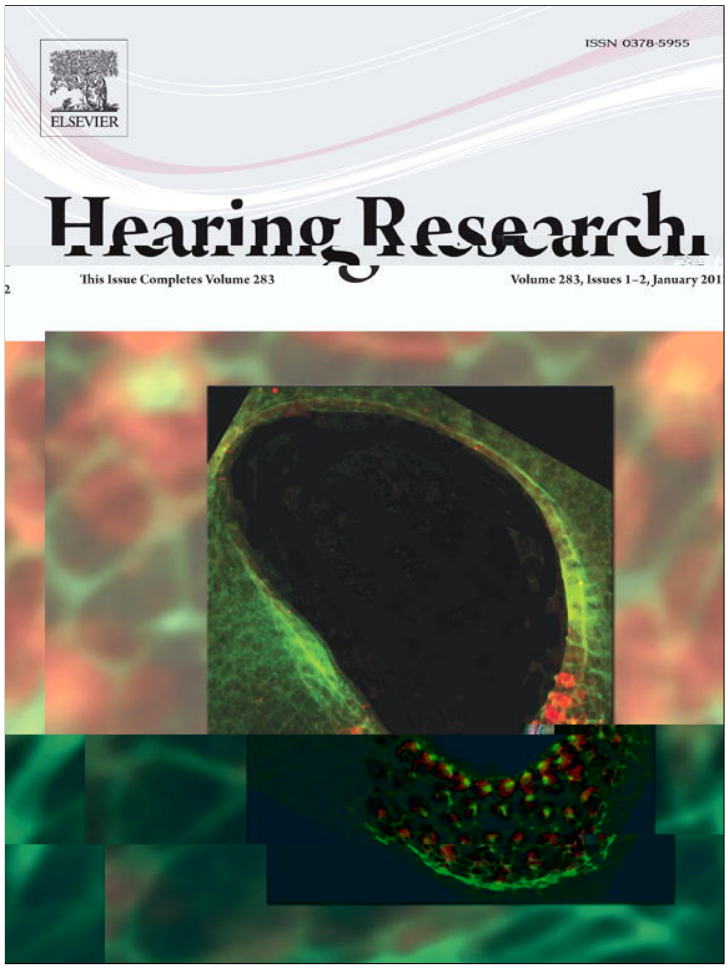

Hearing Research 283 (2012) 136–143

Contents lists available at SciVerse ScienceDirect

Hearing Research

journal homepage: www.elsevier.com/locate/heares

Research paper

Adding irrelevant information to the content prime reduces the prime-induced unmasking effect on speech recognition

Meihong Wu [a,b], Huahui Li [a,b], Yayue Gao [a,b], Ming Lei [a,b], Xiangbin Teng [a,b], Xihong Wu [a,b,*], Liang Li [a,b,**]

[a] Department of Psychology, Speech and Hearing Research Center, Key Laboratory on Machine Perception (Ministry of Education), Peking University, Beijing 100871, China
[b] Department of Machine Intelligence, Speech and Hearing Research Center, Key Laboratory on Machine Perception (Ministry of Education), Peking University, Beijing 100871, China

ARTICLE INFO

Article history:
Received 22 March 2011
Received in revised form
30 October 2011
Accepted 1 November 2011
Available online 10 November 2011

ABSTRACT

Presenting the early part of a nonsense sentence in quiet improves recognition of the last keyword of the sentence in a masker, especially a speech masker. This priming effect depends on higher-order processing of the prime information during target-masker segregation. This study investigated whether introducing irrelevant content information into the prime reduces the priming effect. The results showed that presenting the first four syllables (not including the second and third keywords) of the three-keyword target sentence in quiet significantly improved recognition of the second and third keywords in a two-talker-speech masker but not a noise masker, relative to the no-priming condition. Increasing the prime content from four to eight syllables (including the first and second keywords of the target sentence) further improved recognition of the third keyword in either the noise or speech masker. However, if the last four syllables of the eight-syllable prime were replaced by four irrelevant syllables (which did not occur in the target sentence), all the prime-induced speech-recognition improvements disappeared. Thus, knowing the early part of the target sentence mainly reduces informational masking of target speech, possibly by helping listeners attend to the target speech. Increasing the informative content of the prime further improves target-speech recognition probably by reducing the processing load. The reduction of the priming effect by adding irrelevant information to the prime is not due to introducing additional masking of the target speech.

