Discussion

Two points were clarified: (1) quite similar patterns of results were obtained over two distinctly different languages, i.e., Japanese and British English, and (2) the analyses of British English were highly replicable over two independent speech databases (NTT-AT and ATR). To assess boundaries of frequency bands, we took the crossover frequencies of the curves in the figures. The boundaries were 510, 1880, and 2700 Hz in Japanese, and 550, 1800, 3300 Hz in British English of the NTT-AT database. Informal listening tests by the authors showed that either set of boundaries could yield intelligible noise-vocoded speech in both languages. Therefore, these four frequency bands should represent fundamental processing units along frequency axis related to speech perception.

Acknowledgements

This research was supported by Grants-in-Aid for Scientific Research Nos. 14101001, 19103003, and 20330152 from the Japan Society for the Promotion of Science and by a Grant-in-Aid for the 21st Century COE program.

References

Dorman, M. F., Loizou, P. C., & Rainey, D. (1997). Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. *Journal of the Acoustical Society of America, 102*, 2403-2411.

Fletcher, H. (1940). Auditory patterns. *Reviews of Modern Physics, 12*, 47-65.

Hillenbrand, J. M., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America, 97*, 3099-3111.

Hillenbrand, J. M., & Nearey, T. M. (1999). Identification of resynthesized /hVd/ utterances: Effects of formant contour. *Journal of the Acoustical Society of America, 105*, 3509-3523.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America, 29*, 98-104.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24*, 175-184.

Plomp, R. (1976). *Aspects of Tone Sensation: A Psychophysical Study*. London: Academic Press.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270*, 303-304.

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature, 416*(7 March 2002), 87-90.

Ueda, K., & Nakajima, Y. (2007). Critical-band filter analysis of speech sentences, *The 23rd Annual Meeting of the International Society for Psychophysics* (pp. 503-508). Tokyo.

Verbrugge, R. R., & Rakerd, B. (1986). Evidence of talker-independent information for vowels. *Language and Speech, 29*, 39-57.

Zwicker, E., & Terhardt, E. (1980). Analytical expressions for critical-band rate and critical bnadwidth as a function of frequency. *Journal of the Acoustical Society of America, 68*, 1523-1525.

SPEECH- SYNCHRONIZED VISUAL CUES RELEASE SPEECH FROM INFORMATIONAL MASKING

Mengyuan Wang, Jingyu Li, Liang Zhang, Yanhong Wu, Xihong Wu, and Liang Li
Department of Psychology, Peking University, Beijing, China, 100871
motoluto@163.com

Abstract

Visual speech information, such as lipreading cues, assists listeners to segregate a target voice from competing voices. It is not clear whether a simple visual cue, such as the light flash that is synchronous to the onset of each syllable in target speech, is sufficient to release target speech from noise or speech masking. In this study, when target speech was of a constant rate, the speech-synchronized light flash had no unmasking effects. However, when the rate of target speech was artificially manipulated, the speech-synchronized light flash improved speech recognition when the two-talker speech masker but not the speech-spectrum noise masker was co-presented. Thus, under certain conditions, speech-synchronized visual cues can play a role in helping listeners attend to the target voice and follow the stream of target speech, leading to a release of target speech from informational masking.

People often participate in conversations in noisy environments with noise sounds and person talking. Under such adverse conditions, listeners with normal hearing can use some perceptual cues to segregate target speech from the noise background. For example, viewing a speaker's articulatory movements (e.g., lipreading) substantially improves a listener's recognition of the speaker's speech especially under noisy conditions. Helfer and Freyman (2005) have recently reported that the effect of lipreading on speech recognition is masker-type dependent. Lipreading can only release speech from speech masking but not noise masking, suggesting that visual cues help listeners overcome informational masking but not energetic masking.

However, lipreading information is very complicated. This study investigated whether a single-dimensional signal in lipreading, the speech-synchronized light flash (which temporally matched the onset of each syllable in a target speech sentence) is sufficient to unmask speech.

Participants

Thirty-six young university students participated in this study, twelve in Experiment 1 and twenty-four in Experiment 2 (twelve in each part of Experiment 2). They had normal and symmetrical hearing (no more than 15 dB difference between the two ears, pure-tone hearing thresholds < 25 dB HL between 0.125 and 8 kHz). Their first language was Mandarin Chinese.

Apparatus

The participant was seated at the center of an anechoic chamber (Beijing CA Acoustics). All acoustic and visual signals were digitized using the 24-bit Creative Sound Blaster PCI128 and audio editing software (Cooledit Pro 2.0). The acoustic analog outputs were delivered from a loudspeaker (Dynaudio Acoustics, BM6 A) 200 cm in front of the participant. The flash was delivered from a light-emitting diode (LED) at the center of the loudspeaker.

