al
D 2/p

Y a ch f
i a

Y G ong Wu,[1] Haifen

[1] tr Research
R ot Beiji
a e
a N Medic
A in d Television
B n

s a cocktail party, listeners can us /or cognitive cues
articularly informational masking. Pr shown that tem
tion of target speech against speech oise mas
ive become target-voice ges
be used b sults
yo the
by irrelevant two-talker speech. W son's
rning, temporally prepresentin voice
n masking, particularly for key ech-
priming condition was sig relat The results
cial in on talker identity pla nt ro e and facilitating
o the target-sp h stream against th ch st

Ke ace priming; face- ciati ma priming

de ce: Dr. Liang gy, P 871, China. Email: liangli@pku.edu.cn

Received 16 May 2013. A

such as a cock-
ceptual and/or c
of target speec
ergetic mas
masking
ed by
sker

ple, when the masker is speech, pro-
interferes
tic
a seg
very
, Li
eyn
, W
Ho
use
pers

largely by strengthening their selective attention to target speech, leading to improved recognition of target speech (for a review see Du, Kong, Wang, Wu, & Li, 2011). Some of the cues do not (substantially) change energetic masking of the target speech. These cues include knowledge/familiarity of the target-talker's voice (Brungart, 2001; Helfer & Freyman, 2005; Huang, Xu, Wu, & Li, 2010; Newman & Evers, 2007; Yang et al., 2007), which facilitates the listeners' selective attention to the target speech and improves recognition of the target speech in a speech masker that particularly induces informational masking (Huang et al., 2010; Yang et al., 2007). In more detailed studies, normal-hearing young-adult listeners were presented with a priming sentence in a quiet environment immediately before the copresentation of a target-speech sentence with a masker (either steady-state speech-spectrum noise or two-talker speech). The priming sentence is always recited using the same voice as the target sentence, but has different content from the target sentence. Compared with the no-priming condition, the voice-priming sentence significantly improves recognition of the target sentence when the masker is speech but not noise (Huang et al., 2010; Yang et al., 2007). It is suggested that voice cues, which act at the perceptual level, can be used by listeners to facilitate selective attention to the vocal characteristics of the target stream, leading to a release of speech from informational masking (Huang et al., 2010).

It is well known that the talker's voice contains not only speech-content information, but also talker's identity information. Similarly, face cues also contain information of talker identity (Belin, Bestelmeyer, Latinus, & Watson, 2011; Belin, Fecteau, & Bédard, 2004; Campanella & Belin, 2007) and interact strongly with voice cues in identifying talkers (Joassin et al., 2011). In the present study, we investigated whether static face images that are associated with voices reciting target sentences by associative learning can improve recognition of target speech against informational masking.

Materials and methods

Participants

Thirty-two Mandarin Chinese speaking university students (11 female and 21 male, with a mean age of 21.7 years and ranging from 17 to 26 years) participated in this study. All the participants had both normal or corrected-to-normal visual acuity and normally symmetrical hearing (no more than a 15-dB difference between the two ears, and pure-tone hearing thresholds no more than 25 dB hearing level between 0.125 and 8 kHz). The participants gave their written informed consent and were paid a modest stipend for their participation.

Apparatus and stimuli

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 antialiasing 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, CA, USA). The audio stimuli were presented through a pair of matched Sennheiser earphones (HD 265). The sound pressure level at each ear was set at 56 dBA, calibrated by a Larson Davis Audiometer Calibration and Electroacoustic Testing System (Audit and System 824, Larson Davis, NY, USA). The visual stimuli (static face images against a black background) were displayed on a Dell UltraScan P780 computer monitor with a spatial resolution of 1024 × 768, and a refresh rate of 75 Hz. The participants were seated at a distance of 68 cm from the monitor screen.

