An Empirical Method for Causal Inference of Constructs for QoE in Haptic–Audiovisual Communications

SHUJI TASAKA, Nagoya Industrial Science Research Institute

This article proposes an empirical method for inferring causal directions in multidimensional Quality of Experience (QoE) in multimedia communications, noting that causation in QoE is perceptual. As an example for modeling framework, we pick up a Bayesian structural equation model (SEM) previously built for haptic audiovisual interactive communications. The SEM includes three constructs (Audiovisual quality, Haptic quality, and User experience quality), which are latent variables each representing a group of observed variables with similar characteristics. In the SEM, the causal directions of the constructs were assumed by resorting to the domain knowledge. This article aims at proposing a methodology for inferring causal directions of constructs in general by verifying the assumption of causal directions in the SEM through their observed data alone. For that purpose, we compare six SEMs each with different causal directions of constructs, one of which is the one from the domain knowledge. The proposed method is based on QoE prediction by a Bayesian approach with Markov chain Monte Carlo (MCMC) simulation. Setting observed scores to the indicators of exogenous variables in each SEM, we predict values of all the indicators; we then assess the mean square error (MSE) between predicted QoE and mean opinion score (MOS) from observed scores and estimate the probability distribution of the MSE in each SEM. We can compare any two SEMs to find which is more plausible by examining the probability that the MSE for one SEM is smaller than or equal to that for the other. These probabilities are estimated with MCMC simulation. The method indicates that the causal directions thus inferred for the haptic audiovisual interactive communications adequately support the original ones drawn from the domain knowledge. In addition, we demonstrate that QoE can behave like the “impact-perceive-adapt” model of the effects of delayed haptic and visual feedback on performance in a collaborative environment, which Jay, Glencross, and Hubbold proposed in 2007, and that it accompanies reversal of plausible causal directions like a flip–flop.

CCS Concepts: • **Computing methodologies** → **Modeling methodologies**; • **Networks** → *Network performance modeling*; • **Human-centered computing** → *User models*;

Additional Key Words and Phrases: Causal inference, Quality of Experience (QoE), media synchronization, causation, SEM, construct, latent variables, Bayesian modeling, haptic–audiovisual interactive communications, MCMC, OpenBUGS

This work was supported by JSPS KAKENHI (Grant-In-Aid for Scientific Research of Japan Society for the Promotion of Science) grant number 20K04495. The author thanks Eiichi Isomura and Prof. Toshiro Nunome for providing the experimental data.

Author’s address: S. Tasaka, Nagoya Industrial Science Research Institute, 1-13 Yotsuya-dori, Chikusa-ku, Nagoya, 464-0819, Japan; emails: tasaka@nisri.jp, s.tasaka@nitech.jp.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2022 Association for Computing Machinery.

1551-6857/2022/01-ART17 \$15.00

<https://doi.org/10.1145/3473986>

ACM Reference format:

Shuji Tasaka. 2022. An Empirical Method for Causal Inference of Constructs for QoE in Haptic–Audiovisual Communications. *ACM Trans. Multimedia Comput. Commun. Appl.* 18, 1, Article 17 (January 2022), 24 pages. <https://doi.org/10.1145/3473986>

1 INTRODUCTION

Quality of Experience (QoE) is an essentially important measure of multimedia communications services, since it reflects the ultimate goal of the services, namely, the end-users’ perceived quality. Effective enhancement and control of QoE requires understanding its *cause-effect relationships*, i.e., *causality* [1–3].

This article supposes QoE and causality of *continuous media* such as video, audio, and haptic media. It does not include discrete computer data like those used in networked games [4] since its QoE and causality are assessed in a different manner from that for continuous media.

Causation in multimedia communications QoE is distinguished from traditional causation such as those treated in [1–3] by two features of QoE: (a) *temporal distortion of event observation*, and (b) *human perceptual nature*. The former means variation in temporal order and intervals of packets arriving at the receiving-end from those at the information source; it occurs owing to packet loss, network delay, and delay jitter produced by the Internet, which offers the great majority of contemporary multimedia communications services with best-effort. The latter is due to the fact that QoE is perceptually judged by individual end-users and therefore depends upon the individuality of human beings (i.e., observers) and their environment; this property does not exist in typical causation in natural science, most of which can be regarded as deterministic cause-effect relationships. In other words, causality in QoE is a sort of *perceptual causality* [5].

Regarding feature (a), the receiving-end (receiver) usually copes with it by exerting *media synchronization control* [6] over the **media unit (MU)**, which is defined as the transmission unit at the application layer, such as a video frame (a video MU) and a constant number of audio samples (an audio MU). Its basic principle is simple; the receiver can absorb network delay jitter completely if it stores arriving MUs in the buffer during the maximum end-to-end delay time before outputting the MUs. The output is carried out according to their generation intervals in the source by referring to timestamps attached to the MUs. This is so-called *playout buffering control*. Note that the playout buffering control produces additional end-to-end delay due to the buffering.

As a matter of fact, however, the exact end-to-end delay is variable and unknown to the receiver. Even if it were known, setting the playout buffering time to the maximum delay might lead to loss of good interactivity in communications. Consequently, the practical setting is an expected end-to-end delay not so as to lose the interactivity. With this setting, MUs arriving beyond the playout buffering time may be either dropped or output, depending on the media synchronization algorithm adopted. Therefore, even if the receiver performs media synchronization control over MUs, it cannot always recover the temporal structure of MUs perfectly; temporal distortion of MUs can still remain in addition to the increase in the end-to-end delay due to playout buffering. Thus, observational order and intervals of MUs for users at the receiver can be different from those at the source. Researches on traditional causality [1–3] usually do not suppose such a situation.

Feature (b) implies that causal directions in QoE must be formulated in a probabilistic manner. When we compare two models of causal directions in QoE, for example, we should derive the probability that one model is more plausible than the other.

We have thus learned that study on causal directions in QoE has to take into consideration the two features. However, we can find no study of this kind in the QoE literature. This article presents

a trial of establishing a methodology for modeling causal directions in QoE, especially focusing on causal directions of *constructs* in QoE.

Multimedia communications QoE has a variety of facets; therefore, it is usually represented by a multidimensional vector with components such as video temporal quality, video spatial quality, audio quality, and user satisfaction. For QoE assessment, we collect subjective scores (typically, on a five-point scale) from users for each component variable, which is referred to as the *observed variable* or *indicator* since it is directly observed by users. In the case of multidimensional QoE, the number of possible cause-effect combinations of observed variables (e.g., video temporal quality and user satisfaction) is so many, which necessitates much effort of the calculation, and more importantly, it is difficult to get a good grasp of the whole causal structure.

In order to reduce the dimension, statistics often introduces a *hypothetical* random variable aggregating two or more observed variables with similar characteristics in the form of *construct*, which is also called a *factor* in *factor analysis* [7]. Since the construct (factor) is unobserved, it is a *latent variable* [7, 8].

A useful statistical method for inferring causation is **structural equation models (SEMs)** [9], which has been developed in non-engineering areas such as social sciences and biometrics; only few publications of SEM applied to multimedia communications can be found in literature [10–13].

The SEM method can explore cause-effect relationships between random variables by formulating the *data generating process*. A SEM with both constructs and observed variables is referred to as the *LISREL type* [9, 14]. A model of LISREL type consists of two submodels: the *structural model* and the *measurement model*. The structural model describes cause-effect relationships between constructs as a *regression* model where a response variable is regarded as the effect, and the explanatory variables are the causes. The measurement model relates each construct to the corresponding observed variables (indicators).

Since a construct is a hypothetical variable, *a construct does not possess a natural metric*: its *location* (origin), *scale*, *orientation* are indeterminate [15]; they must be resolved in model formulation. This problem is referred to as *identifiability* [16] or *indeterminacies* [15] in Bayesian analysis, and *identification* [9] in “frequentist” (classical) approach including **maximum likelihood (ML)** or **least squares** estimation.

Structural models in SEM are often built on the basis of *domain knowledge (expert knowledge)* of the target for modeling. However, it is desirable that the regression model thus obtained from the domain knowledge can be verified mainly by observed data. Many methodologies are available for inference of cause-effect relationships from observed data [1, 2]. The majority of the methodologies aim at discovery of cause-effect relationships between observed variables themselves, not including ones between constructs (latent variables).

This article proposes a method for inferring cause-effect relationships between constructs in a SEM of LISREL type by utilizing observed data of indicators for constructs. It takes a *Bayesian modeling* approach [15–17] with *Markov chain Monte Carlo (MCMC)* simulation [18, 19]. As a material of the methodology development, we employ the same collected scores in a haptic audiovisual interactive communication system [20] as those that were used to build the SEM in [13], where domain knowledge is utilized to formulate the structural model. This article builds five different structural models from the one in [13] while the totally six SEMs have the same measurement model; the new five SEMs are compared with the SEM in [13] to demonstrate the plausibility of the latter.

The remainder of this article is organized as follows. Section 2 presents related work to this article, focusing on media synchronization and causal inference. Section 3 briefly reviews the experimental system and QoE measures for haptic audiovisual communications which is used in [13], and it further describes two playout buffering control schemes one of which is not treated in [13].

Section 4 presents the six SEMs for comparison and shows that they are not distinguishable from each other only by the method used in [13]. In Section 5, we propose a new method for inferring causal directions of constructs. Section 6 demonstrates MCMC simulation results of the six SEMs and discusses the effect of playout buffering time on QoE. Section 7 concludes the article.

2 RELATED WORK

In this article, we investigate QoE by taking into consideration media synchronization and try to infer causal directions of constructs. We briefly describe previous researches on these two topics.

2.1 Media Synchronization

Media synchronization has a long history of study and a variety of proposals for control schemes [6, 21, 22]. In this article, however, we treat only two simple control schemes for continuous media (i.e., MAB and BUF introduced in Section 3), and comparison of many schemes from a QoE point of view is beyond the scope of the current study. Therefore, we mention only research results closely related to later discussion. Huang, Nahrstedt, and Steinmetz give an extensive up-to-date survey of media synchronization in [22], to which the reader is referred for details.

