INVESTIGATION OF SYSTEM LATENCY DETECTION THRESHOLD OF VIRTUAL AUDITORY DISPLAY

Satoshi Yairi, Yukio Iwaya and Yôiti Suzuki

Research Institute of Electrical Communication and
Graduate School of Information Sciences, Tohoku University,
Katahira 2–1–1, Aoba-ku, Sendai 980–8577, Japan
{yairi@ais, iwaya@fir, yoh@ais}.riec.tohoku.ac.jp

ABSTRACT

It is important in a virtual auditory display (VAD) system to reproduce not only static sound information, but also dynamic variation of sound. Thus, to achieve a highly precise virtual auditory display system, the system should be responsive to a listener's head movement. However, system latency (SL), in which the listener's head movement is reflected in the sound, certainly exists. If SL is detectable to the listener, it results in incongruousness. Consequently, the detection threshold (DT) of SL must be well investigated and SL should be sufficiently smaller than it. However, there have been relatively few studies on the DT of SL. Moreover, as inter-subject differences have been reported, it is necessary to examine DT in more detail. In this study, the DT and difference limen (DL) were investigated using two kinds of experiments and compared. As a result, averaged DT and DL over listeners were estimated to be 94 ms and 70 ms, respectively. Moreover, a strong correlation between the DT and DL (r=0.81 ($p < .01$)) was observed. This may mean that DL can be regarded as DT when the minimum system latency of the system is sufficiently small. Therefore, by taking the average of our results and previous studies, DT of SL was estimated as being around 75 ms.

1. INTRODUCTION

We can localize a sound position using HRTFs (Head Related Transfer Functions) empirically [1]. In virtual auditory display (VAD), a perceived sound position can be arbitrarily controlled by convolving of HRTFs to a sound source. To prevent cross-talk between two channels, headphones are usually used in a VAD system.

In sound localization, dynamic change of HRTFs is one of the most important cues [2]. Especially, when we move our head while listening, the accuracy of localization is markedly enhanced [3, 4, 5, 6, 7, 8, 9, 10]. That is, it is important in a VAD system to reproduce not only static sound information, but also dynamic variation of sound.

In the production of a virtual sound image via headphones, however, the whole virtual world moves according to the head movement if a listener's head movement is not reflected in the control of the sound position. This does not occur in the real world, and leads to the listener's perception of front-back error and/or lateralization [1, 11] as a result. Therefore, in a VAD system with headphones, a three-dimensional position sensor is usually employed to obtain information on the listener's head position and movement [5, 12, 13]. Appropriate HRTFs are then set according to the position of sound relative to head direction and position. This means that when a listener's head moves, a virtual sound image is simulated so as to move in the opposite direction to express a fixed real sound source. On the other hand, system latency (SL) or total system latency (TSL), in which the listener's movement is reflected in the sound, necessarily arises. SL is the sum of the times, after position data is updated, to interpolate HRTFs, to convolve the HRTFs to a sound source, and to output the data through buffers.

To render the virtual world more realistic by VAD, SL should be as short as possible. In particular, it is crucial to design and examine SL considering the detection threshold (DT), which is a minimum delay time to distinguish that the output is delayed. If the SL of VAD is longer than DT, the delay caused by SL is detected by the listener. As a result, the listener feels that the virtual sound image is not fixed in the virtual world, but moves according to head movement with a delay after his/her head has stopped. This does not occur for a fixed sound source in the real world. Therefore, it is important to investigate DT and make SL sufficiently smaller than DT.

Few researchers, however, have examined DT of SL. Besides, in most reported studies, the difference limen (DL), which is a threshold to distinguish two different delay times, was examined instead of DT and was regarded as DT. For example, Kimura *et al.* [14] applied paired comparison in their experiments and estimated DL to be around 80 ms. Brungart *et al.* [15] reported that the average listener is able to reliably detect an SL greater than 82 ms by comparison, the shortest SL being 11.7 ms. They also reported that the values of DL for nine listeners ranged from 60 ms to 120 ms. This indicates that there are large inter-subject differences in DL.