The speech stimuli were Chinese nonsense sentences that were 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, Helfer, McCall, and Clifton (1999), Freyman, Balakrishnan, and Helfer (2004), Li et al. (2004), and Ezzatian et al. (2011). Each sentence had 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, "That corona may remove my crest" (the keywords are underlined). This nonsense subject–predicate–object structure provided no contextual support for recognizing keywords. The development of the Chinese nonsense sentences is described elsewhere (Yang et al., 2007). To maintain consistency with our previous investigations since the Li et al. (2004) study, target, masker, priming, and training speech sentences were recited by three different young female voices.

During the target/masker copresentation, three different nonsense sentences (one was the target one and the others were the masking ones) with the same sound pressure level

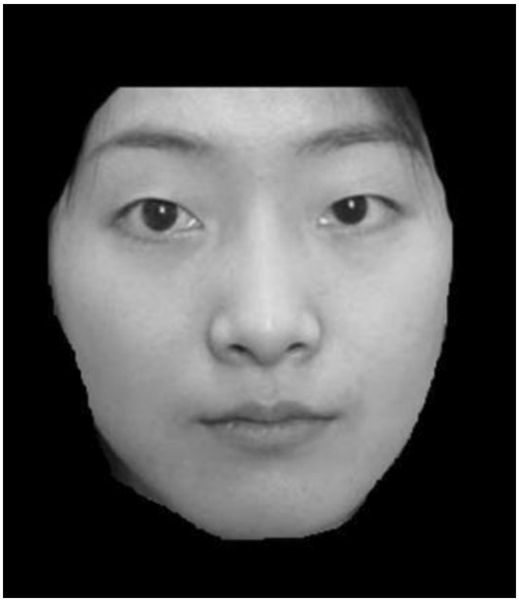


Figure 1. An example of the facial stimuli used in the study (obtained from the Chinese Affective Picture System, CAPS; Bai, Ma, Huang, & Luo, 2005).

were presented simultaneously and recited by different voices (one for the target sentence and two for the masking sentences). Thus, the signal-to-masker ratio (SMR) was −3 dB.

Black and white photographs of the faces of three young Chinese females were selected from the Chinese Affective Picture System (CAPS; see Figure 1 for an example; Bai, Ma, Huang, & Luo, 2005) with 370 × 556 pixel resolution and presented in the center of the monitor with a black background. To avoid irrelevant influences from headwear and/or hair style, the female models' hair was masked in the photograph.

Procedure

There were three types of priming condition (basic/no-priming, voice-priming, and face-priming), the presentation of which was arranged with a block design through partial counterbalance across 32 participants with 48 trials per condition for 12 participants (who participated in the study earlier) and 30 trials per condition for 20 participants (who participated in the study later).

In the phase of voice-photograph-association training, participants were trained to establish a one-to-one association between a particular voice and a particular face photograph by associative learning. Specifically, in a training trial for example, a sentence recited by a particular voice was copresented with a particular face photograph and participants were told that the voice and the photograph belonged to the same talker. After the training, each of the participants passed an examination confirming that the connection of each of the three voice-face pairs was well established.

In the testing phase of the experiment, the participant pressed a button of a response box to start a trial. Before the copresentation of the target and masking sentences one of the three primes was presented: (1) a sentence recited by the target voice under the voice-priming condition, (2) a face photograph (with a presentation duration of 2000 ms) that was associated with the voice reciting the target sentence under the face-priming condition, and (3) nothing presented under the basic (no-priming) condition. Three hundred milliseconds after the presentation of either the voice or face-photograph prime, or after the button press under the basic (no-priming) condition, the target and masker sentences were presented simultaneously and then terminated at the same time.

Under the basic (no-priming) condition, to unambiguously specify the target sentence among the three copresented sentences, only the target sentence was started with the characters 这个 'this' or 这些 'these', and the participants were instructed that in the coming trial the three sentences would start simultaneously after the button press but only the target sentence would start with the characters 这个 or 这些. Under the voice-priming condition, as the priming voice was specifically associated with the target sentence, participants were instructed that in the coming trial the voice reciting the prime was identical to that reciting the target sentence. Under the face-priming condition, as the priming face was specifically associated with the voice reciting the target sentence, participants were instructed that in the coming trial the voice reciting the target sentence was associated with the face prime. Immediately after the stimulus presentation, participants were instructed to loudly repeat the whole target sentence as best they could.

