Volume xx (200y), Number z, pp. 1–12

Audio-Visual Resource Allocation for Bimodal Virtual Environments

E. Doukakis[†‡], K. Debattista, C. Harvey, T. Bashford-Rogers, and A. Chalmers

WMG, University of Warwick, UK.

Abstract

Fidelity is of key importance if virtual environments are to be used as authentic representations of real environments. However, simulating the multitude of senses that comprise the human sensory system is computationally challenging. With limited computational resources it is essential to distribute these carefully in order to simulate the most ideal perceptual experience. This paper investigates this balance of resources across multiple scenarios where combined audio-visual stimulation is delivered to the user. A subjective experiment was undertaken where participants (N=35) allocated five fixed resource budgets across graphics and acoustic stimuli. In the experiment, increasing the quality of one of the stimuli decreased the quality of the other. Findings demonstrate that participants allocate more resources to graphics, however as the computational budget is increased, an approximately balanced distribution of resources is preferred between graphics and acoustics. Based on the results, an audiovisual quality prediction model is proposed and successfully validated against previously untested budgets and an untested scenario.

Keywords: Multi-Modal, Cross-Modal, Bi-modal, Sound, Graphics

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

1. Introduction

While computer graphics has improved significantly in recent years and is still the focus of much research, visual stimuli are seldom presented in isolation and for most applications are complemented by other senses. Such applications, termed Virtual Environments (VEs), provide the opportunity to simulate a wide range of applications, from training to entertainment, in a safe and controlled manner. For applications which require high levels of authenticity, for example commercial driving simulators, the VEs need to provide multiple, physically accurate sensory stimuli [GGB05]. However, computing and delivering multiple sensory stimuli in high-fidelity requires significant computing resources. Previous work has shown that humans are not able to fully attend to all the sensory stimuli in a real environment [SM93]. While many limitations of the Human Sensory System (HSS) have been previously used to reduce computational requirements [HHD*12], precisely how best to allocate a computational budget in the simulation of multiple stimuli, has not, to our knowledge, been investigated previously.

The work presented in this paper examines how the budget of available computational resources affects the perceived fidelity of an audio-visual experience. The goal is to consider how to best distribute a limited computational budget across the senses of vision and hearing. A model is presented that can be used for generic scenarios. The model is based on a subjective experiment whereby participants' quality preferences when allocating resources were captured. This model is then evaluated with different, previously untested budgets and previously tested and untested scenarios. Rendering systems could make use of such a model to direct resource allocation given a computational budget (which can be calcu-

† e.doukakis@warwick.ac.uk

‡ Submitted: March 2016; Acccepted: June: 2017.

lated on the fly) thus providing a superior cost-performance ratio based on perceptual functionality. Figure 1 presents an overview of the work.

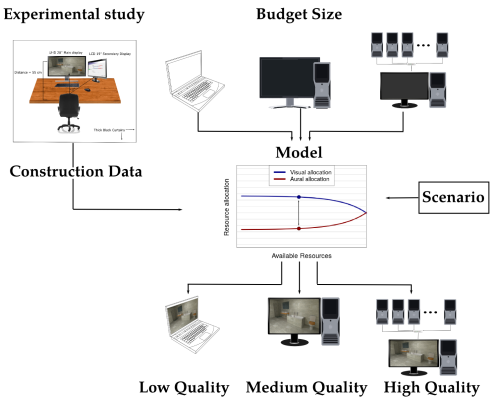

Figure 1: Diagram that describes the inputs and output of the prediction model. The model was constructed from data obtained from the experimental study.

Based on the phenomenon of visual dominance over the other senses [PNK76] in many tasks, we hypothesise that, generally, graphics will dominate the resource allocation. However, we investigate whether this is also the case when the presented auditory stimuli contain meaningful information (e.g., human voice). In this work, we find evidence that people tend to favour graphics quality; however as more computational resources become available to the users, a balanced distribution is preferred. It also appears that the content of the auditory stimulus biases the human allocation criteria.

The current allocation results are limited on the selection of quality metrics for audio-visual stimuli. This study focuses on a small number of variables, however, different selection of metrics for the two senses might infere different results, a question that will be investigated in future work.

The main contributions of this work are:
- A novel methodology for investigating the perceived contribution of the senses of sight and hearing within bimodal virtual environments,
- An investigation into the perceptual importance of vision and audio modalities under different amounts of available computational resources,
- Evidence that generally participants prefer to allocate more resources for increasing visual quality,
- An indication that as the available resources increase the devoted percentage to audio quality increases significantly and the audio-visual resource distribution is balanced, and,
- A validated prediction model that can be used to guide resource allocation in rendering systems.

2. Background and Related Work

Multi-sensory VEs aim to provide accurate representations of real world environments by simulating multiple senses at the same time. The coexistence of many diverse stimuli is able to provide a realistic experience and increase users' overall level of immersion [DM95]. Applications that make use of VEs with multiple senses range across many different industrial and educational areas. Examples include video games [MBT*07], concert hall [Cat] and architectural design [Nay93] etc.

Research in psychology has considered the perception of individual senses separately [Bro58, BS06, Pyl06, Sch01], and across different modalities [DS98, Duf99, BdG04]. Although, understanding of the perception of individual senses is important, in reality, exposure to stimuli affecting solely one modality is rare. Furthermore, the impact of one sense on another can be significant, for example, the ventriloquism effect in which the viewer is fooled into thinking that the sound source emanates from a visual cue [HT66].

The computation of physically accurate sensory stimuli requires the devotion of a large amount of resources and it is still unachievable in real time despite the significant algorithmic and hardware advances. Many limitations of the HSS have been exploited in the past in order to reduce the computation complexity of bi-modal rendering systems with little or no perceived quality difference of the delivered visual or auditory stimuli, for example [SZ00]. Perceptual interactions between the senses of vision and hearing result due to the different spatial and temporal properties of the respective sensory systems. Particularly, the Human Visual System (HVS) is more sensitive to spatial variations while the Human Auditory System (HAS) prevails in temporal oriented tasks [FN05].

Many selective rendering frameworks have utilized bimodal interactions in order to achieve computational gains without any perceived loss in quality either on the visual or auditory domains. In the following we present previous work where the influence of one modality has resulted in a reduction of the perceived quality in the other sense, termed the 'target' modality. In many cases these cross-modal interactions lead to computational savings in the target modality while the overall virtual experience is perceived unchanged.

2.1. Auditory influence on Vision

An early study on auditory-visual cross-modal interactions demonstrated that quality in VEs depends on both auditory and visual components [Sto98]. The author showed that high-quality audio increases the perceived quality of high-quality video. Furthermore, high-quality video decreases the perceived quality of low quality audio.

