

**Distribution Of Multi-view Entertainment using content aware
DElivery Systems**

DIOMEDES

Grant Agreement Number: 247996

D3.4

**Report on the Quality of Experience model and the audio and
visual attention models**

Document description	
Name of document	Report on the Quality of Experience model and the audio and visual attention models
Abstract	This report will describe the Quality of Experience model, and the audio and visual attention models. It will compare the results obtained with the models to those obtained in the viewer tests, reported in this deliverable. A strong correlation with the viewer tests will be shown. As well as a description of the algorithms, a brief analysis of the computational complexity of the algorithms will be provided.
Document identifier	D3.4
Document class	Deliverable
Version	1.0
Author(s)	N. Just, P. tho Pesch (IRT) / K. Kunze (IDMT), J. Liebetrau, T. Korn (IDMT) / Xiyu Shi, Hemantha Kodikara Arachchi, Chunggeun Kim (UNIS)
QAT team	P. Kovacs (HOL) / M. Tekalp, S. Savas (KOC)
Date of creation	2011-08-01
Date of last modification	2011-10-31
Status	Final
Destination	European Commission
WP number	WP3

TABLE OF CONTENTS

1	INTRODUCTION	7
1.1	Purpose of the document	7
1.2	Scope of the work.....	7
1.3	Achievements.....	7
1.4	Structure of the document.....	7
2	QUALITY OF EXPERIENCE MODEL	8
2.1	Video QoE Modelling	8
2.1.1	KPI metadata models	8
2.1.2	QoE model for video + depth 3D content	9
2.1.3	QoE model for stereoscopic contents	13
2.2	Perceptual Evaluation Methodologies	16
2.2.1	Audio-only experiments	16
2.2.2	Audio-visual experiments	23
3	ATTENTION MODELLING	31
3.1	Visual Attention Modelling	31
3.1.1	Model Concept	31
3.1.2	Functionality	32
3.1.3	Evaluation.....	37
3.2	Audio Attention Modelling	40
3.2.1	Introduction and subjective test procedure	40
3.2.2	Results and analysis.....	40
3.2.3	Principles for application.....	44
3.2.4	Validation experiment.....	44
3.2.5	Summary and conclusion	46
3.2.6	Loudness Analysis Model.....	47
4	CONCLUSIONS	56
	REFERENCES	57
	APPENDIX A: GLOSSARY OF ABBREVIATIONS	59

LIST OF FIGURES

Figure 1 - Architecture for deploying the DIOMEDES QoE model	8
Figure 2 - MOS vs. QP of the leftover region for different attention area QP settings for (a) ballet, (b) champagne tower, and (c) music test sequences	11
Figure 3 - Proposed 3D QoE metric and VQM vs. QP of the leftover region for different attention area QP settings for the ballet test sequence	12
Figure 4 - Proposed 3D QoE metric and VQM vs. QP of the leftover region for different attention area QP settings for the champagne tower test sequence	12
Figure 5 - Proposed 3D QoE metric and VQM vs. QP of the leftover region for different attention area QP settings for the music test sequence	13
Figure 6 - MOS vs. QP of the leftover region for different attention area QP settings for (a) fencing, (b) music, (c) lecture sequences, (d) band, (e) cafe, and (f) Poznan street.....	14
Figure 7 - Average VQM vs. QP of the leftover region for different attention area QP settings for (a) fencing, (b) music, (c) lecture sequences, (d) band, (e) cafe, and (f) Poznan street	15
Figure 8 - Stimulus presentation in the ACR method [P.910].....	16
Figure 9 - Mean values for items and conditions.....	19
Figure 10 - Relation of quality terms	21
Figure 11 - Boxplots for ratings of all stimuli of the respective bit rate	22
Figure 12 - Schematic representation of test set-up, as system with reduced loudspeakers only the ones with one dot and a circle are used.....	25
Figure 13 - Mean Overall Quality Rating for the two contents, Error bars show the 95% confidence intervals.....	27
Figure 14 - Mean Overall Quality Rating, averaged over content and reproduction system, Error bars represent 95% confidence intervals	28
Figure 15 - Mean Impairment Ratings of the Angular displacement of audio and video objects	29
Figure 16 - Mean Impairment Ratings for parameters bitrate and angle	30
Figure 17 - Visual Attention Framework	31
Figure 18 - Fencing Scene Attention Map	34
Figure 19 - Attention Model Script.....	36
Figure 20 - Mean Opinion Scores for combinations of quality degradations in stimuli - music and speech at normal loudness and wide at the centre. The quality degradation in speech seems to be perceived more salient than music (as indicated by the circle on the plot).....	41
Figure 21 - Mean Opinion Scores for combinations of quality degradations in stimuli - speech louder than music, both wide at the centre.....	41
Figure 22 - Mean Opinion Scores for combinations of quality degradations in stimuli - music louder than speech, both wide at the centre.	42
Figure 23 - Mean Opinion Scores for combinations of quality degradations in stimuli - both at the same	

