Speaking rhythmically improves speech recognition under “cocktail-party” conditions

Mengyuan Wang, Lingzhi Kong, Changxin Zhang, Xihong Wu, and Liang Li

Citation: The Journal of the Acoustical Society of America **143**, EL255 (2018); doi: 10.1121/1.5030518
View online: https://doi.org/10.1121/1.5030518
View Table of Contents: http://asa.scitation.org/toc/jas/143/4
Published by the Acoustical Society of America

Speaking rhythmically improves speech recognition under "cocktail-party" conditions

Mengyuan Wang,[1] Lingzhi Kong,[2] Changxin Zhang,[3] Xihong Wu,[4] and Liang Li[5,a)]

[1]*Beijing Key Lab of Applied Experimental Psychology, School of Psychology, Beijing Normal University, Beijing 100875, China*
[2]*Allied Health School, Beijing Language and Culture University, Beijing 10083, China*
[3]*Faculty of Education, East China Normal University, Shanghai 200062, China*
[4]*Department of Machine Intelligence, Peking University, Beijing 100871, China*
[5]*School of Psychological and Cognitive Sciences, Beijing Key Laboratory of Behavior and Mental Health, Key Laboratory on Machine Perception (Ministry of Education), Peking University, Beijing 100080, China*
wangmengyuan@bnu.edu.cn, konglingzhi@blcu.edu.cn, changxin_zhang@126.com, wxh@cis.pku.edu.cn, liangli@pku.edu.cn

Abstract: This study examines whether speech rhythm affects speech recognition under "cocktail-party" conditions. Against a two-talker masker, but not a speech-spectrum noise masker, recognition of the last (third) keyword in a normal rhythmic sentence was significantly better than that of the first keyword. However, this word-position-related speech-recognition improvement disappeared for rhythmically hybrid target sentences that were constructed by grouping parts from different sentences with different artificially modulated rhythms (rates) (fast, normal, or slow). Thus, the normal rhythm with a constant rate plays a role in improving speech recognition against informational speech masking, probably through a build-up of temporal prediction for target words.

[MC]
Date Received: December 31, 2017 **Date Accepted:** March 19, 2018

1. Introduction

Under a noisy "cocktail-party" listening condition with multiple talkers, it is difficult for listeners to attend to and recognize target speech. Perceptual cues, for example, the perceived spatial separation and target-speech primes, facilitate selective attention to target speech and improve target-speech recognition (Freyman *et al.*, 1999; Huang *et al.*, 2008; Huang *et al.*, 2009; Huang *et al.*, 2010; Li *et al.*, 2004; Li *et al.*, 2013; Yang *et al.*, 2007).

Speech is inherently rhythmic with the temporal regularity in the range of 3–20 Hz (Greenberg and Arai, 2004). Particularly, the time interval between syllables is roughly equal with the averaged syllable rate around 4 Hz (Steeneken and Houtgast, 1980). The temporal regularity of the naturally spoken speech is a common feature shared by most languages in the world (Pike, 1945; Lin and Wang, 2007). As the temporal regularity makes the occurrence of syllables more predictable, the question raised here is whether the speech rhythm benefits our daily communication under "cocktail-party" listening conditions with multiple talkers.

Listeners can selectively attend to the sound around the time it is expected to occur while ignoring the masking background (Wright and Fitzgerald, 2004). The temporal rhythm and predictability have been suggested to be a grouping cue during auditory scene analyses (Rajendran *et al.*, 2013). Thus, it is of interest and importance to know whether the rhythm in a target sentence can be used as an unmasking cue.

It has also been known that recognition of the keyword in a target sentence is a function of the position in the target sentence under speech-masking (informational masking) but not noise-masking (energetic masking) conditions (Ezzatian *et al.*, 2012). This study was to investigate whether the temporal rhythm of target speech contributes to the build-up of an unmasking effect by examining the recognition of the keywords at different positions in the target sentence whose presenting rhythm is either naturally constant or artificially varied.

[a)]Also at: Beijing Institute for Brain Disorders, Beijing, China. Author to whom correspondence should be addressed.

2. Methods

2.1 Participants

Twelve university students (6 females and 6 males, mean age = 21.3 years) with normal hearing (pure-tone thresholds less than 25 dB hearing level between 0.25 and 8 kHz) participated in experiment 1 in which the recognition of target sentences with normal constant rate was measured against either a noise or speech masker. Another 12 university students (7 females and 5 males, mean age = 22.7 years) with normal hearing participated in experiment 2, in which the recognition of target sentences with an either normally constant rate (the same as the speech-masking condition in experiment 1) or varied rate was measured against the speech masker. They gave written informed consent and were paid a modest stipend for their participation.

