LOCALIZATION OF VIRTUAL SOUND CREATED USING INDIVIDUALIZED AND NON-INDIVIDUALIZED HRTF FOR DIRECT AND REFLECTED SOUND

Ryouichi Nishimura

National Institute of Information and Communications Technology / ATR Cognitive Information Sciece Laboratories
2-2-2 Hikaridai 619-0288, Japan
ryou@nict.go.jp

Hiroaki Kato

National Institute of Information and Communications Technology / ATR Cognitive Information Sciece Laboratories
2-2-2 Hikaridai 619-0288, Japan
kato.hiroaki@nict.go.jp

ABSTRACT

Good sound localization is an essential factor required for virtual auditory display (VAD) systems. These systems especially those based on the Head-Related Transfer Function (HRTF) often encounter the problem where the locations of virtual sound images are perceived at different locations to those that have been assumed. Considering the fact that reflected sound enhances the reality of virtual space, the accuracy of sound localization in a VAD system might be improved by presenting not only direct but also reflected sound. Therefore, we investigated what effect the presence of a single reflected sound had on the accuracy of the azimuthal localization of a virtual sound image. The results of subjective tests revealed that reflection created using a listener's own HRTF (individualized) is more effective for localizing sound than that created using someone else's HRTF (non-individualized). However, the performance was comparable with cases where only direct sound was presented.

1. INTRODUCTION

Listeners occasionally perceive a sound source that appears within their frontal hemisphere as if it were located in their rear hemisphere, and vice versa. This front-back confusion most likely occurs in VAD systems that are both based on HRTF and presenting sound through headphones because of ambiguity in the cues of primary interaural differences, particularly in interaural time differences. Although this problem becomes more remarkable when non-individualized HRTF rather than the individualized HRTF is used [1], it can be resolved by allowing listeners to move their heads because this can provide the information necessary to resolve the ambiguity.

Another solution to avoid front-back confusion is VAD systems that reconstruct the sound field itself. Wave Field Synthesis [2] and Boundary Surface Control [3] are two typical examples. Good sound localization with these methods can be expected because the head movements of the listener reproduce interaural differences that resolve ambiguity. While these types of systems are promising, they are not suitable for personal use because both Wave Field Synthesis and Boundary Surface Control need numerous speakers, amplifiers, D/A converters, as well as a special room where sound characteristics can be precisely determined. Hence, VAD systems with presentation with headphones are still necessary. While it is inevitable that these types of VAD systems will use HRTF data appropriate to the listener, measurements of HRTF usually need special equipment, such as an anechoic room and A/D and D/A converters, which involve a time-consuming process. Consequently, we need to be able to create good localization of virtual sound images without having to be bothered by such troublesome measurements.

It is inherently well known that the presence of reverberation enhances the reality of virtual space better than when nothing else but direct sound is presented. Reverberation usually consists of two parts; early reflections and late reverberation. According to a paper recently published [4], the boundary point between them is 70 to 300 ms . Also, early reflections contribute to the localization of the sound source in conjunction with direct sound more than late reverberation. One auditory phenomenon demonstrating this contribution is referred to as the precedence effect. According to the review by Litovsky *et al.* [5], the precedence effect can be classified into three phenomena; fusion, localization dominance, and lag-discrimination suppression. Fusion is the phenomenon whereby two temporally consecutive sound signals, i.e., leading and lagging sound stimuli, are perceived as a single sound image rather than two separate images. Localization dominance is the phenomenon whereby the localization of the sound source is dominated by either a leading or a lagging stimulus depending on the relative relations between their acoustical properties. Usually, the leading sound stimulus contributes to sound localization more strongly than the lagging sound stimulus. This superiority of leading sound in sound localization has been called the "law of the first wave front" as well as localization dominance. Finally, discrimination suppression is the phenomenon where the human ability to discriminate a change in the acoustical characteristics of the leading or lagging sound stimulus, is affected by the presence of other stimuli.

In this paper, the effect a single early reflection has on the sound localization of the perceived sound image on a virtual auditory display is investigated taking into consideration the localization dominance of the three types of precedence effect. Moreover, we especially focus on individualized and non-individualized HRTFs with which direct and reflected sounds are created. The findings we discovered in this investigation should help in designing virtual sound that is to be presented by a VAD system when the individualized HRTFs of users are only available for a limited number of directions.

