The Effect of Voice Cuing on Releasing Speech From Informational Masking Disappears in Older Adults

Ying Huang, Lijuan Xu, Xihong Wu, and Liang Li

Objective: To investigate whether older adults can use voice information to unmask speech.

Design: Under a voice-priming condition, before a target-speech sentence was presented with a noise or speech masker, one or two voice-priming sentences were recited with the same voice reciting the target sentence. Eighteen younger adults and 12 older adults with clinically normal hearing were instructed to loudly repeat the target sentence.

Results: Presenting the voice-priming sentence(s) improved target-speech identification only when the masker was speech in younger adults but not older adults.

Conclusion: For older adults, the inability to use voice information to reduce informational masking contributes to their speech-recognition difficulties in "cocktail-party" environments.

(Ear & Hearing 2010;31;579–583)

INTRODUCTION

In (simulated) cocktail-party environments, younger adult listeners with normal hearing are able to use various cues to facilitate their recognition of target speech against masking (Cherry 1953; Freyman et al. 1999, 2004; Li et al. 2004; Yang et al. 2007). It is well known that the talker's voice contains not only speech content information but also talker's identity information and affective information. One particular cue, which is associated with this study, is knowledge and/or familiarity of the voice of target speech. Specifically, in the study by Yang et al. (2007), immediately before the copresentation of a target speech sentence with a masker (either steady-state speech-spectrum noise or two-talker speech), normal-hearing young adult listeners were presented with a priming sentence in quiet. This priming sentence was always recited using the same voice as the target sentence but had different content with the target sentence. Compared with the no-priming condition, the voice priming sentence significantly improved recognition of the target sentence when the masker was speech but not noise. It is suggested that, voice cues, which act at the perceptual level, can be used by young adult listeners to facilitate selective attention to the voice characteristics of the target stream, leading to a release of speech from informational masking (for the concepts of informational masking and energetic masking see Freyman et al. 1999; Li et al. 2004). In the study by Yang et al. (2007), however, because only one target voice was used, long-term familiarity of the target voice might influence the results.

Older listeners often find it difficult to understand speech in "cocktail-party" environments (Gelfand et al. 1988; Helfer & Wilber 1990; Cheesman et al. 1995; Huang et al. 2008). To our knowledge, no studies have been conducted to investigate whether the voice-priming effect releasing speech from speech masking occurs in older adults. The unmasking effect of voice priming must depend on auditory processing of acoustic details of the cuing voice. Our previous studies have confirmed that older adults have reduced abilities to temporally maintain acoustic details (Huang et al. 2009b; Li et al. 2009). In addition, previous studies by other investigators have shown that older adults have reduced abilities to discriminate talkers' voices (Helfer & Freyman 2008), remember talkers' voices (Yonan & Sommers 2000), and take advantage of the voice distinctiveness in target message identification (Rossi-Katz & Arehart 2009). This study examined whether there are age-related changes in the ability to use the perceptual-level voice-priming cue to unmask speech. Because older adults have reduced abilities to perceive and remember voice signals (Yonan & Sommers 2000; Helfer & Freyman 2008; Rossi-Katz & Arehart 2009), presenting only one single priming sentence may not be able to ensure that older adults become sufficiently familiar with the target voice. In this study, both single and double presentations of priming sentence, each recited by the target voice, were used.

MATERIALS AND METHODS

Participants

Eighteen younger university students (19 to 27 yrs, mean age = 22.0 yrs, 10 women and 8 men) and 12 older adults (60 to 80 yrs, mean age = 63.0 yrs, nine women and three men) participated in this study. Their first language was Mandarin Chinese. The participants gave their written informed consent to participate in the experiment and were paid a modest stipend for their participation.

All the participants had symmetrical hearing (no more than 15 dB difference between the two ears), younger participants had pure-tone hearing thresholds no more than 25 dB HL between 125 and 8000 Hz, and the older participants had pure-tone hearing thresholds no more than 25 dB HL between 125 and 500 Hz and no more than 40 dB HL between 1000 and 4000 Hz. Although hearing thresholds at 8000 Hz were also measured in older participants, they were not used for screening purposes.