amplitude, speech from the right channel with music wide at the centre.....	42
Figure 24 - Mean Opinion Scores for combinations of quality degradations in stimuli - speech louder than music, speech from the right channel with music wide at the centre.....	43
Figure 25 - Mean Opinion Scores for combinations of quality degradations in stimuli - music louder than speech, speech from the right channel with music wide at the centre.....	43
Figure 26 - Means and standard deviations of subjective evaluation scores of the provided auditory scene where various objects were individually degraded in terms of bitrate.....	45
Figure 27 - Suggested processes for audio attention modelling	46
Figure 28 - Integration of the Loudness Analysis Model	47
Figure 29 - Functional model of loudness measurement according to Zwicker.....	48
Figure 30 - Loudness measurement according to EBU R128	48
Figure 31 - Schematic representation of the Loudness Analysis Model.....	50
Figure 32 - User interface of the loudness analysis Matlab programme	52
Figure 33 - Measurement setup for determining the loudness of a time variant object based audio scene in the anechoic chamber	53
Figure 34 - Deviation with all calculation processes running.....	54
Figure 35 - Deviation with directivity of sources turned off	54
Figure 36 - Deviation with sound dissipation turned off.....	55
Figure 37 - Deviation with two source objects.....	55

LIST OF TABLES

Table 1 - QP combinations used for encoding attention area and leftover region for subjective experiments	10
Table 2 - Validation of the proposed 3D visual QoE model under attention area based coding.....	11
Table 3 - Validation of the proposed stereoscopic QoE model under attention area based coding	15
Table 4 - Attributes of the Selected Test Items	17
Table 5 - Screenshots and descriptions of the used test items	24
Table 6 - Visual Attention Filter Category.....	32
Table 7 - Visual Attention Output Format	33
Table 8 - Visual Attention evaluation results for 2D.....	38
Table 9 - Visual Attention evaluation results for 3D.....	38
Table 10 - Acoustic attributes varied in combination for the subjective listening test	40
Table 11 - Sound objects and attributes controlled for the validation experiment	45
Table 12 - Description parameters used in the DIOMEDES audio metadata	49
Table 13 - Physical based audio processes implemented into the scene model	51

1 INTRODUCTION

1.1 Purpose of the document

This deliverable provides documentation of the Quality of Experience model and the audio and visual attention models developed within WP3. This includes the developed concepts and algorithms as well as the subjective and objective test that were carried out. This document provides updates regarding the work described in chapter 2 of deliverables D3.2 and D3.3.

1.2 Scope of the work

The summary of the WP3 work given in this document will be used as an input for the content encoding process within the developed system.

1.3 Achievements

This deliverable presents the developed QoE model as well the audio and visual attention models. This also includes a detailed description of the underlying concepts, the implementation and the evaluation of the models.

1.4 Structure of the document

Chapter 2 shows the developed Quality of Experience model in detail. This also includes a description of the KPI metadata models, the QoE model for video + depth 3D as well as for stereoscopic content. Furthermore the perceptual evaluation methodologies that are used are presented.

Chapter 3 gives a description of the audio and visual attention models. Their functionality is explained in detail and evaluation results based on user tests show the quality of the models.

2 QUALITY OF EXPERIENCE MODEL

In the following the developed Quality of Experience (QoE) model will be shown.

