Te a e-e ed ead c e e ease eec f f a a a s

C a W , S a Ca , X W , a d L a L [a)]
Department of Psychology, Department of Machine Intelligence, Speech and Hearing Research Center, Key Laboratory on Machine Perception, Ministry of Education, Peking University, Beijing 100871, China
wuchaowfy@pku.edu.cn, daidaicsy@gmail.com, wuxihong.pku@gmail.com, liangli@pku.edu.cn

Ab ac : Listeners can use temporally pre-presented content cues and concurrently presented lipreading cues to improve speech recognition under masking conditions. This study investigated whether temporally pre-presented lipreading cues also unmask speech. In a test trial, before the target sentence was co-presented with the masker, either target-matched (priming) lipreading video or static face (priming-control) video was presented in quiet. Participants' target-recognition performance was improved by a shift from the priming-control condition to the priming condition when the masker was speech but not noise. This release from informational masking suggests a combined effect of working memory and cross-modal integration on selective attention to target speech.

PACS be 43.72.Dv, 43.66.Dc, 43.71.+m, 43.71.Gv [QJF]
Da e Rece ed: December 21, 2012 **Da e Acce ed:** February 26, 2013

1. I d c

Listeners can use various perceptual/cognitive cues to facilitate perceptual segregation between the target speech and the masker by strengthening their selective attention to the target speech (Du *et al.*, 2011). These cues include both viewing a speaker's movements of the speech articulators (simultaneous lipreading) (Summerfield, 1979; Rosenblum *et al.*, 1996; Grant and Seitz, 2000; Rudmann *et al.*, 2003; Helfer and Freyman, 2005) and prior knowledge about part of the target-sentence content (i.e., temporally pre-presented content prime) (Freyman *et al.*, 2004; Yang *et al.*, 2007).

Speech information contained in lipreading is both redundant and complementary to cues provided acoustically (Summerfield, 1979). For example, lipreading contains important phonetic information of speech, including that of vowels, diphthongs, and the place of articulation of consonants (Summerfield, 1992). In addition, the degree of mouth opening of the target talker is associated with the overall amplitude contour of target speech (Grant and Seitz, 2000; Summerfield, 1992). Moreover, the visual-auditory temporal synchrony (temporally co-modulated visual and auditory information) indicates the distinctive rate and dynamic phase of target-speech syllables, facilitates listeners' selective attention to the time windows that contain target syllables, and form the temporal expectation of the forthcoming components of the target stream (Wright and Fitzgerald, 2004).

On the other hand, for the content prime-induced unmasking of target speech, when either a noise masker or a speech masker is co-presented with a meaningless target sentence (with three keywords), recognition of the last (third) keyword in the target sentence is improved if the content prime, a segment from the start of the same sentence (e.g., including the first two keywords), is temporally pre-presented in quiet

[a)]Author to whom correspondence should be addressed.

before the co-presentation of the target and the masker (Freyman *et al.*, 2004; Yang *et al.*, 2007). Since the target sentences used in this line of studies are meaningless ("nonsense"), listeners receive no contextual support for recognizing the last keyword. As suggested by Freyman *et al.* (2004), the pre-presented content prime helps listeners focus attention more quickly on the target, thereby facilitating recognition of the last keyword in the target stream against speech *informational masking* that is caused by confusion between the target and masker and/or uncertainty regarding the target. Because the temporally pre-presented prime-content information needs to be held during the target/masker presentation, the content-priming effect must depend on a working memory resource.

Previous studies have shown that the recognition of target speech under either noise-masking or speech-masking conditions can be improved by either simultaneously presenting lipreading cues (e.g., Helfer and Freyman, 2005) or temporally pre-presenting verbal content cues (e.g., Freyman *et al.*, 2004; Yang *et al.*, 2007), even though the unmasking effect is larger under the speech masking condition than that under the noise-masking condition. Thus, either of the two types of unmasking cues is not completely specific for releasing target speech from informational masking. Since informational masking involves more higher-order and top-down processes, this study examined whether temporally pre-presented lipreading cues (whose processing involves both working memory and cross-modal translation/integration) are more specific for improving the speech recognition under speech-masking conditions. To date, this issue has not been addressed in the literature.

2. Me dS

2.1 Participants

Eighteen Mandarin Chinese–speaking university students (11 females and 7 males, mean age = 22.2 years with the range from 18 to 26 years) participated in this study. All the participants were right-handed, 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.

