Audiovisual integration in the human perception of materials

Waka Fujisaki
Human Technology Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan

Naokazu Goda
Division of Sensory and Cognitive Information, National Institute for Physiological Sciences, Okazaki, Japan

Isamu Motoyoshi
NTT Communication Science Laboratories, Nippon Telegraph & Telephone Corporation, Atsugi, Japan
Present Address: Department of Life Sciences, The University of Tokyo, Tokyo, Japan

Hidehiko Komatsu
Division of Sensory and Cognitive Information, National Institute for Physiological Sciences, Okazaki, Japan

Shin'ya Nishida
NTT Communication Science Laboratories, Nippon Telegraph & Telephone Corporation, Atsugi, Japan

Interest in the perception of the material of objects has been growing. While material perception is a critical ability for animals to properly regulate behavioral interactions with surrounding objects (e.g., eating), little is known about its underlying processing. Vision and audition provide useful information for material perception; using only its visual appearance or impact sound, we can infer what an object is made from. However, what material is perceived when the visual appearance of one material is combined with the impact sound of another, and what are the rules that govern cross-modal integration of material information? We addressed these questions by asking 16 human participants to rate how likely it was that audiovisual stimuli (48 combinations of visual appearances of six materials and impact sounds of eight materials) along with visual-only stimuli and auditory-only stimuli fell into each of 13 material categories. The results indicated strong interactions between audiovisual material perceptions; for example, the appearance of glass paired with a pepper sound is perceived as transparent plastic. Rating material–category likelihoods follow a multiplicative integration rule in that the categories judged to be likely are consistent with both visual and auditory stimuli. On the other hand, rating-material properties, such as roughness and hardness, follow a weighted average rule. Despite a difference in their integration calculations, both rules can be interpreted as optimal Bayesian integration of independent audiovisual estimations for the two types of material judgment, respectively.

Introduction

One fundamental function of perception is allowing one to interact with objects in the environment. To decide whether a given object should be avoided, can be eaten, or is worthwhile to buy, one needs to estimate, above all, what material the object is made from. Material perception appears to involve complex sensory computation, but it is functionally important, and we are very good at it. Recently, there is a growing interest in human visual material perception (e.g., Adelson, 2001; Doerschner, Boyaci, & Maloney, 2010; Kim, Marlow, & Anderson, 2012; Motoyoshi, Nishida,

Citation: Fujisaki, W., Goda, N., Motoyoshi, I., Komatsu, H., & Nishida, S. (2014). Audiovisual integration in the human perception of materials. *Journal of Vision, 14*(4):12, 1–20, http://www.journalofvision.org/content/14/4/12, doi:10.1167/14.4.12.

doi: 10.1167/14.4.12 Received November 1, 2013; published April 17, 2014 ISSN 1534-7362

Sharan, & Adelson, 2007; Nishida & Shinya, 1998; Wijntjes & Pont, 2010; see also Anderson, 2011; Fleming, 2014; Maloney & Brainard, 2010; Zaidi, 2011 for review) and its neural correlates (e.g., Cant & Goodale, 2007, 2011; Cavina-Pratesi, Kentridge, Heywood, & Milner, 2010a, 2010b; Goda, Tachibana, Okazawa, & Komatsu, 2014; Nishio, Goda, & Komatsu, 2012). Although the focus of many studies has been on visual perception of material-related properties (in particular, gloss), considerable recent interest has also been directed toward a higher processing goal of material perception, perception of material category.

Past studies on visual material-category perception showed that human observers are extremely adept at material categorization, even when materials are presented very briefly (Sharan, Rosenholtz, & Adelson, 2009), and outperform up-to-date computer algorithms for material recognition (Sharan, Liu, Rosenholtz, & Adelson, 2013). It is suggested that the perception of material categories is systematically related to the perception of basic material-related properties (Fleming, Wiebel, & Gegenfurtner, 2013). Nevertheless, whether material recognition is based on a combination of several basic properties or on critical image features specific to each material category remains an open question. The neural process underlying material-category perception has been also studied. Cortical activity related to material categories is found in a higher-order area of the ventral visual cortex: the fusiform gyrus in humans (Hiramatsu, Goda, & Komatsu, 2011) and the inferior temporal cortex in monkeys (Goda et al., 2014).

Useful material category information is provided not only by vision, but also by audition, touch, smell, and taste. In particular, auditory material-category perception has received researchers' attention as much as or more than visual material-category perception (Aramaki, Besson, Kronland-Martinet, & Ystad, 2011; Giordano & McAdams, 2006; R. Klatzky, Pai, & Krotkov, 2000; Lemaitre & Heller, 2012; Lutfi & Oh, 1997; Wildes & Richards, 1988). For instance, Giordano and McAdams investigated people's ability to identify object materials from an impact sound (i.e., something striking the object) and demonstrated that listeners could almost perfectly discriminate between gross material categories, such as steel-glass versus wood-Plexiglas. Such studies have shown that acoustic features relevant to material discrimination include decay and spectral component of impact sounds. An fMRI study suggests that a subregion in a ventro-medial pathway is specialized for auditory-based material perception (Arnott, Cant, Dutton, & Goodale, 2008).

In daily life, we naturally attempt to obtain reliable information about material by combining multiple sensory modalities. When uncertain about an object's material from its visual appearance alone, for example, we cannot help hitting the object to hear its impact sound. The way in which material-category information is combined across different sensory modalities, however, remains unknown. What kind of material do we perceive when the visual appearance of one material is combined with the impact sound of another? In the present study, we aimed to identify the rules of such multimodal integration of material information by showing participants numerous audiovisual material combinations and asking them to judge the likelihood that a stimulus was made of a specific material.

While scant research has focused on multisensory integration in material-category perception, several studies have addressed the way in which information about a material-related property, such as roughness, is integrated across sensory modalities (see R. L. Klatzky & Lederman, 2010, for a recent review). It has been shown that multisensory integration of roughness information follows a weighted average rule with a higher weight given to the more reliable modality for the task (Lederman, Thorne, & Jones, 1986)—a rule most frequently found for multichannel integration of perceptual properties (Alais & Burr, 2004; Ernst & Banks, 2002; Landy, Maloney, Johnston, & Young, 1995; Yuille & Bülthoff, 1996). The fuzzy logical model (Massaro, 1987, 2004; Massaro & Stork, 1998) is another computational model of multimodal integration, originally developed for audiovisual speech perception. The model assumes an optimal integration of multiple sources of sensory information with each being independently evaluated to give the continuous degree to which that source specifies various alternative interpretations.

With these previous studies in mind, using the same set of stimuli as were used for material-category perception, we also examined audiovisual integration for a wide range of material-related properties. Some were visual (e.g., gloss), some auditory (e.g., high-pitched), and some tactile (e.g., hard). We also tested properties concerning subjective value (e.g., expensive). There were several purposes for our inclusion of material-property judgments. One was to assess whether the weighted average is a general rule of multisensory integration of material-property perception. Another was to examine whether audiovisual integration of material information could produce a change in visual appearance by an unmatched impact sound or a change in impact sound quality by an unmatched visual appearance. The main and final purpose was to see whether material-category perception follows the same rules as material related–property perception.

To anticipate, we found that multisensory integration of material-property perception could be described as weighted average. With regard to our second purpose, we found no clear evidence of perceptual property changes; that is, the weight was nearly exclusively given to vision for visual properties and to audition for

auditory properties. Finally, we found that the rule of multisensory integration for material category judgments was not one of weighted average, but of multiplication. We will discuss how this discrepancy may be ascribed to task difference and how both the multiplicative integration of the material-category judgments and the weighted average of the material-property judgments can result from the same computational principle (i.e., optimal Bayesian integration of independent visual and auditory signals; Ernst, 2006, 2012; Landy et al., 1995; Massaro, 1987, 2004; Massaro & Stork, 1998; Yuille & Bülthoff, 1996).

Methods

Participants

Participants consisted of 16 paid volunteers (age range: 20–40 years old), who were blind to the purpose of the experiment. All had normal or corrected-to-normal vision and hearing. This experiment was approved by the Institutional Review Board of the National Institute of Advanced Industrial Science and Technology (AIST) and was performed in accordance with the Declaration of Helsinki.

Stimuli

Visual stimuli

Visual stimuli were computer-generated movies of a scene in which a human hand hit an object with a small stick. Only a right arm and hand were visible, which was modeled by Poser 2010 SR3 (SmithMicro software) using the position and pose data captured from a video of actual human movement. Figure 1 shows example images for the six material categories we used: glass, ceramic, metal, stone, wood, and bark. We chose these material categories, their textures, and the cylindrical object shape following Hiramatsu et al. (2011). Glass had a transparent body with a glossy surface. The other materials were opaque. The ceramic surface had pure diffuse reflectance, and metal had pure specular reflectance. Stone, wood, and bark were created by texture mapping using texture data prepared by LightWave 3D (NewTek). The scene images were rendered under appropriate global illuminations with a grey background using Vue 9 (e-on software). The spatial resolution of the final output was 600 × 800 pixels.

Auditory stimuli

For auditory stimuli, we used the impact sounds of eight real objects (glass, ceramic, metal, stone, wood, vegetable (pepper), plastic, and paper; Figure 2a).

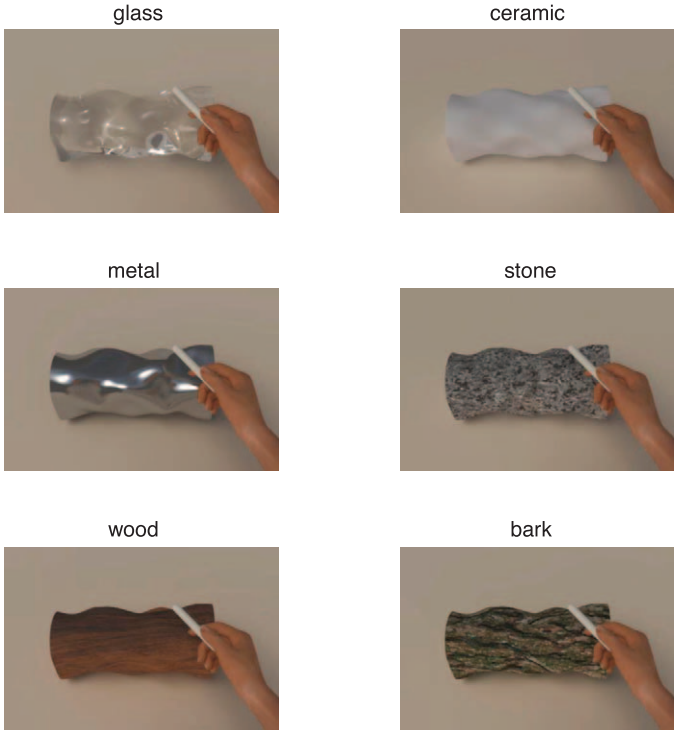

Figure 1. Snapshot images for six visual stimuli used in the experiment: glass, ceramic, metal, stone, wood, and bark.