Performance was scored as the number of correctly identified syllables for each keyword. To ensure that all the participants fully understood and correctly followed the experimental instructions, there was one training session before formal testing. To avoid the content priming effect, the sentences used in the training, priming, and testing were different for each participant.

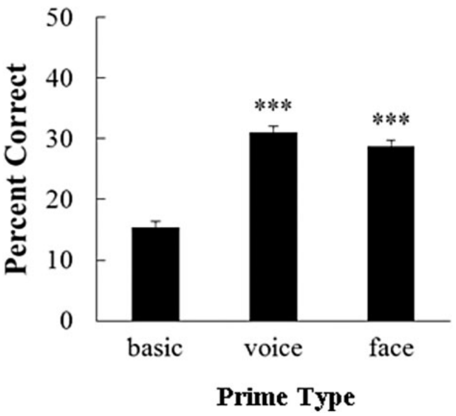
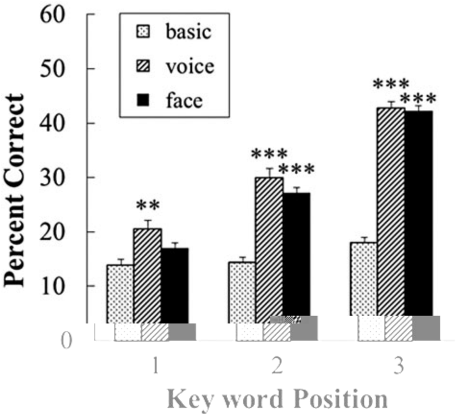


Figure 2. Group mean percentage correct identification of all the keywords in target speech across 32 participants under each of the three conditions: basic (no-priming) condition, voice-priming condition, and face-priming condition. The error bars represent the standard errors of the mean. ***$p < .001$ against the basic condition.



Re l

C m a i i eech ec g i j be ee he h ee imi g c di j

Figure 2 plots the group mean percentage correct identification of target sentences under each of the three priming conditions. Obviously, participants' performance under the voice- or face-priming condition was better than that under the basic (no-priming) condition. A one-way ANOVA confirmed that the effect of priming condition was significant, $F(2, 30) = 144.416$, $p < .001$. Bonferroni-corrected post hoc analysis indicated that performance under the basic condition was significantly worse than that under both the voice-priming ($MD = 0.156$, $SE = 0.009$, $p < .001$) and face-priming conditions ($MD = 0.133$, $SE = 0.011$, $p < .001$). There was no significant difference between the performance under the voice-priming and the face-priming conditions ($MD = 0.023$, $SE = 0.010$, $p = .085$).

We also analyzed the keyword position effect. Figure 3 shows the group mean percentage correct identification of each of the three keywords under each of the three priming conditions. A 3 (priming condition: basic, voice, face) × 3 (keyword position: first, second, third) within-subject ANOVA showed that all main effects were significant (all ps $< .001$), and the interaction between priming condition and keyword position was significant, $F(4, 124) = 32.014$, $p < .001$.

Further analyses were conducted using multiple t-tests (Bonferroni corrected). For the first keyword, the performance under the voice-priming condition was significantly better than that under the basic condition, $t(31) = -4.116$, $p = .010$, but there was no significant difference for the face-priming from basic-priming and voice-priming conditions (both $p > .521$). For the second keyword, the performance under the basic condition was significantly worse than under the voice-priming, $t(31) = -10.235$, $p < .001$, and face-priming conditions, $t(31) = -8.546$, $p < .001$, but there was no significant difference between the voice-priming and face-priming conditions ($p = 6.228$). For the third keyword, the performance under the basic condition was significantly worse than under the voice-priming, $t(31) = -19.040$, $p < .001$, and face-priming conditions, $t(31) = -17.915$, $p < .001$, but there was no significant difference between the voice-priming and face-priming conditions ($p = 21.725$).

