Journal of Information Hiding and Multimedia Signal Processing ©2017 ISSN 2073-4212
Ubiquitous International Volume **8**, Number **6**, November **2017**

Influence of Visual Depth and Vibration on the High-level Perception of Reality in 3D Contents

Zhenglie Cui[1], Shuichi Sakamoto[1], Jiro Gyoba[2] and Yôiti Suzuki[1]

[1]Research Institute of Electrical Communication, Tohoku University
2-1-1 Katahira, Aoba-ku, Sendai, 980-8577, Japan
{sai@ais., saka@ais., yoh@}riec.tohoku.ac.jp

[2]Graduate School of Arts and Letters, Tohoku University
27-1 Kawauchi, Aoba-ku, Sendai, 980-8576, Japan
gyoba@sal.tohoku.ac.jp

Received January 2017; revised September 2017

Abstract. *This study investigated the influence of stereoscopic visual depth and body vibration on the high-level affective perception that concerns the senses of presence and verisimilitude. The multisensory content used in our experiment consisted of the 3D video recording of a passing train as seen from the side of the railroad, binaural recording of the 3D spatial sound, and the ground vibrations caused by the train. The results showed that the augmented stereoscopic depth information enhanced the affective perception of both the presence and verisimilitude. Moreover, both the presence and verisimilitude saturated with an increase in the intensity of body vibration under the augmented perceived depth condition. However, compared to the presence, the verisimilitude tended to saturate at a lower vibration-intensity. This tendency is consistent with that reported in our previous research.*
Keywords: Sense of presence, Sense of verisimilitude, Perceived reality, Body vibration, 3D vision, 3D sound

1. **Introduction.** With the advancement of information technology in recent years, it is now possible to synthesize and present multisensory content consisting of not only audio-visual information, but also information of other sensory modalities such as touch, smell, and vibration. Owing to such advancement, there is an increase in the demand for ultra-realistic systems that enable users to easily share "sense as if being in another space" and "sense as if it is a true experience." In order to achieve such realism, it is important to understand how humans process multisensory information to establish high-level affective perception.

Among the measurements of high-level affective perception, the sense of presence is the most common and popular index used to evaluate multisensory information such as that provided by virtual reality (VR) systems and environments[1, 2, 3, 4]. The sense of presence is usually defined as the subjective experience of being in one place or environment even when the observers are physically located in another place[5]. Therefore, this sense mainly concerns the place where the observers are present and it evaluates the realism of the space and environment of the scenes; in other words, the entire information of the scene including the background[6]. Meanwhile, we sought another type of affective perception that concerns focused information as "figure" of the scene. For such an index, we proposed "sense of verisimilitude," which is defined as the trueness of appearance and

quality of the presented objects. The sense of verisimilitude thus concerns the foreground and target objects in the scene and it mainly evaluates the realism of the principal components of the scene. We have successfully demonstrated that the sense of verisimilitude displays the different characteristics of the sense of presence[7, 8]. The reality evaluation of the scenes can be regarded as a combination of affective perception of space or environment, which corresponds to the sense of presence, and that of substantial objects and events, which correspond to the sense of verisimilitude[9]. Based on this idea, we are able to investigate the higher-level process for experiencing multisensory content by using both these metrics.

Following this approach, we investigated the influence of body vibration on high-level affective perception of multisensory content, because body vibration is known to be effective in enhancing the sense of presence[10, 11]. More specifically, we examined how the intensities of sound and body vibration affect the senses of presence and verisimilitude of multisensory content that consists of not only audio-visual information, but also body vibration. The results indicated that the sense of presence increases monotonically with the increase in intensities of sound pressure and vibration up to values which are much higher than the actual ones. This tendency is also observed for other physical parameters; e.g. the sense of presence monotonically increases as a function of the angle of view[12]. In contrast, the verisimilitude saturates at intensities around the actual values, and then shows a tendency to decrease[7, 10]. Thus, it is seen that the senses of presence and verisimilitude have different characteristics, suggesting that a balance between both the senses would be important for creating a natural and realistic virtual reality environment.