These were selected from 16 sounds based on the data of a preliminary experiment (see Appendix). Auditory stimuli were created by hitting real objects with a wooden (maple) mallet in a soundproof chamber (Figure 2c). Some objects were selected from the Shitsukan sample set (Takei Scientific Instruments Co., Ltd., Japan) and others from daily items. The impact sound of hitting each object was recorded through a microphone (AKG, C1000S) and an audio interface (M-audio, 410) by using audio editing and recording software (Audacity 1.3.12) via USB with a computer (MacBook Air, Apple). To prevent any impact of floor vibration on recording, antivibration sponges and rubber were placed underneath the material before it was struck (Figure 2b). Inside the chamber, one of the authors struck the object and another operated the computer to ensure that the sound was recorded properly; the sound of each material was recorded at least 10 times, after which we checked the sound spectrograms and sound waves and selected a sound that we subjectively judged to be the "best" representative of the samples.

Audiovisual stimuli

Seven visual stimuli (including a blank image for auditory-only conditions) and nine auditory stimuli (including a silent sound for visual-only conditions)

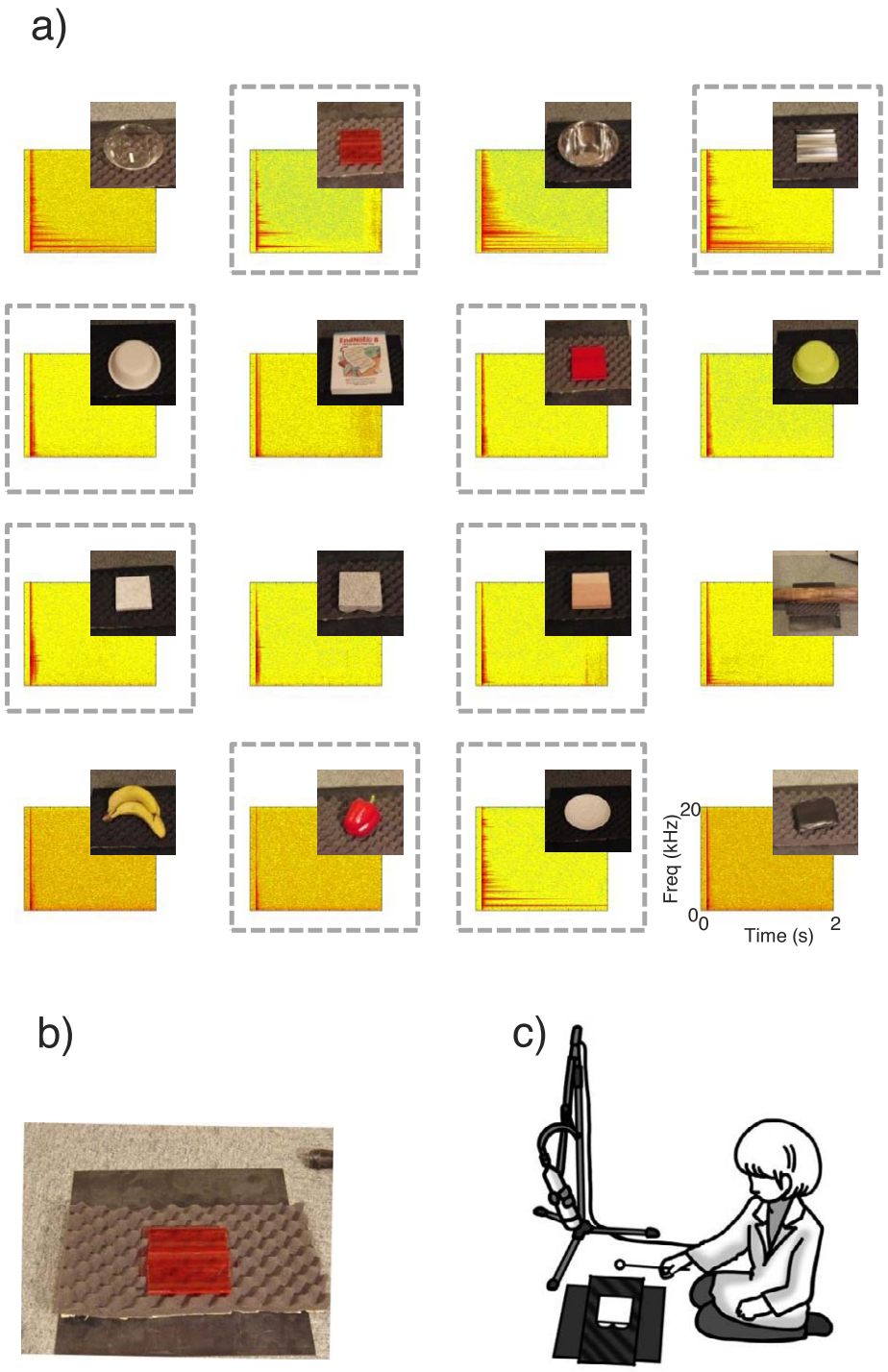

Figure 2. Auditory stimuli used in the experiment. (a) Sound spectrograms of the impact sounds of hitting real objects. Eight sounds selected for the main experiment are indicated by square dotted lines. (b) The setup for impact sound recording, including antivibration sponges and rubber being placed underneath the material. (c) Illustration of hitting the real object by a wooden (maple) mallet for the recording of the impact sound in a soundproof chamber.

were combined to create 62 combinations of audiovisual stimuli (excluding the combination of a blank image and a silent sound). Adobe Premiere Pro was used to create MP4-format movie clips from a series of rendered images combined with audio tracks of impact sounds. Each striking movement took about 1 s to complete and was repeated five times in one sequence.

(a) Material-category rating	Material-property rating			(e) Audiovisual combination naturalness rating
	(b) Visual properties	(c) Auditory properties	(d) Other properties	
Glass	Dark surface–bright surface	Soft sound–loud sound	Smooth-rough	Unnatural-natural
Ceramic	Uniform surface–textured surface	Low-pitched sound–high-pitched sound	Cold-warm	
Metal	Colorless surface–colorful surface	Dampened sound–ringing sound	Soft-hard	
Stone	Matte surface–gloss surface	Dull sound–sharp sound	Light-heavy	
Wood	Opaque looking–transparent looking	Mixed sound–pure sound	Dry-wet	
Vegetable		Narrow sound–broad sound	Hollow-solid	
Plastic		Mild sound–intense sound	Cheap-expensive	
Paper		Poor sound–rich sound	Dirty-clean	
Vinyl			Old-new	
Rubber				
Cloth				
Clay				
Leather				

Table 1. Words selected for the experiment. *Notes*: (a) Thirteen words selected for the material category–rating experiment. (b) Five visual properties used for the material property–rating experiment. (c) Eight auditory properties used for the material property–rating experiment. (d) Ten other properties, consisting of tactile, thermal, cross-modal, and other properties. (e) Audiovisual combination naturalness rating.

Questionnaires

Material-category rating

Table 1a shows the 13 material names that were used in the material category–rating experiment. Participants used seven-point scales to rate how well these words applied to each of the 62 stimulus combinations, ranging from 1 ("this could not in any way be material X," where "X" is the given material) to 7 ("this is almost certainly material X"). Although our rating method took more time than the alternative of asking participants to choose the best material category for each stimulus, it could reveal a more detailed profile of material perception. Specifically, we could tell whether participants judged a given stimulus as belonging to only one category or whether it fit into other categories as well.

Material-property rating

A set of 23 bipolar adjective pairs were used for the material property–rating experiments, most of which were selected by referring to previous literature (Cunningham, Wallraven, Fleming, & Strasser, 2007; Fujisawa, Iwamiya, & Takada, 2004; Gabrielsson & Sjogren, 1979; Osgood & Anderson, 1957; Solomon, 1958; von Bismarck, 1974a, 1974b). Table 1b, 1c, and 1d shows the five, seven, and 10 adjective pairs describing the visual, auditory, and other properties (e.g., tactile, thermal, cross-modal) rated by participants, respectively. Participants also rated the naturalness of each audiovisual combination (Table 1e). In the analysis of the material properties, "1" was assigned to the first adjective in each pair and "7" to the second.

Procedure

The experiment was divided into three sessions: the material-category rating, visual and auditory material-property rating, and other property rating and audiovisual naturalness rating. Each session consisted of three blocks: six visual-only trials (Block 1), eight auditory-only trials (Block 2), and 48 audiovisual combined trials (Block 3). Within each session, participants completed the visual-only (Block 1) and auditory-only (Block 2) trials first. Each block took about 5–10 min to complete. Participants then took a break for about 15 min. They then performed the audiovisual trials (Block 3), which took around 40 min to complete. We used this block order to reduce the participants' load by starting each session with short and easy blocks. This also ensured that participants rated visual and auditory stimuli before rating the audiovisual stimuli, which contained some unexpected combinations (although our procedure did not exclude prior exposure to audiovisual stimuli for the second and third sessions). Within each block, the order of stimulus presentations was randomized. The order of sessions was also randomized across participants. Participants rested sufficiently between sessions. The

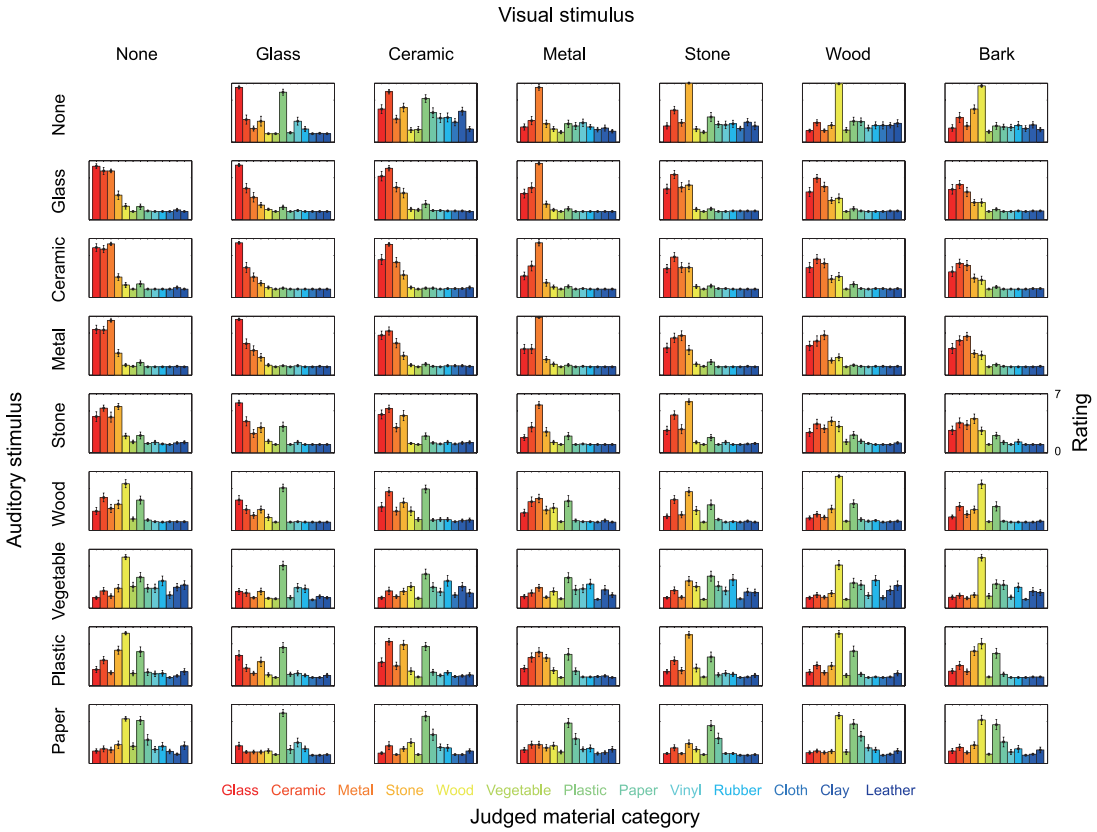

Figure 3. Material category rating for all 63 stimulus conditions. In each panel, rating values averaged over 16 participants are shown for each material-category judgment (shown by different color). While the actual rating ranges from one to seven, the ordinate ranges from zero to seven. Error bar: ±1 *SEM* across participants.

whole experiment took about 5–6 hr in total (including breaks).

