Journal of Information Hiding and Multimedia Signal Processing ©2022 ISSN 2073-4212
Ubiquitous International Volume **13**, Number **4**, December **2022**

Determination of optimal levels of whole-body vibration using audio-visual information of multimodal content

Shota ABE[1], Shuichi SAKAMOTO[1], Zhengile CUI,[2] Yôiti SUZUKI[3]

[1]Research Institute of Electrical Communication and Graduate School of Information Sciences, Tohoku University
2–1–1 Katahira, Aoba-ku, Sendai, 980–8577 Japan
{shota.abe.r4@dc, saka@ais.riec}.tohoku.ac.jp

[2]Aichi University of Technology, 50–2 Manori, Nishihasama-cho, Gamagori, Aichi, 443–0047 Japan
yamataka-masahiro@aut.ac.jp

[3]Tohoku Bunka Gakuen University, 6–45–1 Kunimi, Aoba-ku, Sendai, 981–8551 Japan
s-yoiti@ait.tbgu.ac.jp

Received September 2022; revised November 2022

Abstract. *The effectiveness of adding whole-body vibration information to audio-visual content has been demonstrated. However, current content generally consists only of audio-visual information and does not include whole-body vibrations. Therefore, it is necessary to establish a method for generating effective vibration information. In our previous study, we showed that some characteristics related to the amplitude of the vibration have a greater impact on perceived reality of multimodal content than those related to the frequency of the vibration. Therefore, in this study, we focused on the amplitude of whole-body vibration and investigated the relationship between the optimal amplitude of whole-body vibration and audio-visual information provided by the content. The results suggest that a multiple regression model using acoustic features can explain the optimal vibration amplitude.*

Keywords: Multimodal information, Whole-body vibration,Synthesis of vibration, Acoustic features, Visual features

1. **Introduction.** Recently, people have gained access to various audio-visual content in daily life, and have obtained several experiences accordingly. To enrich such experiences, various studies have been conducted on the relationship between audio-visual information and perceived reality[1–3]. Furthermore, recent advances in information technology have made it possible to experience content with various types of sensory information, such as touch, smell, taste, and vibration. As multisensory information may generate more realistic experiences than audio-visual information alone, the relationship between multisensory information and perceived reality should be examined. Among the numerous types of sensory information, whole-body vibration, which vibrates the entire body, is closely related to perceived reality[4]. For example, the addition of whole-body vibration information increases the sense of presence and verisimilitude[5][6]. The addition of whole-body vibration information also improves the quality of the concert experience[7]. Furthermore, a recent study showed that the presence of whole-body vibration information reduces motion sickness in VR experiences[8]. Motion sickness is known to negatively affect the sense of presence[9]. Therefore, these studies indicate that whole-body vibration is essential for enhancing the perceived reality of multimodal content.

However, current content generally consists only of audio-visual information and does not include whole-body vibrations. Therefore, to use whole-body vibration information more easily, it is necessary to somehow generate vibration information from audio-visual information. For example, as a method for generating vibration from audio information, a low-pass filtered sound was regarded as a time wave of the vibration and provided to the users[7][10]. A method to generate vibration from visual information has also been proposed[11]. In this method, camera motion in first-person content is regarded as the motion of a user. According to this idea, the velocity of the camera's shaking from a first-person perspective is converted into the vibration amplitude. These studies have shown that the quality of the experience induced by the content can be improved by adding the generated vibrations.

Although methods for generating effective vibrations from sensory information have been proposed, the proposed methods need "additional information" to generate effective vibration. When the method is applied, optimal amplitude of the vibration cannot be decided without using the users' preferences. To determine the amplitude of the vibration by using the method, the vibration which is actually recorded in the environment of the content is required. However, the recorded vibration obviously cannot be obtained in every content. The method[11] can be only applied to content having camera motion with a first-person perspective. In addition, as these papers generate vibration from either audio or visual information, they cannot generate vibration for content that does not contain that sensory information. To handle vibration information more easily, it is important to construct a generation method that can be applied to any audio-visual content, which includes "audio-only" or "visual-only" content, without requiring any other information.

In this study, we propose a novel method to generate vibration in any of audio-visual contents. In the method, only audio-visual information in the contents is used without requiring additional restriction. To do this, we examined how acoustic and visual features can explain the optimal vibration amplitude. In the experiment, observers watched/listened to a wide variety of audio-visual content. During this experiment, they adjusted the vibration amplitude to what they felt was appropriate. To clarify the relationship between presented sensory information and vibration amplitude, three experimental conditions were prepared regarding audio-visual information presented with vibration: an audio-only presentation; a visual-only presentation; and an audio-visual presentation. The audio and visual-only presentations were included to consider not only contribution of acoustic and visual features but also possible mutual interaction between acoustic and visual features in determining the optimal vibration level.

2. Methods.

2.1. Contents.

To prepare a variety of content, we selected audio-visual contents (videos) from the video-sharing site "Vimeo[12]". We selected the contents used as the stimuli of the experiment according to the following processes. First, from the 91 categories and subcategories provided by Vimeo, similar categories were summarized and reclassified into six categories: "Animation," "Life," "Sports," "Performance," "Machine operation," and "Talk." Next, two to six kinds of contents in each classified category were selected in order of views for each classified category according to the following criteria:

- Indicates a Creative Commons license[13] (Attribution, ShareAlike and Non-commercial)
- There are only ambient sounds in the location where the video was made.
- Videos are less than a minute long.