On the other hand, Sasaki *et al.* [16] examined DT and DL in two experiments and reported that with values of about 50 ms, the two generally agreed. However, since the number of listeners in studies by Sasaki *et al.* [16] was small, the inter-subject differences could not be well examined. If the DT is in good agreement with the DL, there must be large inter-subject differences in DT, too. It thus seems necessary to investigate DT in more detail, including the relationship between DT and DL.

In the present study, therefore, two kinds of experiments with absolute judgement and with paired comparison were applied to estimate DT and DL, respectively. Then, the relationship between DT and DL as well as their inter-subject differences were examined.

2. VAD SYSTEM

A software VAD system that we developed [17] was used in the experiments. The system consisted of a pair of headphones, a magnetic position sensor, and a personal computer (3.06 GHz Pentium 4 CPU, 2 GByte memory), as shown in Figure 1. The operating system of the personal computer was Linux (Kernel 2.6). Electrostatic open–back type headphones were used (STAX SRS-2020, earspeaker: SR-202 and driver unit: SRM-212). Polhemus FASTRAK was used as the magnetic position sensor. The receiver detects the magnetic field generated by a transmitter and calculates its relative x, y, z-position and yaw, pitch, roll-rotation to the transmitter. In our system, one receiver was installed on the top of the STAX headphones and 120 samples per second of position data could be acquired. The system latency of this VAD system was 9.93 ms, including the latency of the position sensor.

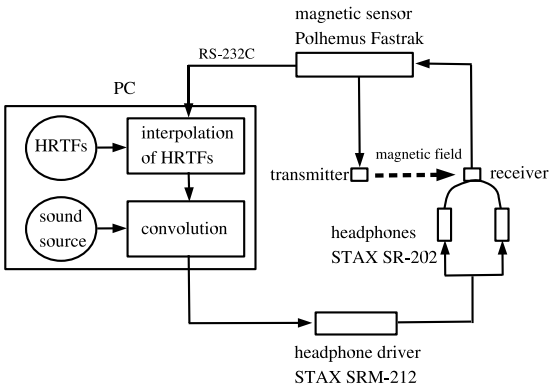

Figure 1: *System configuration.*

Impulse responses corresponding to HRTFs, *i.e.*, HRIRs (Head Related Impulse Responses), were measured with a spherical speaker array installed in an anechoic room of the authors' institute (Figure 2). HRIRs for sound sources located 1.5 m from the center of the spherical array were measured with an equal interval angle of 5 degrees in the horizontal plane and 10 degrees in the medium plane. As a result, HRIRs for 1154 (*i.e.*, 72×16+2) directions were measured for each listener. To realize smooth rendering, measured HRIRs were interpolated to obtain those for any directions with a resolution of 0.1 degrees [18]. HRIRs were temporally divided into a main response part and an initial delay part, so that the peak point in every direction of HRIRs becomes same at the main response part. The length of the main response of each HRIR was set at 256, with a sampling frequency of 48 kHz. On the other hand, the length of the initial delay part was variable and was recorded as integer value in terms of samples. Then, the main response part and the length of the initial delay part of each HRIR were separately loaded into the PC memory. Interpolated HRIRs were derived from four adjacent HRIRs, using linear weighting of the four main response parts and the four lengths of the initial delay part, respectively.

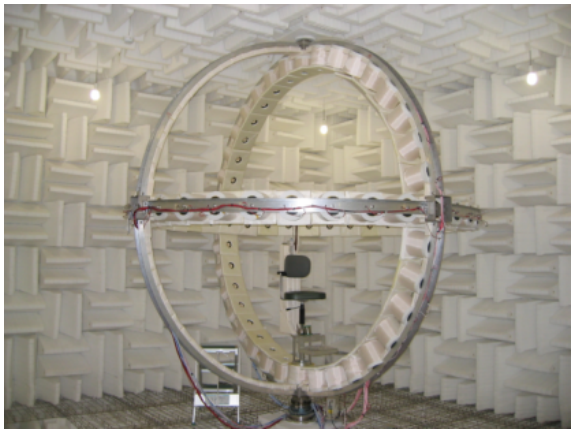

Figure 2: *Spherical speaker array to measure HRIRs.*

3. EXPERIMENTS

3.1. Measurement of detection threshold (DT) based on absolute evaluation

3.1.1. Method

A virtual sound source rendered with five kinds of different latencies was presented to a listener, who was then asked to judge whether he/she was aware of the delay of the sound stimulus or not.