The experiment was conducted in a well-lit room. The audiovisual stimuli, saved as MP4 movies, were presented on a web browser (Mozilla Firefox) through a PHP script running on a computer (Sony, VAIO VPCSE). The movie frame rate was 30 fps, and audio sampling frequency was 44 kHz. Visual stimuli were presented on a liquid crystal, vertically positioned monitor (Nanao, FlexScan L997). In each trial, a visual stimulus appeared in the upper half of the monitor screen. Participants wore headphones (Sennheiser HDA 200) through which the auditory stimuli were presented, using an amplifier (Audio-Technica, AT-HA2) from an onboard audio device (Realtek High Definition Audio). The amplitude of each sound was not normalized across stimuli because we thought that the amplitude itself might convey information about the properties of a material. The participants were instructed to set the overall volume of the amplifier to a comfortable level. The seven-point scale questionnaires were presented in the lower half of the monitor screen. Participants responded by clicking buttons on the monitor screen corresponding to their answers on the scales. Each movie lasted for 5 s, but the participants could replay it as many times as they wanted until they had completed all the questionnaires.

Results

Cross-modal interactions in material-category perception

Figure 3 shows the material category ratings for all 63 stimulus conditions. In each panel, rating values averaged over 16 participants were shown for each material-category judgment. Panels in the same row show the ratings for the same auditory stimuli, and panels in the same column show the ratings for the visual stimuli. Auditory-only conditions are shown in the leftmost column, and visual-only conditions are shown in the top row. The rating correlation between participants was 0.5474 (average of 120 individual pairs) with a standard deviation of 0.0798.

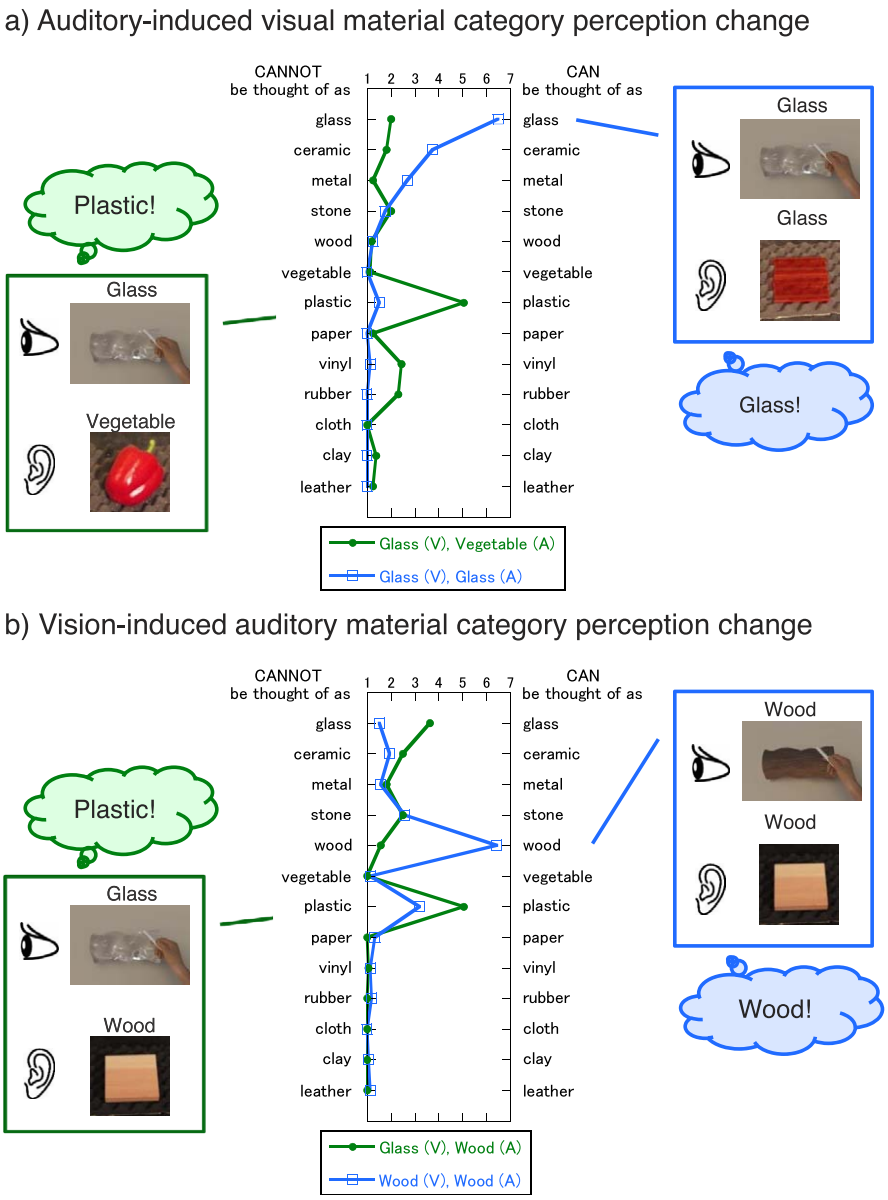

Figure 4. Examples of the obtained profiles in the material category experiment (average of 16 participants). (a) An auditory-induced visual material category–perception change. Different material categories were perceived for the same visual stimulus with different sounds. (b) A vision-induced auditory material category–perception change. Different material categories were perceived for the same auditory stimulus with different visual stimuli.

We observed strong audiovisual interactions in material-category perception. Figure 4 shows two examples of the obtained response profiles from the material-category rating. Figure 4a shows an auditory-induced perception change in a visual material category—namely, participants perceived different audiovisual material categories for the same visual stimulus when it was coupled with different sounds. For example, when a "glass" image was coupled with a "glass" sound (Movie 1), participants perceived the object as "glass," but when the same visual stimulus was coupled with the "vegetable" sound (Movie 2), participants tended to perceive the material category as "plastic." In Figure 3, this type of auditory-induced perception change is represented as a change in response profile along a vertical column. A clear change is observed in the second column (that includes the pair of conditions shown in Figure 4a) as well as in some other columns.

Figure 4b shows the opposite perception change: a vision-induced change in the perception of an auditory material category. In this case, when a "wood" sound and images were congruent, participants perceived the object as "wood," but the same sound coupled with a "glass" image made participants perceive "plastic." In Figure 3, this type of visual-induced perception change

Movie 1. A "glass" image coupled with a "glass" sound (Figure 3a, profile shown in blue). Participants tended to perceive the object's material category as "glass."

Movie 2. A "glass" image coupled with a "vegetable (pepper)" sound (Figure 3a, profile shown in green). Participants tended to perceive the object's material category as "plastic."

is represented as a change in response profile along a horizontal row. A clear change is observed in the sixth row (that includes the pair of conditions shown in Figure 4b) as well as in some other rows.

Integration of audiovisual information in material-category perception

To understand the mechanism underlying audiovisual interactions in material-category perception, we analyzed how visual and auditory information was integrated into multimodal perception. Figure 5 shows how we compared the response profiles of the material category rating (average of 16 participants) among visual-only, auditory-only, and audiovisual conditions. In Figure 5a, both visual and auditory stimuli were "ceramic." For the visual-only stimulus, high ratings were given to "ceramic" and "plastic." For the auditory-only stimulus, high ratings were given to "glass," "ceramic," and "metal." When the stimuli were combined, the highest rating was given to "ceramic." In Figure 5b, the auditory stimulus was replaced by "paper." For the auditory-only stimulus, high ratings were given to "wood" and "plastic." When it was paired with visual "ceramic," the highest rating was given to "plastic." In both cases, the materials rated highly for the audiovisual stimuli were those rated highly for both visual-only and auditory-only stimuli. This suggests that the integration rule of audiovisual information in material-category perception is similar to an AND operation rather than another rule such as a weighted average.

To test this conjecture, we conducted a multiple regression analysis using all the data, including a multiplicative interaction term. The regression equation was as follows:

$$\mathrm{VA} = \beta_0 + \beta_1 V_{\mathrm{only}} + \beta_2 A_{\mathrm{only}} + \beta_3 V_{\mathrm{only}} \times A_{\mathrm{only}}$$

In this equation, VA indicates the predicted audiovisual rating; V_{only} and A_{only} the obtained visual-only and auditory-only ratings, respectively; and $V_{\mathrm{only}} \times A_{\mathrm{only}}$ the multiplication of the two obtained ratings. The rating value was normalized to [0, 1]. Regression weights (β_1, β_2, β_3) for visual-only, auditory-only, and the interaction terms are shown in Figure 6. If VA can be explained by a simple weighted average of visual and auditory information, the regression weight of the interaction term, β_3, would be very small or zero. However, our result clearly showed that the value of this interaction term was significantly higher than zero. Scatter plots in Figure 7 show the relationships of the obtained audiovisual ratings with the three terms of the regression analysis: (a) visual-only rating, (b) auditory-only rating, and (c) the interaction. As shown in Figure 7c, multiplying the unimodal ratings predicted the multimodal rating with a reasonable degree of accuracy. The prediction is better described by a second-order polynomial regression ($R^2 = 0.8753$, AIC [Akaike information criterion] $= -1889.5$) than by a linear regression ($R^2 = 0.7739$, AIC $= -1520.3$).

