Multi-Modal Perception for Selective Rendering

Carlo Harvey, Kurt Debattista, Thomas Bashford-Rogers and Alan Chalmers

WMG, University of Warwick, UK.

Abstract

A major challenge in generating high-fidelity virtual environments (VEs) is to be able to provide realism at interactive rates. The high-fidelity simulation of light and sound is still unachievable in real-time as such physical accuracy is very computationally demanding. Only recently has visual perception been used in high-fidelity rendering to improve performance by a series of novel exploitations; to render parts of the scene that are not currently being attended to by the viewer at a much lower quality without the difference being perceived. This paper investigates the effect spatialised directional sound has on the visual attention of a user towards rendered images. These perceptual artefacts are utilised in selective rendering pipelines via the use of multi-modal maps. The multi-modal maps are tested through psychophysical experiments to examine their applicability to selective rendering algorithms, with a series of fixed cost rendering functions, and are found to perform significantly better than only using image saliency maps that are naively applied to multi-modal virtual environments.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Picture/Image Generation—Viewing Algorithms I.4.8 [Computer Graphics]: Image Processing and Computer Vision—Scene Analysis - Object Recognition I.4.8 [Computer Graphics]: Image Processing and Computer Vision—Scene Analysis - Tracking

Keywords: Multi-Modal, Cross-Modal, Saliency, Sound, Graphics, Selective Rendering

1. Introduction

A major research challenge of Virtual Environments (VEs) is to accurately simulate a real world environment. This is motivated by the increasing use of VEs in a wide range of applications such as concert hall and architectural design [Dal, Nay93] and immersive video games [MBT*07, RLC*07, GBW*09]. Multi-modal VEs aim to deliver more sensory information than from a sole domain and yield an increased sense of immersion over single modality environments [DM95]. Furthermore, such multi-modal VEs can aid object recognition and placement; identification and localisation; and generating conclusions pertaining to the scale and shape of the environment [Bla97].

Limitations of the human sensory system have been used in order to improve the performance of perceptually-based rendering systems. Examples of this used to decrease the auditory [TGD04, MBT*07] or visual [CCL02, RFWB07, RBF08] rendering complexity with little or no perceivable quality difference to a user have been implemented and verified. Moreover, it has been shown that it is possible to increase the perceptual quality of a stimulus in one modality by directing gaze due to the introduction of another modality [MDCT05a]. This can be used for improving the perception of a material's quality [BSVDD10], Level-of-Detail (LOD) selection [GBW*09] or for increasing the spatial [MDCT05a, HHT*11] and temporal [MDCT05b, HHT*11,HDAC10] quality of visuals by coupling them with corresponding auditory stimuli.

Straightforward applications that take advantage of the human sensory system have attempted to predict gaze direction and attention. However, spatial sound has been shown to be important in the perception of a scene in VR and should thus not be ignored; furthermore, spatial visual saliency is not necessarily the best predictor of visual attention [MD02]. In addition to this, an estimator of the sound intensity alone is not enough [KPLL05]. Kayser *et al.* showed quantitatively that an auditory saliency map extracts a measure of saliency which cannot be obtained from sound intensity alone. Although typically attention can be controlled, a strong enough novel cue can take our attention [Pet99]. A novel auditory stimulus can attract visual attention as this distraction aids the detection and spatial localisation of objects [DS98]. In this paper we propose a general algorithm to concatenate sound saliency with visual saliency in the spatio-visual domain. In particular, this paper considers the effect of directional sound on a user's visual attention towards rendered images. Based on a sound transport simulation, multi-modal maps are derived. These are used to reduce render times while maintaining perceptual equality. This is validated in a further user study. Specifically we make the following contributions:

- Construction of a sound map which encodes directional sound saliency information. These are produced through a sound simulation based on tracing phonons [BDM*05].
- Utilising the sound maps to represent saliency of a directional sound signal and using density estimation to construct a saliency map for the spatial domain.
- Combination of the traditional visual saliency map with the directional sound saliency map into a multi-modal saliency map. This is used to reduce rendering time in this paper. The technique could be more broadly applied to any audio-visual interface requiring a degrade function, such as compression of video.
- A user study which validates the use of the multi-modal maps to reduce rendering time, but maintaining similar perceptual quality to reference images computed at higher sampling rates.

2. Background and Related Work

2.1. Images

Saliency models have been used previously in computer graphics, and more so in computer vision applications. Yee *et al.* [Yee00, YPG01] adapted Itti and Koch's [IKN98] model of visual saliency in order to speed up the rendering process. For each frame a spatiotemporal error tolerance map [Dal98] was created based on velocity dependant contrast sensitivity, and a saliency map [IKN98]. The two maps were combined to create a new map, termed *aleph map*. The aleph map was used to determine where computational resources were to be directed in screen space.

Marmitt *et al.* [MD02] examined how Itti and Koch's [IKN98] model performed when predicting visually salient features in virtual scenes. The model had been shown to perform accurately on real imagery [PS00], however the analysis showed that the correlation between human saccades and model predicted saccades was quite low. Marmitt *et al.* [MD02] hypothesised that the lack of correspondence between real and predicted views occurred due to the absence of a memory module in the artificial model. The human brain has temporal memory and remembers what it has seen.

More recent work by Koulieris *et al.* [KDCM14] showed a method to extend a recent saliency model, incorporating effects such as object context, uniqueness of objects and temporality. This allowed an attention based level-of-detail manager to constrain material quality in presented images whilst maintaining frame rate. The benefit of this technique in a proof of concept was to incorporate parallax occlusion mapping on a mobile device. For a full overview on perception in graphics please see [MMG11].

2.2. Sound

The computational bottlenecks in sound rendering can be grouped into two broad types: the cost of acoustic spatialisation and the cost per sound source. The processing of complex sound scenes is composed of spatialisation and per source information. This can take advantage of perceptually-based optimisations in order to reduce both the necessary computer resources and the amount of audio data to be stored and processed. The MPEG I Layer 3 (mp3) standard [PS00] is one such example of this which exploits Perceptual Audio Coding (PAC), where prior work on auditory masking [Moo97] had been successfully utilised. This is implicitly used together with masking to discard information of audio content deemed perceptually irrelevant from the original sound. The missing audio content is not perceived in the resultant sound.

The auditory saliency map presented by Kayser *et al.* [KPLL05] has been used to predict the parts of a sound source that will attract human attention, so that more resources in the acoustic rendering process could be assigned for their computation.

This method was adapted by Moeck *et al.* [MBT*07] for acoustic rendering by integrating saliency values over frequency subbands. They suggested using auditory saliency as a heuristic for the clustering stage of multiple audio sources. Recent work on the synthesis of sound, showed that combining the instantaneous energy of the emitted signal and attenuation is also a good criteria [GLT05, Tsi05].

The presence of many sensory stimuli, including sound, may influence the amount of cognitive resources available to a viewer to perform a visual task, this is termed as the *modality appropriateness hypothesis* [WW80]. Research has investigated the influence of auditory cues on visual attention and visual cues on audition. Mastoropoulou *et al.* [MDCT05a, MDCT05b] showed that a selective rendering technique for Sound Emitting Objects (SEO) can be used to render animations, and can decrease the rendering time required. Considering the angular sensitivity of the Human Visual System (HVS) and inattentional blindness, the visual region that contained the SEO was rendered in high quality at an appropriate angle, whilst low quality visuals were displayed for the rest of the scene and the viewer failed to notice the quality difference.

