Journal of Information Hiding and Multimedia Signal Processing ©2017 ISSN 2073-4212
Ubiquitous International Volume **8**, Number **5**, September **2017**

Expanded Estimation Model for Instantaneous Presence in Audio-visual Content Incorporating Binaural Information

Masaaki Ito, Kenji Ozawa, Masanori Morise and Yuichiro Kinoshita

Integrated Graduate School, University of Yamanashi
4-3-11 Takeda, Kofu, Yamanashi 400-8511, Japan
ozawa@yamanashi.ac.jp

Received December 2016; revised June 2017

Abstract. *The sense of presence is a key component of the performance of multimedia content and systems. Our previous studies have shown that the sense of presence in audio-visual (AV) content has two elements: content presence and system presence. We constructed an estimation model of content presence as a time series. The accuracy of this model is compromised because it does not consider audio system presence. Therefore, a model that takes auditory system presence into account is needed. To construct such a model, we first conducted an experimental evaluation of instantaneous presence for 40 AV content items, using two auditory-reproduction methods of binaural and diotic reproduction techniques. Based on the experimental results, we constructed a neural network-based model that uses 19 AV features, extracted from the content items in 500-ms intervals, considering binaural information. The 19 features consist of 7 audio and 12 visual features. The audio features include two interaural information-related measures which are introduced to represent auditory system presence, i.e. the spatial impression of a sound. The visual features are basically the same as those in our previous model. A generalization test of the expanded model confirms that it is sufficiently accurate to estimate time series presence.*
Keywords: Sense of presence, Content and system presence, Audio-visual content, Audio reproduction method, Neural network.

1. **Introduction.** Advanced audio-visual (AV) systems, including high-definition (HD) television (TV), 3-dimensional (3D) TV and ultra-HD (UHD) TV with audio systems suitable for a UHD TV (such as the 22.2-multichannel system [1, 2]), have been developed and some of them have achieved widespread consumer adoption. The sense of presence is a key performance attribute of advanced multimedia systems. Therefore, a standardized methodology is needed for evaluating the sense of presence in AV systems. Although a number of studies have examined the properties of the sense of presence [3, 4, 5, 6, 7, 8], a complete model that evaluates the degree of presence has yet to be demonstrated. Our research group has been developing a model that estimates the sense of presence, based on the properties of AV content items [9, 10, 11, 12]. Our ultimate goal is to create a *presence meter*, which estimates the sense of presence accurately based on the physical properties of an AV stimulus. A presence meter would be useful for both creators of AV content and consumers of AV equipment, in order to optimize the performance of created content and to evaluate equipment for consumers, respectively.

When an advanced TV system is installed in a home, its visual condition is mostly determined by the size and resolution of its visual display and by the seated position of

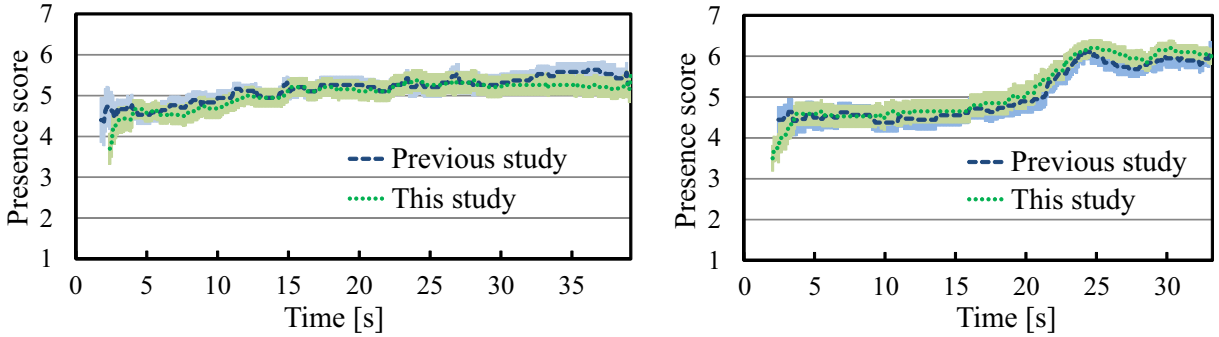
the viewer. On the other hand, audio condition has several degrees of freedom. Depending on the loudspeaker setup in the home, a 22.2-channel audio signal [1, 2] can be reproduced fully, or it may be reproduced by down mixing to a 5.1-channel or a 2-channel audio signal. The viewer might use headphones instead of loudspeakers. These imply that an accurate evaluation of the effects of binaural information is important to the accuracy of presence models. Therefore, this study focuses on the effects of binaural information on the sense of presence.