As a result, 20 contents were selected in total. The selected contents are presented in Table 1.

2.2. **Stimuli.** The 20 selected contents were processed and controlled as follows. The visual signal was re-encoded to a resolution of 1920 × 1080 pixels and a frame rate of 30 fps. The sampling frequency and quantization bit rate of the audio signal were set to 48 kHz and 16 bits, respectively. As the actual sound volume in the environment in which the contents were recorded was unknown, all contents were controlled to have the same sound pressure level. The total time length of the 20 contents were varied, and some of them had non-stationary sound information. Therefore, the sound pressure level was controlled using a statistical measure, percentile A-weighted sound pressure level ($L_{A20} = 75[dB]$). For vibration information, we applied a method called vibration from low-frequency audio (ViLA), which was proposed in our previous study [10], to virtually generate vibration from the sound signals of the content. In this method, a low-pass filter with a cut-off frequency of 70 Hz was applied to the audio signal of each content after monophonic conversion[10]. The International Organization for Standardization (ISO) standard for human vibration perception, ISO 2631-1:1997 "Mechanical vibration and shock -Evaluation of human exposure to whole-body vibration. Part 1: General requirements," was considered as the reason for selecting 70 Hz as the low-pass filter[14]. The signal was regarded as the vibration amplitude waveform, and the percentile vibration levels of each content were adjusted to the same level ($L_{A20} = 60[dB]$). The sampling frequency and quantization bit rate of the vibration signal were set at 8 kHz and 16 bits, respectively.

2.3. **Observers.** 27 healthy adults with normal or corrected-normal vision and normal hearing (22 males, 7 females, mean age = 22.4 years, SD = 2.2) participated in this experiment.

2.4. **Experimental Setup.** Figure 1 shows the experimental setup. All experiments were conducted in an audio-proof room. The observers were asked to stand on a motion platform and watch the content. The visual stimulus was presented using a digital light processing (DLP) projector (PDG-DHT100JL, SANYO Co. Ltd.) on a screen (Stewart Audio Screen) set 2.5 m in front of the observer. The field of horizontal and vertical views were 90° and 50°, respectively. The audio stimulus was presented via headphones (HDA-200, Sennheiser Electronic), and the whole-body vibration stimulus was provided via the motion platform (D-BOX Mastering Motion). Only one degree of freedom (1DOF) for vibration (in the vertical direction) was utilized during the experiment. A throttle-lever controller (Throttle Quadrant, SAITEK) was attached to the dominant hand side (right hand: 25, left hand: 2) of the observers. The controller moves smoothly only in the vertical direction and can be kept stationary. The data acquisition sampling rate was 10 Hz. The movable range was 0 - 90°, and the angle resolution was 0.2°.

Table 1. Obtained contents

Categories	Title expressing content
Animation (4 kinds)	Robot actions, Gunfights, Object rotations, River current
Life (6 Kinds)	Squirrel movement, Balloon launch, Wave, Lightning, Church activities, Everyday scenery
Performance (3 Kinds)	Hand games with sticks, Drumming, Dancing
Sports (3 Kinds)	Skateboarding, Paramotoring, Futsal
Machine operation (2 kinds)	Printer operation, Pythagoras device
Talk (2 kinds)	Speech (male), Speech (female)

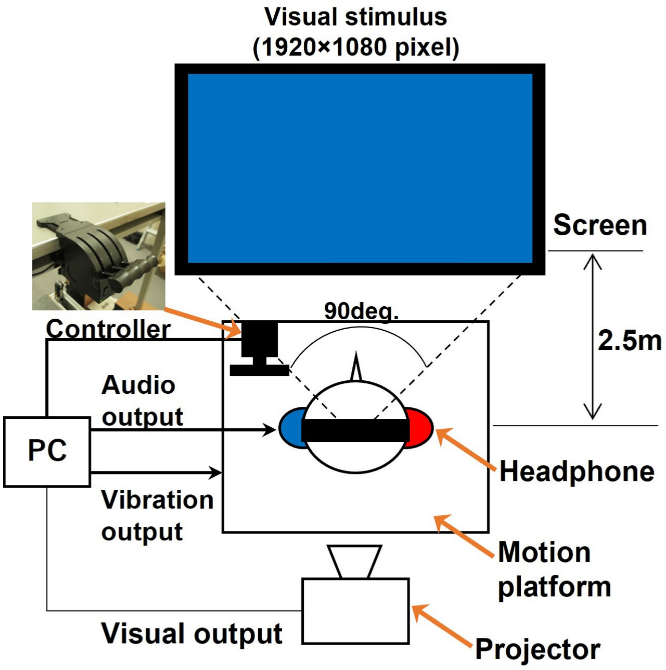

Figure 1. Experimental Setup

2.5. **Experimental Procedure.** Three conditions (Audio & Visual, Audio-only, Visual-only) were prepared in this experiment. In the Audio-only or Visual-only conditions, experimental stimuli were presented with only audio information or visual information. In the Audio & Visual condition, experimental stimuli were presented with both audio and visual information. Each observer experienced all content once in only one of the three conditions. The experimental conditions assigned to the content were counterbalanced so that all contents were experienced the same number of times in all conditions. Consequently, experimental data from each content was obtained for nine individuals under each experimental condition.