Mastoropoulou *et al.* [MDCT05] studied the effect of sound emitting objects (SEO) on rendering animations. Low quality graphics were presented in unattended areas while

high quality graphics were delivered in the area surrounding the SEO. The results indicated that users failed to notice the quality difference while the overall rendering times were decreased. Further work in this area by Harvey *et al.* [HWBR*10] included the application of spatialized sound in the VE context. They showed via eye tracking that visual attention is significantly affected by the direction of the incoming auditory stimulus.

Further experimental studies showed that pairing an animation with congruent or incongruent sound effects can partially affect viewer's attention and temporarily distract them from the visual domain [HAC08]. This example allowed a reduction in frame rate without the user perceiving any quality degradation [HCD*09].

2.2. Visual influence on Audition

The sense of vision can also decrease the perception of auditory stimuli. This is used to reduce auditory rendering complexity. In a study by Moeck *et al.* [MBT*07], hierarchical clustering of all the available sound sources is applied while only the salient features of the audio signal are considered for rendering sound. The concurrent application of both vision and audio stimuli has been shown to influence the quality perception of materials when adjusting the quality level of the delivered audio stimulus [BSVDD10]. In another study, Grelaud *et al.* [GBW*09] accelerated audio rendering by using an audio-visual Level-of-Detail (LOD) selection algorithm. This algorithm can detect when virtual object collisions occur in a VE and dynamically adjust the computation resources needed for the two senses depending on the number of collisions in the VE. In this study the variability of the available resources was user defined and not shown to change with different tasks.

2.3. Multisensory Integration

Burr and Allais [BA06] proposed a framework in which bimodal information can be combined as a sum of all individual stimuli estimates weighted appropriately. The estimate can be calculated as $\hat{S} = w_A\hat{S}_A + w_V\hat{S}_V$, where w_A and w_V are weights by which the individual stimuli are scaled, and $\hat{S}_A$ and $\hat{S}_V$ are independent estimates for audition and vision respectively. This has been tested using different visual stimuli with different levels of blurriness [AB04]. An example where audio dominates the overall estimate $\hat{S}$ occurs when visual stimuli are corrupted by blurring the visual target over a large region. The blurring, however, has to be significant i.e. over about 60°, which makes most scenes unrecognisable.

The relative importance of audio and visual stimuli has also been investigated when assessing the quality of multimedia applications. In a study by Pinson *et al.* [PIW11], the authors reviewed a series of experiments that were used to evaluate the importance of each of the two senses in a range of different source materials, video resolutions, audio and video compression schemas. Each of these experimental studies produces a model that predicts audio-visual quality based on the measured visual, auditory and combined audio-visual mean opinion scores (MOS). In another study by Belmudez and Möller [BM13], the effect of audiovisual quality is investigated in the context of interactive communication services. The authors test the effect of bimodal interactions in a range of different experimental conditions including both passive and interactive contexts (viewing and listening). They propose an integration function that predicts audio-visual quality assessment depending on the type of communication application. Although these methods investigate the relative importance of one sense over the other in a range of case studies they do not explicitly consider how this relationship changes when the computational budget varies.

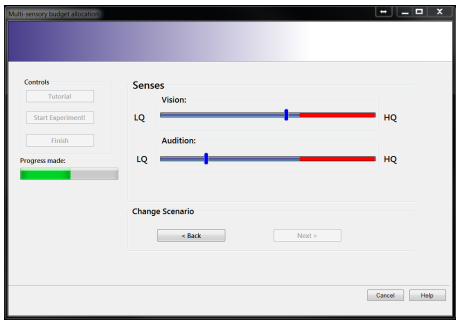

Figure 2: Snapshot of the experimental software used. It shows an instance of budget size B_3 and all the GUI controls.

None of the methods discussed in the related work have explicitly looked into the required fidelity per sense as computational resources change. Techniques have investigated the benefit of resource fluctuation within a multi-sensory context, based upon perceptual observations, but no work has looked into quantifying this process.

3. Experimental Framework

This section introduces the design for this experimental study followed by the material preparation. Details for the participants and the experimental procedure conclude the framework description.

3.1. Design

The objective of this work is to study users' allocation preferences when distributing visual and aural resources while the available computational budget varies. Resource allocation is measured over a series of experimental scenarios whereby participants are asked to assign as much of the budget they want to the rendered audio and visual stimulations.

Our hypothesis is that high quality imagery is generally

appreciated more, compared to audio, when experiencing a virtual world [Col74]. We investigate how this visual-audio importance relationship scales across different budget sizes that can be available to the user and when the delivered auditory stimulus might contain meaningful information (e.g., dialogue). We make use of the collected experimental data to build a prediction model that can estimate resource allocation depending on the budget and the scenario.

This experiment was designed using rendered images and audio where the quality level of each of the two senses could be adjusted interactively using the controls of a graphical user interface (GUI); physically-based simulations were used for the computation of both aural and visual stimuli. The assigned budget corresponds directly to quality improvements. In turn, these quality improvements in vision and audio directly correspond to a higher computational cost that is allocated from the available budget. The experiment requires the users to move interlinked sliders that modify the quality of displayed graphics and acoustics interactively. When one is increased the other is automatically decreased. Figure 2 shows an example of the GUI used for modifying the sliders.

The experiment was conducted with five distinct budgets across five different scenarios. The experimental design is within participants and each participant is asked to compute the best perceived quality for all budgets and scenarios. The presentation of the 25 possibilities was randomized to avoid any potential ordering bias.

The rest of this section explains what metrics are used for creating the quality levels for vision and audio, and provides the definitions of computational cost and computational budget as they are used in this work.

3.1.1. Visuals

For visuals, resolution was chosen as the variable that modifies quality. A number of standards already exist for resolution and it is possible to be abstracted from the underlying algorithm used for image synthesis. The computational cost of image rendering changes approximately linearly with resolution for many ray-tracing methods and its derivatives. Other variables that can admittedly modify quality such as lighting models, texture mapping, shaders, etc. are kept fixed in the following experiments; they have not been chosen for modification in order to avoid complex scenarios that would result in a combinatorial explosion of possibilities.

All the images were computed using path tracing [Kaj86] due to its accuracy and straightforward nature of the computation. The objective was to create realistic scenes that present physically correct illumination and material properties, and may be representative of future rendering systems. The images are rendered to convergence to generalize this work to any possible algorithm. 240 images were computed, which varied in resolution from 16×9, to the highest resolution at Ultra HD (UHD) (3840×2160). The lowest resolution was chosen to reflect the level at which humans find it difficult to identify images [Tor09], and a fixed aspect ratio of 16:9 was used for all the intermediate images.