2.1 Video QoE Modelling

DIOMEDES content is organised into independently encoded video and depth map streams. These video streams are used by the video renderer to generate appropriate views for the display device. The aim of the video Quality of Experience (QoE) model is to provide the framework for generating Key Performance Indicator (KPI) metadata at the sender side and computing the QoE at the receiver side as described in the deliverable D4.1. This scenario is illustrated in Figure 1. In order to align the model into this architecture, the quality metric proposed for this project focuses on determining KPI metadata and the way of combining the metadata to produce the QoE score. Hence, the video QoE development activity is performed in two stages. In the first stage, the KPI models were identified. Subsequently, the QoE models, which define the way of computing the final QoE combining the KPI values for the given application scenario (e.g., stereoscopic application, video+depth application, etc.), are defined in the second stage. The KPI models are used by the sender to extract KPI metadata while the QoE models are used by the receiver to compute the QoE value of the received content. The following subsections introduce the KPI and QoE models.

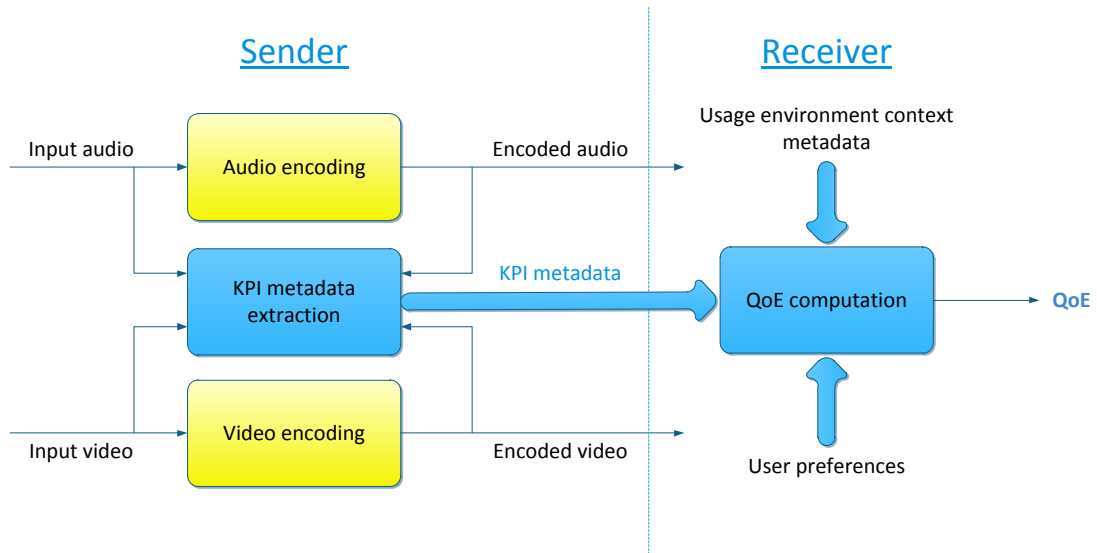


Figure 1 - Architecture for deploying the DIOMEDES QoE model

2.1.1 KPI metadata models

The video QoE models two KPIs, namely the Image Quality (IQ) and Depth Perception (DP). IQ provides a measure of perceptual texture distortions while DP measures the perceptual depth distortions. As described in the previous deliverable D3.3, IQ is measured by using Video Quality Model (VQM) [1]. Moreover, the DP is measured using the VQM and Disparity Distortion Metric (DDM) [4]. Therefore,

$$IQ = VQM \quad (1)$$

$$DP = (1 - VQM)^\alpha \cdot MDDM^\beta \quad (2)$$

where $\alpha = 1.5$ and $\beta = 1$, and $MDDM$ is defined as:

$$MDDM(X, Y) = \frac{1}{M} \sum_{j=1}^M DDM(x_j, y_j) \quad (3)$$

The effectiveness of the above mentioned depth quality measurement technique is evaluated in [2].

2.1.2 QoE model for video + depth 3D content

In order to obtain the overall video QoE metric, the IQ and the DP have to be combined. This is achieved by defining a weighting factor, w , as shown below [3]:

$$QoE_v = (1 - w) \cdot IQ + w \cdot DP \quad (4)$$

where QoE_v is the video QoE.

