Bimodal perception of audio-visual material properties for virtual environments

Nicolas Bonneel[1]
REVES/INRIA Sophia-Antipolis

and
Clara Suied
and
Isabelle Viaud-Delmon
CNRS-UPMC UMR 7593

and
George Drettakis
REVES/INRIA Sophia-Antipolis

High-quality rendering of both audio and visual material properties is very important in interactive virtual environments, since convicingly rendered materials increase realism and the sense of immersion. We studied how the level of detail of auditory and visual stimuli interact in the perception of audio-visual material rendering quality. Our study is based on perception of material discrimination, when varying the levels of detail of modal synthesis for sound, and Bidirectional Reflectance Distribution Functions for graphics. We performed an experiment for two different models (a Dragon and a Bunny model) and two material types (Plastic and Gold). The results show a significant interaction between auditory and visual level of detail in the perception of material similarity, when comparing approximate levels of detail to a high-quality audio-visual reference rendering. We show how this result can contribute to significant savings in computation time in an interactive audio-visual rendering system. To our knowledge this is the first study which shows interaction of audio and graphics representation in a material perception task.

Categories and Subject Descriptors: I.3.3 [**Computer Graphics**]: Picture/Image

General Terms: Algorithms, Experiments

Additional Key Words and Phrases: Audio-visual rendering, perception, crossmodal

1. INTRODUCTION

Interactive audio-visual virtual environments are now commonplace, ranging from computer games with high-quality graphics and audio, to virtual environments used for train-

ing, car and flight simulation, rehabilitation, therapy etc. In such environments, synthetic objects have audio-visual material properties, which are often based on physical measurements of real objects. Realistic, high-quality rendering of these materials is a central element for the overall realism and the sense of immersion offered by such virtual environments: This is true both for graphics and audio. Real-time or interactive performance is central to such systems. One way these systems handle the ever-increasing complexity of the graphics and the sounds is to use *level-of-detail* (LOD) rendering. This approach consists in rendering lower quality versions of entities in the virtual environment, which require lower computation time. As a result, more complex environments can be rendered. In what follows, we will use the general term *material* to mean the audio-visual material properties which are physically measurable. The goal of our study is twofold. First we ask whether audio and graphics mutually interact in the perception of material rendering quality, and in particular when independantly varying the LOD for both audio and graphics in an interactive rendering context. Second, if such an interaction exists, we want to see whether it can be exploited to improve overall interactive system performance. Therefore, we hope both to identify a perceptual effect of the influence of audio and graphics on material perception *and* to achieve algorithmic gain.

Given the interactive audio-visual context of our work, we will concentrate on choices of stimuli which are feasible in the context of such systems. This inevitably leads to the use of approximations to create LOD for both audio and graphics, so that practical algorithmic benefit can be achieved. In the virtual environments of this study audio is rendered in realtime and graphics rendering runs at 29 frames per second.

To our knowledge, no study exists on the mutual interaction of audio and graphics on material perception. Nonetheless, earlier work [Storms and Zyda 2000] has found some improvement in overall perception of visual image quality in the presence of better sound.

This experimental study of static images and sounds showed that the perceived quality of a high quality visual display evaluated alone was enhanced when coupled with high quality sound. The study further showed that the perceived quality of a low quality auditory display evaluated alone was reduced when coupled with a high quality visual display. Visual degradations were varied by resampling images or adding noise, while audio degradations was varied by changing sampling rates or by adding Gaussian noise.

We have designed an experiment to evaluate whether there is a mutual influence of audio and graphics on the perception of materials. Since we are interested in optimizing the perception of material quality in an interactive rendering setting, we chose to perform a material similarity experiment (see [Klatzky et al. 2000] for a similar experiment on material perception of contact sounds for audio only).

Stimuli vary along two dimensions: graphic LOD and audio LOD. Participants are asked to compare these to a hidden audio-visual reference. This reference is rendered at the highest possible quality given the constraints of the interactive system. Stimuli are synthetic objects falling onto a table. Audio for contact sounds is provided by using modal synthesis [van den Doel and Pai 2003], and LOD result from choosing a progressively larger number of modes [2]. Graphics are rendered using an environment map and Bidirectional Reflection Distribution Functions (BRDF) [Cook and Torrance 1982]. A BRDF describes how a material reflects light, and can be measured from real materials [Matusik et al. 2003].

[2]Please watch and listen the accompanying video in which we have captured the entire set of audio-visual LOD for a Gold Bunny and a Plastic Dragon.

To provide visual LOD, we project the BRDF onto a Spherical Harmonic basis [Kautz et al. 2002], and increase the number of coefficients to obtain progressively better visual quality. Increasing the number of modes or spherical harmonic coefficients in our LOD improves the mathematical approximation, i.e., the error of the approximations with respect to the high-quality reference diminishes. This is illustrated in the accompanying video.

Results of the present experiment show that this also results in better perceived quality for each of audio and visuals independantly. Interestingly, we also show that for this context there is a mutual influence of sound and graphics on the perception of material similarity. We highlight how this result can be directly used to significantly improve overall rendering performance in an interactive audio-visual system. To our knowledge, this study is the first to demonstrate a combined effect of graphics and audio on a task related to material perception.

2. METHODS

2.1 Participants

Ten participants (7 men) from 23 to 46 years old (mean age 30.8 years, Standard deviation: 7.7 years) participated in the experiment. All had normal or corrected to normal vision and all reported normal hearing. All were naive to the purpose of the experiment.

2.2 Stimuli

We next present a detailed description of both visual and auditory stimuli as presented to the participants, and the corresponding LOD mechanisms used to create the stimuli.

2.2.1 *Visual LOD.* In order to make our method applicable for realtime rendering, we interactively render realistic materials with measured BRDF. This rendering can then be used to generate visual stimuli for the experiment and is also usable in the context of interactive audio-visual applications (computer games, virtual environments, audiovisual simulations etc.). Using realtime rendering in the experiment simplifies the potential application of the results in interactive systems, since the conditions are the same. To achieve realtime rendering in complex environments, various approximation have been proposed which include infinite light sources [Ramamoorthi and Hanrahan 2002; Kautz et al. 2002], static viewpoint, [Ben-Artzi et al. 2006] and/or static geometry [Sloan et al. 2002; Kristensen et al. 2005]. We assume infinite light sources through the use of an environment map which gives the illumination from distant sources (eg., a panoramic photograph of the sky). This approximation is reasonable since the environment map was captured at the true location of the object and the motion of the object is not large compared to the size of the environment.