In previous studies [14, 15, 16], the DL/DT was measured by applying the same range of delay times for all listeners. However, considering the observed inter-subject difference in DL [15], it must be better to select an appropriate range of delay times for each listener. Thus, in the present study, the range of delay times was set individually for each listener based on a preliminary experiment.

The maximum delay time, $SL_{\max}$, was determined for each listener as the time at which he/she could clearly distinguish a delayed sound stimulus from the minimum system latency, $\mathrm{SL}_{\min}$ (9.93 ms), by paired comparison. Then, five kinds of delay time between $\mathrm{SL}_{\min}$ and $SL_{\max}$ were prepared for each listener, namely, $\mathrm{SL}_{\min}$, $\frac{SL_{\max}+3\mathrm{SL}_{\min}}{4}$, $\frac{SL_{\max}+\mathrm{SL}_{\min}}{2}$, $\frac{3SL_{\max}+\mathrm{SL}_{\min}}{4}$, and $SL_{\max}$. In one trial, one of these five kinds of latencies was randomly selected and sound rendered with this latency was presented for 4 s. Each of the five kinds of latency randomly appeared six times in one session. Each listener participated in five sessions. As a result, one specific latency was presented 30 times to one listener in all the sessions.

The experiment was performed in a soundproof room. Listeners were six male and four female young adults with normal hearing. A listener sat on a chair in the room. Stimuli were generated by convolving the listener's own HRTFs to a sound source with the VAD mentioned in Sec. 2. The original sound source was a steady sound consisting of 100 harmonic tones with a fundamental frequency of 100 Hz. Thus the frequency spectrum ranged from 100 Hz to 10 kHz. Moreover, the envelope of the frequency spectrum decayed at the rate of -3 dB per octave. At the beginning of a trial, a virtual sound source was first presented in front of a listener. To regulate head movement, a listener was asked to move his/her head only one cycle (front-right-left-front or front-left-right-front)

during presentation of a stimulus.

3.1.2. Experimental results and estimation of detection threshold (DT)

Figure 3 shows the relationship between the system latency and the rate of a listener's awareness of the delay (positive answers). This figure shows that the rate of positive answers monotonically increased as the system latency increased. Such a monotonic relationship was observed for all of the listeners.

Detection thresholds are usually defined as the stimulus value when the rate of positive answers crosses the central rate between the chance level and 100%. It is typically 50% in absolute judgement and 75% in two–interval judgement. In the present study, however, since a certain positive rate was observed for the shortest system latency, the following procedure was applied to estimate the detection threshold from the experimental results.

1. A logistic function was fitted to the experimental results (solid lines in the panels of Figure 3).
2. b was defined as the rate at which the fitting line crossed the y-intercept.
3. DT was estimated as the system latency which corresponded to $(b + \frac{100-b}{2}) = c\%$ of the rate on the fitting line.

Table 1 shows the DT of each listener. The average DT for ten listeners was calculated as 94 ms. This result corresponds to the difference limen (DL) reported by Kimura *et al.* [14] and Brungart *et al.* [15]. Moreover, as also reported by Brungart *et al.*, a large inter-subject difference was observed, *i.e.*, the DT ranged from 42 ms to 132 ms as shown in Table 1.

3.2. Measurement of difference limen (DL) based on paired comparison

3.2.1. Method

Another experiment was performed using paired comparison. Conditions were almost same as in the case of absolute evaluation. A pair of different system latencies, which were selected from the five kinds of latencies stated in the previous section, was presented to a listener, who was asked to judge which sound stimulus in a pair had the larger delay time. Each stimulus interval in a pair was 4 s, respectively, and the inter-stimulus interval was 2 s as shown in Figure 4. Each of the ten kinds of combinations ($_5C_2$) for latencies randomly appeared two times in one session. Each listener participated in 15 sessions. Therefore, one specific combination was presented 30 times to one listener through all sessions. If the listener found it difficult to distinguish the difference, he/she could listen to the same pair of stimuli as many times as he/she liked. Actually, however, all the listeners could answer within one or two repetitions in almost all cases.

