

The effect of target speech distance on spatial auditory attention under multi-talker environment *

○ Florent Monasterolo, Shuichi Sakamoto, César D. Salvador, Zhenglie Cui, Yôiti Suzuki
(RIEC & GSIS, Tohoku University)

1 Introduction

Humans have the capacity to detect and concentrate on a desired sound or conversation, no matter how noisy the auditory environment is (Cocktail party effect [1]). This acuity is attributed to auditory selective attention, especially, to auditory spatial attention. A number of results have highlighted the benefits of spatial selective attention such as directional separation of sound sources and enhancement of sound intelligibility for a particular direction (e.g. [2, 3]). Yet, the effects of source distance on human auditory attention remain unclear.

The space within 1 m from the listener's head has a special status for distance perception. In this space, listeners become capable of relatively accurate distance judgement [4, 5]. In addition, the space within 1 m also corresponds to the adult peripersonal space (PPS), within which processes tend to change [6].

By working with virtual sound sources presented from within 1 m, Shinn-Cunningham *et al.* [7] and Brungart & Simpson [8] reported the benefits of distance separation of sound sources on speech reception threshold (SRT) and the importance of interaural differences in these benefits. This suggests effects of spatial unmasking for source separation along near distances within PPS.

We aim to investigate the effects of peripersonal distances on human auditory attention. The first experiment investigates the effects of the position of a target speech sound on reaction time (RT) when simultaneously presented with a distracting speech signal uttered by a speaker of the same gender. Both egocentric distance and source distance separation are studied. In the second experiment, with a similar test design, we examine the capacities of spatial auditory attention on distance by implicitly orienting the focus of attention on a predefined distance. In these experiments, we generated head-related transfer functions (HTRFs) for various distances by relying on distance-varying filters (DVF) [9, 10].

2 Methods

2.1 Apparatus

The sound stimuli were presented through Sennheiser HDA-200 headphones binaurally. The headphone transfer function was compensated for by convolving a 2048 point inverse filter calculated from the headphones' impulse responses. The sound stimuli were convolved with a set of individualized HTRFs in order to obtain spatialized virtual sound sources.

Individual head-related transfer functions (HTRFs) were measured for 1.5 m with a 5° azimuth resolution for each listener. Then, HTRFs for other distances were generated by relying on distance-varying filters (DVF) [9, 10]. Here, a 512 point DVF is applied to listeners' HTRFs for 1.5 m to generate near-distance HTRFs for 1 m, 0.5 m, 0.25 m, 0.13 m.

2.2 Spatial configurations

Two spatial configurations as described below are considered. They are illustrated in Fig. 1.

- **Same distance.** In this condition, both the target and distracter are set at the same distance. The distance from the head center to the virtual sound sources was either one of the following distances: 1 m, 0.5 m, 0.25 m, and 0.13 m.
- **Distance separation.** In this second condition, the distance of the target and the distracter sound was different. The distracter sound was always presented at 1 m from the center of the head and the target was set at either one of the following four distances: 1 m, 0.5 m, 0.25 m, and 0.13 m.

In both conditions, target and distracter were presented from the same direction, either from the front ($\theta = 0^\circ$), the left ($\theta = -90^\circ$) or the right ($\theta = +90^\circ$) side. These azimuths are chosen to diversify the effects of auditory parallax and of interaural level differences (ILD). In addition, sound stimuli with and without sound intensity cue expressed by the inverse square law were prepared. The A-weighted output sound pressure level of continuous sounds at the headphones measured with a BK 4153 artificial

* 標的音声の距離が競合音声存在下における聴覚の空間的注意に及ぼす影響。

Monasterolo Florent, 坂本修一, Salvador César D., 崔正烈, 鈴木陽一 (東北大).





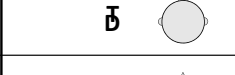
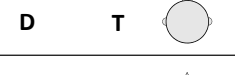
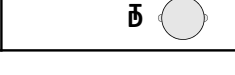
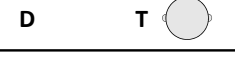
	Same distance	Distance separation
	$-90^\circ, 0^\circ, +90^\circ$	$-90^\circ, 0^\circ, +90^\circ$
1 m		
0.5 m		
0.25 m		
0.13 m		

Fig. 1 Spatial configurations considered in this experiment. In Same distance conditions (left), the distracter (D) and the target (T) are both presented from the same position. In Distance separation conditions (right), the distracter is fixed at 1 m and the target can be at any of four distances.

ear was set to 65 dB for each ear for the virtual sound source at 1 m and $\theta = 0^\circ$.