The thresholds of older participants were generally higher than those of younger participants, and the group difference in thresholds increased with frequency. Particularly, for frequencies of 4000, 6000, and 8000 Hz, the thresholds of older adults exceeded 25 dB HL. Thus, these two age groups of participants were different not only in age but also in hearing sensitivity. Although these older adults were clinically normal in hearing, they were best characterized as being in the early stages of presbycusis.

Departments of Psychology, and Machine Intelligence, Speech and Hearing Research Center, Key Laboratory on Machine Perception (Ministry of Education), Peking University, Beijing, China.

0196/0202/10/3104-0579/0 • Ear & Hearing • Copyright © 2010 by Lippincott Williams & Wilkins • Printed in the U.S.A.

Apparatus and Stimuli

Participants were seated in a chair at the center of an anechoic chamber with a size of 560 × 400 × 193 cm (Beijing CA Acoustics Co. Ltd., Beijing, China). All acoustic signals were digitized at 22.05 kHz using a 24-bit Creative Sound Blaster PCI128 (Creative Technology, Ltd., Singapore) (which had a built-in antialiasing filter) and audio editing software (Cooledit Pro 2.0, Syntrillium Software Corporation, Phoenix, AZ) under the control of a computer with a Pentium IV processor. The acoustic signals were delivered to a loudspeaker (Dynaudio Acoustics, BM6 A, Dynaudio, Risskov, Denmark) located at the frontal position with a height of 106 cm (which was approximately ear level for a seated listener with average body height) and distance to the participant's head of 160 cm.

Speech stimuli were Chinese "nonsense" sentences that are syntactically correct but not semantically meaningful (Yang et al. 2007). Direct English translations of the sentences are similar but not identical to the English nonsense sentences that were developed by Helfer (1997). Each of the Chinese nonsense sentences has three keywords (subject, predicate, and object) with two characters for each used for scoring recognition performance. The "nonsense" sentence frame does not provide any contextual support for recognition of the keywords.

Target speech sentences were spoken by three young adult female talkers (talker A, B, or C), whose fundamental frequencies were 235, 229, and 225 Hz, respectively. Masking speech with different content from target speech was a combination of continuous recordings of Chinese "nonsense" sentences spoken by two other young adult female talkers (talkers D and E), whose fundamental frequencies were 225 and 228 Hz, respectively. Talker D and talker E spoke different sentences. The noise masker was a stream of steady-state speech-spectrum noise (Yang et al. 2007).

Under a single-priming condition, one priming sentence was produced by the same talker reading the target sentence. Under a double-priming condition, two priming sentences were produced successively also by the same talker reading the target sentence. Priming sentences were always different from each other and different from target sentences in content. Under a nonpriming condition, no priming sentence was presented.

All the stimuli were calibrated using a B&K sound level meter (Type 2230) (Yang et al. 2007). Both target and priming speech sounds were presented at 52 dBA. For younger participants, sound pressure levels of maskers were adjusted to produce four signal-to-noise ratios (SNRs) of −12, −8, −4, and 0 dB (Yang et al. 2007). For older participants, to minimize the floor effect caused by age-related hearing loss, sound pressure levels of maskers were adjusted to produce another four SNRs: −8, −4, 0, and 4 dB. In our previous study (Li et al. 2004), older listeners exhibited comparable ability to younger listeners to use perceived spatial separation to release target speech from either speech masking or noise masking as long as the SNR was improved by 2.8 dB.

Design and Procedure

For each participant group, there were three within-subject variables: (1) priming type (no priming, single priming, and double priming), (2) masker type (speech and noise), and (3) SNR. Each of the 24 conditions contained 18 target sentences, and 6 target sentences were assigned to each of talkers A, B, and C in a random order under each condition.

The six (3 × 2) priming/masker combinations were partially counterbalanced across 18 younger participants or 12 older participants using a Latin square design, and the four SNRs were arranged randomly for each priming/masker combination.