Our previous investigations demonstrated that perceived presence depends on the reproduced content items, even if the same audio system is used for recording and reproduction [13, 14]. Thus, the sense of presence has two aspects: *system presence*, which is determined by the characteristics of the AV system used, and *content presence*, which depends on the characteristics of the reproduced content item [14]. However, we consider that it is difficult to split the sense of presence exactly into these two aspects. The difference in content presence can be observed when different content items are evaluated using the same recoding and reproduction system, and the difference in system presence is able to be seen when different recording and reproduction systems are used for a specific content item. Our first study focused on content presence and developed a model to estimate the overall presence of an AV content item (the presence of the entire item) [9]. In subsequent work, we observed the effects of binaural information on the sense of presence and developed a model that takes system presence into consideration [10]. This evaluates the overall presence of a content item, but not instantaneous presence. In a more recent study, we developed a content presence meter for an AV content item, which estimates instantaneous presence based on the AV features of the item [11]. We then constructed an audio presence meter for an audio content item, taking system presence into consideration [12]. Based on the AV presence meter [11], this study aims to expand the estimation model of instantaneous AV presence by adding the effects of binaural information on system presence [12]. This is because the sense of audio presence is evaluated finally by a person with his or her two ears for all audio systems.

2. **Presence evaluation experiment.**

2.1. **Overview of the experiment.** As a pre-requisite to constructing an estimation model, we created a dataset comprised of the measured instantaneous presence, on a seven-point scale, for 40 content items. We used the same experimental procedure as we did for our previous study [15]. However, binaural reproduction [16] was used in the previous investigation [15], while diotic reproduction was used in the present study. Diotic reproduction means that exactly the same sound signals are presented to both ears. We focused on the binaural and diotic reproduction methods because, of the five methods tested in [10], they are the most and least effective methods, respectively, to create the sense of presence [10]. In this experiment, diotic sounds were synthesized by averaging the left and right channels of a binaural sound and presented diotically to subjects via headphones.

2.2. **Experimental methods.** The experimental methods described in detail in [15], are summarized briefly as follows. Forty content items, each lasting about 20 to 40 s, were recorded using a full-HD video camera (Sony, PMW-EX1R). Sound was binaurally recorded using a dummy head (Koken, SAMRAI). The visual stimuli were reproduced on a 65-inch full-HD display (Sharp, LC-65GX5). The distance between the display and a subject was 2.4 m, which corresponds to 3H (Height of a display) in the ITU Recommendation [17]. The auditory stimuli were reproduced diotically or binaurally to subjects by headphones (Sennheiser, HD650): the 40 content items were diotically reproduced, and



(a) *Passing vehicles.* (b) *Passing roller coaster A (front).*

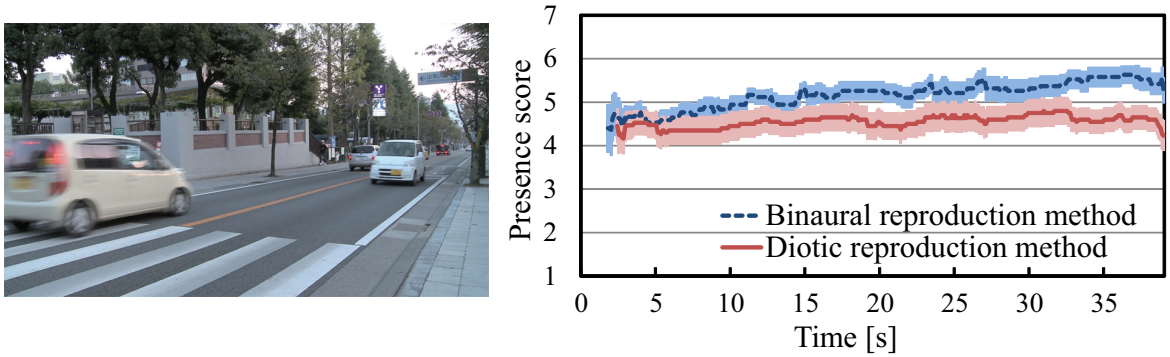
FIGURE 1. Examples of the evaluation results compared to the previous experiment [15]. Error bars indicate the standard errors calculated at 100-ms intervals.

five out of the 40 content items were also reproduced binaurally, resulting in a total of 45 stimuli. Because the subjects were different from whom had participated in the previous experiment [15] in which all stimuli were reproduced binaurally, we needed to check the the consistency between the different subject groups by using the same binaurally reproduced stimuli. These stimuli were presented in random order. The experimental subjects were twenty undergraduate students (five female and 15 male).