We focus here on studies of human perception of skew between two or more media streams; lip sync [23] (namely, between audio and video) is a classical example. Media synchronization between haptic media and video/audio (e.g., [24–27]) is of particular interest to the current study.

In [24], Jay, Glencross, and Hubbard report an intriguing finding of user behavior for delayed haptic and visual feedback in a collaborative virtual environment. They proposed the “impact-perceive-adapt” model of user behavior. In the experiment, a pair of users did a collaborative target acquisition task in a sort of 3D virtual space by delaying haptic and visual updates from the other user. In the virtual space, each user moved his/her own object, staying in touch with the surface of the other user’s object, to reach a specified target. Experimental runs were repeated at ten levels of additional end-to-end latency between 0 and 400 ms without any media synchronization control. Jay et al. collected users’ subjective score on a ten-point scale (the smaller the better) for task difficulty and disruption of haptic and visual feedback, in addition to recording the object movement time and rate of object positioning error (not touching with each other) over the 10 levels of latency. Examining the data thus obtained, they found that increasing latency gives detrimental effects on user performance, which changes at three latency thresholds: (1) Until the *impact threshold* (25 ms), errors increase rapidly, whereas users are unaware of the latency, (2) At the *perception threshold* (50 ms), users begin to perceive detrimental effects of latency and start to slow down the movement, though the decrease in speed is not enough to stop the increase in error rates, (3) At the *adaptation threshold* (100 ms) and above, users slow their movements in direct proportion to latency to suppress the rise in error rates.

We will observe QoE’s behavior similar to the impact-perceive-adapt one later in this article, which will be explained by change in causal direction.

2.2 Causal Inference

When we want to examine how a random variable X is related to another random variable Y , we often calculate the correlation coefficient between X and Y . Since its calculation is easy, it is widely used as a statistical measure. However, *correlation is not necessarily causation* as is well-known [1].

Correlation between X and Y appears because of not only simple causation but also many other reasons: a simple causal effect of X on Y ($X \rightarrow Y$), its reverse causal effect ($X \leftarrow Y$) and a common cause C of both X and Y ($X \leftarrow C \rightarrow Y$) and others including the combinations of the previous three cases. In particular, the causal directions $X \rightarrow Y$ and $X \leftarrow Y$ are not distinguishable only by the

value of the correlation coefficient. Consequently, for effective QoE control, we have to distinguish between causation and correlation.

In [13], Tasaka compares three statistical models for a haptic audiovisual interactive communication system in order to explore the issue of causation and correlation: a SEM with three constructs (3C-SEM) two of which are correlated and give causal effects on the other one ($X \rightarrow Z \leftarrow Y$, where X and Y are correlated), a *confirmatory factor analysis model* (3C-CFA) in which the three constructs are only correlated with each other, and another confirmatory factor analysis model (1C-CFA) in which the three constructs degenerate into a single construct. He demonstrated that the 3C-SEM is most plausible among the three models. However, the causal directions of constructs in 3C-SEM were drawn from domain knowledge, and the correspondence of each construct to observed variables (i.e., the measurement model) was specified according to prior information, which is a kind of expert knowledge. Therefore, we can say that the 3C-SEM has been built heuristically.

We can find more systematic and theoretical methodologies for inferring causal directions between random variables in literature [1–3, 28]. They aim at discovering causal directions by using observed data alone. Many of them employ SEMs as the base.

The most well-known methodology of this type is a graphical one by means of *directed acyclic graphs* (DAGs) which utilize *conditional independence* and the *v-structure* of nodes (i.e., random variables) with *no assumption* on probability distributions [1, 2]. The conditional independence is exemplified by X and Y given Z in $X \rightarrow Z \rightarrow Y$, X and Y given Z in $X \leftarrow Z \leftarrow Y$, X and Y given Z in $X \leftarrow Z \rightarrow Y$, whereas the *v-structure* is such that X and Y are independent in $X \rightarrow Z \leftarrow Y$. The *v-structure* can also be interpreted as a kind of conditional independence since conditioning on Z produces mutual dependence of X and Y , though it is the opposite change of conditional independence.

A SEM can be represented by a DAG, which is referred to as a *causal graph*. Graphs produced by algorithms utilizing only the conditional independence, which includes the *v-structure*, such as the PC algorithm, **Fast Causal Inference (FCI)**, and **Greedy Equivalence Search (GES)** often belong to the same *Markov equivalence class*, where its members of the graphs have the *same conditional independence* and *v-structures*, and edges whose directions cannot be determined are left. This implies, for example, that causal directions $X \rightarrow Y$ and $X \leftarrow Y$ are indistinguishable.

Imposing some restrictions on probability distributions of nodes seems to improve the identifiability of causal graphs.

A SEM usually assumes linear equations and *Gaussian (normal)* distributions for random variables including observed, latent, and error variables [9, 17]. This is also the case with structural models (construct models) used later in this article (see Figure 3), though the whole SEM is non-linear because of categorization of score in the measurement model as shown in Figure 2. Under the linear–Gaussian condition, however, causal graphs can be provided as a Markov equivalence class which has *the same set of probability distribution* for the graphs, and therefore the graphs belonging to the class are not identifiable [28, Chapter 7]; especially, $X \rightarrow Y$ and $X \leftarrow Y$ may be indistinguishable by observed data alone.

In order to obtain identifiable models only with observed data, several methods utilizing *non-Gaussian (non-normal)* distributions have been proposed; Wiedermann and von Eye have edited recent articles related to this issue [29]. Among them, *LiNGAM* [30] is a powerful approach assuming non-Gaussian independent error terms in a SEM.

Even in the case of Gaussian SEMs with linear functions, we can find several studies that can identify a unique causal graph by putting some restrictions on the joint probability distribution of the nodes. For example, Peters and Bühlmann [31] proved that linear Gaussian SEMs with equal error variances can recover a unique causal graph from the joint Gaussian distribution.

Fig. 1. Experimental system configuration.

Note that all the methodologies mentioned so far suppose observed variables without latent variables. Although we can find algorithms for causal discovery of latent variable models (e.g., [32]), they are not suitable to the purpose of this article.

The SEM in Figure 2, which is our target model, is a Bayesian SEM including (latent) constructs and categorical indicators. To the best of the author's knowledge, no theoretical method for providing a unique causal graph in such a case (namely, covering "Bayesian", "constructs", and "categorical" together) is not yet available in literature. Thus, this article resorts to an *empirical method utilizing the controlled experimental condition* of score collection, which will be proposed in Section 5.

3 EXPERIMENTAL SYSTEM AND QOE MEASURES

This section provides a minimal amount of information about previous studies of the experiment on score collection for Bayesian modeling [11–13] so that the article can be self-contained at a minimum level. See [11, 13] for more details.

3.1 Experimental System

The baseline SEM we utilize in this article has been built in [13] with experimental data of *five-point* scores for haptic audiovisual *interactive* communications in *real* space (not in virtual space) [20]. Figure 1 illustrates the experimental system configuration.

Terminals 1 and 2 have the same structure; each terminal is equipped with a haptic device (PHANTOM Omni; hereafter, *PHANTOM* for simplicity), a Web camera, an LCD monitor and a headset, and has an identical workspace, where a subject carries out the task. The workspace, PHANTOM's stylus and subject's forearm are put in the visual field of the Web camera. The two terminals are located in different rooms. Each terminal transmits/receives haptic media, audio and video to/from the other terminal as *three separate UDP streams*.

Poisson load traffic consisting of UDP datagrams each with a payload of 1,472 bytes at an average bit rate of 6.0 Mb/s is transmitted between Load Senders and Receivers as interference with audio-video and haptic streams, which are impaired, thus yielding packet loss, delay and

delay jitter mainly owing to packet buffering and overflow at the two routers (both Cisco 2811) [20].

A pair of *subjects* conducts an interactive task. We have prepared two kinds of task, *object movement* and *castanets hitting*; see the left and right panels, respectively, in Figure 1 and further refer to [13] for more details. In the experiment, the receiver exerts playout buffering control for intra-stream media synchronization, which we call *Msync* here. No inter-stream synchronization control is performed. Two kinds of *Msync* have been installed into the receiver: the **Media Adaptive Buffering (MAB)** and an ordinary playout buffering scheme, which is referred to as **Buffering (BUF)**. The BUF is not treated in [13]. Note that the same kind of *Msync* is selected at both terminals in each experimental run for simplicity of the experiment. Thus, from a QoE-assessment point of view, we have four pairs of (task, *Msync*), which for convenience, we denote hereafter as (*object*, MAB), (*castanets*, MAB), (*object*, BUF), and (*castanets*, BUF).

3.1.1 Tasks. In doing a task, one subject is assigned to the role of the *instructor*, and the other is the *manipulator*. The manipulator receives an audio instruction over the network from the instructor. According to the instruction, the manipulator *physically manipulates* an object or a pair of castanets on the instructor's side over the network by using two PHANToM's. Note that a force feedback loop between the two PHANToM's allows the user to manipulate the stylus of the other side remotely by his/her own stylus [33, 34].

During an experiment run (i.e., 30 s), the manipulator watches the instructor's workspace displayed on his/her LCD monitor; the instructor holds his/her own stylus in a hand and surrenders him/herself to the manipulator's movement not so as to impede it, just as if he/she were a marionette remotely operated by the manipulator. In other words, the manipulator can regard the two PHANToM's in this experiment as a single virtual PHANToM connected via the network and the instructor, which serves as a sort of mechanical actuator, and therefore as a single virtual arm.

The two subjects alternate the roles in each experimental run.

In the task of object movement, the manipulator moves an object (a circle, a triangle, or a square) in the center circle to a destination (a circle, a triangle, or a square) at the corners of the instructor's workspace (e.g., move the circle object to the square destination). As an objective measure of efficiency, the number of objects whose movement to the destination is completed during an experimental run (30 s) is counted.