In each trial, the participant pressed a button of a response box to start the priming sound. Under priming conditions, either the single-priming or double-priming sentences were presented in quiet. Immediately after the priming phase, either the two-talker speech masker or the noise masker was presented, and about 1 sec later, a single target sentence was presented along with the masker. Then, the masker was gated off simultaneously with the target. Under a no-priming condition, the masker was presented immediately after the button press.

Participants were informed of both the masking condition and the priming type for a testing session and instructed to loudly repeat the whole target sentence as best as they could immediately after stimuli terminated. The experimenters, who sat outside the anechoic chamber, scored whether characters of the three keywords had been identified correctly. The number of correctly identified keyword characters was tallied later. There was one training session before the formal experiment, using different sentences from those used in the formal testing.

RESULTS

Figure 1 presents the group mean percentages of correctly identifying keyword characters for younger participants (top panels) and older participants (bottom panels) when the masker was speech (left panels) or noise (right panels) for the three priming conditions: (1) no priming (open circles), (2) single priming (filled circles), and (3) double priming (filled squares). The smooth curves drawn through symbols are logistic functions fitting the results (Yang et al. 2007). As shown by Figure 1, for younger participants, but not older participants, both single priming and double priming improved speech recognition under speech-masking, but not noise-masking, conditions.

For younger participants, a 2 (masker type) × 3 (priming type) × 4 (SNR) three-way within-subject analysis of variance (ANOVA) shows that the interaction between priming type and masker type was significant ($F[2,34] = 4.698$, $p = 0.016$), the interaction between masker type and SNR was significant ($F[3,51] = 11.229$, $p < 0.001$), but the interaction between priming type and SNR was not significant ($F[6,102] = 1.325$, $p = 0.253$) and the three-way interaction was not significant ($F[6,102] = 0.645$, $p = 0.694$). For the noise-masking condition, a 3 (priming type) × 4 (SNR) two-way within-subject ANOVA shows that the interaction between priming type and SNR was not significant ($F[6,102] = 1.866$, $p = 0.094$), the main effect of priming type was not significant ($F[2,34] = 0.976$, $p = 0.387$), and the main effect of SNR was significant ($F[3,51] = 874.005$, $p < 0.001$). For the speech-masking condition, a 3 (priming type) × 4 (SNR) two-way within-subject ANOVA shows that the main effect of priming type was significant ($F[2,34] = 15.857$, $p < 0.001$), the main effect of SNR was significant ($F[3,51] = 502.316$, $p < 0.001$), but the interaction between priming type and SNR was not significant ($F[6,102] = 0.248$, $p = 0.959$). Follow-up t tests indicate that for the speech-masking condition, performance under the

