Development of an Estimation Model for Instantaneous Presence in Audio-Visual Content*

Kenji OZAWA^{†a)}, Member, Shota TSUKAHARA^{††}, Yuichiro KINOSHITA[†], Nonmembers,
and Masanori MORISE[†], Member

SUMMARY The sense of presence is often used to evaluate the performances of audio-visual (AV) content and systems. However, a *presence meter* has yet to be realized. We consider that the sense of presence can be divided into two aspects: *system presence* and *content presence*. In this study we focused on content presence. To estimate the overall presence of a content item, we have developed estimation models for the sense of presence in audio-only and audio-visual content. In this study, the audio-visual model is expanded to estimate the instantaneous presence in an AV content item. Initially, we conducted an evaluation experiment of the presence with 40 content items to investigate the relationship between the features of the AV content and the instantaneous presence. Based on the experimental data, a neural-network-based model was developed by expanding the previous model. To express the variation in instantaneous presence, 6 audio-related features and 14 visual-related features, which are extracted from the content items in 500-ms intervals, are used as inputs for the model. The audio-related features are loudness, sharpness, roughness, dynamic range and standard deviation in sound pressure levels, and movement of sound images. The visual-related features involve hue, lightness, saturation, and movement of visual images. After constructing the model, a generalization test confirmed that the model is sufficiently accurate to estimate the instantaneous presence. Hence, the model should contribute to the development of a presence meter.

key words: sense of presence, instantaneous presence, audio-visual content, neural networks, time-series data

1. Introduction

The term “sense of presence” is used to evaluate the performance of audio-visual (AV) equipment as well as the AV content such as a movie or a sports program. To date, many advanced AV systems have been proposed to enable the perception of a higher degree of presence, including an auditory display using a head-related transfer function [1], the 22.2-multichannel sound system [2], a 3D television, and the ultra-high definition television (UHDTV) [3]. Methodologies to evaluate the sense of presence evoked from these systems also play important roles in the evolution of AV content and systems. Although the sense of presence has been measured using psychological [4]–[12] and physiolog-

ical methods [13]–[15], a *presence meter* has yet to be realized. Such a meter would be useful for creators of AV content because the sense of presence evoked by the created content could be estimated. Moreover, it would be useful for purchasers of AV equipment because the degree of the sense of presence evoked by an AV system, which could be quantitatively measured using a set of standard content items, could be indicated in the system specifications.

The literal meaning of presence is a subjective experience of being in one place or environment while physically being situated in another [4]–[6]. Accordingly, the sense of presence is determined by how accurately an AV system reproduces the physical characteristics of the original field. However, our preliminary studies on auditory presence, which is the sense of presence evoked by auditory stimuli, revealed that the perceived presence depends on the reproduced content even if one recording and reproduction system is used [8]–[10]. Thus, we have examined the sense of presence by dividing it into two aspects: *system presence* and *content presence* [10]–[12]. System presence is determined by the characteristics of the AV system, while content presence depends on the characteristics of the contents reproduced by a system. Although both types of presence must be considered to develop a presence meter, a set of standard content items must be prepared to measure system presence. Thus, this article focuses on content presence using a single recording and reproduction system as the first step.

In our previous paper [11], we developed models to estimate the overall presence of a content item in audio-only and audio-visual conditions. Then we expanded the audio-only model to exhibit the instantaneous presence of an audio content item [16]. Herein we aim to expand the audio-visual model to express the instantaneous presence of an AV content item based on the results of our previous experiment [17].

The *system presence* of an AV system can be measured as an “overall” value using a set of standard content items as stimuli. On the other hand, the *content presence* is to be measured in both “overall” and “instantaneous” values. An overall value is useful for the overall evaluation of a content item such as a commercial film or a movie. Instantaneous values are useful for information retrieval with respect to the sense of presence; for example, if a *presence meter* is installed in a video camera, automatic summarizing of a long video is capable by extracting scenes with a higher degree

Manuscript received March 30, 2015.

Manuscript revised August 10, 2015.

Manuscript publicized October 21, 2015.

[†]The authors are with the Interdisciplinary Graduate School, University of Yamanashi, Kofu-shi, 400–8511 Japan.

^{††}The author is with the Faculty of Engineering, University of Yamanashi, Kofu-shi, 400–8511 Japan.