A commonly used method to interactively render measured BRDFs with environment maps is the projection of the BRDF or visibility and the environment map into a set of basis functions [Kautz et al. 2002]. This is performed by computing the scalar product of the BRDF and each basis element. Rendering is performed by computing the dot product of these coefficients. Choices of the basis functions include Spherical Harmonics (SH) [Ramamoorthi and Hanrahan 2002; Kautz et al. 2002; Green 2003; Kristensen et al. 2005; Sloan et al. 2002], Wavelets [Ng et al. 2003], Zonal harmonics [Sloan et al. 2005] or any other orthogonal basis.

We have chosen BRDF rendering with SH projection because it allows a relatively

smooth increase of material quality when increasing the number of coefficients (i.e., number of basis functions used in the calculation). Additionally, the increase in number of coefficients is directly related to the specularity (or glossiness) of the material: higher degree SH basis functions correspond to higher frequencies and thus well represent glossier materials or lighting.

In what follows, we will assume (θ_i, ϕ_i) to be the incoming light direction, (θ_o, ϕ_o) the outgoing view direction, Y_k the k'th basis function, f_k the projection of the function onto a SH basis function, and N the number of SH bands.

Spherical Harmonics form a basis of spherical functions. A BRDF can thus be approximated by the sum of N^2 of these function bases as:

$$f(\theta_i, \phi_i, \theta_o, \phi_o) \cos\theta_i = \sum_{k=1}^{N^2} Y_k(\theta_i, \phi_i) f_k(\theta_o, \phi_o)$$

The environment map can also be decomposed into N^2 SH :

$$L(\theta_i, \phi_i) = \sum_{k=1}^{N^2} Y_k(\theta_i, \phi_i) L_k,$$

where L_k is the coefficient of lighting for the kth basis function. Because of SH orthogonality, rendering a point x consists in computing the dot product to find the outgoing radiance $L(x)$:

$$L(x) \;=\; \sum_{k=1}^{N^2} f_k(\theta_o, \phi_o) L_k$$

Note that, following standard practice, we included the $\cos\theta_i$ term which takes into account the attenuation due to incident angle into the BRDF without loss of generality.

In practice, the coordinate system (CS) of the environment map (world CS) has to be aligned with the BRDF CS (local CS). Thus, to efficiently render the scene, we precomputed an environment map with SH rotations into a $128 * 128 * N^2$ cubemap containing multiple rotations of the environment map's SH, and a $128 * 128 * N^2$ cubemap for the BRDF containing SH for multiple outgoing directions. This method is similar to [Kautz et al. 2002], except for SH rotations which are all precomputed and tabulated for efficiency. We did not take visibility into account since our stimuli consisted in a single object falling with no occlusion of light coming from the environment map. Self occlusion with respect to the environment map was also negligible.

For visual rendering, previous studies [Fleming et al. 2003] show that using natural outdoors illumination of the object can aid in material perception. We thus chose a configuration (outdoor summer scene, with occlusion of the sky by trees) where light had relatively high frequencies to avoid impairing material appearance by having a "too diffuse" look. We acquired a High Dynamic Range (HDR) environment map and integrated the stimuli into a HDR photo consistent with the environment map, with a method similar to ([Debevec 1998]). Shadows were computed with a Variance Shadow Map ([Donnelly and Lauritzen 2006]).

We also chose glossy materials to be able to get a sufficient number of levels of detail when increasing the number of SH basis functions. Lambertian surfaces are already very well approximated with 3 SH bands [Ramamoorthi and Hanrahan 2001b; 2001a].

Rendering was performed using deferred shading to floating point render targets (High Dynamic Range rendering) and Reinhard et al's global tonemapping operator [Reinhard et al. 2002] was applied to account for low dynamic light intensity range of monitors and human eye sensitivity. Interactive rendering (29 frames per second (fps)) was achieved for up to 12 SH bands.

The visual rendering time was kept constant between LODs by adding idle loops in order to slow down low visual LODs to avoid subjects being disturbed by varying framerate.

2.2.2 *Sound LOD.* Contact sounds of rigid objects can be realistically generated in several ways. The context of interactive audio-visual rendering precludes the use of physical models such as the one used in [McAdams et al. 2004] applicable for bars only or recorded sounds in [Giordano and Mcadams 2006]. In contrast, tetrahedral finite elements methods provide an accurate simulation of object deformations [O'Brien et al. 2002], for complex object shapes such as those used in computer games. The method is used to solve the linear elasticity problem of objects of general shapes, under small deformations (Hooke's law) which is suitable for vibrating objects. This approach results in a set of vibrational modes which are excited with a force at each contact. Each mode results in an audio stream, given as a sine wave of the modes' frequency modulated by an exponential decay and a constant amplitude. In this way, computing a contact sound $s(t)$ over time t consists in computing a sum of N modes:

$$s(t) = \sum_{n=1}^{N} a_n e^{-\alpha_n t} \sin(\omega_n t)$$

where a_n is the mode amplitude which is computed in realtime, α_n is the decay (in seconds^{-1}) which indicates how long the sound of mode will last, and ω_n is the frequency (in radians per second). Sound radiation amplitudes of each mode were estimated using a far-field radiation model (see Eq. 15 in [James et al. 2006]).

Varying the sound LOD consists in varying the number of excited modes N, or *mode culling* [Raghuvanshi and Lin 2007]. In our case, we order modes by energy [Bonneel et al. 2008]. We found that this ordering provided good quality sounds, in particular when small numbers of modes are used. A pilot experiment was performed for a given set of mode budgets, and the best sounding values were selected. This pilot experiment also guided our choice of sorting by energy compared to other possible orderings (e.g., by amplitude [Doel et al. 2002]).

2.2.3 *Comparison of audio and visual stimuli.* Both SH and Modal synthesis refer to a projection of a scalar field (the directional reflectance, the incident radiance and the displacement of each node of the tetrahedral mesh) into a set of functional basis. The common point of these bases is that they both refer to the eigenvalues of a Laplacian operator either over a sphere (which gives Spherical Harmonics) or over the mesh (which gives vibrational modes). Thus, they both lead to the same type of reconstruction errors: fewer basis functions results in a smoother reconstructed function, if basis functions are sorted by their frequency (mode frequency or SH band). The combined choice of these two methods for audio and visual is thus consistent.

2.2.4 *Objects, materials and LOD choices.* Shapes were carefully chosen to facilitate material recognition. We use two objects identified by the study of [Vangorp et al. 2007]: the Bunny and the Dragon (see Fig. 1, 2). According to this study, both of these shapes

convey accurate perception of the material of the objects.

Material classification has been studied by [Giordano and Mcadams 2006] where subjects had to determine the material an object (wood, plexiglass, steel or glass) was made from. They show that two main categories of material were correctly classified (wood and plexiglass vs. steel and glass).

We thus used the following materials: Gold (similar to steel in that study) and plastic (similar to plexiglass).