3 Experiment 1

3.1 Stimuli

In this first experiment, the distracter sound was synthesized using 6 streams consisting of continuously spoken words randomly chosen from the familiarity-controlled Japanese word corpus FW03 [11]. The words were spoken by a single male speaker and the word streams were overlapped for 8 seconds with random delays. The resulting sound was a meaningless 8 s long speech-like sound.

The target consisted of a single 4 mora word spoken by the same male speaker as the distracter. It was 1 to 1.2 seconds long. A distracter was always started first, followed by an additional target sound with a delay of random period ranging from 2 to 6 seconds.

3.2 Procedure

The experiment consisted of a series of trials, in which the above mentioned conditions were fully randomized. Each configuration (3 azimuths and 4 target distances) was heard 10 times for all conditions, that is, for Same distance and Distance separation conditions, and for the two intensity cue conditions. Therefore, the total amount of trials is 480. These were divided into 5 sessions whose lengths are less than 15 minutes so as to preserve the listener's attention as strong as possible.

The listeners were asked to respond as fast as possible via a gamepad button once they judged that they heard the target sound, which was informed

at the beginning of each session. If the measured RT fell outside of the interval 0 ms–2000 ms, this particular trial was considered as failed and was repeated later on during the session.

Listeners were 7 young and healthy adults with normal hearing (6 male, 1 female. Ages 21–24). They all were students belonging to the authors' laboratory and were unfamiliar to this type of reaction task before. All of these listeners also took part in an evaluation of localization accuracy using their own DVF filtered HRTF prior to this experiment. Moreover, all the listeners participated in a training session in prior. This training session consisted of 20 trials picked up at random from conditions included in the experiment.

3.3 Results

The mean RT for each listener for each condition are calculated and presented in Fig. 2. Subfigures a and b, and subfigures c and d, respectively, show the results of Same distance and Distance separation conditions. Moreover, subfigures a and c shows results for sources on the interaural axis ($\pm 90^\circ$), b and d show results for sources on the median plane (0°).

Target distance consistently acted to reduce mean RT for conditions including intensity ($F(3, 72) = 70, p < .001$) and excluding intensity ($F(3, 72) = 5.7, p < .005$) for Distance separation conditions and only when including intensity ($F(3, 72) = 7.9, p < .001$) for Same distance conditions. This suggests that distance separation of competing sound sources consistently leads to faster RT to target stimuli. This reduction could be as much as 44 ms when excluding the intensity cue. Azimuth had no consistent effect on results ($F(2, 12) = 1.12, p = 0.36$).

However, when comparing results for the interaural axis with those for the median plane, results show no statistically significant effects of distance for sounds from the median plane in Same distance condition (with intensity cue : $F(3, 18) = 1.3, p = 0.32$; without intensity cue : $F(3, 18) = 0.17, p = 0.91$) nor for Distance separation condition without intensity cue ($F(3, 18) = 1.02, p = 0.41$). On the other hand, distance affected RT significantly in every condition for sources on the interaural axis ($F(3, 18) = 41, p < .001$). This suggests that auditory parallax is not a usable cue for faster RT in any condition, whereas ILD are usable in both conditions. This also suggests that egocentric distance of both target and distracter signals affects to reduce RT by 27 ms when excluding intensity cue when they are presented from the interaural axis, revealing PPS effects on RT.