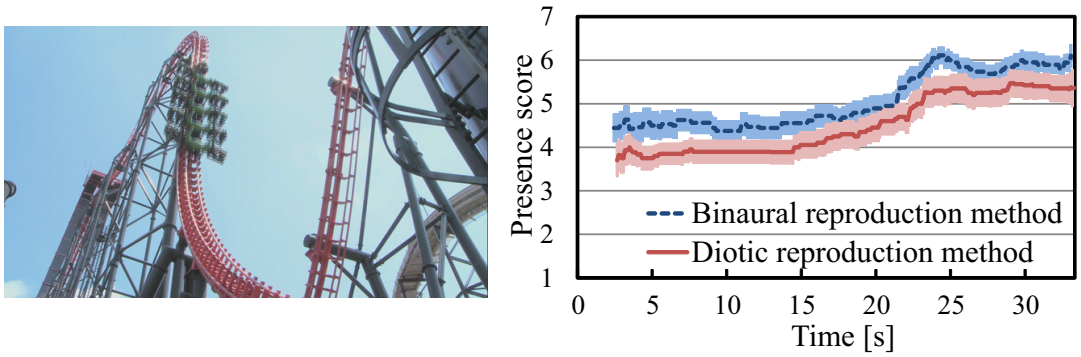
Prior to the experiment, the sense of presence was defined as "a feeling that you are actually in the situation/location." No request was given to the subjects how they judged the sense of presence. The subject responded instantaneously to the sense of presence, using the method of continuous judgment by category [18] by pressing one of seven keys (1–7) on a computer keyboard. The 1 key indicates "No Sense of Presence" and 7 indicates a "Strong Feeling of Presence." Each subject was instructed not to press a key until he or she could evaluate presence after the stimulus began. After each item was presented, the subjects were asked to indicate the overall presence score using the seven-point scale.

2.3. **Experimental results.** We first confirmed the consistency between the two subject groups. Figure 1 shows the average presence scores and the standard errors (SEs) among the subjects, for two examples out of the five stimuli that were common between the present and the previous experiment [15]. Because latency (the time interval between the start of stimulus presentation and the start of the response) differs among the subjects, the averages and SEs were calculated after half (ten) subjects began their responses. The figures show strong consistency between the two subject groups for the two stimuli shown; this consistency is similar for the other three common stimuli (*Passing train*, *Passing roller coaster A (rear)* and *View of a university building*). Thus, we were able to use the result dataset from the previous experiment [15] to represent binaural reproduction in this estimation model.

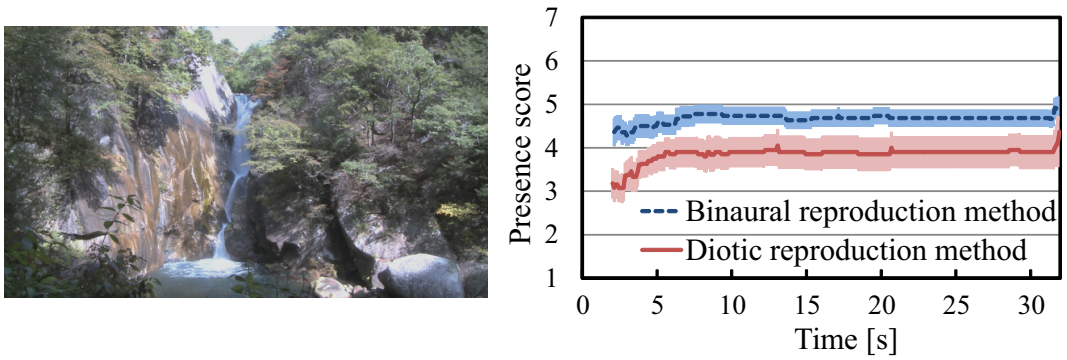
We then observed the effect of the two reproduction methods on presence. Figure 2 shows results for three stimuli, with their corresponding still pictures. As shown in the figure, diotic reproduction has less presence than binaural reproduction. This tendency is consistent with our previous results [10, 12]. The different reproduction methods have less effect for stimuli with moving sound sources (such as the *vehicles* and the *roller coaster* items shown in Fig. 2(a) and (b), respectively), than those when only auditory stimuli were presented [12]. Moreover, the difference of reproduction methods is larger in the static content such as the *scene of waterfall*. Therefore, an estimation model of instantaneous presence must consider the effect of visual stimuli.



(a) *Passing vehicles.* Still picture is the scene approximately 12 s after the stimulus began.



(b) *Passing roller coaster A (front).* Still picture is the scene approximately 21 s after the stimulus began.



(c) *Scene of a waterfall.* Still picture is the scene approximately 15 s after the stimulus began.

FIGURE 2. Effects of binaural information on the instantaneous presence for three example items (Left panel: still picture, Right panel: instantaneous presence values.) Error bars indicate the standard errors calculated at 100-ms intervals.