In order not to introduce any bias due to the presented size, equally-dimensioned images of different quality were preferred, thus every one of these images was resized back to the UHD resolution used by the display hardware. This resizing process was implemented using a bi-cubic interpolation kernel while anti-aliasing and colour dithering filters were used in order to keep the quantization error to a minimum. Bi-cubic interpolation was preferred over other image scaling methods because it produces smooth images and its application has no significant computational cost relative to the overall rendering costs.

The computation time needed for an image of this sequence can be generally estimated as follows:

$$C_k^V \approx P_k \cdot L, \quad k = 1, 2, \cdots, 240, \tag{1}$$

where $P_k = 16 \cdot 9 \cdot k^2$, is the number of pixels of the k-th image and $L = C_{240}^V / P_{240}$ is the time needed for computing an individual pixel. We do note that this estimation, strongly depends on the available hardware, the algorithm used and the applied software optimisations. Other parameters that affect the computation time include, but are not limited to, the scene complexity and material and texture properties. However, in order to generalize the results of the experiment we make the assumption that the computational cost is varying linearly with resolution. Furthermore, we decouple the problem from real time measurement and consider the budget in terms of normalized cost that can be expressed by equation (1) using the normalisation factor $1/C_{240}^V$. This results in visual levels with costs independent of the underlying algorithm used for the computation. Therefore, in what follows, we will use the term *visual cost* to denote the quantity given by:

$$C_k^{N_V} = \left(\frac{k}{240}\right)^2, \quad k = 1, 2, \cdots, 240. \tag{2}$$

For the correct implementation of the experiment, we considered only the subset of images that perceptually differed in quality. Pairs of images that elicit the same perceptual response and have different costs might result in false conclusions. The Human Visual System (HVS) is more capable to distinguish large quality differences at lower resolutions rather than at higher quality levels. In this work we made use of a visual perceptual metric that can discard similar quality image levels from the original sequence.

The High Dynamic Range Visible Difference Predictor (HDR-VDP) [NMPDSLC14] is a widely used objective metric for detecting the perceptual differences between High Dynamic Range (HDR) or Low Dynamic Range (LDR) image pairs. The model provides the Q correlation measure, a

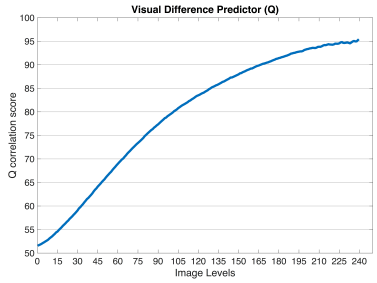

Figure 3: Average values of all Q correlation results obtained from all six scenarios that considered in the experimental study.

numeric score that ranges between 0-100. Low Q values indicate apparent visible differences between the input images while two images with maximum Q score are considered perceptually indistinguishable. In this study, we performed pairwise comparisons between the highest resolution image (UHD) and the 239 other rescaled images using the LDR mode of the metric. The averaged Q scores for all the six scenes used in the experiment are depicted in figure 3. As expected, the results follow a logarithmic trend indicating that at higher levels participants would struggle to find apparent visual differences.

Varshney and Sun [VS13] show the internal representation of a stimulus scales in a logarithmic fashion for increasing stimulus intensities. They argue that any range of stimulus physical intensities is mapped through this log-curve to a finite set of perceptual points that constitute distinct sensory levels for the observer. This mechanism resembles signal quantization where the internal perception points are uniformly spaced and very high intensity values have little or no impact to the internal representation (upper quantization boundary). This argument is true not only for the visual domain but also for other human sensory systems [VS13].

We followed this research outcome in order to discretise the Q values to 80 discrete levels that elicit noticeable perceptual differences to the user. This number is the minimum possible number of levels in order that the resulting set of images includes the majority of common resolution standards that are frequently used in High Definition Television (HDTV) and Standard Definition Television (SDTV) applications. These are:

- quarter High Definition (qHD or 960×540),
- High Definition (HD or 1280×720),
- Full High Definition (FHD or 1920×1080),
- Ultra High Definition (UHD or 3840×2160).

3.1.2. Auditory

For the auditory part of this study, we computed a *Room Impulse Response* (RIR) for the same set of scenes as used in the visuals. A RIR encodes delays and attenuations of the sound waves, for a given sound source, as they interact with the environment's surfaces before reaching a listener. The process of *Auralization* includes computing the RIR, and convolving it with an anechoic stream to produce an auditory response. This response encodes frequency dependent characteristics and reverberations of the sound in the environment.

The anechoic stream is a recording of the sound without any reverberation effects. These are typically recorded in an anechoic chamber, or computed from a recording in an environment with a known RIR, and then extracted via deconvolution. The auralization of an anechoic stream is completed by convolving with the generated RIR. This process generates physically-based audio as being listened in the actual environment.

A ray-tracing approach was also adopted for the acoustic rendering [STKS07]. In this study, we rendered a B-format (Ambisonics) RIR at 352800 Hz sampling frequency as the highest temporal resolution for sound. Such high sampling rates are used in Digital extreme Definition (DxD) applications for editing high-resolution audio tracks.

Similarly to visuals, 240 RIRs were computed at 240 different sampling rates for each of the scenes. Every RIR is interpolated up to the maximum sampling rate of 352800 Hz for playback using bandlimited interpolation and specifically a *sinc* interpolation kernel. This process, effectively, oversamples the RIR signal as the target sampling rate is significantly higher than the Nyquist frequency. The oversampling operation does not affect the quality of the RIR and it is implemented so as to have RIRs which have the same sampling rate as the anechoic streams before performing convolution of the two signals. The anechoic sound was also oversampled using a *sinc* interpolation kernel at 352800 Hz before convolving with the RIR.

After the sampling rate conversion every RIR is convolved with a two-channel head related impulse response (HRIR) signal using the CIPIC HRTF database [ADTA01]. The HRIRs capture all source localisation cues and they encode elevation and azimuth effects of the incoming sound signal to listener's ears. These cues are affected by human anatomical characteristics (head size, shoulder, pinnae shape, etc.) and the HRIR content is used to capture these interactions.

Estimating the computation time needed for rendering the k-th RIR of sampling rate $f_k < f_{240}$, where $f_{240} = 352800$ Hz, was computed as:

$$C^{N_A}_{f_k} \approx \frac{f_k}{f_{240}}, \quad k = 1, 2, \cdots, 240. \tag{3}$$

The normalisation factor $1/C^{A}_{f_{240}}$ was directly applied to indicate that this quantity represents normalised and not physical computational *audio cost*. The non-normalised version of equation (3) reflects the fact that a RIR with half the

number of samples per second needs approximately half the computation time to be rendered.