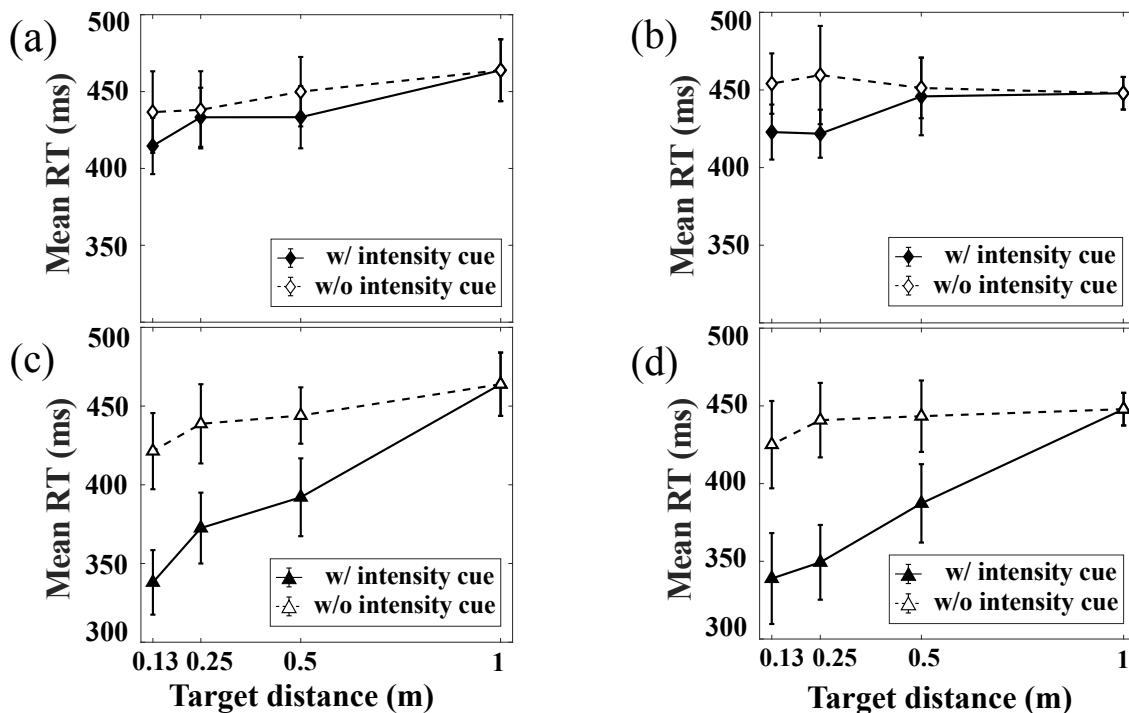


Fig. 2 Results of experiment 1: (a) and (b) Same distance condition, (c) and (d) Distance separation conditions, (a) and (c) results for sources on the interaural axis, (b) and (d) results for sources on the median plane. Filled and open symbols respectively correspond to results with and without the intensity cue. Vertical bars indicate \pm standard error.

4 Experiment 2

4.1 Stimuli

In experiment 2, a continuous 3 minute distracter sound was presented. During these 3 minutes, of the target sound were presented several times. Between two succeeding target sounds, 0, 1, 2 or 3 fake targets were presented from a random distance. The fakes were 4 mora words, randomly chosen from the word corpus, that differed from the target word. When presented with 2 or 3 fakes, the same fake word could not be heard twice.

4.2 Procedure

Only distance separation conditions without intensity cue were considered. Sounds were only presented on the interaural axis. The main difference from the previous experiment is that this experiment uses the probe-signal method [12] to implicitly direct the listener's focus on a predefined distance. The target is presented from the specified focus distance 80% of the time, and the remaining 20% from any other distances. The target is presented 12 times from each non-focus distance and 144 times from the focus distance resulting in a total of 180 trials per focus distances.

Two different focus distances were considered: 1 m corresponding to the limit of peripersonal space and 0.13 m being the closest considered distance. A session with no focus distance, consisting of stim-

uli with a uniform probability distribution for distance, was also included. The listener was asked to respond to each occurrence of only the target sound as fast as possible.

For experiment 2, 4 of the listeners participated in the previous experiment.

4.3 Results

For each listener, the average RT and false alarm (FA) rate are calculated. The mean RT and FA rate for all listeners are plotted in Fig. 3. Since the number of listeners is only 4, the standard error are quite large (13.4 ms to 91 ms) and are not plotted in Fig. 3 for better readability.

Results apparently show an effect of attending to a specified distance on both RT and FA rate. Focusing on 1 m leads to a reduction of 48 ms in RT to sounds presented from 1 m, and an increase of 40 ms for sounds presented from 0.13 m. It also reduces the FA rate for both 0.13 m and 0.25 m. Focusing on 0.13 m leads to an overall drop of RT by an average of 23 ms and a 1.7% increase of FA rate for sounds presented from 0.13 m. These results suggest that attending to near distance sound sources leads to faster reactions to all sounds presented from 0.13 m regardless of their nature, whereas attending to 1 m leads to a more selective form at the attended distance.

