The effect of target speech distance on reaction time under multi-talker environment

Florent MONASTEROLO[†‡], Shuichi SAKAMOTO[†‡], César D. SALVADOR[†], Zhenglie CUI[†‡], Yôiti SUZUKI[†‡]

†Research Institue of Electrical Communications, and ‡Graduate School of Information Sciences, Tohoku University
2-1-1 Katahira, Aoba-ku, Sendai, 980-8577 Japan
E-mail:{fmonaste@dc., saka@ais.riec., salvador@ais.riec., sai@ais.riec., yoh@riec.}tohoku.co.jp

Abstract This study aims to investigate whether distance cues affect reaction time (RT) in a target sound search task. To do this, we investigated the effects of three cues, respectively, when the sound object was close to the listener. The cues considered in this study were the perceived source intensity as well as two binaural cues, interaural level differences (ILD) and auditory parallax. In the psychoacoustic experiments, a target and a distracter, both speech sounds, were spatialized using head-related transfer functions (HRTF). In the first experiment, a target and a distracter were presented to the listener from the same distance, and this distance was varied. Second, the position of the distracting speech sound was fixed at a far distance and that of the target speech soundwas moved closer to the listener. The results of the first experiment suggest that the sound intensity constantly influences to shorten RT. The results of the second experiment suggest that the main contributor to faster RT is the target to distracter ratio at both ears,and that spatial unmasking using binaural cues also shows a positive effect.

Keywords: auditory spatial attention, sound source distance, reaction time, auditory scene analysis, spatial unmasking

1. Introduction

In a complicated sound environment, sounds with certain characteriscs tend to attract human auditory attention. This type of human auditory attention is designated as bottom-up attention. It is often compared to top-down attention, the human capacity to focus on one sound source among many, increasing one's sensitivity to this source. Both these capacities are included in what Cherry [1] first named the Cocktail Party effect.

In psychoacoustic research, This issue has been intensively studied in the past years and many results based on psychoacoustic experiments have been accumulated concerning which sounds attract the most of human attention. Hearing one's name, for example, is known to involuntarily attract attention even when not attended to [2]. Asemi et Al. showed an auditory search asymmetry between temporal fluctuating sounds and pure tones [3], and between speech sounds and time-reversed speech sounds [4] suggesting that the nature of sounds play an important role in bottom-up attention.

The physical properties of sounds tend to affect this phenomenon. There is, for example, ample evidence that higher sound intensity leads to stronger and faster processes [5-7], and that sounds containing intense high-frequency components and high temporal contrast are perceived as more urgent and salient sounds [8].

Additionally, the sound source position is an important cue to auditory attention. As the source position changes, various physical properties of the sound reaching the ears change. Among them, the intensity reaching the listener varies with source distance following an inverse square law if there are no reflections, the head shadow effect leads to interaural level/intensity differences (ILD/IID). As sounds are moved in space at close distances, scattering on the pinna, head and torso of the listener systematically change depending on the sound source positions relative to the listener's ears, resulting in a modification of the spectrum of the sound reaching the eardrum. Moreover, parallax angles become larger for closer sounds.

We have focused on the effect of distance of sources, especially in space near to the listener for the following two reasons. First, in space within roughly 1 m, the head shadow effect, scatterings on the pinna, head and torso and auditory parallax become valuable localization cues [9-11]. Second, the space within 1 m also corresponds to the adult peripersonal space (PPS). This is the space within reachable grasping distance, within which processes tend to change [12]. Shin-Cunningham et Al. [13] evaluated source unmasking as a function of target-distracter distance separation for nearby speech sounds and found that binaural cues are a great contributor to reduction of speech reception threshold (SRT). Following this, Brungart and Simpson [14] showed that the cues to speech segregation for near distances depend on the nature of the sounds used and that for same-sex target and distracter, binaural cues can contribute to a reduction of up to 4-5 dB in SRT as sounds were separated in distance. They also showed that even when the target to distracter intensity ratio (TDR) at the better ear is kept

constant, distance separation leads to unmasking.

In this study, we investigated the effect 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. This research aims to clarify the following questions:

(1) Does distance of presented speech sound stimuli affect RT in a target detection task?

Unmasking of sounds along distance has been proved to lead to lower SRT. (2) Does a similar unmasking in distance affect RT?