*Part of this paper was presented at the 4th IEEE International Conference on Consumer Electronics - Berlin (ICCE-Berlin 2014) in Sept. 2014.

a) E-mail: ozawa@yamanashi.ac.jp

DOI: 10.1587/transinf.2015MUP0014

of instantaneous presence.

2. Evaluation Experiment

Since the details of experimental methods and results were described in our previous article [17], herein they are summarized briefly.

2.1 Experimental Methods

Forty content items, each lasting about 20 to 40 s, were recorded using a video camera while the sounds were binaurally recorded using a dummy head. Figure 1 shows sample still images captured from the AV content.

Each AV content item was reproduced in an office room as a stimulus in the experiments. Video stimuli were presented with a 65-inch full-HD display while auditory stimuli were reproduced through headphones.

The subject responded instantaneously to the sense of presence using the method of continuous judgment by category [18]. Namely, the subject pressed one of seven keys (1–7) on a computer keyboard, where 1 indicates “not feeling presence at all” and 7 indicates a “large feeling of presence.” Twenty undergraduate students participated in the experiment. The subjects were instructed not to press a key until he or she could evaluate the presence after the stimulus began.

2.2 Evaluation Result

Figure 1 shows four examples of the experimental results, in which solid lines represent the average scores of all subjects’ responses averaged in 500-ms intervals. Because the latency, which is the time interval between the start of stimulus presentation and the start of response, differs among the subjects, the averages were calculated after half (ten) subjects began their responses. Error bars show 95 % confidence intervals based on the t distribution, and they are calculated every 1 s.

The sense of presence significantly increases when the movement in the sound and/or visual images is large (e.g., an approaching train around 23 s in Fig. 1 (b) or a rapidly flying arrow around 15 s in Fig. 1 (c)). On the other hand, the sense of presence is steady when the scene is monotonous (Figs. 1 (a) and 1 (d)).

3. Construction of the Presence Estimation Model

3.1 Overview of Model Construction

As described in Sect. 1, we previously developed a model to estimate the overall presence of an AV content item using neural networks [11]. Herein the model is expanded to estimate the instantaneous presence.

Our previous model needs 6 audio-related features and 14 visual-related features as inputs, and these features are extracted over the entire period of a content item to estimate

overall presence. To expand the model to estimate instantaneous presence, herein these features are calculated in 500-ms intervals. These input features are processed by artificial neural networks to obtain an estimated score of the instantaneous presence.

3.2 Audio-Related Features

Our previous model [11] has the following six input features:

- (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 95- and 5-percentile sound pressure levels
- (A5) Standard deviation of the aforementioned dynamic ranges calculated in 500-ms intervals
- (A6) Movements of sound images

Because the sounds are recorded using two channels, each input feature, except for feature (A6), has two derived values. This model adopts the larger value among the right and left channels and normalizes the value between 0–1, where 0 and 1 correspond to the minimum and maximum values for all of the content items, respectively. With respect to feature (A6), a value of 0 is assigned to content items without sound-image movements, while a value of 1 is assigned to those with sound-image movements.

As mentioned above, these six features are extracted over the entire period in the previous model. Herein these features are calculated in 500-ms intervals. Feature (A4) and (A5) are calculated for the most recent 3-s interval. Feature (A6) was subjectively determined by the authors in the previous model. This expanded model automatically determines movement; the sound-pressure-level difference between the right and left channels is initially calculated in 500-ms intervals, and if the level difference changes more than 2 dB between succeeding intervals, a movement occurs.

3.3 Visual-Related Features

In addition to the audio-related features, the following 14 visual-related features are used as inputs to our audio-visual content presence model [11]:

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

(a) *Fireworks (large quantity)*. Still picture is the scene approximately 10 s after the sound began.

(b) *Passing train*. Still picture is the scene approximately 23 s after the sound began.

(c) *Scene of playing Japanese archery*. Still picture is the scene approximately 16 s after the sound began.

(d) *View of lakeside*. Still picture is the scene approximately 10 s after the sound began.

Fig. 1 Examples of the evaluation results and model outputs. (Left panel: still picture, Right panel: instantaneous presence values.) Error bars indicate 95 % confidence intervals calculated every 1 s.