Multiplication can be considered an AND operation, which identifies the material category most consistent with both visual-only and auditory-only stimuli. However, what if there was no category consistent with both visual-only and auditory-only stimuli? Figure 8a shows an example of such a case. Although the auditory "metal" stimulus gained high ratings for the "metal," "ceramic," and "glass" categories, the visual "wood" gained a high rating for the "wood" category

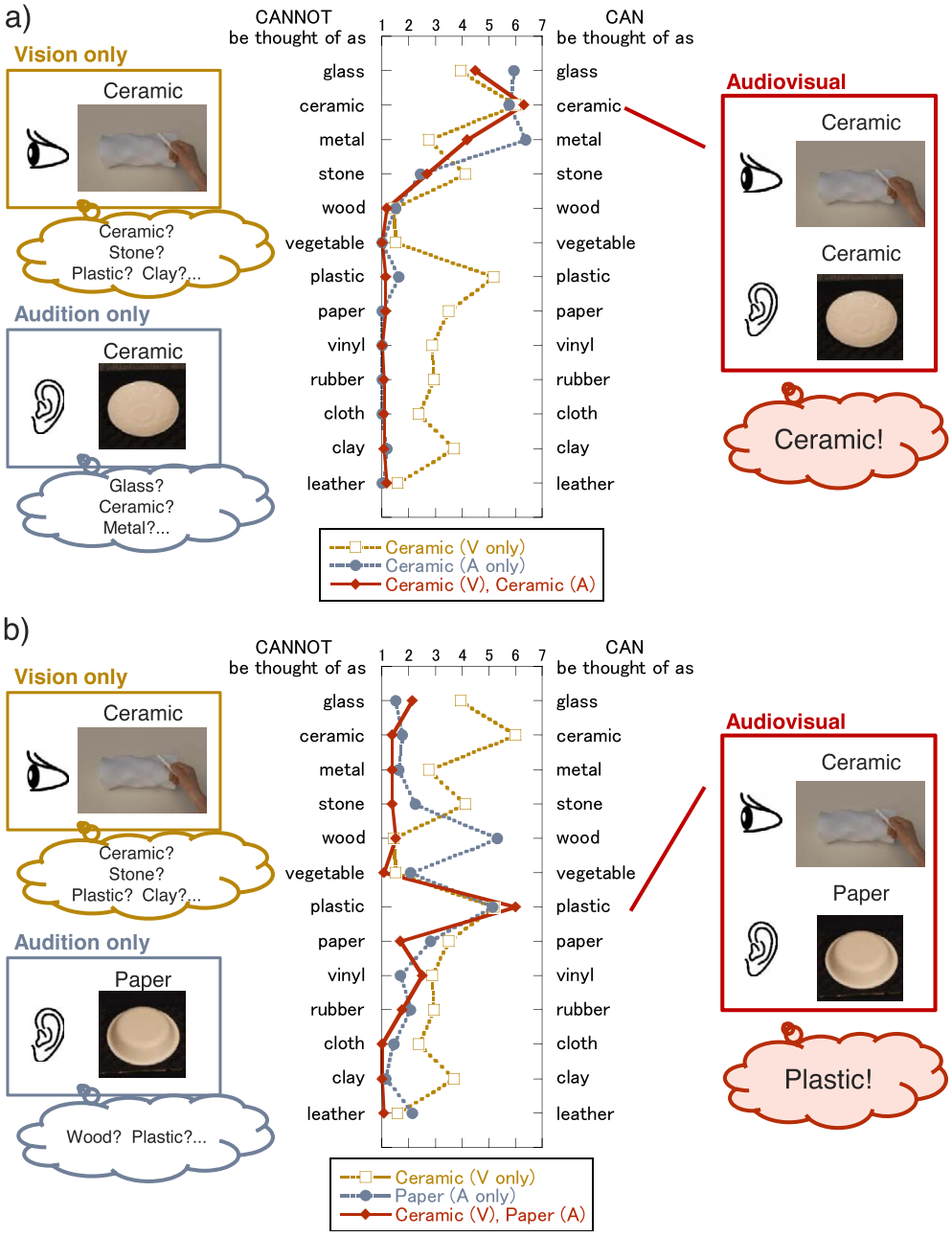

Figure 5. Examples of audiovisual integration in the material category–rating experiment (average of 16 participants) for the (a) congruent and (b) incongruent conditions. Audiovisual material rating is high when both visual-only and auditory-only ratings are high, and it is low when either the visual-only or auditory-only rating is low.

only. When these two were combined, relatively high scores were given to "metal," "ceramic," and "glass" categories, despite their low ratings in the visual-only stimulus. Note that the rating of audiovisual combination naturalness was low for this combination ($M = 1.3$). To show the generality of this observation, we made separate scatter plots for three different ranges of the naturalness rating (Figure 8b). The multiplication model strongly predicted the obtained audiovisual rating when the naturalness rating was high, but this relationship gradually collapsed as naturalness decreased. Furthermore, the regression analysis indicated that weight of the auditory term relative to the visual term increased for unnatural combinations, supporting auditory dominance (Figure 8c).

Integration of audiovisual material-property perception

Next, we consider the audiovisual integration of material properties and compare the associated inte-

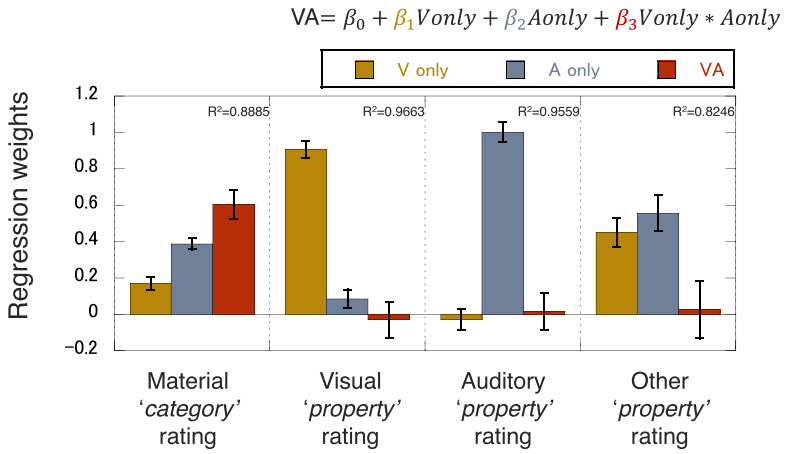

Figure 6. The results of the regression analysis. Regression weights for visual-only, auditory-only, and audiovisual terms are shown separately for the material-category rating and three types of material-property ratings. Error bars indicate 95% confidence intervals.

gration rule with that of material categories. Figures 9, 10, and 11, respectively, show the ratings for the visual, auditory, and other properties in the same format as Figure 3. The rating correlation between participants (mean ± 1 *SD*) was 0.6543 ± 0.1391 for visual properties, 0.5968 ± 0.1115 for auditory properties, and 0.4546 ± 0.1042 for other properties.

We found that even when the auditory stimulus was changed from "ceramic" to "paper" while the visual stimulus remained "ceramic," the audiovisual ratings for visual properties remain unchanged, continuing to follow the visual-only rating (Figure 12a). When an auditory stimulus changed from "ceramic" to "paper" while the visual stimulus remained "ceramic," the audiovisual ratings for auditory properties remain unchanged, continuing to follow the auditory-only rating (Figure 12b). For the "other" property rating, audiovisual ratings reflected both visual-only and auditory-only ratings. When the auditory stimulus changed from "ceramic" to "paper" while the visual stimulus remained "ceramic," the audiovisual rating was located about midway between the visual-only and auditory-only ratings (Figure 12c).

As with material-category ratings, we conducted a multiple regression analysis, including an interaction term. Regression weights (β_1, β_2, β_3) for visual, auditory, and audiovisual terms are shown in Figure 6. For visual properties, the regression weight for vision (β_1) was high, and that for audition (β_2) was very low. As shown by scatter plots in Figure 13, audiovisual ratings agreed strongly with vision-only ratings (Figure 13a) but not with auditory-only ratings (Figure 13b). For auditory properties, the regression weight for audition (β_2) was high, and that for vision (β_1) was nearly zero. As shown by scatter plots in Figure 14, audiovisual ratings agreed strongly with auditory-only ratings (Figure 14b) but not with visual-only ratings (Figure 14a). For other properties, the regression weights were comparable to those for audition and vision. As shown by scatter plots in Figure 15, audiovisual ratings had some correlations with both visual-only ratings (Figure 15a) and auditory-only ratings (Figure 15b). Importantly, for all types of material properties, the regression weight for the interaction term (β_3) was nearly zero, contrary to those of the material-category ratings. It should be noted that in other property ratings, the participants had to use two sources of information to infer the property values from a potentially ambiguous stimulus as in the case of material-category ratings. Thus, audiovisual material-property ratings appear to follow a weighted average rule. The weight was given almost exclusively to vision and audition for visual and auditory properties, respectively. The weight for "other properties" was similar for the two modalities when averaged over the nine properties we used. Looking at each property separately, however, we found a variety of patterns. For instance, the auditory weight was much higher for the mechanical properties of the object material ("soft-hard," "light-heavy," and "hollow-filled").

Discussion

In this study, we investigated how humans integrate material information from different sensory modalities, focusing on the interactions between object appearance and impact sounds. We observed strong interactions between these two types of information in a pattern that suggests an AND-like operation. That is, the likelihood rating for an audiovisual stimulus pair can be approximately predicted by a multiplication of the ratings for the visual-only and auditory-only stimuli. In

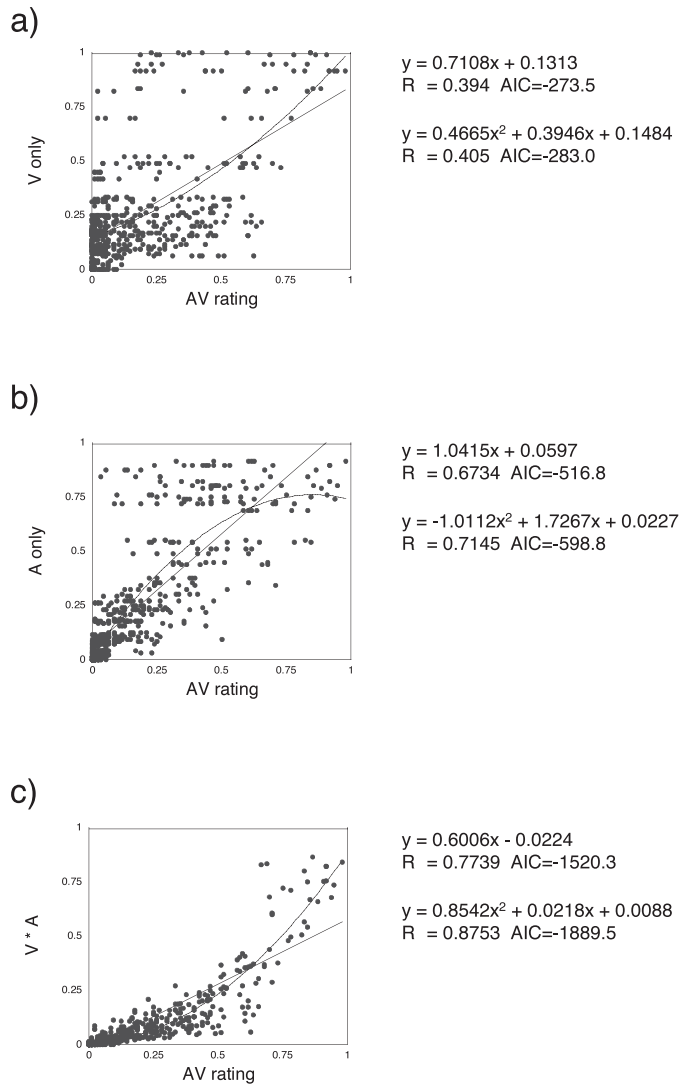

Figure 7. Scatter plots of the normalized material-category ratings for audiovisual stimuli versus (a) those for visual-only stimuli, (b) those for auditory-only stimuli, and (c) multiplication of the ratings for visual-only and auditory-only stimuli. Two continuous lines in each panel show linear and second-order polynomial regressions, the numeral values of which are shown on the right. AIC: Akaike information criterion. Multiplication of visual-only and auditory-only ratings best describes the audiovisual material-category ratings.

contrast, the integration rule for material-property judgments follows a weighted average rule.