In our previous studies, we used 3D sounds using binaural recording with a dummy head; however, we did not consider the stereoscopic depth of the moving picture while changing the angle of view and the distance of the 2D display[7, 9]. Okui, however, showed that the depth and stereoscopic information contributed in improving the sense of presence[13]. Therefore, certain aspects concerning the role of stereoscopic information on multisensory affective perception need to be clarified. In particular, the influence of depth information on the senses of presence and verisimilitude need to be investigated. In this study, we investigated the influence of depth of visual stimulus and amplitude of body vibration on the subjective perception of presence and verisimilitude using multimodal content consisting of 3D video, 3D sound, and ground vibration of a train passing by an observer.

2. **Experiment.**

2.1. **Stimulus.**

The multisensory content was recorded beside a JR line (Tōhoku Honsen, connecting Tokyo and Aomori) near Iwakiri station in Sendai city, when a local train passed by the recording point. Figure 1 shows one scene of the recorded video . This 3D stereoscopic video was newly recorded for this study using a 3D full HD video camera (AG-3DA1, Panasonic Inc.). The sound and vibration stimuli used for this study were those which we measured at the same place in our previous study[14]. Figure 2 shows the recording setup used in the previous study. The 3D spatial sound was recorded binaurally with a dummy head (SAMRAI, Koken Co. Ltd.) with two microphones (4101, Brüel & Kjær) located inside its two ears[15]. This recording was obtained using a DV camera (AG-DVX100A, Panasonic Inc.); the outputs of the binaural microphones were connected to the audio inputs of the DV camera via an amplifier (2639, Brüel & Kjær). The sampling frequency of the binaural signals was 48 kHz. Two acceleration pickups (VM-80, RION Co. Ltd.) were fixed on both sides of a wooden board on the ground.

FIGURE 1. One scene of the multimodal video used in the experiment

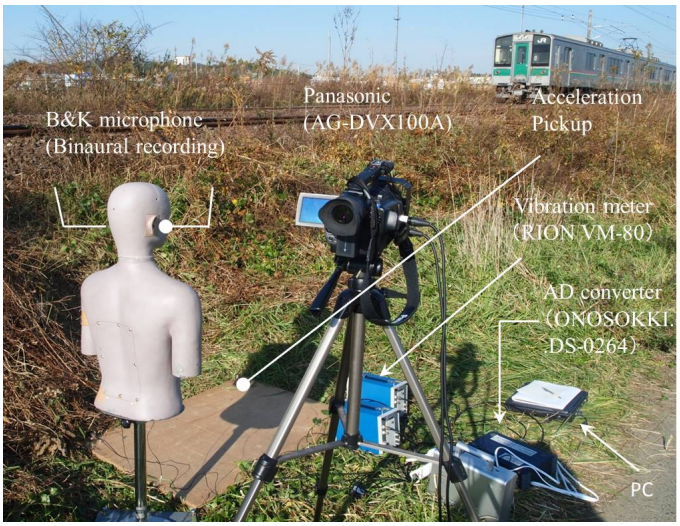

FIGURE 2. Recording setup in previous studies[14]

While the vibration information was being recorded, a weight of approximately 60 kg was placed on the board.

2.2. **Observers.** Sixteen young adults (12 male and 4 female) participated in the experiment. Their ages ranged from 21 to 26 with an average of 24 years (SD=2.6). All of them had normal vision including corrected vision and normal hearing.

We edited these videos, sounds, and ground vibrations to create the multisensory content to be used as stimulus. While we used a dummy head to record the audio signal in the previous study to obtain a binaural sound, this time, we used microphones equipped with 3D stereoscopic video camera to record the sound. In the present study, to present the rich spatial auditory information, we used the binaural sound signal recorded in the previous study. Therefore, we used the newly recorded sound signal only to synchronize the newly recorded 3D stereoscopic video and the binaural sound signals. Specifically, the images of the recorded video and the previously recorded sound and ground vibration were combined as the experimental stimuli. The video and sound/vibration were synchronized as follows. First, the cross correlation between the waveform of the acoustic data included in the new 3D video file and that of the previous study were calculated. The synchronization timing was then estimated as the point that showed the highest correlation coefficient (107.1 s, Fig. 3). With this delay time, we synchronized the newly recorded 3D stereoscopic video and the binaural sound signals.