Visual rendering was performed using measured BRDFs “gold-metallic-paint3" and “specular-green-phenolic" from the database of [Matusik et al. 2003].

We selected five different levels of visual quality, and five levels of sound quality. They were chosen so that perceptual degradations were as close as possible to uniformly distributed; note that given the discrete nature of the BRDF LOD, the choices were very limited. Some of the visual stimuli can be seen in the first two rows of Fig. 1 and 2.

The LOD used in the experiment correspond to budgets given in table 3. Budgets represent the number of modes mixed for sound, or the number of SH bands for graphics.

Given the interactive rendering context, the highest quality or “reference" solution is still an approximation. To verify how far these are from the “ground truth" we computed static offline references by sampling the rendering integral over the environment map. The use of SH stored in a spherical parametrization leads to distortion near the singularity, and given that we can handle at most 12 bands in realtime, the reference renderings are not exactly the same as the 12 bands stimuli which serve as references in the experiment. No particular treatment has been applied to limit ringing since it would result in a reduction of high frequencies (low pass filtering) which is undesirable for glossy materials. As noted by [Kautz et al. 2002], ringing is also masked by bumpy complex models.

Overall, as can be seen by comparing the middle and last rows of Fig. 1 and 2, the differences between our highest quality interactive rendering and the offline reference are overall acceptable. This comparison is provided to give some evidence that the highest quality interactive rendering is close to the true offline reference.

2.3 Procedure

In each trial, two sequences are shown to the participant. In each sequence an object falls onto a table and bounces twice. One of the two sequences is a reference (i.e., highest quality) rendering both in audio and graphics. The participant is unaware of this. The objects falls for 0.5 seconds, while the total length of the sound varies from 0.5 et 1.5 seconds starting at the time of the impact. The duration of the sound is of course shorter for plastic and longer for gold, and also depends on the object shape. Participants were asked to rate on a scale from 0 to 100 the perceived similarity of the materials of the two falling objects “A" and “B". Since the subjects rate similarity to a high-quality reference rendering of the material, this can also be seen as an implicit material quality test. Printed instructions were given before starting the experiment. An initial pair was presented separately to the participant showing the worst quality audio *and* visual with the highest quality audio *and* visual. Participants were told that this pair should be considered as the most different pair and they were explicitly told that this pair represented the lowest score (i.e., the “weakest feeling of the same material”). Participants were asked to attend to both modalities simultaneously. They were also asked not to pay attention to the shadows and the motion of the object itself. Each trial was completed in 8 seconds on average. A short training was performed at the beginning of the experiment.

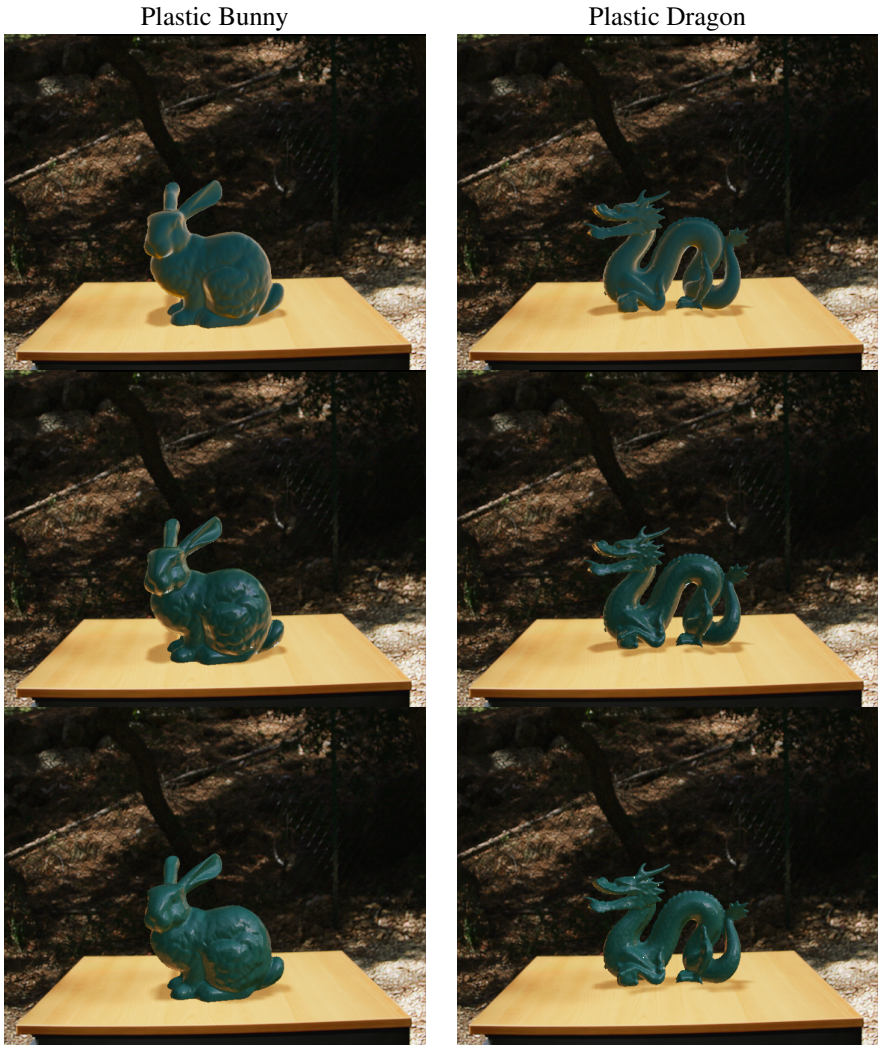

Fig. 1. First row: lowest visual LOD. Second row: highest visual LOD. Both were rendered interactively in the experiment. The last row shows the offline reference rendering of each object for Plastic.

The experiment is naturally divided into four blocks based on the combination of *Object* and *Material* Participants passed the blocks in counterbalanced order. For each of these blocks, two main parameters vary: Sound LOD (the quality, or LOD of the contact sounds of the falling objects was controlled by varying the number of modes used for modal synthesis) and BRDF LOD (the quality of the material rendering was controlled by varying the number of spherical harmonic coefficients used for each LOD). Each trial was repeted three times. For each block and each trial, we measure first the similarity rating and the time spent to give the rating (reaction time).