Moreover, under the basic condition, there was no significant difference between the recognition of the first, second, and third keywords ($p > .261$). Under both the voice-priming condition and the face-priming condition recognition of the first keyword was significantly worse than that of the second keyword, $t(31) = -6.441$, $p < .001$ for voice-priming and $t(31) = -6.352$, $p < .001$ for face-priming, as well as that of the third keyword, $t(31) = -12.438$, $p < .001$ and $t(31) = -13.283$, $p < .001$, and recognition of the second keyword was significantly worse than that of the third keyword under the two conditions, $t(31) = -7.662$, $p < .001$ and $t(31) = -9.342$, $p < .001$.

C ela j be ee v ice imi g a d face imi g

Figure 4 shows the percentage correct of recognizing target speech for individual participants under the voice-priming

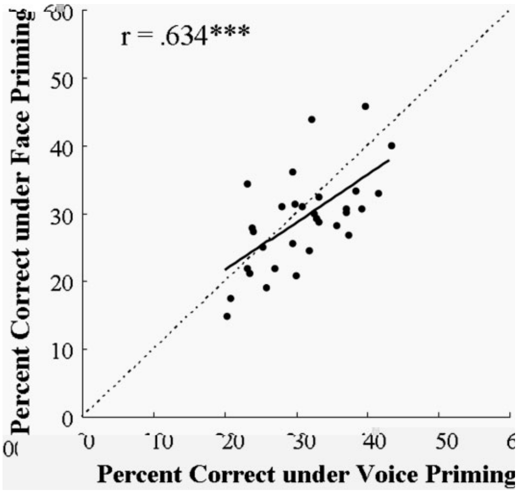


Figure 4. Percentage correct identification of all the keywords in target speech in each of the 32 participants under the voice-priming condition (values along the abscissa) and that under the face-priming condition (values along the ordinate). The solid line represents the linear regression between the two priming conditions. The Pearson's correlation coefficient is indicated at the top of the figure. ***$p < .001$.

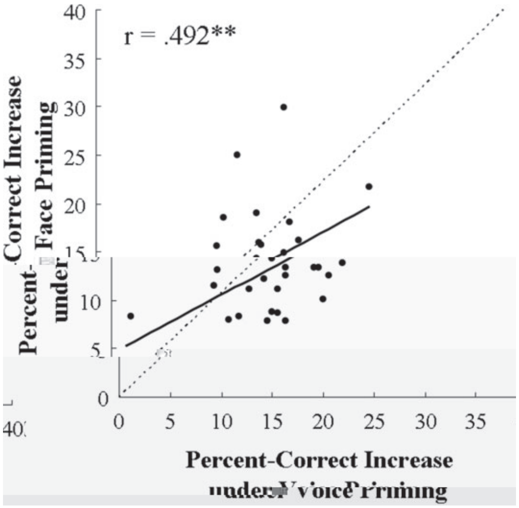


Figure 5. Prime-induced percentage correct increase in recognition of target speech (against the basic condition) in each of the 32 participants under the voice-priming condition (values along the abscissa) and that under the face-priming condition (values along the ordinate). The solid line represents the linear regression of individuals' value between the two priming conditions. The Pearson's correlation coefficient is indicted at the top of the figure. **$p < .01$.

condition (values along the abscissa) and that under the face-priming condition (values along the ordinate). A linear regression was conducted between the two sets of data. As shown in Figure 4, there was a significant correlation in target-speech recognition between the two priming conditions (Pearson's $r = .643$, $p < .001$), suggesting that if a participant performed better under the voice-priming condition, they had a high chance to perform better under the face-priming condition.

We also investigated the correlation in the prime-induced gain (against the basic condition) between the voice-priming condition and the face-priming condition. Specifically, we used the percentage-correct increase under each of the priming conditions against that under the basic condition as the index and calculated the correlation between the two indices across 32 participants. As shown in Figure 5, there was a significant correlation in the prime-induced gain between the voice-priming condition and the face-priming condition (Pearson's $r = .492$, $p = .004$), suggesting that if a participant benefited more from introducing the voice prime, they had a greater chance of benefitting more from introducing the face prime.