Stimuli

Speech stimuli were Chinese "nonsense" sentences, which 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 by Freyman et al. (1999, 2001, 2004) and Li et al. (2004). Each of the Chinese sentences has three keywords, with two syllables for each. The development of the Chinese nonsense sentences is described by Yang et al. (2007).

In Experiment 1, target speech was spoken by a young female talker with a normal, rate-constant speech. In Experiment 2, the rate of target sentences was artificially modulated. The rate-varied target speech sentences contained three different rates (1.5, 1, and 0.5) which were randomly assigned in a single sentence. The speech masker was recording of nonsense sentences spoken by other two young females. The noise masker was a stream of steady-state speech-spectrum noise. Both target and masker presented from the same loudspeaker.

Three kinds of flash lights were used: syllable-synchronized, stable, and random.

Design and Procedure

Both Experiment 1 and Experiment 2.2 had three within-subject factors: (1) masker type, (2) SNR, and (3) light flash type. Experiment 2.1 also had three different within-subject factors, but the third factor was different from others: (1) masker type, (2) SNR, and (3) target speech rate type. The listener's task was to loudly repeat the whole target sentence as best as he/she could immediately after sounds were completed. A training session was provided to participants before formal experiments. Each of the two syllables for a key word was scoring individually.

Results

Experiment 1: Speech recognition of stable-rate target speech with and without speech-synchronized light flash cue under noise or speech masking

For each of the 12 listeners, a logistic psychometric function,

$$p(y) = \frac{1}{1 + e^{-\sigma(x-\mu)}}$$

was fit to the mean data across trials under each condition, where y is the probability of correct identification of key words, x is the SNR corresponding to y, μ is the SNR corresponding to 50% correct on the psychometric function, and σ determines the slope of the psychometric function.

Figure 1 shows group-mean percent correct of speech identification as a function of SNR, along with the best-fitting psychometric functions (curves) under noise and speech masking conditions. Clearly, presenting the syllable-synchronized light flash did not affect speech recognition under either noise masking or speech masking.

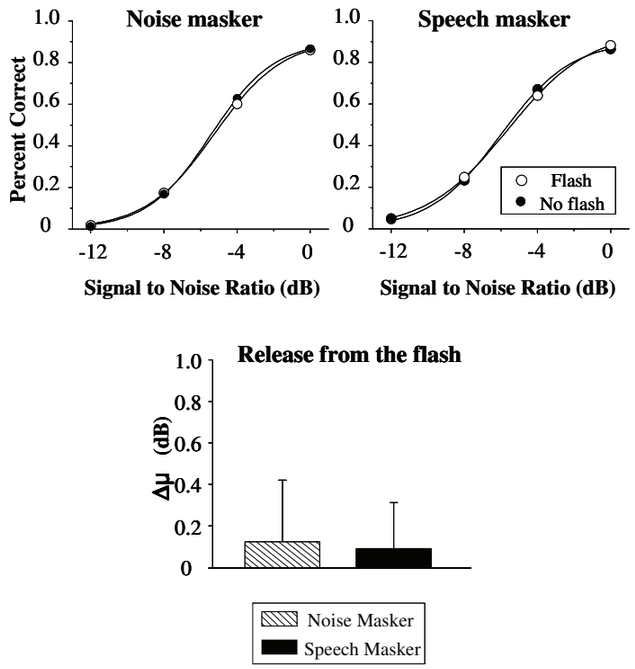

Figure 1. Top panels: Mean percent-correct word identification as a function of SNR, along with the best-fitting psychometric functions (curves) in noise (left panel) and speech masker conditions (middle panel) across two different light flash types, speech-synchronized light flash (open circle) and no flash (closed circle). Bottom panel: Amount of benefit from the speech-synchronized flash cues. Benefit is expressed as the difference in SNR in noise masker versus speech masker conditions for 50%-correct recognition of speech.

Experiment 2: Speech Recognition of rate-varied target speech with or without speech-synchronized light flash under noise or speech masking

Experiment 2.1 Comparison of speech recognition between constant-rate speech and varied-rate speech

Figure 2 shows the group-mean percent correct of speech identification as a function of SNR, along with the best-fitting psychometric functions (curves) for noise and speech masking conditions, when target speech was of constant rate or varied rate.

As indicated by the bottom panel of Figure 2, there was marked difference in the 50% correct threshold (μ) between constant-rate-speech recognition and varied-rate-speech recognition. A two-way ANOVA shows that the main effect of masker type was significant [F (1, 11) = 11.986, MSE = 13.486, p = 0.005], the main effect of target speech types was significant [F (1, 11) = 55.653, MSE = 42.147, $p < 0.001$], but the interactive between two factors was not significant [F (1, 11) = 0.005, MSE = 0.002, p = 0.945].