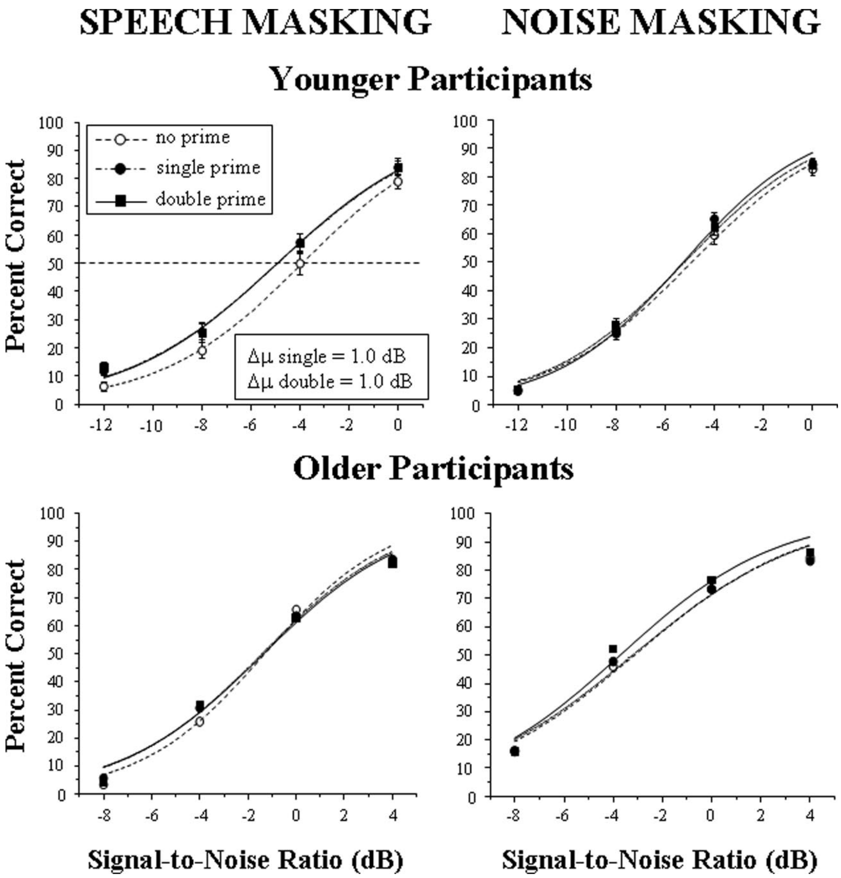

Fig. 1. Group mean percent correct recognition of keyword characters as a function of the signal-to-noise ratio for younger participants (top panels) and older participants (bottom panels) when the masker was speech (left panels) or noise (right panels) under each of the three priming conditions: (1) no priming (open circles), (2) single priming (filled circles), and (3) double priming (filled squares). The smooth curves drawn through symbols are logistic functions fitting the results (Yang et al. 2007). In the top left panel (for younger participants under speech masking), the horizontal, straight broken line indicates the 50% correct performance level, and the length (in dB) of the section in this horizontal line between the no-priming curve and single-priming (one-priming-sentence) curve and that between the no-priming curve and the double-priming (two priming sentence) curve represent single-priming sentence-induced benefit ($\Delta\mu$ single) and double-priming sentence-induced benefit ($\Delta\mu$ double), respectively. The error bars represent the standard errors of the mean.

no-priming condition was significantly poorer than both that under the single-priming condition ($t[17] = -4.385$, $p = 0.001$, $\alpha = 0.017$ with a Bonferroni adjustment) and that under the double-priming condition ($t[17] = -5.000$, $p < 0.001$), but there was no significant difference between the two priming conditions ($p = 0.903$). Moreover, 2 (masker type) × 4 (SNR) two-way within-subject ANOVAs show that under the no-priming, single-priming, and double-priming conditions, the main effect of SNR was significant ($p < 0.001$ for all the three conditions), but the main effect of masker was not significant ($p \geq 0.020$ for all the three conditions, $\alpha = 0.017$ with a Bonferroni adjustment) and the interaction between masker type and SNR was not significant ($p \geq 0.020$ for all the three conditions). These results suggest that in younger participants, the difference in masking effect between the noise masker and the speech masker was not significant, and presenting either one or two priming sentences significantly released target speech from speech masking but not from noise masking. Arcsine transformation of individual participants' data in the widely spanning percent correct values was also conducted (Studebaker 1985). ANOVAs of the arcsine-transformed data in arcsine units show the same statistical conclusions.