The experimental flow was as follows: first, a crossed gazing point was displayed on the screen in front of the observer. After the gazing point disappeared, one of the contents was presented. During the presentation, the observers were asked to adjust the vibration amplitude in real-time to "that which they felt appropriate for the content (scene)" by operating a lever in their hand. Temporal changes in the adjusted vibration amplitude were recorded. At the end of the content playback, the observers could choose "proceed to the next content" or "adjust again". The observers could repeat the adjustment until satisfied. If the user selected "adjust again," the vibration adjusted in the previous trial was presented as the reference vibration. The experiment resulted in an average of 3.9 ± 1.46 adjustments per one content.

2.6. **Relationship between the adjusted angle and the presented vibration amplitude.** The increase in the vibration level $R_n(t)$[dB], which varies with the lever adjustment, can be described as follows:

$$R_n(t) = R_{n-1}(t) + d \tag{1}$$

$$R_0(t) = 0 \tag{2}$$

where t is the time [s], n is the number of trials, and R_n is $0 \leq R_n(t) \leq 40$ due to constraints in the hardware used to reproduce the vibration. d is the amount of change [dB] adjusted

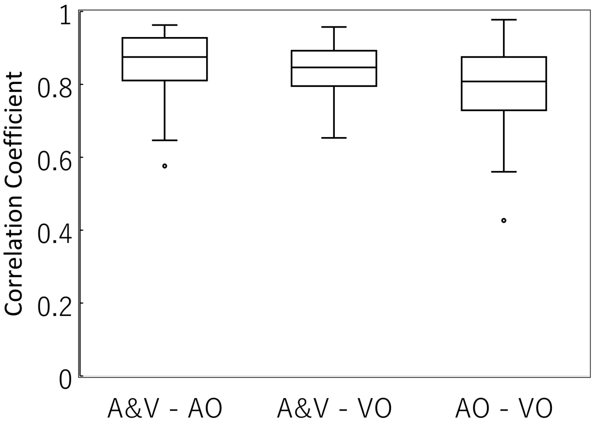

FIGURE 2. Correlation coefficients between each experimental condition for the temporal change of the calculated optimal vibration level of each content

TABLE 2. Correlation coefficients of the optimal overall vibration levels for each of the three combinations

Dependent variable - explanatory variable	A&V - AO	A&V - VO	AO - VO
Correlation coefficients	0.88	0.81	0.77

by the lever. At the beginning, the lever was set at the position of 45°, which corresponded to $d = 0$. When the lever is adjusted upward, d increases, and vice versa. The degree of the lever and vibration amplitude range can be varied to allow for precise vibration adjustments after each adjustment. For the first and second adjustments, ±1 dB/degree was set; for the third and fourth adjustments, ±0.5 dB/degree was set; and for the fifth and subsequent adjustments, ±0.25 dB/degree was set. Note that when $R_n(t)$ fell below zero due to lever adjustments, an $R_n(t) = 0$ value was set. The experimental results showed that optimal vibration level was adjusted to $R_n(t) = 40$ with an upper limit of 2.6% in all trials.

3. Results.

3.1. Overall tendencies of the obtained optimal vibration level in the three conditions.

The overall tendencies in vibration levels obtained by experiment (hereafter referred to as "optimal vibration levels") for Audio & Visual (A&V), Audio-Only (AO), and Visual-Only (VO) were analyzed using two approaches: temporal level change and overall level.

First, we analyzed the effect of the experimental conditions on the temporal level change of the optimal vibration level for each content using three combinations: 1. AV - AO; 2. AV - VO; and 3. AO - VO. The temporal level change of the optimal vibration level for each of the nine observers was firstly calculated every 0.1 s. Then, the data for every 0.1 s obtained from the nine observers was averaged. For all three combinations of the experimental conditions, correlation coefficients were obtained for the temporal level changes of the optimal vibration level. This process was performed for the vibration level of each content, separately. The obtained correlation coefficients are summarized in Fig. 2 with a box plot. These results indicate that the temporal variation of the optimal vibration level among all experimental conditions is similar.

Next, we analyzed the effect of the experimental conditions on the tendency of the overall optimal vibration level for each content. The overall optimal vibration level for