2. PERCEPTION OF TWO SOUND SOURCES

2.1. Precedence effect

According to several reports on the precedent effect [5] [6] [7], fusion echo thresholds are generally 30 to 50 ms for speech signals and 5 to 10 ms for clicks. In addition, the longer the duration of noise, the longer the threshold of the fusion echo [8]. For example, the threshold is 5 to 6 ms for a 20-ms duration, 12 ms for a 50-ms duration, and 22 ms for a 100-ms duration. In localization dominance, Litovsky *et al.* reported that the lead location was chosen from 75% of trials when the interval between the leading and lagging sound stimuli was about 10 ms.

We often experience localization dominance in our daily lives. That is, the location of a sound source can correctly be identified even when several reflections are heard at the same time, implying that the reflections contribute relatively little to directional information. While the leading and lagging stimuli we used in the experiments were often clicks of equal amplitude and identical waveforms, there have been some reports on the effect of cross frequency and uncorrelated leading-lagging stimuli on the suppression of lag discrimination. The results from these reports have suggested that the suppression of spatial information contained in the lag is strongest when the lead and lag have a similar spectrum. There has also been a study suggesting that the suppression of discrimination depends on the relative locations of the lead and lag.

2.2. Plausibility theory

While many studies have implied that precedence effects originate from the suppression process by the leading stimulus to the lagging stimulus, there is a theory called "plausibility theory" that has tried to account for parts of the large variations in the time-intensity trading ratio within the context of sound localization [9]. In short, the theory suggests that decisions on sound localization are made by integrating the plausibility of cues from the interaural time difference (ITD) and the interaural level difference (ILD). Extending plausibility theory and considering contributions of lag stimuli to localization that can not be ignored, the reliability of the perceived direction of virtual sound would increase if reflections were created by individualized HRTF even when direct sound was created by non-individualized HRTF. This speculation motivated us to conduct subjective tests to clarify whether this assumption was correct.

3. VIRTUAL AUDITORY DISPLAY SYSTEM

3.1. System configuration

To carry out the experiments, an auditory display system that had capabilities of creating multiple virtual sound images and presenting them to a listener without perceptible time delay was required. Although there are commercial virtual auditory display systems currently available, we used one developed by a group at Tohoku University because of its low latency. Since the original system had no capabilities for creating multiple virtual sound sources, we implemented this capability for the experiments.

3.1.1. Hardware configuration

The system was constructed with the following hardware. Head movement was tracked using an InterSence IS-900 SimTracker, which is an ultrasonic sensor with a position resolution of 0.75 mm and an angle resolution of 0.05 degree. The static accuracy was 1.0 to 3.0 mm for position, 0.25 degree for the pitch and roll angle, and 0.50 degree for the yaw angle. The tracked position data were obtained at a refresh rate of 180 Hz and transmitted to a host computer connected via an RS-232C cable with a latency of 4 ms. The API for the sound driver was obtained from the Advanced Linux Sound Architecture (ALSA) project. The VAD program read the position data every 1 ms and generated sound signals for the left and right ears to present a virtual sound image. The HRTF to be convolved with a source signal was calculated by interpolating the HRTF data from four adjacent positions observed every 1 ms. The HRTF data were extracted from an HRTF database that contained free-field transfer functions [10] for several listeners including the listener who participated in the experiments, and measured every 5 degrees for azimuth and every 10 degrees for elevation within a distance of 1.5 m in an anechoic room.

The HRTFs were interpolated using the method proposed by Watanabe *et al.* [11] to achieve smooth change in the perceived sound as listeners moved their head.. The time lags to the onset of the transfer function for all directions were preliminarily calculated after up sampling and then clipped from the raw data. These clipped data and the time-lag information were stored separately. For each of the clipped signals and the time-lag information, those corresponding to the four adjacent positions were lineally weighted and summed separately, and then combined to obtain the transfer function for the present position of the virtual sound source relative to the listener. The sound signals were generated at a sampling rate of 48 kHz and converted to an analogue signal using a Roland UA-101 connected to the control PC via a USB cable. The control PC was a DELL PRECISION 650, that contained four 2.80-GHz Intel(R) Xeon(TM) CPUs.