In the task of castanets hitting, the instructor tells the number of hitting which he/she randomly chooses among 1 through 3 to the manipulator as the instruction. The manipulator hits both his/her castanets and the instructor's one the number of times specified by the instruction. As shown in Figure 1(b), the manipulator swings his/her own stylus down to his/her castanets; this action is, in a sense, copied at the instructor's side to hit the castanets by virtue of the force feedback (i.e., the single virtual arm). Each subject can listen to not only his/her own sound of hits but also the other's one, since the sound is transmitted to each other. Therefore, the two subjects can notice asynchrony between the two kinds of sound if any (Castanet Sync). The number of instructions given by the instructor in an experimental run is employed as an objective measure of efficiency.

Note that the two tasks have the opposite characteristics of the response speed and necessary precision of PHANToM operation. The object movement requires the precise operation with video monitoring, while response is not necessarily fast. The castanets hitting needs fast response, but the PHANToM operation does not have to be so precise as long as the castanets are hit.

3.1.2 Msync. The MAB sets the playout buffering time at the receiver for the haptic MU to a value smaller than the value of audio and video MUs. A haptic MU is current positional information of 320 bits of the stylus with a rate of 1,000 MU/s. An audio MU is 320 audio samples, and a video MU is a video frame. A video frame is divided into 15 slices, each of which forms an IP packet.

Table 1. QoE Measures, Score Criteria, Score Variables, and Constructs Along with Latent Variables

Subjective QoE measure for observed variable	Score criteria (5-point scale) 1 (leftmost), 2, 3 (neutral), 4, 5 (rightmost)	Castanets hitting task	Object movement task	Construct (construct variable)
		score (latent variable)	score (latent variable)	
Video Temporal quality	Video is jerky $\leftrightarrow$ smooth	$z_1(v_t = z_1^*)$	$z_1(v_t = z_1^*)$	AVQ (f_1)
Video Spatial quality	Video is corrupt $\leftrightarrow$ clear	$z_2(v_s = z_2^*)$	$z_2(v_s = z_2^*)$	
Video Usefulness	Image is hard to grasp $\leftrightarrow$ easy	$z_3(v_u = z_3^*)$	$z_3(v_u = z_3^*)$	
Inter-stream sync between video and haptic	Out of sync $\leftrightarrow$ In sync	$z_4(v_h = z_4^*)$	$z_4(v_h = z_4^*)$	
Audio Naturalness	Audio is artificial $\leftrightarrow$ natural	$z_5(a_n = z_5^*)$	$z_5(a_n = z_5^*)$	
Castanet Sync	Out of sync $\leftrightarrow$ In sync	$z_6(c_s = z_6^*)$	Not applicable	
Haptic Smoothness	Movement is awkward $\leftrightarrow$ smooth	$z_7(h_m = z_7^*)$	$z_6(h_m = z_6^*)$	HQ (f_2)
Haptic Operability	Manipulation is heavy $\leftrightarrow$ light	$z_8(h_o = z_8^*)$	$z_7(h_o = z_7^*)$	
Haptic Stability	Manipulation is unstable $\leftrightarrow$ stable	$z_9(h_s = z_9^*)$	$z_8(h_s = z_8^*)$	
Response	Response is slow $\leftrightarrow$ quick	$z_{10}(r = z_{10}^*)$	$z_9(r = z_9^*)$	UXQ (f_3)
Communicability	Hard to communicate $\leftrightarrow$ Easy	$z_{11}(c_m = z_{11}^*)$	$z_{10}(c_m = z_{10}^*)$	
System comfort	Uncomfortable $\leftrightarrow$ Comfortable	$z_{12}(c_f = z_{12}^*)$	$z_{11}(c_f = z_{11}^*)$	
Work difficulty	Difficult to work $\leftrightarrow$ Easy	$z_{13}(w_d = z_{13}^*)$	$z_{12}(w_d = z_{12}^*)$	
Overall satisfaction	Unsatisfied $\leftrightarrow$ Satisfied	$z_{14}(s_o = z_{14}^*)$	$z_{13}(s_o = z_{13}^*)$	
Objective measure	Number of count	$z_{15}(n_c = z_{15}^*)$	$z_{14}(n_c = z_{14}^*)$	

The point of MAB is to utilize lower sensitivity of audiovisual streams to delay than that of haptic media by allowing asynchrony between the former streams and the latter to a degree with which the asynchrony is not recognizable or at least not annoying for the user [11, 20].

In this experiment, the buffering time of the haptic media (B_h) for MAB is fixed to 10 ms (i.e., $B_h = 10$ ms), while the audio and video streams adopt the same buffering time B_{av} , which is set to either 20, 40, 60, 100, or 150 ms for the task of object movement, while the castanets hitting further employs 300 ms in addition to the five values.

Because of $B_{av} > B_h$, the MAB yields two kinds of temporal distortions: One is inter-stream skew of output time between audio-video streams and haptic stream, and the other is the haptic rendering delay due to the end-to-end delay, which usually produces a larger displacement between the two styli; the reaction force is proportional to it [33] and therefore becomes stronger.

The BUF sets the same buffering time for the three media in both tasks; i.e., $B_{av} = B_h = 20, 40, 60, 100$, and 150 ms. Consequently, the BUF does not have the first temporal distortion (between received audio-video streams and haptic stream) but has the haptic rendering delay due to the end-to-end delay, which leads to a stronger reaction force than that of MAB because of $B_h = B_{av} \geq 20$.

In both MAB and BUF, MUs arriving beyond the playout buffering time are immediately output if they are in order and dropped otherwise.

Thus, we can see that the two kinds of playout buffering control schemes (MAB and BUF) enables us to examine the effects of the end-to-end delay and temporal skew between audio-video and haptic media on QoE. From a subjective assessment point of view, we have one *stimulus* corresponding to one value of B_{av} . Therefore, we have five stimuli for the task—Msync pair of (object, MAB), (object, BUF), and (castanets, BUF), while having six stimuli for the pair of (castanets, MAB).

3.2 QoE Measures

For QoE assessment, we observe 14 subjective measures (scores on a five-point scale) for the castanets hitting and 13 ones for the object movement, while a single objective measure of efficiency is employed for each of the two tasks as already defined in Section 3.1.1.

Table 1 lists the subjective QoE measures, the corresponding score criteria, the *score (discrete) variables* z_j ($j = 1, \dots, 13$ or 14) along with *score (continuous) latent variables* z_j^* ($j = 1, \dots, 13$ or 14) in parentheses, and the three constructs with construct latent variables in addition to the objective measure of efficiency in the last row. Score latent variables with similar characteristics are classified into one of three *constructs*, whose latent variables are denoted by f_1, f_2 , and f_3 as displayed in the last column; the corresponding constructs are named **Audiovisual quality (AVQ)**, **Haptic quality (HQ)**, and **User experience quality (UXQ)**, respectively.

The score variables z_j ($j = 1, \dots, 14$ or 15) are discrete observed variables (*ordinal categorical scores*: 1, 2, 3, 4, 5), while z_j^* ($j = 1, \dots, 14$ or 15) are continuous latent variable underlying z_j [13, 35]. We have introduced z_j^* since the same score $z_j = k$ (say) does not necessarily mean the identical sensation for all subjects but the sensation can vary slightly among individual subjects; this is also the case with stimuli. The subscript j ($j \geq 6$) of z_j and z_j^* for the object movement is smaller by 1 than that for the castanets hitting because the former task lacks the QoE measure of Castanet Sync.

The QoE for z_j is defined as its expectation: $qoe[j] = E[z_j]$.

Note that the objective measure of efficiency z_{15} or z_{14} is defined as the number of count between 1 and 5; for simplicity of modeling, the original count n_c which exceeds 5 is set to 5 [13]. Therefore, n_c can be regarded as a sort of latent variable.

The subjects who participated in this experiment are 42 Japanese people (16 male and 26 female); their ages ranged from teens to fifties. We made 21 pairs of subjects with the same gender for the interactive test.¹ According to the score criteria specified in the second column of the top row in Table 1, each subject evaluated the five or six stimuli given in random order during an experimental run with a pair of (task, Msync) randomly chosen out of the four. The network traffic condition was controlled over the experimental system to remain the same in every experimental run.

4 STRUCTURAL EQUATION MODELS

In order to infer causal directions of the three constructs (AVQ, HQ, and UXQ), we utilize the SEM in [13], which is called *3C-SEM*, as the baseline and build five more SEMs with different structural models, keeping the same measurement model. In our study on causality, we also refer to the structural model as the *construct model* [13] to emphasize that our purpose is to explore the causal relationship among constructs.

4.1 A Generic Form of SEM with a Specific Measurement Model

Figure 2 illustrates a generic form of the SEM for our study, specifying the measurement model in the case of castanets hitting on the left and right-hand sides. The middle part represents a generic form of the construct model where the three constructs (AVQ, HQ, and UXQ) are supposed to be connected through arrows (causal effects) or double headed curved arrows (correlation); specific forms of the connection will be shown in the next subsection.

The parameters λ_{jn} ($j = 1, \dots, 15; n = 1, 2, 3$) in Figure 2 denote the *loading coefficient* of $E[z_j^*]$ on the construct f_n ; λ_{jn} are unknown parameters to be estimated. See [13] for details of the measurement model and its equations.

4.2 Construct Models for Inferring Causal Directions

The purpose of this article is to explore causal directions of the three constructs (AVQ, HQ, and UXQ). There exist many possible causal directions for them. Although listing them up exhaustively seems to be ideal, the way is not practical because it incurs very heavy computational burden, and

¹The primary purpose of pairing the same gender was to collect data that can reflect some gender effect. For simplicity of modeling, however, this article builds models that are limited with respect to exploration of detailed gender effects.

Fig. 2. A generic form of SEM with the measurement model of 3C-SEM for castanets hitting.

many of the directions are considered implausible from a domain-knowledge point of view. Instead, we focus on the directions drawn from domain knowledge [13] (i.e., those of 3C-SEM), its variants and their reversed directions, since we are interested in indistinguishability between two opposite directions (say $X \rightarrow Y$ and $X \leftarrow Y$).

In [13], we have selected causal directions of constructs in 3C-SEM by considering that components affect the whole and that *cross modality* between audio-video and haptic can be expected. This implies that both AVQ (f_1) and HQ (f_2) affect UXQ (f_3) (i.e., $f_1 \rightarrow f_3$ and $f_2 \rightarrow f_3$) and that f_1 and f_2 are correlated. That is, f_1 and f_2 are *exogenous variables*, while f_3 is the *endogenous variable*. From a regression point of view, exogenous variables are explanatory variables (causes), and the endogenous one is the response variable (effect).