Figure 3 shows the averages of the overall presence scores of all 40 content items. As demonstrated in [15], overall presence is highly correlated with the upper 10th-percentile exceeded presence score, S_{10} (the score that is exceeded for 10 percent of the time during the instantaneous responses). This correlation holds for the present experiment, as the correlation coefficients are 0.97 and 0.99 for the binaural and diotic presentation methods, respectively. Figures 2 and 3 together show that the effects of binaural information on overall presence are dependent on the content item. Although the stimuli are categorized into four groups according to the movements of sound and visual images [15], we cannot find any specific relation between the kind of the stimuli and the effects of the different reproduction methods. This indicates that the relation between the AV features of the

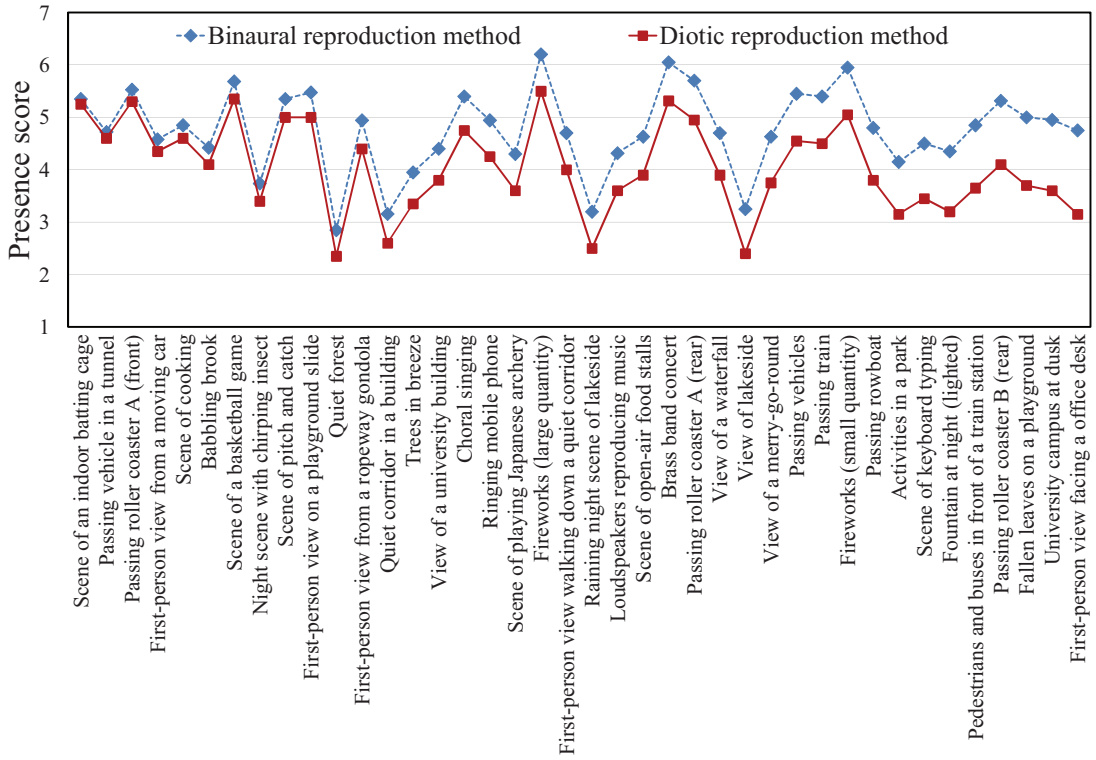


FIGURE 3. Average scores of the overall presence evaluations for each content item reproduced by the two different methods. Items are arranged in the ascending order of the difference between the two methods.

stimuli and perceived presence is so complex that a nonlinear structure will be required for an estimation model of presence. Therefore, we decided to expand our previous estimation model based on artificial neural networks [11].

3. **Expansion of the presence estimation model.**

3.1. **Overview of model construction.** As described in Sect. 1, we previously developed a content presence estimation model for an AV content item in which the instantaneous presence is estimated based on the audio-visual features of the item [11]. The inputs to this model comprise six audio-related features and fourteen visual-related features, calculated in 500-ms intervals. These features are not sufficient to estimate the effect of binaural information on instantaneous presence. We propose an expanded model, to estimate instantaneous presence taking into account the effects of binaural information [12].

3.2. **Audio-related features.** The present model has the following seven audio-related input features, calculated at 500-ms intervals:

(A1) Loudness estimated by Fastl and Zwicker [19]
(A2) Sharpness estimated by Fastl and Zwicker [19]
(A3) Roughness estimated by Vassilakis [20]
(A4) Dynamic range: difference between 95th- and 5th-percentile sound pressure levels
(A5) Standard deviation of the dynamic ranges calculated for the most recent 3-s interval
(A6) Interaural level difference
(A7) Interaural correlation coefficients

Features (A1) to (A5) were also used in the previous model [11]. Feature "(A6) Movements of sound images" in the previous model was discarded, and new features "(A6) Interaural level difference" and "(A7) Interaural correlation coefficients" were introduced. The previous (A6) was a 1-bit feature in which a value of '0' was assigned to content items without sound-image movement, while a value of '1' was assigned to those with sound-image movement. Although this value was determined based on the interaural level difference, the 1-bit resolution was not enough to represent the spatial impression of a sound. The new features (A6) and (A7) are expected to represent spatial impression because they are known to be cues of spatial sound perception such as sound localization and envelopment [21].