3.2.2. Experimental results and estimation of difference limen (DL)

Thurstone's law of comparative judgment (case V) [19] was used to calculate the interval scale from the experimental results. In this calculation to derive the interval scales, no correlations and equal discriminal variances between any of the stimuli were assumed. Such interval scales can be interpreted as relative psychometric distances between the stimuli. The scale values were then described in units of JND (Just Noticeable Difference). When the

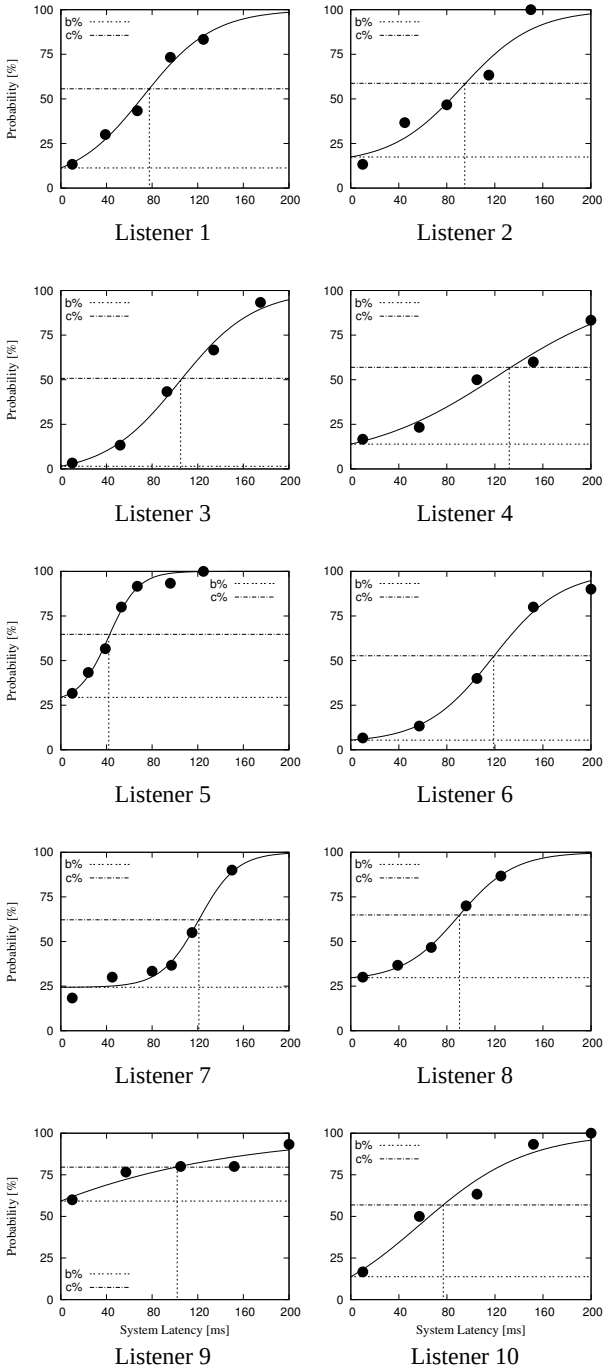

Figure 3: *Rate of positive answers as a function of system latency.*

Table 1: Detection threshold (DT) of each of the ten listeners and its average.

Listener	1	2	3	4	5	6	7	8	9	10	average
DT [ms]	76	95	105	132	42	119	121	91	92	73	94

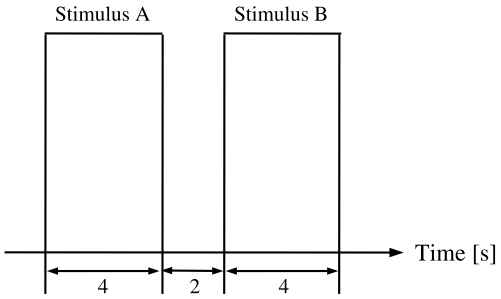

Figure 4: *Time chart of stimuli.*

distance between the scales for two stimuli is larger than unity, it means that a listener can distinguish the stimuli. Figure 5 shows the relationship between the system latency and the calculated rating scale.