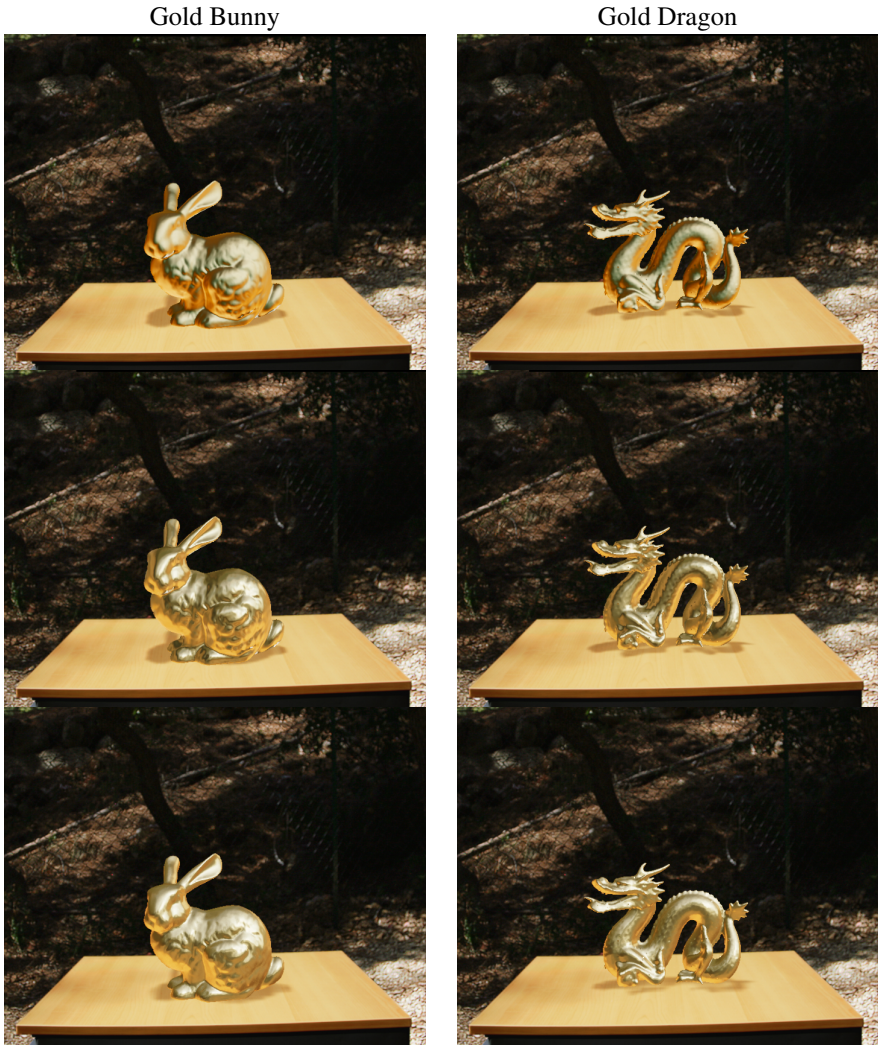

Fig. 2. First row: lowest visual LOD. Second row: highest visual LOD. Both were rendered interactively in the experiment. The last row shows the offline reference rendering of each object for Gold.

2.4 Apparatus

Audio was rendered on headphones and spatialized with stereo panning in front of the participant. The visual algorithms were implemented in a game-oriented rendering engine (Ogre3D), with a high quality graphics card (GeForce 8800GTX). Screen resolution was 1600x1200 on a 20.1 inch screen (DELL 2007FP), and the rendering ran at about 29 fps in a 700x700 screen (700x561 being devoted to the stimuli, the rest for the interface, see Fig. 4). Responses were given on a standard keyboard. Two keys were selected to switch between stimuli with the letters “A" and “B" being highlighted respectively on the top left

	Bunny				Dragon			
	Gold		Plastic		Gold		Plastic	
LOD	BRDF	Sound	BRDF	Sound	BRDF	Sound	BRDF	Sound
1	3	8	2	4	3	8	2	17
2	4	20	3	23	4	26	3	34
3	5	28	4	34	5	39	4	62
4	9	81	7	58	9	109	7	103
5	12	409	12	233	12	439	12	346

Fig. 3. LOD used for the experiment. BRDF represents the number of SH bands, while Sound represent the number of modes.

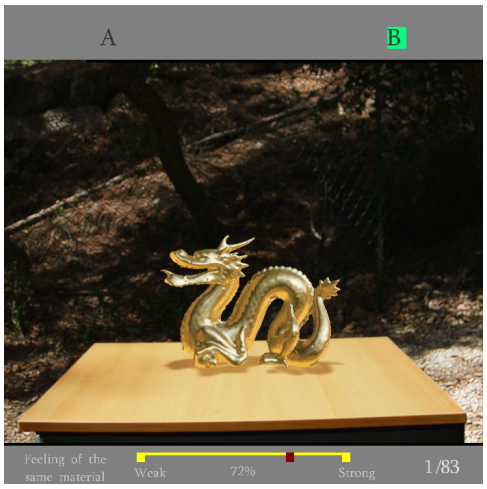

Fig. 4. Screenshot of the user interface.

or right of the interface (Fig. 4). Rating was performed on a finely discretized scale from 0 to 100: the cursor could be moved by 0.5% using the left and right arrows. The Return key was used to validate the choice and go to the next trial. Ratings were recorded, as well as the number of times each of the two stimuli was seen (hidden reference and degraded LOD - see Procedure) and the time to perform each trial.

3. RESULTS

3.1 Similarity ratings

We performed four repeated-measures analysis of variance (ANOVA) on similarity for each block of the experiment with *BRDF LOD*, *Sound LOD* and *Repetition* as within-subjects factors for each. $p < 0.05$ was considered to be statistically significant.

3.1.1 *BRDF and Sound.* For each material and object, the ANOVA revealed a significant main effect of BRDF LOD (Gold Bunny: $F_{4,36} = 72.02$; $p < 0.0001$, Plastic Bunny: $F_{4,36} = 128.84$; $p < 0.0001$, Gold Dragon: $F_{4,36} = 22.40$; $p < 0.0001$, Plastic Dragon: $F_{4,36} = 38.20$; $p < 0.0001$). These results show that an increase in the quality of the BRDF gives improved ratings of similarity with the reference (see Fig. 5).