3.1.2. Software configuration

The VAD program, which ran on the Linux 2.6.22 kernel, had the capability of simultaneously generating two independent virtual sound sources without any loss of performance due to its multi-thread programming and synchronization protocol using pthread_mutex. Listeners heard the signal through a Sennheiser HDA200 headphones. They were asked to move their arm toward the direction of the perceived sound image holding a position sensor in their hand. The direction they pointed was calculated from the relative locations of two position sensors: the first in the hand and the second on the headphones. The main loop of the program for the virtual auditory display used in the experiment is outlined in the flowchart in Fig. 1.

Due to the independent thread for each sound source, it was possible to apply different source signals to sources at different angles. However, in the current experiment the same source signal was applied to generate direct and reflected sound.

3.2. Correction of headphone characteristics

The influence of wearing headphones was compensated for using the method proposed by Iida *et al.* [12]. Using this, the transfer function of the ear canal in wearing headphones was equalized to that assuming free-field listening. The transfer functions between the microphones at the entrance of the ear canal and at the ear drum of a dummy head were first measured under all condition outlined in Fig. 2.

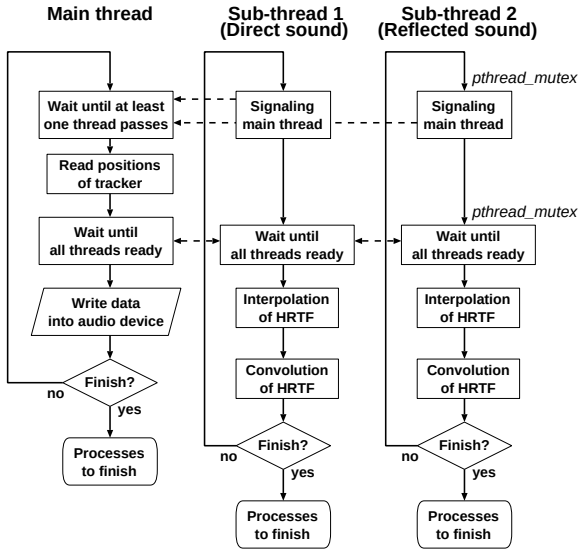


Figure 1: *Flowchart of the main loop of the program for the VAD used in the experiment.*

An Optimized Aoki's Time Stretched Pulse (OATSP) [13] of 2048 points at a sampling rate of 48 kHz was generated in the measurement and presented 20 times through Sennheiser HDA200 stereo headphones, which are the same ones we used in subjective experiments we conducted later, or we used a DIATONE DS-107V loud speaker for presentation.When measuring the sound pressure signal at the ear drum, the microphones at the entrances of both ear canals were removed. The responses to the 20 stimuli under all conditions were synchronized and averaged to increase the signal to noise ratio, and then convolved with the time-reversed OATSP to obtain the impulse response. This measurement was repeated 10 times for each condition after the headphones or ear microphones were removed once and then put on again. This resetting of measurement conditions was to decrease unintended variations in transfer functions that could have been caused by misalignment of the headphones or microphones [14]. The ten impulse responses measured were transformed in the frequency domain by FFT and averaged by taking the geometric mean.

In Fig. 2, $p_e(n)$ and $p_d(n)$ are sound pressure levels at the entrance of the ear canal and at the ear drum of the dummy head under headphone listening conditions. Similarly, $q_e(n)$ and $q_d(n)$ are sound pressure levels at the entrance of the ear canal and at the ear drum under free-field conditions. Assume that the Fourier transform of the signals at these four positions is represented in turn as, $P_e(j\omega)$, $P_d(j\omega)$, $Q_e(j\omega)$, and $Q_d(j\omega)$. Using these notations, the compensation function, $C(j\omega)$, can be represented as

$$C(j\omega) = \frac{Q_d(j\omega)}{Q_e(j\omega)} / \frac{P_d(j\omega)}{P_e(j\omega)}. \quad (1)$$

The compensation functions obtained using Eq. (1) for both ears are shown in Fig. 3. We employed

$$C'(j\omega) = \frac{Q_d(j\omega)}{Q_e(j\omega) \cdot P_d(j\omega)} \quad (2)$$

instead of Eq. (1) to cancel out the headphone characteristics.