Harvey *et al.* showed via eye tracking that human visual attention is distracted by spatial sound, even when related objects are omitted [HWBR*10]. Hulusic *et al.* [HDAC10] investigated how the perceived quality threshold for renderings is influenced by audio. The authors examined how related and unrelated audio influences visual perception. This showed that incongruent sound can be used for increasing the perceived temporal smoothness of graphics, while congruent audio has no significant effect on the perceived quality threshold. Grelaud *et al.* [GBW*09] developed a model to detect when many instances of vibratory and contact synthesis were occurring and to fluctuate resources accordingly from the visual domain to the auditory domain dynamically. As a result, visuals were poorer when many objects collided and audio was deemed more important. This directly attempted to exploit the *modality appropriateness hypothesis*. However, even though the result showed the technique worked and perception was unaltered, no empirical technique was used and the variability of resources was user defined and for a specific task. A generic model for bi-modal scenarios (auditory-visual interaction) has yet to be considered.

3. Modal Map Generation

A novel temporal acoustic algorithm for spatial visual saliency prediction is presented in this section. The algorithm is based upon sound-level-detection on the image plane modulated by the auditory salience feature vector of an asynchronous acoustic stimulus. Blended with conventional visual saliency predictors this enables spatial heuristics to guide sample count for rendering to be employed. In Section 4 we show that this provides better perceptual responses than previous image synthesis sampling strategies.

3.1. Algorithm: Intensity Map

A two step approach to generate the directional intensity of the sound wave on the image plane is used. The first step utilises the algorithm employed by the sonel and phonon mapping techniques for sound rendering [KJM04, BDM*05]. This is a particle tracing based method where starting points and directions of paths are generated on the sound source and propagate around the scene. Information at each hit point is stored and used in a second stage to reconstruct the sound at the listener position.

Initially, the sound source is approximated by a set of frequencies, and for each sound-carrying particle one frequency is sampled. Then the starting point and direction of the sound particle is selected according to the emission distribution of the sound source. This sound particle is then traced into the scene, and at each intersection (including the initial point on the sound source) a set of information is stored. This consists of sound intensity values attributed to each hit point: pressure (P), frequency (F), incoming direction (ω_i) and world space position (x'). These points are stored in a KD-Tree for fast searching in the second step. This could also be implemented through a splatting approach, akin to Progressive Photon Mapping [HOJ08]. Whilst splatting is fast, our approach avoids heuristically setting a splat radius, which leads to a smooth intensity map regardless of the number of sound particles traced. Information such as world space position is used from the KD-Tree later when generating the binaural audio as the sound paths need to be connected to the Head Related Impulse Response (HRIR) for evaluation. The reflection type at the intersection is then sampled, the intensity of the sound particle is appropriately modulated, and the tracing continues. This process continues until the tracing process is terminated stochastically via Russian Roulette. This process is shown in Algorithm 1.

Algorithm 1: Sound Particle Tracing

```
KD-Tree kd
for each sound particle do
    Sample frequency $F$
    Sample sound source emission
    Store sound particle on sound source in kd
    Generate ray starting at sound source
    while Path is not terminated do
        Store sound particle at intersection in kd
        Sample surface reflection and generate ray
        Apply Russian Roulette
    end while
end for
```

The second step generates the auditory intensity map through density estimation using the previously stored sound particles. This is a view-dependent map which encodes the intensity of the sound at points visible in the scene in the view direction. A ray is generated for each pixel using jittered sampling, and density estimation is performed at each primary hit point x using a balloon estimator for a KNN-search. This expands an initial search radius in world space until N sound particles are located. This process is accelerated using the KD-Tree which stores the sound particles. Once the N nearest sound particles to the primary ray hit point are found, a density estimate is performed according to the following equation:

$$S_o = \frac{1}{\pi r^2} \sum_{i=1}^{N} P(i) fr(x'(i), \omega_i(i), \omega_o, F(i)) \qquad (1)$$

where S_o is the pressure at the primary hit point x, r is the radius from the KNN-search, $P(i)$ is the pressure associated with the i'th nearest sound particle, $fr(x'(i), \omega_i(i), \omega_o, F(i))$ is the frequency dependent surface auditory reflectance function (see Siltanen et al. [SLKS07]) at point x' parameterised by the i'th incoming sound particle direction $\omega_i(i)$, the direction to the listener ω_o and frequency F_i. This function encodes how sound of a certain frequency F reflects off a surface. The value N is a user defined value (we use 50). This step is shown in Algorithm 2.

This novel concept of pressure flux through screen space to represent the directionality component of sound is shown in Figure 1 along with a visualisation of the cache point storage in the scene for one octave band.

```
Algorithm 2: Auditory Intensity Map Generation
  for each pixel $p$ do
    Sample pixel and generate ray
    Calculate hitpoint $x$
    Find $N$ nearest sound particles
    Calculate pressure at $x$ (Equation 1)
    Store pressure in map at $p$
  end for
```

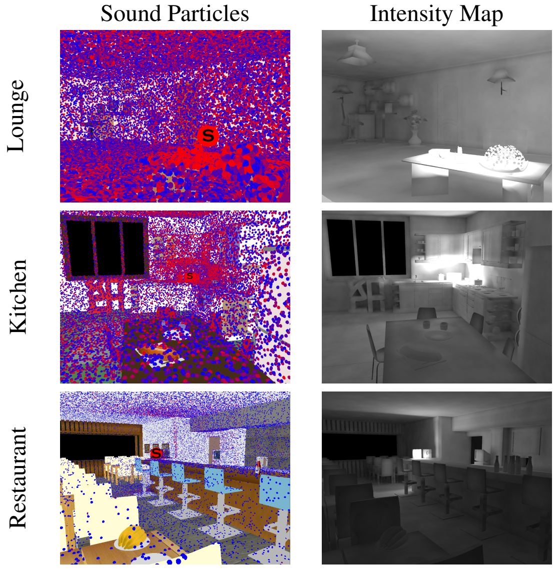


Figure 1: Sound Particles Visualisation and Intensity Maps. The Sound Particles are shown as points and are coloured by stored pressure, red to blue, 1 to 0. The sound sources are denoted by an "S" character in the images. The lounge sound is a phone ringing, the kitchen sound a microwave starting up and running and the restaurant sound is a sample of music emanating from the speakers.

3.2. Temporal Map

Hearing is substantially weaker than vision in spatially related tasks. However, the temporal resolution of the Human Auditory System (HAS) is higher than the visual temporal resolution. According to Fujisaki et al. it is 89.3Hz [FN05]. In order to make the auditory intensity map applicable temporally to the spatial domain it is necessary to weight the importance of the generated spatial sound intensity map. The work in auditory saliency maps can do this by predicting important regions of the acoustic profile into temporal feature vectors. These feature vectors sit as a weight between using the sound intensity maps and visual saliency maps in a perceptual selective temporal renderer.