Why do these two tasks follow different integration rules? Do the two rules arise from different computational principles? Are they specific to material perception? We suggest that the rule difference mainly reflects task differences and that both rules can be interpreted as an optimal integration of independent audiovisual estimations for each task. We use Bayesian inference as a computational framework.

Material-category rating

Consider first the computational mechanism underlying material-category rating. According to the Bayes theorem, the posterior probability that a perceived object falls into the nth material category, C_n, given the ith visual evidence, V_i, is proportional to the product of likelihood and prior probability, that is,

$$P(C_n|V_i) = \frac{P(V_i|C_n)P(C_n)}{P(V_i)}.$$

Similarly, the posterior probability that a perceived object falls into C_n given the jth auditory evidence, A_j, is

$$P(C_n|A_j) = \frac{P(A_j|C_n)P(C_n)}{P(A_j)}.$$

When the evidence of the two modalities, V_i and A_j, are available, the posterior probability that a perceived object falls into C_n is

$$P(C_n|V_i\&A_j) = \frac{P(V_i\&A_j|C_n)P(C_n)}{P(V_i\&A_j)}.$$

In all cases, finding the material category, C_n, of the highest posterior probability is an optimal Bayesian estimate (MAP estimate) for the material recognition. However, our material-rating task was designed such that the participants did not choose the best category consistent with the stimulus but judged to what degree the stimulus (V_i, A_j, or $V_i\,\&A_j$) could be thought of as a given material category (C_n). This is not an index of the subjective posterior probability because the participants did not have to take into account the prior probability of each category, $P(C_n)$. Rather, our material-category rating can be considered primarily as the participants' estimation of the likelihood of obtaining the evidence (V_i, A_j, or $V_i\&A_j$) given that the object falls in the material category C_n. In other words, $P(V_i|C_n)$, $P(A_j|C_n)$, and $P(V_i\&A_j|C_n)$ for the visual-only, auditory-only, and audiovisual conditions, respectively, although the estimated probability of each stimulus, $P(V_i)$, $P(A_j)$, or $P(V_i\&A_j)$, might have an additional effect, see below.

According to this view, the computational implication of the multiplicative integration rule is straightforward. If V_i and A_j are conditionally independent given C_n, then

$$P(V_i\&A_j|C_n) = P(V_i|C_n)P(A_j|C_n) \quad (1)$$

and

$$P(C_n|V_i\&A_j) = \frac{P(V_i|C_n|C_n)P(A_j|C_n)P(C_n)}{P(V_i\&A_j)}. \quad (2)$$

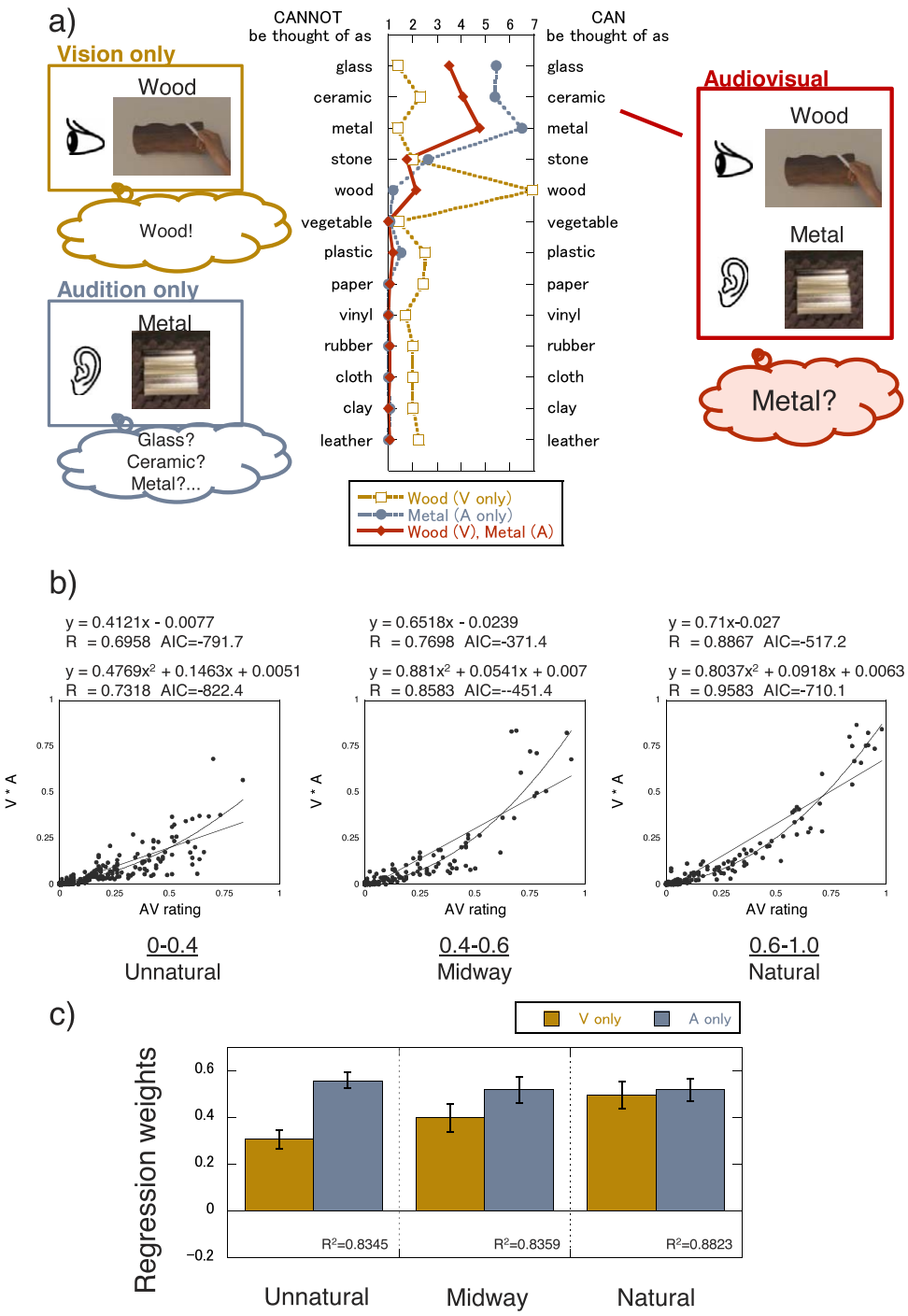

Figure 8. Examples of audiovisual integration in the material category–rating experiment (average of 16 participants) (a) when there was no category consistent with both visual-only and auditory-only stimuli. (b) Scatter plots for three different ranges of the naturalness rating. $V \times A$ strongly predicts the obtained audiovisual rating when the naturalness rating was high, but this relationship gradually collapsed as naturalness rating decreased. (c) The results of the regression analysis for unnatural, midway, and natural combinations. The regression weight of the auditory term relative to the visual term increased for unnatural combinations, supporting auditory dominance.

Equation 1 (conditional independence) is exactly what we found for the multiplicative integration. That is, our finding is consistent with Bayesian-like integration of visual and auditory information with the assumption of intermodality independency, in which the likelihoods of visual and auditory stimuli being consistent with a given material category are independently estimated and multiplied to obtain the likelihood

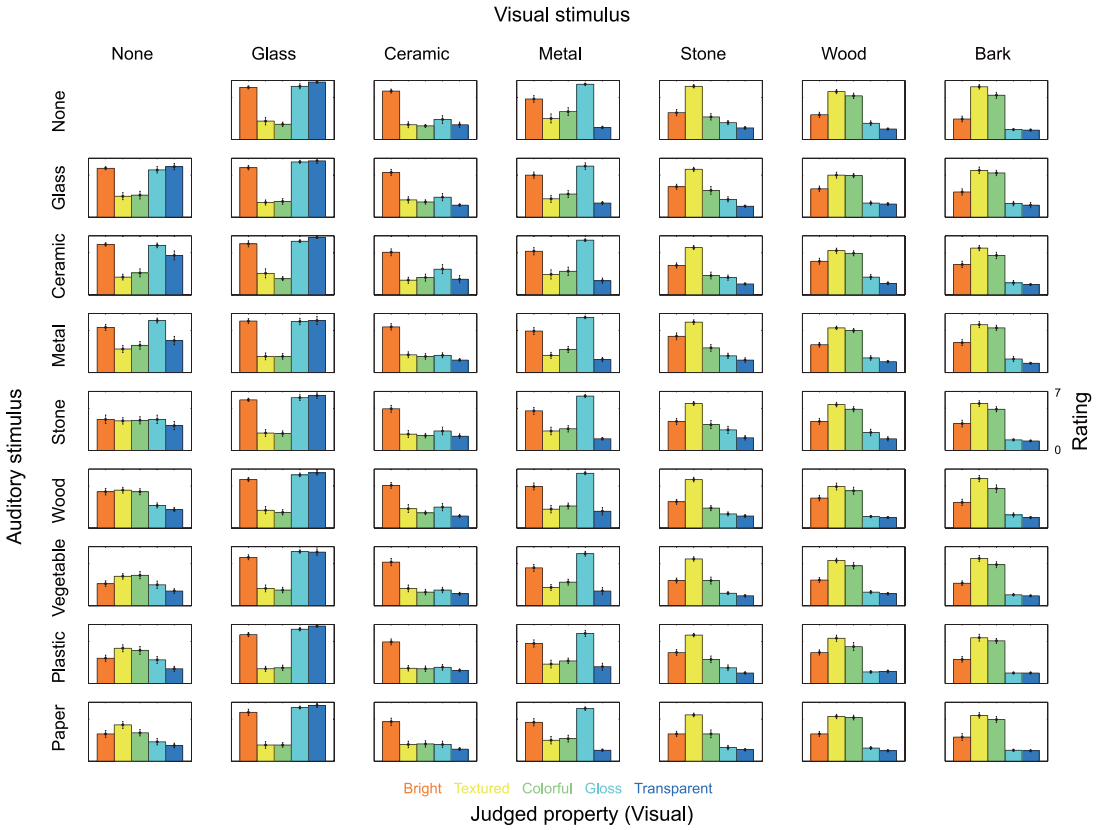

Figure 9. Visual property ratings for all 63 stimulus conditions. See Figure 4 legend for other details.