2.2 Apparatus and stimuli

Semantically anomalous but syntactically correct Chinese sentences (Chinese nonsense sentences) were used. The English translations of these sentences are similar but not identical to the English nonsense sentences used by Freyman *et al.* (1999). For example, the English translation of a Chinese nonsense sentence is "Those directions always understand my gate" (the keywords are underlined). Each of the nonsense sentences has three keywords: subject, predicate, and object, with two characters for each keyword and one syllable for each character (Yang *et al.*, 2007).

Target sentences with a naturally stable rate (rhythm) were spoken by a young female talker (talker A) at an average rate of 5.4 syllables/s (with the standard deviation of 0.7 syllables/s). To create sentences with unstable rhythms, the rate of some normal sentences was modulated using the Synchronized Overlap-Add, Fixed Synthesis algorithm, which is a variation of the Synchronized Overlap-Add Algorithm algorithm (Hejna and Musicus, 1991) but requires significantly reduced computing time. It modulates the speech rate without introducing substantial changes in pitch, speaker identity, and intelligibility.

Three different rates were used for forming a rate-unstable sentence: 50% (slow, about 2.6 syllables/s), 100% (normal, about 5.0 syllables/s), and 150% (fast, about 7.7 syllables/s) of the stable rate. These rhythmically hybrid sentences with unstable rates were constructed by grouping three parts from three different sentences with different rates (slow, normal, fast) (Fig. 1). For each of the rhythmically hybrid target sentences, the order of the three speech rates was arranged randomly. Before the experiments, several experienced experimenters examined the subjective quality of the rhythmically hybrid target sentences and confirmed that both the naturalness and continuity of the artificially modified target sentences were the same as the naturally constant-rate sentences.

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 different young female talkers (talkers B and C). The noise masker was a stream of steady-state speech-spectrum noise.

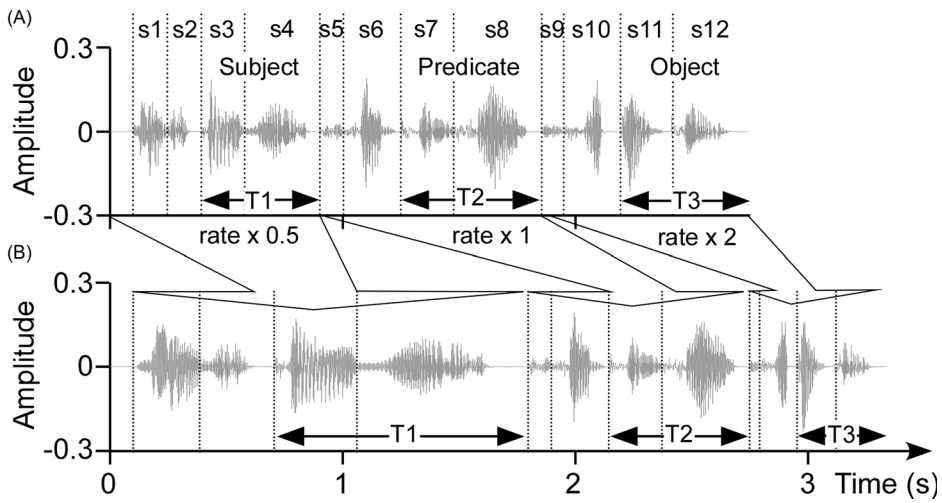

Fig. 1. Demonstration of a constant rate target sentence (A) and a varied rate target sentence (B). s1–s12 represent the 12 syllables in the target sentence. T1 (s3, s4) is the first keyword and T3 (s11, s12) is the last keyword.

The acoustic signals were delivered to a loud-speaker (Dynaudio Acoustics, BM6A) directly in front of a participant in an anechoic chamber (Beijing CA Acoustics, details see Yang *et al.*, 2007). Sound pressure levels (SPLs) were calibrated using a B&K sound level meter (type 2230). Target-speech sounds were presented at a SPL of 56 dBA. The SPLs of the maskers were adjusted to produce four signal-to-noise ratios (SNRs) (−12, −8, −4, 0 dB).