These features are valid for all audio systems including a multichannel audio system which does not have binaural information. This is because any audio system is finally evaluated by listeners with their two ears. Binaural signals can be measured for the listener or a dummy head. These signals are also able to be calculated using room transfer functions and head-related transfer functions [21].

3.3. **Visual-related features.** In addition to the audio-related features, the present model uses the following twelve visual-related features calculated in 500-ms intervals:
Hue — Average number of pixels with the following values:

(V1) Hue values 36–107 (yellow)
(V2) Hue values 108–179 (green)
(V3) Hue values 180–251 (blue)
(V4) Hue values 252–323 (purple)
(V5) Hue values 324–359 and 0–35 (red)

Lightness:

(V6) Average lightness value
(V7) Standard deviation of lightness values
(V8) Skewness of lightness values

Saturation:

(V9) Average saturation value
(V10) Standard deviation of saturation values
(V11) Skewness of saturation values

Number of pixels corresponding to moving objects:

(V12) Average number of pixels corresponding to moving objects per frame

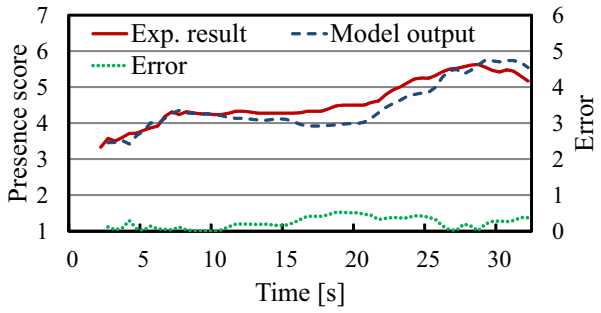
Our previous model [11] used fourteen features. In addition to (V1) to (V12), it included two further features, "(V13) Difference between 95th- and 5th-percentile numbers of moving pixels in a frame," and "(V14) Standard deviation of (V13)," each calculated at 500-ms intervals. These two features were omitted here, because their contributions to estimating presence were relatively small [11].

These features are extracted from video signals. In this experiment, the subjects sat in accordance with the ITU Recommendation [17]. If the visual condition is different from this standard condition, the visual features must be compensated for the difference. When the visual system presence is discussed in the future, this compensation will be mandatory.

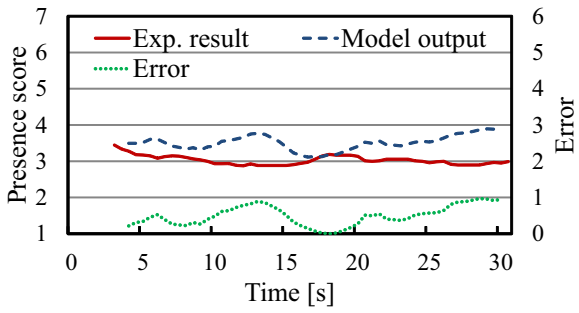
3.4. **Model construction based on neural networks.** The seven audio-related and twelve visual-related features are used as inputs to the model by normalizing each of them to a range of 0–1, where 0 and 1 correspond to the minimum and maximum values for all of the content items, respectively. The basic structure of the neural network is the

(a-1) Passing train (Still image at 23 s)

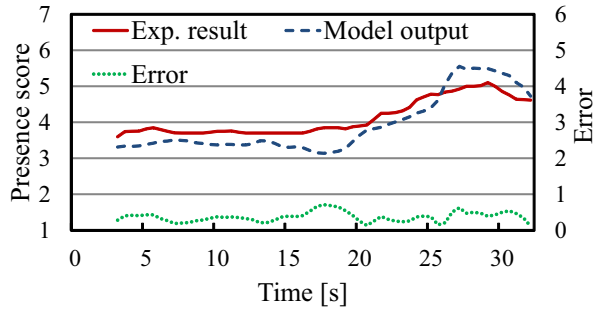
(b-1) View of lakeside (Still image at 10 s)



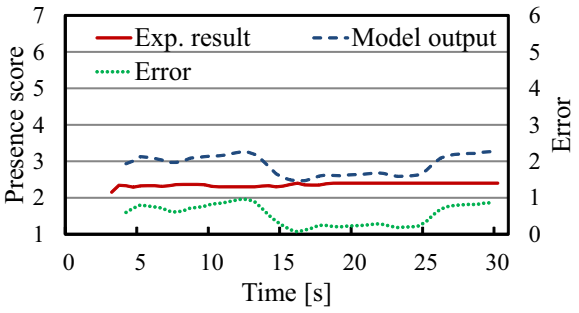
(a-2) Passing train (Binaural method)



(b-2) View of lakeside (Binaural method)



(a-3) Passing train (Diotic method)



(b-3) View of lakeside (Diotic method)

FIGURE 4. Comparisons between the experimental results and the model outputs for two example items with each reproduction method.

same as in our previous model [11]. The expanded model consists of input, hidden, and output layers, which have 19, 15, and 1 units, respectively. The number of hidden layer units was determined by a preliminary examination. The neural networks are trained by back-propagation, using the average evaluation scores obtained by the experiment. The evaluation scores are also normalized to a range of 0–1, where 0 and 1 correspond to 1 and 7 on the seven-point presence scale, respectively.