The construct model for 3C-SEM is depicted in the top left panel of Figure 3, where newly built five models are also displayed. The parameters γ_1, γ_2 , and γ_{nm} ($n, m = 1, 2, 3$) are *regression coefficients*, and ρ is the *correlation coefficient* between f_1 and f_2 ; these parameters are unknown to be estimated. Also, e_1, e_2 , and e_3 are residual error terms of f_1, f_2 , and f_3 , respectively, when they are endogenous.

For building the models in Figure 3, we have assumed that *each exogenous variable has an arrow headed to every endogenous variable* and added necessary arrows to complete the DAG. Under this assumption, the meaning of each model is self-explanatory from the DAG.

For convenience of reference, the name of each model is also given in the box above the DAG; the head part starting with “Ex.” specifies the exogenous variable(s), which indicates arrow’s tail(s), and the following parenthesized part is additional information to complete the DAG. The 3C-SEM, for instance, is denoted as Ex. $f_1 f_2 (f_1 \leftrightarrow f_2)$, where $f_1 \leftrightarrow f_2$ means that f_1 and f_2 are correlated.

We have selected Ex. f_3 since it has causal directions reversed from those of 3C-SEM; recall that the reversal is our great concern in this study. Ex. f_3 can be interpreted as a causal situation where the subject first gets a whole impression and then analyzes it into components in contrast to the synthetic nature of 3C-SEM. In Ex. f_3 , however, the double headed arrow between f_1 and f_2 has been removed, since correlation is supposed only between exogenous variables in SEMs. It should be noted that correlation between two random variables implies the existence of some hidden

Fig. 3. Construct models for inference of causal directions: orange ovals represent exogenous variables.

common factors outside the model affecting the two variables (i.e., *confounders*), whereas causation is not related to outside factors but is connected only to the two variables. In order to examine whether f_1 and f_2 have causal relations, we have added $\text{Ex}.f_3(f_1 \rightarrow f_2)$ and $\text{Ex}.f_3(f_2 \rightarrow f_1)$.

Similarly, we have selected $\text{Ex}.f_1(f_2 \rightarrow f_3)$ and $\text{Ex}.f_2(f_1 \rightarrow f_3)$; comparing them with $\text{Ex}.f_1f_2(f_1 \leftrightarrow f_2)$ (namely, 3C-SEM), we can examine whether causation between f_1 and f_2 (either $f_1 \rightarrow f_2$ or $f_1 \leftarrow f_2$) is more plausible than correlation $f_1 \leftrightarrow f_2$.

4.3 Structural Equations for Construct Models

We now derive the structural equation for each construct model in Figure 3. Note that the measurement equations for the SEM with a construct model in Figure 3 are the same as those of 3C-SEM, which are given in [13].

In Bayesian modeling, we can reflect the individuality of subjects in random variables, which enhances the model's capability of expressing variations in subjects. For that purpose, we add a subscript i , which represents subject i ($i = 1, \dots, N$; $N = 42$), to the latent and observed variables; this leads to notations $f_{i1}, f_{i2}, f_{i3}, e_{i1}, e_{i2}, e_{i3}, z_{ij}^*, z_{ij}$ ($j = 1, \dots, J$; $J = 15$ for the castanets hitting, and $J = 14$ for the object movement). We assume that the loading coefficient λ_{jn} and regression coefficients γ_1, γ_2 , and γ_{nm} do not depend on i . Also, when the intended meaning is the same for all subjects as in the name of a construct model (e.g., $\text{Ex}.f_1(f_2 \rightarrow f_3)$), we omit the subscript i .

In this study, we assume that the random variables e_{i1}, e_{i2}, e_{i3} follow normal distributions with zero means and variances σ_1^2, σ_2^2 , and σ_3^2 , respectively, which are independent of i . Furthermore, unknown parameters $\lambda_{jn}, \gamma_1, \gamma_2$, and γ_{nm} are assumed to be normally distributed in the same way as those in [13].

From now on in this article, we also use the names of the construct models given in Figure 3 to indicate the whole SEM, which consists of the construct model and measurement model, unless the usage is misleading.

Letting $N(\mu, \sigma^2)$ denote the normal distribution with mean μ and variance σ^2 , we will show the structural equations of the six models in turn.

4.3.1 3C-SEM. The structural equation of 3C-SEM has been driven in [13] as follows:

$$f_{i3}|f_{i1}, f_{i2} \sim N(v_i, \sigma_3^2); \quad v_i = \gamma_1 \cdot f_{i1} + \gamma_2 \cdot f_{i2}, \quad E[f_{i1}e_{i3}] = E[f_{i2}e_{i3}] = 0, \quad (1)$$

$$f_{i1} \sim N(0, \sigma_1^2), \quad f_{i2}|f_{i1} \sim N((\sigma_2/\sigma_1)\rho f_{i1}, (1-\rho^2)\sigma_2^2) \quad i = 1, 2, \dots, N, \quad (2)$$

where the joint distribution of f_{i1} and f_{i2} is assumed to be a bivariate normal distribution with zero marginal means, marginal variances σ_1^2 and σ_2^2 , respectively, and correlation $\rho = \text{Cov}[f_{i1}, f_{i2}]/(\sigma_1\sigma_2)$.

The *standardized coefficient* of the regression coefficient γ_n ($n = 1, 2$), which is the regression coefficient of a *standardized variable*, is calculated as $\gamma_{s,n} = \gamma_n\sigma_n/\sqrt{\text{Var}[f_{i3}]}$ ($n = 1, 2$) according to [9]. We can obtain the unconditional variance of the endogenous variable (i.e., $\text{Var}[f_{i3}]$) by means of the *path tracing rule* for SEM [36, pp. 127–134] as well as ordinary probabilistic calculation as

$$\gamma_{s,n} = \gamma_n\sigma_n/\sqrt{\gamma_1^2\sigma_1^2 + \gamma_2^2\sigma_2^2 + 2\gamma_1\gamma_2\sigma_1\sigma_2\rho + \sigma_3^2}; \quad n = 1, 2. \quad (3)$$

Let $C_e^{(nm)}$ ($n \neq m; n, m = 1, 2, 3$) denote the causal effect of f_{im} on f_{in} . In 3C-SEM, we have two causal effects: $C_e^{(3n)} = \gamma_{s,n}; n = 1, 2$.

The correlation coefficients between f_{in} and f_{im} ($n < m$), which are denoted by ρ_{nm} ($n = 1, 2; m = 2, 3$), are expressed as

$$\rho_{12} = \rho, \quad \rho_{13} = \gamma_{s,1} + \rho\gamma_{s,2}, \quad \rho_{23} = \gamma_{s,2} + \rho\gamma_{s,1}. \quad (4)$$

4.3.2 *Ex.f₃*. The variable f_{i3} is exogenous; we assume $f_{i3} \sim N(0, \sigma_3^2)$. Then, we get

$$f_{i1}|f_{i3} \sim N(\gamma_{13}f_{i3}, \sigma_1^2), \quad f_{i2}|f_{i3} \sim N(\gamma_{23}f_{i3}, \sigma_2^2); \quad E[f_{i3}e_{i1}] = E[f_{i3}e_{i2}] = 0, \quad (5)$$

$$\gamma_{s,n3} = \gamma_{n3}\sigma_3/\sqrt{\gamma_{n3}^2\sigma_3^2 + \sigma_n^2}; \quad n = 1, 2, \quad (6)$$

The causal effect $C_e^{(n3)}$ of f_{i3} on f_{in} ($n = 1, 2$) is given by $\gamma_{s,n3}$. The correlation coefficients are

$$\rho_{13} = \gamma_{s,13}, \quad \rho_{23} = \gamma_{s,23}, \quad \rho_{12} = \gamma_{s,13}\gamma_{s,23}, \quad (7)$$

where ρ_{12} is pseudo correlation.

4.3.3 *Ex.f₁(f₂ → f₃)*. The variable f_{i1} is exogenous, and we assume $f_{i1} \sim N(0, \sigma_1^2)$.

$$f_{i2}|f_{i1} \sim N(\gamma_{21}f_{i1}, \sigma_2^2), \quad f_{i3}|f_{i1}, f_{i2} \sim N(\gamma_{31}f_{i1} + \gamma_{32}f_{i2}, \sigma_3^2); \quad E[f_{i1}e_{i2}] = E[f_{i1}e_{i3}] = E[f_{i2}e_{i3}] = 0, \quad (8)$$

$$\gamma_{s,21} = \gamma_{21}\sigma_1/\sqrt{\gamma_{21}^2\sigma_1^2 + \sigma_2^2}, \quad (9)$$

$$\gamma_{s,31} = \gamma_{31}\sigma_1/\sqrt{\gamma_{32}^2\sigma_2^2 + (\gamma_{31}^2 + \gamma_{32}^2\gamma_{21}^2 + 2\gamma_{31}\gamma_{32}\gamma_{21})\sigma_1^2 + \sigma_3^2}, \quad (10)$$

$$\gamma_{s,32} = \gamma_{32}\sqrt{\gamma_{21}^2\sigma_1^2 + \sigma_2^2}/\sqrt{\gamma_{32}^2\sigma_2^2 + (\gamma_{31}^2 + \gamma_{32}^2\gamma_{21}^2 + 2\gamma_{31}\gamma_{32}\gamma_{21})\sigma_1^2 + \sigma_3^2}, \quad (11)$$

$$C_e^{(31)} = \gamma_{s,31} + \gamma_{s,32}\gamma_{s,21}, \quad C_e^{(21)} = \gamma_{s,21}, \quad C_e^{(32)} = \gamma_{s,32}, \quad (12)$$

$$\rho_{13} = \gamma_{s,31} + \gamma_{s,32}\gamma_{s,21}, \quad \rho_{12} = \gamma_{s,21}, \quad \rho_{23} = \gamma_{s,32} + \gamma_{s,31}\gamma_{s,21}, \quad (13)$$

where the second term on the right-hand side of ρ_{23} represents pseudo correlation.