The difference limen (DL) of each listener was then estimated by the following procedure:

1. Calculated scale values were connected by segments (solid segments in the panels of Figure 5).
2. The time x at which a segment crossed 1.0 JND was derived from each panel.
3. DL was estimated as $(x - 9.93)$ ms.

Table 2 shows the estimated DL of each listener. For example, in the case of Listener 6, x of about 100 was derived and then DL was estimated to be about 90 ms. For Listener 9, however, the rating scale at $SL_{\max}$ was only slightly larger than 1.0. This means that the latency difference between the two extreme conditions was just noticeable for this listener. Therefore, Listener 9 was excluded from the average calculation and the average was calculated for the other nine listeners.

Thus, the average DL for nine listeners was 70* ms. This value is consistent with the values reported by Kimura *et al.* (*ca.* 80 ms) and Brungart *et al.* (*ca.* 82 ms). A large inter-subject difference was again observed, *i.e.*, the DL ranged from 30 ms to 118 ms (or 174 ms) as shown in Table 2.

3.3. Comparison of detection threshold (DT) and difference limen (DL)

As a result of the two experiments, detection thresholds (DT) and difference limens (DL) were derived. In this section, we examine the relationships between them.

Averaged DT and DL were 94 ms and 70 ms, respectively. Since there were large inter-subject differences in both DT and DL,

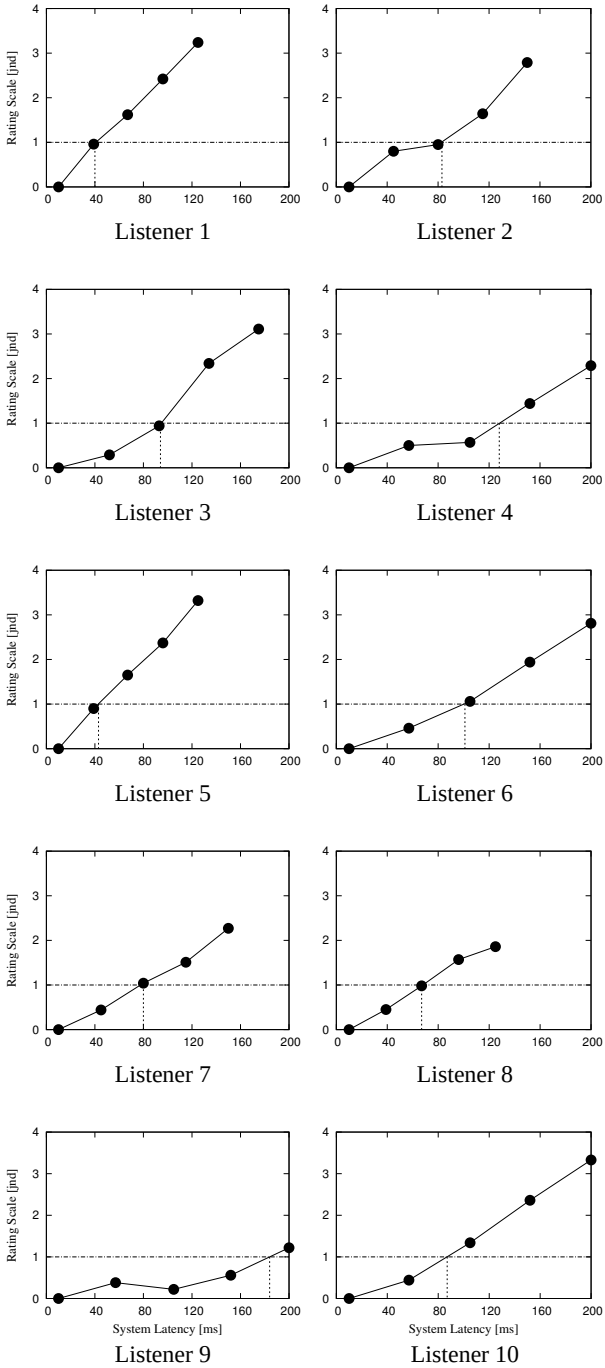

Figure 5: *Rating scale as a function of system latency.*

Table 2: Difference limen (DL) of each of the ten listeners as well as its average.