As with visuals, this formula gives a coarse estimation of the computation time and other factors may affect the estimation, such as, the reflection properties of the environment surfaces under a multitude of different frequencies, the relative position of the listener and receiver inside the geometry, and the method used to compute the RIR.

The sampling rates of all the audio levels were determined using the inverse of equation (3) and applying, as costs, the values given by equation (2). Specifically, the k-th RIR is given a sampling rate of:

$$f_k(C_k^{N_V}) = \lceil C_k^{N_V} \cdot f_{240} \rceil, \quad k = 1, 2, \cdots, 240. \qquad (4)$$

where the $\lceil \cdot \rceil$ operator was used to obtain the closest higher sampling rate value. In practice, the application of equation (4) makes the two senses comparable in terms of their normalized cost; a practical property that allows visual-audio cost interactions without worrying about which particular algorithm was used for computing stimuli of either sense.

The resulting sequence of audio levels included RIRs of higher sampling rate until the last rendered RIR. Lower auditory quality levels could be heard as "dim" sounds because only low frequencies are present in the respective signals while higher auditory levels were rich in harmonics making the sound perceived brighter and clear.

This is analogous to the visual domain quality differences in resolution. The progressive existence of high frequency components makes the process of distinguishing differences between higher quality audio pairs quite difficult. As was the case with visuals, it was important to retain audio levels that elicit different perceptual stimulations.

To the best of our knowledge, there is no analogue of the HDR-VDP in Acoustics that can be used to estimate perceptual differences between two audio tracks. There are many parameters that affect the process of finding audible differences, such as: loudness, pitch, duration, room clarity, instantaneous sound energy, content (e.g. music or speech), sound reflection properties of the room, etc. It is also important to note that identical audio stimuli do not necessarily elicit the same perceptual response on a person by person basis [Est76], making the development of such an objective metric more complicated.

However, the frequency Just Noticeable difference (JND) in Acoustics is 1Hz for complex sounds in the range of 500–1000 Hz while it progressively increases for sounds of higher frequencies [KBM08]. This means that it is easier for a human listener to distinguish two different sounds when their content is at low and mid-frequencies rather than at higher frequencies of the audible spectrum.

In this study, we kept again 80 auditory levels using the same log spacing distribution as we applied in the visual domain using the VDP results. This log-spacing keeps the majority of the low sampling rate RIRs and as a result the convolved sounds have low frequency components that satisfy the JND requirement of 1 Hz apart. Therefore, they are perceptually distinguishable to the average user.

3.1.3. Visual-Auditory cost interactions

In this study we used normalised cost functions for both visual and auditory levels. Also the two senses follow a similar cost distribution as the costs for visual stimuli were used to find the sampling frequencies of the 240 original auditory levels. These assumptions overcome the problem of the unbalanced physical computation time needed for rendering the high quality image compared to that of the high quality RIR. These assumptions do not affect the generic scope of the study which is to find the relative importance of each of the two senses in terms of the percentage of the total budget that is devoted to each of them.

The notion of "budget" as it is used throughout this study adapts to the idea of the normalised cost and represents a theoretical quantity that is distributed amongst the costs of a visual and an audio quality level. In this study, five different theoretical budgets are used; these are shown in Table 1 along with their notation letters. The number of different budget sizes allows investigation on how the allocation preference scales for very small budget sizes (B_1), where quality improvements are constrained, up to larger budget sizes (B_4, B_5) where the user is able to select relatively high quality levels in both of the senses.

Table 1: All the theoretical budgets used in this experimental study along with their notation letters and the number of levels remained when a budget is applied.

Budget letter	**B_1**	**B_2**	**B_3**	**B_4**	**B_5**
Budget	0.0625	0.11	0.25	1	1.12
Total Number of Levels	28	38	48	80	48
Vision (resolution)	qHD	HD	FHD	UHD	UHD
Audio (kHz)	22.05	38.8	88.2	352.8	352.8

The budget selection was determined such that the maximum quality level in vision or audio corresponded to one of the commonly used resolution and sampling rate technology standards. These are listed in Table 1 along with the total number of quality levels that remain when one of the five budgets is used in the experiment.

Budgets B_1, B_2, B_3 use a subset of the original set of quality levels while budget B_4 contains the original number of distinct levels. Budget B_5 does not add new quality levels but instead the user starts improving the quality from medium visual-auditory quality to compensate for the extra budget

amount, $B_5 - B_4$, that is given. B_5 is the only budget size that allows the user to select the maximum quality in either of the two senses and at the same time receive a medium quality stimulus from the other sense.

3.2. Materials

This study used a $28''$ Samsung U28D590 ultra HD LED monitor to display the LDR images at resolution 3840×2160 while a Dell UltraSharp 2007WFP $19''$ LCD monitor was used to display the GUI. All the sounds were delivered binaurally using a set of Sennheizer HD 380 pro headphones.

The distance from the participants' heads to the main display and the rotation angle needed to watch the contents of the secondary display are shown in Figure 4. These follow the guidelines of the ITU-R BT.500-13 standard [ITU12] for adjusting viewing conditions in subjective evaluations using HDTV and SDTV display panels. The experimentation procedure was conducted in a dark and silent room for reducing external disruptions during the experiment.

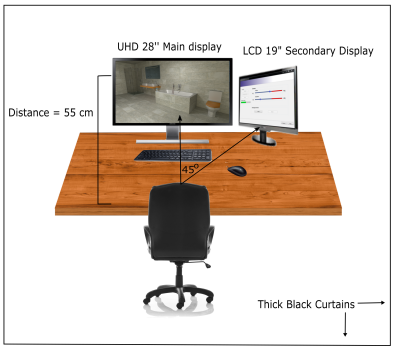

Figure 4: Hardware setup for the experimental study. The viewing distance to the main monitor and the angle needed to see the contents of the secondary monitor are also given.

A total of six scenarios were used for the experiment. These are labeled: Bathroom, Car, Kitchen, Kitti, Restaurant and Yard. The Bathroom scene was used for training before the formal experimental session commenced and it was not used for collecting experimental data during this experiment. Images of the six scenarios are given in Figure 5.

The sound stimulus in the Kitchen scenario was originating from a coffee maker that was visible to the subjects. The Yard scenario consisted of the sound of the engine ignition of a visible lawnmower at the right hand side of the screen. The Kitti sound stimulus was a hymn where the sound source originated from behind the main church tableau. For the Restaurant scenario the sound of an employee's voice delivering food behind the wall was used while there was clatter noise coming from the same position. The Car scenario sound includes both engine ignition and the putting on of a seat belt. The Bathroom scenario included the sound of water flowing in the shower. The objective was to present both indoors and outdoors scenes where the audio stimulus varied from mechanical everyday sound to human speech and melody.