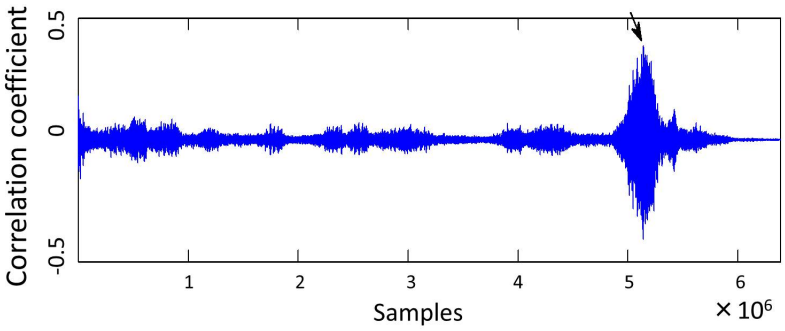

FIGURE 3. Cross-correlation between the sound signal binaurally recorded in the previous study and that recorded as simple stereo sound in the present study

As the sound of the new video was replaced by the former video, the possible difference of the train speeds may have caused an artifact. Thus, we confirmed the time lag of the re-recorded video with the original video accompanying the sound and vibration by a frame-by-frame observation. As a result of this observation, the time difference between the videos was about 500 ms. We measured the durations from the first appearance of the train until the moment when it passed the screen side. Given that our previous study indicated the time-window size of the integration between audio?visual and vibration information, which is estimated to be around 600 ms[14], we believe that the difference is allowable. The total duration of the stimulus was set to 15 s.

2.3. **Experimental setup.** Figure 4 shows the experimental setup. All experiments were conducted in a sound-proof room. The observers wore 3D glasses (3D VISION 2, NVIDIA) and were asked to stand directly on the motion platform and observe the video. The stereoscopic visual stimulus was presented using a DLP projector (Mirage WU7K-M, CHRISTIE Co. Ltd.) on a 150-inch screen (Stewart Sound Screen) that was set in front of an observer. The distance between the observer and the screen was 3.5 m. The field of view was 39? (horizontal) × 25? (vertical). The spatial audio stimulus was presented binaurally via headphones (HDA-200, Sennheiser Electronic). The A-weighted sound pressure level was set to 60 dB (L_{Aeq}). The body vibration stimulus was provided via a motion platform (D-BOX Mastering Motion). Only 1-DOF vibration (perpendicular direction) was presented during the experiment.

As shown in Fig. 4, we used two PCs in the experiment; a pc is used to present the audio visual information (PC for AV) and the other pc is used to control the motion platform system to present full-body vibration (PC for vib). To synchronize these two computers, we introduced the trigger signal from the audio visual system interface (UA-25EX, Roland Corp.) used with the PC for AV. The trigger signal is sent to the audio interface of the PC for vibration. As a result, the time difference between the inputs of the motion platform and that of the audio visual signal was about 51 samples (= 1.1 ms), which we regard as a synchronization that is good enough between sound and vibration based on our previous investigation[16].

