2014

ISSN : 0974 - 7435

Volume 10 Issue 24

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(24), 2014 [16347-16351]

The auditory priming effect on the unmasking of mandarin Chinese speech

Meihong Wu[1], Wentao Ying[1], Junyu Wang[1], Jingfei Yang[1], and Zhiling Hong[2*]

[1]School of Information Science and Technology, Xiamen University, Fujian, P.R., (CHINA)

[2]Softwar School, Xiamen University, Fujian, P.R., (CHINA)

Email: hongzl@xmu.edu.cn

ABSTRACT

Human listeners are able to extract word sequences from running speech. Priming is an implicit memory effect in which exposure to one stimulus influences a response to another stimulus. In a noise environment, the auditory priming effect can significantly release speech from masking. There are two various kinds of masking that lead to speech-recognition difficulty under noise environment, including informational masking and energetic masking. Quantitative evaluating the auditory priming effect in the noise environment still poses a problem that the introduction of acoustic background noise induced stress causes speech recognition algorithms to fail. This study investigated a new quantitative computational method for the auditory priming effect in the speech recognition under noise conditions. This new computational method can help to improve the understanding of human auditory intelligence and then further provide more effective means to facilitate speech recognition of the selective target message in noise.

KEYWORDS

Energetic masking; Speech perception; Priming; Speech unmasking.

INTRODUCTION

Speech is the extremely important method of communication between humans. However, the transmission of speech can be affected by numerous factors, such as background noise, room reverberation, and distortions in the communication devices[1].

The priming paradigm provides excellent control over the effects of individual stimuli on cognitive processing and associated behavior because the same target stimuli can be presented with different primes.

The performance levels of most current speech recognizers degrade significantly when environmental noise occurs during use[2]. Recent years much effort has been directed to reducing this deficit in auto speech processing. In order to further understand the nature of speech processing, it is critical to find experimental paradigms to build an effective computational method for evaluating speech intelligibility in the noisy environment.

Humans usually feel it difficult to attend to target speech when there was multi-people talking. Processing of masking speech interferes with processing of target speech leading to impaired processing of target speech. Previous study has found that humans are able to take advantage of various perceptual cues to facilitate their selective attention to target speech and follow the target stream against masker inputs[3]. Priming is an implicit memory effect in which exposure to one stimulus influences a response to another stimulus. Auditory priming effect can significantly release target speech from not only informational masking but also energetic masking conditions[4].

Since features derived from speech have proven to be the most effective in automatic systems, in order to automatically extract information transmitted in speech signal, figuring out how the auditory system processes speech in noisy environment is crucial.

The analytic recognition strategies can be disrupted since parts of a sound can be masked by a background sound, and the background sound can also provide additional acoustic features that need to be discounted. Thus, building an effective computational method for evaluating speech intelligibility can help to understand how speech is processed and which parts of the speech signal are important for the successful recognition of the target message.

In this study, the auditory priming effect in the artificially synthesized speech recognition under the noisy environment was investigated. The speech corpus was generated by the speech-synthesis method based on Hidden Markov Model. The Hidden Markov Model[5] is one of widely used statistical models to model sequences of speech parameters by well-defined algorithms, and has successfully been applied to speech recognition systems. The two masking condition including energetic masking and informational masking was introduced in the design of this study. After gathering and analyzing the data according to the new subjective computational method, a detailed description to the results including data distribution was given before the conclusion section of this paper.

EXPERIMENTAL

Participants

Sixteen Mandarin-speaking young university students (10 females and 6 males, range from 19 to 23 years old) participated in this study. All the participants had normal (< 25 dB HL between 0.125 and 8 kHz) and bilaterally balanced (< 15 dB difference between the two ears) pure-tone-hearing thresholds.

Speech materials

The participant was seated at the center of an anechoic chamber (Beijing CA Acoustics Co. Ltd)[8]. The acoustic analog outputs were delivered to a loudspeaker (Dynaudio Acoustics, BM6 A, Dynaudio, Risskov, Denmark) in the central front of the participant. Speech stimuli were Chinese "nonsense" sentences and each of the sentences has 6 words including three key components: Subject+ Verb + Object.

Both target speech and priming speech were recited by artificially synthesized young-female voices, and acoustic signals of both target speech and priming speech for each of the three target young-female voices were generated by the Hidden Markov Model based speech-synthesis system.

About 1056 sentences from the Chinese nonsense sentences database was used for training. Speech signal was sampled at 16 KHz, windowed by a 25-ms Blackman window with a 5-ms shift, and then mel-cepstral coefficients were obtained by the mel-cepstral analysis technique[6]. A 5-state left-to-right HMMs with single diagonal Gaussian output a distribution was adapted. By considering relationship between static and dynamic features during parameter generation, smooth spectral sequences are generated according to the statistics of static and dynamic parameters modeled by HMMs, resulting in natural sounding speech[6].