4.3.4 *Ex.* $f_2(f_1 \rightarrow f_3)$. The variable f_{i2} is exogenous; we assume $f_{i2} \sim N(0, \sigma_2^2)$.

$$f_{i1}|f_{i2} \sim N(\gamma_{12}f_{i2}, \sigma_1^2), \quad f_{i3}|f_{i1}, f_{i2} \sim N(\gamma_{31}f_{i1} + \gamma_{32}f_{i2}, \sigma_3^2); \quad E[f_{i2}e_{i1}] = E[f_{i2}e_{i3}] = E[f_{i1}e_{i3}] = 0, \quad (14)$$

$$\gamma_{s.12} = \gamma_{12}\sigma_2 / \sqrt{\gamma_{12}^2\sigma_2^2 + \sigma_1^2}, \quad (15)$$

$$\gamma_{s.31} = \gamma_{31}\sqrt{\gamma_{12}^2\sigma_2^2 + \sigma_1^2} / \sqrt{\gamma_{31}^2\sigma_1^2 + (\gamma_{32}^2 + \gamma_{31}^2\gamma_{12}^2 + 2\gamma_{31}\gamma_{32}\gamma_{12})\sigma_2^2 + \sigma_3^2}, \quad (16)$$

$$\gamma_{s.32} = \gamma_{32}\sigma_2 / \sqrt{\gamma_{31}^2\sigma_1^2 + (\gamma_{32}^2 + \gamma_{31}^2\gamma_{12}^2 + 2\gamma_{31}\gamma_{32}\gamma_{12})\sigma_2^2 + \sigma_3^2}, \quad (17)$$

$$C_e^{(31)} = \gamma_{s.31}, \quad C_e^{(12)} = \gamma_{s.12}, \quad C_e^{(32)} = \gamma_{s.32} + \gamma_{s.31}\gamma_{s.12}, \quad (18)$$

$$\rho_{13} = \gamma_{s.31} + \gamma_{s.32}\gamma_{s.12}, \quad \rho_{12} = \gamma_{s.12}, \quad \rho_{23} = \gamma_{s.32} + \gamma_{s.31}\gamma_{s.12}. \quad (19)$$

4.3.5 *Ex.* $f_3(f_1 \rightarrow f_2)$. The variable f_{i3} is exogenous; we assume $f_{i3} \sim N(0, \sigma_3^2)$.

$$f_{i1}|f_{i3} \sim N(\gamma_{13}f_{i3}, \sigma_1^2), \quad f_{i2}|f_{i1}, f_{i3} \sim N(\gamma_{21}f_{i1} + \gamma_{23}f_{i3}, \sigma_2^2); \quad E[f_{i3}e_{i1}] = E[f_{i3}e_{i2}] = E[f_{i1}e_{i2}] = 0, \quad (20)$$

$$\gamma_{s.21} = \gamma_{21}\sqrt{\gamma_{13}^2\sigma_3^2 + \sigma_1^2} / \sqrt{\gamma_{21}^2\sigma_1^2 + (\gamma_{23}^2 + \gamma_{13}^2\gamma_{21}^2 + 2\gamma_{13}\gamma_{23}\gamma_{21})\sigma_3^2 + \sigma_2^2}, \quad (21)$$

$$\gamma_{s.13} = \gamma_{13}\sigma_3 / \sqrt{\gamma_{13}^2\sigma_3^2 + \sigma_1^2}, \quad (22)$$

$$\gamma_{s.23} = \gamma_{23}\sigma_3 / \sqrt{\gamma_{21}^2\sigma_1^2 + (\gamma_{23}^2 + \gamma_{13}^2\gamma_{21}^2 + 2\gamma_{13}\gamma_{23}\gamma_{21})\sigma_3^2 + \sigma_2^2}, \quad (23)$$

$$C_e^{(13)} = \gamma_{s.13}, \quad C_e^{(21)} = \gamma_{s.21}, \quad C_e^{(23)} = \gamma_{s.23} + \gamma_{s.13}\gamma_{s.21}, \quad (24)$$

$$\rho_{13} = \gamma_{s.13}, \quad \rho_{12} = \gamma_{s.21} + \gamma_{s.13}\gamma_{s.23}, \quad \rho_{23} = \gamma_{s.23} + \gamma_{s.13}\gamma_{s.21}. \quad (25)$$

4.3.6 *Ex.* $f_3(f_2 \rightarrow f_1)$. The variable f_{i3} is exogenous; we assume $f_{i3} \sim N(0, \sigma_3^2)$.

$$f_{i2}|f_{i31} \sim N(\gamma_{23}f_{i3}, \sigma_2^2), \quad f_{i1}|f_{i2}, f_{i3} \sim N(\gamma_{12}f_{i2} + \gamma_{13}f_{i3}, \sigma_1^2); \quad E[f_{i3}e_{i1}] = E[f_{i3}e_{i2}] = E[f_{i2}e_{i1}] = 0, \quad (26)$$

$$\gamma_{s.12} = \gamma_{12}\sqrt{\gamma_{23}^2\sigma_3^2 + \sigma_2^2} / \sqrt{\gamma_{12}^2\sigma_2^2 + (\gamma_{13}^2 + \gamma_{23}^2\gamma_{12}^2 + 2\gamma_{13}\gamma_{23}\gamma_{12})\sigma_3^2 + \sigma_1^2}, \quad (27)$$

$$\gamma_{s.13} = \gamma_{13}\sigma_3 / \sqrt{\gamma_{12}^2\sigma_2^2 + (\gamma_{13}^2 + \gamma_{23}^2\gamma_{12}^2 + 2\gamma_{13}\gamma_{23}\gamma_{12})\sigma_3^2 + \sigma_1^2}, \quad (28)$$

$$\gamma_{s.23} = \gamma_{23}\sigma_3 / \sqrt{\gamma_{23}^2\sigma_3^2 + \sigma_2^2}, \quad (29)$$

$$C_e^{(13)} = \gamma_{s.13} + \gamma_{s.23}\gamma_{s.12}, \quad C_e^{(12)} = \gamma_{s.12}, \quad C_e^{(23)} = \gamma_{s.23}, \quad (30)$$

$$\rho_{13} = \gamma_{s.13} + \gamma_{s.23}\gamma_{s.12}, \quad \rho_{12} = \gamma_{s.12} + \gamma_{s.13}\gamma_{s.23}, \quad \rho_{23} = \gamma_{s.23}. \quad (31)$$

Fig. 4. DIC for the six SEM models.

4.4 Comparison of the Six SEMs

Before we propose a method for resolving uncertainty of causal directions of the three constructs, we demonstrate how the 3C-SEM is not distinguishable from the other SEMs with different causal directions of the three constructs.

We make a comparison of the six SEMs with the construct models in Figure 3 in the way proposed in [13], where we resort to MCMC simulation with OpenBUGS [19]. In each model, the **Deviance Information Criterion (DIC)** [16, 18] is obtained for model comparison (*the smaller the better*), and the **Posterior Predictive p -value (PPp-value)** is calculated for model checking (values around 0.5 are plausible). The DIC is a kind of penalized log-likelihood criterion.

For the six models of this article, we conducted MCMC simulations with $B_{av} = 20, 40, 60, 100$, and 150 for the task—Msync pairs of (object, MAB), (object, BUF), and (castanets, BUF), and with an additional B_{av} of 300 for the pair of (castanets, MAB), to obtain the DIC and PPp-value of each model given the B_{av} values.² Three chains of each model were run for a 81,000-iteration period following a 1,000-iteration burn-in with thinning of 10.

Figure 4 plots DIC as a function of B_{av} . Note that the DIC differences of more than 10 might definitely rule out models with higher DIC, and differences between 5 and 10 are substantial [18]. A red figure above a bar indicates the difference between the DIC value of the bar and that of 3C-SEM when its absolute value is not smaller than 5. Out of the 126 DIC values in Figure 4, we find

²The DIC values of 3C-SEM in Figure 4 are either equal to or slightly smaller than those in [13] because this article has used a revised version of the OpenBUGS codes in the supplementary material of [13]. The difference does not affect the conclusion of the superiority of the 3C-SEM over the 3C-CFA and 1C-CFA. We have also calculated WAIC [37] of the three models; the calculation again supports the superiority of the 3C-SEM, which gives the smallest WAIC value among the three.

a single absolute value larger than 10 (i.e., -11) and 14 values between 5 and 10. Approximately 88 % of the values are smaller than 5. We can thus see that it is difficult to decide which model has the smallest DIC for all B_{av} values in each task-Msync pair.

With respect to the PPp-values, we have found that there are no substantial differences among the six models, though they vary to small extent over the models with the same B_{av} value, and that the PPp-values exhibit satisfactory fit since they are in a range of $[0.43, 0.49]$; note that the PPp-value of 0.5 implies a plausible model, and mid-range values (between 0.2 and 0.8) indicate satisfactory models [16]. A figure of the PPp-values can be found as Supplemental-Figure 1 in the supplemental material accompanying this article.

In addition to DIC for comparison of the six SEMs, we have further calculated **Widely Applicable Information Criterion (WAIC)** [37], which calls researchers' attention in statistics and machine learning. WAIC is applicable to a wider classes of statistical models than DIC, including singular models as well as regular ones. As a result of the calculation, we noticed no substantial differences in WAIC among the six models. We have thus learned that the six SEMs are not distinguishable (identifiable) in the cause-effect relationship only by the DIC, PPp-values, and WAIC.

4.5 QoE for 3C-SEM and Ex. f_3

We compare three QoE measures (the larger the better) between 3C-SEM and Ex. f_3 , since these two models have opposite causal directions. Figure 5 plots *posterior means* of work difficulty (WkDiffi), number of completed tasks (NCmpl) as an objective measure, and Castanet Sync along with their *95% credible intervals* (represented by vertical lines with top and bottom bars) as a function of B_{av} for the four pairs of task-Msync. From this figure, we see that the QoE values for 3C-SEM and that for Ex. f_3 are almost the same. Although we have shown only the three QoE measures here, we have confirmed the same results with respect to the other measures.³ This result is compatible with our finding of indistinguishability among the six models in the previous subsection.