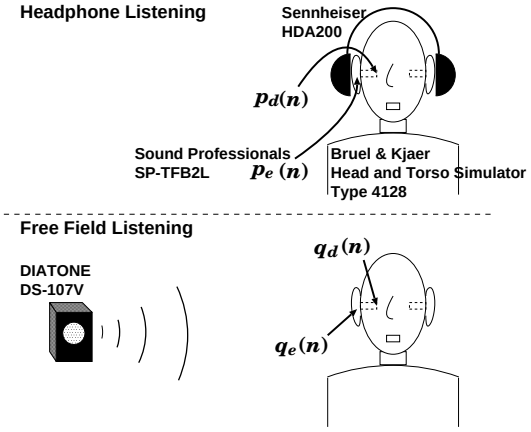


Figure 2: *Configuration of measurements for the compensation function for the influence on wearing headphones.*

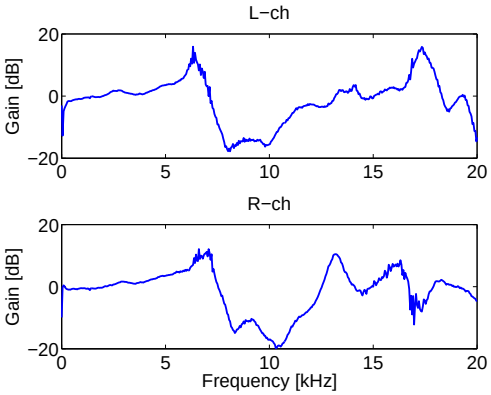


Figure 3: *Compensation functions for the transfer functions of the path from the entrance of the ear canal to the ear drum.*

4. EXPERIMENT

4.1. Experimental conditions

Taking plausibility theory into account, the hypotheses we wanted to test in the subjective experiments were twofold:

- Whether it became easier for listeners to accurately identify the location of the virtual sound source if reflected sound as well as direct sound were provided.
- Whether it also became easier for listeners to accurately identify the location of the virtual sound source if the reflected sound was created using individualized HRTF even when the direct sound was created using non-individualized HRTF.

As described in the previous section, front-back confusion often occurs in VAD systems. Some researchers have suggested that this problem could be resolved by properly adapting the interaural time difference of the virtual sound source in the horizontal plane as listeners move their head [15]. Extending this method to three-dimensional space, a virtual sound source should remain unchanged in its spatial position regardless of the motion of the listener's head. As localization of the perceived virtual sound source

undergoes noticeable change due to head tracking, we decided to examine cases both with and without this tracking.

To test the hypotheses, subjective experiments on sound localization with several combinations of HRTF data were carried out. The combinations of HRTFs to create the direct and reflected sound are summarized in Table 1.

Table 1: Combinations of HRTFs for subjective evaluation

Direct sound	Reflection	Case
Head Tracking ON		
Individualized HRTF	-	I-a
	Non-individualized	I-b
	Individualized	I-c
Non-individualized HRTF	-	II-a
	Non-individualized	II-b
	Individualized	II-c
Head Tracking OFF		
Individualized HRTF	-	III-a
	Non-individualized	III-b
	Individualized	III-c
Non-individualized HRTF	-	IV-a
	Non-individualized	IV-b
	Individualized	IV-c

Sound stimuli were generated referring to those used in the sound localization test by Wightman and Kistler [16] where a train of bursts of Gaussian noise was presented as a source signal. All bursts were 250 ms in duration and had squared cosine ramps with 20 ms at their onsets and offsets. A silent interval of 300 ms was inserted between two adjacent bursts. The Gaussian noise was bandpassed using a 10th-order FIR filter that passed a signal within a frequency band of 200 to 14000 Hz. In their original paper [16], Gaussian noise was further processed to add random tonal variations. Although these variations could prevent listeners from determining the direction based on the memory of tonal information presented previously for the same direction, tonal changes in sound are also important for humans to determine the direction sound is arriving from. Taking this into account, no additional processes to provide tonal variations to the sound source were included in the present experiment.

