ACOUSTICAL LETTER

Temporal characteristics of auditory spatial attention on word intelligibility

Ryo Teraoka[1,2,*], Shuichi Sakamoto[1], Zhenglie Cui[1], Yôiti Suzuki[1] and Satoshi Shioiri[1]

[1] *Research Institute of Electrical Communication and Graduate School of Information Sciences, Tohoku University,*
2–1–1 Katahira, Aoba-ku, Sendai, 980–8577 Japan
[2] *Research Fellow of the Japan Society for the Promotion of Science, Japan*

(*Received 30 September 2018, Accepted for publication 18 April 2019*)

Keywords: Auditory attention, Auditory selective attention, Cocktail-party effect, Auditory scene analysis

PACS number: 43.66.Lj, 43.71.Bp [doi:10.1250/ast.41.394]

1. Introduction

Humans have a remarkable ability to extract a specific sound from crowds of other sounds. Recent studies have revealed that auditory selective attention in the spatial domain (hereafter, "auditory spatial attention") plays an important role in this ability in a complex sound environment [1–3]. In fact, the authors have shown that listening performance in terms of speech intelligibility is enhanced when the auditory spatial attention is directed to a specific location from which target sounds are presented [4]. However, the whole mechanism behind this ability remains unclear.

To direct auditory spatial attention in the direction where a cue sound is presented, the listener must first localize the cue sound. Because it is known that the process of sound localization takes a certain amount of time [5], growth in terms of auditory spatial attention may be immature if the target sound is presented immediately after a cue sound is finished. However, the time course regarding how auditory spatial attention grows and is established has yet to be clarified. To investigate this issue, we examined whether the difference in time required to focus one's auditory spatial attention affects the word intelligibility of a target speech sound.

2. Methods

2.1. Listeners

Six male volunteers participated (mean age: 23.2 years). All listeners were naïve as to the purposes of the experiment. All listeners were native Japanese speakers with normal hearing acuity.

2.2. Apparatus and stimuli

The experiment was conducted in an anechoic room at the Research Institute of Electrical Communication, Tohoku University. The sound stimuli were presented through 12 loudspeakers, circularly distributed from $-180°$ to $+180°$ (positive value on the right side of the listener) with 30° separations at a distance of 1.6 m from the listener (see Fig. 1). The target sound was only presented from one of the five loudspeakers located at 0°, ±60°, and ±120°. Distractors were then presented from the other 11 loudspeakers.

The speech sounds comprised of four Japanese moras uttered by male and female speakers, extracted from "Familiarity-controlled word lists 2003" (FW03 [6]). One thousand words ranked as having the highest level of familiarity were selected from these lists. From these words, the target speech sounds were selected. These target words were those recorded from "Familiarity-controlled word lists 2007" (FW07 [7]), which is a compressed version of FW03 for clinical use. The total number of the target words was 400 (20 lists, with 20 words per list). The other 600 words were used as distractors. In this study, one female (fhi) and one male (mya) voice were assigned as the target and distractor, respectively. The amplitudes of the target and distractor utterances were adjusted such that the equivalent continuous A-weighted sound pressure level (L_{Aeq}) of the target and distractors were 70 and 65 dB, respectively, at the position corresponding to the center of the listener's head in the absence of the listener.

2.3. Procedure

One target word, uttered by the female speaker, was presented from one of twelve loudspeakers, and eleven distractors, uttered by a male speaker, were presented from the other 11 loudspeakers. The listener's task was to focus on the target speaker (i.e., the female speaker) and write down the uttered words on a response sheet exactly as they were heard. The heads of the listeners were not restrained, but the listeners were asked to keep their head stationary and face straight ahead at 0° during the entire session.

During the test session, the loudspeaker where the next target utterance would be presented was indicated beforehand. To indicate the direction of this loudspeaker, 50 ms white noise (a cue sound) was delivered from the loudspeaker prior to the presentation of the speech sounds. The duration between the cue sound and the presentation of speech sounds, namely, the inter-stimulus interval (ISI) was varied at 100, 200, 500, or 1,000 ms. The listeners were asked to direct their attention to the loudspeaker where the cue was presented. During the experiment, one list (20 words) was assigned as the target speech in each target direction and under each ISI condition. The session consisted of 100 trials, namely, five target speech directions (0°, ±60°, and ±120°) × four ISI duration conditions (100, 200, 500, and 1,000 ms) × five repetitions. That is, five lists (i.e., 100 words) were assigned as the target speech sound, and 1,100 words, which were selected from the 600 distractor candidates, were assigned as the distractors. The orders of the loudspeakers from which the target sound was presented and the words were randomized.