1. Introduction

To improve their recognition of target speech in a noisy environment with multiple people talking, listeners use perceptual/cognitive cues to facilitate perceptual segregation of the target and masker, largely by strengthening their selective attention to the target speech. Some of the cues do not (substantially) change energetic masking of the target speech. Energetic masking is produced when the masker occupies peripheral resources for processing the target (see Helfer and Freyman, 2009). Cues that do not affect energetic masking include precedence-effect-induced spatial separation between the target image and the masker image (Freyman et al., 1999; Huang et al., 2008, 2009; Li et al., 2004; Rakerd et al., 2006; Wu et al., 2005), prior knowledge about where and/or when the target speech will occur (Best et al., 2007, 2008; Kidd et al., 2005), knowledge/familiarity of the target talker's voice (Brungart et al., 2001; Helfer and Freyman, 2009; Huang et al., 2010; Newman and Evers, 2007; Yang et al., 2007), prior knowledge about the topic of the target sentence (Helfer and Freyman, 2008), and visual information from the talker's face (Grant and Seitz, 2000; Helfer and Freyman, 2005; Rosenblum et al., 1996; Rudmann et al., 2003; Sumby and Pollack, 1954; Summerfield, 1979). It appears that any perceptual/cognitive cues, if they facilitate the listeners' selective attention to the target speech, can improve recognition of the target speech in a masker, especially for speech-masker-induced informational masking, which is caused by confusion between the target and masker and/or uncertainty regarding the target (Helfer and Freyman, 2009) (for further discussion of the concept of informational masking, see Arbogast et al., 2002; Agus et al., 2009; Freyman et al., 1999; Kidd et al., 2005; Schneider et al., 2007).

In addition to the cues mentioned above, presenting the early part of a target sentence (called the content prime) improves listeners' recognition of the later part of the target sentence in

* Corresponding author. Department of Machine Intelligence, Speech and Hearing Research Center, Key Laboratory on Machine Perception (Ministry of Education), Peking University, Beijing 100871, China.
** Corresponding author. Department of Psychology, Speech and Hearing Research Center, Key Laboratory on Machine Perception (Ministry of Education), Peking University, Beijing 100871, China. Tel.: +1 905 569 4628; fax: +1 905 569 4850
E-mail addresses: wxh@cis.pku.edu.cn (X. Wu), liangli@pku.edu.cn (L. Li).

0378-5955/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.heares.2011.11.001

a masker. More specifically, when either a noise or speech masker is present, recognition of the last (third) keyword in a three-keyword semantically anomalous (nonsense) target sentence is improved if the content prime, an early segment of this sentence (including the first and second keywords of the target sentence), is presented in quiet (Ezzatian et al., 2011; Freyman et al., 2004; Yang et al., 2007). Since the target sentences used in these studies are meaningless (nonsense), listeners receive no contextual support from the content prime for recognizing the last keyword. Moreover, the priming benefit is much larger when the masker is speech than when it is noise (Ezzatian et al., 2011; Freyman et al., 2004; Yang et al., 2007). Thus, Freyman et al. (2004) suggest that the prime helps the listener to extract the target auditory "object" out of the mixture of three talkers and makes it easier to attend to the target words and ignore the jumbled utterances of the other two talkers.

It should be noted that the priming effect depends on both a memory resource that holds the prime information during the target/masker co-presentation and a perceptual process comparing the prime content with the content of the later part of the sentences. Thus, adding irrelevant syllables (which do not appear in the target speech) to the prime may increase the processing load and/or introduce a disruption of the relevant information in the prime, resulting in a reduction of the priming effect. This study investigated whether introducing irrelevant content information in the prime reduces the priming effect. This issue has not been addressed in previous studies of the content-priming effect (Ezzatian et al., 2011; Freyman et al., 2004; Yang et al., 2007).

2. Methods

2.1. Participants

Twenty-four Mandarin Chinese-speaking university students (15 females and 9 males, mean age = 24.0 yrs, range 20–27) participated in this study. All the participants had symmetrical hearing (no more than a 15-dB difference between the two ears) and pure-tone hearing thresholds no more than 25 dB HL between 0.125 and 8 kHz. The participants gave their written informed consent and were paid a modest stipend for their participation.

The participant was seated at the center of an anechoic chamber (Beijing CA Acoustics Co. Ltd, Beijing, China), which was 560 cm in length, 400 cm in width, and 193 cm in height. All acoustic signals were digitized at a sampling rate of 22.05 kHz using a 24-bit Creative Sound Blaster PCI128 with a built-in anti-aliasing filter (Creative Technology, Ltd., Singapore) and were edited using Cooledit Pro 2.0, under the control of a computer with a Pentium IV processor (Intel Corporation, Santa Clara, California, USA). The acoustic analog outputs were delivered to a loudspeaker (Dynaudio Acoustics, BM6 A, Dynaudio, Risskov, Denmark) at 0° azimuth and elevation relative to the participant. 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.