These results suggest that changing the speech rate within a target sentence significantly affected the speech identification performance under either noise or speech masking conditions.

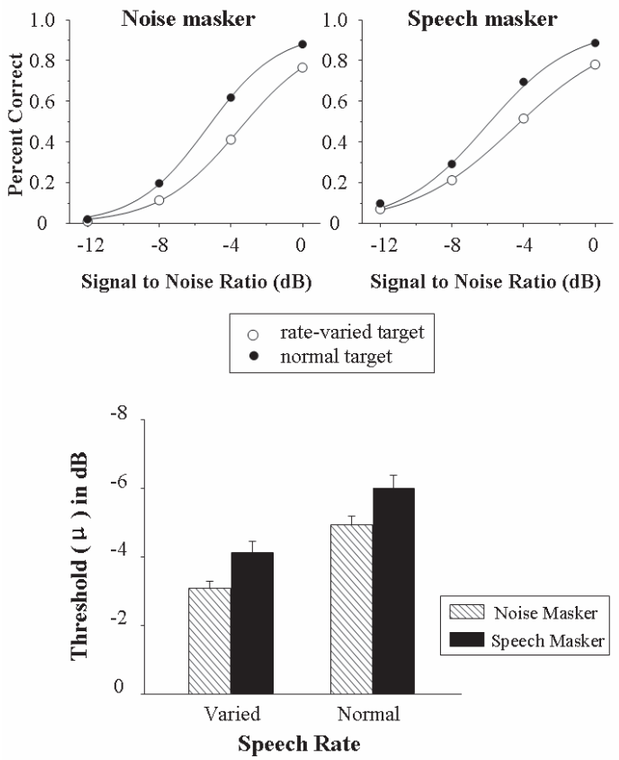

Figure 2. Top panels: Mean percent-correct word identification as a function of SNR, along with the best-fitting psychometric functions (curves) in noise (top left panel) and speech masker conditions (top right panel) across two different target speech types, rate-varied speech (open circle) and constant-rate speech (closed circle). Bottom panel: The SNR corresponding to 50% correct on the psychometric function across each condition.

Experiment 2.2 Speech recognition of varied-rate target speech with and without speech-synchronized light flash cue under noise or speech masking

Figure 3 shows the group-mean percent correct of speech identification as a function of SNR, along with the best-fitting psychometric functions (curves) for noise and speech masking conditions, when rate-varied target speech was combined with no flash, random flash, or syllable-synchronized flash.

As indicated in the two top panels of Figure 3, the "no flash" curve and the "random flash" curve almost overlap with each other under both the noise-masking condition and the speech-masking condition. At the same time, the "synchronized flash" curve clearly separated from other two curves under the two masking conditions.

A two-way within-subject test shows the interactive effect between masker type and flash type is significant [F (1, 11) = 3.921, MSE = 2.419, $p = 0.037$]. Separate one-way ANOVAs show that the flash type effect was significant under both noise-masking conditions [F (1, 11) = 35.746, MSE = 20.803, $p < 0.001$] and speech-masking conditions [F (1, 11) = 40.913, MSE = 21.346, $p < 0.001$]. Post hoc tests indicate that under either noise masking or speech masking, the threshold for the synchronized-flash condition was significantly different from that for the no-flash condition and that for the random-flash condition, but there was no significant difference the between the no-flash condition and the random-flash condition.

Relative to that under the no-flash condition, presenting the syllable-synchronized flash caused a 0.94-dB improvement under the noise masking condition and a 2.12-dB improvement under the noise masking condition.

These results suggest that when the rate of target speech is not constant, speech-synchronized flash cues can help listeners follow the target speech stream and reduce both energetic masking and informational masking.