The speech masker was a 47-s loop of digitally-combined continuous recordings for Chinese nonsense sentences and the noise masker was a stream of steady-state speech-spectrum noise.

Both priming and target speech sounds were presented at a level of 60 dBA. The sound pressure levels of maskers were adjusted to produce four signal-to-noise ratios (SNRs): -8, -4, 0, and 4 dB.

Experiment design and procedure

There were three types of priming conditions: (1) 0-priming condition: No priming sentence; (2) 1/3-sentence-priming condition: 1/3 length of the whole target sentence; (3) 2/3-sentence-priming condition: 2/3 length of the whole target sentence.

For a testing session, participants were informed of both the masking condition and the priming type. Each trial was started with the priming phase. A single target sentence and an either noise masker or speech masker were both started and terminated simultaneously. Participants' task was to determine which stimulus was the target sentence and the performance for each participant was scored on the number of correctly identified last keyword (the sentence's object part) of target sentences. Six whole-course target sentences out of 18 in a testing session were recited by each of the three target voices. About 384 priming sentences and 768 target sentences were used in this study.

RESULT AND DISSCUSS

A logistic psychometric function,

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

was fit to each of the 16 participants' data, using the Levenberg-Marquardt method[7], where y is the probability of correct identification of last keywords in target sentences, 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 illustrates group-mean percent-correct word identification as a function of SNR, along with the group-mean best-fitting psychometric functions (curves) under the syllable-correct scoring scheme when the masker was noise (left panel) or speech (right panel). As shown in Figure 1, under either energetic-masking or informational-masking conditions, word identification performance was substantially improved as the SNR increased from -8 to 4 dB. Also, under energetic-masking conditions, the word-identification performance was not affected by the priming-stimulus type. However, under informational-masking conditions, performance following the presentation of the 2/3-sentence prime was remarkably better than that following the presentation of 0-sentence prime or that following the presentation of the 1/3-sentence prime.

The differences in threshold μ between conditions were examined. Figure 2 shows that when under energetic masking condition, the thresholds were very similar across the three priming conditions. A one-way Analysis of variance (ANOVA) confirms that the effect of priming type was not significant ($F_{2, 30} = 1.039$, p=0.216). However, when the masker was human-like speech (informational masking),

the threshold under the 2/3-sentence-priming condition was much more negative than those under other priming conditions. A one-way ANOVA shows that the effect of priming type was significant ($F_{2, 30}$ = 7.521, p = 0.002<0.01). Post hoc analyses show that the threshold under the 2/3-sentence-priming condition was significantly better than that under the 0-sentence-priming condition (t_{15} = 3.453, p = 0.003; α was adjusted to 0.0167) and that under the 1/3-sentence-priming condition (t_{15} = 2.582, p = 0.011). There was significant difference in threshold between the 0-sentence-priming condition and the 1/3-sentence-priming condition (t_{15} = 4.698, p = 0.016).

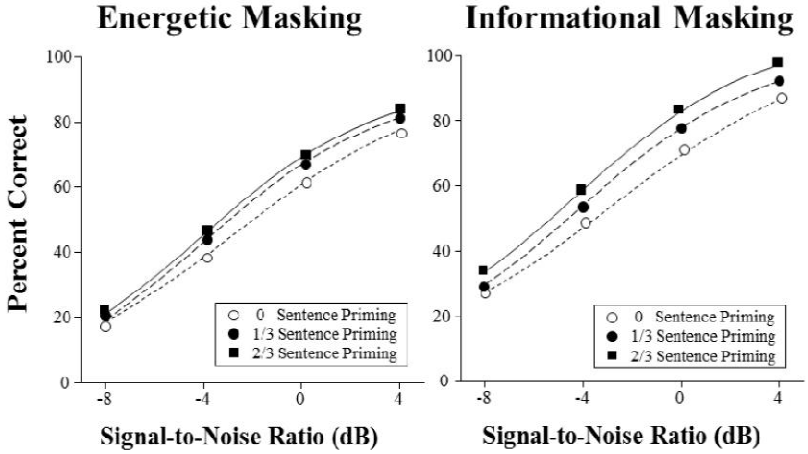

Figure 1: The mean percent-correct identification of the last target keyword across 16 listeners as a function of SNR for the three priming conditions when the masker was noise (energetic masking) (left panel) and when the masker was two-talker speech (informational masking) (right panel). The two panels also show the best-fitting psychometric functions (curves) under this three priming conditions