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

(V13) Difference between 95- and 5-percentile numbers of moving pixels in a frame

(V14) Standard deviation of the aforementioned differences calculated in 500-ms intervals

In our current model, these 14 features are calculated in 500-ms intervals, and features (V13) and (V14) are calculated for the most recent 3-s interval.

The pixels are classified into five color categories in the HLS color system [21], and the number in each category is counted for each frame in 500-ms intervals. For each category, the average number of pixels per frame is normalized between 0–1, where 0 and 1 correspond to 0 pixels and the largest number of pixels throughout all categories and content items, respectively. These normalized values are adopted as input features. The average lightness/saturation value is given by the average value for the pixels and frames in 500-ms intervals. For the standard deviation of the lightness/saturation values, the average lightness/saturation value is calculated for each frame, and then the standard deviation of all the average values is derived for every 500-ms interval. The skewness of the lightness/saturation values is also calculated on the basis of the average lightness/saturation value for each frame in 500-ms intervals, and are normalized such that 0 and 1 correspond to the minimum and maximum values throughout all content items, respectively. The average number of pixels corresponding to the moving objects is derived by employing the frame difference method to extract the moving object, and the number of pixels corresponding to the objects is counted for each frame for every 500-ms interval.

3.4 Model Construction Based on Neural Networks

The 20 features of the 6 audio-related and the 14 visual-related ones are used as inputs to the model. The basic structure of the neural networks is the same as the previous model [11]. Namely, the expanded model consists of input, hidden, and output layers, which have 20, 15, and 1 units, respectively. A preliminary examination determined the number of hidden layer units. Training of the networks is executed by the backpropagation method. Table 1 shows the parameters for the network, where η represents a training parameter for backpropagation. WR_{ih} and WR_{ho} are the maximum values of the absolute values to randomly set default values of weights for the input layer to the middle layer and the middle layer to the output layer, respectively. (The default values of these weights are set randomly between plus and minus where the range is shown as the absolute values in Table 1.) T is the temperature coefficient used in

Table 1 Parameters of the neural network.

Parameter	Value
η	0.1
WR_{ih}	5
WR_{ho}	5
T	1
MSE_{min}	1×10^{-3}

the sigmoid function, while MSE_{min} is the minimum value of MSE between the teacher signal and the output of the system (if MSE is less than or equal to MSE_{min} , the training iteration stops).

Using the average evaluation scores obtained by the experiment, back-propagation learning trains the network to an accuracy of 1.0×10^{-3} mean square error, which is equivalent to 0.27 on the seven-point presence scale. The evaluation scores are also expressed as a range between 0–1, where 0 and 1 correspond to 1 and 7 on the seven-point presence scale, respectively.

Although the number of content items is 40, the total number of learned and tested items is 2,395 because every 500-ms interval is treated as a learned or tested sample. By the way, the performance of a neural network depends on a set of default weight values in the network learning. We prepared 50 sets of random initial-weight-values, and the set giving the best performance was adopted as a model. This learning method improved the performance of the model compared to that obtained by a trial-and-error based method [22].

4. Evaluation of the Model

4.1 Comparison between the Evaluation Results and the Model Outputs

After constructing the model, performance tests were conducted. Because the number of stimuli is limited to 40, testing the neural network models employs a leave-one-out cross-validation to increase the number of content items used to train the network. In the validation, one item selected from the 40 content items tests the generalization performance of the network trained by the remaining 39 content items. This generalization test was conducted for all 40 possible combinations of testing and training.

In Fig. 1, the dashed lines denote the model outputs, which are given as the three-point moving averages of raw outputs, while the dotted lines indicate the absolute errors between the model outputs and the average scores. Except for Fig. 1 (d), the results show that the errors are within 0.6 point of the measured presence, and the model outputs follow the changes in the evaluation scores. Figure 2 shows the average error and ± 1 standard deviation for each content item. Although the 95 % confidence intervals of the experimental results are about 0.6 for most cases as shown in Fig. 1, the average errors are less than 0.6, except for *view of lakeside* (Fig. 1 (d)). Even with this exception, the mean of the average errors for the 40 items is 0.21, indicating that

Fig. 2 Average error between the experimental results and model outputs for each of the 40 content items. Error bars indicate ± 1 standard deviations over the presentation time. Underline indicates the results shown in Fig. 1. AV content items are arranged in the ascending order in the magnitude of error. Alphabets (a) through (d) for each item indicate group names with respect to the movements of sound and visual images as follows: (a) sound: moving, visual: moving; (b) sound: moving, visual: stationary; (c) sound: stationary, visual: moving; and (d) sound: stationary, visual: stationary.

the model sufficiently estimates the instantaneous presence.