2.4. **Experimental procedures.** We set three conditions for the visual stimulus: 2D, depth-enhancing, and pop-up conditions as shown in Fig. 5. We used stereoscopic player software (3dtv.at[17]) to modify the parallax of the 3D video. Under the depth-enhancing condition, the parallax of the 3D video was shifted in the outward direction and, therefore, the train appeared at an infinite distance from the viewing point, and it was adjusted to

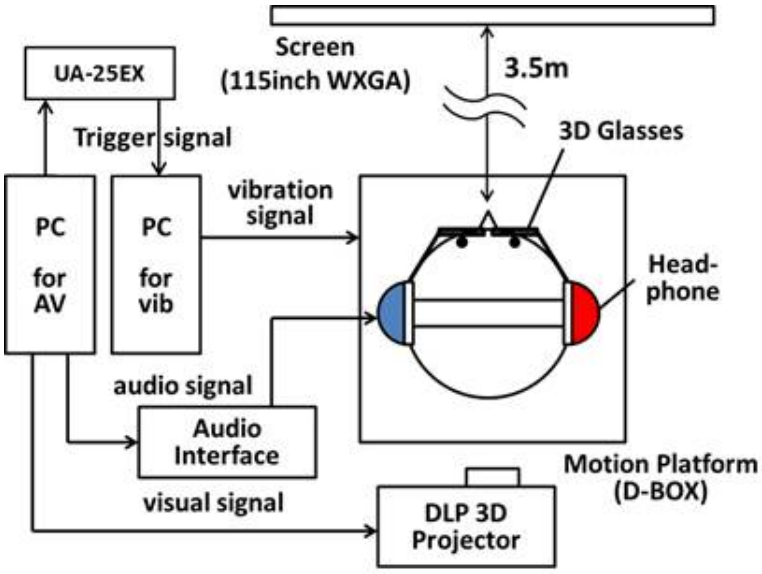

FIGURE 4. Experiment setup

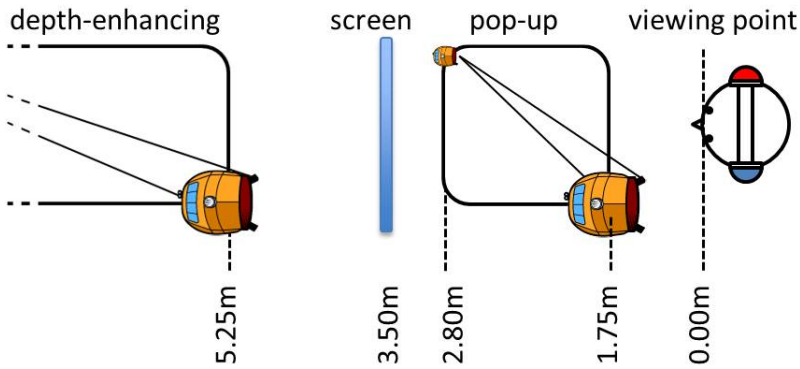

FIGURE 5. Image diagram for the two 3D conditions

appear to pass at a distance of 5.25 m in the screen. Under the pop-up condition, the parallax of the 3D video was shifted in the inward direction, and as a result the train appeared at a distance of approximately 2.8 m from the viewing point and passed by the screen side at a distance about 1.75 m. Besides these two 3D conditions, 2D condition was used as the control condition. The 2D condition used only the left eye images of the stereoscopic video.

For the vibration condition, the recorded amplitude was defined as the standard (0 dB), and the following five vibration amplitude levels were prepared: −12, −6, 0, +6, +12 dB.

The observers were asked to judge the senses of presence and verisimilitude and rate them on a scale of 7 (from 0 (low) to 6 (high)). To avoid confusion or misunderstanding between presence and verisimilitude, the observers were shown the definitions of these terms, which were as follows:

sense of presence: sensation as you feel like being there, and

sense of verisimilitude: sensation as you feel what you see looks real

These definitions were similar to those given to the observers in our earlier studies. In the experiment, different sessions were set for each of the three visual conditions, and the five vibration conditions were repeatedly presented, four times in each session. Therefore,

one session consisted of 20 trials (4 × 5). The observers participated in three sessions for each of the visual stimulus condition, resulting in twelve trials per condition. We have randomized the order of trials in a session for each observer and the order of the three sessions.