The weighting factor, w , is content dependent. Activity in the depth map increases the importance of the depth quality for the overall quality. Therefore, Z-direction (depth direction) Motion Activity (ZMA) of the 3D video is used to model the weighting factor [3]. Assume that the standard deviation measured over the temporal dimension for a given pixel location (i, j) , $\sigma_{Y_{i,j}}^t$, is:

$$\sigma_{Y_{i,j}}^t = \left[\frac{1}{N} \sum_{k=1}^N (Y_{i,j}^k - \mu_{Y_{i,j}}^t)^2 \right]^{\frac{1}{2}} \quad (5)$$

where N is the number of depth map frames considered and $Y_{i,j}^k$ and $\mu_{Y_{i,j}}^t$ are the pixel value and temporal mean of the pixels. ZMA is the average temporal standard deviation over the video sequence as defined below:

$$ZMA = \frac{1}{M \cdot H \cdot W} \sum_{i=1}^M \sum_{j=1}^H \sum_{i=1}^W \sigma_{Y_{i,j}}^t \quad (6)$$

where M is the number of temporal segments of length N of depth map in the video sequence. H and W denote the frame height and width respectively. Since the above defined ZMA is dependent on the bit-depth of the depth map, ZMA is normalised to eliminate this dependency.

$$nZMA = \frac{ZMA}{2^n - 1} \quad (7)$$

where n is the bit-depth of the depth map. To model the weighting factor w , the relative importance of the subjective ratings (Mean Opinion Score - MOS) for image and depth quality with respect to the subjective ratings (MOS) for overall 3D quality are analysed for the Interview, Orbi and Breakdancers test sequences. A model for the weighting factor is derived as functions of $nZMA$, such that they correlate with the subjectively evaluated relative importance of image quality and depth perception with regards to overall 3D experience [3]:

$$w = 0.997 \cdot nZMA^{0.2393} \quad (8)$$

Therefore, from (4) the final QoE model for video+depth content is defined as follows:

$$QoE_v = (1 - 0.997 \cdot nZMA^{0.2393}) \cdot IQ + 0.997 \cdot nZMA^{0.2393} \cdot DP \quad (9)$$

Subjective experiments

The aim of the subjective experiments is to assess the above mentioned model for measuring the quality of attention area based encoded contents considered in the DIOMEDES project. The attention area defines the area where users would mostly look at while the leftover region defines the area where users pay less attention to. As detailed in the deliverable D4.4, the attention area is encoded at an increased quality over the leftover region for these subjective experiments. The Quantisation Parameter (QP) values used for encoding the attention area and leftover area are explained in Table 1. The attention and leftover areas were detected using the visual attention model presented in Section 3.1.

The subjective experiments were performed on Phillips WOWvx auto-stereoscopic display using video+depth contents according to the Double Stimulus Continuous Quality Scale (DSCQS) method specified in ITU-T BT500 standard [4]. Each test sequence was 10 s long. 15 subjects were attended for this subjective experiment. The results obtained from this experiment are shown in Figure 2.

Reference	QP over attention area	QP over leftover region
1	20	22
2	20	24
3	20	30
4	20	40
5	30	32
6	30	35
7	30	40
8	40	44
9	40	48

Table 1 - QP combinations used for encoding attention area and leftover region for subjective experiments