In addition, the application of a spline introduces the weight in advance of the predicted onset and thus allows the selective renderer pipeline to not be reactive to attentional models but to be proactive and sample an area in advance of predicted attention towards that area. More information on this spline can be found in Section 4.3.4.

Fusion for audio-visual inputs can be performed at two distinct levels: *low-level*, at the extracted saliency; or *high-level*, at the original feature vector level. Given a video stream, audio-visual salience would be construed as a temporal sequence of audio-visual saliency values. This would have each value represent a measure of importance of the multi sensory stream at every frame m. This may result in some form of fusion of the two features, which may be non-linear, have some form of memory or vary with time. For example, for the purposes of the experiments presented in Section 4, this paper proceeds with a linear and memoryless schema for this audio-visual fusion:

$$S[m] = w_A \cdot S_A[m] + w_V \cdot S_V[m] \qquad (2)$$

where the weights w_A and w_V assigned to the fusion can be perceptually guided based upon a high level load balancing framework. However, in the case of the experiment presented in Section 4 these are assigned values: $w_A = F_V[m]$ and $w_V = 1 - F_V[m]$ where $F_V[m]$ is the sound saliency feature vector trace for the relevant audio sample corresponding to frame m. S_A and S_V are the selective guidance mechanism for the image plane for the different modalities, audio and visual respectively. This acts as a temporal slider between the standard visual saliency map and the *auditory intensity map* based upon the salient features of the acoustic information.

Figure 2 shows the original visual saliency, the sound intensity map, and a weighted combination for frame number $m = 180$, time = 3s. This is shown for the Lounge, Kitchen and Restaurant scenes with the respective value of $F_V[m]$ for the relevant audio on a scene by scene basis guiding the weighting.

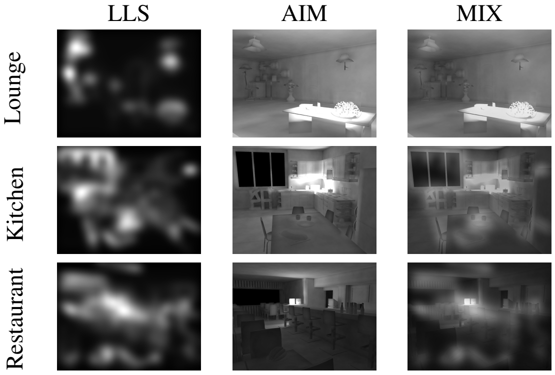


Figure 2: Comparison of visual and auditory intensity render mixes. LLS: Low Level Saliency Map, AIM: Auditory Intensity Map, MIX: The Mix strategy which combines both techniques together.

4. Psychophysical Experimental Layout and Procedure

The psychophysical experiment outlined in this section intends to validate two frequently used selective rendering operators against the algorithm reported in Section 3.1 and also a non-temporal intensity-only version of this.

4.1. Method

Four rendering strategies are used to evaluate the two methods presented in the previous section and compare with more traditional methods: *Uniform* (Uni), each pixel is sampled uniformly; *Visual Saliency* (Sal), each pixel is sampled based on the visual attention prediction map; *Sound Intensity Map* (SoSal), each pixel is weighted based upon the auditory saliency trace and the intensity map from acoustic simulation only; and, *Temporal Map* (Mix), each pixel is sampled by a weighted combination of the visual saliency map and the temporal auditory saliency trace weighted by the directional intensity acoustic simulation.

Pairwise comparisons amongst all the renderings for three scenes were used to judge the proposed methods. Pairwise comparisons were chosen as there were not that many techniques to compare against so a comparison of all methods against each other was feasible within a reasonable time. Participants were asked to always choose one of the two pairs (forced choice).

The experiment used 18 image pairs, six comparisons for each of three scenes presented in Section 4.3.1. The conditions investigate the effect of various maps against one another in a pairwise performance test. The rendering strategies are governed, not by the algorithm, but by the pixel sampling strategy and the render function cost.

4.2. Participants

A total of 28 participants took part in this experiment, 21 males and 7 females. Participants reported no hearing difficulties and normal or corrected-to-normal vision. The age range of participants was between 21 and 42, with an average age of 27. Each participant was presented with all of the scenes, thus looking at a total of $\frac{t(t-1)}{2} \times 3 = \frac{4(3)}{2} \times 3 = 18$ image pairs.

4.3. Materials

The participant sat on a chair, with the backrest of the chair 115 cm from the display. Binaural headphones were used for audio delivery and the monitor used was a 37" LCD panel display. The resolution of the LCD panel was 1024×768 with a refresh rate of 60 Hz and images displayed corresponded to this resolution so no up or down scaling was necessary and the images were displayed natively. The 2 channel audio streams encoded the attenuation and delays of the HRIR for every sound contribution path reaching the user in the simulation. The convolved sound was represented as a two channel lossless 24-bit *.wav* file.

A significant number of materials and parameters have been used for the user study so they are discussed in detail in the following subsections.

4.3.1. Scenes

Three different scenes of varying complexity were used, each with a static camera. Figure 4 demonstrates renders (and saliency maps) of the three scenes termed: Lounge, Kitchen and Restaurant. The sound sources used in each of these scenes were congruent to the scene and were representative of an object in the scene. In the Lounge scene there was a phone ringing, in the Kitchen scene the microwave was turned on and, finally, in the Restaurant there was some music playing from the speakers. Each of the sounds was spatialised for that point and the listener was positioned at the same position as the camera.

4.3.1.1. Render Cost Function: In a selective rendering pipeline given a map as a heuristic to weight the sampling strategy; a cost to compute an image in terms of the degree of sampling used can be given as:

$$V = \sum_{x=0}^{w} \sum_{y=0}^{h} s_{\min} + ((s_{\max} - s_{\min}) \cdot sal(x,y)) \qquad (3)$$

where V is the number of samples required to compute an image, $s_{\min}$ is the minimum number of samples used to calculate radiance through a pixel, $s_{\max}$ is the most number of samples used to calculate this, $sal(x,y)$ is the weighting coefficient for a specific pixel in image space, and x and y are pixel coordinates.

Scene	Sampling	s_{min}	s_{max}	Avg. SPP
Lounge	Uni	1690	1690	1690
Lounge	Sal	1200	5000	1700
Lounge	SoSal	500	5000	1719
Kitchen	Uni	630	630	630
Kitchen	Sal	200	2105	627
Kitchen	SoSal	200	2000	613
Restaurant	Uni	75	75	75
Restaurant	Sal	70	90	73
Restaurant	SoSal	70	95	73

Table 1: Render Cost Function Across Scenes. Avg. SPP (Samples per pixel) is the average number of samples used to generate an image, dictating complexity.

To investigate the perceptual difference between two selective rendering strategies it is necessary to control this render cost function so that, given a number of samples to generate each image, V, an optimisation process starts to vary s_{min} and s_{max} such that $(S \approx V) \pm f$ where f is some user defined control of sufficient leeway to compensate for the fact that s_{min} and s_{max} are restricted to integers in the optimisation process and S is the actual number of samples used in the generation of the image.

Varying render cost functions were used to investigate if the technique was applicable generically or not. Table 1 shows the various s_{min}, s_{max} for the various sampling methods used.