2. Methods

To address the above-mentioned research questions, we conducted a psychoacoustic experiment in this study. This experiment comprises two condition. In one condition, a target speech sound was presented to measure reaction time (RT), for various distances including peripersonal ones, with a distracting multi-talker speech sound presented at the same distance as that of the target (Same distance condition). In the second condition, RT for a target speech sound was measured for various distances when a distracting multi-talker speech sound presented at a different distance from that of the target (Separate condition). For condition 1, we consider if listeners associate near-field sounds to intrusion of PPS and thus more urgent sounds. If this would be true, it should arouse more urgent processes, resulting in a faster reaction as sounds get closer. For condition 2 we expected to consider possible spatial unmasking by separating the two sounds along distance on RT. If such spatial unmasking would help separate the auditory streams, it should lead to faster reaction to the task.

2.1. Apparatus

The sound stimuli were generated using an RME BabyfacePro exterior sound card and headphone amplifier connected to a Dell Precision Tower 7910 desktop PC with 32GB RAM and Windows 10 installed. The sounds were presented through Sennheiser HDA-200 headphones diotically. The headphone transfer function was compensated for by convolving a 2048 point inverse filter calculated from the headphones' impulse responses using a B&K 4153 artificial ear and repetitive time-stretched pulse signals [15].

2.2. Stimuli

In this experiment, the distracter sound was synthesized using 6 word streams spoken by a male speaker randomly chosen from the familiarity-controlled Japanese word corpus called FW03 [16]. These word streams were overlapped with random delays. The resulting sound was a meaningless 8 seconds long speech-like sound.

The target consisted of a single 4 mora word spoken by the same male speaker as the distracter chosen from the familiarity-controlled Japanese word corpus FW07 [17]. It was 1 to 1.2 seconds long. A distracter was always started first, followed by a target sound with a delay of random period ranging from 2 to 6 seconds.

Accurate measurement of individual HRTFs is often difficult for distances within 1 m due to technical limitations. To overcome this limitation and conduct psychoacoustic tests in the PPS, we used a method based on circular harmonics to generate HRTF of peripersonal distances. This method is called distance-varying filters (DVFs) [18-19]. By applying DVFs to HRTFs measured in far distances, near-distance HRTFs are approximated. For test subjects to participate in the experiment, individual HRTFs were measured for 1.5 m with a 5° azimuth resolution. For each individual, a 512 point DVF filter could then be calculated for every centimeter closer than 1.5 m and with the same azimuth resolution as the initial measured HRTF set. These filters were applied to measured HRTFs constitute the set of synthesized HRTFs used during this psychoacoustic test. Virtual sound sources, to be spatialized at specified positions, were then synthesized by convolving digital sound signals with these HRTFs.

2.3. Spatial configurations

Condition 1 (Same distance condition)

In this condition, both the target and distracter were set at the same distance, where the distance from the head center to the virtual sound source was either one of the following distances: 1 m, 0.5 m, 0.25 m, and 0.13 m. Both target and distracter are presented from the same distance, either one of the following distances from the head: 1 m, 0.5 m, 0.25 m, 0.13 m. Target and distracter were simultaneously presented from the same direction, either from the front ($\theta = 0°$), the left ($\theta = -90°$) or the right ($\theta = 90°$) side. These azimuths were chosen to separate the effects of auditory parallax and of ILD. When sounds come from the median plane, ILD hardly vary with distance. On the other hand, when sounds are presented on the interaural axis, auditory parallax effects disappear. All configurations are illustrated in Figure 1.

Condition 2 (Separation condition)

In this second condition, a distance difference between the target and the distracter and this difference was varied. 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. Like in Same distance conditions, both target and distracter were presented simultaneously from the same direction. All configurations are illustrated in Figure 2.

In both conditions, sound stimuli with and without sound intensity cue which is expressed by the inverse square law were

prepared. To remove the sound intensity cue, for each sound source position, the intensity was normalized by the following value after convolving HRTF:

$$\sqrt{mean(x_{left}(t))^2 + mean(x_{right}(t))^2}$$

This value was used to equalize the total sound intensity reaching at the left and right ears. With this normalization, regardless of the distance and of the direction of the sound source, the overall sound intensity reaching the listener's two ears is expected to be fixed. To include the sound intensity cue, on the other hand, the result of the previous normalization was then multiplied by a factor following the inverse square law of sound intensity to distance, using the distance between the head center to the virtual sound source. The A-weighted output sound intensity level (sound pressure level) of the headphones measured with an artificial ear was set to present 65 dB for each ear for the virtual sound source at 1 m at θ = 0°.