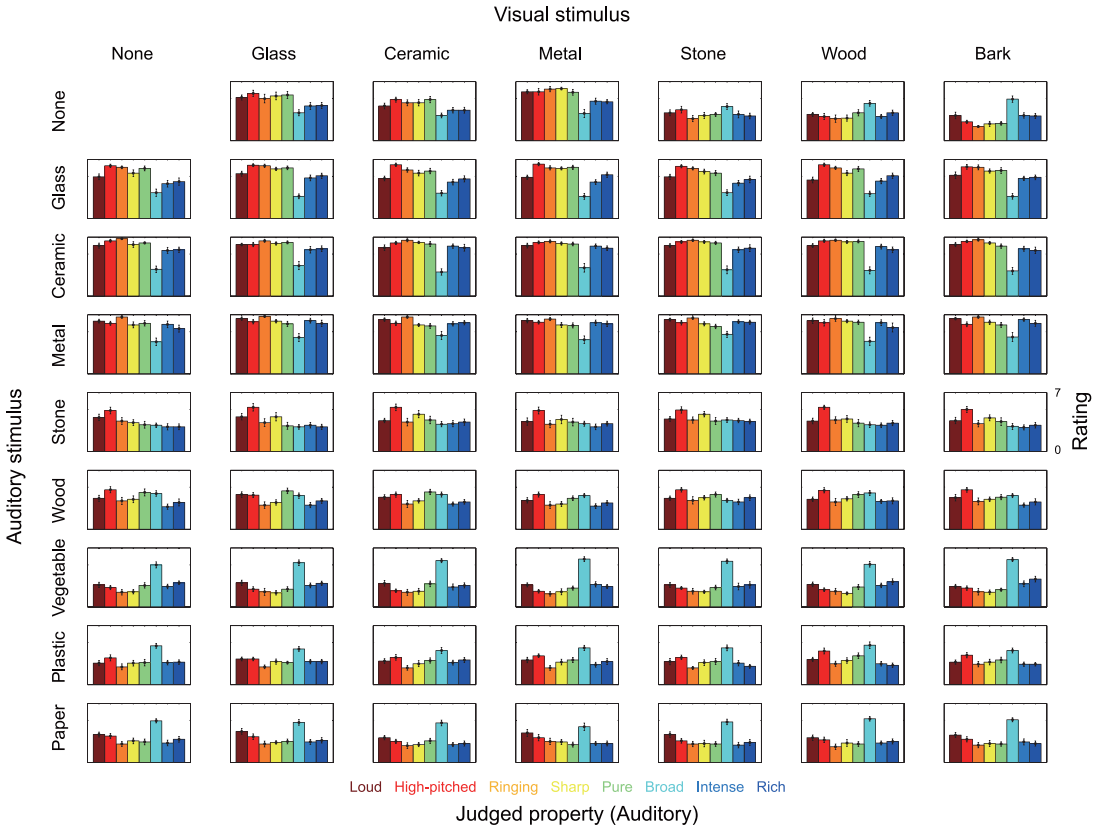

Figure 10. Auditory property ratings for all 63 stimulus conditions. See Figure 4 legend for other details.

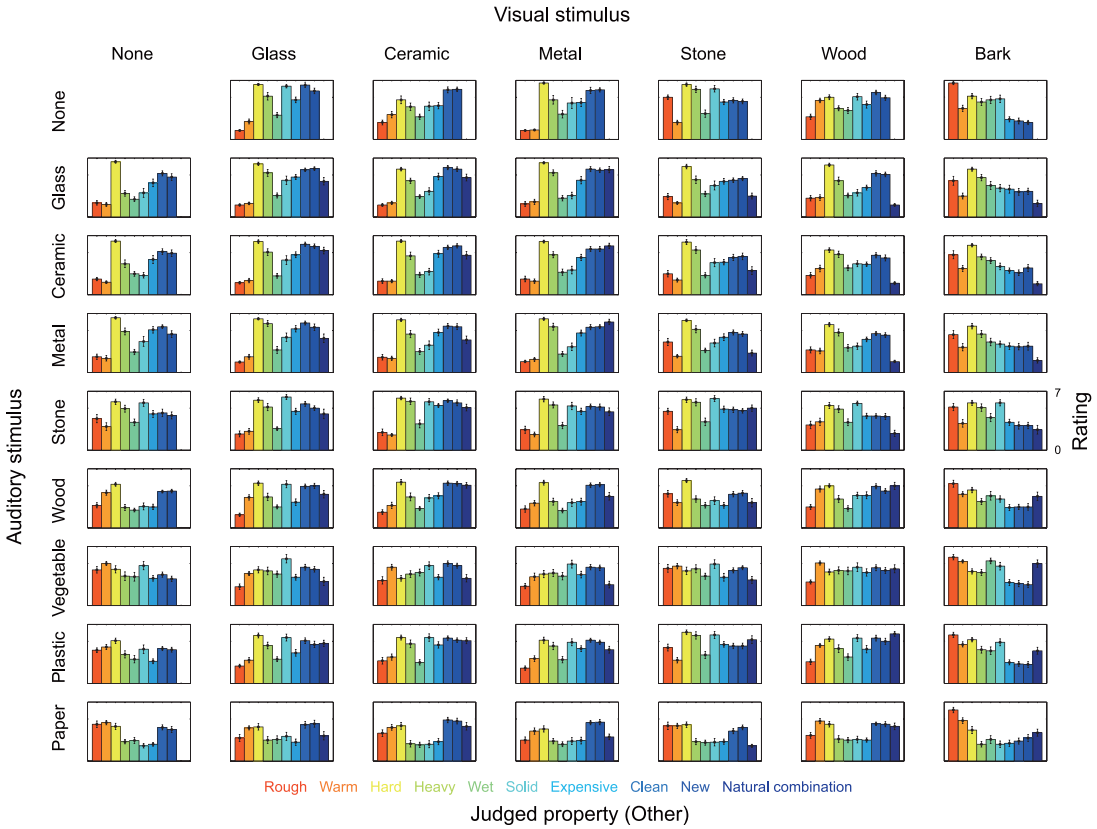

Figure 11. Other property ratings for all 63 stimulus conditions (including combination naturalness rating for 48 audiovisual stimulus conditions). See Figure 4 legend for other details.

of the audiovisual combination being consistent with that material category.

The results of the material category rating also suggest that the multiplication rule does not hold for unnatural audiovisual combinations. When there is no plausible candidate resulting from a multiplication of audio and visual information, the auditory evidence appears to dominate even when visual evidence seems to suggest otherwise. In many cases, integration of two sources breaks down, and one source begins to dominate when there is a large discrepancy between them (Landy et al., 1995; Ernst, 2012). In the present case, the auditory dominance may be interpreted as resulting from the perception of objects as being multilayered (i.e., objects could have a surface coating different from their underlying material). In the example shown in Figure 8, the participants could have judged that the object was "metal" but the surface was painted like "wood." This perceptual interpretation is unnatural but logically possible. We consider that the participants know, explicitly or implicitly, that visual material judgments are deceptive and auditory material judgments are harder to fake and that this asymmetry is the source of auditory dominance. In the Bayesian framework, this situation can be described as $P(V_i|C_n) > 0$ for nearly any combinations of visual appearance and material category, and $P(A_j|C_n) \approx 0$ for many combinations of impact sound and material category: The category consistent with the auditory stimulus has a nonzero likelihood even when it is inconsistent with the visual stimulus. This mechanism, however, is insufficient to explain the present results because visual-only ratings suggest $P(V_i|C_n) \approx 0$ for many cases, including "glass" and "metal" ratings for "wood" appearance. We speculate that an unnatural audiovisual combination forced observers to switch the mental object model from a single-layered object to a multilayered one and that this object model switch was accompanied by an increase in the internal estimation of $P(V_i|C_n)$.

In sum, audiovisual material-category perception can be explained by a mechanism in which visual and auditory likelihoods are independently estimated and multiplied to yield the audiovisual estimation. The fuzzy logical model of perception (Massaro, 1987, 2004; Massaro & Stork, 1998) proposed a similar mechanism to account for audiovisual integration of speech signals, including emotional expression and the McGurk effect (McGurk & MacDonald, 1976). Critically, all of these tasks comprise category classification. The integration rule we found for the material-category perception may

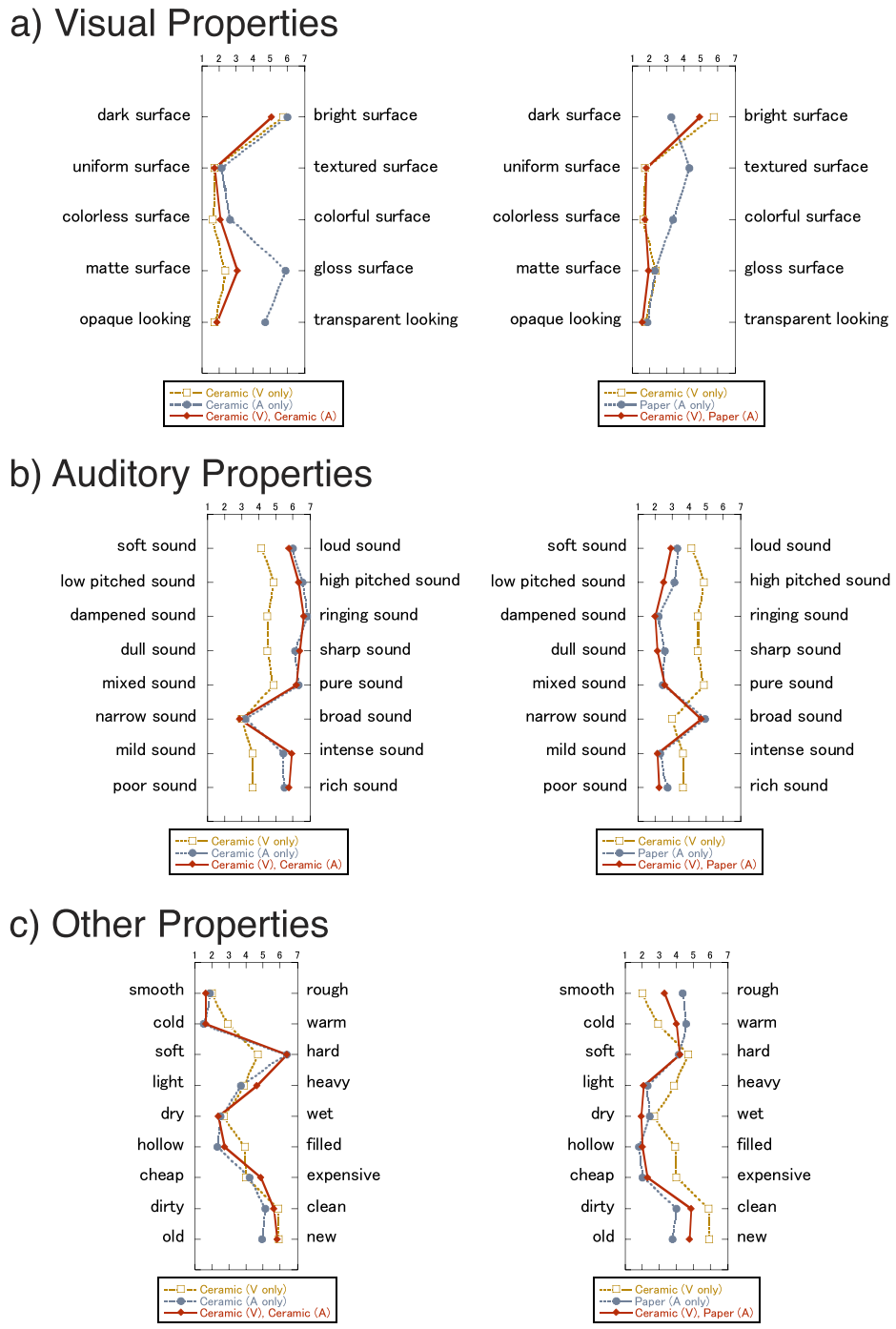

Figure 12. Example of response profiles (average of 16 participants) for (a) visual (e.g., dark-bright, uniform-textured) (b) auditory (e.g., quiet-loud, low-pitched–high-pitched), and (c) other property ratings (e.g., smooth-rough, cold-warm).

be a universal rule of multimodal integration for categorical judgments.