3. **Results.** For each observer, the rating data were normalized based on the average of the observers' ratings, and standard deviation and were converted to z-scores to enable comparison considering the subjective reality of each observer. In the present experiment, the range of judgment was largely different for each observer and thus the positions of the peaks of distribution range widely. Therefore, we were afraid that if the data would be pooled from the observer's data, it could not satisfy the assumption of the normal distribution. Furthermore, we consider it useful to retain the availability of the analysis of the individualities of the data. Therefore, we normalized the data by calculating the z-score individually. On the other hand, given that all data are normalized using the z-score, the average values and the variances become 0 and 1 for all observers. Therefore, the two-way ANOVA was carried out separately for the senses of presence and verisimilitude with three visual conditions and five vibration conditions as factors.

Figures 6 to 8 show the average value for each visual condition as a function of the vibration amplitude level. The error bars denote the standard errors. The open squares and circles represent z-scores for the senses of presence and verisimilitude, respectively. The results of the 2D (Fig. 6) and pop-up (Fig. 8) conditions show that the sense of presence increased monotonically, whereas the sense of verisimilitude saturates around the amplitude level of +6 dB. In the depth-enhancing condition (Fig. 7), both presence and verisimilitude increase up to 6 dB, and then tend to saturate. When the vibration is small (from −12 to −6 dB), the sense of verisimilitude was higher than that of the presence under the 3D conditions (depth-enhancing and pop-up conditions), while the two senses show a similar tendency under the 2D condition.

To compare the results among the three visual conditions, the z-scores for the three conditions are drawn in Figs. 9 and 10 with the vibration amplitude level as a parameter. The comparison of the results depicted in these figures show that the z-scores for verisimilitude under the two 3D conditions are generally larger than those for the 2D condition and that the dependence on the visual stimuli condition is generally smaller for the evaluation of presence than that of verisimilitude. Moreover, when the evaluations of the two 3D conditions were compared, the evaluation under the 3D pop-up condition was found to be smaller than that under the depth-enhancing condition for a weak vibration (−12 dB).

Thus, we have separately carried out a two-way ANOVA for the sense of presence and the sense of verisimilitude. The three visual conditions and five vibration conditions were treated as factors and the observer as the repeated measure. Results for the sense of presence revealed significant main effects of the vibration conditions ($F(4, 60) = 31.433$, $p < .001$). The results of the multiple comparisons (Ryan's method, $p < .05$) on the main effects of vibration revealed that the evaluation scores under the vibration amplitude level of −12 dB is statistically lower than that of all other levels, and that of −6 dB show lower values than those for 0 dB, 6 dB, and 12 dB. Furthermore, the evaluation scores under 0 dB exhibited lower values than those for 6 dB and 12 dB. However, no significance was observed between the vibration amplitude levels of 6 dB and 12 dB.

Meanwhile, for the sense of verisimilitude, the result revealed significant main effects of the visual conditions ($F(2, 30) = 5.910$, $p < .01$) and vibration conditions ($F(4, 60) = 6.659$, $p < .001$). The results of the multiple comparisons (Ryan's method, $p < .05$) on the main effect of visual condition revealed that evaluation scores under the 3D conditions