In the pair of (Castanets, MAB), we find that as B_{av} increases from 40 to 60, WkDiffi improves, while NCmpl and Cast Sync both deteriorate slightly. As B_{av} further increases from 60 via 100 to 150, values of all the three measures fall at $B_{av} = 100$ and then rise at $B_{av} = 150$ again. As B_{av} increases to 300, WkDiffi and Cast Sync fall at slow rates, while NCmpl keeps approximately the same value. This behavior of QoE is very similar to that of the “impact-perceive-adapt” model mentioned in Section 2.1; $B_{av} = 60, 100$, and 150 approximately correspond to the three thresholds.

We will elucidate this observation from a causality point of view in Section 6.

5 A METHOD FOR INFERRING CAUSAL DIRECTIONS

This section proposes a Bayesian method for inferring causal directions of constructs; we predict QoE of the six SEMs in Figure 3 by means of MCMC simulation with OpenBUGS [19]. The probability of the prediction accuracy is derived for each SEM. Comparing the degrees of accuracy between two SEMs provides the posterior probability that one model is more plausible than the other. Note that we do not make the comparison not in units of directions of individual edges between two SEMs but in units of the set of directions for the three edges in a construct model.

The procedure for the evaluation can be stated as follows:

- (1) A SEM in Figure 3 has one or two exogenous variables, which can be identified as a cause or causes, while an endogenous variable corresponds to an effect.

³For reference, figures of haptic-related QoE measures are given in the supplemental material as Supplemental- Figure 2.

Fig. 5. WkDiffi, NCmpl, and CastSync versus B_{av} for 3C-SEM and Ex.f3.

- (2) In each SEM, we select a B_{av} value, which we call the *model- B_{av} value* being denoted as B_{av}^M (say $B_{av}^M = 40$). Using the dataset for B_{av}^M , we then estimate unknown parameters including the regression and loading coefficients of the SEM in a way similar to [13].
- (3) Utilizing the estimates of unknown parameters for B_{av}^M , we can make prediction of the indicators of the endogenous variables for the target B_{av} value, which is denoted as B_{av}^t (say $B_{av}^t = 20$). The prediction can be made by setting observed scores for B_{av}^t to the indicators of the exogenous variables, while **Not Available (NA)** to the endogenous ones in each SEM. We can then evaluate the accuracy of the predicted values of all indicators by comparing them to the corresponding observed ones if available for the B_{av}^t in each SEM. Note that setting $B_{av}^M = B_{av}^t$ can be considered the most accurate. However, this is not always the case; we will discuss this issue later.
- (4) As the accuracy measure, we adopt the **Mean Square Error (MSE)** between the posterior mean of predicted *qoe* and **Mean Opinion Score (MOS)**. We did not employ individual scores because of their fluctuations but the expectations of scores; in particular, MOS was chosen as the baseline for the comparison since it is independent of any SEM. The MSE is obtained by taking an average of the individual MSE between the two over all the 14 or 15 QoE measures. We thus get the posterior probability distribution of MSE.
- (5) Changing B_{av}^t in turn from 20 to 150 (or 300) exhaustively, we follow the same procedure for each B_{av}^t : $B_{av}^t = 20, 40, 60, 100$, and 150 for the pairs of (object, MAB), (object, BUF), and (castanets, BUF), and additionally 300 for the pair of (castanets, MAB).
- (6) Once we have completed the evaluation of the posterior probabilities of MSEs for all the SEMs, we compare them between two SEMs (say M_1 and M_2) and calculate the posterior probability that the MSE of M_1 is smaller than or equal to that of M_2 , which is denoted as $P_{MSE}(M_1 \leq M_2)$. We can consider this the probability that causal directions of constructs in M_1 are more plausible than those in M_2 . In OpenBUGS, we can get the probability $P_{MSE}(M_1 \leq M_2)$ with many iterations of the unit step function $\text{step}(\text{MSE of } M_2 - \text{MSE of } M_1)$,

where $\text{step}(x) = 1$ if $x \geq 0$, $\text{step}(x) = 0$ otherwise. Also, $P_{MSE}(M_2 < M_1) = 1 - P_{MSE}(M_1 \leq M_2)$.

Note that the dataset of observation for both exogenous and endogenous is available for all the B_{av}^t values in this study since the scores have been collected in a controlled experimental environment (see Section 3.2). The stimuli were given to the subjects in random order under the same traffic conditions in a way stratified by the pair of (task, MSync). In public Internet environments, however, data collection like this is impossible since we cannot control the measurement conditions.

6 SIMULATION RESULTS

We carried out MCMC simulation of the six SEMs based on the construct models in Figure 3, utilizing OpenBUGS. Priors of unknown parameters were set in the same manner as those in [13]. For details, see OpenBUGS codes in the supplemental material accompanying this article.

Three chains of each model were run for a 41,000-iteration period following a 1,000-iteration burn-in with thinning of 10 for each B_{av} value.

Prediction of $qoe[j]$ ($j = 1, 2, \dots, 14$ or 15) by a SEM with $B_{av}^M = B_{av}^t$ was made for each B_{av}^t by providing observed scores to the indicators of the exogenous variable(s) (in the task of object movement, for example, $z_1, \dots, z_8$ in 3C-SEM, but $z_9, \dots, z_{14}$ in Ex. f_3) and further assuming *missing scores* (NA) for the indicators of the endogenous variable(s).

6.1 Posterior Median of MSE

Figure 6 plots the posterior median of MSE between predicted qoe and MOS as a function of the audio-video buffering time B_{av} for the four pairs of task and Msync: (object, MAB), (object, BUF), (castanets, MAB), and (castanets, BUF). We have used the median instead of the mean since the distributions of the MSE are asymmetric; two examples are given in Figure 7.

In Figure 6, we first find that for each model of a given Msync (MAB or BUF), the MSE of the castanets hitting task is generally smaller than that of the object movement task. We next see that the 3C-SEM model achieves the minimum MSE among the six models except for $B_{av} = 20$ in the pairs of (object, MAB), (object, BUF), and (castanets, MAB), and $B_{av} = 60$ in the pair of (castanets, MAB). We further observe that for $B_{av} = 150$ in the pair of (castanets, BUF), the MSE of 3C-SEM is close to that of Ex. $f_1(f_2 \rightarrow f_3)$.

In the cases of $B_{av} = 20$, the most plausible causal direction seems to be ambiguous. This is because $B_{av} = 20$ is not long enough to absorb network delay jitter, and therefore, the output media quality degrades. This implies that estimates of unknown parameters in SEMs with $B_{av} = 20$ are not accurate. We refer to this finding as the *causal confusion due to the degradation of output media quality*, on which we will enlarge in the next subsection.

When $B_{av} = 60$ in the pair of (Castanets, MAB), the MSE of Ex. f_3 is slightly smaller than that of 3C-SEM, which suggests that plausible causal directions of the constructs in this case ($f_1 \leftarrow f_3 \rightarrow f_2$) are opposite to those in many other cases (i.e. $f_1 \rightarrow f_3 \leftarrow f_2$). This finding is referred to as *causal confusion due to interstream asynchrony by MAB*. Recall that $B_{av} = 60$ in the pair of (Castanets, MAB) can be regarded as the impact threshold introduced by Jay, Glencross, and Hubbard [24] as stated in Section 4.5. We can therefore make a conjecture that the “impact-perceive-adapt” behavior of QoE is related to changes in plausible causal directions.

As we have seen, it is difficult to understand how one model (M_1) is more plausible than another (M_2) with respect to causal directions of constructs in a quantitative way only through the posterior median of MSE. A solution to the problem is the derivation of $P_{MSE}(M_1 \leq M_2)$, which is defined in Section 5 as the posterior probability that the MSE of M_1 is smaller than or equal to that of M_2 .

Fig. 6. Posterior median of MSE versus B_{av} : MSE between predicted qoe and MOS.

6.2 Posterior Probability of Causal Direction

Since we have set the 3C-SEM as the baseline in this study, we evaluate $P_{MSE}(3C-SEM \leq M)$ and $P_{MSE}(M < 3C-SEM) = 1 - P_{MSE}(3C-SEM \leq M)$, where M is either $Ex.f_3$, $Ex.f_1(f_2 \rightarrow f_3)$, $Ex.f_2(f_1 \rightarrow f_3)$, $Ex.f_3(f_1 \rightarrow f_2)$ or $Ex.f_3(f_2 \rightarrow f_1)$.

6.2.1 BGR Diagnostic. When the probability $P_{MSE}(3C-SEM \leq M)$ is larger than approximately 0.9, or equivalently when $P_{MSE}(M < 3C-SEM)$ is smaller than about 0.1, the *BGR diagnostic*, which is a metric of convergence in MCMC simulation [18], was not obtained. However, the convergence of simulation has been confirmed by means of other output data including the ratio of **Monte Carlo standard error (MCSE)** to the posterior standard deviation less than 0.05, posterior probabilities, and dynamic trace of simulation runs.

The BGR diagnostic, which is denoted as $\hat{R}$, was not obtained because (1) the unit step function, $\text{step}(\text{MSE of } M - \text{MSE of } 3C-SEM)$, utilized for calculation of $P_{MSE}(3C-SEM \leq M)$ is a binary function, which takes 0 or 1, and (2) $\hat{R}$ is defined by B/W where W and B are within- and between-chain variability, respectively, for simulation of multiple chains (say M chains, each of length $2T$) [18, p.75]. The empirical $100(1 - \alpha)\%$ credible interval for each chain is calculated from the final T iterations as a measure of variability. The average width of these intervals across the M chains is W , while all MT samples pooled together gives B . OpenBUGS sets $\alpha = 0.2$ [19], i.e., the 80% credible interval is calculated. When the probability is larger than 0.9, the relative frequency of output “1” during the last T iterations ($T = 20,000$ here) is also approximately 0.9; consequently,

Fig. 7. Posterior probabilities of MSE and causal directions, and dynamic trace of three-chain simulation for 3C-SEM and $\text{Ex.f}_1(f_2 \rightarrow f_3)$ of (object, BUF) with $B_{av} = 60$, where the BGR diagnostic cannot be calculated.

the 80% credible interval covers only output “1” and leads to length 0 owing to continuity of real numbers. We thus get $B = W = 0$, and $\hat{R} = B/W$ becomes indeterminate.