4.2 Discrepancy between the Evaluation Results and the Model Outputs

There is a possible explanation for the discrepancy between the evaluation results and the model outputs for the exceptional item, *view of lakeside*. As easily imagined from the still picture in Fig. 1 (d), the original sound field of *view of lakeside* is so quiet that the audio-only presence score is low when only audio stimulus is presented [17]. On the other hand, its visual-only presence score is rather high when only its visual stimulus is presented [11], [17]. Even though the neural network was trained with the observed data under the condition of audio-visual presentation, the visual stimulus with high presence might prevent the model to recall the low presence in audio-visual condition. This explanation is applicable for the second worst item in error, *quiet for-*

est. This suggests that more examples with low auditory-presence should be considered when preparing a set of standard content items.

4.3 Delay of Responses

Figure 1 (c) suggests that the psychological response to the sound is delayed from the physical event such as a flying arrow. Although the arrow began flying about 15 s after the stimulus began, the presence score increases slowly after that. Such a delay is also observed in another content item, *ringing mobile phone*, which includes a sudden change in its sound pressure level. As shown in Fig. 1 (c), this time delay increases the magnitude and standard deviation of the errors.

This delay is referred to as the *reaction time*; the reaction time is about 1 s with respect to loudness judgment [18]. It is also indicated, however, that the length of the reac-

(a) Scanned feature: (A1) Loudness.

(b) Scanned feature: (A6) Movements of sound images.

Fig. 3 Examples of the effects of the parameter scan for the item *passing train*. The parameter values are the values of the scanned feature. The difference between the maximum and the minimum values is calculated for every 500-ms interval.

tion time depends on various factors, such as the listening attitude of the subject and the predictability of the stimulus [18]. These factors may also influence the reaction time in judging the sense of presence. Thus, further consideration is necessary to introduce the proper delay in displaying the estimated sense of presence when constructing a presence meter.

4.4 Effect of the Movements of Images on the Magnitude of Errors

In our previous paper [17], the AV content items were classified into four groups with respect to the movement of sound and visual images relative to a listener as follows:

- (a) Sound and visual images are moving.
- (b) Sound images are moving but visual images are stationary.
- (c) Sound images are stationary but visual images are moving.
- (d) Sound and visual images are stationary.

Here *movement* does not mean a moving picture, but a change in position relative to the observer [17], and these

Fig. 4 Box plots of the distributions of the average differences calculated for each content item. The abscissa denotes the scanned feature, while the average differences were derived by the parameter scan for that feature as shown in Fig. 3. The bottom and top of the box are the first and third quartiles, and the band inside the box is the median. The lower and higher ends of the whiskers represent the lowest datum still within 1.5 IQR (interquartile range) of the lower quartile, and the highest datum still within 1.5 IQR of the upper quartile, respectively. The data not included between the whiskers are plotted as outliers with circles.

classifications were due to the authors' subjective judgments. In Fig. 2, the group in which each item is belonging is indicated with (a) to (d) and by color (a: red, b: blue, c: green, and d: black). As seen in the figure, no specific relation can be found between the magnitude of error and the item groups. This suggests that the movements of sound and visual images are not the key factor of the errors.

4.5 Relevance of the Features to Presence

We examined the relevance of the features to the estimated presence by using the parameter scan method [23]. In the method, the values of a specific input-feature are changed in 0.1 steps and the outputs of the neural network are observed. Figure 3 shows examples of the effects of the parameter scan for the item *passing train*. A set of values of the input features including a scanned one is fed into the derived model for every 500-ms interval, and the outputs from the model are arranged along the time axis using the three-point moving averages of raw outputs. When the values of the feature “(A1) Loudness” are scanned, the outputs from the model vary significantly (Fig. 3 (a)). This suggests that (A1) is a dominant feature. On the other hand, the effects of the scan for the feature “(A6) Movements of sound images” are relatively small (Fig. 3 (b)), indicating that this feature is less dominant. By the way, an interesting nonlinear change is observed in Fig. 3 (b): the largest value (1.0) of the scanned feature does not always result in the highest presence.