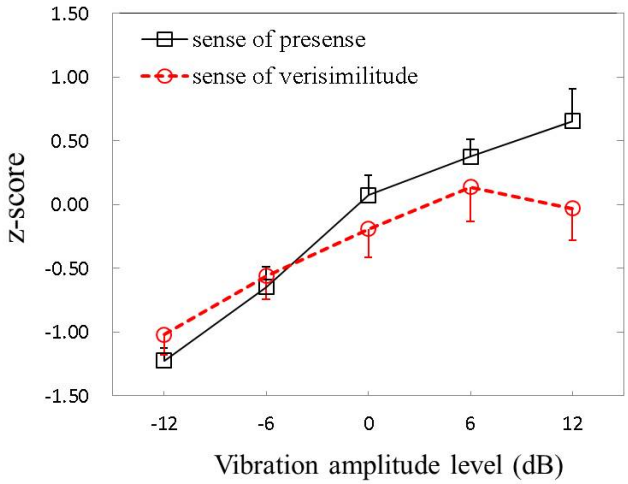

FIGURE 6. Results of the experiment for 2D condition

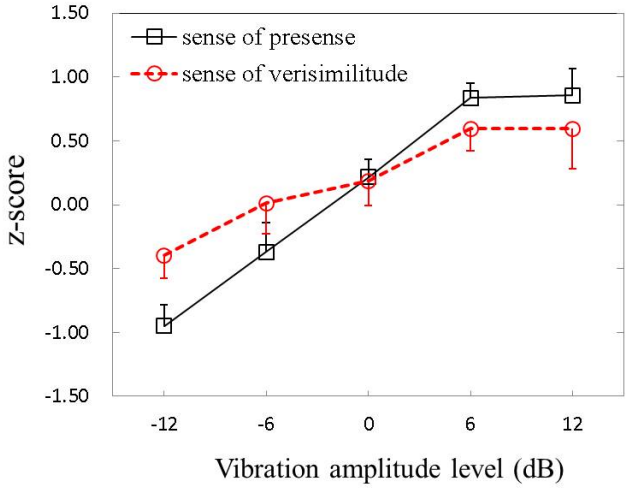

FIGURE 7. Results of the experiment for 3D depth-enhancing condition

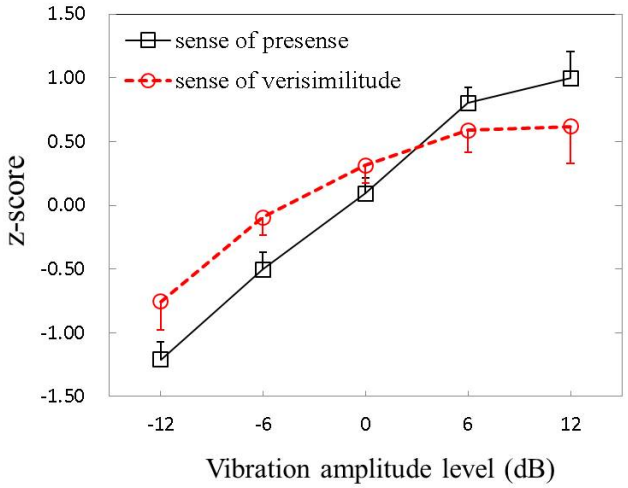

FIGURE 8. Results of the experiment for 3D pop-up condition

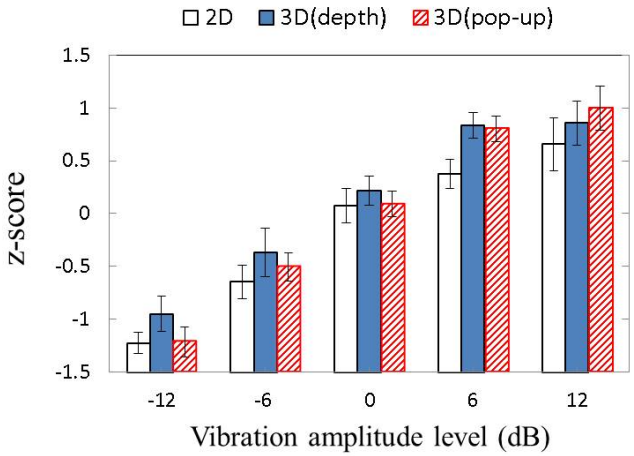

FIGURE 9. Comparison of the sense of presence for each visual condition

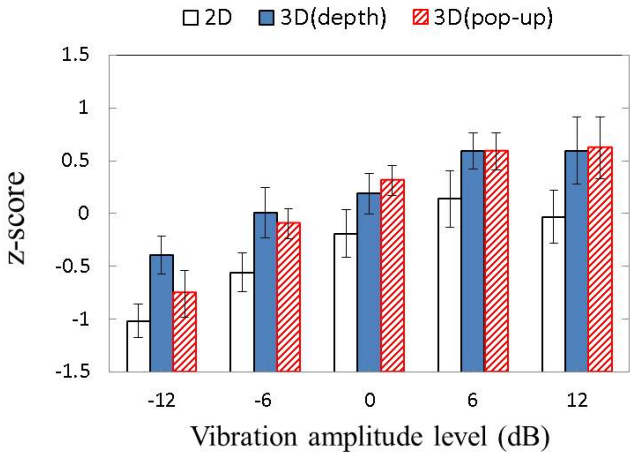

FIGURE 10. Comparison of the sense of verisimilitude for each visual condition

are significantly higher than those of the 2D condition; however, the difference between the two 3D conditions is not significant. The results of the multiple comparisons on the main effects of vibration revealed that the scores for the vibration amplitude level of -12 dB is significantly lower than those for 0 dB, 6 dB and 12 dB. However, no significant difference was observed between the vibration amplitude levels of -12 dB and -6 dB.

4. **Discussion.** First, we discuss the results for the sense of presence. As shown in Figs. 6 - 8, the evaluation (z-score) for presence increased monotonically as a function of the vibration amplitude under the 2D and the pop-up conditions. This tendency is consistent with the results reported in previous studies, which state that the sense of presence increases as the intensity of physical stimuli increase beyond the actual values[18]. However, the analysis of variance does not show significant difference between the vibration amplitude level of 6 dB and 12 dB. From these facts, it is considered that the sense of presence saturates when the vibration is around 10 dB larger than the actual value. An interesting difference is observed between the trends of the results of the two 3D conditions for the vibration amplitude levels from 6 dB to 12 dB: the sense of presence is almost same under the depth-enhancing condition, while it becomes higher for 12 dB under the pop-up condition. This difference might be attributable to the apparent distance of the

train; under the depth-enhancing condition, the distance is perceived as far as infinity, which naturally causes small vibrations, while under the pop-up condition it is perceived to be near, which may cause large vibrations, resulting in the difference in the trends from 6 to 12 dB. Since this difference is not statistically significant in this study, this may be an interesting issue to be confirmed in future studies.