2.2. Apparatus and stimuli

The speech stimuli were Chinese "nonsense" sentences. These sentences are syntactically correct but not semantically meaningful. 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, 2004), Li et al. (2004), and Ezzatian et al. (2011). The sentences have a *subject–predicate–object* structure and provide no contextual support for recognizing keywords. Each sentence has 12 characters (also 12 syllables) including the subject (first), predicate (second), and object (third) keywords with two characters (syllables) for each. For example, the English translation of one Chinese nonsense sentence is "*This polyester will expel that stomach*" (the keywords are underlined). The development of the Chinese nonsense sentences has been described elsewhere (Yang et al., 2007).

In the present study, a large number of nonsense-sentence stimuli were required. To satisfy this requirement, and to guarantee both high quality and uniformity of the acoustical features of the stimuli, both target and priming speech were spoken by three different artificially synthesized young-female voices. The speech masker was a 47-s loop of digitally combined continuous recordings of Chinese nonsense sentences (whose keywords did not appear in the target sentences) spoken by two other young-female talkers, and the noise masker was a steady speech-spectrum one (Yang et al., 2007).

Sounds were calibrated using a Larson Davis Audiometer Calibration and Electroacoustic Testing System (AUDit and System 824, Larson Davis, USA) whose microphone was placed at the center position of the participant's head when the participant was absent, using a "slow"/"RMS" meter response. The levels of both prime and target sounds were set to 60 dBA, and the masker pressure level was adjusted to produce four signal-to-noise ratios (SNRs): −8, −4, 0, and 4 dB.

2.3. Speech synthesis

Speech synthesis based on the Hidden Markov Model (HMM) has been successfully used for text-to-speech transformation (converting written text into audible speech) (Yoshimura et al., 1999; Cao et al., 2011; Zen et al., 2007a,b). Since some speech-acoustical parameters can be modeled and modulated using the HMM, voice characteristics can also be added into the artificially synthesized speech signals. In this study, the target speech and priming speech for the three prime/target voices were generated using a HMM-based system. A Chinese corpus including 6000 sentences with a news-broadcast style, which was both phonetically and prosodically rich, was downloaded from the website (http://www.synsig.org/index.php/BlizzardChallenge2009, see King and Karaiskos, 2009) and sampled for model training with a sampling rate of 16 kHz. Using the method developed by Zen et al. (2007a,b), some critical parameters of speech features (including the mel-cepstrum, log F_0, and band aperiodicity measures) were extracted. At the synthesis stage, the speech-parameter sequence for each sentence stimulus was generated from the corresponding HMMs. Then, using the method developed by Fukada et al. (1992), a speech waveform was synthesized using the algorithm of the Mel Log Spectrum Approximation Filter with the generated parameters. Finally, the initial acoustical model was established by a training procedure using the speech corpus with the voice of a selected female Talker (Talker O).

Speech samples (about 600 sentences and lasting 40 min) of each of the three young-female speakers (t_1, t_2, and t_3) were processed by the initial acoustic model as described above to obtain the acoustical model for each of the three prime/target voices. Consequently, for each of the prime/target voices, using the resultant target-voice acoustical models, written nonsense sentences were transformed into signals with the speaker's vocal characteristics. Finally, to equalize speech rates across the three synthesized voices, rate and other temporal information from Talker O were used to modulate the prime/target-voice models of the three prime/target voices, resulting in speech rates that were identical across the different synthesized speech samples.

2.4. Testing procedures

There were 32 (2 masker types: noise, speech; 4 priming types: no-priming, four-syllable-priming, true eight-syllable-priming, mixed eight-syllable-priming; 4 SNRs: −8 dB, −4 dB, 0 dB, 4 dB) conditions and 18 target sentences (one list) were used for each condition. The 8 masker/prime combinations were counterbalanced across 24 participants using a Latin square design, and the four SNRs were arranged randomly for each masker/prime combination.

To balance the amount of information across stimulus conditions, the amount of information of a keyword in a sentence was calculated as

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

where f is word frequency. The amount of information in a sentence was the sum of that for the three keywords. The sentences in each list were chosen in such a way that the amount of information in each list was roughly constant (Yang et al., 2007).