[*] e-mail: terar@dc.tohoku.ac.jp

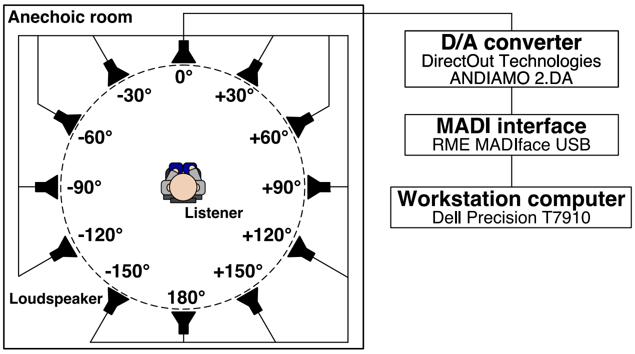

Fig. 1 Experimental setup.

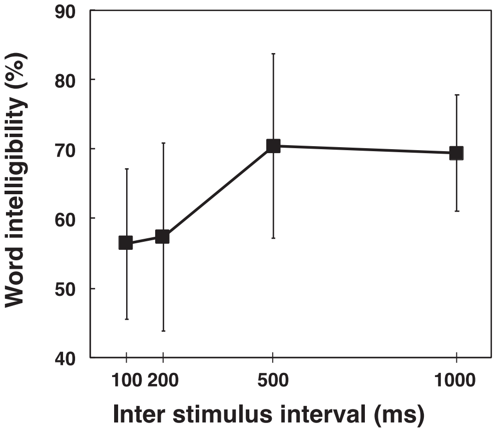

Fig. 2 Mean word intelligibility score as a function of ISI duration condition. Squares represent the mean word intelligibility scores across all listeners. Error bars denote the standard error of the mean.

3. Results and discussion

Figure 2 show the mean word intelligibility scores across all listeners as a function of the ISI duration condition. The squares and error bars represent the mean and standard error, respectively, across all listeners.

Seemingly, the word intelligibility increases until 500 ms, and then, over 500 ms, the word intelligibility decreases. A one-way repeated measures analysis of variance (ANOVA) was performed on the mean data using the ISI duration condition (100, 200, 500, and 1,000 ms) as a factor. The main effect is not significant ($F(3,15) = 1.25$, $p = 0.328$, $\eta^2_G = 0.006$). The results suggest that the improvement of the word intelligibility depends on the time required to maintain auditory spatial attention toward a specific location (i.e., the ISI).

In previous studies, the temporal characteristics of auditory spatial attention have demonstrated. Monder and Zatorre [8] measured the reaction time to a target sound presented from one of thirteen spatially distributed loudspeakers. The location where the target sound would be presented was indicated beforehand. The duration between the cue and the target sound (ISI) was varied (150, 600, 1,050, or 1,500 ms). The results show that reaction time declined with increasing the ISI. The authors hypothesized that a certain amount of time is required to fully engage the listener's attention at a specific spatial location. Consistent with the previous study, with ISI of 100 to 500 ms, the present findings suggest that such process to direct the listener's attention might affect the word intelligibility scores.

It is noteworthy that, in the present study, the word intelligibility score did not increase monotonically: the score declined with ISI of 1,000 ms ISI conditions. This finding is inconsistent with that in the previous study by Monder and Zatorre. One possible reason for this finding may be a blurring of auditory spatial attention. This will be an interesting topic of future study.

Acknowledgment

This research was supported by JSPS KAKENHI, Grant-in-Aid for Challenging Exploratory Research (No. 17K19990) and Grant-in-Aid for JSPS Research Fellow (No. 18J13203).

References

[1] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Act. Acust. united Ac.*, **86**, 117–128 (2000).

[2] A. W. Bronkhorst, "The cocktail-party problem revisited: Early processing and selection of multi-talker speech," *Atten. Percept. Psychophys.*, **77**, 1465–1487 (2015).

[3] M. Ebata, "Spatial unmasking and attention related to the cocktail party problem," *Acoust. Sci. & Tech.*, **24**, 208–219 (2003).

[4] R. Teraoka, S. Sakamoto, Z. Cui and Y. Suzuki, "Effects of auditory spatial attention on word intelligibility performance," *Proc. Int. Workshop Nonlinear Circuits, Communications and Signal Processing* (*NCSP'17*), pp. 485–488 (2017).

[5] S. M. Abel and P. J. Banerjee, "Accuracy versus choice response time in sound localization," *Appl. Acoust.*, **49**, 405–417 (1996).

[6] S. Sakamoto, Y. Suzuki, S. Amano, K. Ozawa, T. Kondo and T. Sone, "New lists for word intelligibility test based on word familiarity and phonetic balance," *J. Acoust. Soc. Jpn.* (*J*), **54**, 842–849 (1998) (in Japanese).

[7] T. Kondo, S. Amano, S. Sakamoto and Y. Suzuki, "Development of familiarity-controlled word-lists (FW07)," *IEICE Tech. Rep.*, **107**, 43–48 (2008) (in Japanese).

[8] T. A. Monder and R. J. Zatorre, "Shifting and focusing auditory spatial attention," *J. Exp. Psychol. Hum. Percept. Perform.*, **21**, 387–409 (1995).