For the sense of verisimilitude, the evaluation for the vibration amplitude level of −12 dB produced statistically lower values than that of all the other levels, whereas no significant difference was observed for the other combinations. Figures 6 to 8 show that verisimilitude shows a peak around the vibration amplitude level of 6 dB, which is several dB larger than the actual vibration intensity. This can be compared to our previous results[14], which showed that verisimilitude exhibits saturation when the real and stimulus intensities have similar values.

Furthermore, the main effect of the visual conditions is observed only for the senses of verisimilitude. Similarly, the tendencies of the statistical analyses show certain difference between the two senses. Thus, we examined the significance of the interaction relating to the two senses by using three-way ANOVA with the two senses, three visual conditions, and vibration conditions as factors and the observer as the repeated measure. The results exhibited significant interaction between the two senses and the vibration conditions ($F(4, 60) = 7.129$, $p < .001$), which strongly suggests that the different tendencies found between the sense of presence and the sense of verisimilitude are meaningful.

The statistical test results related to the visual conditions show that the significant effect of the depth of visual information for the sense of verisimilitude, i.e., the evaluation scores under the two 3D conditions were significantly higher than that under the 2D condition. This can be considered as the consequence of additional depth information in the visual presentation, suggesting that depth information is effective in enhancing high-level affective perception of multisensory content. The results shown in Figs. 10 suggest that the depth-enhancing condition would derive the highest affective perception, particularly for verisimilitude, when the vibration intensity is small. Although this tendency is not clearly confirmed by the statistical analyses, it shows a possibility that realism of multisensory content is effectively enhanced by appropriate spatial information without the use of strong vibration stimuli. This possibility, i.e. whether the presentation of rich spatial information of a specific sensory modality would reduce the necessity of information of other modalities to obtain the same level of realism, seems to be an interesting topic for future studies.