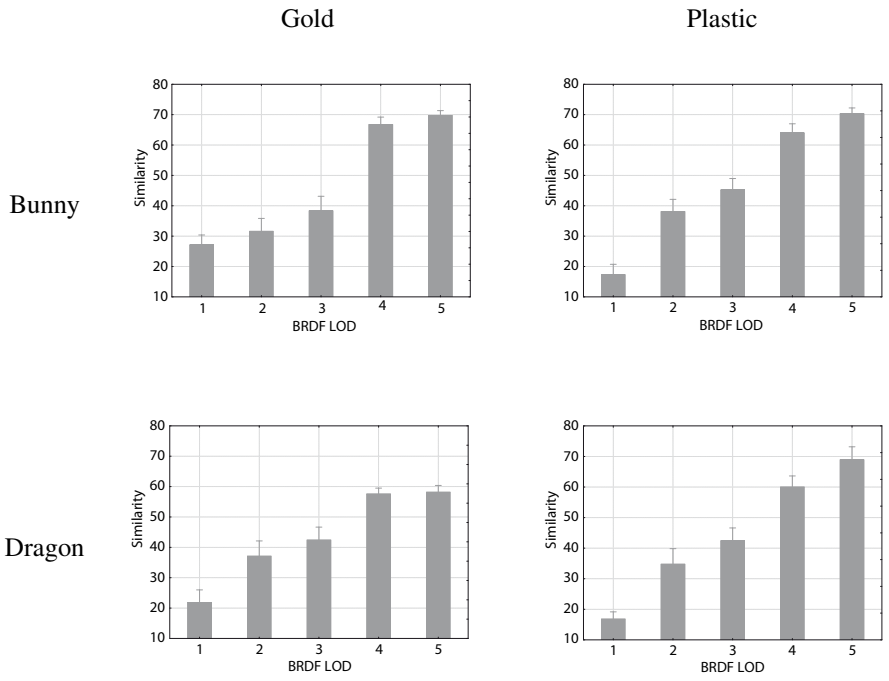

Fig. 5. Average mean ratings of material similarity depending on visual LOD for each object and material. Error bars represent the Standard Error of Mean (SEM). When increasing the number of SH bands for the graphics, we indeed observe a better perception of material quality.

Similarly, the ANOVA revealed a significant main effect of Sound LOD (Gold Bunny: $F_{4,36} = 144.81$; $p < 0.0001$, Plastic Bunny: $F_{4,36} = 55.80$; $p < 0.0001$, Gold Dragon: $F_{4,36} = 62.94$; $p < 0.0001$, Plastic Dragon: $F_{4,36} = 44.94$; $p < 0.0001$). In a manner similar to BRDF LOD, increasing the LOD of the modal synthesis improves the similarity rating with respect to the reference (see Fig. 6).

3.1.2 *Interaction between BRDF and Sound.* The most interesting aspect for this work is the *interaction* between *BRDF LOD* and *Sound LOD*. The ANOVAs also revealed that for each material and object, a significant interaction between BRDF LOD and sound LOD exists (Gold Bunny: $F_{16,144} = 14.94$; $p < 0.0001$, Plastic Bunny: $F_{16,144} = 17.10$; $p < 0.0001$, Gold Dragon: $F_{16,144} = 9.14$; $p < 0.0001$, Plastic Dragon: $F_{16,144} = 5.11$; $p < 0.0001$); see Fig. 7. This indicates that the quality of sound and the quality of BRDF rendering mutually interact on the judgment of similarity.

3.1.3 *Repetition.* The *Repetition* factor did not reach a significant level ($p > 0.2$) except for the Gold Bunny case ($F_{2,18} = 5.51$; $p = 0.014$). However, although significant, this effect was due to very small disparities between the three repetitions (mean similarity rating for Repetition 1: 48.6; mean similarity rating for Repetition 2: 47.5; mean similarity rating for Repetition 3: 44.1). Overall, the fact that there is almost no significant effect on

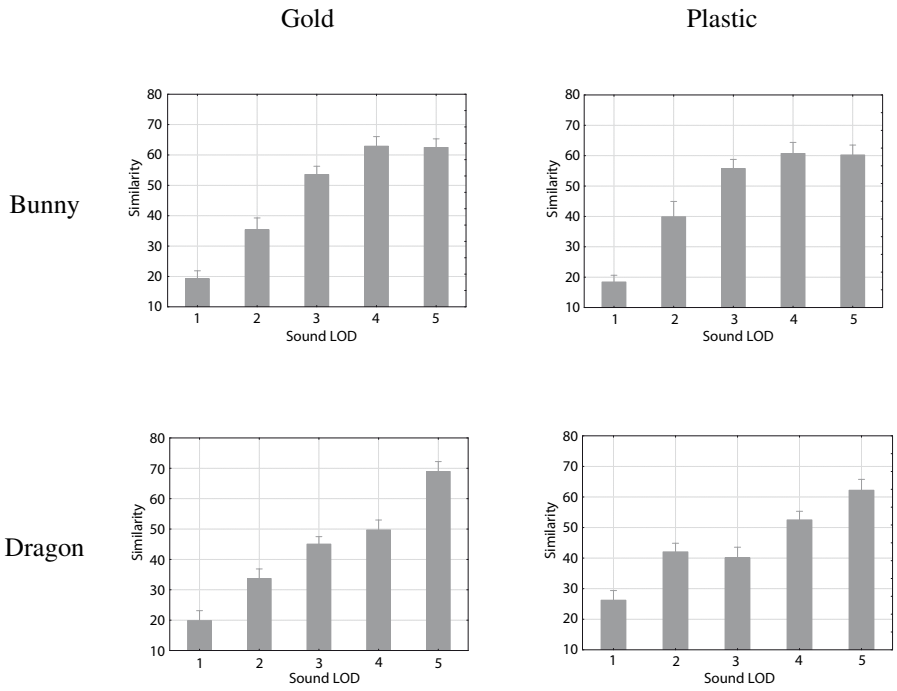

Fig. 6. Mean ratings and SEM of material differences depending on sound LOD for each object and material. When increasing the number of modes for the sound, we indeed observe a better perception of material quality.

the Repetition factor strongly indicates that participants performed the task well and were stable in their judgment across one experimental block.

3.1.4 *Other Interactions.* As discussed earlier (Sect. 2.2.4), *BRDF LOD* and *Sound LOD* did not take the same values between the two different materials and the two objects. As a consequence, the previous ANOVAs were conducted on each material and object to identify the effect of these specific LOD on similarity ratings.

We are also interested in exploring the potential differences between the two objects or the two materials in more detail. To do this, a repeated-measures ANOVA including *Object, Material, BRDF LOD, Sound LOD* and *Repetition* as within-subjects factors was also performed. This ANOVA revealed, as could be expected from the preceding analysis, a main effect of *BRDF LOD* ($F_{4,36} = 89.57$; $p < 0.0001$), *Sound LOD* ($F_{4,36} = 154.34$; $p < 0.0001$), and an interaction effect between *BRDF LOD* and *Sound LOD* ($F_{16,144} = 29.46$; $p < 0.0001$). It also revealed an interaction effect between *Material* and *BRDF LOD* ($F_{4,36} = 6.53$; $p < 0.001$), (see Fig. 8), *Material* and *Sound LOD* ($F_{4,36} = 4.86$; $p < 0.005$), and *Object* and *Sound LOD* ($F_{4,36} = 14.75$; $p < 0.0001$). No significant main effects of *Material* or *Object* were shown.

Interaction between *Material* and *Sound LOD* as well as *Object* and *Sound* was due to very small differences between similarity ratings for Gold or Plastic.