For each testing session, participants were informed of both the masker type (noise or speech) and the priming condition (no-priming, four-syllable-priming, true eight-syllable-priming, or mixed eight-syllable-priming). Under the condition with prime presentation, the participant pressed a button on a response box to start a trial. A masker (noise or speech) was then gated on about 200 ms after the prime presentation. To approximately equalize the onset-to-onset interval between the prime and masker presentations, the four-syllable prime was followed by a silent period with a duration of about four syllables (800–900 ms) plus 200 ms before the masker started (Fig. 1). Under the no-priming (baseline) condition, a masker (r)16(o)29(xi)9inr-16(ing)1meitasy a120(3(er)-244(t1617(e)-328(f1n)28(t)15(on)-263(o1428(es)13(sss)-2.((r)16Thr)-285(TJ8(f1-35ed)-

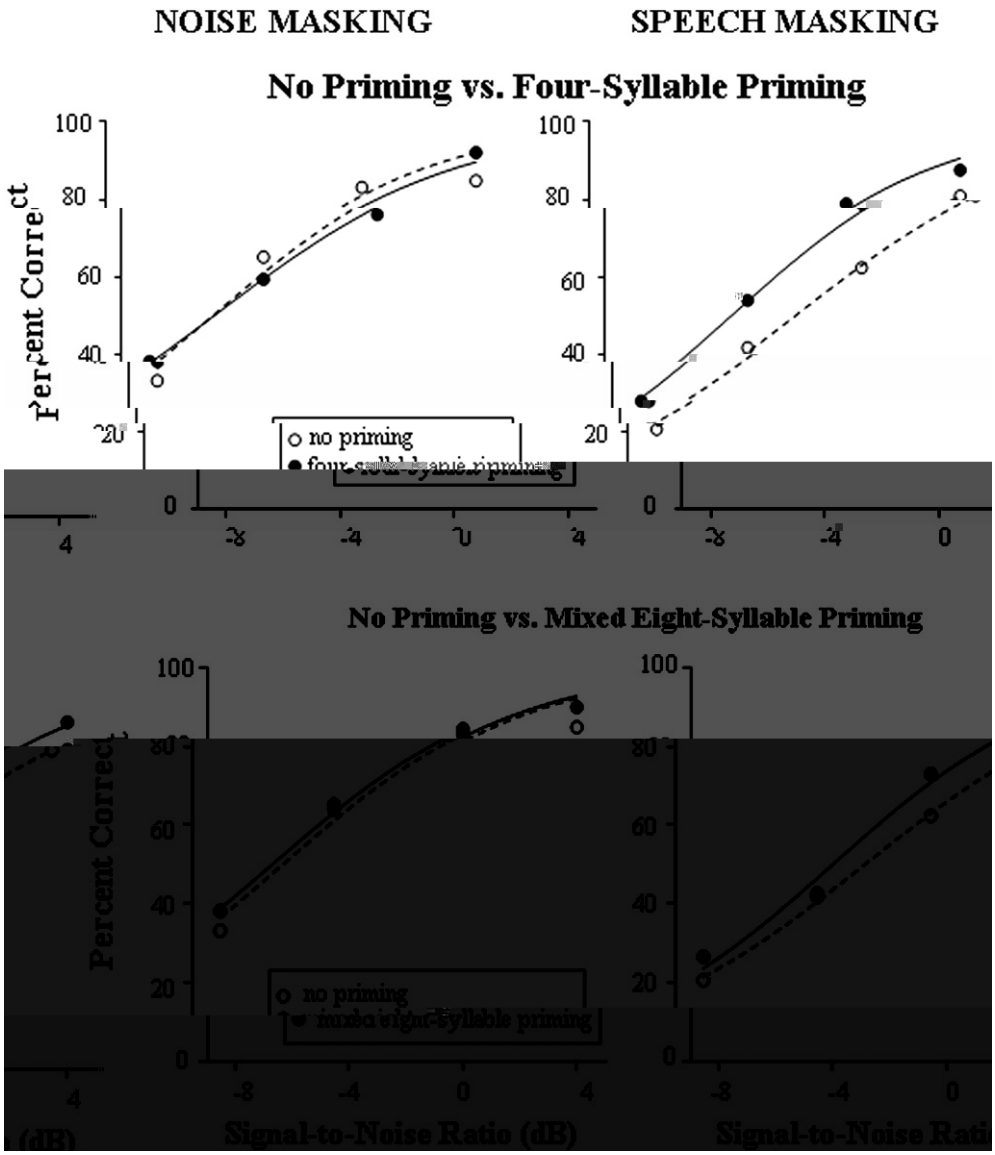

Fig. 2. Group-mean percent-correct identification of the second target keyword across 24 participants along with the best-fitting psychometric functions as a function of SNR for the three priming conditions (no-priming, four-syllable-priming, mixed eight-syllable-priming) under which the second keyword did not occurr in the prime. The masker was either steady speech-spectrum noise (left panels) or two-talker speech (right panels).

(SNR) within-subject ANOVA showed that all main effects were significant (all $p < 0.05$), all the interactions between masker type and other factors were significant (all $p \leq 0.001$), but the interaction between priming condition and SNR was not significant ($p = 0.320$).