5. **Conclusions.** We investigated how stereoscopic visual depth and body vibration influence the high-level affective perception concerning senses of presence and verisimilitude. The results show that augmented stereoscopic depth information enhances the affective perception for sense of verisimilitude. Moreover, it was observed that the sense of verisimilitude tends to saturate at lower vibration intensity than the sense of presence does. This tendency is consistent with the observations reported in our previous research.

Acknowledgment. A part of this work was supported by the National Institute of Information and Communications Technology (NICT) of Japan and JSPS KAKENHI Grant Numbers JP26280067, JP26330306, and JP16H01736. We appreciate the support from Ms. Emi TAKAHASHI for her technical assistance in the experiment.

REFERENCES

[1] Slater, M., et al., The influence of body movement on presence in virtual environments, *Human Factors*, vol.40, pp.469-467, 1998.

[2] Sheridan, T. B., Musings on telepresence and virtual presence, *Presence: Teleoperators and Virtual Environments*, vol.1, pp.120-126, 1992.
[3] Lessiter, J., et al., A cross-media presence questionnaire: The ITC-Sense of Presence Inventory, *Presence: Teleoperators and Virtual Environments*, vol.10, pp.282-298, 2001.
[4] Lombard, M., et al., At the heart of it all: The concept of presence, *Journal of Computer Mediated Communication*, vol.3-2, pp.1-43, 1997.
[5] Witmer, B. G., et al., Measuring presence in virtual environments: A presence questionnaire, *Presence: Teleoperators and Virtual Environments*, vol.7, no.3, pp.225-240, 1998.
[6] Teramoto W, et al., What is "sense of presence?" A non-researcher's understanding of sense of presence, *Journal of the Virtual Reality Society of Japan*, vol.15, no.1, pp.7-16, 2010. (in Japanese)
[7] W. Teramoto, et al., Spatio-temporal characteristics responsible for high vraisemblance, *Journal of the Virtual Reality Society of Japan*, vol.15, no.3, pp.483-486, 2010. (in Japanese)
[8] Akio Honda, et al., Senses of presence and verisimilitude of audio-visual contents: Effects of sounds and playback speeds on sports video, *Interdisciplinary Information Sciences Journal*, vol.21, no.2, pp.143-149, 2015.
[9] Honda A, et al., Determinants of senses of presence and verisimilitude in audio-visual contents, *Journal of the Virtual Reality Society of Japan*, vol.18, no.1, pp.93-101, 2013. (in Japanese)
[10] Sakamoto S, et al., Body vibration effects on perceived reality with multi-modal contents, *ITE Transactions on Media Technology and Applications*, vol.2, no.1, pp.46-50, 2014.
[11] Woszczyk W, et al., Getting immersed in multisensory, interactive music via broadband networks, *J. Audio Eng. Soc.*, vol.53, pp.336-344, 2005.
[12] J.D. Prothero, et al., Widening the Field-of-view Increases the sense of presence within immersive virtual environments, *Human Interface Technology Laboratory Tech. Rep.*, R-95-4, Seattle: University of Washington, 1995.
[13] OKUI Makoto, Integral three-dimensional television, *The Journal of Institute of Electronics, Information and Communication Engineers*, vol.93, no.5, pp.377-381, 2010. (in Japanese)
[14] Shuichi SAKAMOTO, et al., Effects of vibration information on the senses of presence and verisimilitude of audio-visual scenes, *INTER-NOISE*, 2016.
[15] B. R. Schröder, et al., Computer simulation of sound transmission in rooms, *IEEE Inter. Conv. Rec.*, vol.7, pp.150-155, 1963.
[16] T. Okamoto, et al., Implementation of a high-definition 3D audio-visual display based on higher-order Ambisonics using a 157-loudspeaker array combined with a 3D projection display, *Proc. IEEE IC-NIDC*, pp.179-183, 2010.
[17] http://www.3dtv.at/
[18] Mayuko Mitsuki, et al., Reproduced sound pressure level yielding the maximum auditory presence: Further study on effects of reproduced SPLs on auditory presence, *Acoustical Science and Technology*, vol.26, no.1, pp.79-81, 2005.