2.2 Equipment and materials

The participant was seated at the center of a sound-attenuated chamber (EMI Shielded Audiometric Examination Acoustic Suite). Acoustic signals, calibrated by a sound-level meter (AUDit and System 824, Larson Davis, USA), were transferred from a notebook computer sound card (ATI SB450 AC97) and bilaterally presented to participants using headphones (Model HAD 200) with a sound pressure level (SPL) of 60 dBA at each ear. The SPL of the masker was adjusted to produce four signal-to-masker ratios (SMRs): −8, −4, 0, and 4 dB.

Speech stimuli were Chinese nonsense sentences, which are syntactically correct but not semantically meaningful (Yang *et al.*, 2007). For example, the English translation of a Chinese nonsense sentence is "One *appreciation* could *retire* his *ocean*" (keywords are in italic). Each of the Chinese sentences has 12 syllables (also 12 characters) including three keywords with two syllables for each. The sentence frame cannot provide any contextual support for recognizing the keywords. The development of the Chinese nonsense sentences was described elsewhere (Yang *et al.*, 2007).

Target speech was spoken by three young adult female talkers (Talkers A, B, and C). In a trial, either the true priming-stimulus lip movement (whose content was identical to the target sentence) or the static face picture (priming control) was presented before the co-presentation of the target and the masker. To minimize the influence of gazing and/or facial identity processing, only the lower half of the face was presented.

The noise masker was a stream of steady-state speech-spectrum noise (Yang *et al.*, 2007). The speech masker was a 47-s loop of digitally combined continuous

recordings for Chinese nonsense sentences (whose keywords did not appear in target sentences) spoken by two other young adult female talkers (Talkers D and E) (Yang *et al.*, 2007).

2.3 Procedures

There were three within-subject variables: (1) masker type, (2) priming condition, and (3) SMR. Twelve trials (also 12 target sentences) were used in each condition. To avoid potential voice priming, each of the three target voices had the equal chance to be used for a condition, and the presentation order for the three target voices was arranged in random manner. The presentation order for the maskerltw7(cha2497.3erltin)-3erderDeachin therango

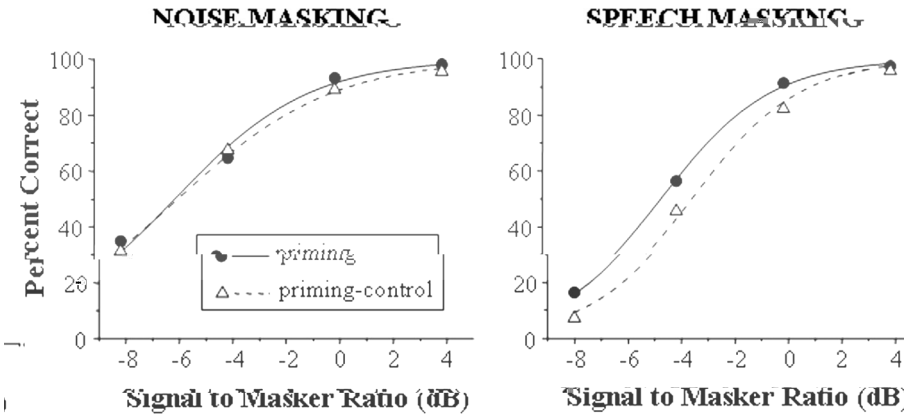

Fig. 1. Group-mean percent-correct recognition of the target keywords as a function of signal-to-masker ratio (SMR) along with the group-mean best-fitting psychometric functions under the priming-control condition (open triangles and dash curve) and the priming condition (filled circles and solid curve) when the masker

$p < 0.001$), and the interaction between prime type and SMR was significant ($F_{3, 136} = 3.244$, $p = 0.024$). Thus, presenting the lipreading prime significantly released target speech from speech masking but not noise masking.

Figure 2 shows group-mean percent-correct recognition of each of the three keywords as a function of the SMR along with the group-mean best-fitting psychometric functions when the masker was speech. Clearly, the lipreading-induced improvement of keyword recognition mainly occurred for the first keyword, especially when the SMR was low (−8 dB).

4. D[illegible]

The results of this study showed that recognition of the keywords, particularly the first keyword, in the target sentence was significantly better under the priming condition than that under the priming-control condition only when the masker was speech. Thus, this study for the first time presents evidence that listeners are able to use temporally pre-presented lipreading cues to facilitate recognition of target sentences (particularly in the early part of target sentences) against informational masking when the SMR was low.