The Visual Difference Predictor (VDP) [Dal93] results of comparisons between the three sampling strategies are presented in Table 2 and shown in Figure 3. The VDP results show there are distinct differences between the selectively rendered images for the same computational costs; effectively indicating that without sound there are clear differences between the methods.

Scene	Uni vs. SoSal	Uni vs. Ref	SoSal vs. Ref
Lounge	28.8546%	45.8543 %	13.8315%
Restaurant	7.17%	75.6775%	80.0777%
Kitchen	8.4064%	63.4815%	63.0377%

Table 2: Selective Render VDP Analysis at P>75%

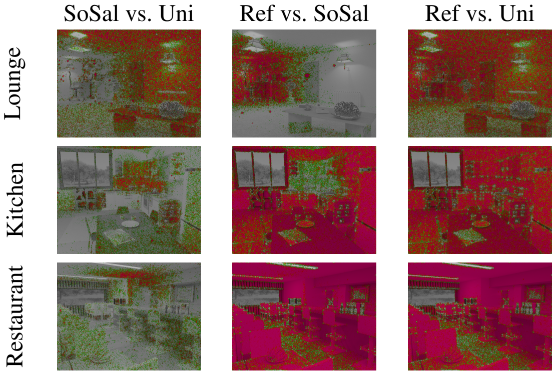


Figure 3: Selective Render VDP Image Comparisons. From top to bottom by row; Lounge, Kitchen and Restaurant scenes respectively. This probability of detection map displays how likely a difference between two images is noticeable. Red denotes high probability, green - low probability.

4.3.2. Vision

Visual saliency predictor maps are computed using Itti et al's method [IKN98]. The saliency maps are shown in Figure 4. This step used the reference uniform path traced image as the input to the saliency generation. However a GPU snapshot of the scene could just as easily be used in a real time implementation of this pipeline, as suggested by Longhurst et al. [LDC06] and by Yee *et al.* [YPG01].

4.3.3. Audio

The binaural format was chosen to reproduce acoustic spatialisation features within the multi-modal VR environment. The pipeline calculates the Room Impulse Response (RIR) in the environment for a particular sound source location and listener position. This RIR encodes how the sound paths travel from the source to the listener in the environment. To convert this to binaural, a modelled Head-Related Transfer Function (HRTF) was implemented using the structural models of the Inter-Aural Time Difference (ITD) and Inter-Aural Level Difference (ILD) equations. This was done per N paths in the acoustic simulation for each azimuth and elevation to the listener position. These delays and attenuations are convolved with the RIR to provide a binaurally encoded impulse response. Figure 5 shows the RIRs from the scenes used.

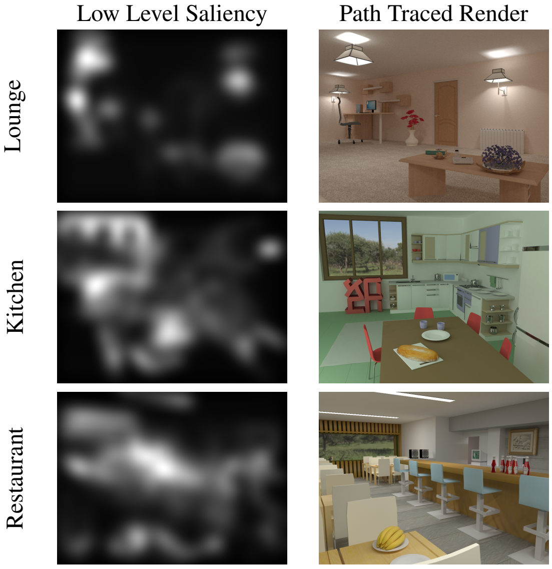


Figure 4: Low level Saliency Maps and the Path Traced Renders of the three scenes. The audio reproduced in each room; Lounge - Phone, Kitchen - Microwave, Restaurant - Speaker playing music.

Simulations were performed on monaural anechoic sound and using the pipeline described above, rendered to encode spatial features of the presented environments: phones in the lounge, microwave in the kitchen and speaker playing music in the restaurant. In order to generate the RIRs, accurate material absorption coefficients had to be used in the environment to accurately encode the frequency responses of different material absorption rates. Common material α_f values per frequency were used appropriately throughout the scenes on correspondent surfaces, after [Sur12].

4.3.4. Auditory Saliency: Feature Vector Curves:

Using the auditory saliency model suggested by Coath et al. [Coa05, CDS*09] to extrapolate salient identities of the transient onsets provides a coarse grained acoustic feature vector for an arbitrary wave form. The attention curve for the audio signal is constructed from the saliency values, provided by the set of audio features (transient onsets, cochlear response and spectral change). Conceptually, salient information is modelled through source excitation and average rate of spectral and temporal change. The simplest scenario

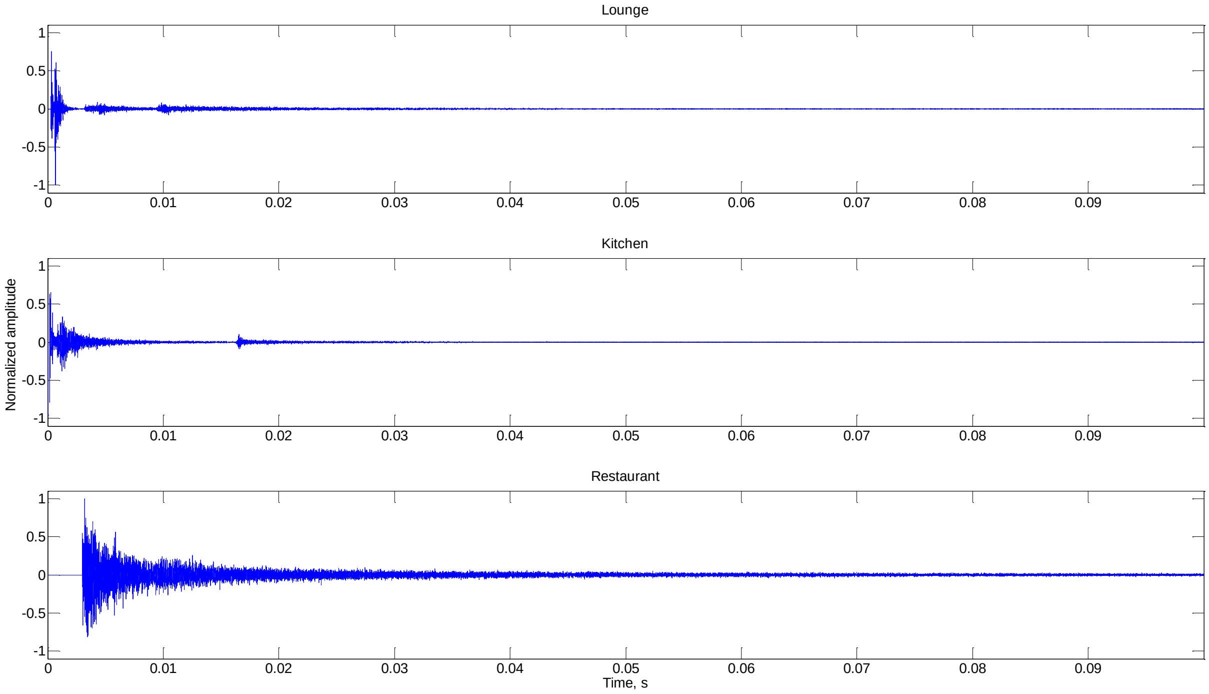


Figure 5: Scene Room Impulse Responses; (top) Lounge, (middle) Kitchen, (bottom) Restaurant. A room impulse response is the response of an environment to an input signal, in this case an approximation to an ideal Dirac Delta function.

of an audio saliency curve is a weighted linear combination of the normalised features. A perceptually motivated approach is a non-linear fusion technique, based on time varying weights. Temporal variation information is extracted by the onset and offset portions, while spectral change is calculated from the intermediate sustain periods. Energy measurement has previously been used to detect speech event boundaries [KPLL05] and as such is used as an index to an event of a transitional point.