The dominance of the feature can be expressed by

Fig. 5 Example of outputs of the previous model when the same input features as the present model were fed. The experimental results are taken from Fig. 1.

the difference between the maximum and minimum values evoked by scanning (Fig. 3 (a)). These differences are calculated for every 500-ms interval and averaged over the time axis. The average differences are 2.00 and 0.39 for Fig. 3 (a) and (b), respectively. These average differences are calculated for every content item and feature, and their distributions are plotted in Fig. 4. Although the most and least dominant features seem to be “(A1) Loudness” and “(A6) Movements of sound images” respectively, it can be said that every feature has some relevance to the presence. The reason why feature (A6) is the least dominant might be due to its definition. As described in Sect. 3.2, the movement value is either 0 (without movement) or 1 (with movement) depending on the change in an interaural level difference (ILD). In a future research, we need to examine the dominance if the values of the ILD change is adopted as feature (A6). On the contrary, there is a possibility that the movement of a sound source is denoted in the change in loudness, and thus feature (A1) seems dominant.

4.6 Effect of the Expansion of the Model

As mentioned above, we previously developed a model to estimate the overall presence of an AV content item [11]. In the current paper this model was expanded to estimate the instantaneous presence. Here the effect of the expansion is

examined.

Because the structure of these models are common, the 20 features used in the present model can be inputted to the previous model. Figure 5 shows the outputs from the previous model for the items *fireworks (large quantity)* and *passing train* as examples. By comparing these to Figs. 1 (a) and 1(b), the discrepancies between the experimental results and model outputs are significantly larger. The average differences are 0.84 and 0.70 for Fig. 5 (a) and (b), respectively. The mean of the average errors for the 40 items is 0.71, which exceeds the criterion of 0.6 shown in Sect. 4.1. This indicates that the expansion was effective. In other words, it is important to train the neural network using instantaneous feature values for obtaining good estimation of the instantaneous presence. Moreover, we consider that the increase in the number of learned sample gave a better performance of the model: the previous overall model was learned with only 39 samples while the present model was trained with more than 2000 samples as stated in Sect. 3.5.

As described in the introduction, a *presence meter* has yet to be realized. Although we initially expected that our previous model estimating the overall presence [11] could be applied to estimate the instantaneous presence, the results in Fig. 5 show that it does not function as a presence meter. Thus, in this paper we decided to expand the previous model. Because our paper [17] indicated that the overall presence can be estimated using the instantaneous presence, the developed model in this paper can be applied to the estimation of the overall presence.

5. Conclusion and Future Study

This study focuses on the instantaneous *content presence* evaluated by the method of continuous judgment by category. Based on the experimental results, we expanded our previous model to estimate the instantaneous presence in AV content. The constructed model successfully estimates the instantaneous presence. Hence, the model should contribute to the development of a presence meter.

However, it is suggested that more stimuli are required when considering a set of standard content items. As the nature of an artificial neural network, it is expected that the more contents will result in improvement of the performance of the network. The next step is to examine system presence by observing the instantaneous presence in various recording and reproduction systems including the advanced AV systems mentioned in the introduction.