The differences in recognizing the second keyword across priming conditions can be more concisely examined by analyzing the differences in μ when the masker was either noise or speech. When the masker was noise, a one-way ANOVA showed that the priming effect on μ was not significant ($F_{2,\ 46} = 1.008$, $p = 0.373$), suggesting that presenting either the four-syllable prime or the mixed eight-syllable prime did not affect recognition of the second keyword under the noise-masking condition. When the masker was speech, a one-way ANOVA showed that the priming effect on μ was significant ($F_{2,46} = 7.008$, $p = 0.002$). Post hoc analyses with the adjusted α of 0.05/3 showed that the threshold under the four-syllable-priming condition was significantly different from that under the no-priming condition ($p < 0.001$), but there was no significant difference in μ between the no-priming condition and the mixed eight-syllable-priming condition ($p = 0.048$). Also, the threshold under the four-syllable-priming condition was not significantly different from that under the mixed eight-syllable-priming condition ($p = 0.267$). The results suggest that presenting the four-syllable prime improved recognition of the second keyword for the speech masker but not for the noise masker. The improvement in threshold induced by the four-syllable prime was 2.2 dB. However, adding four irrelevant syllables to the four-syllable prime (giving the mixed eight-syllable prime) appeared to reduce the four-syllable-prime-induced improvement in recognizing the second keyword, because the mixed eight-syllable prime did not induce significant improvement relative to the no-priming condition.

3.2. Recognition of the third keyword

The third (last) keyword did not occur in any of the primes. Fig. 4 illustrates group-mean percent-correct syllable identification for the third keyword as a function of SNR under the no-priming condition (open circles) or one of the three conditions with prime

presentation (filled circles), when the masker was noise (left panels) or speech (right panels). The group-mean best-fitting psychometric functions are also presented. Fig. 5 compares the group-mean μ values for recognizing the third keyword across priming conditions for each of the two masker types.

When the masker was noise (left panels), only the true eight-syllable prime appeared to induce an improvement in recognition of the third keyword. When the masker was speech, both the four-syllable prime and the true eight-syllable prime, but not the mixed eight-syllable prime, improved recognition. A 2 (masker type: noise, speech) $\times$ 4 (priming condition: no-priming, four-syllable-priming, eight-syllable-priming, mixed eight-syllable-priming) $\times$ 4 (SNR) within-subject ANOVA showed that the interaction between masker type and priming condition was significant ($F_{3,69} = 4.757$, $p = 0.005$), but no other interactions were significant ($p \geq 0.164$). The differences in μ across priming conditions were examined for each of the masker types.

When the masker was noise, a one-way ANOVA showed that the priming effect on μ was significant ($F_{3,\ 69} = 4.563$, $p = 0.006$). Post hoc analyses with the adjusted α of 0.05/6 showed that the threshold under the true eight-syllable-priming condition was significantly different from those under the no-priming ($p = 0.008$), four-syllable-priming ($p = 0.007$), and mixed eight-syllable-priming ($p = 0.006$) conditions. There were no significant differences in μ between the no-priming, four-syllable-priming, and mixed eight-syllable-priming conditions ($p > 0.100$ for all). The results indicate that the eight-syllable prime, but not the four-syllable prime or the mixed eight-syllable prime, significantly improved recognition of the third keyword under noise masking ($\Delta\mu = 1.0$ dB).

When the masker was speech, a one-way ANOVA showed that the priming effect on μ was significant ($F_{3,\ 69} = 16.905$, $p < 0.001$). Post hoc analyses with the adjusted α of 0.05/6 showed that the threshold under the true eight-syllable-priming condition was significantly different from those under the no-priming ($p < 0.001$) and mixed eight-syllable-priming ($p < 0.001$) conditions. Also, the threshold under the four-syllable-priming condition was significantly different from those under the no-priming ($p = 0.001$) and mixed eight-syllable-priming ($p = 0.007$) conditions. There was no significant difference in μ between the no-priming condition and mixed eight-syllable-priming condition ($p = 0.075$), and there was no significant difference between the true eight-syllable-priming condition and four-syllable-priming condition ($p = 0.042$). The results indicate that presenting either the four-syllable prime or the true eight-syllable prime improved recognition of the third keyword under speech masking. The improvements induced by the four-syllable prime and the true eight-syllable prime were 1.3 and 2.0 dB, respectively. However, under the mixed eight-syllable-priming condition, no significant priming effect occurred.

4. Discussion

4.1. The content-priming effect was both masker-type and content-amount dependent

In the present study, both the target sentence and masker (noise or speech) were presented from the same loudspeaker in front of the participant. When the participant tried to attend to the target speech, spatial attention could not be used to focus on the target. In other words, due to the co-location, the target and masker shared the listeners' spatial attention.