Although the number of stimuli is 80 consisting of the 40 binaurally presented items plus their 40 diotic counterparts, the total number of learned and tested content items is 4,630 because every 500-ms interval is treated as a learned or tested sample. In the learning process, we made the evaluated presence score of a 500-ms frame responsive to the features preceded by 2 frames. This is because the latency of presence judgment evoked by a physical stimulus is about 1 s [15].

4. **Evaluation of the model.**

4.1. **Evaluation method.**

The performance of the model was tested by cross-validation. Because the number of stimuli is limited to 80, we used the following testing method to

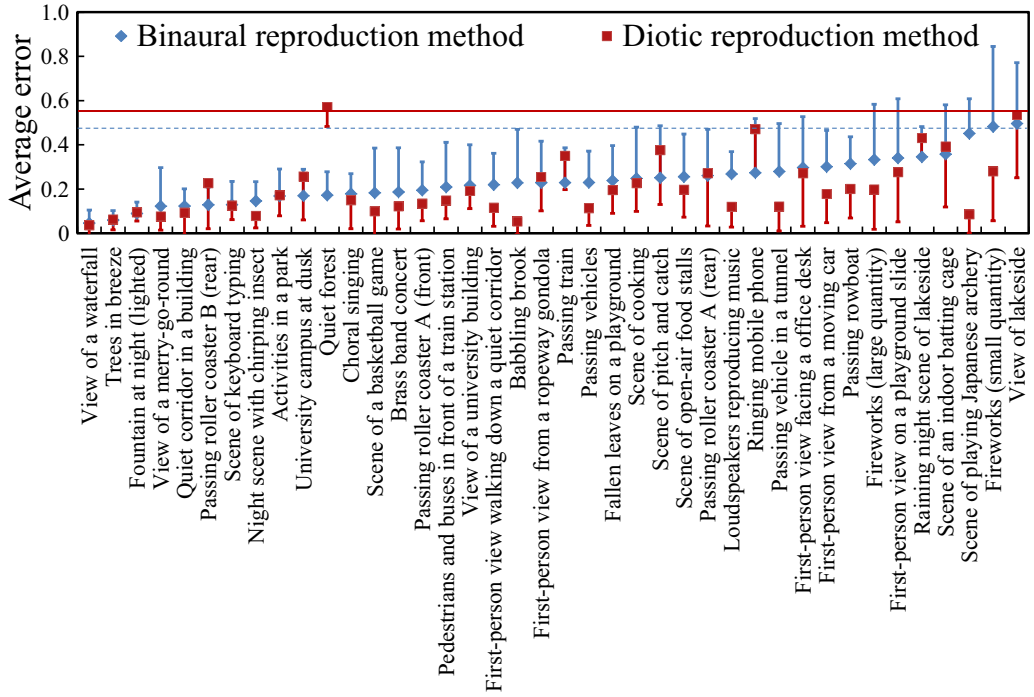


FIGURE 5. Average error of each content item between the experimental result and the model output. Items are arranged in the ascending order of the average errors for the binaural reproduction method. Horizontal solid and dashed lines represent the averages of the 95% confidence intervals in presence evaluation, for the stimuli with the diotic and binaural reproduction methods, respectively.

increase the number of content items used to train the network. Each of the content items, with both reproduction methods, was used, in turn, to test the generalization performance of the network trained by the remaining 78 stimuli. This generalization test was conducted for all 40 possible combinations of testing and training.

4.2. **Comparison between the evaluation results and the model outputs.** Figure 4 shows results for two examples of content items. The *passing train* item has large movement of a sound source, while the *lakeside* item has no moving sound source. The solid and dashed lines in the figure denote the experimental results and the three-point moving averages of the model output, respectively. The dotted lines show the absolute errors between the experimental results and the model outputs.

For the *passing train* item, the magnitudes of errors are less than 1 for both reproduction methods. Although the *lakeside* item shows relatively greater errors, the magnitudes of errors are still less than 1 for both reproduction methods. The *lakeside* item also showed the largest error with the content AV model [11], yet the error was not large in the auditory presence meter [12]. This suggests that the large error is due to visual-related features and we need further consideration for visual-related features.