Considering the study on fusion and localization dominance described in the previous section, the direction of reflection was set to 30 degrees above the direct sound and delayed by 10 ms from the direct sound, as schematically outlined in Fig. 4. The

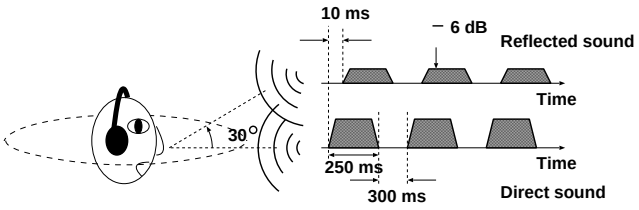


Figure 4: *Configuration of the direct and reflected sound*

virtual sound sources were presented from one of 12 directions in the azimuthal plane and 30 degrees apart. Six trials were tested for each direction. Consequently, the listeners assessed 180 trials under each condition, which were divided into three sessions to enable listeners to take two breaks between them. It took approximately ten minutes to administer each session. The sound pressure level was set to 54 dB(L_{Aeq}) for the direct sound and 6 dB lower for the reflected sound when the virtual sound source was located just in front of the listener.

One well-trained male person in his thirties whose HRTF data were in the HRTF database served as the listener in the experiments. He held one head tracker in his hand and was asked to point out the direction of the perceived sound image by stretching his arm holding the tracker in that direction. When he had determined the direction, he held that posture for a while and uttered a set phrase to signal the experimenter. Hearing the phrase, the experimenter pressed a key on a keyboard and obtained position information on the head trackers from the IS-900 processor. The angle information was calculated from the relative positions between the head trackers in the listener's hand and on his head. The system setup for the experiment is schematically outlined in Fig. 5.

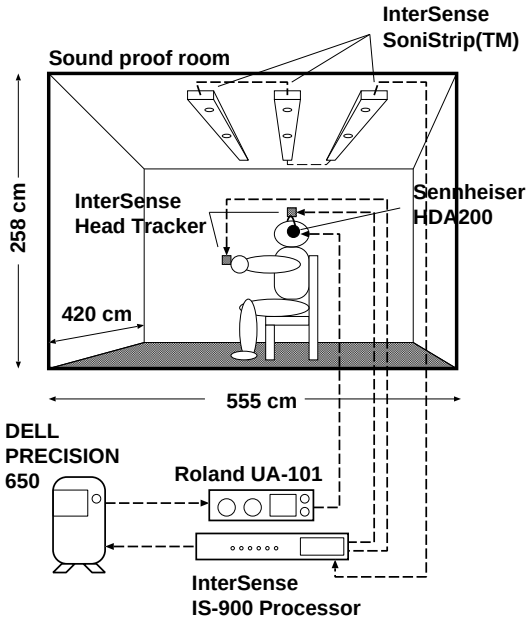
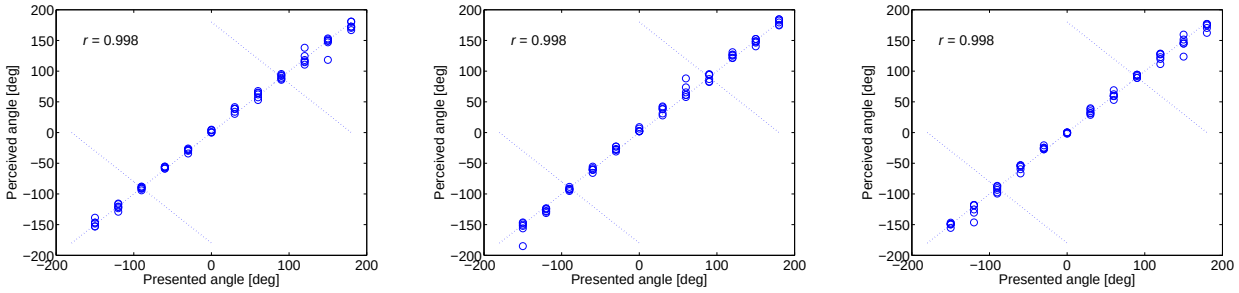