As mentioned in Sec. 1, lipreading contains important phonetic information of speech (Summerfield, 1992), overall amplitude contour of speech (Grant and Seitz, 2000; Summerfield, 1992) and temporal dynamic information of speech (Wright and

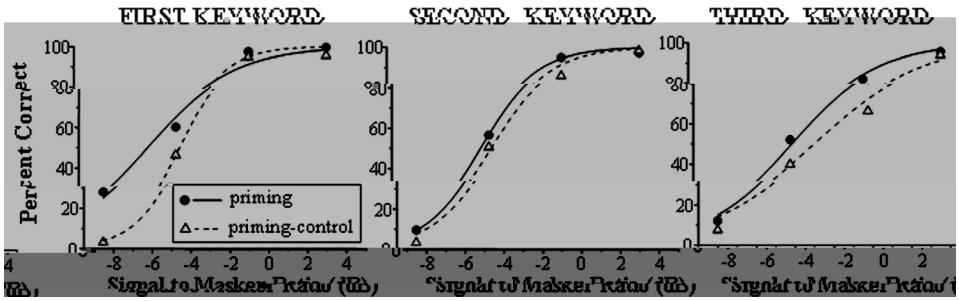

Fig. 2. Group-mean percent-correct recognition of each of the three target keywords as a function of SMR along with the group-mean best-fitting psychometric functions under the priming-control condition (open triangles and dash curve) and priming condition (filled circles and solid curve) when the masker was speech.

Fitzgerald, 2004). Also, temporally pre-presented content prime can release target speech from masking, especially speech masking (Freyman *et al.*, 2004; Yang *et al.*, 2007), indicating that the content information about part of the target sentence is retained in the working memory and helps listeners attend to the target stream. Similar to the unmasking effect of content primes (Freyman *et al.*, 2004; Yang *et al.*, 2007) the information on phonetics (Summerfield, 1992), overall amplitude contour (Grant and Seitz, 2000; Summerfield, 1992), and/or temporal dynamic features (Wright and Fitzgerald, 2004) of target speech is provided by the temporally pre-presented lipreading prime, retained in the working memory, and used by listeners to attend to the target speech that is co-presented with the speech masker. The larger facilitation of the recognition of the first keyword in the target sentence suggests that the working memory of the visual speech information continues to fade during the co-presentation of the target speech and the speech masker.

As mentioned in Sec. 1, neither the simultaneously presenting lipreading cue nor the temporally pre-presenting verbal content cue is specific for releasing target speech from informational masking. However, temporally pre-presenting the lipreading cue releases target speech from speech masking but not noise masking. Thus, the temporally pre-presenting lipreading cue is more useful for isolating informational masking from energetic masking. This masking-specific effect is based on the co-operation of working memory and audiovisual integration, involving higher-order and top-down processes.

Ac[illegible]ed[illegible]e[illegible]

This work was supported by the "973" National Basic Research Program of China (2009CB320901; 2013CB329304), the National Natural Science Foundation of China (31170985; 91120001, 61121002), the Ministry of Science and Technology of China (2010DFA31520), the National Health Public welfare Industry Research Project of China (201202001), and "985" grants from Peking University.

Refe[illegible]ce[illegible]a[illegible]d[illegible]

Du, Y., Kong, L., Wang, Q., Wu, X., and Li, L. (**2011**). "Auditory frequency-following response: A neurophysiological measure for studying the 'cocktail-party problem,'" Neurosci. Biobehav. Rev. **35**, 2046–2057.

Freyman, R. L., Balakrishnan, U., and 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.

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

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

Rosenblum, L. D., Johnson, J. A., and Saldana, H. M. (**1996**). "Point-light facial displays enhance comprehension of speech in noise," J. Speech. Lang. Hear. Res. **39**, 1159.

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

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

Summerfield, Q. (**1992**). "Lipreading and audio-visual speech perception," Philos. Trans. R. Soc. London Ser. B **335**, 71–78.

Wright, B. A., and Fitzgerald, M. B. (**2004**). "The time course of attention in a simple auditory detection task," Atten. Percept. Psychophys. **66**(3), 508–516.

Yang, Z., Chen, J., Huang, Q., Wu, X., Wu, Y., Schneider, B. A., and Li, L. (**2007**). "The effect of voice cuing on releasing Chinese speech from informational masking," Speech Commun. **49**, 892–904.