For use in the selective rendering pipeline the values from the saliency trace need to be absolute values and then normalised so the trace lies in the range $x \in 0,1 \colon x = ||\hat{x}||$. Fitting a spline to this in order to smooth map transition states helps to mitigate flickering as weights are altered. Other metrics may be chosen for this process of regularisation, such as a solid blur. Whilst the HVS supports an element of flicker fusion at the rate of 26Hz [HS36], temporal discrepancies are still picked up. Due to the *temporal sensitivity* of the HAS the important parts of the vector are the most salient, and thus highest values. The saliency trace, x, was processed into ten bins of max values. These values were used as derivatives for a 1-D bicubic spline interpolation. The bicubic interpolation problem consists of determining the 1000 coefficients a_{ij} to upsample the vector back to the appropriate size. Figure 6 shows the original absolute normalised saliency feature vectors (trace), $\hat{x}$ and the bicubic spline fit version for the microwave sound in the Kitchen scene. This process is shown for a 1-D spline fit $p(x,y)$ where a_{ij} are constants:

$$p(x,y) = \sum_{i=1}^{1000} \sum_{j=1}^{1} a_{ij} x^i y^j \tag{4}$$

4.3.5. Temporal Modal Map

In the absence of animations and just single image exposure it was possible to blend the Sal render and the SoSal render temporally guided by the relevant feature vectors for Mix. A timer kept track of which feature value in the vector was appropriate for the current time. A shader read this from a text file of precomputed feature values and two textures were blended to create the final temporal composite.

4.4. Procedure

The display's update frequency was controlled to 60Hz to allow v-sync within the experimental code to easily derive current time and the correct auditory feature vector value for that time. The distance between equipment was standardised and controlled. The experiment was conducted in a dark room to avoid any effects of ambient lighting and participants were allowed five minutes in order to adjust to the environment before commencing the actual experiment. Video presentation order was randomised. The participants

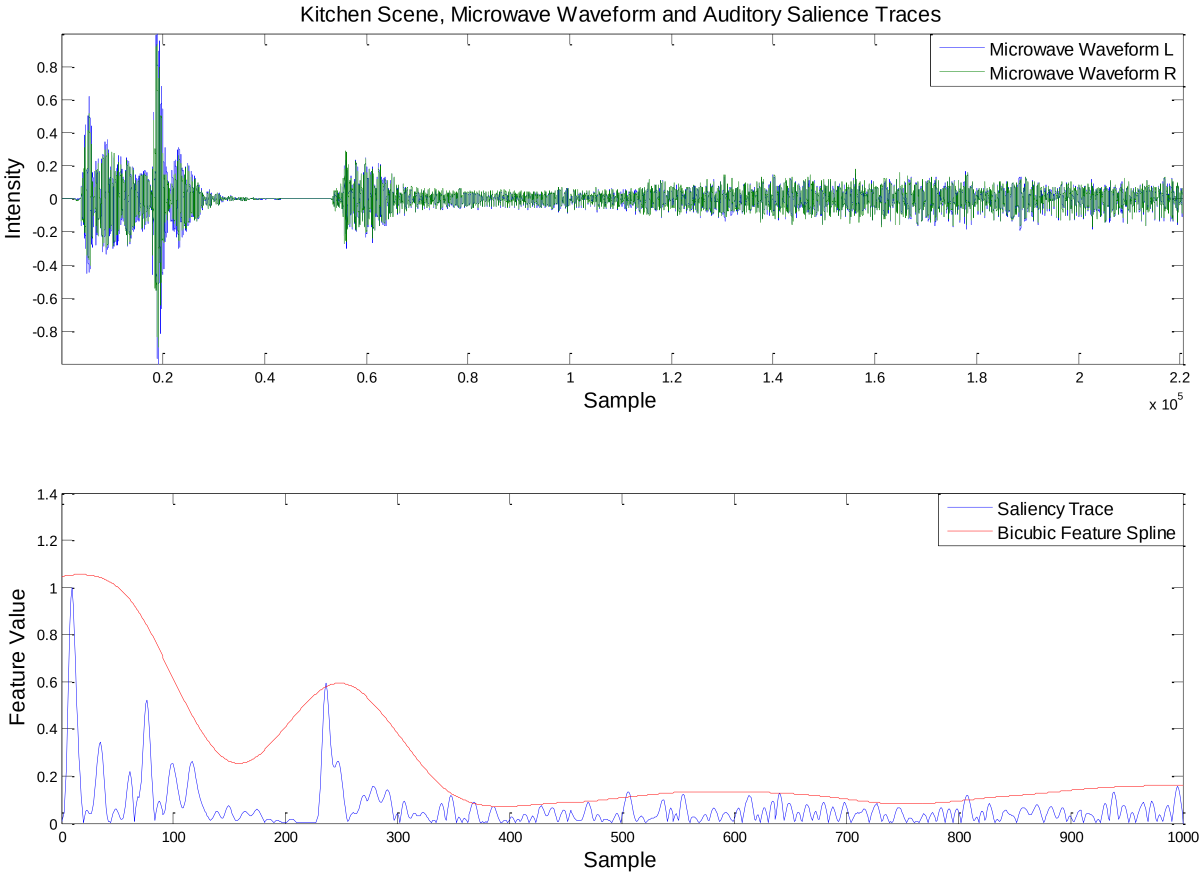


Figure 6: Microwave Waveform and Saliency Traces

were briefed prior to the experiment in order to gain a clear understanding of their task.

Each participant was assigned the 18 image sets in random order and the A or B image was randomised within that image (slide) set. Participants were presented videos in the order A→G→B with a decision slide that waited for an input: *"which video A or B is closest, in your opinion, to slide G?"*, where G is the gold standard reference, enforcing two alternative forced choice assessment. This ordering was chosen as opposed to side-by-side because the sound would be mismatched spatially if the participant had to look between screens. A→G→B→G heuristic could have been chosen but repeated exposure to the same material could introduce more bias and was deemed less appropriate. Image videos were presented asynchronously with the relevant modality for a total of five seconds each. Buffer slides, providing a visual cue (displaying A or B) as to the current video to be shown were presented for two seconds before the advent of the respective video. The decision slide halted the experiment and waited indefinitely for a response on the pairwise comparison (input A or B on the keyboard). Spatial sound congruent to the object in the scene was delivered to the participant for the full duration of the relevant videos.