A custom GUI application was developed for this subjective study. The interface included two main windows, one displayed the images in full screen at the main monitor and the other contained a window with the basic controls for adjusting the quality of the two senses. Two slider thumbs were used to dynamically change the quality of the presented visual and audio stimulus by taking values that correspond to different quality levels.

The effect of different budget sizes is shown in the GUI by colouring red the portions of the slider bar which contain quality levels that cannot be presented with the given budget. An example of the GUI configuration is given in Figure 2. This image presents an instance of the medium budget size B_3 used while it depicts all the controls that composed the GUI. The visual slider bar, along with its label, was randomly interchanged with the auditory at different experimental trials to avoid participants' adaptation to control only one of the bars for resources allocation.

At the beginning of each trial the two sliders were located at a "null" stimulus mode. This configuration includes the presentation of a zero cost image and zero cost audio to the subjects. A grey image was utilised as a zero cost image while a silent track was the zero cost audio track. Grey images are frequently used in experimental studies as a mean of neutralising participant's eyes before the next experimental trial [ITU12]. The two thumbs always start from the "null" stimulus (grey image/silent track) at the beginning of each trial. This is not the case when the trial includes the distribution of budget B_5. In that case, both thumbs start from an audio-visual level that has theoretical cost equal to $B_5 - B_4$ and corresponds to a medium quality audio and image. Using this thumb configuration the user was able to start exploring quality levels from the beginning before deciding the desired quality.

When a trial starts, the two sliders are independent of each other until the first time the user tries to exceed the budget given for the trial. In that case, the slider that is not currently controlled by the user is adjusted so as the two costs sum up to the total budget. For example, in a trial where the budget is $B_3 = 0.25$ and the 86-th and 75-th levels are selected for vision and audio respectively the two thumbs are still independent because the total cost is $0.225 < B_3$. When the user increases the visual quality to the 95-th level, the total cost exceeds the given budget and the audio slider is automatically moved to the 73-rd level making the total cost equal to the budget. The dependency of the sliders remains until the user continues to the next trial. This dependency allows the user to realise the constraint of the budget in the resource allocation task. If the sliders were always independent on each other, the task of allocating resources would be pointless as the users could maximise the qualities for both senses, in that

case, the total cost of the audio-visual stimuli would always go beyond the budget of the trial. The independence of the sliders at the beginning serves to not bias the participants' allocation preference.

3.3. Participants

A total of 35 participants, 18 men and 17 women with ages between 19 and 53 years and from various academic and working disciplines took part in the experiment. The average age of all the participants was 31.6 years. All the participants had normal or corrected to normal vision and reported no hearing problem.

3.4. Procedure

Figure 5: The visual scenarios used at this experimental study, from left to right and top to bottom: Bathroom, Car, Kitchen, Kitti, Restaurant, Yard.

Each participant undertook every possible combination of budget size and scenario for a total of 25 experimental trials. The tutorial scenario was also combined with all budgets for five test trials where the users could familiarise themselves with the GUI, the budget configurations and the task of allocating the resources. There was no constraint on the training time and participants could repeat any of the test trials as many times as they thought necessary. Before the main session, the participants were asked if they thoroughly understood the task of allocating the available resources to the two senses in order to create a satisfactory virtual experience based upon what they saw and heard.

4. Results

In this section the results of the experimental study are presented. The percentage devoted to graphics over the total budget is considered as the dependent variable. Audio allocation is deterministically given from the graphics allocation percentage.

Graph 6 depicts a summary of the experimental data. As budget size increases, vision allocation percentage decreases. In other words, vision quality is considered critical when limited resources are given while this trend gradually changes when more resources become available to the users. At higher budget sizes participants appear to prefer a more balanced distribution of resources for the two senses. The same Figure also shows how graphics allocation percentage varies at different experimental scenarios for a given budget. Vision allocation percentages for Kitti and Restaurant were generally lower than the other two scenarios, indicating a preference for allocating less visual resources in favour of higher aural quality improvements for these scenarios.

The analysis of the data included the application of a 5 (budget) $\times$ 5 (scenario) repeated measures ANOVA as every participant was tested at all possible experimental conditions. This design allows to separate the variability due to the individual subjects from the variation that is explained due to the independent variables.

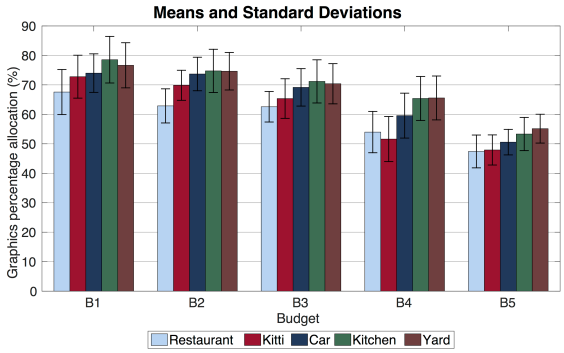

Figure 6: Means and Standard deviations for the graphics allocation percentage across all five budgets and experimental scenarios.

For the overall effect of budget, Mauchly's test showed a violation of sphericity against the budget size ($W(9) = 0.149, p < 0.05$) so Greenhouse-Geisser correction with $\varepsilon = 0.56$ was applied. The results revealed that budget size was statistically significant, $F(2.27, 77.2) = 37.5$ and $p < 0.05$. The overall effect of the scenario was also found significant, $F(1.87, 63.6) = 7.27$ and $p < 0.05$; sphericity test was violated ($W(9) = 0.154, p < 0.05$) and Greenhouse-Geisser correction with $\varepsilon = 0.46$ was applied. The mutual interaction of budget and scenario did not yield statistical significance, $F(16, 544) = 37.53$ and $p > 0.05$ suggesting that their common interaction does not affect a user's allocation preferences. This can be explained due to the process of normalisation as the scenario complexity does not depend anymore on the budget devoted for its computation.

Table 2: Contrast comparisons between budgets at every scene and across all scenes. Budgets with no significant differences are grouped together.

Scenario	Budget size					pvalue
Restaurant	B_1	B_2	B_3	B_4	B_5	< 0.05
Kitti	B_1	B_2	B_3	B_4	B_5	< 0.05
Car	B_1	B_2	B_3	B_4	B_5	< 0.05
Kitchen	B_1	B_2	B_3	B_4	B_5	< 0.05
Yard	B_1	B_2	B_3	B_4	B_5	< 0.05
All	B_1	B_2	B_3	B_4	B_5	< 0.05

Table 3: Contrast comparisons between scenarios at every budget and across all budgets. Scenarios with no significant differences are grouped together.

