Fusion Function, Content Classification, and Perceived Quality of Audiovisual Media Content

Ulrich Reiter

Norwegian University of Science and Technology, Trondheim, Norway, email: reiter@q2s.ntnu.no

Introduction

In today's application scenarios like IPTV, teleconferencing, or transmission of media content over packet-based networks in general, perceived quality estimation is done using mono-modal objective metrics, as no truly cross-modal metrics are available. After estimating auditory and visual qualities separately, an overall audiovisual quality is determined by applying a so-called fusion function, usually looking something like $AVQ = a_0 + a_1 \cdot AQ + a_2 \cdot VQ + a_3 \cdot AQ \cdot VQ$. Going through the literature it becomes apparent that there is little agreement on the magnitude of weighting factors a_1, a_2, and a_3 to be used in the fusion function.

It is suspected that this is because different content used in different experiments draws users' attention towards different quality attributes or features. These can be located in either the auditory or the visual domain, or can be truly cross-modal. In this paper we look for evidence that content itself is actually the driver of the fusion function's parametrization.

Perceptual Features and Content Classification

Content classes can be derived based on many different types of features. Both technical descriptors (e.g. spatial or temporal information index, motion vector) as well as perceptual features can be used. Here, a set of features originally suggested by Woszczyk et al. [3] is used. These features were originally suggested to measure the overall audiovisual quality of home cinema setups. Woszczyk et al. derive a set of 'perceptual dimensions of audiovisual experience', with each dimension itself 'characterized by attributes which equally address visual and auditory modalities', resulting in a $4x4$ matrix of *Dimensions* and *Attributes*, see Figure 1.

Fusion Function Experiment

The different weighting factors of the fusion function have been compiled in [1], see Table 1. In order to check whether differences in weighting factors of the fusion function actually stem from the different content used in the experiments summarized in Table 1, we chose 7 different audiovisual media contents for an experiment. Each clip contained a meaningful segment of audio and video information, and start and end points were chosen such that semantic structures were maintained. The total length of each clip was around 30 seconds. All clips had a CIF video resolution (352x288 pixels) downsized from original high quality, high resolution video sequences.

Dimension	Attribute
Space dimension: the illusion of being in a projected space	**Quality** (distinctness, clarity, and detail of the impression) – 'How distinct is the sensation?'
Motion dimension: the illusion of physical flow and movement	**Magnitude** (the strength of the impression) – 'How powerful is the sensation?'
Mood dimension: the articulation and density of atmosphere	**Involvement** (the emotional effect on the viewer) – 'How involving is the sensation?'
Action dimension: the sensation of dynamic intensity and power	**Balance** (relative contribution of auditory and visual sensations) – 'How balanced are modalities: stronger sound or stronger picture?'

Figure 1: $4x4$ matrix of *Dimensions* and *Attributes* as suggested by Woszczyk et al. [3].

Table 1: Overview of weighting factors for audio-visual quality assessments, excerpt from [1].

Lab	a_0	a_1	a_2	a_3	Corr.
KPN	1.12	0.007	0.24	0.088	0.98
	1.45	0	0	0.11	0.97
Bellcore	1.07	0	0	0.111	0.99
	1.295	0	0	0.107	0.99
ITS	-0.677	0.217	0.888	0	0.978
	1.514	0	0	0.121	0.927
	0.517	-0.0058	0.654	0.042	0.98
NTT	1.17	-0.144	0.186	0.154	0.96
	0.908	-0.192	0.258	0.193	0.96
ICRFE	-0.9222	0.5691	0.5064	0.1697	0.911
	-0.6313	0.2144	0.0124	0.1184	0.902
BT	1.15	0	0	0.17	0.85
	0.95	0	0.25	0.15	0.83
	4.26	0.59	0.49	0	0.97
	-3.34	0.85	0.76	-0.01	0.99

Video display was provided on a standard non-glossy 19" consumer LCD computer monitor, resulting in a video size of approximately 11cm across on the monitor (mid grey desktop background color). Subjects were sitting at a viewing distance of 40cm. In this experiment, audio streams were 16bit 44.1kHz stereo wav files extracted from the original sequences. Audio was played back using a professional grade sound card with external converters, and circumaural, open headphones. The laboratory used for the experiment was AURA lab at NTNU, an acoustically treated listening room with a controlled lighting situation created for audiovisual quality tests. A total of 18 subjects (13 male, 5 female) between 24 and 40 years of age participated in the experiment. More details can be found in [2].