5. Results & Analysis

This section presents the results of the experiment.

The overall similarity results for the 3 scenes are shown in Tables 3, 4 and 5. The paired comparison data is provided in Table 7 and coloured rings highlight that no significant difference resulted in between the selective rendering strategies on a per scene and/or overall basis.

5.1. Statistical Analysis

The null hypothesis is given as H_0', that all conditions are equal under testing ($H_0' : \pi_i = \frac{1}{2}$). The alternative being that not all the conditions π_i are equal. p_{ij} is the number of times that an image i is preferred to image j by a participant. The sum of this result per participant, excluding the condition where $i = j$, is given as Σ:

$$\Sigma = \sum_{i \neq j}^{t(t-1)} \binom{p_{ij}}{2} \tag{5}$$

where, t is the number of selective rendering strategies to be considered. Σ is the sum of the number of agreements between pairs. Kendall and Babington-Smith [KBS40] proposed a coefficient of agreement (also termed concordance) amongst the experiment participants defined as:

$$u = \frac{2\Sigma}{\binom{s}{2}\binom{t}{2}} - 1 \qquad (6)$$

where, s is the number of participants and $u = 1$ if all s participants made identical choices during the experiment. The less participants agree in their choices, the smaller u becomes.

If u is statistically significant then there are differences between the conditions and the null hypothesis can be rejected. The significance test of summed scores aims to find a value R' such that the probability $P(R \geq R') \leq \alpha$, where α is an arbitrary value $\alpha \in [0,1]$ and is typically assigned 0.05. The Σ for each condition presented which have differences of less than $\pm R$ is deemed to not be significantly different and the conditions can be perceptually grouped into the same categories. However, the conditions with different perceptual groups are declared to be significantly different when $\Sigma \pm R$ does not fall in range with other values of Σ and the condition is awarded a separate perceptual grouping.

If the score difference for a given scene between two rendering conditions is larger than R^+ (the smallest integer greater than R'), the conclusion is that there is a statistically significant difference between the two conditions presented and this indicates that one is perceptually closer to the ideal reference image than the other. A more complete write up of this statistical process is included as part of the supplementary material for the interested reader.

Preference Tables

Results for the computation are based on preference tables, which can be viewed as matrices in which one method was better than the other. The preference tables for each scene are presented in Tables 3, 4, 5 and combined in Table 6.

	Uni	Sal	SoSal	Mix	Score
Uni	*	10	8	6	24
Sal	18	*	11	9	38
SoSal	21	17	*	5	43
Mix	21	19	23	*	63

Table 3: Preference matrix for the lounge scene.

	Uni	Sal	SoSal	Mix	Score
Uni	*	13	5	7	25
Sal	15	*	13	4	32
SoSal	23	16	*	9	48
Mix	21	23	19	*	63

Table 4: Preference matrix for the kitchen scene.

	Uni	Sal	SoSal	Mix	Score
Uni	*	10	10	5	25
Sal	18	*	8	5	31
SoSal	18	20	*	8	46
Mix	23	23	20	*	66

Table 5: Preference matrix for the restaurant scene.

	Uni	Sal	SoSal	Mix	Score
Uni	*	33	23	18	74
Sal	51	*	32	18	101
SoSal	62	53	*	22	137
Mix	65	55	62	*	192

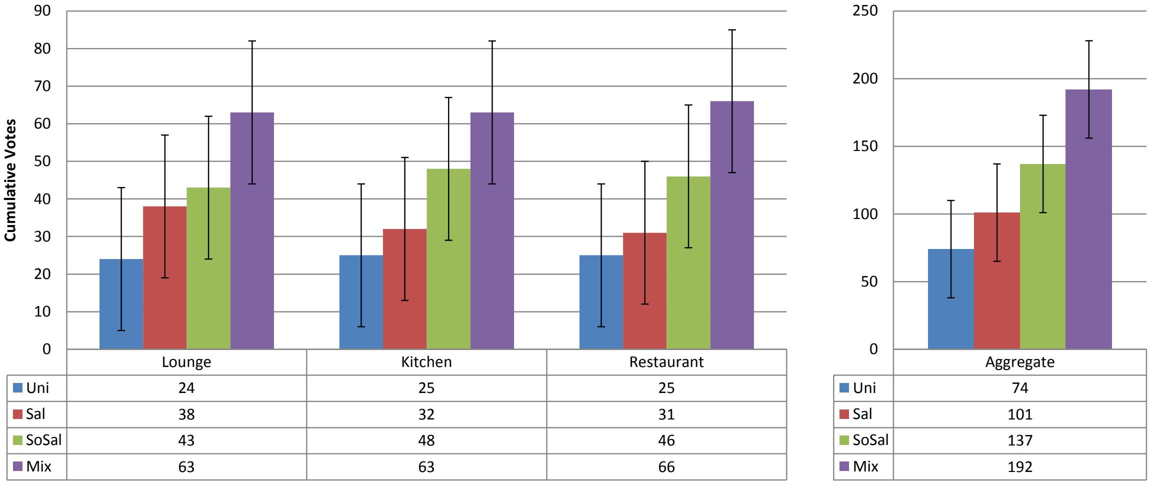
Table 6: Preference matrix for all the scenes combined.

5.2. Results

The results are shown in Table 7. In all scenes where there was ambiguity in the grouping, more than one technique was grouped together. The Mix operator came top of every preference table, had fewer discrepancies and was perceptually distinguishable statistically in two cases of testing whilst being in the top group of the third case. It is also first in the overall and statistically significantly better than the other methods.

The coefficient of consistency in this experiment was $\zeta_{average} \approx 0.75$ and as such the participant's consistency was deemed to be good and can all be included in the paired comparison study. The results provided an R^+ (the smallest integer greater than R') of 19. $\chi^2_{df=3,p<0.05} = 7.82, \chi^2_{df=3,p<0.01} = 11.35, \chi^2_{df=3,p<0.001} = 16.27$. This is shown against cumulative votes for each scene and aggregated in Figure 7. In addition the average coefficients of agreement and consistency are presented. As can be seen the results were perceptually distinguishable to a degree in all scenes. In all cases the null hypothesis is rejected and the multiple comparison range test can be used to find any pairwise difference scores equal to or greater than R^+ to be significant.

The Mix model described in this paper performed best in pairwise comparisons in every scene and render cost function. An ordering of Mix $\rightarrow$ SoSal $\rightarrow$ Sal $\rightarrow$ Uni prevailed throughout. The Lounge scene has $\chi^2_{3,0.001} < 32.8571$, Kitchen scene $\chi^2_{3,0.001} < 35.5714$, Restaurant scene $\chi^2_{3,0.001} < 38$ and H'_0 is rejected for all cases. The perceptual grouping was not clear cut in all cases, however the multiple comparison range test has the conspicuous property of making it difficult for true differences to show themselves. Yet the method allows comparisons to be performed after the initial inspection of experimental results and preference matrix generation. In addition, the probability of any incorrect declaration of grouping differences is controlled at the



	Lounge	Kitchen	Restaurant
🟦 Uni	24	25	25
🟥 Sal	38	32	31
🟩 SoSal	43	48	46
🟪 Mix	63	63	66

	Aggregate
🟦 Uni	74
🟥 Sal	101
🟩 SoSal	137
🟪 Mix	192

Figure 7: Method Preference; Error bars indicate the range R^+.