2.4. Experimental procedures

The experiment consisted of a series of trials, where the above mentioned conditions were fully randomised. This series was divided into 5 sessions during which a specific configuration of the conditions could be heard twice. Each session lasted less than 15 minutes so as to preserve the subject's attention as best as possible. The subjects 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. This procedure was processed through a Matlab response interface and a computer gamepad connected to the desktop computer. 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. The response delay of the gamepad was not taken into account and was assumed to be constant.

Subjects were 9 young and healthy adults with normal hearing (8 male, 1 female. Ages 21~24). They all are students belonging to the authors' laboratory but had relatively little experiences of psychophysical tests. All of these subjects also participated in an evaluation of localization accuracy using their own DVF filtered HRTF prior to this experiment. They had been however unfamiliar to this type of reaction task before. For each subject, a prior training session was held with 20 trials picked up at random from the conditions included in the experiment.

All experiments were conducted in a double-walled sound-proof room in the Advanced Acoustic Information Systems Laboratory in the Research Institute of Electrical Communication of Tohoku University.

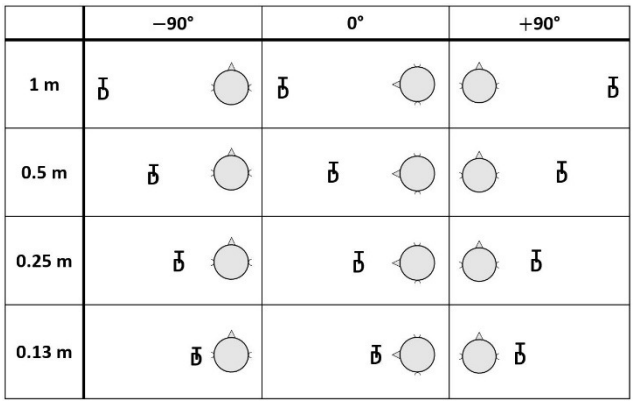

Fig. 1: Schema of Same distance condition (Condition 1). Both target (T) and distracter (D) are presented from the same position at either one of 4 distances (1 m ; 0.5 m ; 0.25 m ; 0.13 m) and at 3 possible azimuths (–90° ; 0° ; 90°).

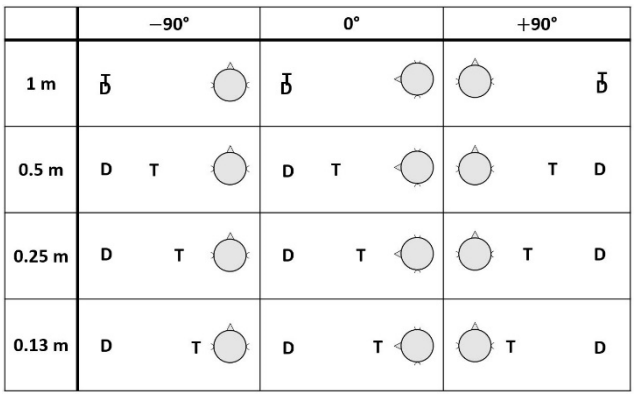

Fig. 2: Schema of Separation condition (Condition 2). The distracter (D) is always presented from 1 m and the target (T) is presented from either one of 4 distances (1 m ; 0.5 m ; 0.25 m ; 0.13 m). Both are presented from the same direction, one of 3 possible azimuths (–90° ; 0° ; 90°).