Listener	1	2	3	4	5	6	7	8	9	10	average*
DL [ms]	30	73	84	118	33	91	70	57	(174)	77	70*

(* Listener 9 was excluded when calculating the average)

a correlation coefficient between them was calculated to compare them for each of the listeners. Figure 6 shows the relationships between DT and DL of the nine listeners except for Listener 9. Correlation coefficient r is 0.81 ($p < .01$). Therefore, there was a strong correlation between DT and DL.

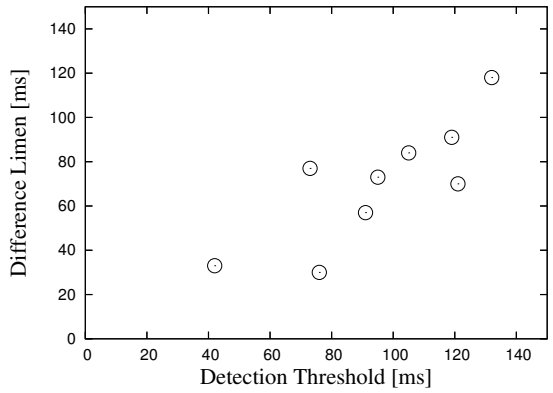

Figure 6: *Relationships between DT and DL (r=0.81).*

"To detect a delay" must have almost the same meaning as *"to distinguish a delay condition from a no-delay condition."* On the other hand, the DL estimated in the present study is that which corresponds to the latency distinguishable from minimum system latency. The minimum system latency (SL_{min}) of VAD which we applied was 9.93 ms, much shorter than the estimated DT and DL. Thus, if the SL_{min} condition can be regarded as a (almost) no-delay condition, the estimated DL in this study could be considered to be equivalent to DT.

Brungart *et al.* estimated DL to be 82 ms with a VAD system of which the minimum system latency was 11.7 ms. Moreover, Sasaki *et al.* applied a VAD system with a minimum system latency of 17 ms and estimated DL to be 50 ms. These TSLs are much shorter than the estimated DLs, and therefore we believe that these DLs can be regarded as directly corresponding to DTs. Given the overall average of these values, we conclude that the detection threshold of TSL in VAD can be estimated to be around 75 ms.

Kimura *et al.* applied a VAD with a minimum system latency of 56 ms and estimated DL as being around 80 ms. It is difficult to regard this TSL of 56 ms as a no-delay condition. Nevertheless, if a linear relationship between system latency and rating scale can be assumed, DL values can be regarded as representing DT even if the minimum system latency cannot be regarded as a no-delay condition. Actually, as shown in Figure 5, the rating scales of most of the listeners could be well fitted by linear functions. Thus, it is reasonable that the DL estimated by Kimura *et al.* (80 ms) is close to the detection threshold estimated in the present study (75 ms).

4. CONCLUSIONS

To investigate the detection threshold and difference limen, two types of experiments were performed. The average DT and DL for listeners were estimated to be 94 ms and 70 ms, respectively. These values are in good agreement with the DLs reported in previous studies.

A strong correlation was observed between DT and DL (r=0.81 ($p < .01$)). Since the minimum system latency of the VAD system which we applied (9.93 ms) is far smaller than these values, we regard our estimated DL to be equivalent to DT. Our results indicate that DT can be estimated by DL when the minimum system latency of the system is sufficiently small. Therefore, we conclude that the DL estimated by Sasaki *et al.* [16] and Brungart *et al.* [15] can be considered to be DT. Thus, by taking the average of our result of DT, DL and the DL estimated in two previous studies, we estimated the detection threshold of the system latency in VAD as being around 75 ms.