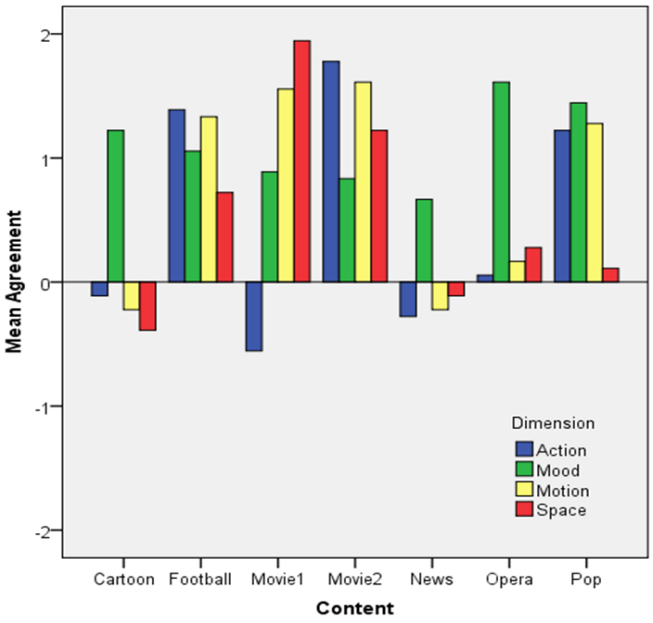

Figure 2: Mean agreement with assumed high importance of dimensions Action, Mood, Motion, and Space for each clip. Agreement scale: 2=Strongly agree, 1=Agree, 0=Neither agree nor disagree, -1=Disagree, -2=Strongly disagree

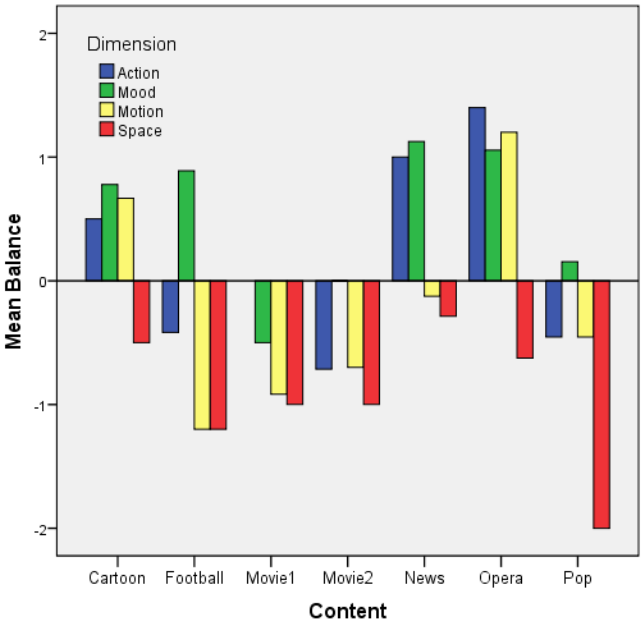

Figure 3: Mean *Balance* attribute for each dimension and clip. Agreement scale: 2=Much stronger sound, 1=Stronger sound, 0=Both are equally strong, -1=Stronger video, -2=Much stronger video

Test subjects were asked how much they thought the different dimensions were important for their overall experience of each clip. Figure 2 gives an overview on subjects' mean agreement with the statement "The [*insert definition of dimension here*] is very important for my overall perceptual experience of this clip." Interestingly, the Mood dimension received rather constant agreement across all clips, indicating that the importance of the Mood dimension seems to be independent of content type. For all other dimensions, there is at least one clip for which no agreement to the above statement could be identified.

Figure 3 shows subjects' mean answer to the question whether they experienced each of the *Dimensions* stronger in the visual or in the auditory domain. For the Balance attribute in the Mood dimension, the question was then: "How balanced are the modalities: stronger sound or picture for the articulation and density of atmosphere?". Again, significant differences between content could be substantiated.

Conclusions

The experiment outlined here has clearly shown that the parametrization of the fusion function is content dependent. Significant differences in the importance of perceptual dimensions have been found among the seven different test clips used. A classification of content can be multi-faceted. The features used here are in no way all-embracing, more features are sure to exist. This will be explored in our future research. For the parametrization of the fusion function, especially *Balance* is interesting, as it describes the relative contribution of auditory and visual sensations to the overall perceived quality.

However, the evaluation of perceptual features requires subjective assessments. This is impractical for most application purposes and bars the approach from being used in quality monitoring scenarios. The question is therefore whether perceptual features could be measured or estimated in an automated way? Either based on physiological measures like heart-rate, skin conductivity, or EEG, or by successfully correlating perceptual features with content-immanent technical descriptors like spatial or temporal information index and motion vector. We are planning to look into these aspects in our future work.

References

[1] You, J., Reiter, U., Hannuksela, M., Gabbouj, M., Perkis, A.: Perceptual-based Quality Assessment for Audio-visual Services - A Survey. Signal Processing: Image Communication. Vol. 25, No. 7, Aug. 2010.

[2] Reiter, U.: Towards a Classification of Audiovisual Media Content. In Proceedings of the 129th Convention of the Audio Engineering Society. AES, San Francisco, USA, Nov. 2010.

[3] Woszczyk, W., Bech, S., Hansen, V.: Interactions Between Audio-Visual Factors in a Home Theater System: Definition of Subjective Attributes. AES Audio Engineering Society 99th Convention, Preprint 4133, New York, USA, 1995.