The results of this study showed that when the masker was speech, relative to the no-priming condition, presenting the four-syllable prime in quiet significantly improved recognition of the second and third keywords, with a reduction of the threshold μ by 2.2 dB for the second keyword and 1.3 dB for the third keyword. However, no significant improvement for recognizing either of the two keywords occurred when the masker was noise. Also, relative to the no-priming condition, when the length of the content prime increased from four syllables to eight syllables (under the true eight-syllable-priming condition), recognition of the third keyword was significantly improved with a reduction of μ by 2.0 dB when the masker was speech and 1.0 dB when the masker was noise.

These results confirm that prior knowledge (memory) of the early part of a sentence facilitates listeners' selective attention to the later part of the sentence, thereby improving their recognition of speech in a masker (Ezzatian et al., 2011; Freyman et al., 2004; Yang et al., 2007). Since presenting the prime does not influence the acoustical signals during target/masker co-presentation, the prime-induced release of target speech from either speech masking or noise masking is dependent on higher-order processes.

In this study, presenting the four-syllable prime led to a significant improvement in recognizing the second and third keywords under the speech-masking condition but not the noise-masking condition. Also, presenting the true eight-syllable prime caused a larger improvement in recognizing the third keyword under the speech-masking condition than under the noise-masking condition. Thus, the results support previous findings that content priming mainly releases target speech from informational masking (Ezzatian et al., 2011; Freyman et al., 2004; Yang et al., 2007). Since presenting the true eight-syllable prime, but not the four-syllable prime, significantly reduced the noise-masking effect, the content amount of the prime appears to be a factor determining the degree and selectivity of the unmasking effect of the content prime.

Finally, in this study, the voice speaking the prime and the voice speaking the target were always different, and the potential voice-priming effect (Yang et al., 2007; Huang et al., 2010) was minimized. However, it should be noted that in a trial, after the participant