Sex difference

The results also showed that female participants performed better on average than male participants. ANOVAs confirmed that a group mean sex difference significantly occurred for each of the priming conditions: basic condition, $t(30) = -2.699$, $p = .012$; voice-priming condition, $t(29.726) = 2.288$, $p = .029$; and face-priming condition, $t(30) = 2.235$, $p = .033$.

Discussion

The voice prime releases target speech from informational masking

The results of this study showed that for younger participants with normal hearing, prepresenting the priming sentence recited with the target talker's voice in a quiet environment before the masker/target copresentation significantly released target speech from speech masking. The results are consistent with those reported by Yang et al. (2007) and Huang et al. (2010), indicating that younger listeners with normal hearing are able to use their short-term familiarity with a particular target voice as a cue to facilitate their selective attention to the target stream when other disruptive talking is presented. These previous studies have also shown that this unmasking effect does not occur when the masker is steady-state speech-spectrum noise (Huang et al., 2010; Yang et al., 2007). As steady-state speech-spectrum noise predominately provides energetic masking and two-talker speech provides both energetic and informational masking, the improvement of speech recognition induced by voice priming only under the speech-masking condition reflects a release specifically from informational masking.

The voice prime speeds up the build-up of speech-recognition improvement

Under the basic condition there was no significant difference in recognition between the three keywords. The results suggest that, when the prime is absent, there is a slow improvement of target-speech recognition with the listener's successive exposures to the target sentence. Under the voice-priming condition, recognition of the third keyword was significantly better than that of both the first and second keywords, and recognition of the second keyword was also significantly better than that of the first keyword. Thus, when the voice prime is present, the build-up of the speech-recognition improvement is speeded up.

The face priming effect

It has been well known that voice processing interacts closely with face processing in person identification (e.g., Ellis, Jones, & Mosdell, 1997; Latinus & Belin, 2011; Sai, 2005; Schweinberger, Robertson, & Kaufmann, 2007; von Kriegstein et al., 2008; von Kriegstein, Eger, Kleinschmidt, & Giraud, 2003; von Kriegstein, Kleinschmidt, Sterzer, & Giraud, 2005). This study for the first time provides evidence that when a static face image becomes perceptually linked with a particular voice through associative learning, temporally prepresenting this face image releases target speech from speech masking when the target speech is recited by the associated voice. Thus, face priming unmasks the target speech just like voice priming. Because in this study the face image was static without articulatory movements, the unmasking effect revealed in this study was based on talker identity. It is plausible that the face-priming effect is mediated by a chain of processes including the following: (a) visual perception of the face image, (b) identification of the face image that has been learned, (c) activation of the central representation of the voice associated with the face image, (d) recognition of the voice that recites target sentences, (e) facilitation of selective attention to the target talker's voice due to the priming effect, and (f) improved recognition of the target sentence.

The results of the current study also showed that, similar to the effect of the voice prime, the face prime improved recognition in tandem with the listener's successive exposures to the target sentence. Specifically, under the face-priming condition, recognition of the third keyword was significantly better than that of both the first and second keywords, and recognition of the second keyword was significantly better than that of the first keyword. As recognition of the third keyword (a noun) was always better than recognition of the second keyword (a verb), the potential interaction of word order and word type should not be ruled out.

Correlation between voice priming and face priming

More importantly, the results of this study showed that participants' speech-recognition performance under the voice-priming condition was significantly correlated to that under the face-priming condition. In addition, the performance improvement caused by introducing the voice prime (against the performance under the basic [no-priming] condition) was significantly correlated to that caused by introducing the face prime. The results indicate that the abilities to process the face information for talker identification and to process the voice information for talker identification are functionally correlated. Clinical evidence also supports the view that deficits in face cognition and those in voice cognition are associated (Garrido et al., 2009; Hailstone, Crutch, Vestergaard, Patterson, & Warren, 2010).