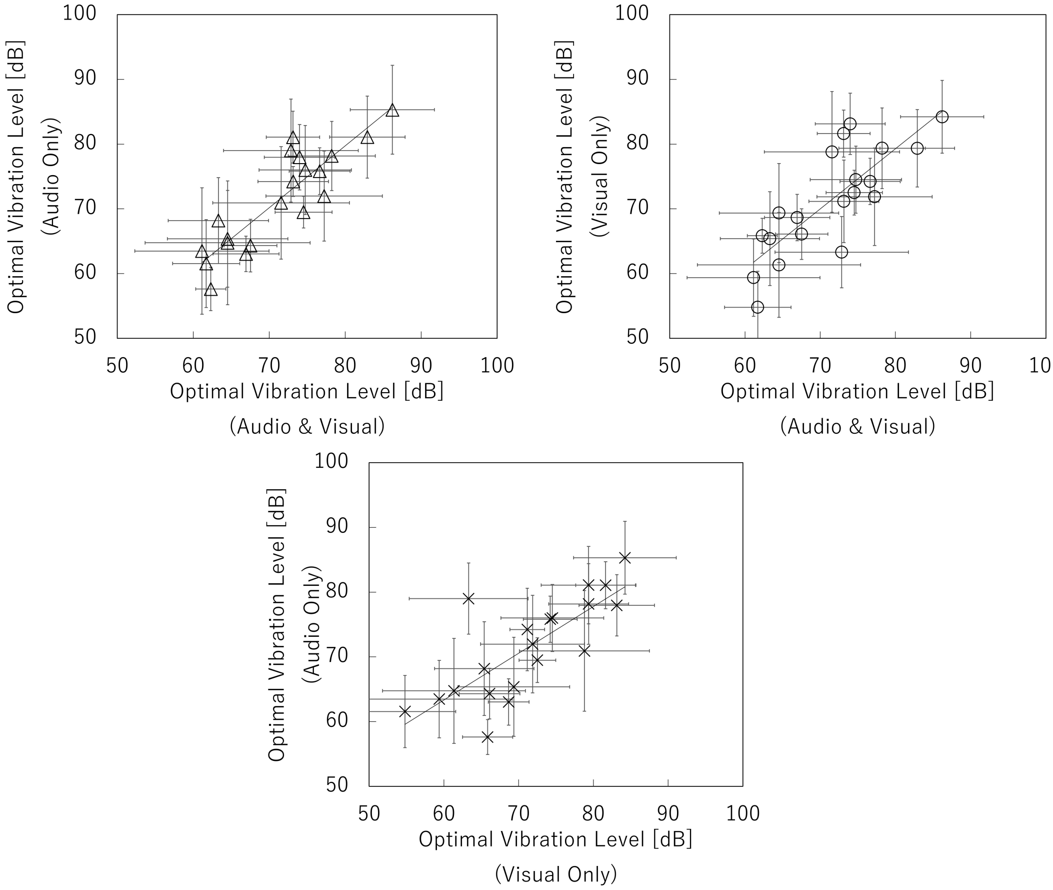

FIGURE 3. Comparison of the optimal overall vibration levels for each of the three combinations

each content was obtained by averaging the temporal level change of the optimal vibration level of each content across the time length. Scatter plots were then drawn for all combinations of the experimental conditions using the overall optimal vibration level of the 20 contents (Fig. 3). A regression analysis was performed for each combination of the experimental conditions to examine the differences between the three regarding the tendency of the overall optimal vibration level for each content. The regression lines resulting from the regression analysis are shown in Fig. 3, and the correlation coefficients for each experimental condition are shown in Table 2. These results indicate that the overall optimal vibration levels for each content among each experimental condition have a similar tendency. These results are almost the same as the analysis of the temporal level change of the optimal vibration level.

3.2. **The relationship between the results of each experimental condition and audio-visual information.** Analysis of the overall tendency of the optimal vibration levels showed a relationship between the three experimental conditions. If the optimal vibration level can be estimated from the audio-visual information, it will be possible to add effective vibration to any audio-visual content. Therefore, we conducted a multiple regression analysis of the temporal level change of the optimal vibration levels obtained from the experiment in section 3.1. As already mentioned in section 3.1, the temporal level changes of the optimal vibration level examined were obtained for each experimental condition and for each content every 0.1s. For each of the optimal vibration levels obtained at 0.1s, multiple regression analysis was performed using the acoustic and visual features

Table 3. Correlation coefficients for each explanatory variable combination used in the multiple regression analysis in the AO condition

	Roughness	Sharpness	MFCC (1st)	MFCC (2nd)	ViLA
Loudness	0.073	-0.040	0.010	-0.087	0.30
Roughness	-	-0.042	0.010	-0.046	0.14
Sharpness	-	-	-0.88	-0.29	-0.26
MFCC (1st)	-	-	-	0.094	0.28
MFCC (2nd)	-	-	-	-	-0.013

at that point. Multiple regression analysis was performed for each experimental condition. In the analysis, we include the vibration level of ViLA as an explanatory variable because the ViLA used in this experiment as the pre-adjustment vibration was also considered to affect the optimal vibration level.

For acoustic features, three representative psychometric quantities (loudness[15], roughness[16], and sharpness[17]), and two parameters related to speech sound, the primary and secondary Mel-frequency spectrum coefficients (MFCC), were used. The correlation coefficients between each explanatory variable are shown in Table 3. The variable inflation factor (VIF) for each explanatory variable was less than 10. This means that there was no multicollinearity effect.

For visual features, three features were used: the ratio of salient objects occupying the screen[18], the sum of the magnitude of the optical flow within salient objects, and the magnitude variance of the optical flow in all pixels. The features for salient objects were selected based on the assumption that the motion and size of the salient objects that attract people's attention affect the optimal vibration level. The variance of the magnitude of the optical flow in all pixels was also selected with the expectation that it would be related to the object's motion. The score of these variances is small when there is no motion on the screen or when the entire screen is shaking in the same direction. However, the variance score is large when individual objects on the screen show some movement. For the optical flow acquisition method, we used the Farneback method, a one type of gradient method, that can obtain the optical flow from all pixels[19]. The luminance value of a pixel and its surroundings in each frame is approximated by a polynomial equation, and the amount of movement is estimated by comparing the polynomial equation between each frame. By doing this for each pixel, optical flow can be obtained for all pixels. However, in some contents, there were some situations where optical flow could not be calculated. In these situations, the surface of objects did not have stable surface features (e.g., water or lightning). This means that these situations do not satisfy the properties assumed by the gradient method, such as luminance invariance and motion constancy in the neighboring region[20]. As the results, following four contents are excluded from the analysis: two contents in the animation category, "River current" and "Object rotation," and two contents in the life category, "Waves" and "Lightning." Therefore, in the analysis of the VO and A&V conditions, only 16 of the 20 contents were used. However, since the AO condition is not related to the problem of acquiring visual features, the analysis was conducted on all 20 contents. The correlation coefficients between each explanatory variable are shown in Table 4. The variable inflation factor (VIF) for each explanatory variable was less than 10. This means that there was no multicollinearity effect.