2.3 Design and procedure

Experiment 1 had two within-subject factors: masker type (noise, speech) and SNRs (−12, −8, −4, 0 dB) and experiment 2 also had two within-subject factors: target-rhythm type (constant rate, varied rate) and SNR (−12, −8, −4, 0 dB). Thirteen target sentences were used in each condition.

In each trial, the participant pressed a button to start the masker sound. About 1 s (mean = 1.14 s, standard deviation = 0.09 s) later a single target sentence was presented with the masker. Then the masker was gated off with the target. Participants were instructed to vocally repeat the whole target sentence as best as they could immediately after the target speech stopped. The number of correctly identified syllables in keyword one (T1) and keyword three (T3) (using the syllable-correct scoring scheme) was tallied later. There was one training session before the formal experiment.

3. Results

A logistic psychometric function [Eq. (1)] was fit to the mean data across the four SNR levels for each participant, where y is the probability of correct identification of keywords, 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,

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

Figure 2(A) illustrates group-mean percent-correct identification of T1 and T3 in target sentences with the naturally constant rate as a function of SNR (experiment 1), along with the group-mean best-fitting psychometric functions (curves). Figure 2(B) shows that the threshold μ of T3 recognition was significantly lower (more negative) than that of T1 under the speech-masking ($t = 3.453$, $p < 0.01$) but not noise-masking ($t = -0.183$, $p = 0.858$) condition.

The effect of speech rhythm was tested by comparing the recognition of T1 and that of T3 when the rhythm of the target sentence was either naturally constant or artificially varied under the speech-masking condition [experiment 2, Fig. 3(A)]. The T3 recognition was significantly better than the T1 recognition when the speech rate of target sentence was constant [$t = 3.274$, $p < 0.01$, Fig. 3(B)], confirming the speech recognition improvement by the shift from T1 to T3 shown in Fig. 2(B). In contrast, there was no significant improvement from T1 to T3 recognition when the speech rate of the target sentence was varied [$t = -1.314$, $p = 0.216$, Fig. 3(B)].

4. Discussion

In a "cocktail-party" listening environment, listeners use various perceptual cues to release target speech from informational speech masking. For example, the recognition of target speech is improved when there is a perceived spatial separation between the target speech and masking speech (Freyman *et al.*, 1999). The results of this study showed a release from speech masking when the spatial cue was absent. The listeners' performance of recognizing target keywords in natural speech improved as a function

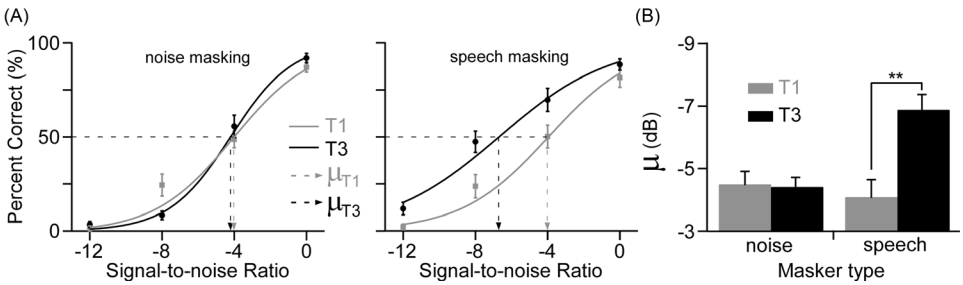

Fig. 2. Effect of masker type on the recognition of T1 and T3 in target sentences with normal constant rate. The error bars represent the standard error of the mean (SEM).

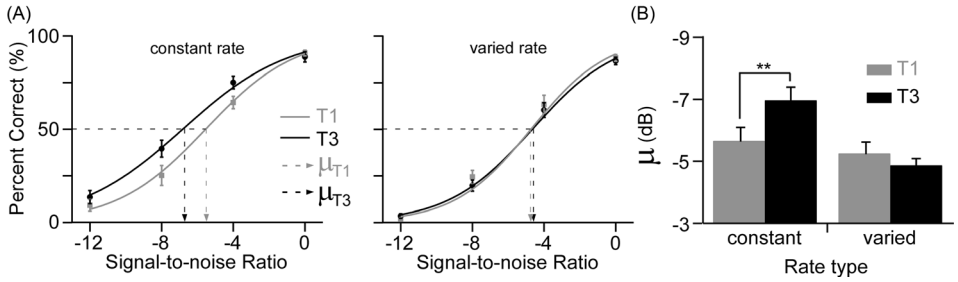

Fig. 3. Effect of rate type on the recognition of T1 and T3 in target sentences against speech masker. The error bars represent the SEM.

of keyword position: the recognition was lower for the first keyword and higher for the last one. This finding was consistent with the findings showing the build-up effect (Ezzatian *et al.*, 2012).