	$u_{average}$	$\zeta_{average}$	χ^2	P, df=3	Rank 1	Rank 2	Rank 3	Rank 4
Lounge	0.1658	0.7143	32.86	<0.05	Mix	SoSal	Sal	Uni
Kitchen	0.1825	0.7571	35.57	<0.05	Mix	SoSal	Sal	Uni
Restaurant	0.1975	0.7571	38.00	<0.05	Mix	SoSal	Sal	Uni
All	0.1819	0.7428	35.47	<0.05	Mix	SoSal	Sal	Uni

Table 7: Overall similarity study conclusion for the various scenes presented with spatialised acoustic stimuli and visual congruency's.

significance level reported in *P*. As such, any declaration of grouping is stringently correct. In this case, whilst no clear cut group exists for every set, the fact that under the same render costs different sampling strategies report a perceptual difference is an important result. What is also interesting is that the *auditory intensity map* presented in this paper also performs well, in the mid-range grouping with the Sal set. This is likely a result of the spatially encoded directional features of the audio. The answer may lie in temporal sensitivity of the HAS. However, this would require further investigation.

6. Discussion

The performance expected from the auditory attention models is limited by the features used in the models: intensity, frequency contrast, temporal contrast and cochlear response. Conventional auditory saliency models fail to perform tasks that require features which are not considered. For example, the model used in this paper uses monaural signals, and spatial cues are not considered. As a result, while the model is successful for tasks which are represented by at least one of the features of the model, it fails at the tasks which require spatial cues, such as localisation and sound source separation. This is accounted for by synergy between the presented auditory intensity map and auditory saliency feature vectors. The intensity map encodes features of saliency, localisation and separation in the visual domain that are not considered by the auditory saliency model alone. A combination of the feature vectors, intensity map and visual saliency map formed the temporal hybrid auditory-visual domain model. As this model proposed in this paper is a bottom up model, assuming no task driven intervention, this means that the method should be effective regardless the sensory content, however further work would have to investigate the impact of tasks on this scenario.

7. Conclusions and Future Work

This paper has presented a novel temporal hybrid multi-sensory saliency detection algorithm for use in the spatial visual domain. This method exploits the HSS's bottom-up approach. The results extend previous work of sound's combination with graphics and confirms the impact that the inclusion of sound into a high-fidelity virtual environments has for selective rendering. The algorithm meshes visual saliency and auditory saliency in a selective rendering pipeline to temporally and dynamically load balance computation. The algorithm is psychophysically evaluated on a number of scenes, and across a number of render cost functions to evaluate its performance. It is shown to perform significantly better than simple image saliency or acoustic intensity maps when they are used as a rendering strategy and is generic in its formation and application. In visual-auditory VR environments the presented algorithm accounts for visually important information when the auditory information presented is not deemed to be important and vice versa.

Future work will investigate the variability of the weighting function used across the different maps, especially investigating the effect varying frequencies have upon sound and directional attentional capture. This type of map cannot, currently, be used for realtime processing in virtual environments. The aim of the approach was to generate the auditory intensity map as a bi-product of the acoustic simulation step such that when hardware is more able to simulate closer to realtime the technique is more feasible. The type of simulation may be changed in the future to account for wave based effects of more recent sound simulation models. Phonon tracing was deemed appropriate for the inherent practicality of the sound cache schema, but especially low frequency diffraction effects need to be better accounted for. The scene types could be more varied in order to draw more general conclusions, however in terms of the technique presented, outdoor type scenes should be invariant to the results. It would be a logical progression to look into dynamic scenes with moving camera sequences and varying frequency lighting. Sound could be tested presented spatially, decoding ambisonics to 5.1 instead of binaural. Indeed even stereoscopic imagery could be an interesting tangent to the research. In evolutionary terms, certain sounds are more salient, and in fact, the pinna has the effect of amplifying these mid band frequencies down the auditory canal. In addition it is necessary to study the effect multiple sound sources have on visual attention. A first hypothesis would be the more salient source in the temporal domain would dominate spatially in the visual domain. A similar avenue of research is to study the intensity of the sound sources, specifically at which decibel level does the effect on visual attention come into play. Whilst the human ear can detect sound, the threshold of audibility remains true, but the effect to which salience takes precedence may not necessarily be linear in scale.

8. Acknowledgements

The authors would like to acknowledge EPSRC CASE studentship award number 07001392 without which this work would not have been possible. Also, ARUP for keen industrial support and in particular Steve Walker. Finally, the authors would like to acknowledge Elmedin Selmanovic for his valuable input to this work.

References

[BDM*05] Bertram M., Deines E., Mohring J., Jegorovs J., Hagen H.: Phonon tracing for auralization and visualization of sound. *In Proceedings of IEEE Visualization* (2005), 151–158. 2, 3

[Bla97] Blauert J.: *Spatial Hearing : The Psychophysics of Human Sound Localization*. M.I.T. Press, Cambridge, MA, 1997. 1

[BSVDD10] Bonneel N., Suied C., Viaud-Delmon I., Drettakis G.: Bimodal perception of audio-visual material properties for virtual environments. *ACM Trans. Appl. Percept. 7*, 1 (2010), 1–16. 1

[CCL02] Cater K., Chalmers A., Ledda P.: Selective quality rendering by exploiting human inattentional blindness: looking but not seeing. In *VRST '02* (New York, NY, USA, 2002), ACM, pp. 17–24. 1

[CDS*09] Coath M., Denham S., Smith L., Honing H., Hazan A., Holonowicz P., Purwins H.: An auditory model for the detection of perceptual onsets and beat tracking in singing. *Connection Science 21*, 2 (2009), 193–205. 6

[Coa05] Coath M.: *A Computational Model of Auditory Feature Extraction and Sound Classification*. PhD thesis, Centre for Theoretical and Computational Neuroscience, University of Plymouth, 2005. 6

[Dal] Dalenbäck B.-I.: CATT-Acoustic, Gothenburg, Sweden. www.netg.se/catt. 1

[Dal93] Daly S.: The visible differences predictor: an algorithm for the assessment of image fidelity. *Digital images and human vision* (1993), 179–206. 6

[Dal98] Daly S.: Engineering observations from spatiovelocity and spatiotemporal visual models. *Human Vision and Electronic Imaging III* (1998), 180–191. 2

[DM95] Durlach N., Mavor A.: *Virtual Reality Scientific and Technological Challenges*. Tech. rep., National Research Council Report, National Academy Press, 1995. 1

[DS98] Driver J., Spence C.: Attention and the crossmodal construction of space. In *Trends in Cognitive Sciences* (1998), vol. 2, pp. 254–262. 2

[FN05] Fujisaki W., Nishida S.: Temporal frequency characteristics of synchrony-asynchrony discrimination of audio-visual signals. *Exp Brain Res 166*, 3-4 (October 2005), 455–464. 4