The results of the multiple regression analysis are shown in Table 5, which shows that the coefficient of determination is greater than 0.6 for all experimental conditions. Table 5 also shows that all explanatory variables except for the first-order MFCC were significant

Table 4. Correlation coefficients for each explanatory variable combination used in the multiple regression analysis in the VO condition

	Sum of magnitude of the optical flow within salient objects	Variance of magnitude of the optical flow in whole pixels	ViLA
Ratio of salient objects in the screen	0.49	0.011	0.20
Sum of magnitude of the optical flow within salient objects	-	0.57	0.31
Variance of magnitude of the optical flow in whole pixels	-	-	0.35

Table 5. Results of the multiple regression analysis for each condition with optimal vibration levels as the dependent variable

Explanatory variables	AO	VO	A&V
	Standard partial regression coefficient	Standard partial regression coefficient	Standard partial regression coefficient
Loudness	0.23*	-	0.21*
Roughness	0.081*	-	0.093*
Sharpness	-0.19*	-	-0.24*
MFCC (1st)	0.015	-	-0.085*
MFCC (2nd)	-0.28*	-	-0.21*
Ratio of salient objects occupying the screen	-	-0.16*	-0.057*
Sum of magnitude of the optical flow within salient objects	-	0.18*	0.13*
Variance of magnitude of the optical flow in all pixels	-	0.20*	0.026*
ViLA	0.66*	0.62*	0.60*
Adjusted determination coefficient : R^2	0.68*	0.63*	0.65*
$*$: $p < .05$	n = 7009	n = 5697	n = 5697

in the AO condition ($p < .05$), and all explanatory variables were significant in the VO and A&V conditions ($p < .05$). In the A&V condition, the trend of the standard partial regression coefficients for each explanatory variable is similar to those in the AO condition. However, the standard partial regression coefficients of the visual features in the A&V condition, especially the variance of magnitude of the optical flow in all pixels, are lower than those in the VO condition.

Table 6. AIC for each regression model

	Acoustic features &Visual features & ViLA	Acoustic features &ViLA	Visual features &ViLA	only ViLA
AO	-	1.15×10^4	-	1.37×10^4
VO	-	-	1.03×10^4	1.16×10^4
A&V	0.97×10^4	0.99×10^4	1.10×10^4	1.12×10^4

3.3. **Prediction accuracy comparisson for each model.** To analyse the improvement in the prediction accuracy by introducing acoustic and visual features, Akaike's information criterion (AIC), an index with high prediction accuracy, was calculated for the multiple regression equation in Table 5 and for the single regression equation with ViLA-only in Table 6. The AIC was also calculated for the models using each feature together with ViLA to consider the respective effects obtained by introducing acoustic and visual features in the A&V condition. Table 6 shows that the models using acoustic and visual features with ViLA have the lowest AIC in the A&V condition. In addition, the AIC of the model using acoustic features with ViLA is lower than that of the model using visual features with ViLA in the A&V condition. The AIC is a relative index, with lower values indicating higher predictive accuracy. Therefore, it was shown that the accuracy of prediction could be improved by introducing acoustic and visual features in addition to ViLA in the A&V condition and that the contribution of acoustic features to the accuracy of prediction is larger than the contribution of the visual features.

For a more practical evaluation, the root-mean-square error (RMSE) of the estimated optimal vibration level is used to evaluate the improvement in accuracy by introducing acoustic and visual features. The RMSE was calculated for each content between the optimal vibration level estimated by the models shown in Table 6 and the temporal level change of the optimal vibration level (Section 3.1). Figure 4 shows the RMSEs obtained for the 20 contents with a box-plot. Figure 4 shows that the variance of the RMSE for the 20 contents is smaller, especially in the AV condition. However, the RMSE of the multiple regression equation is almost the same as that of the ViLA single regression model for most contents.

4. **Discussion.** In this study, to examine the method of determining the optimal vibration level, we conducted an experiment in which observers were asked to adjust the optimal vibration levels for audio-visual content. The relationship between the optimal vibration level and audio-visual information was analyzed using multiple regression analysis, suggesting that observers adjusted the optimal vibration levels using several acoustic and visual features.

First, we discuss how acoustic features affect the optimal vibration levels. In the AO and A&V conditions, the two explanatory variables with large positive standard partial regression coefficients were loudness and ViLA. The ViLA used as the pre-adjustment vibration in this experiment was generated by applying a low-pass filter to the sound. This means that temporal changes in vibration levels would be closely correlated to temporal changes in the low-frequency band level of the sound. Therefore, the positive values of these two standard partial regression coefficients suggests that the observers regarded the amplitude related with the sound amplitude as optimal. Generally, there are multiple cases in everyday life where loud sounds are generated when large vibrations occur, due