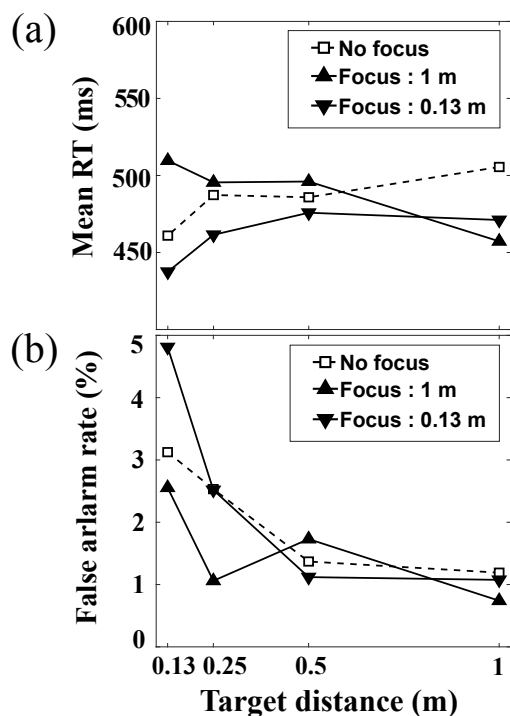


Fig. 3 Results of experiment 2: (a) mean RT, (b) mean FA rate. Open squares show the results when no focus on distance is attempted. Filled triangles show results for focus on 1 m (upper triangle) and 0.13 m (lower triangle).

5 Discussion and conclusion

Results in this study suggest that the properties of a target sound source near the listener's head affect to accelerate target detection tasks even with no increase in target sound intensity reaching the listener's ears. This acceleration is more consistently observed when the competing sound sources are separated in distance and when presented from the interaural axis. The present study shows that detection of targets becomes faster when competing sound sources are separated in distance than when all sounds are at the same distance. This can be attributed to benefits of sound source separation on spatial unmasking [7, 8].

Whereas previous studies [13, 14] showed faster RT for real sound sources located within PPS when using audio-tactile stimuli, the current study suggests that similar effects of source distance on RT are observable even in uni-modal virtual sound source presentation. This suggests that the properties of the sounds alone lead to PPS effects. The boundaries of auditory PPS using virtual sources remain unclear, and the interactions between direction and distance within PPS remain to be studied.

This study also found that listeners can attend to an auditory distance. This capability seems to shorten RT to sound sources presented from the dis-

tance of focus and longer RT to sounds presented from outside of the focused distance. This observation is similar to the effects of auditory attention to direction [2, 3]. Results suggest that the form of auditory attention along distance depends on the attended distance. Attending to very near distances might lead to faster reaction to all distances and to higher false alarm rate, whereas focusing on 1 m leads to faster reaction only for sounds presented from 1 m. Sounds in very near distances seem to be treated differently [6], and results from this study suggest that this difference is accentuated by voluntary auditory attention on their location. Further investigation on the distance at which processes change, as well as the mechanisms of distance auditory attention remain to be investigated.

Acknowledgment This research was supported in part by JSPS KAKENHI No. 17k19990 and the MEXT JASSO scholarship.

References

- [1] Cherry, J. *Acoust. Soc. Am.*, 25 (5), 975-979, 1953.
- [2] Ebata *et al.*, *J. Acoust. Soc. Am.*, 43 (2), 289-297, 1968.
- [3] Kidd *et al.*, *J. Acoust. Soc. Am.*, 118 (6), 3804-3815, 2005.
- [4] Brungart *et al.*, *J. Acoust. Soc. Am.*, 106 (4), 1956-1968, 1999.
- [5] Kim *et al.*, *Applied Acoustics*, 62, 245-270, 2001.
- [6] Graziano *et al.*, *Nature*, 397 (6718), 428, 1999.
- [7] Shinn-Cunningham *et al.*, *J. Acoust. Soc. Am.*, 110 (2), 1118-1129, 2001.
- [8] Brungart & Simpson, *J. Acoust. Soc. Am.*, 112 (2), 664-676, 2002.
- [9] Salvador *et al.*, *Acoust. Sci. Tech.*, 38 (1), 2017.
- [10] Salvador *et al.*, *Proc. Audio Eng. Soc. Int. Conf. Spatial Reproduction*, 2018.
- [11] Amano *et al.*, *NII Speech Resources Consortium*, 2003.
- [12] Greenberg & Larkin, *J. Acoust. Soc. Am.*, 44 (6), 1513-1523, 2016.
- [13] Canzoneri *et al.*, *PloS one*, 7(9), e44306, 2012.
- [14] Teneggi *et al.*, *Curr. Bio.*, 23(5), 406-411, 2013.