Figure 7 shows an example of the output data when the BGR diagnostic cannot be calculated: posterior probabilities and dynamic traces of three-chains (represented by green, blue, and red lines) for 3C-SEM and $\text{Ex.f}_1(f_2 \rightarrow f_3)$ of (object, BUF) with $B_{av} = 60$.

6.2.2 Probability of Causal Direction. Figure 8 plots the posterior probabilities of causal direction as a function of B_{av} for the four pairs of task and Msync. The models compared with 3C-SEM are Ex.f_3 , $\text{Ex.f}_3(f_1 \rightarrow f_2)$, and $\text{Ex.f}_3(f_2 \rightarrow f_1)$ in the left panels, whereas $\text{Ex.f}_1(f_2 \rightarrow f_3)$ and $\text{Ex.f}_2(f_1 \rightarrow f_3)$ in the right panels. By the left panels, we can make comparison of the probabilities of $f_1 \rightarrow f_3 \leftarrow f_2$ and $f_1 \leftarrow f_3 \rightarrow f_2$ along with comparison of the correlation and causation between f_1 and f_2 . The right panels provide information on how the correlation between f_1 and f_2 , namely, $(f_1 \leftrightarrow f_2)$ is more plausible than the causation $f_1 \rightarrow f_2$ and $f_2 \rightarrow f_1$.

6.2.3 Causal Confusion. We first consider the causal confusion due to the degradation of output media quality at $B_{av} = 20$. When $B_{av} = 20$ in the left panels of Figure 8, we notice that $P_{MSE}(3C-SEM \leq \text{Ex.f}_3)$ is not larger than $P_{MSE}(\text{Ex.f}_3 < 3C-SEM)$ in the pairs (object, BUF) or (castanets, MAB), though the pair (object, MAB) exhibits $P_{MSE}(3C-SEM \leq \text{Ex.f}_3)$ larger than $P_{MSE}(\text{Ex.f}_3 < 3C-SEM)$ as well as the pair (castanets, BUF).

As an output video quality measure, we introduce the *video slice arrival ratio*, which is defined as the ratio of the number of output video slices to the total number of transmitted video slices. Recall that a video frame is divided into 15 slices, each of which forms an IP packet.

Figure 9(a) plots the video slice arrival ratio versus B_{av} for the four pairs of task and Msync. From this figure, we learn that the output video quality degrades largely at $B_{av} = 20$, while $B_{av} \geq 40$ improves the quality. Therefore, we can expect more accurate SEMs for $B_{av} = 20$ if we use estimates of unknown parameters for $B_{av} \geq 40$ in building SEMs for $B_{av} = 20$, i.e., $B_{av}^M \geq 40$.

Figure 9(b) demonstrates $P_{MSE}(3C-SEM \leq \text{Ex.f}_3)$ and $P_{MSE}(\text{Ex.f}_3 < 3C-SEM)$ for $B_{av}^M = 20$ versus the pair of task and Msync in the case of $B_{av}^M = 40$. We observe larger or much larger $P_{MSE}(3C-SEM \leq \text{Ex.f}_3)$ than $P_{MSE}(\text{Ex.f}_3 < 3C-SEM)$.

We next examine the causal confusion due to interstream asynchrony by MAB, looking at $P_{MSE}(3C-SEM \leq \text{Ex.f}_3)$ and $P_{MSE}(\text{Ex.f}_3 < 3C-SEM)$ at $B_{av} = 60$ for the (castanets, MAB) pair of Figure 8, where we can confirm that the former is smaller than the latter. However, comparing 3C-SEM with $\text{Ex.f}_3(f_1 \rightarrow f_2)$ and $\text{Ex.f}_3(f_2 \rightarrow f_1)$ (see the left panel), and with $\text{Ex.f}_1(f_2 \rightarrow f_3)$ and $\text{Ex.f}_2(f_1 \rightarrow f_3)$ (in the right panel), we find that $P_{MSE}(3C-SEM \leq M)$ is larger than

Fig. 8. Posterior probability of causal direction versus B_{av} . The bar marked with $M_1 \leq M_2$ displays the probability that MSE of M_1 is smaller than or equal to that of M_2 ; i.e., $P_{MSE}(M_1 \leq M_2)$.

$P_{MSE}(M \leq 3C\text{-SEM})$, where M denotes the other four SEMs. This suggests that Ex. f_3 has the most plausible causal directions at $B_{av} = 60$. Increasing B_{av} from 60 to 100, or decreasing it from 60 to 40 in the (castanets, MAB) pair, we observe that $P_{MSE}(3C\text{-SEM} \leq M)$ becomes larger than $P_{MSE}(M \leq 3C\text{-SEM})$ again. Therefore, we see that the more plausible causal directions reverse at

Fig. 9. Video slice arrival ratio versus B_{av} and posterior probability of causal direction for $B_{av}^t = 20$ with $B_{av}^M = 40$. The left panel was drawn by referring to a figure in [20].

$B_{av} = 60$ from $f_1 \rightarrow f_3 \leftarrow f_2$ to $f_1 \leftarrow f_3 \rightarrow f_2$ and vice versa. We can thus validate the conjecture that the “impact-perceive-adapt” behavior in QoE accompanies reversal of plausible causal directions like a flip-flop.

It is interesting to note that we may observe the “impact-perceive-adapt” behavior in the other three pairs of task and Msync if we take finer observational intervals of B_{av} . We need further investigations to have full understanding of the behavior.

6.2.4 Effect of End-to-End Delay. In the case of BUF ($B_h = B_{av}$), we can examine the effect of only the end-to-end delay not including interstream skew due to $B_h \neq B_{av}$, which is the case with MAB, on the probability of causal direction.

First, let us look at the left panel of (object, BUF) with $B_{av} \geq 40$. We then find that $P_{MSE}(3C-SEM \leq M)$, where M is $Ex.f_3$, $Ex.f_3(f_1 \rightarrow f_2)$, or $Ex.f_3(f_2 \rightarrow f_1)$, monotonically increases as B_{av} increases; this indicates that *the increasing end-to-end delay helps the users identify plausible causal directions with higher probabilities* (i.e., $f_1 \rightarrow f_3 \leftarrow f_2$ is more likely to be recognized than $f_1 \leftarrow f_3 \rightarrow f_2$).

The pair of (castanets, BUF) exhibits a similar tendency of the relationship between $f_1 \rightarrow f_3 \leftarrow f_2$ and $f_1 \leftarrow f_3 \rightarrow f_2$, though it is not so clear as that of (object, BUF). The effect of end-to-end delay on the causal directions seems to depend on the kind of task; slow movement and fine control of PHANToM may be clearly affected by end-to-end delay. This needs further study.

Next, let us examine how the end-to-end delay affects the correlation or causation between f_1 and f_2 under the condition of $f_1 \rightarrow f_3 \leftarrow f_2$ in the right panels of (object, BUF) and (castanets, BUF). In the pair of (castanets, BUF), we notice that $P_{MSE}(3C-SEM \leq Ex.f_1(f_2 \rightarrow f_3))$ monotonically decreases as B_{av} increases, whereas $P_{MSE}(3C-SEM \leq Ex.f_2(f_1 \rightarrow f_3))$ tends to increase and stays near 1.0 for $B_{av} \geq 60$. In the pair of (object, BUF), $P_{MSE}(3C-SEM \leq Ex.f_1(f_2 \rightarrow f_3))$ has the maximum at $B_{av} = 60$, while $P_{MSE}(3C-SEM \leq Ex.f_2(f_1 \rightarrow f_3))$ for $B_{av} \geq 40$ is almost constant (close to 1.0) for $B_{av} \geq 40$.

From these observations, we guess that the assumption of confounders affecting f_1 and f_2 (the correlation between them) outside the model is more reasonable than the causation $f_2 \rightarrow f_1$ as long as B_{av} is long enough to provide accurate estimates of unknown parameters. The assumption of confounder is also more reasonable than the causation $f_1 \rightarrow f_2$, but the relationship between the two depends on the kind of task. Clarification of the relationship requires more studies.

7 CONCLUSIONS

We proposed an empirical method based on QoE prediction with Bayesian structural equation modeling for deriving probabilities of causal directions for three constructs in haptic audiovisual

interactive communications. The probability of causal direction was obtained by comparison of two SEMs in terms of the MSE between predicted QoE and MOS in each SEM: The probability that MSE of SEM M_1 is smaller than or equal to MSE of SEM M_2 , $P_{MSE}(M_1 \leq M_2)$, is interpreted as the probability that causal directions in M_1 is more plausible than M_2 . Six constructs models (structural models) including that of 3C-SEM treated in [13] were built for that purpose, while keeping the measurement model the same. MCMC simulation was carried out extensively with OpenBUGS. As a result, we noticed that the causal directions in 3C-SEM, which was drawn from the domain knowledge, is the most plausible among the six. We also found that QoE can behave like the “impact-perceive-adapt” model proposed by Jay, Glencross, and Hubbard, which accompanies reversal of plausible causal directions in a flip-flop fashion, and that increasing end-to-end delay helps the users identify plausible causal directions with higher probabilities.

This study has many limitations. First, we confirmed the identifiability of the SEMs only through MCMC simulation and did not provide theoretical proof. The comparison of possible construct models was not exhaustive, though we could carry out exhaustive comparison in principle. We evaluated plausible causal directions of constructs from observed data and gave a piece of evidence to support the causal directions drawn from domain knowledge; however, we still rely on the domain knowledge in various aspects such as the number of constructs, selection of possible construct models, and formulation of the measurement model (e.g., the non-overlapping structure and informative priors) [13]. Thus, the proposed method cannot be replacement of established systematic methodologies for causal inference [1–3, 28–30] but be a supplementary one to them.