However, inter-subject differences which were observed in DT and DL were large. Thus, further investigation of details of inter-subject differences seems to be an important problem for future study.

5. ACKNOWLEDGEMENT

This study was partially supported by a Grant-in-Aid for Young Scientists (B) from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government (17700090).

6. REFERENCES

[1] J. Blauert, "Spatial Hearing," The MIT Press, 1983.

[2] H. Wallach, "On sound localization," *J. Acoust. Soc. Am.*, vol. 10, pp. 270-274, 1939.

[3] W.R. Thurlow and P.S. Runge, "Effect of induced head movement in localization of direction of sound," *J. Acoust. Soc. Am.*, vol. 42, pp. 480-488, 1967.

[4] W.R. Thurlow, J.W. Mangels and P.S. Runge, "Head movements during sound localization," *J. Acoust. Soc. Am.*, vol. 42, pp. 489-493, 1967.

[5] J. Kawaura, Y. Suzuki, F. Asano and T. Sone, "Sound localization in headphone reproduction by simulating transfer function from the sound source to the external ear," *J. Acoust. Soc. Jpn. (E)*, vol. 12, pp. 203-216, 1991.

[6] S. Perrett and W. Noble, "The effect of head rotations on vertical plane sound localization," *J. Acoust. Soc. Am.*, vol. 102, pp. 2325-2332, 1997.

[7] S. Perrett and W. Noble, "The contribution of head motion cues to localization of low-pass noise," *Percept Psychophys*, vol. 59, pp. 1018-1026, 1997.

[8] F.L. Wightman and D.J. Kistler, "Resolution of front-back ambiguity in spatial hearing by listener and source movement," *J. Acoust. Soc. Am.*, vol. 105, pp. 2841-2853, 1999.

[9] M. Kato, H. Uematsu, M. Kashino and T. Hirahara, "The effect of head motion on the accuracy of sound localization," *Acoust Sci & Tech*, vol. 24, pp. 315-317, 2003.

[10] Y. Iwaya, Y. Suzuki and S. Takane, "Effects of listener's head movement on the accuracy of sound localization in virtual environment," in *Proc. of the 18th International Congress on Acoustics*, 2004.

[11] M. Morimoto and Y. Ando, "On the simulation of sound localization," *J. Acoust. Soc. Jpn. (E)*, vol. 1, pp. 167-174, 1980.

[12] N. Asahi, H. Aoyama and S. Matsuoka, "Headphone hearing system to reproduce natural sound localization," *Technical Report of IEICE (in Japanese)*, EA79-24, 1979.

[13] M. Ohuchi, Y. Iwaya, Y. Suzuki and T. Munekata, "Training effect of a virtual auditory game on sound localization ability of the visually impaired," in *Proc. of the 11th Meeting of the International Conference on Auditory Display*, 2005.

[14] D. Kimura and Y. Suzuki, "An effect of delay time in an auditory display system on the perception of a sound image," *Technical Report of IEICE (in Japanese)*, EA2000-67, 2001.

[15] D.S. Brungart, B.D. Simpson and A.J. Kordik, "The detectability of headtracker latency in virtual audio displays," in *Proc. of the 11th Meeting of the International Conference on Auditory Display*, 2005.

[16] H. Sasaki, Y. Iwaya and Y. Suzuki, "Estimation of the detection threshold of latency of localization in a virtual auditory display," in *Proc. of Spring Meeting of Acoustical Society of Japan (in Japanese)*, 1-5-20, pp. 531-532, 2003.

[17] S. Yairi, Y. Iwaya and Y. Suzuki, "Relationship between head movement and total system delay of virtual auditory display systems," *Technical Report of IEICE (in Japanese)*, EA2005-38, 2005.

[18] K. Watanabe, S. Takane and Y. Suzuki, "A novel interpolation method of HRTFs based on the common-acoustical-pole and zero model," *Acustica Acta Acustica*, vol. 91, pp. 958-966, 2005.

[19] L.L. Thurstone, "A law of comparative judgement," *Psychol. Rev.*, vol. 35, pp. 273-286, 1927.