Material-property rating

On the other hand, the material property rating falls into the other class of multimodal integration task involved in estimating the value of a given perceptual dimension. Redundant estimations are obtained from two modalities, and the participant has to integrate them to obtain the final estimation. For this class of task, the optimal integration is the weighted average of the two estimations with the weight of each modality being proportional to the reliability of the signal (Cochran, 1937; Landy et al., 1995; Yuille & Bülthoff, 1996). Previous studies indicated that human participants actually integrate multimodal signals in this

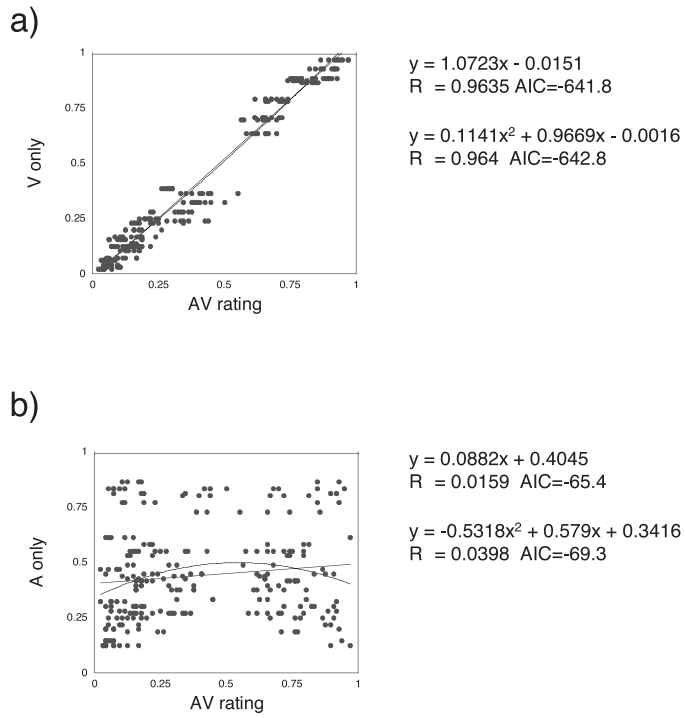

Figure 13. Scatter plots of the normalized visual property ratings for audiovisual stimuli versus (a) those for visual-only stimuli and (b) those for auditory-only stimuli. The ratings for visual-only stimuli best describe the ratings for audiovisual stimuli.

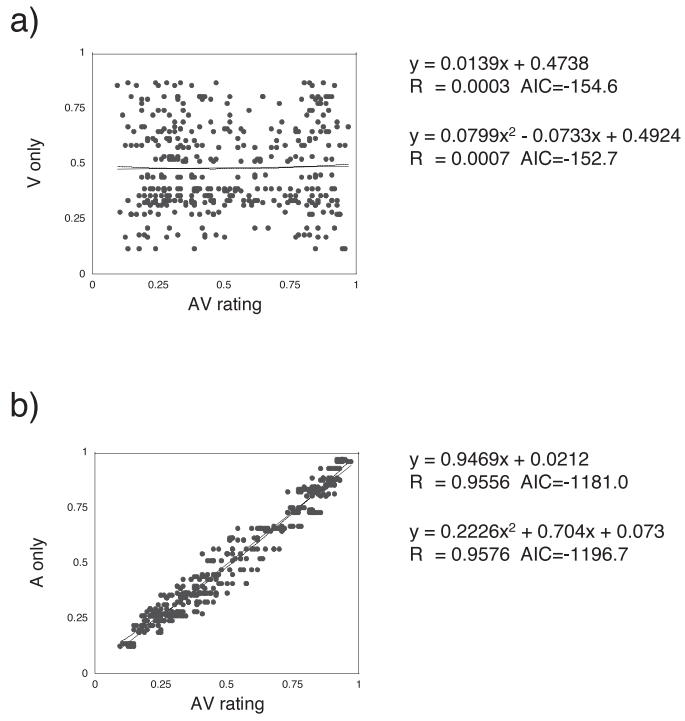

Figure 14. Scatter plots of the normalized auditory property ratings for audiovisual stimuli versus (a) those for visual-only stimuli and (b) those for auditory-only stimuli. The ratings for auditory-only stimuli best describe the ratings for audiovisual stimuli.

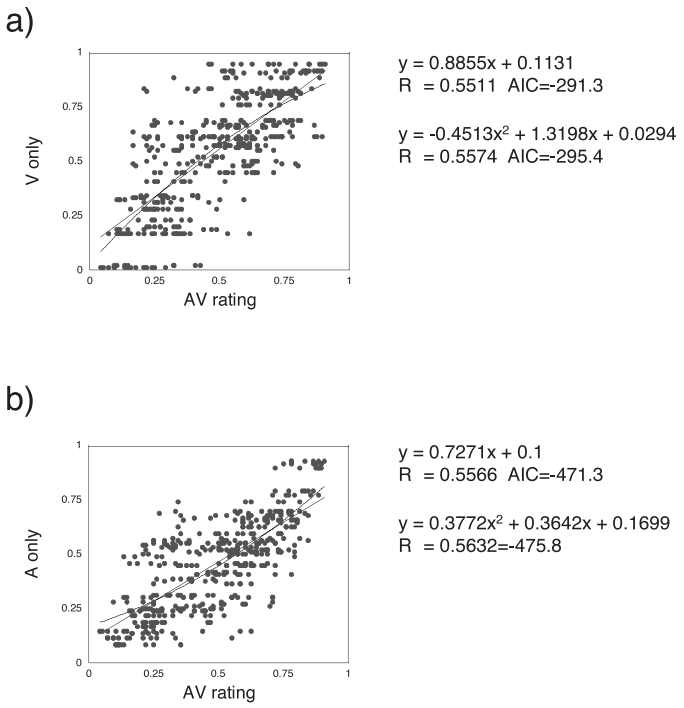

Figure 15. Scatter plots of the normalized other property ratings for audiovisual stimuli versus (a) those for visual-only stimuli and (b) those for auditory-only stimuli. Both visual-only and auditory-only ratings have moderate correlations with the other property ratings obtained with audiovisual stimuli, consistent with the notion that it can be best described by the weighted average of visual and auditory ratings.

manner in various situations (Alais & Burr, 2004; Ernst & Banks, 2002).

The weighted average is an optimal Bayesian estimation, again with the assumption of intermodality independency (Ernst, 2006, 2012). It can be derived from an equation similar to Equation 2, that is,

$$P(X_V \& X_A | V \& A) = \frac{P(V|X_V)P(A|X_A)P(X_V \& X_A)}{P(V \& A)},$$

where X_v and X_A are the estimated values of the property in question from the visual evidence, V, and the auditory evidence, A, respectively. Assume that the likelihoods $P(V|X_V)$ and $P(A|X_A)$ form a 2-D Gaussian distribution with its vertical and horizontal spreads being inversely proportional to the reliabilities of the visual and auditory inputs and that the prior, $P(X_V \& X_A)$, is nonzero only when $X_V = X_A$. Then, the peak of the product, which indicates the MAP estimate, agrees with the weighted average.

In agreement with this optimal integration strategy, our participants gave higher weights to the more reliable modality, for example, vision for color, audition for pitch and hardness. However, to strictly test the Bayesian integration with modality independency in material-property judgments, we should

systematically manipulate the simulated property variable and signal reliability.

Thus, the multiplicative integration of the material-category judgments and the weighted average of the material-property judgments share a computation principle, that is, optimal Bayesian integration of independent visual and auditory signals. The critical task difference is that the material-category task rates the likelihood, and the property task rates the value along the designated dimension. If the participants rate the likelihood of a stimulus as having a specific property value (e.g., very smooth), they may show a multiplicative integration rule.

While we used real sounds for auditory stimuli, we used computer-generated images for visual stimuli so as to precisely control the shape and size of the object and the hitting movement of the operator. Nonetheless, we cannot exclude the possibility that participants distrusted the synthetic visual stimuli and reduced the weight given to them. The relative audiovisual weights of our data should be interpreted as such. Nevertheless, we find no good reason to believe that the use of synthetic visual stimuli affects the rules of integration.

Multimodal material perception

In conclusion, multiplicative integration of multimodal signals is an effective cognitive strategy, particularly when different modalities carry independent information about different aspects of the same recognition target, thus complementing each other. This is presumably the case for audiovisual integration of material-category information, in which vision is a useful modality for understanding the surface properties of an object, and audition is a useful modality for understanding its internal properties (R. Klatzky et al., 2000).

We did not examine the tactile modality in multimodal material-category perception. Because vision provides information about surface properties and audition provides information about internal properties—and touch provides information about both—it would be interesting to test, in the future, whether visuo-tactile or auditory-tactile category perceptions also follow the integration rule.

We note that we examined material-category and property perception separately. However, material-category and property judgments are likely related (Fleming et al., 2013). In the present data, although the rule of integration was different, there were no inconsistencies between material-category ratings and property ratings. For instance, a "glass" (V) × "vegetable" (A) combination that yielded a "plastic" perception (Movie 2) was judged to be a transparent object with a low-pitched impact sound, being neither warm nor hot and neither soft nor hard. Material-property perception could be the precursor to material-category perception; in other words, audiovisual material-category perception could arise directly from the integration of auditory and visual material-property judgments rather than auditory and visual material-category judgments. That is, the brain may combine visual judgments, such as an object being glossy and transparent, with auditory judgments, such as the impact sound being high-pitched and sharp, and then reason that the object is likely to be glass. If the mapping from visual and auditory properties to material category follows an AND rule, the present results are consistent with this possibility as well. In addition, material-category judgments might affect some material-property judgments. For instance, if an audiovisual combination indicates that the object is made of glass, it might be judged more fragile than what could be perceived with only visual or auditory information. These arguments, however, remain speculative. The relationship between the two types of material judgments is a fundamental problem that awaits further study.

Keywords: material perception, audio-visual integration, Bayesian integration, surface texture, impact sound

Acknowledgments

This work was supported by Grant-in-Aid for Scientific Research on Innovative Areas (Nos. 22135004, 22135007) from the Ministry of Education, Science, Culture, Sports and Science, Japan. We thank Junichi Nakamura for help in stimulus generation and Satohiro Tajima for the helpful comments.

Commercial relationships: none.
Corresponding author: Waka Fujisaki.
Email: w-fujisaki@aist.go.jp.
Address: Human Technology Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), Japan.

References

Adelson, E. H. (2001). On seeing stuff: The perception of materials by humans and machines. *Proceedings of the SPIE, Vol. 4299, Human Vision and Electronic Imaging VI*, San Jose, CA.

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology, 14*(3), 257–262.

Anderson, B. L. (2011). Visual perception of materials and surfaces. *Curr Biol*, *21*(24), R978–R983.

Aramaki, M., Besson, M., Kronland-Martinet, R., & Ystad, S. (2011). Controlling the perceived material in an impact sound synthesizer. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(2), 301–314.

Arnott, S. R., Cant, J. S., Dutton, G. N., & Goodale, M. A. (2008). Crinkling and crumpling: An auditory fMRI study of material properties. *NeuroImage*, *43*(2), 368–378.

Cant, J. S., & Goodale, M. A. (2007). Attention to form or surface properties modulates different regions of human occipitotemporal cortex. *Cerebral Cortex*, *17*(3), 713–731.

