On Saliency and Interactivity in Human Audiovisual Perception

Ulrich Reiter

Q2S - Centre for Quantifiable Quality of Service in Communication Systems, NTNU Trondheim, Norway, reiter@q2s.ntnu.no

Introduction

Most of today's interactive application systems aim at simulating an accurate representation of the real world by focusing on the (arguably) most important human sense, vision. Auditory stimuli are used in these systems to enhance the overall impression of realism. Still, the stimuli of the two modalities are rendered and presented mostly independently from the other modality. The level of detail in the respective (visual or auditory) simulation is kept as high as possible with regard to computing power available, independently from the level of detail in the other modality and independently from the user's current focus.

This approach apparently contradicts our real world experiences. In the real world, we experience a simultaneous stimulation of all our senses, providing us with a wealth of information about objects surrounding us. We deliberately or not choose the object or event which is of most interest to us, and perceive its characteristics multi-modally. Yet, not all stimuli that are perceived by our senses are equally important in the generation of an overall impression. Depending on a number of factors our perceptual processor subconsciously selects those stimuli that are important - and downgrades or completely discards others of less importance.

Saliency

Unfortunately, these factors and the related selection mechanisms are not fully understood. Yet, if it was possible to provide a salience model describing what the importance of all singular percepts in a multi-modal perceptual situation is, then we could design audiovisual application systems in such a way that they offered an optimum quality / cost ratio.

A salience model would thus mainly identify (and ideally quantify) the influence factors that control the level of saliency of the perceived objects contained in audiovisual presentations. For this it is necessary to get away from a generalized salience model. A generalized salience model would be too complex and the influence factors too manifold to cope with at this state of knowledge. Instead, it is reasonable to focus on a salience model valid for a singular, well-defined application scenario only. Here, a salience model for interactive audiovisual applications of moderate complexity, e.g. like those that are enabled by the MPEG-4 standard (ISO/IEC 14496, [1]), is introduced.

Experimental Results

A number of subjective assessments have been performed to verify these factors that were previously identified for a typical prototype reproduction setup of interactive audiovisual content. This exemplary setup makes use of a large projecting screen, a multichannel loudspeaker setup and real-time room-acoustic simulation rendered on a standard PC. The results of these experiments reveal a number of interesting points:

1. The first two assessments [2, 3] focused on a possible reduction of algorithmic complexity for the bimodal case compared to the unimodal case. The simplifications assessed were directly related to the computational load that the real-time rendering of audio imposes on the processor. It has been shown that the number of loudspeakers necessary in interactive audiovisual application systems of moderate complexity using a VBAP panning approach depends on the content itself. As a rule of thumb, the well-known five-channel setup defined in ITU-R BS.775 [4] is suitable for such systems. The Perceptual Approach algorithm as specified in MPEG-4 Scene Description [5] can be simplified to use only four internal workchannels without degrading the overall perceived quality in the audiovisual context.

2. The next three assessments [6, 7, 8] focused on the effect that user interaction with the audiovisual application or scene might have on the perceived overall quality. Here the general assumption was that by offering an attractive interactive content or by assigning the user a challenging task, the user would become more involved and thus experience a subjectively higher overall quality. As the results of the experiments show, this is not generally the case. However, when both task and main varying (or salient) quality attribute were located in the same modality, such an effect could be substantiated. Apparently, inner-modal influence is significantly greater than cross-modal influence. This is also suggested by the common theories of capacity limits in human attention.

3. The last assessment [9] showed that a cross-modal influence of interaction is possible when stimuli and interaction are carefully balanced. Yet, at this time it is not possible to determine or quantify that balance *a priori*.

Salience Model

However, some of the influence factors that contribute to this balance have been identified in a salience model