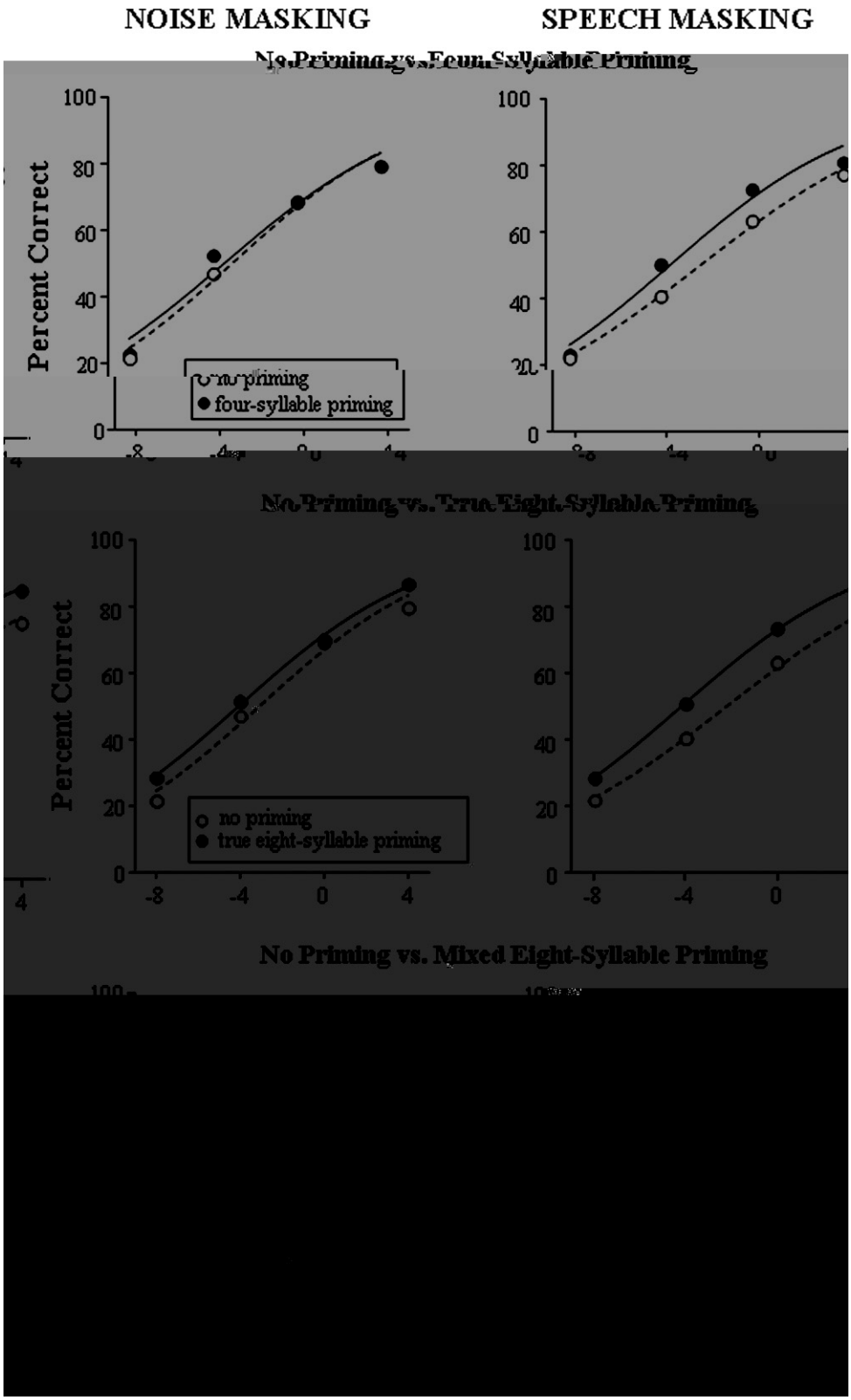

Fig. 4. Group-mean percent-correct identification of the third keyword across 24 participants along with the best-fitting psychometric functions as a function of SNR for the four priming conditions (no-priming, four-syllable-priming, true eight-syllable-priming, and mixed eight-syllable-priming), when the masker was either steady speech-spectrum noise (left panels) or two-talker speech (right panels).

participant using the target talker's vocal characteristics for interpreting later parts of the target speech cannot be discounted. In addition to vocal characteristics, the prosodic cues in spoken sentences help listeners to track a particular utterance over time (e.g., Darwin and Hukin, 2000a,b). Thus, since the speech stimuli used in this study were both phonetically and prosodically rich, it is

also possible that participants used prosodic cues provided by the prime to facilitate their attention to syllables at different temporal points in the target speech.

Based on these results, we propose that when the prime content is short, it mainly helps listeners attend to the target speech in the target/masker complex. Since the target speech and masking noise have distinctively different acoustical characteristics, listeners are able to quickly notice the difference. When the masker is noise, listeners can use both lower-order acoustic features and higher-order content knowledge to attend to the target speech. The absence of priming effects under the noise-masking condition when the prime was short suggests that listeners tended to use lower-order acoustic cues to follow speech in a noise masker. Thus, the four-syllable prime did not release the target speech from noise masking. However, when the masker was speech, since the young-female masking voices were similar to the young-female target voice, listeners judged it difficult to determine which voice belonged to the target. Presenting the prime helped listeners attend to the target stream (i.e., the target-orienting function), thereby reducing the speech-on-speech masking effect.

As suggested by Freyman et al. (2004), one of the possible mechanisms underlying the priming effect is that the prime decreases the memory load required for the words occurring in the prime and allows more resources to be brought to processing of the later words. Thus, in this study, when prime was made longer so that it contained both the first and second keywords, in addition to its target-orienting function, the prime may have reduced the processing load for recognizing and repeating the first 8 syllables in the sentence. Consequently, the participant was able to assign more processing resources to the last keyword. Note that the effects of increasing the prime length were general for both speech masking and noise masking.

4.2. Adding irrelevant information to the prime reduces the prime-induced unmasking effect

Importantly, the results of this study showed that relative to the no-priming condition, presenting the mixed eightthetltcmayedocesq4(eiuieech

References

Agus, T.R., Akeroyd, M.A., Gatehouse, S., Warden, D., 2009. Informational masking in young and elderly listeners for speech masked by simultaneous speech and noise. J. Acoust. Soc. Am. 126, 1926–1940.

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

Baddeley, A., 1981. The concept of working memory – a view of its current state and probable future-development. Cognition 10, 17–23.

Best, V., Ozmeral, E.J., Shinn-Cunningham, B.G., 2007. Visually-guided attention enhances target identification in a complex auditory scene. J. Assoc. Res. Otolaryngol. 8, 294–304.

Best, V., Ozmeral, E.J., Kopco, N., Shinn-Cunningham, B.G., 2008. Object continuity enhances selective auditory attention. Proc. Natl. Acad. Sci. U. S. A. 105, 13174–13178.

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. Am. 110, 2527–2538.

Cao, S.Y., Li, L., Wu, X.H., 2011. Improvement of intelligibility of ideal binary-masked noisy speech by adding background noise. J. Acoust. Soc. Am. 129, 2227–2236.