Cant, J. S., & Goodale, M. A. (2011). Scratching beneath the surface: New insights into the functional properties of the lateral occipital area and parahippocampal place area. *The Journal of Neuroscience*, *31*(22), 8248–8258.

Cavina-Pratesi, C., Kentridge, R. W., Heywood, C. A., & Milner, A. D. (2010a). Separate channels for processing form, texture, and color: Evidence from fMRI adaptation and visual object agnosia. *Cerebral Cortex*, *20*(10), 2319–2332.

Cavina-Pratesi, C., Kentridge, R. W., Heywood, C. A., & Milner, A. D. (2010b). Separate processing of texture and form in the ventral stream: Evidence from fMRI and visual agnosia. *Cerebral Cortex*, *20*(2), 433–446.

Cochran, W. (1937). Problems arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society (Supplement)*, *4*(1), 102–118.

Cunningham, D., Wallraven, C., Fleming, R., & Strasser, W. (2007). *Perceptual reparameterization of material properties. Proceedings of the Computational Aesthetics in Graphics, Visualization, and Imaging*, Banff, Alberta, Canada.

Doerschner, K., Boyaci, H., & Maloney, L. (2010). Estimating the glossiness transfer function induced by illumination change. *Journal of Vision*, *10*(4):8, 1–9, http://www.journalofvision.org/content/10/4/8, doi:10.1167/10.4.8. [PubMed] [Article]

Ernst, M. O. (2006). A Bayesian view on multimodal cue integration. In G. Knoblich, I. M. Thornton, M. Grosjean, & M. Shiffrar (Eds.), *Human body perception from the inside out* (pp. 105–131). Oxford, UK: Oxford University Press.

Ernst, M. O. (2012). Optimal multisensory integration: Assumptions and limits. In B. E. Stein (Ed.), *The new handbook of multisensory processes* (pp. 1084–1124). Cambridge, MA: MIT Press.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433.

Fleming, R. W. (2014). Visual perception of materials and their properties. *Vision Research, 94,* 62–75.

Fleming, R. W., Wiebel, C., & Gegenfurtner, K. (2013). Perceptual qualities and material classes. *Journal of Vision*, *13*(8):9, 1–20, http://www.journalofvision.org/content/13/8/9, doi:10.1167/13.8.9. [PubMed] [Article]

Fujisawa, N., Iwamiya, S., & Takada, M. (2004). Auditory imagery associated with Japanese onomatopoeic representation. *Journal of Physiological Anthropology and Applied Human Science*, *23*(6), 351–355.

Gabrielsson, A., & Sjogren, H. (1979). Perceived sound quality of sound-reproducing systems. *Journal of the Acoustical Society of America*, *65*(4), 1019–1033.

Giordano, B. L., & McAdams, S. (2006). Material identification of real impact sounds: Effects of size variation in steel, glass, wood, and plexiglass plates. *Journal of the Acoustical Society of America*, *119*(2), 1171–1181.

Goda, N., Tachibana, A., Okazawa, G., & Komatsu, H. (2014). Representation of the material properties of objects in the visual cortex of non-human primates. *Journal of Neuroscience*, *34*(7), 2660–2673.

Hiramatsu, C., Goda, N., & Komatsu, H. (2011). Transformation from image-based to perceptual representation of materials along the human ventral visual pathway. *Neuroimage*, *57*(2), 482–494.

Kim, J., Marlow, P. J., & Anderson, B. L. (2012). The dark side of gloss. *Nature Neuroscience*, *15*(11), 1590–1595.

Klatzky, R., Pai, D., & Krotkov, E. (2000). Perception of material from contact sounds. *Presence: Teleoperators & Virtual Environments*, *9*(4), 399–410.

Klatzky, R. L., & Lederman, S. J. (2010). Multisensory texture perception. In J. Kaiser & M. J. Naumer (Eds.), *Multisensory object perception in the primate brain* (pp. 211–230). New York: Springer.

Landy, M. S., Maloney, L. T., Johnston, E. B., & Young, M. (1995). Measurement and modeling of depth cue combination: In defense of weak fusion. *Vision Research*, *35*(3), 389–412.

Lederman, S. J., Thorne, G., & Jones, B. (1986). Perception of texture by vision and touch: Multidimensionality and intersensory integration. *Journal of Experimental Psychology-Human Perception and Performance, 12*(2), 169–180.

Lemaitre, G., & Heller, L. M. (2012). Auditory

perception of material is fragile while action is strikingly robust. *Journal of the Acoustical Society of America*, *131*(2), 1337–1348.

Lutfi, R. A., & Oh, E. L. (1997). Auditory discrimination of material changes in a struck-clamped bar. *Journal of the Acoustical Society of America*, *102*(6), 3647–3656.

Maloney, L. T., & Brainard, D. H. (2010). Color and material perception: Achievements and challenges. *Journal of Vision*, *10*(9):19, 1–6, http://www.journalofvision.org/content/10/9/19, doi:10.1167/10.9.19. [PubMed] [Article]

Massaro, D. (1987). Speech perception by ear and eye: A paradigm for psychological inquiry. Hillsdale, NJ: Erlbaum.

Massaro, D. (2004). From multisensory integration to talking heads and language learning. In G. Calvert, C. Spence, & B. E. Stein (Eds.), *Handbook of multisensory processes* (pp. 153–176). Cambridge, MA: MIT Press.

Massaro, D., & Stork, D. (1998). Speech recognition and sensory integration: A 240-year-old theorem helps explain how people and machines can integrate auditory and visual information to understand speech. *American Scientist, 86*(3), 236–244.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748.

Motoyoshi, I., Nishida, S., Sharan, L., & Adelson, E. H. (2007). Image statistics and the perception of surface qualities. *Nature*, *447*(7141), 206–209.

Nishida, S., & Shinya, M. (1998). Use of image-based information in judgments of surface-reflectance properties. *The Journal of the Optical Society of America A, Optics, Image Science, and Vision*, *15*(12), 2951–2965.

Nishio, A., Goda, N., & Komatsu, H. (2012). Neural selectivity and representation of gloss in the monkey inferior temporal cortex. *The Journal of Neuroscience, 32*(31), 10780–10793.

Osgood, C. E., & Anderson, L. (1957). Certain relations among experienced contingencies, associative structure, and contingencies in encoded messages. *American Journal of Psychology*, *70*(3), 411–420.

Sharan, L., Liu, C., Rosenholtz, R., & Adelson, E. (2013). Recognizing materials using perceptually inspired features. *International Journal of Computer Vision, 103,* 348–371.

Sharan, L., Rosenholtz, R., & Adelson, E. H. (2009). Material perception: What can you see in a brief glance? *Journal of Vision, 9*(8): 784, http://www.journalofvision.org/content/9/8/784, doi:10.1167/9.8.784. [Abstract]

Solomon, L. (1958). Semantic approach to the perception of complex sounds. *The Journal of the Acoustical Society of America*, *30*(5), 421–425.

von Bismarck, G. (1974a). Sharpness as an attribute of the timbre of steady sounds. *Acustica, 30,* 159–172.

von Bismarck, G. (1974b). Timbre of steady sounds: A factorial investigation of its verbal attributes. *Acustica, 30,* 146–159.

Wildes, R. P., & Richards, W. A. (1988). Recovering material properties from sound. In W. A. Richards (Ed.), *Natural computation* (pp. 356–363). Cambridge, MA: The MIT Press.

Wijntjes, M. W. A., & Pont, S. C. (2010). Illusory gloss on Lambertian surfaces. *Journal of Vision*, *10*(9):13, 1–12, http://www.journalofvision.org/content/10/9/13, doi:10.1167/10.9.13. [PubMed] [Article]

Yuille, A., & Bülthoff, H. (1996). Bayesian theory and psychophysics. In D. Knill & W. Richards (Eds.), *Perception as Bayesian inference* (pp. 123–161). New York: Cambridge University Press.

Zaidi, Q. (2011). Visual inferences of material changes: Color as clue and distraction. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(6), 686–700.

Appendix: Preliminary experiment for auditory stimulus selection

Participants were 16 university students who were blind to the purpose of the experiment. None of them participated in the main experiment. Auditory stimuli used were impact sounds of 16 real objects of 10 material categories (glass, ceramic, metal, stone, wood, vegetable, fruit, plastic, leather, and paper) (Figure 2a). The stimuli were presented to the participants in a classroom through a set of speakers. Each sound was repeated 15 times, and the participants had to make the material-category rating for each of the 13 categories (Table 1a) by the end of the stimulus repetition. The rating task required a seven-point likelihood rating, similar to that used in the main experiment but for participants scoring the ratings on a questionnaire sheet. From multidimensional scaling of the squared Euclidean distance of rating vectors, we estimated the subjective similarity map of the 16 sounds (Figure A1). We then selected one sound from each of eight different material categories (fruit and leather categories were dropped). The eight selected sounds (indicated by square dotted lines in Figure 2a) were broadly distributed on the similarity map although some of them were fairly close to each other (e.g., glass and ceramic).

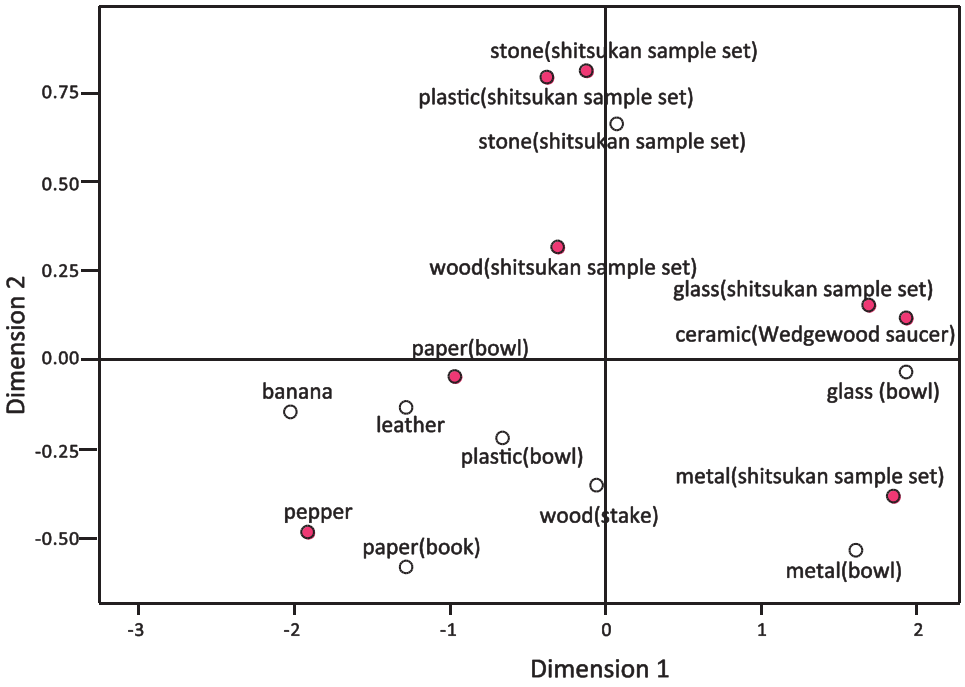

Figure A1. Subjective similarity map of the 16 sounds estimated from multidimensional scaling of the squared Euclidean distance of rating vectors.