Figure 5: *Setup for the experiment. A listener wore headphones with a position sensor on them and held another position sensor in his hand to point the direction of the perceived sound image.*

4.2. Results

The results of the experiments are plotted in Figs 6 under conditions with and 7 without head tracking. The correlation coefficient between the presented and perceived sound directions is also given in all six figures. A bias toward one direction with constant amount, which might occur by pointing the perceived direction by moving listener's arm, does not affect correlation coefficients. When the perceived azimuth was less than -150 degrees for the presentation azimuth of 180 degrees, 360 degrees was added to resolve the problem of discontinuity at 180 degrees. Similarly, when

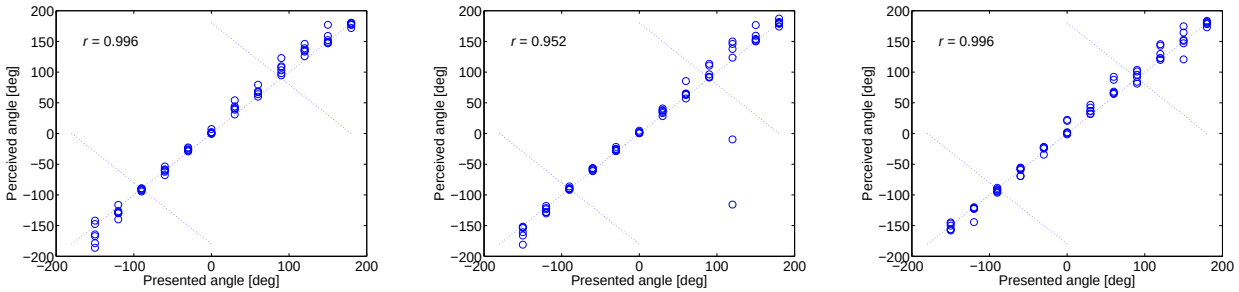


(a) Only direct sound was presented.

(b) Reflection created using non-individualized HRTF was added.

(c) Reflection created using individualized HRTF was added.

Case I: Individualized HRTF was used for creating the direct sound.



(a) Only direct sound was presented.

(b) Reflection created using non-individualized HRTF was added.

(c) Reflection created using individualized HRTF was added.

Case II: Non-individualized HRTF was used for creating the direct sound.

Figure 6: *Localization of virtual sound when head tracking was on.*

the perceived azimuth was greater than 150 degrees for the presentation azimuth of −150 degrees, 360 degrees was reduced from the obtained angle. These operations helped to prevent unacceptably small correlation coefficient.

The correlation coefficients between the presented and the received azimuthal angles were calculated to quantitatively evaluate localization and these are plotted in Fig. 8 with their confidence intervals.