Conclusion

Both temporally prepresented content primes (Aydelott, Baer-Henney, Trzaskowski, Leech, & Dick, 2012; Ezzatian et al., 2011; Freyman et al., 2004; Jones & Freyman, 2012; Sheldon, Pichora-Fuller, & Schneider, 2008; Wu, Cao et al., 2012; Wu, Li, Gao et al., 2012 Wu, Li, Hong et al., 2012; Yang et al., 2007) and temporally prepresented voice primes (Huang et al., 2010; Yang et al., 2007) can unmask target speech at the cognitive level (with knowledge about part of the target-sentence content) and the perceptual level (with knowledge about the voice reciting the target speech). Based on the perceptual association between the face images and the voices reciting target speech, it is plausible that the face-priming effect is mediated by a chain of processes, including visual perception of the face image, identification of the face image, activation of the central representation of the voice associated with the face image, recognition of the voice that recites target sentences, facilitation of selective attention to the target talker's voice, and improved recognition of the target sentence against the irrelevant (masking) speech background. The strategy of the face-input-induced unmasking of speech is valuable for developing the top-down unmasking function of the computer speech-recognition system used in cocktail party environments.

Ack ledgme

This work was supported by the National Natural Science Foundation of China (31170985), the Research Special Fund for Public Welfare Industry of Health (201202001), the “973” National Basic Research Program of China (2011CB707805), and “985” grants from Peking University.

Refe e ce

Aydelott, J., Baer-Henney, D., Trzaskowski, M., Leech, R., & Dick, F. (2012). Sentence comprehension in competing speech: Dichotic sentence-word priming reveals hemispheric differences in auditory semantic processing. *Language and Cognitive Processes*, *27*, 1108–1144. doi:10.1080/01690965.2011.589735

Bai, L. 白露, Ma, H. 马慧, Huang, Y. 黄宇霞, & Luo, J. 罗跃嘉 . (2005). 中国情绪图片系统的编制 在46名中国大学生中的试用 [The development of native Chinese affective picture system—A pretest in 46 college students]. *Chinese Mental Health Journal* [中国心理卫生杂志], *19*(11), 719–722. doi:10.3321/j.issn:1000-6729.2005.11.001

Belin, P., Bestelmeyer, P. E. G., Latinus, M., & Watson, R. (2011). Understanding voice perception. *British Journal of Psychology*, *102*, 711–725. doi:10.1111/j.2044-8295.2011.02041.x

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences*, *8*, 129–135. doi:10.1016/j.tics.2004.01.008

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *Journal of the Acoustical Society of America*, *109*, 1101–1109. doi:10.1121/1.1345696

Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences*, *11*, 535–543. doi:10.1016/j.tics.2007.10.001

Du, Y., Kong, L., Wang, Q., Wu, X., & Li, L. (2011). Auditory frequency-following response: A neurophysiological measure for studying the “cocktail-party problem.” *Neuroscience and Biobehavioral Reviews*, *35*, 2046–2057. doi:10.1016/j.neubiorev.2011.05.008

Ellis, H. D., Jones, D. M., & Mosdell, N. (1997). Intra- and inter-modal repetition priming of familiar faces and voices. *British Journal of Psychology*, *88*, 143–156. doi:10.1111/j.2044-8295.1997.tb02625.x

Ezzatian, P., Li, L., Pichora-Fuller, K., & Schneider, B. (2011). The effect of priming on release from informational masking is equivalent for younger and older adults. *Ear and Hearing*, *32*, 84–96. doi:10.1097/AUD.0b013e3181ee6b8a

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2004). Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *Journal of the Acoustical Society of America*, *115*, 2246–2256. doi:10.1121/1.1689343

Freyman, R. L., Helfer, K. S., McCall, D. D., & Clifton, R. K. (1999). The role of perceived spatial separation in the unmasking of speech. *Journal of the Acoustical Society of America*, *106*, 3578–3588. doi:10.1121/1.428211

Garrido, L., Eisner, F., McGettigan, C., Stewart, L., Sauter, D., Hanley, J. R., . . . Duchaine, B. (2009). Developmental phonagnosia: A selective deficit of vocal identity recognition. *Neuropsychologia*, *47*, 123–131. doi:10.1016/j.neuropsychologia.2008.08.003