Budget	Scenarios					pvalue
B_1	Restaurant	Kitti	Car	Kitchen	Yard	< 0.05
B_2	Restaurant	Kitti	Car	Kitchen	Yard	< 0.05
B_3	Restaurant	Kitti	Car	Kitchen	Yard	< 0.05
B_4	Restaurant	Kitti	Car	Kitchen	Yard	< 0.05
B_5	Restaurant	Kitti	Car	Kitchen	Yard	< 0.05
All	Restaurant	Kitti	Car	Kitchen	Yard	< 0.05

Contrast comparisons between groups of budgets were conducted using post-hoc t-tests to reveal which pairs were statistically significant. All the p-values were adjusted using Bonferroni correction at a significance level $\alpha = 0.05$. Table 2 summarizes all budget contrast comparisons for every scenario and across all scenarios. Budgets with no significant differences are grouped together. These results indicate that at very small budget sizes, participants follow a similar allocation strategy while as the budget size increases, its effect on distributing resources is more evident.

We also conducted scenario contrast comparisons for every budget, as shown in Table 3 to see if there are statistical significant differences between groups of scenarios. The results indicate that the scenarios Restaurant and Kitti were found significantly different from each one of the other three scenarios presented for the majority of the budgets, meaning that participants generally tend to prefer better audio stimulus quality in these two scenarios.

The human ear is most sensitive to the range 1-4 kHz, which is the range of typical human speech [Sal12]. Motivated by the HAS's tendency to attend to speech above other sounds [Dar08], an analysis was conducted to determine the size of the difference between those scenarios that had some element of human speech {Restaurant, Kitti} with the others. A check was conducted to see if there is significant difference amongst the groups {Restaurant, Kitti} and {Kitchen, Car, Yard} across all the budgets using the contrast:

$$L = \sum_{i \in \{\text{scenario}\}} \alpha_i \mu_i$$

where $\alpha_{\text{Restaurant}} = \alpha_{\text{Kitti}} = \frac{1}{2}$ and $\alpha_{\text{Kitchen}} = \alpha_{\text{Car}} = \alpha_{\text{Yard}} = -\frac{1}{3}$ and μ_i denotes the mean allocation for the scenario i. Conducting the hypothesis testing $H_0 : L = 0$ vs $H_1 : L \neq 0$ it was found that at significance level 0.05, H_0 is rejected suggesting that the two groups are different ($F(4,874) = 5.092, p = 0.015 < 0.05$).

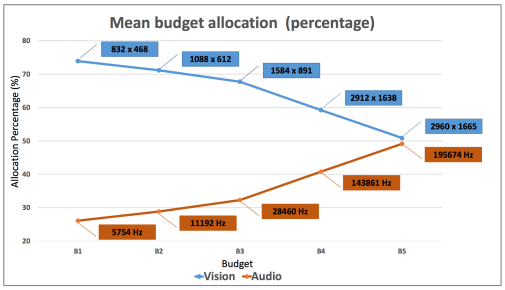

Figure 7: Average percentage allocation for vision and audio for every budget size and across all experimental scenes. The boxes depict the resolution and sampling rate that correspond to these averages.

Figure 7 depicts the resolution and sampling rate that corresponds to the mean allocation for every one of the five budgets. Also the same Figure shows that audio-vision distribution becomes more balanced as the budget size increases.

5. Proposed Model and Validation

In order to be able to make use of the obtained results in actual applications, we designed a model that takes the computational budget as input and provides an audio-visual ratio estimation. A second experiment was conducted with a new set of participants and untested budgets to validate the proposed statistical model.

5.1. Model

The data obtained from the first experiment was used to construct two regression models that could estimate allocation for the two senses. The first model ($\mathbf{M_1}$) takes into account just the budget size while the second ($\mathbf{M_2}$) depends both on the budget and the type of the scenario. This is motivated by the results discussed in section 4, whereby allocation strategy is significantly affected when a scenario includes human voice {Restaurant, Kitti} than when it does not. This distinction led us to construct $\mathbf{M_2}$ to adapt depending on the scenario type. The two models are given in equation (5).

$$\begin{aligned} \mathbf{M_1}: &\quad \hat{Y_1} = \hat{\beta_0} + \hat{\beta_1} \cdot \text{budget} \\ \mathbf{M_2}: &\quad \hat{Y_2} = \hat{\gamma_0} + \hat{\gamma_1} \cdot \text{budget} + \hat{\gamma_2} \cdot \mathbb{1}_{\text{H}} \end{aligned} \tag{5}$$

where $\hat{Y_1}$ and $\hat{Y_2}$ are the graphics allocation estimations and $\hat{\beta_0}, \hat{\beta_1}$ and $\hat{\gamma_0}, \hat{\gamma_1}, \hat{\gamma_2}$ are the least squares regression estimates of $\mathbf{M_1}$ and $\mathbf{M_2}$ respectively. The indicator function $\mathbb{1}_{\text{H}}$ takes

the unity value when the presented audio stimuli includes human voice and zero otherwise.

A power (Box-Cox) transform for the power parameter λ was conducted for every one of the two models to examine if transformation of the response variable (allocation) is needed. In both cases the resulting confidence interval contained the value $\lambda = 1$ indicating that no transformation of the response variable was necessary. Normal Q-Q and residual plots both confirmed that the residual values are normally distributed and homoscedastic for both the proposed models.

Pairs of hypothesis tests for the unknown parameters of the linear regression models (see equation 5) were performed. These tests aim to examine whether the variables budget for $\mathbf{M_1}$ and budget, scenario type are significant for the graphics allocation percentages $\hat{Y}_1$ and $\hat{Y}_2$ respectively. In essence, the intercept and slope parameters of each model are tested to see whether they are significantly different from 0 using the following sets of hypothesis tests:

$$H_0 : \beta_i = 0 \text{ vs } H_1 : \beta_i \neq 0, \quad i \in \{0,1\}$$
$$H_0' : \gamma_j = 0 \text{ vs } H_1' : \gamma_j \neq 0, \quad j \in \{0,1,2\}$$

The results show that all the regression parameters are statistically significant (reject H_0 and H_0') for every i and j suggesting their importance for predicting audio-visual allocation. Specifically, the parameter γ_2 for the scenario type was found to be statistically significant with $t(\text{df} = 873) = 5.53$ and $p = 4.2 \cdot 10^{-6}$.

Table 4 summarises the regression coefficient estimations and the coefficients of determination for both statistical models. Coefficients of determination (R^2 and R^2_{adj}) give an indication of how much variance in the resource allocation (dependent variable) can be predicted from the budget and scenario (independent variables). Model M_2 gives higher R^2 and R^2_{adj} values, a fact that indicates a better fitting for the data obtained from the first experiment.

Table 4: Estimated values for the regression coefficients of the models $\mathbf{M_1}$ and $\mathbf{M_2}$. The coefficients of determination demonstrate how well the experimental data fit the proposed models.