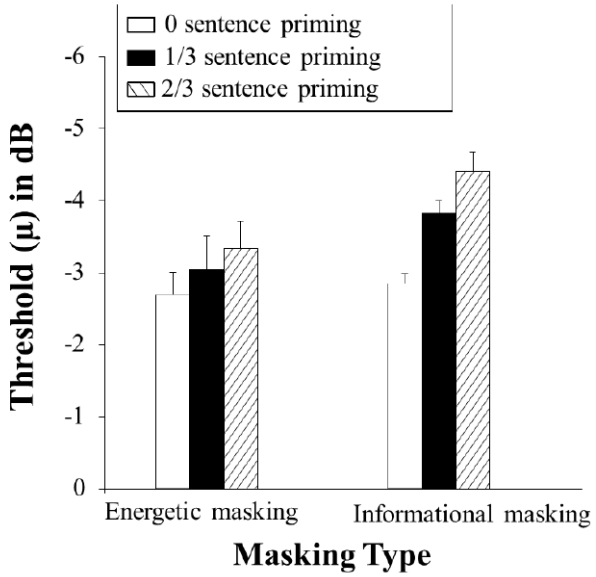

Figure 2: Average threshold values (μ) as a function of the type of masker and type of prime for recognizing the last target keyword. Error bars indicate the standard errors of the mean

This paper has introduced a subjective computational method as a tool for evaluating the relative contributions that auditory priming effect makes to the overall perception of complex speech stimuli. The results of this study show that auditory priming effect significantly improved speech recognition when the masker was human-like speech. Under each of the stimulus conditions, percent-correct word scores increased monotonically with the increase of SNR from -8 dB to 4 dB, without displaying

plateaus. The absence of non-monotonicity is in agreement with the results reported by Freyman et.al, 2004[4] and yang et.al, 2007[9].

CONCLUSIONS

Auditory priming effect is useful in specifically reducing informational masking in the noise environments. However, there was not an effectively computational method to evaluate the amount of unmasking effect under the adverse listening condition. This new proposed method can easily to find the helpful paradigm to facilitate the target stream recognition

It is well known that the introduction of acoustic background noise causes speech recognition algorithms to fail in the auto speech recognition process[2]. Subjective assessment can provide strong basis for objective assessment and this subjective computational method can provide a useful method for improving the automatic speech recognition systems.

ACKNOWLEDGEMENT

This work is supported by National Natural Science Foundation of China (31200769), the 2014 Program for New Century Excellent Talents in Fujian Province University, and the Open Funding Project of Zhejiang Key Laboratory for Research in Assessment of Cognitive Impairments (PD11001005002009).

REFERENCES

[1] A.W.Bronkhorst; The cocktail party phenomenon: A review of research on speech intelligibil-ity in multiple-talker conditions, Acustica, **86(1)**, 117-128 **(2000)**.

[2] Y.Gong; Speech recognition in noisy environments: A survey. Speech communication, **16(3)**, 261-291 **(1995)**.

[3] E.C.Cherry; Some experiments on the recognition of speech, with one and with two ears, J.Acoust.Soc.Am., **25**, 975-979**(1953)**.

[4] R.L.Freyman, U.Balakrishnan, K.S.Helfer; Effect of number of masking talkers and auditory priming on informational masking in speech recognition, J.Acoust.Soc.Am., **115**, 2246-2256 **(2004)**.

[5] T.Yoshimura, K.Tokuda, T.Masuko, T.Kobayashi, T.Kitamura; Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis, Proc.Eurospeech, **5**, 2347-2350 **(1999)**.

[6] Tokuda, Keiichi, Heiga Zen, Alan W.Black; An HMM-based speech synthesis system applied to English, Speech Synthesis, 11-13**(2002)**.

[7] S.Wolfram; Mathematica: A System for Doing Mathematics by Computer, Addison-Welsey, New York **(1991)**.

[8] X.H.Wu, C.Wang, J.Chen, H.W.Qu, W.R.Li, Y.H.Wu, B.A.Schneider, L.Li; The effect of perceived spatial separation on informational masking of Chinese speech, Hearing Research, **199**, 1-10 **(2005)**.

[9] Z.G.Yang, J.Chen, X.H.Wu, Y.H.Wu, B.A.Schneider, L.Li; The effect of voice cuing on releasing Chinese speech from informational masking, Speech Communication, **49**, 892-904 **(2007)**.

[10] S.Y.Cao, L.Li, X.H.Wu; Improvement of intelligibility of ideal binary-masked noisy speech by adding background noise, Journal of the Acoustical Society of America, **129**, 2227-2236 **(2011)**.

[11] T.Masuko, K.Tokuda, T.Kobayashi, S.Imai; Speech synthesis from HMMs using dynamic features, in Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP-96, Atlanta, GA, 389–392 **(1996)**.