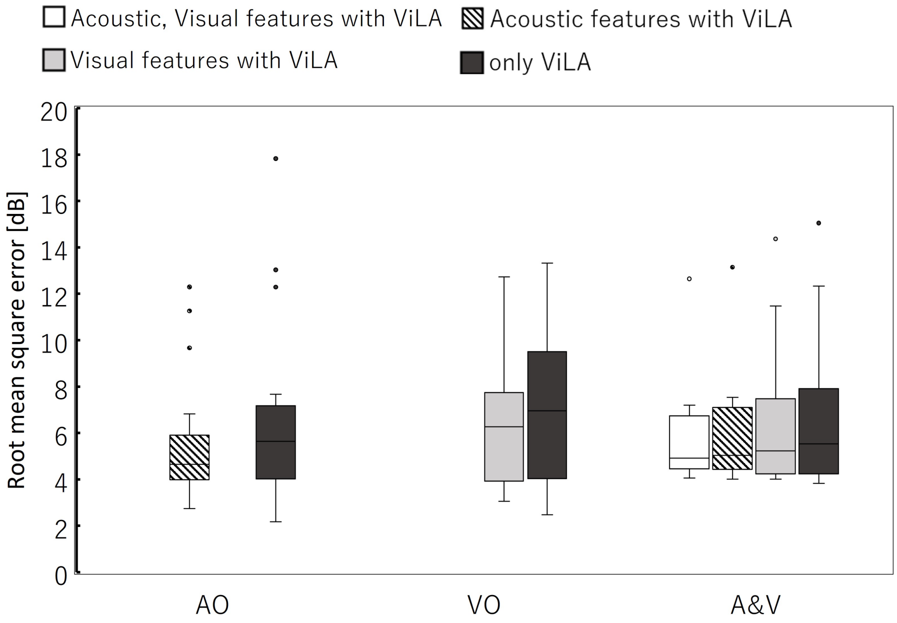

FIGURE 4. RMSEs obtained for the 20 contents for each experimental condition

to colliding objects, etc. Therefore, in our experiments, loud sounds might have caused the observers to expect larger vibrations.

Acoustic features such as sharpness and low-order MFCC negatively influenced the optimal vibration levels. Sharpness corresponds to a higher sound pitch. The fact that this feature has a negative partial regression coefficient could explain the tendency of an observer to adjust the vibration to a smaller level when the sound has a solid high-frequency component. The "heaviness of sound" is related to the low frequency of sound[21], and therefore, observers might adjust their vibration to a greater degree by feeling the power from the low frequency of the sound. Another reason could be that low sounds are often perceived as vibrations in everyday life, and this experience may have caused the observers to expect vibrations for low-frequency sounds. MFCC is a feature that is strongly related to voice [22]. There are very few situations in daily life in which floor vibrations occur in conjunction with speech sound. Therefore, the observers may have concluded that the vibration presentation for the voice was inappropriate. This decision may have resulted in a lower optimal vibration level, and ultimately, a negative value for the standard partial regression coefficient of MFCC, which indicates a voice-specific value.

Second, we discuss how visual features affect the optimal vibration levels. In the VO and A&V conditions, the two visual features related to salient objects had opposite effects. The ratio of salient objects occupying the screen negatively affected the optimal vibration level, and the sum of the magnitudes of the optical flows within salient objects positively affected the optimal vibration level. These results may be related to the fact that salient objects do not necessarily include motion. For example, if a large building or statue is shown on the screen, it is likely to be a prominent object, but no movement exists. This suggests that whether or not the object is moving is an essential factor for the optimal vibration level.

The variance in the magnitude of the optical flow in all pixels positively affected the optimal vibration level. The effect was large for the VO condition and small for the A&V condition. As this feature is related to the motion presented throughout the screen, it is still possible that the motion of the object is strongly related to the optimal vibration level. A possible reason for the different influence of the variance of the magnitude of the optical flow in all pixels in the two conditions could be whether or not audio information is presented. Because object motion is often accompanied by sound, audio information could provide the observer with events related to the motion present in the content. There is no sound in the VO condition. Therefore, the observers searched the entire screen for events related to motion, which could have resulted in a strong association with the magnitude of the optical flow in all pixels. This observation is based on the comparison of the results of the AV and VO conditions and we speculate that some mutual interaction effects may exist between audio and visual information. Such mutual interaction between audio and visual information would have occurred for the other features but we could not clearly observe it. This point seems us an interesting future study topic to deepen our insight how observers use audio-visual information to determine the optimal vibration level.

In terms of whether acoustic or visual features had more influence on the optimal vibration level, Table 6 suggested that the influence of acoustic features on the optimal vibration level of the AV conditions was greater than that of the visual features. This reason may also be related to the fact that audio information conveys motion information of events better than visual information (e.g., audio information can convey off-screen motion information and motion information inside machine, etc). Due to these sound characteristics, it is possible that the observer adjusted the optimal vibration level by placing more emphasis on the audio information.

Next, we discuss how ViLA, which was employed as an explanatory variable, influenced the optimal vibration level. ViLA had a large standard partial regression coefficient in the VO condition as well as in the AO and A&V conditions. It was somewhat surprising that it had a high partial regression coefficient even in the VO condition, where no audio information was presented. In general, audio and visual information is often synchronized, so the observers may not have had any reason to be confused by the temporal amplitude changes in ViLA. This consideration may also explain the relationship of the tendency of optimal vibration levels between the AO and VO conditions. Although only one of the sensory information is presented in the AO and VO conditions, the tendency of the optimal vibration levels would be similar, due to the synchronization of audio and visual information.