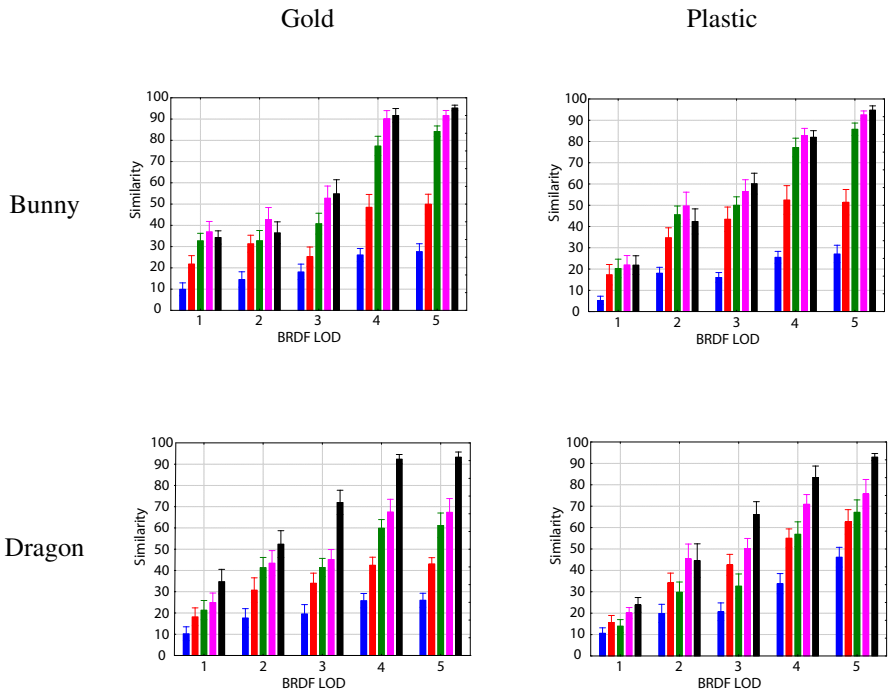

Fig. 7. Interaction between BRDF and sound: Mean similarity ratings and SEM of the different BRDF LOD and Sound LOD, for the two different objects and the two different materials. Blue, red, green, pink and black bars represent increasing sound LOD while the main horizontal axis represents increasing BRDF LOD. Greater perceived differences when varying sound quality can be seen at high BRDF quality than for low BRDF quality.

3.2 Reaction Time

We also analysed the reaction times (RT) needed for participants to rate similarity. We thus performed a repeated-measure ANOVA with *Object*, *Material*, *BRDF LOD*, *Sound LOD* and *Repetition* as within-subjects factors. The ANOVA revealed a significant main effect of *Object* ($F_{1,9}$=6.58; p<0.05), *BRDF LOD* ($F_{4,36}$=8.83; p<0.0001), *Sound LOD* ($F_{4,36}$=9.86; p<0.0001) and *Repetition* ($F_{2,18}$=11.63; p<0.001). Mean RT were longer for the Dragon (M = 8346 ms) than for the Bunny (M = 7755 ms). Mean RT were also longer for the first repetition (M = 8538 ms) compared to the two following repetitions (M = 7816 ms for the second repetition and M = 7798 ms for the third repetition).

This main effect of *Repetition* shows a small learning effect of the task during the experiment. Importantly, this learning effect did not affect the main results (no significant interaction between *Repetition* and all other factors). The data also revealed a significant interaction between *Object* and *Sound* ($F_{4,36}$=4.48; p<0.01).

Figure 9 shows the reaction time for a response, as a function of the BRDF LOD (A) and the sound LOD (B). This pattern of results indicates that when the similarity between the two sequences was high, reaction time was longer. In contrast the shortest reaction time

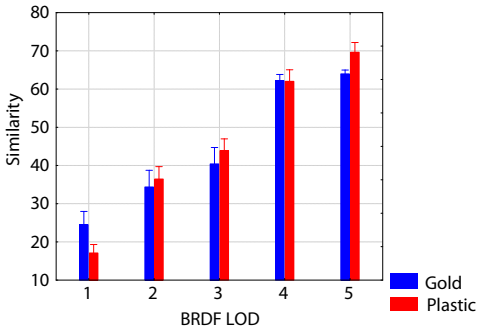

Fig. 8. Interaction between BRDF and Material. Mean similarity ratings and SEM of the BRDF LOD per material. We were obliged to choose different BRDF LODs for each material to accomodate the differences of the perceived materials. Given the discrete nature of the BRDF LODs we see that the choice of LOD results in similar similarity ratings.

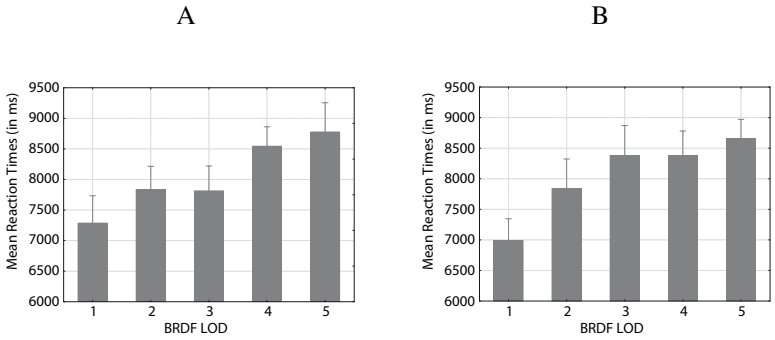

Fig. 9. Reaction time as a function of *BRDF LOD* (A) and *Sound LOD* (B). Overall reaction time increases with LOD, since the stimuli are more similar to the hidden reference, making the task harder.

was observed for the cases where similarity was lowest i.e., the materials were perceived to be very different, since the LOD used is low.

4. DISCUSSION

This experiment demonstrated an interaction between *BRDF* and *Sound LOD*, which has significant algorithmic consequences. We also discuss the validity of stimuli, reaction times and some potential avenues for generalization of this work.

4.1 Stimuli Validation

A first validation of the choice of stimuli can be performed by observing perceived material similarity with the reference for increasing *BRDF LOD* alone and increasing *Sound LOD* alone; this is shown in Figures 5 and 6. With the exception of sound levels 2 to 3 for the Plastic Dragon, material quality was overall rated as increasing when the quality of one modality alone increases. This is a strong indication that the choice of stimuli (see Sect. 2.2) is valid and allows us to have confidence in our results.

As can be seen in Figs. 5, 6, BRDF levels 4 and 5 for Gold Bunny and Dragon, and Sound levels 4 and 5 for Gold and Plastic Bunny are rated approximately the same. In this case the choice of stimuli could potentially have been better. However, BRDF levels are discrete and quite limited and thus we believe that our choice was reasonable.