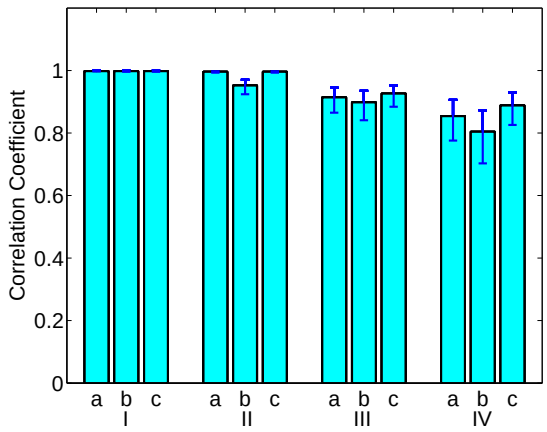


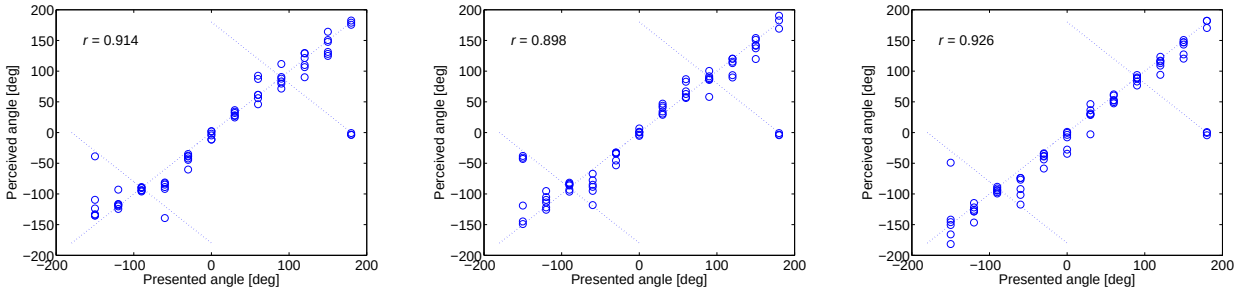
Figure 8: *Correlation coefficients for all conditions. Error bars indicate confidence intervals of 95%.*

From these figures, it is obvious that the conditions with head tracking outperformed those without as has been suggested by many preceding research reports. Moreover, adding a single reflection created using individualized HRTF tended to yield better performance in terms of correlation coefficients between presented and perceived sound directions than where the reflected sound was created using non-individualized HRTF. This tendency seemed clear in cases where a low correlation coefficient was obtained, i.e., under conditions where head tracking was off. While adding reflected sound created using individualized HRTF seemed better than adding reflected sound created using non-individualized HRTF, the performance was still comparable where no reflected sound was presented.

5. DISCUSSION

The listener reported that he sometimes heard the direct and reflected sound signals as a single sound source as a result of fusion but sometimes he heard two sound signals. Even when a single source image was perceived, it was more blurred compared with where only direct sound was presented. This uncertainty of localization might have caused degradation of localization in the present experiments, resulting in comparable performance between cases where only direct sound was presented and where reflected sound as well as direct sound were presented. It is certain that the comparable performance in Cases I and II was due to a ceiling effect.

From the results of Case IV in Fig. 7, where non-individualized

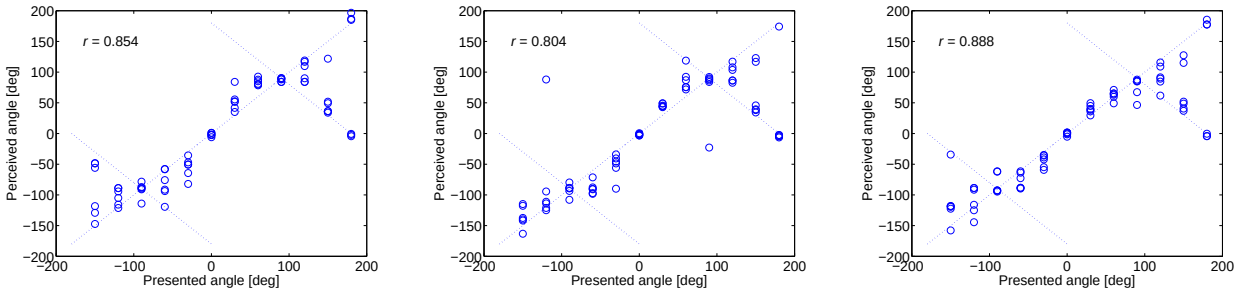


(a) Only direct sound was presented.

(b) Reflection created using non-individualized HRTF was added.

(c) Reflection created using individualized HRTF was added.

Case III: Individualized HRTF was used for creating the direct sound.



(a) Only direct sound was presented.

(b) Reflection created using non-individualized HRTF was added.

(c) Reflection created using individualized HRTF was added.

Case IV: Non-individualized HRTF was used for creating the direct sound.

Figure 7: *Localization of virtual sound when head tracking was off.*

HRTF was used for creating direct sound, and where head tracking was off, the listener perceived sound signals presented at 30 degrees in front of or behind the side direction as if they had almost arrived exactly from the lateral direction. This could be interpreted as sound images appearing at both the front and rear due to front-back confusion and fussing because of their similarities in sound characteristics, resulting in a signal sound image in the lateral direction.