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}_0$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	R^2	R^2_{adj}
$\mathbf{M_1}$	73.68	−17.83	*	*	*	0.61	0.59
$\mathbf{M_2}$	*	*	76.60	−17.83	−7.31	0.72	0.75

5.2. Validation

This section describes the experiment used to validate the statistical models. The two models have subtle differences and it is desirable to see how the presence of the extra term $\hat{\gamma}_2 \cdot \mathbb{1}_{\text{H}}$ affects the prediction compared to real allocation preferences. For this purpose a second experiment was conducted in order to investigate the performance of the two proposed models with three untested budgets and one untested scenario.

5.2.1. Method

The general design of the experiment follows on from the experimental procedure introduced in Section 3. Three scenarios were used for this experiment: Restaurant, Yard and Bathroom, while the Kitti scenario was selected for the training session. The scenario Bathroom was used in the original experiment as a tutorial and no data was collected from this scenario, and therefore, this scene was not used in the generation of the two models.

Three test budgets were used in the experiment, each one different from the previously used budgets. The budgets were chosen to lie in the mid-point between B_1 and B_2, B_2 and B_3 and B_3 and B_4. All materials were prepared in the same way as the first experiment. The procedure was also the same except for the different number of trials. All the participants were tested against all three scenarios and budgets resulting in 9 experimental trials in total. None of the participants in the second experiment had taken part in the first experiment. A total of 10 people, 7 men and 3 women, participated and the average age of all the participants was 28.2 years old. Participants had normal or corrected to normal vision and reported no hearing deficiency.

5.2.2. Results

Figure 8 shows how the two regression models perform against actual human preferences collected from the validation experiment. The mean allocation preferences for the test budgets are compared with the models' output for the graphics quality. In all three cases, $\mathbf{M_2}$ gives better estimations than the model $\mathbf{M_1}$. Across all the validation scenarios used, the average error of absolute difference for $\mathbf{M_1}$ is 3.73% while for the model $\mathbf{M_2}$ is 0.69%. Specifically, for the newly introduced scenario Bathroom, results are largely comparable with 3.12% for $\mathbf{M_1}$ and 0.59% for $\mathbf{M_2}$. The error for model $\mathbf{M_1}$ is relatively small on the whole, but compared to $\mathbf{M_2}$ it underestimates the graphics allocation for the Yard while it provides an overestimate in the case of the Restaurant scenario. Crucially, even though none of the models was designed using data from the Bathroom scene, the results indicate that the models give accurate estimations for this new scenario, broadly validating the models.

6. Discussion

The results of the first experiment yielded a number of potentially interesting findings related to how participants chose to allocate resources. The first relevant finding is that, overall, as expected, under our experimental conditions there was a preference for improving the quality of graphics over the quality of audio. This confirms the hypothesis and is supported by previous research [Col74,PNK76]. Furthermore, a second relevant finding is that the allocation is dependent on

Figure 8: Comparison between model estimations with the data collected from the validation study. The means and standard deviations of the three test budgets are depicted in black colour while the colour curves are the model predictions. Log-spacing was applied to the x-axis for better visualisation.

the amount of resources available; when the amount of available resources is increased there is an increase in the amount of resources dedicated to audio. This increase is clear and is statistically significant. This indicates that participants become more concerned with the quality of the acoustic stimulation when paired with higher quality visual stimuli.

The scenarios also demonstrate a significant difference between them. It is presumed that allocation may be scenario dependent and there are indications that this may be the case. As expected [Sal12], the grouping of the scenarios which include human voices provided significant differences compared to those that do not, with participants allocating more resources to audio in the former.

The validation experiment provides evidence that the presented models can be used in general rendering frameworks. Model $\mathbf{M_1}$ provides relatively good results and $\mathbf{M_2}$ even stronger as long as the scenarios can be correctly categorised. The two models predicted equally well in untested experimental conditions (budgets, scenario), indicating a broader application to other inputs. Rendering frameworks can consult the models for a given computational load or achievable frame rate and identify how much of that time is to be dedicated to graphics and acoustics.

6.1. Limitations

Further work with more scenarios is required to provide an understanding of whether the scenarios can be classified in the manner described above, or what other characteristics can be used to further predict correct utilisation of resources. A possible limitation of the work may be due to the static nature of the graphics. The use of resolution to evaluate quality of the visuals can be considered simplistic, but the complexity of the problem required to keep other parameters that can affect the image quality fixed, hence resolution was chosen as the primary of the possible quality altering variables. Future work will consider changing combinations of quality variables, as can be found in various Virtual Reality studies, although this will increase the number of combinations to be tested in an experimental study significantly.

It is also important to verify whether these observations hold when videos or interactive environments are used for the visuals. The normalisation of the two stimuli may also be considered a limitation in terms of generalisation of the results; however, the alternative of working with real, measured, computational budgets requires fixing the acoustic and graphical rendering to particular, distinct algorithms (and these can vary significantly in terms of computation, possibly by several orders of magnitude; for example, path traced and rasterised images). This would be difficult to generalise further and limit the scope of the current work.

The goal of this work was to identify the right amount of resources to dedicate to each sense from an available computational budget. In this respect our proposed model can be used as an initial guideline for audio-visual resource allocation.

7. Conclusion and Future Work

This work has introduced a new method for analysing bimodal interactions with a direct relevance to how the resources can be allocated. The results suggest that, as expected, graphics are typically considered better than audio although the difference declines significantly as more resources become available. The model proposed in this paper can be used as a guideline for budget allocation for VE simulation on single and also parallel rendering systems. In parallel systems the resource allocation could be used to drive load balancing decisions. This work is a first attempt at understanding the potential of models to provide efficient resource allocation to VE simulation systems. Future work will look to further validate these results with a number of additional applications, including dynamic scenes, task-based scenarios and including the simulation of more senses.