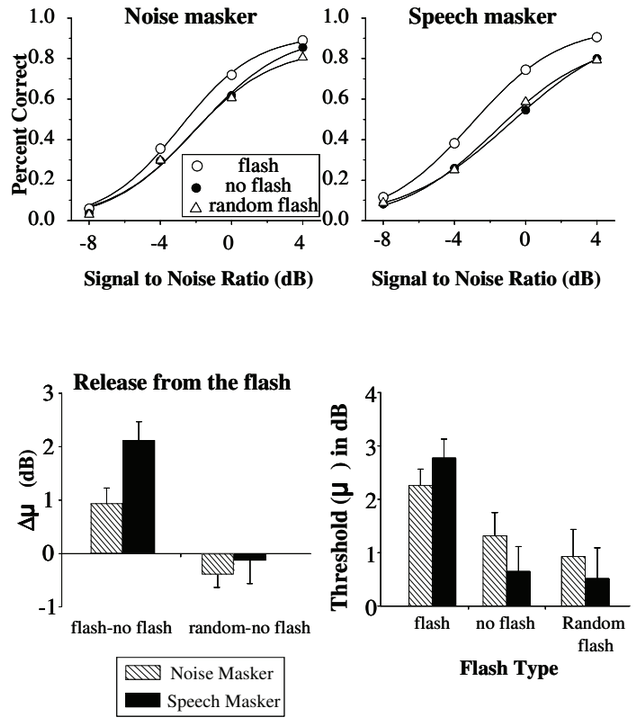

Figure 3. Top panels: Mean percent-correct word identification as a function of SNR, along with the best-fitting psychometric functions (curves) under noise masking conditions (top left panel) and speech masking conditions (top right panel) across three different light flash conditions: speech-synchronized flash (open circle), random flash (open triangle), and no flash (closed circle). Right bottom panel: The SNR corresponding to 50% correct on the psychometric function for each condition. Left bottom panel: The amount of benefit (Δμ) relative to the no-flash condition under noise masking or speech masking.

Conclusion

When target speech is of a constant rate, the speech-synchronized light flash had no unmasking effects. However, when the rate of target speech is artificially manipulated, the speech-synchronized light flash improves speech recognition particularly when the masker is the two-talker speech. Thus, when the speech cannot be predicted, speech-synchronized visual cues can play a role in releasing speech from masking, especially informational masking.

References

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

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

Helfer, K.S. 1997. Auditory and auditory-visual perception of clear and conversational speech. *J. Sp. Lan. 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.

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.

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 Communication, 49,* 892-904.

PARALLELS AND TRANSVERSAL SUBJECTIVE CONTOURS IN THE POGGENDORFF ILLUSION

Andrea Spoto, Alessia Bastianelli, Roberto Burro* and Giulio Vidotto
Department of General Psychology, University of Padua, Italy
** University of Verona, Italy*
andea.spoto@unipd.it

Abstract

When the continuity of an obliquely oriented line is broken by a vertically oriented pair of parallels, the position of the line segment on the other side of the interruption does not seem to be collinear, but vertically shifted (i.e. the Poggendorff illusion). Evidences from literature proved that the Poggendorff illusion is still present when there are no parallels but Kanizsa-like subjective contours. The present study attempts to verify whether the Poggendorff illusion persists when both the transversal segment and the parallels consist of Kanizsa-like subjective contour. Eight participants were tested using the method of constant stimuli a) on a number of horizontal subjective parallels and transversal line segments, b) on both horizontal parallels and transversal segment Kanizsa-like subjective contours. The response bias in the direction of the classical effect and the threshold values for the two patterns are discussed.

Geometrical optical illusions have always aroused interest. Since the 19th century, scientists have engaged in systematic investigations with the aim of revealing something more about human perceptual limitations. However, this is not the only reason, beginning with the Gestalt school, illusions have often been used as an instrument for testing the theory. Some illusions by now are considered to be classic, they are taught in schools and they are widely used and recognized. The Poggendorff effect (1860) is a robust illusion usually observed when the continuity of an obliquely oriented line is broken by a vertically oriented pair of parallels. Although several explanations have been proposed to account for this effect, this remains one of the most controversial geometrical illusions. In the past, the account most frequently proposed that the illusion arises from a misperception of the angles in the stimulus (Blackmore, Carpenter & Georgeson, 1970; Burns & Prichard, 1971). According to this explanation an overestimation of the acute angles in the standard stimulus (and an underestimation of the obtuse ones) probably has an effect on the apparent orientation of the line segments. Gilliam (1971) formulated the depth-processing theory based on the hypothesis that the geometrical information in the retinal projection might be mistakenly detected by the observers. The depth-processing theory suggested that geometrical illusions arise from the tendency of the perceptual system to process a two-dimensional figure as a representation of a three-dimensional scene. In regard to the Poggendorff illusion, the oblique lines in the stimulus configuration are interpreted as line which extend in depth and are therefore perceived to be non-collinear (Gregory, 1963; 1997).

According to Morgan (1999) the Poggendorff illusion arises because of retinal and cortical processes involved in the processing of relative position, orientation, and the collinearity of spatial separated lines and the object in general. This model suggests that the collinearity in the Poggendorff configuration is judged by comparing the orientation of the visible oblique lines with that of the virtual line joining the points of their intersection with the verticals. Morgan (1999) stated that the orientation of the virtual line can only be estimated from its endpoints and he proposed that the Poggendorff alignment is carried out as