Figure 5 shows the average errors during the instantaneous responses for each content item, with the two reproduction methods. In the figure, the horizontal solid and dashed lines represent the averages of the 95% confidence intervals of the evaluated presence scores for the stimuli with the diotic and binaural reproduction methods, respectively. Although the *lakeside* item has the greatest errors for both reproduction methods in average, the errors are almost the same as the averages of the 95% confidence intervals of subjects' judgments. The *quiet forest* item shows the largest error only for diotic reproduction but

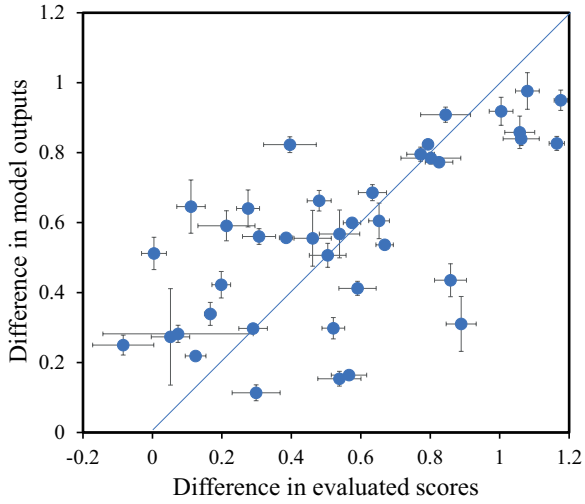


FIGURE 6. Relation of the differences of presence scores for 40 content items between the experimentally evaluated scores and the model outputs. The difference was defined by subtracting the score of diotic reproduction from that of binaural reproduction. Error bars show the 95% confidence intervals.

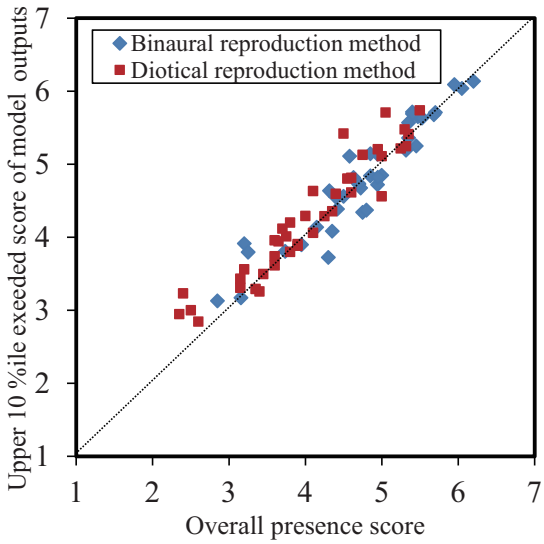


FIGURE 7. Relation between the evaluated overall presence scores and the upper ten-percentile exceeded scores of the model outputs.

the error is almost the same as the average of the 95% confidence interval. The means of the average errors for the 40 content items are 0.24 and 0.21 for the binaural and the diotic methods, respectively. This indicates that the model sufficiently estimates instantaneous presence. As the model is based on an artificial neural network, training it with more content will improve its performance.

The accuracy of estimation is considered from the viewpoint of comparison between the differences of reproduction methods. Here the difference is defined by subtracting the score of diotic reproduction from that of binaural reproduction. Figure 6 exhibits the relation of the differences between the experimentally evaluated scores and the model outputs.

In the figure, the average over stimulus presentation was plotted and the 95% confidence interval is shown as an error bar. The correlation coefficient is 0.65 ($t_{38} = 5.32, p < 0.001$). In the model learning, the raw scores of presence were used as teaching signals. Thus the error in the difference of the model outputs between the two reproduction methods is an accumulation of two errors shown in Fig. 4. If we can take the difference between two reproduction methods into consideration in the learning process, the accuracy will be improved. This must be an important future work.

4.3. **Comparison between the overall presence scores and the model outputs.** As described in Sect. 2, the subjects rated the overall presence for each item, with results as shown in Fig. 3. Figure 7 shows the correlation between the overall presence of each item using both reproduction methods, and the upper 10th-percentile exceeded scores, S_{10}, of the corresponding model outputs. The correlation coefficients for the two reproduction methods are both 0.95. This means that the present model is applicable to estimating not only instantaneous presence, but also overall presence, irrespective of the audio reproduction method used.

5. **Conclusions.** This study has focused on the effects of binaural information on instantaneous presence evaluation. First, an evaluation experiment was conducted using binaural and diotic reproduction methods. Diotic reproduction has less presence than binaural reproduction. Based on experimental results, we expanded our previous model to estimate instantaneous presence using audio and visual features, considering binaural information. The improved model should contribute to the development of an AV presence meter.

More stimuli are required for further improvement of the accuracy of the model. The next step is to examine *visual system presence* by observing the effects of various video recording and reproduction systems (including HD TV, 3D TV and UHD TV) on system presence.