[GBW*09] Grelaud D., Bonneel N., Wimmer M., Asselot M., Drettakis G.: Efficient and practical audio-visual rendering for games using crossmodal perception. In *I3D '09* (New York, NY, USA, 2009), ACM, pp. 177–182. 1, 3

[GLT05] Gallo E., Lemaitre G., Tsingos N.: Prioritising signals for selective real-time audio processing. In *Proceedings of Intl. Conf. on Auditory Display (ICAD) 2005, Limerick, Ireland* (July 2005). 2

[HDAC10] Hulusic V., Debattista K., Aggarwal V., Chalmers A.: Maintaining frame rate perception in interactive environments by exploiting audio-visual cross-modal interaction. *The Visual Computer* (2010), 1–10. 1, 2

[HHT*11] Hulusic V., Harvey C., Tsingos N., Debattista K., Walker S., Howard D., Chalmers A.: Acoustic Rendering and Auditory-Visual Cross-Modal Perception and Interaction. In *EG 2011 - State of the Art Reports* (2011), John N., Wyvill B., (Eds.), Eurographics Association, pp. 151–184. 1

[HOJ08] Hachisuka T., Ogaki S., Jensen H. W.: Progressive photon mapping. *ACM Transactions on Graphics (TOG) 27*, 5 (2008), 130. 3

[HS36] HECHT S., SHLAER S.: Intermittent stimulation by light: The relation between intensity and critical frequency for different parts of the spectrum. *Gen. Physiol. 19*, 6 (jul 1936), 965–77. 7

[HWBR*10] HARVEY C., WALKER S., BASHFORD-ROGERS T., DEBATTISTA K., CHALMERS A.: The Effect of Discretised and Fully Converged Spatialised Sound on Directional Attention and Distraction. In *TPCG '10* (2010), Collomosse J., Grimstead I., (Eds.), Eurographics Association, pp. 191–198. 2

[IKN98] ITTI L., KOCH C., NIEBUR E.: A model of saliency-based visual attention for rapid scene analysis, 1998. 2, 6

[KBS40] KENDALL M., BABINGTON-SMITH B.: On the method of paired comparisons. *Biometrika 31* (1940), 324–345. 9

[KDCM14] KOULIERIS G. A., DRETTAKIS G., CUNNINGHAM D., MANIA K.: C-lod: Context-aware material level-of-detail applied to mobile graphics. *Computer Graphics Forum 33*, 4 (2014), 41–49. 2

[KJM04] KAPRALOS B., JENKIN M., MILIOS E.: Acoustic modeling utilizing an acoustic version of phonon mapping. *In Proc. of IEEE Workshop on HAVE* (2004). 3

[KPLL05] KAYSER C., PETKOV C. I., LIPPERT M., LOGOTHETIS N. K.: Mechanisms for allocating auditory attention: An auditory saliency map. *Current Biology 15*, 21 (November 2005), 1943–1947. 1, 2, 7

[LDC06] LONGHURST P., DEBATTISTA K., CHALMERS A.: A gpu based saliency map for high-fidelity selective rendering. In *ARFIGRPAH '06* (New York, NY, USA, 2006), ACM, pp. 21–29. 6

[MBT*07] MOECK T., BONNEEL N., TSINGOS N., DRETTAKIS G., VIAUD-DELMON I., ALLOZA D.: Progressive perceptual audio rendering of complex scenes. In *I3D '07: Proceedings of the 2007 symposium on Interactive 3D graphics and games* (New York, NY, USA, 2007), ACM, pp. 189–196. 1, 2

[MD02] MARMITT G., DUCHOWSKI A.: Modeling visual attention in vr: Measuring the accuracy of predicted scanpaths. *In Eurographics* (2002), 217–226. 1, 2

[MDCT05a] MASTOROPOULOU G., DEBATTISTA K., CHALMERS A., TROSCIANKO T.: Auditory bias of visual attention for perceptually-guided selective rendering of animations. In *GRAPHITE '05* (New York, NY, USA, 2005), ACM Press, pp. 363–369. 1, 2

[MDCT05b] MASTOROPOULOU G., DEBATTISTA K., CHALMERS A., TROSCIANKO T.: The influence of sound effects on the perceived smoothness of rendered animations. In *APGV '05* (New York, NY, USA, 2005), ACM Press, pp. 9–15. 1, 2

[MMG11] MCNAMARA A., MANIA K., GUTIERREZ D.: Perception in graphics, visualization, virtual environments and animation. In *SIGGRAPH Asia 2011 Courses* (New York, NY, USA, 2011), SA '11, ACM, pp. 17:1–17:137. 2

[Moo97] MOORE B. C.: *An introduction to the psychology of hearing*. Academic Press, 4th edition, 1997. 2

[Nay93] NAYLOR J.: ODEON - another Hybrid Room Acoustical Model. *Applied Acoustics 38*, 1 (1993), 131–143. 1

[Pet99] PETTERSSON R.: Attention an information design perspective! International Institute for Information Design (IIID), Vienna, Austria, 1999. 1

[PS00] PAINTER E. M., SPANIAS A. S.: Perceptual coding of digital audio. *Proceedings of the IEEE 88*, 4 (april 2000). 2

[RBF08] RAMANARAYANAN G., BALA K., FERWERDA J. A.: Perception of complex aggregates. In *SIGGRAPH '08: ACM SIGGRAPH 2008 papers* (New York, NY, USA, 2008), ACM, pp. 1–10. 1

[RFWB07] RAMANARAYANAN G., FERWERDA J., WALTER B., BALA K.: Visual equivalence: towards a new standard for image fidelity. *ACM Trans. Graph. 26*, 3 (2007), 76. 1

[RLC*07] RAGHUVANSHI N., LAUTERBACH C., CHANDAK A., MANOCHA D., LIN M. C.: Real-time sound synthesis and propagation for games. *Commun. ACM 50*, 7 (2007), 66–73. 1

[SLKS07] SILTANEN S., LOKKI T., KIMINKI S., SAVIOJA L.: The room acoustic rendering equation. *J. Acoust. Soc. Am. 122*, 3 (Sep 2007), 1624–1632. 3

[Sur12] SURFACES A.: Sound absorbtion coefficients. World Wide Web electronic publication, 2012. 6

[TGD04] TSINGOS N., GALLO E., DRETTAKIS G.: Perceptual audio rendering of complex virtual environments. *ACM Trans. Graph. 23*, 3 (2004), 249–258. 1

[Tsi05] TSINGOS N.: Scalable perceptual mixing and filtering of audio signals using an augmented spectral representation. *Proc. of 8th Intl. Conf. on Digital Audio Effects (DAFX'05), Madrid, Spain* (Sept. 2005). 2

[WW80] WELCH R., WARREN D.: Immediate perceptual response to intersensory discrepancy. In *Psychological Bulletin* (nov 1980), vol. 88(3), pp. 638–667. 2

[Yee00] YEE Y. L. H.: *Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic enivroments*. PhD thesis, Cornell University, Aug 2000. 2

[YPG01] YEE H., PATTANAIK S., GREENBERG D. P.: Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments. *ACM Transactions on Graphics 20*, 1 (January 2001), 39–65. 2, 6