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

Darwin, C.J., Hukin, R.W., 2000b. Effects of reverberation on spatial, prosodic, and vocal-tract size cues to selective attention. J. Acoust. Soc. Am. 108, 335–342.

Ezzatian, P., Li, L., Pichora-Fuller, K., Schneider, B.A., 2011. The effect of priming on release from informational masking is equivalent for younger and older adults. Ear Hear. 32, 84–96.

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. Am. 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. Am. 106, 3578–3588.

Fukada, T., Tokuda, K., Kobayashi, T., Imai, S., 1992. An adaptive algorithm for mel-cepstral analysis of speech. Proc. ICASSP, 137–140.

Grant, K.W., Seitz, P.F., 2000. The use of visible speech cues for improving auditory detection of spoken sentences. J. Acoust. Soc. Am. 108, 1197–1208.

Hasher, L., Zacks, R.T., 1988. Working memory, comprehension, and aging: a review and a new view. In: Bower, G.H. (Ed.), The Psychology of Learning and Motivation: Advances in Research and Theory, vol. 22. Academic Press, San Diego, CA, pp. 193–225.

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

Helfer, K.S., Freyman, R.L., 2005. The role of visual speech cues in reducing energetic and informational masking. J. Acoust. Soc. Am. 117, 842–849.

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

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

Huang, Y., Huang, Q., Chen, X., Qu, T.S., Wu, X.H., Li, L., 2008. Perceptual integration between target speech and target-speech reflection reduces masking for target-speech recognition in younger adults and older adults. Hear. Res. 244, 51–65.

Huang, Y., Huang, Q., Chen, X., Wu, X.H., Li, L., 2009. Transient auditory storage of acoustic details is associated with release of speech from informational masking in reverberant conditions. J. Exp. Psychol. Hum. Percept. Perform. 35, 1618–1628.

Huang, Y., Xu, L.J., Wu, X.H., Li, L., 2010. The effect of voice cuing on releasing speech from informational masking disappears in older adults. Ear Hear. 31, 579–583.

Kidd, G., Arbogast, T.L., Mason, C.R., Gallun, F.J., 2005. The advantage of knowing where to listen. J. Acoust. Soc. Am. 118, 3804–3815.

King, S., Karaiskos, V., 2009. The Blizzard Challenge 2009. In: Proc. Blizzard Challenge Workshop, Edinburgh, U.K.

Li, L., Daneman, M., Qi, J.G., Schneider, B.A., 2004. Does the information content of an irrelevant source differentially affect speech recognition in younger and older adults? J. Exp. Psychol. Hum. Percept. Perform. 30, 1077–1091.

Newman, R.S., Evers, S., 2007. The effect of talker familiarity on stream segregation. J. Phon. 35, 85–103.

Rakerd, B., Aaronson, N.L., Hartmann, W.M., 2006. Release from speech-on-speech masking by adding a delayed masker at a different location. J. Acoust. Soc. Am. 119, 1597–1605.

Rosenblum, L.D., Johnson, J.A., Saldana, H.M., 1996. Point-light facial displays enhance comprehension of speech in noise. J. Speech Hear. Res. 39, 1159–1170.

Rudmann, D.S., McCarley, J.S., Kramer, A.F., 2003. Bimodal displays improve speech comprehension in environments with multiple speakers. Hum. Factors 45, 329–336.

Schneider, B.A., Li, L., Daneman, M., 2007. How competing speech interferes with speech comprehension in everyday listening situations? J. Am. Acad. Audiol. 18, 578–591.

Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. 26, 212–215.

Summerfield, A.Q., 1979. Use of visual information for phonetic processing. Phonetica 36, 314–331.

Wolfram, S., 1991. Mathematica: A System for Doing Mathematics by Computer. Addison-Wesley, New York, pp. 1–961.

Wu, X.H., Wang, C., Chen, J., Qu, H.W., Li, W.R., Wu, Y.H., Schneider, B.A., Li, L., 2005. The effect of perceived spatial separation on informational masking of Chinese speech. Hear. Res. 199, 1–10.

Yang, Z.G., Chen, J., Wu, X.H., Wu, Y.H., Schneider, B.A., Li, L., 2007. The effect of voice cuing on releasing Chinese speech from informational masking. Speech Commun. 49, 892–904.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 1999. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. Proc. Eurospeech 5, 2347–2350.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A.W., Tokuda, K., 2007a. The HMM-based speech synthesis system (HTS) version 2.0. In: Proc. 6th ISCA Workshop Speech Synth. (SSW-6), Bonn, Germany, Aug.

Zen, H., Toda, T., Nakamura, M., Tokuda, K., 2007b. Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005. IEICE Trans. Inform. Systems E90-D (1), 325–333.