As shown in the previous section, the optimal vibration level can be estimated to some extent by using features obtained from audiovisual information. However, Table 5 shows that the coefficient of determination is around 0.6 for all experimental conditions, which means that the optimal vibration levels obtained in the experiments are not perfectly estimated. Furthermore, as shown in Fig. 4, the difference between the estimated vibration level using the multiple regression equation with acoustic and visual features and that using the single regression model withy ViLA is not very large. Inclusion of the information related to the object types appearing in the content might be helpful. To confirm this effect, we analyzed the relationship between the optimal vibration level calculated by the multiple regression model and that obtained by the experiment for each of the reclassified categories. Figure 5 shows the result of the analysis. One plot represents one content, and the error bars represent the standard deviation of the optimal vibration level for each content. In Fig. 5, there are category-specific biases in all experimental conditions. These results suggest that, it would be necessary to build a model that includes the information of object types to estimate the optimal vibration level accurately.

5. **Future studies.** This experiment made it possible to estimate the optimal vibration level for any audio-visual content. The results also indicated that the influence of features such as ViLA, loudness, and sharpness was particularly significant in determining the

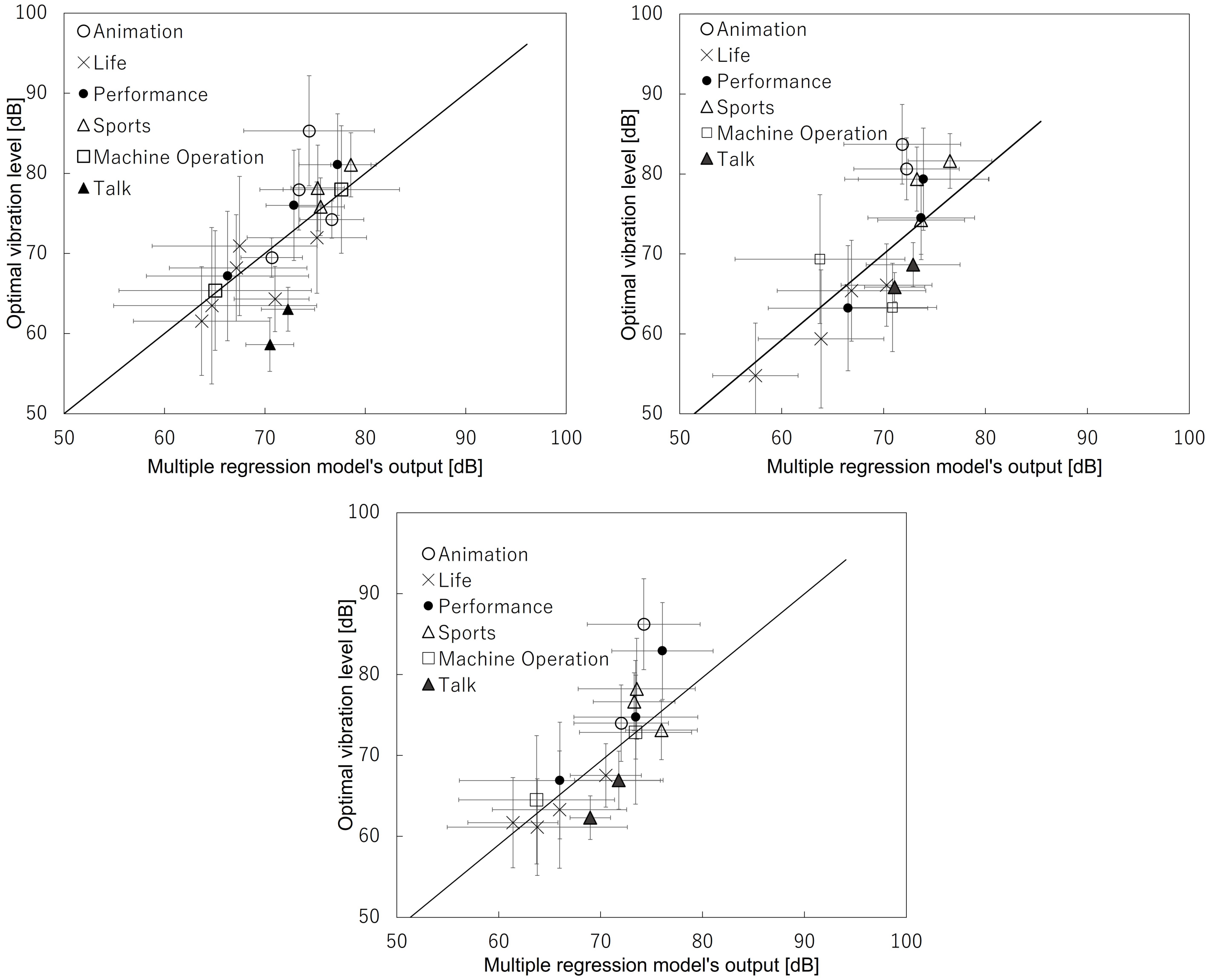

FIGURE 5. The relationship between the optimal vibration level and the output of the multiple regression model : Top left is AO condition, Top right is VO condition, Bottom is A&V condition

vibration amplitude under the A&V conditions. However, the influence of visual features such as optical flow is also statistically significant. Therefore, it would be an interesting future research topic to estimate the optimal level of the vibration amplitude more accurately by adopting audio and visual information.

