masker was noise. Thus, a priori knowledge of the talker's voice and/or the content of the target speech improves speech recognition in a Chinese "cocktail-party" environment.

Acknowledgments

We are grateful to Hua Shu and Yuan-Shan Cheng for insightful comments and discussion, to Xian Liu for technical support, and to Wen-Jie Wang and Meng-Yuan Wang for data collection. This work was supported by the National Natural Science Foundation of China (30670704; 60605016; 60535030; 60435010), the National High Technology Research and Development Program of China (2006AA01Z196; 2006AA010103), the Trans-Century Training Program Foundation for the Talents by the State Education Commission, "985" grants from Peking University, and the Natural Sciences and Engineering Research Council of Canada.

Appendix A

In fitting the psychometric functions we determined the values of μ and σ that minimized the Pearson χ^2 measure of goodness of fit, where

χ^{2}Tj /X1 1 Tf86.4755 0 0 6.4755 316.516oun0235 4eter991

Author's personal copy

where $p1_i$ and $p2_i$ are the probabilities of getting syllables one and two correct, respectively, when the sentences are presented at SNR i. Values of $p1_i$ and $p2_i$ were determined that minimized this χ^2. The number of degrees of freedom at each level i is 1 because there are four mutually-exclusive categories (3 degrees of freedom), and we fit two parameters at each level of SNR leaving 1 degree of freedom for each SNR level, and 4 degrees of freedom in total.

References

Arbogast, T.L., Mason, C.R., Kidd, G., 2002. The effect of spatial separation on informational and energetic masking of speech. J. Acoust. Soc. Amer. 112, 2086–2098.

Assmann, P.F., Summerfield, Q., 1989. Modeling the perception of concurrent vowels – vowels with the same fundamental-frequency. J. Acoust. Soc. Amer. 85, 327–338.

Brungart, D.S., 2001. Informational and energetic masking effects in the perception of two simultaneous talkers. J. Acoust. Soc. Amer. 109, 1101–1109.

Brungart, D.S., Simpson, B.D., Ericson, M.A., Scott, K.R., 2001. Informational and energetic masking effects in the perception of multiple simultaneous talkers. J. Acoust. Soc. Amer. 110, 2527–2538.

Brungart, D.S., Simpson, B.D., 2002. The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal. J. Acoust. Soc. Amer. 112, 664–676.

Darwin, C.J., Hukin, R.W., 2000. Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. J. Acoust. Soc. Amer. 107, 970–977.

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

Durlach, N.I., Mason, C.R., Shinn-Cunningham, B.G., Arbogast, T.L., Colburn, H.S., Kidd, G., 2003. Informational masking: Counteracting the effects of stimulus uncertainty by decreasing target-masker similarity. J. Acoust. Soc. Amer. 114, 368–379.

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

Freyman, R.L., Balakrishnan, U., Helfer, K.S., 2001. Spatial release from informational masking in speech recognition. J. Acoust. Soc. Amer. 109, 2112–2122.

Freyman, R.L., Balakrishnan, U., Helfer, K.S., 2004. Effect of number of masking talkers and auditory priming on informational masking in speech recognition. J. Acoust. Soc. Amer. 115, 2246–2256.

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

Gustafsson, H.A., Arlinger, S.D., 1994. Masking of speech by amplitude-modulated noise. J. Acoust. Soc. Amer. 95, 518–529.

Helfer, K.S., 1997. Auditory and auditory–visual perception of clear and conversational speech. J. Sp. Lan. Hear. Res 40, 432–443.

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

Kang, J., 1998. Comparison of speech intelligibility between English and Chinese. J. Acoust. Soc. Amer. 103, 1213–1216.

Kidd Jr., G., Mason, C.R., Gallun, F.J., 2005a. Combining energetic and informational masking for speech identification. J. Acoust. Soc. Amer. 118, 982–992.

Kidd Jr., G., Arbogast, T.L., Mason, Sopt19T-508.2(Gallu4,)-29292(F.J5l)-30292(b