Hailstone, J. C., Crutch, S. J., Vestergaard, M. D., Patterson, R. D., & Warren, J. D. (2010). Progressive associative phonagnosia: A neuropsychological analysis. *Neuropsychologia*, *48*, 1104–1114. doi:10.1016/j.neuropsychologia.2009.12.011

Helfer, K. S. (1997). Auditory and auditory-visual perception of clear and conversational speech. *Journal of Speech, Language, and Hearing Research*, *40*, 432–443.

Helfer, K. S., & Freyman, R. L. (2005). The role of visual speech cues in reducing energetic and informational masking. *Journal of the Acoustical Society of America*, *117*, 842–849. doi:10.1121/1.1836832

Huang, Y., Xu, L., Wu, X., & Li, L. (2010). The effect of voice cuing on releasing speech from informational masking disappears in older adults. *Ear and Hearing*, *31*, 579–583. doi:10.1097/AUD.0b013e3181db6dc2

Joassin, F., Pesenti, M., Maurage, P., Verreckt, E., Bruyer, R., & Campanella, S. (2011). Cross-modal interactions between human faces and voices involved in person recognition. *Cortex*, *47*, 367–376. doi:10.1016/j.cortex.2010.03.003

Jones, J. A., & Freyman, R. L. (2012). Effect of priming on energetic and informational masking in a same–different task. *Ear and Hearing*, *33*, 124–133. doi:10.1097/AUD.0b013e31822b5bee

Latinus, M., & Belin, P. (2011). Human voice perception. *Current Biology*, *21*, R143–R145. doi:10.1016/j.cub.2010.12.033

von Kriegstein, K., Dogan, Ö., Grüter, M., Giraud, A.-L., Kell, C. A., Grüter, T., . . . Kiebel, S. J. (2008). Simulation of talking faces in the human brain improves auditory speech recognition. *Proceedings of the National Academy of Sciences*, *105*, 6747–6752. doi:10.1073/pnas.0710826105

von Kriegstein, K., Eger, E., Kleinschmidt, A., & Giraud, A. L. (2003). Modulation of neural responses to speech by directing attention to voices or verbal content. *Cognitive Brain Research*, *17*, 48–55. doi:10.1016/S0926-6410(03)00079-X

von Kriegstein, K., Kleinschmidt, A., Sterzer, P., & Giraud, A.-L. (2005). Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience*, *17*, 367–376. doi:10.1162/0898929053279577

Wu, C., Cao, S., Zhou, F., Wang, C., Wu, X., & Li, L. (2012). Masking of speech in people with first-episode schizophrenia and people with chronic schizophrenia. *Schizophrenia Research*, *134*, 33–41. doi:10.1016/j.schres.2011.09.019

Wu, M., Li, H., Gao, Y., Lei, M., Teng, X., Wu, X., & Li, L. (2012). Adding irrelevant information to the content prime reduces the prime-induced unmasking effect on speech recognition. *Hearing Research*, *283*, 136–143. doi:10.1016/j.heares.2011.11.001

Wu, M., Li, H., Hong, Z., Xian, X., Li, J., Wu, X., & Li, L. (2012). Effects of aging on the ability to benefit from prior knowledge of message content in masked speech recognition. *Speech Communication*, *54*, 529–542. doi:10.1016/j.specom.2011.11.003

Wu, X., Wang, C., Chen, J., Qu, H., Li, W., Wu, . . . Li, L. (2005). The effect of perceived spatial separation on informational masking of Chinese speech. *Hearing Research*, *199*, 1–10.doi:10.1016/j.heares.2004.03.010

Yang, Z., Chen, J., Huang, Q., Wu, X., Wu, Y., Schneider, B. A., & Li, L. (2007). The effect of voice cuing on releasing Chinese speech from informational masking. *Speech Communication*, *49*, 892–904. doi:10.1016/j.specom.2007.05.005