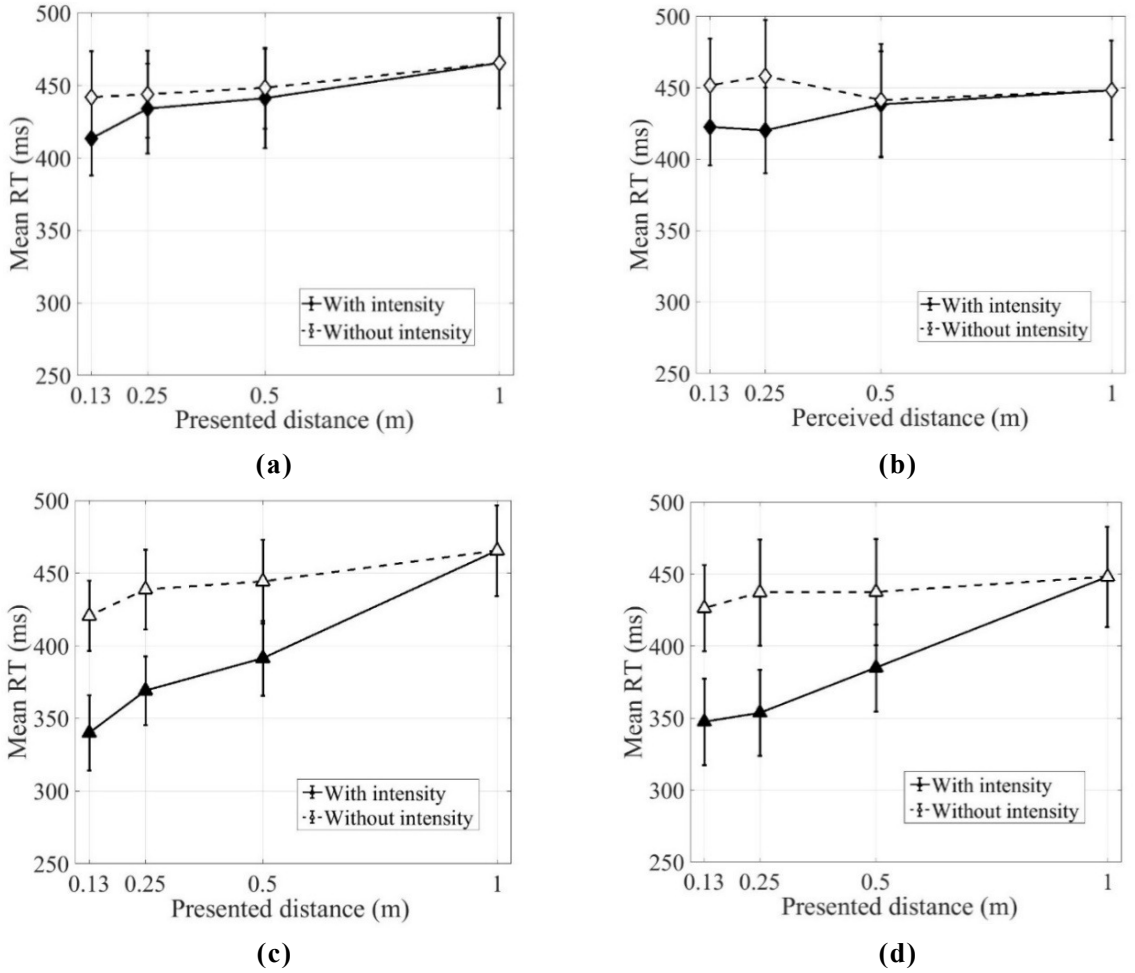

Fig. 3: Results for both Same distance condition and Separation Condition on the interaural axis and the median plane. The individual mean RT are averaged for each configurations and plotted as a function of presented distance. The vertical bar corresponds to the standard deviation for this particular point. Both conditions including the source intensity cue (With intensity) and excluding the intensity cue (Without intensity) are represented. **(a)** and **(b)** show results for Same distance conditions in respectively azimuths $\theta = \pm 90°$ and $\theta = 0°$. **(c)** and **(d)** show results for Separation conditions in respectively azimuths $\theta = \pm 90°$ and $\theta = 0°$.

3. Results

3.1. Analysis method

In each condition, the subjects' mean RT were averaged and the standard deviations were calculated. We then analyzed these values separately along the interaural axis (configurations with $\theta = \pm 90°$) and along the median plane ($\theta = 0°$) as a function of presented distance of the target sound source (Figure 3).

The data was submitted to a repeated-measures analyses of variance (ANOVA) with 3 within-subject factors: Condition (4 entries), Distance (4 entries) and Azimuth (3 entries). Results showed that the Condition factor and the Distance factor were significant: Condition ($F(3,32) = 3.736$, $p < 0.05$), Distance ($F(3,32) = 6.64$, $p < 0.01$). The azimuth factor however was not significant ($F(2,16) = 2.45$, $p = 0.095$). Two-way interactions were significant only for Condition×Distance ($F(9,72 = 15.74$, $p < 0.001$) and Distance×Azimuth ($F(6,48) = 2.90$, $p < 0.01$). Condition×Azimuth ($F(6,48) = 0.014$, $p > 0.9$). Three-way interaction was insignificant ($p > 0.9$).