6. CONCLUSIONS

Although the number of listeners used in the present study was quite limited, the results of subjective experiments did enable us to posit some conclusions. It seems that adding reflection created using individualized HRTF is more effective for localizing sound than that created using non-individualized HRTF. This was true for both cases where direct sound was created using individualized and non-individualized HRTFs. Moreover, no noticeable differences were observed under the conditions tested compared with when only direct sound was presented, Therefore, it would helpful to add reflected sound to VAD systems to localize it if individualized HRTF is available. We intend to conduct further experiments using larger numbers of participants from various age groups as well as under a variety of conditions to arrive at more concrete conclusions.

7. ACKNOWLEDGMENT

The authors would like to thank the undergraduate and postgraduate students at the Advanced Acoustic Information Systems Laboratory of Tohoku University for their help with measuring the HRTF data. We would also like to thank Dr. P. Mokhtari for providing us with the HRTF data he simulated using the Finite-Difference Time-Domain (FDTD) method as non-individualized HRTF.

8. REFERENCES

[1] E.M. Wenzel, M. Arruda, D.J. Kistler, and F.L. Wightman, "Localization using nonindividualized head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 94, no. 1, pp. 111–123, Jul. 1993.

[2] A.J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *J. Acoust. Soc. Am.*, vol. 93, no. 5, pp. 2764–2778, May 1993.

[3] S. Ise, "A principle of sound field control based on the Kirchhoff-Helmholtz integral equation and the theory of inverse systems," *Acta Acustica united with Acustica*, vol. 85, no. 1, pp. 78–87, Jun. 1999.

[4] T. Hidaka, Y. Yamada, and T. Nakagawa, "A new definition of boundary point between early reflections and late reverberation in room impulse responses," *J. Acoust. Soc. Am.*, vol. 122, no. 1, pp. 326–332, Jul. 2007.

[5] R.Y. Litovsky, H.S. Colburn, W.A. Yost, and S.J. Guzman, "The precedence effect," *J. Acoust. Soc. Am.*, vol. 106, no. 4, pp. 1633–1654, Oct. 1999.

[6] M. Ebata, T. Sone, and T. Nimura, "On the Perception of Direction of Echo," *J. Acoust. Soc. Am.*, vol. 44, no. 2, pp. 542–547, 1968.

[7] R.L. Freyman, "Dynamic processes in the precedence effect," *J. Acoust. Soc. Am.*, vol. 90, no. 2, Pt. 1, pp. 874–884, Aug. 1991.

[8] E.D. Schubert and J. Wernick, "Envelope versus Microstructure in the Fusion of Dichotic Signals,"*J. Acoust. Soc. Am.*, vol. 45, no. 6, pp. 1525–1531, 1969.

[9] B. Rakerd and W.M. Hartmann, "Localization of sound in rooms, II: The effects of a single reflecting surface," *J. Acoust. Soc. Am.*, vol. 78, no. 2, pp. 524–533, Aug. 1985.

[10] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, MIT Press, pp. 78–93, 1997.

[11] K. Watanabe, S. Takane, and Y. Suzuki, "A new interpolation method of HRTF based on the Common Pole-Zero model," *Proc. 17th ICA*, 2001.

[12] K. Iida, A. Murase, and M. Morimoto, "A method of 3-D sound image localization with externalization through headphones using a new correction filter of headphone-to-eardrum transfer functions," *Proc. 18th ICA*, vol. 2, pp. 993–996, Apr. 2004.

[13] 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," *J. Acoust. Soc. Am.*, vol. 97, no. 2, pp. 1119–1123, Feb. 1995.

[14] A. Kulkarni and H.S. Colburn, "Variability in the characterization of the headphone transfer-function," *J. Acoust. Soc. Am.*, vol. 107, no. 2, pp. 1071–1074, Feb. 2000.

[15] P.A. Hill, P.A. Nelson, O. Kirkeby, and H. Hamada, "Resolution of front-back confusion in virtual acoustic imaging systems," *J. Acoust. Soc. Am.*, vol. 108, no. 6, pp. 2901–2910, Dec. 2000.

[16] F.L. Wightman and D.J. Kistler, "Headphone simulation of free-field listening. II: Psychophysical validation," *J. Acoust. Soc. Am.*, vol. 85, no. 2, pp. 868–878, Feb. 1989.