It is also important to investigate whether the optimal vibration level increases the perceived reality. Previous studies have demonstrated that perceived reality significantly depends on the amplitude of the vibration presented[5]; for example, the sense of presence increases as the vibration amplitude increases, and the sense of verisimilitude peaks at a certain vibration amplitude. In future studies, it is important to investigate the optimal vibration level from the perspective of perceptual reality.

6. **Conclusions.** This study investigated the vibration level perceived as optimal for the audio-visual content by the observers. The results suggest that both audio and visual information determines the optimal amplitude of the vibration for audio-visual content. Furthermore, we examined the relationship between the optimal vibration level and acoustic and visual features using multiple regression analysis. As a result, the optimal vibration level was explained in about 60% of the experimental results. The acoustic features

that particularly contributed to determining the optimal vibration level were loudness, sharpness, and second-order MFCC, while the visual feature was optical flow in salient objects. This research suggests that the optimal vibration level can be estimated for any type of content by extracting the acoustic and visual features.

Acknowledgment. Part of this work was supported by RIEC Nation-wide Cooperative Research Projects and JSPS KAKENHI Grant Numbers 19H04145.

REFERENCES

[1] A. Honda *et al.*, :Determinants of Senses of Presence and Verisimilitude in Audio-Visual Contets, The Journal of the Virtual Reality Society of Japan, Vol. 18, No. 1, pp. 93-101, 2013. (in Japanese)

[2] K. Ozawa *et al.*, : Effects of the Relation between Reproduced Sound Pressure Levels and Viewing Angles on the Sense of Presence in Audio-visual Content, Journal of Information Hiding and Multimedia Signal Processing, Vol. 7, No. 1, pp. 31-40, 2016.

[3] M. Ito *et al.*, : Expanded Estimation Model for Instantaneous Presence in Audio-visual Content Incorporating Binaural Information, Journal of Information Hiding and Multimedia Signal Processing, Vol. 8, No. 5, pp. 1092-1102, 2017.

[4] W. Teramoto *et al.*, : What is "sense of presence?" A non-researcher's understanding of the sense of presence, The Journal of the Virtual Reality Society of Japan, Vol. 15, No. 1, pp. 7-16, 2010. (in Japanese)

[5] S. Sakamoto *et al*, : Body vibration effects on perceived reality with multi-modal contents, The ITE Transactions on Media Technology and Applications, Vol. 2, No. 1, pp. 46-50, 2014.

[6] Z. Cui *et al*, : Influence of Visual Depth and Vibration on the High-level Perception of Reality in 3D Contents, Journal of Information Hiding and Multimedia Signal Processing, Vol. 8, No. 6, pp. 1382-1391, 2017.

[7] S. Merchel *et al*, : The influence of vibration on musical experience, Journal of the Audio Engineering Society, Vol. 62, No. 4, pp. 230-234, 2014.

[8] Y. Sawada *et al.*, : Effects of synchronized engine sound and vibration presentation on visually induced motion sickness, Scientic Reports, Vol. 10, No. 1, pp. 1-10, 2020.

[9] Bob G. Witmer et al., : Measuring Presence in Virtual Environments: A Presence Questionnaire, Presence, Vol. 7, No. 3, pp. 225-240, 1998.

[10] Z. Cui *et al.*, : How can body vibration generated from audio signal in AV content enhance perceived reality?, Information Processing Society of Japan, Vol. 59, No. 11, pp.1986-1994, 2018.(in Japanese)

[11] D. Gongora *et al.*, : Vibrotactile Rendering of Camera Motion for Bimanual Experience of First-Person View Videos, Proceedings of IEEE World Haptics Conference, pp.454-459, 2017.

[12] https://vimeo.com/ (accessed April 16th 2022)

[13] https://creativecommons.org/ (accessed April 16th 2022)

[14] International Organization for Standardization : Mechanical vibration and shock -Evaluation of human exposure to whole-body vibration. Part 1 : General requirements. ISO 2631-1, 1997.

[15] B. C. J. Moore *et al.*, : Testing and refining a loudness model for time-varying sounds incorporating binaural inhibition, The Journal of the Acoustical Society of America, Vol. 143, No. 3, pp. 1504-1513, 2018.

[16] H. Fastl *et al.*, : Psychoacoustics: Facts and Models, 3rd ed. (Springer, Berlin, Heidelberg), pp. 257–264, 2006.

[17] DIN 45692 : Measurement Technique for the Simulation of the Auditory Sensation of Sharpness, German Institute for Standardization, 2009.

[18] X. Qin *et al.*, : U2-Net : Going deeper with nested U-structure for salient object detection, Pattern recognition, Vol. 106, pp. 1-15, 2020.

[19] G. Farneback : Two-Frame Motion Estimation Based on Polynomial Expansion, Proceedings of Scandinavian Conference on Image Analysis, pp. 363–370, 2003.

[20] T. Corpetti et al., : Estimating fluid optical flow, Proceedings 15th International Conference on Pattern Recognition, pp. 1033–1036, 2000.

[21] LN. Solomon *et al.*, : Semantic Approach to the Perception of Complex Sounds, and Research for Physical Correlates to Psychological Dimensions of Sounds, The Journal of the Acoustical Society of America, Vol. 30, pp. 421-497, 1958.

[22] K. Chakraborty *et al.*, : Voice recgnition using MFCC algorithm, International Journal of Innovative Research in Advanced Engineering, Vol.1, No.10, pp.158-161, 2014.