As noted previously, we chose different LOD for the different materials. In Fig.3 we see the different choices for visual LOD for each of the two materials, and in Fig.8their respective ratings. If we had chosen the same LOD for both Gold and Plastic, ratings would significantly differ from Gold to Plastic. Recall that the choice of SH bands is discrete, and thus no intermediate choice was possible.

4.2 Reaction Times

It is interesting to note that when visual and sound differences are obvious, the reaction times appear to be lower (see Fig. 9). For the case when the stimulus and the hidden reference are almost indistinguishable and for intermediate cases, participants re-played a the pairs of stimuli longer. A similar effect of the number of times each audio sample is played when varying decay differences was shown in [Klatzky et al. 2000].

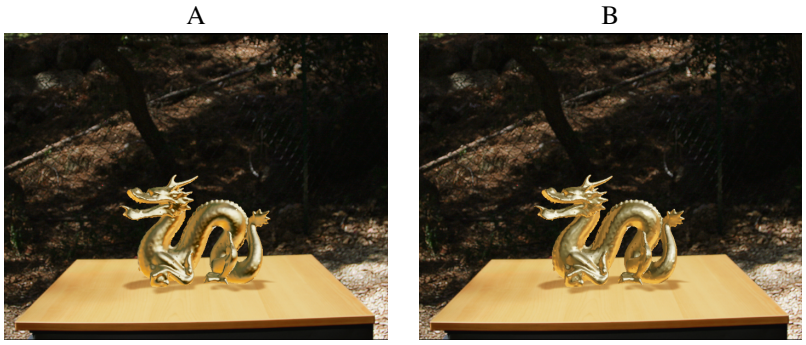

Fig. 10. The Gold Dragon with the third visual quality (A) and the best visual quality (B - visual hidden reference). When using the highest audio quality the approximation A was rated as being very similar to the hidden audiovisual reference (72 on a scale of 100). In contrast, the best visual quality (B) when seen with Sound LOD equal to 3 (intermediate), was actually judged to be *less similar* (61 on a scale of 100) than the audiovisual reference.

4.3 BRDF SH Rendering

Another interesting observation is the lack of perceived differences between BRDF LOD 4 and 5 (see Fig. 5). This means mean that we could easily render 9 SH bands (i.e.,

81 coefficients) for Gold or 7 (i.e., 49 coefficients) for Plastic instead of 12 bands (144 coefficients) without perceivable difference in material similarity. For comparison, the rendering time for 12 bands is 17.8ms (without additional cost), whereas it is 6.6ms for 9 bands and 1.2ms for 7 bands. Previous work on Spherical Harmonics lighting observes that 3 SH bands were enough to render Lambertian surfaces ([Ramamoorthi and Hanrahan 2001b; 2001a]) given the fast decay of SH coefficients for the Lambertian term. In our context, our experimentation indicates that 7 (plastic) or 9 (gold) bands could be enough to render glossy materials like metallic BRDFs.

4.4 Interaction between Sound and Visual Quality

The most interesting result is the interaction between BRDF and sound LOD in perceived quality. If we interpret similarity to the reference as a measure of quality, we see that, for the same BRDF LOD, material quality is judged to be higher when the sound LOD is higher. This can be seen in the different-colored bars for each *BRDF LOD* in Fig. 7.

As an example consider the BRDF level of detail equal to 3 for the Gold Dragon (see Fig. 10, A). With the highest sound LOD (equal to 5), material similarity compared to the reference audio-visual Gold Dragon (see Fig. 10, B) was rated about 72 on a scale of 100. It is important to note that the reference versus hidden reference similarity was rated 92 on a scale of 100. As can be seen in Fig. 10 (without sound), the differences are quite visible.

We thus see that using the above LOD parameters (BRDF 3 and Sound 5) results in higher perceived quality than, for example, BRDF level 5 with sound LOD level 3, which only rates 61% similarity to the reference.

This result is important since the cost of rendering better quality sound is typically much lower than the cost of better quality BRDF rendering. This is because, computing the dot product for BRDF rendering requires $\mathcal{O}(N^2P)$ operations where N is the number of SH bands (representing N^2 SH basis functions) and P is the number of pixels drawn on screen, whereas the sound requires $\mathcal{O}(M)$ operations where M is the number of modes.

Considering the quadratic increase in computational cost for BRDF rendering compared to the linear cost in modal sound rendering, it is more beneficial to reduce graphics quality while increasing audio quality for the same global perceived material difference to the reference.

To get a feeling for the practical implications of this result, the computation time of the third sound LOD is about 0.21ms and for the highest quality 1.95ms. For BRDFs, the computation time (performed on GPU) for the third quality is about 0.5ms (with an additional 16.7ms of constant cost for soft shadows, deferred shading pass and rotations) whereas it is 17.8ms (in addition to the 16.7ms constant cost) for the highest BRDF quality. In this particular case, we have a gain of 15.56ms per frame if we choose our BRDF and sound LOD based on the results of our study. Another way to see this is that the frame rate (assuming this BRDF rendering to be the only cost), would increase from 30fps to 60fps. This gives a very strong indication of the utility of our results.

Besides the very promising algorithmic consequences of our findings, we believe that the actual effect of audio-visual interaction on material perception we have shown could be very promising in a more general setting. Given the interactive rendering context of our work, and the consequent constraints, we were in some ways limited (discrete levels of detail, some parameters which are only loosely related to physical quantities etc). Nonetheless, to our knowledge this is the first study which shows interaction of audio and graphics in a material perception task. We thus are hopeful that our finding will be a

starting point for more general perceptual research, in which the constraints of interactive rendering will not be required. This could allow the use of parameters such as decay for sound synthesis or a continuous visual level of detail parameter, and lead to wider, more perceptually-motivated results.

4.5 Algorithmic Generalization

Evidently, our study is only a first step in determining the combined influence of sound and visual rendering quality on perceiving material similarity, and in particular similarity to a "gold standard" reference. Our study only examined a limited setting with two objects and two materials, although the choice of materials corresponds to hopefully representative classes of material properties.

More extensive studies of different material types and object geometries should be undertaken, including more objects made of several different materials. We believe that extensions to our results could have a significant potential and utility in an algorithmic context, when managing audio-visual rendering budgets with a global approach.

5. CONCLUSION

Our goal was to determine whether the combined quality levels of visual and sound rendering influence the perception of material, and in particular in the context of interactive systems. The constraints of interactive rendering led us to choose Spherical Harmonic-based levels of detail for BRDF rendering, and a mode-culling contact sound synthesis approach. We designed a study in which subjects compare the *similarity* of interactive sequences with a given audio-visual reference (i.e., high-quality sound and graphics).