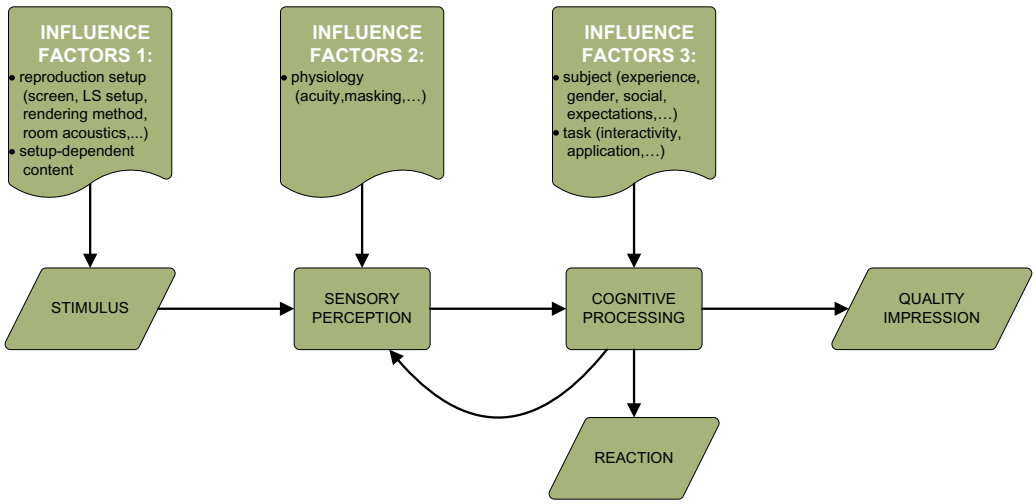

Figure 1: A salience model for interactive audiovisual applications of moderate complexity.

for interactive audiovisual applications of moderate complexity, see fig. 1. They can be grouped into three categories: influence factors (IFs) of level 1 (related to the generation of stimuli), IFs of level 2 (related to the physical perception of stimuli / the physiology of the user), and IFs of level 3 (related to the processing and interpretation of the perceived stimuli).

The salience model describes that sensory perception and cognitive processing are influencing each other: a percept that calls the attention of the cognitive processor will be shifted more toward the focus of the next *Perceptual Cycle*, a concept that has been introduced by Neisser [10]. Thus, sensory perception is to a certain degree under control of cognitive processing.

The cognitive processing results both in a (physical) reaction and a quality impression. It can be assumed that reaction and quality impression are correlated to each other because they are created based on the same processing. Attempts should be made to describe this relationship in more detail, e.g. by analyzing the user's physical reaction.

Although interaction does not generally increase the degree of perceived quality, it is a good indicator of the user's current focus of attention: as interaction requires attention, the user's current focus can be deduced by analyzing his input to the system. This could help in further evolving the suggested salience model.

References

[1] Int. Std. (IS) ISO/IEC 14496-1:2004, Information technology - Coding of audio-visual objects - Part 1: Systems, 3rd Ed., Geneva, Switzerland, 2004.

[2] Reiter, U.: Subjective Assessment of the Optimum Number of Loudspeaker Channels in Audio-Visual Applications Using Large Screens, Proc. AES 28th International Conference, Pitea, Sweden, June 30 - July 2, 2006, pp 102-109.

[3] Reiter, U.; Partzsch, A.; Weitzel, M.: Modifications of the MPEG-4 AABIFS Perceptual Approach: Assessed for the Use with Interactive Audio-Visual Application Systems, Proc. AES 28th International Conference, Pitea, Sweden, June 30 - July 2, 2006, pp 110-117.

[4] Recommendation ITU-R BS.775-1, Multichannel stereophonic sound system with and without accompanying picture, International Telecommunication Union, Geneva, Switzerland, 1994.

[5] Int. Std. (IS) ISO/IEC 14496-11:2004, Information technology - Coding of audio-visual objects - Part 11: Scene description and Application engine, Geneva, Switzerland, 2004.

[6] Reiter, U.; Jumisko-Pyykkö, S.: Watch, Press and Catch - Impact of Divided Attention on Requirements of Audiovisual Quality, 12th International Conference on Human-Computer Interaction, HCI2007, Beijing, PR China, July 22-27, 2007.

[7] Reiter, U.; Weitzel, M.; Cao, S.: Influence of Interaction on Perceived Quality in Audio Visual Applications: Subjective Assessment with n-Back Working Memory Task, Proc. AES 30th International Conference, Saariselkä, Finland, March 15-17, 2007.

[8] Reiter, U.; Weitzel, M.: Influence of Interaction on Perceived Quality in Audio Visual Applications: Subjective Assessment with n-Back Working Memory Task, II, AES 122nd Convention, Vienna, Austria, May 5-8, 2007, Preprint 7046.

[9] Reiter, U.; Weitzel, M.: Influence of Interaction on Perceived Quality in Audiovisual Applications: Evaluation of Cross-Modal Influence, Proc. 13th International Conference on Auditory Displays (ICAD), Montreal, Canada, June 26-29, 2007.

[10] Farris, J. Shawn: "The Human Interaction Cycle: A Proposed and Tested Framework of Perception, Cognition, and Action on the Web", PhD Thesis, Kansas State University, USA, 2003.