In addition, it is necessary to apply the proposed method to different datasets so that we can confirm results consistent with the one obtained in this article, especially the plausibility of the 3C-SEM model and the causal confusion. Regarding this issue, we have conducted another type of experiments with the same experimental system but by different subjects under similar conditions [38], by which we have made datasets that can be utilized for the validation purpose. Using a small portion of the dataset as a trial, we calculated probabilities of causal directions by the proposed method and then confirmed the plausibility of 3C-SEM and the causal confusion. However, models thus evaluated are limited; further investigation is needed to report convinced results.

In this article, we did not examine any issue of the gender effect. In order to investigate the gender effect on causal direction, we require new models according to the gender, which are left for future work. Even the present models, however, can assess the QoE measures for each gender separately and plot the figures. An example can be found in the supplemental figures.

Future work also includes application of the method to other multimedia communications QoE models and theoretical proof of the identifiability of the proposed method.

Finally, a fundamental issue of causality might be added. In this article, we have learned that the causality over the Internet is manipulable unlike causality in natural science. Then, a question arises. How should we validate the manipulation from a public consensus point of view? This might be controversial.

REFERENCES

- [1] Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference*, 2nd edition. Cambridge University Press.
- [2] Peter Spirtes, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search*, 2nd edition, The MIT Press.
- [3] Guido W. Imbens and Donald B. Rubin. 2015. *Causal Inference for Statistical, Social, and Biomedical Sciences, An Introduction*. Cambridge University Press.
- [4] Yutaka Ishibashi and Shuji Tasaka. 2003. Causality and media synchronization control for networked multimedia games: Centralized versus distributed. In *Proceedings of the ACM NetGames'03: 2nd workshop on Network and System Support for Games*. 42–51. DOI: [10.1145/963900.963904](https://doi.org/10.1145/963900.963904)

- [5] Amy Fire and Song-Chun Zhu. 2015. Learning perceptual causality from video. *ACM Transactions on Intelligent Systems and Technology* 7, 2 (Nov. 2015), Article 23, 23:1–23:22. DOI : [10.1145/2809782](https://doi.org/10.1145/2809782)
- [6] Gerold Blakowski and Ralf Steinmetz. 1996. A media synchronization survey: Reference model, specification, and case studies. *IEEE Journal on Selected Areas in Communications* 14, 1 (1996), 5–35. DOI : [10.1109/49.481691](https://doi.org/10.1109/49.481691)
- [7] David Bartholomew, Martin Knott, and Irini Moustaki. 2011. *Latent Variable Models and Factor Analysis: A Unified Approach, 3rd Edition*. John Wiley & Sons.
- [8] John C. Loehlin and A. Alexander Beaujen. 2017. *Latent Variable Models, An Introduction to Factor, Path, and Structural Equation Analysis, 5th edition*. Routledge.
- [9] Kenneth A. Bollen. 1989. *Structural Equations with Latent Variables*. John Wiley & Sons.
- [10] Tatsuya Yamazaki, Takumi Miyoshi, Masato Eguchi, and Kyoko Yamori. 2012. A service quality coordination model bridging QoS and QoE. In *Proceedings of the 2012 IEEE 20th International Workshop on Quality of Service*. 1–4.
- [11] Shuji Tasaka. 2016. Bayesian structural equation modeling of multidimensional QoE in haptic–audiovisual interactive communications. In *Proceedings of the IEEE International Conference on Communications*. 3345–3350. DOI : [10.1109/ICC.2016.7511202](https://doi.org/10.1109/ICC.2016.7511202)
- [12] Shuji Tasaka. 2019. Bayesian categorical modeling of multidimensional QoE in haptic–audiovisual communications. In *Proceedings of the IEEE International Conference on Communications*. 7. DOI : [10.1109/ICC.2019.8761784](https://doi.org/10.1109/ICC.2019.8761784)
- [13] Shuji Tasaka. 2020. Causal structures of multidimensional QoE in haptic–audiovisual communications: Bayesian Modeling. *ACM Transactions on Multimedia Computing, Communications, and Applications* 16, 1 (2020), Article 11, 23. DOI : [10.1145/3375922](https://doi.org/10.1145/3375922)
- [14] Karl G. Jöreskog and Dag Sörbom. 1993. *LISREL 8: Structural Equation Modeling with the SIMPLIS Command Language*. Scientific Software International.
- [15] Roy Levy and Robert J. Mislevy. 2016. *Bayesian Psychometric Modeling*. CRC Press.
- [16] Peter Congdon. 2014. *Applied Bayesian Modelling, 2nd Edition*. John Wiley & Sons.
- [17] Xin-Yuan Song and Sik-Yum Lee. 2012. *Basic and Advanced Bayesian Structural Equation Modeling with Application in the Medical and Behavioral Sciences*. John Wiley & Sons.
- [18] David Lunan, Christopher Jackson, Nicky Best, Andrew Thomas, and David Spiegelhalter. 2013. *The BUGS Book*. CRC Press.
- [19] *OpenBUGS*. 2018. Retrieved November 16, 2018 from <http://www.openbugs.net/w/Downloads>.
- [20] Eiichi Isomura, Shuji Tasaka, and Toshiro Nunome. 2013. QoE enhancement by media adaptive intra–stream synchronization in audiovisual and haptic IP communications. *IEICE Transactions on Communications* (in Japanese), J96–B, 2 (2013), 59–70.
- [21] Yutaka Ishibashi and Shuji Tasaka. 2000. A comparative survey of synchronization algorithms for continuous media in network environments. In *Proceedings of the 25th Annual IEEE Conference on Local Computer Networks*. 337–348. DOI : [10.1109/LCN.2000.891066](https://doi.org/10.1109/LCN.2000.891066)
- [22] Zixia Huang, Klara Nahrstedt, and Ralf Steinmetz. 2013. Evolution of temporal multimedia synchronization principles: A historical viewpoint. *ACM Transactions on Multimedia Computing, Communications and Applications* 9, 1s (2013), Article 34, 34:1–34:23. DOI : [10.1145/2490821](https://doi.org/10.1145/2490821)
- [23] Ralf Steinmetz. 1996. Human perception of jitter and media synchronization. *IEEE Journal on Selected Areas in Communications* 14, 1 (1996), 61–72. DOI : [10.1109/49.481694](https://doi.org/10.1109/49.481694)
- [24] Caroline Jay, Mashhuda Glencross, and Roger Hubbard. 2007. Modeling the effects of delayed haptic and visual feedback in a collaborative virtual environment. *ACM Transactions on Computer–Human Interaction* 14, 2 (2007), Article 8, 8/1–8/31. DOI : [10.1145/1275511.1275514](https://doi.org/10.1145/1275511.1275514)
- [25] Ayano Tatematsu, Yutaka Ishibashi, Norishige Fukushima, and Shinji Sugawara. 2010. QoE assessment in haptic media, sound, and video transmission: Effect of playout buffering control. In *Proceedings of the 2010 IEEE Intern. Workshop Tech. Committee on Commun. Quality Reliability*. 6. DOI : [10.1109/CQR.2010.5619913](https://doi.org/10.1109/CQR.2010.5619913)
- [26] Juan M. Silva, Mauricio Orozco, Jongeun Cha, Abdulmotaleb El Saddik, and Emil M. Petriu. 2013. Human perception of haptic–to–video and haptic–to–audio skew in multimedia applications. *ACM Transactions on Multimedia Computing, Communications and Applications* 9, 2 (2013), Article 9, 9:1–9:16. DOI : [10.1145/2457450.2457451](https://doi.org/10.1145/2457450.2457451)
- [27] Zhenhui Yuan, Ting Bi, Gabriel-Miro Muntean, and Gheorghita Ghinea. 2015. Perceived synchronization of mulse-media services. *IEEE Transactions on Multimedia* 17, 7 (2015), 957–966. DOI : [10.1109/TMM.2015.2431915](https://doi.org/10.1109/TMM.2015.2431915)
- [28] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press.
- [29] Wolfgang Wiedermann and Alexander von Eye, Editors. 2016. *Statistics and Causality: Methods for Applied Empirical Research*. John Wiley & Sons.
- [30] Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. 2006. A linear non–Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research* 7, 12/1/2006 (2006), 2003–2030.

- [31] Jonas Peters and Peter Bühlmann. 2014. Identifiability of gaussian structural equation models with equal variances. *Biometrika* 101, 1 (2014), 219–228. DOI : [10.1093/biomet/ast043](https://doi.org/10.1093/biomet/ast043)
- [32] Ricardo Silva, Richard Scheines, Clark Glymour, and Peter Spirtes. 2006. Learning the structure of linear latent variable models. *Journal of Machine Learning Research* 7, 12/1/2006 (2006), 191–246.
- [33] SensAble Technologies, Inc. 2009. OpenHaptics Toolkit version 3.0, Programmer’s Guide, 2009. Retrieved November 29, 2020 from <https://sivirt.utsa.edu/Documents/Manuals/OpenHapticsToolkit.pdf>.
- [34] Eckehard Steinbach, Sandra Hirche, Marc Ernst, Fernanda Brandi, Rahul Chaudhari, Julius Kammerl, and Iason Vittorias. 2012. Haptic communications. *Proceedings of the IEEE* 100, 4 (2012), 937–956. DOI : [10.1109/JPROC.2011.2182100](https://doi.org/10.1109/JPROC.2011.2182100)
- [35] Shuji Tasaka. 2017. Bayesian hierarchical regression models for QoE estimation and prediction in audiovisual communications. *IEEE Transactions on Multimedia* 19, 6 (2017), 1195–1208. DOI : [10.1109/TMM.2017.2652064](https://doi.org/10.1109/TMM.2017.2652064)
- [36] Stanley A. Mulaik. 2009. *Linear Causal Modeling with Structural Equations*. CRC Press.
- [37] Sumio Watanabe. 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research* 11, 116 (2010), 3571–3591. Retrieved from <http://jmlr.org/papers/v11/watanabe10a.html>.
- [38] Toshiro Nunome, Suguru Kaede, and Shuji Tasaka. 2020. User-assisted QoS control for QoE enhancement in audiovisual and haptic interactive IP communications. *IEICE Transactions on Communication* E103–B, 10 (2020), 1107–1116. DOI : [10.1587/transcom.2019EBP3235](https://doi.org/10.1587/transcom.2019EBP3235)

Received December 2020; revised April 2021; accepted July 2021