Human speech is inherently rhythmic (Greenberg and Arai, 2004). Listeners may be able to extract the rhythm information directly from the target speech sound to predict the timing of occurrence of forthcoming syllables. Indeed, the results of this study showed that the position-related improvement of keyword identification disappeared when the temporal regularity of the target sentence was artificially varied. This finding supports the previous report showing that temporal attention facilitates the detection of stimuli presented at the predicted time (Wright and Fitzgerald, 2004).

In a "cocktail-party" listening environment, perceptual cues that can be used by listeners to release target speech from informational speech masking include the regularity of speech rate which provides temporal cues for predicting the time when target syllables occur. However, it should be noted that the temporal cue is not as robust as spatial cues in unmasking target speech because the build-up effect becomes not significant when there is a perceptual spatial separation between the target and masker (Ezzatian *et al.*, 2012).

Note that in this study the constant-rate stimulation condition used in experiment 2 was the same as the speech-masking stimulation condition used in experiment 1. However, although the recognition of the last (the third) keyword in experiment 1 (with the threshold near −7 dB) was very similar to that in experiment 2 (also with the threshold near −7 dB), the recognition of the first keyword in experiment 1 (with the threshold around −4.0 dB) appears to be worse than that in experiment 2 (with the threshold around −5.5 dB). In addition to the participant variety (participants in experiments 1 and those in experiment 2 were different), it is not clear whether the difference in the coupled stimulation condition between the two experiments (i.e., noise masking and constant target-speech rate in experiment 1; speech-masking and unstable target-speech rate in experiment 2) was the significant cause for the performance difference between the two experiments.

Acknowledgments

This work was supported by the National Natural Sciences Foundation of China (Grant No. 31771252) and the Beijing Municipal Science & Tech Commission (Grant No. Z161100002616017).

References and links

Ezzatian, P., Li, L., Pichora-Fuller, M. K., and Schneider, B. A. (**2012**). "The effect of energetic and informational masking on the time-course of stream segregation: Evidence that streaming depends on vocal fine structure cues," Lang. Cognitive Proc. **27**(7–8), 1056–1088.

Freyman, R. L., Helfer, K. S., McCall, D. D., and Clifton, R. K. (**1999**). "The role of perceived spatial separation in the unmasking of speech," J. Acoust. Soc. Am. **106**, 3578–3588.

Greenberg, S., and Ari, T. (**2004**). "What are the essential cues for understanding spoken language?," IEICE Trans. Inf. Syst. **E87**, 1059–1070.

Hejna, D., and Musicus, B. (**1991**). "The SOLAFS time-scale modification algorithm," technical report (Bolt Beranek & Newman, University of Cambridge, Cambridge).

Huang, Y., Huang, Q., Chen, X., Qu, T.-S., Wu, X.-H., and 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., and 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. Perc. Perf. **35**, 1618–1628.

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

Li, H.-H., Kong, L.-Z., Wu, X.-H., and Li, L. (**2013**). "Primitive auditory memory is correlated with spatial unmasking that is based on direct-reflection integration," PLoS One **8**(4), e63106.

Li, L., Daneman, M., Qi, J. G., and 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. Perc. Perf. **30**, 1077–1091.

Lin, H., and Wang, Q. (**2007**). "Mandarin rhythm: An acoustic study," J. Chin. Ling. Comput. **17**, 127–140.

Pike, K. L. (**1945**). "The intonation of American English," in *Intonation*, edited by D. Bolinger (Penguin, Harmondsworth, UK), pp. 53–83.

Rajendran, V. G., Harper, N. S., Willmore, B. D., Hartmann, W. M., and Schnupp, J. W. H. (**2013**). "Temporal predictability as a grouping cue in the perception of auditory streams," J. Acoust. Soc. Am. **134**(1), EL98–EL104.

Steeneken, H. J. M., and Houtgast, T. (**1980**). "A physical method for measuring speech-transmission quality," J. Acoust. Soc. Am. **67**, 318–326.

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

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