The results of our study show that, for the cases we examined, better quality sound improves the perceived similarity of a lower-quality visual approximation to the reference. This result has direct applicability in rendering systems, since increasing the visual level of detail is often much more costly than increasing the audio level of detail. The examples provided show potential for significant computation time savings, for the same, or even better perceived material quality.

To our knowledge, our study is the first to demonstrate interaction between audio and graphics in a task related to perception of materials. Given our motivation for interactive audio-visual rendering, we were necessarily constrained in our choices of stimuli and the extent of our setup. Nonetheless, we are hopeful that our initial study, which indicates the existence of a potentially cross-modal audiovisual effect on material recognition, will inspire more perceptually oriented studies in a more general context.

Acknowledgments

This research was funded by the EU IST FET OPEN project CROSSMOD. We also thank Nicolas Tsingos for his help on audio rendering, and all the volunteers who participated in the experiment reported here. We are very grateful to Durand Begault for his helpful comments on a previous version of the manuscript.

REFERENCES

BEN-ARTZI, A., OVERBECK, R., AND RAMAMOORTHI, R. 2006. Real-time brdf editing in complex lighting. In *ACM Transactions on Graphics (SIGGRAPH Conference Proceedings)*. ACM Press, New York, NY, USA, 945–954.

BONNEEL, N., DRETTAKIS, G., TSINGOS, N., VIAUD-DELMON, I., AND JAMES, D. 2008. Fast modal sounds with scalable frequency-domain synthesis. *ACM Transactions on Graphics (SIGGRAPH Conference Proceedings) 27,* 3 (August).

COOK, R. L. AND TORRANCE, K. E. 1982. A reflectance model for computer graphics. *ACM Transactions on Graphics 1,* 1, 7–24.

DEBEVEC, P. 1998. Rendering with natural light. In *SIGGRAPH '98: ACM SIGGRAPH 98 Electronic art and animation catalog*. ACM Press, New York, NY, USA, 166.

DOEL, K. V. D., PAI, D., ADAM, T., KORTCHMAR, L., AND PICHORA-FULLER, K. 2002. Measurements of perceptual quality of contact sound models. *Proceedings of the International Conference on Auditory Display*, 345–349.

DONNELLY, W. AND LAURITZEN, A. 2006. Variance shadow maps. In *I3D '06: Proceedings of the 2006 symposium on Interactive 3D graphics and games*. ACM, New York, NY, USA, 161–165.

FLEMING, R. W., DROR, R. O., AND ADELSON, E. H. 2003. Real-world illumination and the perception of surface reflectance properties. *Journal of Vision 3,* 5 (July), 347–368.

GIORDANO, B. L. AND MCADAMS, S. 2006. Material identification of real impact sounds: Effects of size variation in steel, glass, wood, and plexiglass plates. *The Journal of the Acoustical Society of America 119,* 2, 1171–1181.

GREEN, R. 2003. Spherical harmonic lighting: The gritty details. *Archives of the Game Developers Conference*.

JAMES, D. L., BARBIC, J., AND PAI, D. K. 2006. Precomputed Acoustic Transfer: Output-sensitive, accurate sound generation for geometrically complex vibration sources. *ACM Transactions on Graphics 25,* 3 (July), 987–995.

KAUTZ, J., SLOAN, P., AND SNYDER, J. 2002. Fast, arbitrary brdf shading for low-frequency lighting using spherical harmonics. In *Proceedings of the 13th Eurographics workshop on Rendering*. 291–296.

KLATZKY, R., PAI, D., AND KROTKOV, E. 2000. Perception of material from contact sounds. *Presence: Teleoperators and Virtual Environments*, 399–410.

KRISTENSEN, A. W., AKENINE-MÖLLER, T., AND JENSEN, H. W. 2005. Precomputed local radiance transfer for real-time lighting design. *ACM Transactions on Graphics 24,* 3, 1208–1215.

MATUSIK, W., PFISTER, H., BRAND, M., AND MCMILLAN, L. 2003. A data-driven reflectance model. *ACM Transactions on Graphics (SIGGRAPH'03) 22,* 3 (July), 759–769.

MCADAMS, S., CHAIGNE, A., AND ROUSSARIE, V. 2004. The psychomechanics of simulated sound sources: Material properties of impacted bars. *The Journal of the Acoustical Society of America 115*, 1306–1320.

NG, R., RAMAMOORTHI, R., AND HANRAHAN, P. 2003. All-frequency shadows using non-linear wavelet lighting approximation. *ACM Transactions on Graphics 22,* 3, 376–381.

O'BRIEN, J. F., SHEN, C., AND GATCHALIAN, C. M. 2002. Synthesizing sounds from rigid-body simulations. In *The ACM SIGGRAPH 2002 Symposium on Computer Animation*. ACM Press, 175–181.

RAGHUVANSHI, N. AND LIN, M. C. 2007. Physically based sound synthesis for large-scale virtual environments. *IEEE Computer Graphics and Applications 27,* 1, 14–18.

RAMAMOORTHI, R. AND HANRAHAN, P. 2001a. An efficient representation for irradiance environment maps. In *SIGGRAPH 2001, Computer Graphics Proceedings*, E. Fiume, Ed. 497–500.

RAMAMOORTHI, R. AND HANRAHAN, P. 2001b. On the relationship between radiance and irradiance: Determining the illumination from images of a convex lambertian object. *The Journal of the Optical Society of America*.

RAMAMOORTHI, R. AND HANRAHAN, P. 2002. Frequency space environment map rendering. *ACM Transactions on Graphics 21,* 3, 517–526.

REINHARD, E., STARK, M., SHIRLEY, P., AND FERWERDA, J. 2002. Photographic tone reproduction for digital images. *ACM Transactions on Graphics 21,* 3, 267–276.

SLOAN, P.-P., KAUTZ, J., AND SNYDER, J. 2002. Precomputed radiance transfer for real-time rendering in dynamic, low-frequency lighting environments. 527–536.

SLOAN, P.-P., LUNA, B., AND SNYDER, J. 2005. Local, deformable precomputed radiance transfer. *ACM Transactions on Graphics 24,* 3, 1216–1224.

STORMS, R. L. AND ZYDA, M. 2000. Interactions in perceived quality of auditory-visual displays. *Presence: Teleoperators and Virtual Environments 9,* 6, 557–580.

VAN DEN DOEL, K. AND PAI, D. K. 2003. Modal synthesis for vibrating objects. *Audio Anecdotes*.

VANGORP, P., LAURIJSSEN, J., AND DUTRÉ, P. 2007. The influence of shape on the perception of material reflectance. *ACM Transactions on Graphics (SIGGRAPH Conference Proceedings)*.

Received XX and accepted YY.