3.2. Interaural axis

Results from the Same distance condition (Figure 3.a) showed a regular reduction of RT for both including and excluding source intensity sources were presented closer to the listener. For results including source intensity, a drop in RT could be observed between 0.25 m and 0.13 m. This could be attributed to the feeling of dangerous penetration of the PPS. Although conditions excluding source intensity show no statistically significant reduction of RT ($F(3,24) = 2.36$, $p = 0.097$), there is a strong tendency for RT reduction with closer sound sources. In the Same distance condition, sound unmasking could not be done using distance cues, yet a

reduction of RT for closer sound sources was observed. This illustrates the effect of source distance in decision processes. The contributors to this reduction could be the increase in level at the ipsilateral ear and to binaural cues for very near distances.

Separation conditions (Figure 3.c) show a clear change in RT with distance separation of the target and distracter. Both conditions led to a statistically significant evolution of RT indicating that binaural cues are a considerably salient cue to faster unmasking. In addition to the drop in RT between 0.25 m and 0.13 m, a drop in RT could also be observed between 1 m and 0.5 m.

3.3. Median plane

Same distance conditions (Figure 3.b) did not show any obvious contribution of binaural cues on RT reduction. When including source intensity, closer sounds had a minor effect on RT until very close distances. This suggested that even if including source intensity, presenting closer sound sources in the median plane did not affect greatly the target search task.

Separation conditions (Figure 3.d) excluding source intensity in the median plane suggest that parallax alone could be used for stream separation only for the highest distance separation ($t(8) = 2.12$, $p < 0.05$). Otherwise, reductions in RT were not significant. According to the feedback from the subjects, they did not perceive a distance separation between target and distracter in these conditions, but rather a sound image difference, especially for large distance separation. This suggests that, although parallax effects were poor for distance detection in these conditions, difference of sound image led to unmasking. When including source intensity, the main cue to distance unmasking is believed to be TDR at both ears. Indeed, for every halving distance an increase of 6 dB in sound reaching both ears is observed when in an anechoic environment. This leads to a TDR of about 10 dB at the ipsilateral ear and 3 dB at the contralateral ear for the largest distance separation, making it much easier for subjects to detect the target sound.

4. Overall interpretations and discussion

4.1. Sound stream separation

When both point sources are at the same position, spatial information cannot be used to separate target sound and distracter sound. However, in Separation conditions, the TDR increased greatly at both ears for closer sounds, leading the target to stand out compared to the background distracter. This benefited sound stream separation considerably, leading to faster reactions.

On the interaural axis excluding source intensity, there was a bigger RT reduction when the target was moved from 1 m to 0.5 m than when it was moved to distances under 0.5 m. This suggests that the highest benefits of unmasking are obtained by the action of mentally separating sounds into streams, and that once those streams are separated, more distance separation is less beneficial to process speed.

RT reductions could only partly be explained by the TDR. On the interaural axis, the highest variation in TDR occurs between 1 m and 0.13 m. When excluding source intensity, TDR is of 1.3 dB at the ipsilateral ear (better ear) and –8 dB at the contralateral ear. In addition, between 1 m and 0.5 m on the interaural axis, the TDR increase at the ipsilateral ear was minor (+0.3 dB). But, this small increase was enough to lead to considerable unmasking effects on RT. This leads to think that even the slightest increase in TDR coupled with binaural cues is enough to lead to considerable benefits in sound stream separation.

The distance separation condition excluding source intensity in the median plane is interesting. Although there is no statistically significant reduction of RT for closer sounds, a tendency can still be observed. In these conditions, no increase of TDR is consistent, yet we believe that a slight unmasking is possible. This can be attributed to a change in parallax between target and distracter, leading to a separation of perceived sound image and to ease of target detection.

4.2. PPS in a simulated environment

According to results in Same distance conditions, RT reduction due to presentation of near distance sounds was rather small until very close distances. This implies that penetration of the PPS does not accelerate decision times in a target search task until the sounds are presented from the very closest distances. This may also be explained by the limitations in sound space reproduction accuracy at such close distances, or the use of headphones that deform the perception of sound space.

5. Conclusion

The effect of distance of stimuli on simple RT in a target speech sound search task under multi-talker environment was studied.

When sounds were at very close distances, especially on the interaural axis, the RT dropped. This change in behavior could be attributed to a danger zone at very close distances or to the feeling of head penetration, leading to faster processes. However, this result was consistent only when including the source intensity cue.

Distance separation, even at a fixed perceived source intensity, led to faster RT. The main cue is believed to be the increase of TDR at both ears, but conditions excluding source intensity indicated that binaural cues such as different ILD values between target and distracter or parallax differences could also lead to unmasking. In these cases, rather than a perceived distance separation, a sound image differentiation could be the source of unmasking.