For older participants, a 2 × 3 × 4 three-way within-subject ANOVA shows that the main effect of priming type was not significant ($F[2,22] = 1.480$, $p = 0.249$), the interaction between priming type and masker type was not significant ($F[2,22] = 2.099$, $p = 0.146$), the interaction between priming type and SNR was not significant ($F[6,66] = 2.038$, $p = 0.073$), and the three-way interaction was not significant ($F[6,66] = 0.748$, $p = 0.613$). However, the interaction between masker type and SNR was significant ($F[3,33] = 19.780$, $p < 0.001$). Further 2 (masker type) × 3 (priming type) two-way within-subject ANOVAs show the following results: at the SNR of −8 dB, the main effect of masker type was significant ($F[1,11] = 123.833$, $p < 0.001$), the main effect of priming type was not significant ($F[2,22] = 0.285$, $p = 0.755$), and the interaction between masker type and priming type was not significant ($F[2,22] = 0.665$, $p = 0.524$). At the SNR of −4 dB, the main effect of masker type was significant ($F[1,11] = 42.554$, $p < 0.001$), the main effect of priming type was not significant ($F[2,22] = 3.634$, $p = 0.043$, $\alpha = 0.013$ with a Bonferroni adjustment), and the interaction between masker type and priming type was not significant ($F[2,22] = 0.561$, $p = 0.579$). At the SNR of 0 dB, the main effect of masker type was significant ($F[1,11] = 25.121$, $p < 0.001$), the main effect of priming type was not significant ($F[2,22] = 0.356$, $p = 0.704$), and the interaction between masker type and priming type was not significant ($F[2,22] = 2.017$, $p = 0.157$). At the SNR of 4 dB, the main effect of masker type was not significant ($F[1,11] = 1.242$, $p = 0.289$), the main effect of priming type was not significant ($F[2,22] = 0.240$, $p = 0.789$), and the interaction between masker type and priming type was not significant ($F[2,22] = 1.955$, $p = 0.165$). These results indicate that in older participants, the speech masker caused significantly larger masking effect than the noise masker at the SNRs of −8, −4, and 0 dB, but not +4 dB. In addition, presenting either one or two priming sentences did not release target speech from either speech masking or noise masking. ANOVAs of the arcsine-transformed data show the same statistical conclusions.

DISCUSSION

This study adopted the methods used in the study by Yang et al. (2007) to examine the speech-unmasking effect of voice priming in both younger adult listeners with normal hearing and older adult listeners with normal hearing for their age. However, in this study, instead of a single target talker's voice, three different young females' voices were used to recite target speech. In a testing trial, because one of the three voices was used as the target voice, participants depended only on the short-term memory of voice characteristics of this particular target voice for priming target speech. Thus, compared with the study by Yang et al. (2007), in which only one target talker's voice was used, in this study, any potential influences of long-term familiarity of a target voice were reduced.

The results of this study show that for younger participants, presenting either one or two priming sentences in quiet before the masker/target presentation significantly released target speech from speech masking but not noise masking. The results are consistent with those reported by Yang et al. (2007), indicating that younger listeners with normal hearing are able to use their short-term familiarity of a particular target voice as a cue to facilitate their selective attention to the target stream when other disruptive talking is presented. Because steady-state speech-spectrum noise seems to predominately provide energetic masking and two-talker speech provides both energetic and informational masking, the improvement of speech recognition induced by voice priming may reflect a release specifically from informational masking. In this study, because three different target talker voices were used, the voice cuing effect was effective only for a particular testing trial, based on short-term storage of voice information. In addition, compared with presenting one priming sentence, presenting two priming sentences to the younger adult listeners before the target speech presentation did not lead to any additional effects on recognition of target speech under either the noise-masking condition or the speech-masking condition. Thus, under the stimulus conditions used in this study, the familiarity with the target voice induced by a single presentation of the voice-priming sentence in a testing trial is sufficiently effective for younger listeners.

However, for older adult participants, presenting either one priming sentence or two priming sentences did not cause any releases of target speech from either speech masking or noise masking. Thus, older adult listeners with age-corrected normal hearing are not able to use voice information of target speech to improve their recognition of speech under "cocktail-party" environments.