Acknowledgment. This study was partially funded by the National Institute of Information and Communication Technology (NICT), Japan. The authors thank Mr. Shota Tsukahara at the University of Yamanashi for his assistance in constructing the model.

REFERENCES

[1] K. Hamasaki, T. Nishiguchi, R. Okumura, Y. Nakayama, and A. Ando, A 22.2 multichannel sound system for ultrahigh-definition TV, *SMPTE Motion Imaging J.*, vol. 117, no. 3, pp. 40–49, 2008.

[2] Rec. ITU-R BS.2051, Advanced sound system for programme production, ITU, pp. 1–12, 2014.

[3] T. B. Sheridan, Musings on telepresence and virtual presence, *Presence*, vol. 1, no. 1, pp. 120–126, 1992.

[4] T. B. Sheridan, Further musings on the psychophysics of presence, *Proc. of IEEE Int. Conf. on Systems, Man, and Cybernetics (SMC 1994)*, pp. 1073–1077, San Antonio, USA, 1994.

[5] B. G. Witmer and M. J. Singer, Measuring presence in virtual environments: A presence questionnaire, *Presence*, vol. 7, no. 3, pp. 225–240, 1998.

[6] T. Schubert, F. Friedmann, and H. Regenbrecht, The experience of presence: Factor analytic insights, *Presence*, vol. 10, no. 3, pp. 266–281, 2001.

[7] W. Teramoto, K. Yoshida, N. Asai, S. Hidaka, J. Gyoba, and Y. Suzuki, What is "sense of presence": A non-researcher's understanding of sense of presence, *Trans. of the Virtual Reality Soc. of Japan*, vol. 15, pp. 7–16, 2010. (in Japanese)

[8] K. Fukue, K. Ozawa, and Y. Kinoshita, Study on multidimensional structure of the sense of presence in audio-visual content, *Trans. of Japan Soc. of Kansei Engineering*, vol. 11, no. 2, pp. 183–192, 2012. (in Japanese)

[9] Y. Kinoshita, K. Fukue, and K. Ozawa, Development of *Kansei* estimation models for the sense of presence in audio-visual content, *Proc. of IEEE Int. Conf. on Systems, Man, and Cybernetics (SMC 2011)*, Anchorage, USA, pp. 3280–3285, 2011.
[10] K. Ozawa, M. Obinata, and Y. Kinoshita, *Kansei* estimation models for the sense of presence in audio-visual content with different audio reproduction methods, *Proc. of 13th ACIS Int. Conf. on Software Engineering, Artificial Intelligence Networking and Parallel/Distributed Computing (SNPD 2012)*, Kyoto, Japan, pp. 567–573, 2012.
[11] K. Ozawa, S. Tsukahara, Y. Kinoshita, and M. Morise, Development of an estimation model for instantaneous presence in audio-visual content, *IEICE Trans. on Information and Systems*, vol. E99-D, no. 1, pp. 120–127, 2016.
[12] M. Ito, M. Morise, K. Ozawa, and Y. Kinoshita, Construction of an auditory presence estimation model considering audio reproduction methods, *J. Acoustical Soc. of Japan*, vol. 72, no. 6, pp. 306–314, 2016. (in Japanese)
[13] K. Ozawa, Y. Chujo, Y. Suzuki, and T. Sone, Contents which yield high auditory-presence in sound reproduction, *Kansei Engineering International*, vol. 3, pp. 25–30, 2002.
[14] K. Ozawa and Y. Chujo, Content presence vs. system presence in audio reproduction systems, *Proc. of 2nd Int. Symp. on Universal Communication (ISUC2008)*, Osaka, Japan, pp. 50–55, 2008.
[15] K. Ozawa, S. Tsukahara, Y. Kinoshita, and M. Morise, Instantaneous evaluation of the sense of presence in audio-visual content, *IEICE Trans. on Information and Systems*, vol. E98-D, pp. 49–57, 2015.
[16] H. Møller, Fundamentals of binaural technology, *Applied Acoustics*, vol. 36, pp. 171–218, 1992.
[17] Rec. ITU-R BT.710-4, Subjective assessment methods for image quality in high-definition television, ITU, 1998.
[18] S. Kuwano and S. Namba, Continuous judgment of level-fluctuating sounds and the relationship between overall loudness and instantaneous loudness, Psychol. Res., vol. 47, no. 1, pp. 27–37, 1985.
[19] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and models*, Springer, New York, USA, 2006.
[20] P. N. Vassilakis, SRA: A web-based research tool for spectral and roughness analysis of sound signals, Proc. of the 4th Sound and Music Computing Conference, pp. 319–325, Lefkada, Greece, 2007.
[21] J. Blauert, *Spatial Hearing – The Psychophysics of Human Sound Localization (Revised Edition)*, The MIT Press, Cambridge, USA (1997).