Ackowledgment

This research was supported in part by the Grant-in-Aid for Challenging Exploratory Research No. 17k19990 and the MEXT JASSO scholarship.

References

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears" from J. Acous. Soc. Am., Vol. 25 (5), pp. 975-979, 1953.

[2] N. Moray, "Attention in dichotic listening: Affective cues and the influence of instructions" from Quarterly journal of experimental psychology, Vol. 11 (1), pp. 56-60, 1959.

[3] N. Asemi, Y. Sugita, and Y. Suzuki, "Auditory search asymmetry between pure tone and temporal fluctuating sounds distributed on the frontal-horizontal plane" from Acta acustica united with Acustica, Vol. 89 (2), pp. 346-354, 2003.

[4] N. Asemi, Y. Sugita, and Y. Suzuki, "Auditory search asymmetry between normal Japanese speech sounds and time-reversed speech sounds distributed on the frontal-horizontal plane" from Acoustical Science and Technology, Vol. 24 (3),pp. 145-147, 2003.

[5] R. Chocholle, "Variation des temps de réaction auditifs en fonction de l'intensité à diverses fréquences" from Annales de Psychologie, Vol. 41, pp. 65-124, 1945.

[6] R. S. Woodworth, and H. Schlossberg, *Experimental Psychology*, Oxford and IBH Publishing, 1954

[7] M. J. Nissen, "Stimulus intensity and information processing" from Perception and Psychophysics, Vol. 22, pp. 338-352, 1977.

[8] C. Kayser, C. I. Petkov, M. Lippert, and N. K. Logothetis, "Mechanisms for allocating auditory attention: an auditory saliency map" from Current Biology, Vol. 15 (21), pp. 1943-1947, 2005.

[9] D. S. Brungart, and W. M. Rabinowitz, "Auditory localization of nearby sources. Head-related transfer functions" from J. Acoust. Soc. Am., Vol. 106 (3), pp. 1465-1479, Sep. 1999.

[10] D. S. Brungart, N. I. Durlach, and W. M. Rabinowitz, "Auditory localization of nearby sources. II. Localization of a broadband source" from J. Acoust. Soc. Am., Vol. 106 (4), pp. 1956-1968, Oct. 1999.

[11] H-Y. Kim, Y. Suzuki, S. Takane, T. Sone, "Control of auditory distance perception based on the auditory parallax model" from Applied Acoustics, Vol. 62, pp. 245-270, 2001.

[12] M. SA. Graziano, L. AJ. Reiss, and C. G. Gross, "A neuronal representation of the location of nearby sounds" from Nature, Vol. 397 (6718), p. 428, 1999.

[13] B. Shinn-Cunningham, J. Schickler, N. Kopčo, and R. Litovsky, "Spatial unmasking of nearby speech sources in a simulated anechoic environment" from J. Acoust. Soc. Am., Vol. 110 (2), pp. 1118-1129, Aug. 2001.

[14] D. S. Brungart, and B. D. Simpson, "The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal" from J. Acoust. Soc. Am., Vol. 112 (2), pp. 664-676, Aug. 2002.

[15] Y. Suzuki, F. Asano, H. Y. Kim, and T. Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses" from J. Acoust. Soc. Am., Vol.97(2), pp. 1119-1123, 1995.

[16] S. Amano, S. Sakamoto, T. Kondo, Y. Suzuki, "Development of familiarity-controlled word lists 2003 (FW03) to assess spoken word intelligibility in Japanese" from Speech Communication, Vol. 51, pp. 76-82,2009.

[17] S. Amano, T. Kondo, Y. Suzuki, and S. Sakamoto, "Familiarity-controlled word lists 2007 (FW07)" from The Speech Resources Consortium, National Institute of Informatics, 2007.

[18] C. D. Salvador, S. Sakamoto, J. Treviño, and Y. Suzuki, "A new signal processing procedure for stable distance manipulation of circular HRTFs on the horizontal plane" from Proc. Spring Meeting Acoust. Soc. Jpn, Yokohama, Japan: Acoustical Society of Japan , March 2016.

[19] C. D. Salvador, S. Sakamoto, J. Treviño, and Y. Suzuki, "Dataset of near-distance head-related transfer functions calculated using the boundary element method" from Proc. Audio Eng. Soc. Int. Conf. Spatial Reproduction —Aesthetics and Science—, Tokyo, Japan, August 2018.