As mentioned in the Introduction section, both abilities to perceive, remember, and use talkers' voices (Yonan & Sommers 2000; Helfer & Freyman 2008; Rossi-Katz & Arehart 2009) and abilities to temporally store acoustic fine structure details (Huang et al. 2009b; Li et al. 2009) decline in older adults. Because the use of voice cues to unmask speech must depend on processing of acoustic details of the cuing voice, the absence of any voice-priming effects in older adults may be attributed to the decreased ability to discriminate acoustic features of various voices with similar fundamental frequencies and/or the decreased ability to temporally store acoustic details. Moreover, our recent studies have shown that the temporal storage of acoustic details is also critical for releasing speech from informational masking under (simulated) reverberant environments (Huang et al. 2009a). Thus, for older adults, the inability to temporally store acoustic details (Huang et al. 2009b; Li et al. 2009) may mainly contribute to both the absence of voice-priming effects (this study) and the reduction of unmasking effects of perceptual integration (Huang et al. 2008). The results of this study also show that for older participants, but not younger participants, the speech masker caused a larger masking effect than the noise masker when the SNR was 0 dB or less. The augmented detrimental influence of the speech masker may be attributed to both the reduced ability to encode voice characteristics and the reduced ability to take advantage of temporal fluctuations in the speech masker. Thus, this study proposed an explanation as to why older listeners often find it difficult to understand speech under "cocktail-party" environments (Gelfand et al. 1988; Helfer & Wilber 1990; Cheesman et al. 1995; Huang et al. 2008).

In other studies, hearing loss in older adults is not significantly correlated with either performance of voice discrimination (Helfer & Freyman 2008) or temporal storage of acoustic details (Huang et al. 2009b; Li et al. 2009). Furthermore, older adults have the comparable ability to younger adults to use the cue of perceived spatial separation to release target speech from informational masking (Helfer & Freyman 2008; Li et al. 2004). Nonetheless, the role of the age-related hearing loss in impairing voice priming should not be completely ruled out.

ACKNOWLEDGMENTS

This work was supported by the "973" National Basic Research Program of China (2009CB320901), the National Natural Science Foundation of China (30711120563; 30670704; 60535030), and "985" grants from Peking University.

Address for correspondence: Liang Li, Department of Psychology, Peking University, Beijing 100871, China. E-mail: liangli@pku.edu.cn.

Received August 9, 2009; accepted February 9, 2010.

REFERENCES

Cheesman, M. F., Hepburn, D., Armitage, J. C., et al. (1995). Comparison of growth of masking functions and speech discrimination abilities in younger and older adults. *Audiology*, *34*, 321–333.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *J Acoust Soc Amer*, *25*, 975–979.

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., et al. (1999). The role of perceived spatial separation in the unmasking of speech. *J Acoust Soc Am*, *106*, 3578–3588.

Gelfand, S. A., Ross, L., Miller, S. (1988). Sentence reception in noise from one versus two sources: Effects of aging and hearing loss. *J Acoust Soc Am*, *83*, 248–256.

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. (2008). Aging and speech-on-speech masking. *Ear Hear*, *29*, 87–98.

Helfer, K. S., & Wilber, L. A. (1990). Hearing-loss, aging, and speech-perception in reverberation and noise. *J Speech Hear Res*, *33*, 149–155.

Huang, Y., Huang, Q., Chen, X., et al. (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., et al. (2009a). 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., Wu, X.-H., Li, L. (2009b). Detection of the break in interaural correlation is affected by interaural delay, aging, and center frequency. *J Acoust Soc Am*, *126*, 300–309.

Li, L., Daneman, M., Qi, J. G., et al. (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.

Li, L., Huang, J., Wu, X.-H., et al. (2009). The effects of aging and interaural delay on the detection of a break in the interaural correlation between two sounds. *Ear Hear*, *30*, 273–286.

Rossi-Katz, J., & Arehart, K. H. (2009). Message and talker identification in older adults: Effects of task, distinctiveness of the talkers' voices, and meaningfulness of the competing message. *J Speech Lang Hear Res*, *52*, 435–453.

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

Yang, Z.-G., Chen, J., Huang, Q., et al. (2007). The effect of voice cuing on releasing Chinese speech from informational masking. *Speech Commun*, *49*, 892–904.

Yonan, C. A., & Sommers, M. S. (2000). The effects of talker familiarity on spoken word identification in younger and older listeners. *Psychol Aging*, *15*, 88–99.