References

[AB04] ALAIS D., BURR D.: The Ventriloquist Effect Results from Near-Optimal Bimodal Integration. *Current Biology 14*, 3 (February 2004), 257–262. 3

[ADTA01] ALGAZI V. R., DUDA R. O., THOMPSON D. M., AVENDANO C.: The Cipic hrtf database. *Proceedings 2001 IEEE workshop on Applications of Signal Processing to Audio and Electroacoustics* (2001), 99–102. 5

[BA06] BURR D., ALAIS D.: Combining visual and auditory information. *Prog Brain Res 155* (2006), 243–258. 3

[BdG04] BERTELSON P., DE GELDER B.: *Crossmodal Space and Crossmodal Attention*. Oxford University Press, USA, May 2004, ch. The Psychology of Multimodal Perception. 2

[BM13] BELMUDEZ B., MÖLLER S.: Audiovisual quality integration for interactive communications. *EURASIP Journal on Audio, Speech, and Music Processing 2013*, 1 (2013), 1–23. 3

[Bro58] BROADBENT D. E.: *Perception and communication*. Oxford : Oxford University Press, 1958. 2

[BS06] BLAKE R., SEKULER R.: *Perception*, 5th ed. McGraw-Hill Higher Education, 2006. 2

[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), 1:1–1:16. 3

[Cat] CATT-Acoustic, Gothenburg, Sweden. www.netg.se/catt. 2

[Col74] COLAVITA F. B.: Human sensory dominance. *Perception and Phychophysics 16*, 2 (1974), 409–412. 4, 10

[Dar08] DARWIN C.: Listening to speech in the presence of other sounds. *Philosophical Transactions of the Royal Society B: Biological Sciences 363*, 1493 (2008), 1011–1021. 9

[DM95] DURLACH N. I., MAVOR A. S.: *Virtual reality: Scientific and Technological Challenges*. Committee on Virtual Reality Research and Development, National Research Council report, 1995. 2

[DS98] DRIVER J., SPENCE C.: Crossmodal attention. *Curr Opin Neurobiol 8*, 2 (Apr 1998), 245–253. 2

[Duf99] DUFOUR A.: Importance of attentional mechanisms in audiovisual links. *Exp Brain Res 126*, 2 (May 1999), 215–222. 2

[Est76] ESTES K. W.: *Handbook of Learning and Cognitive Processes: Vol. 4, Attention and Memory*. Psychology Press, 1976. page 305. 6

[FN05] FUJISAKI W., NISHIDA S.: Temporal frequency characteristics of synchrony-asynchrony discrimination of audio-visual signals. *Experimental Brain research 166*, 3–4 (2005), 455–464. 2

[GBW*09] GRELAUD D., BONNEEL N., WIMMER M., ASSELOT M., DRETTAKIS G.: Efficient and practical audio-visual rendering for games using crossmodal perception. In *Proceedings of the 2009 Symposium on Interactive 3D Graphics and Games* (New York, NY, USA, 2009), I3D '09, ACM, pp. 177–182. 3

[GGB05] GUTTMAN S. E., GILROY L. A., BLAKE R.: Hearing what the eyes see: auditory encoding of visual temporal sequences. *Psychological Science 16*, 3 (2005), 228–235. 1

[HAC08] HULUSIĆ V., ARANHA M., CHALMERS A.: The influence of cross-modal interaction on perceived rendering quality thresholds. In *Theory and Practice of Computer Graphics* (2008), procceedings W. . F. P., (Ed.), UNION Agency-Science Press, Plzen, Chech Republic, pp. 41–48. 3

[HCD*09] HULUSIĆ V., CZANNER G., DEBATTISTA K., SIKUDOVA E., DUBLA P., CHALMERS A.: Investigation of the beat rate effect on frame rate for animated content. In *Proceedings of the 25th Spring Conference on Computer Graphics* (New York, NY, USA, 2009), SCCG '09, ACM, pp. 151–159. 3

[HHD*12] HULUSIC V., HARVEY C., DEBATTISTA K., TSINGOS N., WALKER S., HOWARD D., CHALMERS A.: Acoustic rendering and auditory–visual cross-modal perception and interaction. *Comput. Graph. Forum 31*, 1 (Feb. 2012), 102–131. 1

[HT66] HOWARD I. P., TEMPLETON W. B.: *Human spatial orientation [by] I.P. Howard and W.B. Templeton*. Wiley, London, New York,, 1966. 2

[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 *Theory and Practice of Computer Graphics* (2010), Collomosse J., Grimstead I., (Eds.), The Eurographics Association. 3

[ITU12] *General viewing conditions for subjective assessment of quality of SDTV and HDTV television pictures on flat panel displays*. ITU-R, August 2012. 7

[Kaj86] KAJIYA J. T.: The rendering equation. *ACM SIGRAPH Computer Graphics 20*, 4 (1986), 143–150. 4

[KBM08] KOLLMEIER B., BRAND T., MEYER B.: *Springer Handbook of Speech Processing*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, ch. Perception of Speech and Sound, pp. 61–82. 6

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

[MDCT05] MASTOROPOULOU G., DEBATTISTA K., CHALMERS A., TROSCIANKO T.: Auditory bias of visual attention for perceptually-guided selective rendering of animations. GRAPHITE '05, ACM, pp. 363–369. 2

[Nay93] NAYLOR G. M.: ODEON–Another hybrid room acoustical model. *Applied Acoustics 38*, 1 (1993), 131–143. 2

[NMPDSLC14] NARWARIA M., MANTIUK R., PERREIRA DA SILVA M., LE CALLET P.: HDR-VDP-2.2: a calibrated method for objective quality prediction of high-dynamic range and standard images: a calibrated method for objective quality prediction of high-dynamic range and standard images. *Journal of Electronic Imaging 24*, 1 (Dec. 2014), 010501. 4

[PIW11] PINSON M. H., INGRAM W., WEBSTER A.: Audiovisual quality components. *IEEE Signal Processing Magazime 28*, 6 (2011), 60–67. 3

[PNK76] POSNER M. I., NISSEN M. J., KLEIN R. M.: Visual dominance: an information-processing account of its origins and significance. *Psychological review 83*, 2 (1976), 157. 2, 10

[Pyl06] PYLYSHYN Z. W.: *Seeing and Visualizing: It's not what you Think*. MIT Press, March 2006. 2

[Sal12] SALVEND G.: *Handbook of Human Factors and Ergonomics*, 4 ed. Wiley, 2012. 9, 11

[Sch01] SCHOLL B. J.: Objects and attention: the state of the art. *Cognition 80*, 1-2 (June 2001), 1–46. 2

[SM93] STEIN B. E., MEREDITH M. A.: *The merging of the senses.* The MIT Press, 1993. 1

[STKS07] SILTANEN S., TAPIO L., KIMINKI S., SAVIOJA L.: The room acoustic rendering equation. *Acoustical Society of America 122*, 3 (2007), 1624–1632. 5

[Sto98] STORMS R. L.: *Auditory-Visual Cross-Modal Perception Phenomena*. {PhD} thesis, Naval Postgraduate School, 1998. 2

[SZ00] STORMS R. L., ZYDA M. J.: Interactions in perceived quality of auditory-visual displays. *Presence: Teleoper. Virtual Environ. 9*, 6 (Dec. 2000), 557–580. 2

[Tor09] TORRALBA A.: How many pixels make an image? *Visual Neuroscience 26*, 1 (2009), 123–131. 4

[VS13] VARSHNEY L. R., SUN J. Z.: Why do we perceive logarithmically? *Significance 10*, 1 (2013), 28–31. 5