Acknowledgments

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

References

- [1] B.R. Schroder and B.S. Atal, “Computer simulation of sound trans-

- mission in rooms," *Proc. of the IEEE*, vol.51, no.3, pp.536–537, March 1963.
- [2] K. Hamasaki, T. Nishiguchi, R. Okumura, Y. Nakayama, and A. Ando, "A 22.2 multichannel sound system for ultrahigh-definition TV," *SMPTE Motion Imaging Journal*, vol.117, no.3, pp.40–49, April 2008.
- [3] Rec. ITU-R BT.2020, "Parameter values for ultra-high definition television systems for production and international programme exchange," ITU, June 2012.
- [4] T.B. Sheridan, "Musings on telepresence and virtual presence," *Presence*, vol.1, no.1, pp.120–126, Winter 1992.
- [5] 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, Oct. 1994.
- [6] B.G. Witmer and M.J. Singer, "Measuring presence in virtual environments: A presence questionnaire," *Presence*, vol.7, no.3, pp.225–240, June 1998.
- [7] T. Schubert, F. Friedmann, and H. Regensbrecht, "The experience of presence: Factor analytic insights," *Presence*, vol.10, no.3, pp.266–281, June 2001.
- [8] K. Ozawa, Y. Chujo, Y. Suzuki, and T. Sone, "Contents which yield high auditory-presence in sound reproduction," *Kansei Engineering International*, vol.3, no.4, pp.25–30, 2002.
- [9] K. Ozawa, S. Ohtake, Y. Suzuki, and T. Sone, "Effects of visual information on auditory presence," *Acoustical Science and Technology*, vol.24, no.2, pp.97–99, March 2003.
- [10] K. Ozawa and Y. Chujo, "Content presence vs. system presence in audio reproduction systems," *Proc. of 2nd Int. Symp. on Universal Communication (ISUC2008)*, pp.50–55, Osaka, Japan, Dec. 2008.
- [11] Y. Kinoshita, K. Fukue, and K. Ozawa, "Development of Kansei estimation models for the sense of presence in audio-visual content," *Proc. of 2011 IEEE Int. Conf. on Systems, Man, and Cybernetics (SMC 2011)*, pp.3280–3285, Anchorage, USA, Oct. 2011.
- [12] 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)*, pp.567–573, Kyoto, Japan, Aug. 2012.
- [13] M. Meehan, B. Insko, M. Whitton, and F.P. Brooks Jr., "Physiological measures of presence in stressful virtual environments," *ACM Trans. on Graphics, SIGGRAPH 2002*, vol.21, no.3, pp.645–652, July 2002.
- [14] M. Emoto, K. Masaoka, M. Sugawara, and F. Okano, "Viewing angle effects from wide video projection images on the human equilibrium," *Displays*, vol.26, no.1, pp.9–14, Jan. 2005.
- [15] M. Clemente, A. Rodriguez, B. Rey, and M. Alcáñiz, "Assessment of the influence of navigation control and screen size on the sense of presence in virtual reality using EEG," *Expert Systems with Applications*, vol.41, no.4, pp.1584–1592, March 2014.
- [16] K. Ozawa, S. Tsukahara, Y. Kinoshita, and M. Morise, "Development of an estimation model for instantaneous presence in audio content," *Proc. of the 4th IEEE Int. Conf. on Consumer Electronics - Berlin (ICCE-Berlin 2014)*, pp.238–242, Sept. 2014.
- [17] 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, no.1, pp.49–57, Jan. 2015.
- [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, April 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, July

2007.

- [21] T. Berk, A. Kaufman, and L. Brownston, "A human factors study of color notation systems for computer graphics," *Communications of the ACM*, vol.25, no.8, pp.547–550, Aug. 1982.
- [22] S. Tsukahara, K. Ozawa, Y. Kinoshita, and M. Morise, "On development of an estimation model for instantaneous presence in audio-visual content," *IEICE Technical Report*, EA2015-17, Aug. 2015. (in Japanese)
- [23] Y. Isu, U. Nagashima, H. Hosoya, and T. Aoyama, "Development of NEural network simulator for structure-activity CORrelation of molecules: Neco," *J. of Chemical Software*, vol.2, no.2, pp.76–94, July 1994. (in Japanese)

Kenji Ozawa received the B.E., M.E., and Ph.D. degrees in Electrical Communication from Tohoku University in 1986, 1988, and 1994, respectively. He is currently a professor of the University of Yamanashi. His research interests include psychoacoustics, signal processing of audio signals, and *Kansei* (Emotion and Sensibility) information processing.

Shota Tsukahara received the B.E. degree in Computer and Media Engineering from the University of Yamanashi in 2013. He is currently with a private company. He engaged a research on the *Kansei* evaluation and modelling of the sense of presence when he was with the University of Yamanashi.

Yuichiro Kinoshita received the B.Eng. and Ph.D. degrees in computer science from Ritsumeikan University in 2002 and 2007, respectively. He is currently an associate professor in the Department of Computer Science and Engineering, University of Yamanashi. His research interests include *Kansei* modelling, *Kansei* information processing and human-computer interaction.

Masanori Morise received the Ph.D. degree in engineering from Wakayama University in 2008. He was a JSPS Research Fellow (DC1) in 2006–2008, a postdoctoral researcher at Kwansei Gakuin University in 2008–2009, and an Assistant Professor at Ritsumeikan University in 2009–2013. He is currently a Project Assistant Professor at University of Yamanashi. His research interests include speech analysis/synthesis and speech perception.