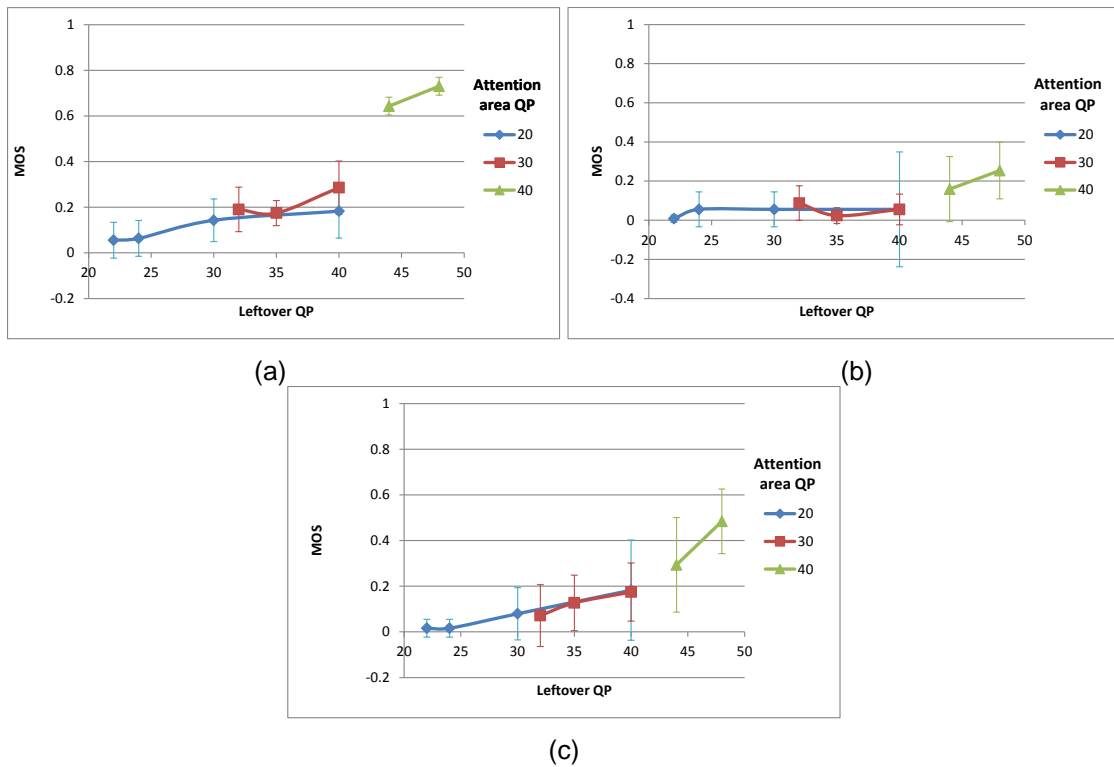


Figure 2 - MOS vs. QP of the leftover region for different attention area QP settings for (a) ballet, (b) champagne tower, and (c) music test sequences

In order to assess the fitness of the proposed objective metric, the correlation study was performed using logistic regression analysis technique using the following logistic function:

$$p = \frac{1}{1 + e^{(D - D_M)G}} \quad (9)$$

The experiment results are summarised in Table 2. The results shown in the table indicates a correlation of over 0.95 for the proposed 3D visual QoE metric. As a comparison, the table also shows the correlation for the VQM measured on the video component of the video+depth content. Figure 3, Figure 4 and Figure 5 present the objective quality values obtained from the VQM and the proposed 3D QoE metric vs. the QP of the leftover region for different attention area QP settings.

Metric	CC	SSE	RMSE
VQM	0.8513	0.4081	0.1229
Proposed 3D QoE metric	0.9564	0.1283	0.0689

Table 2 - Validation of the proposed 3D visual QoE model under attention area based coding

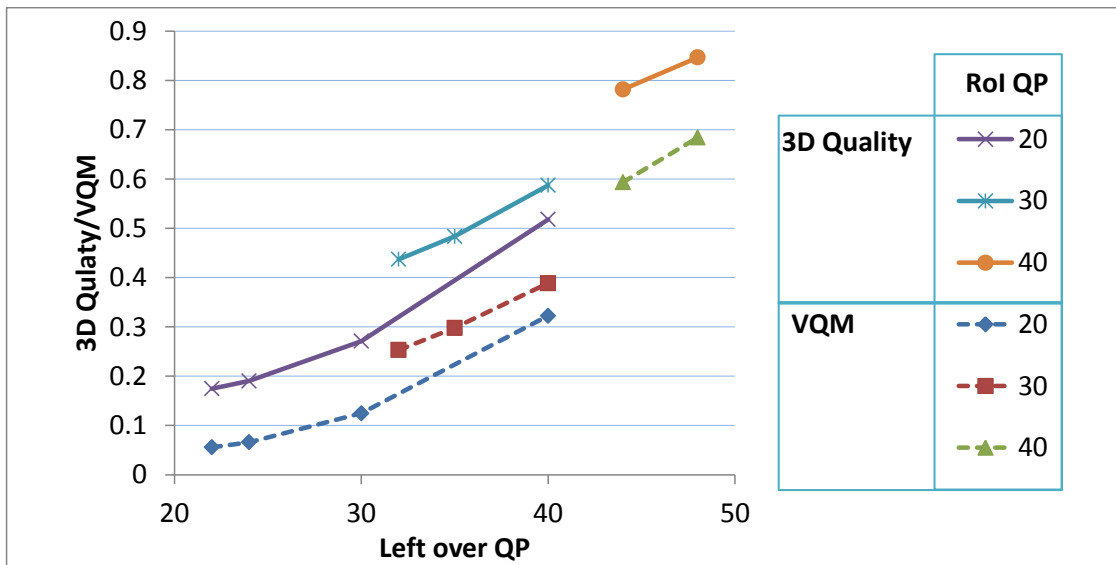


Figure 3 - Proposed 3D QoE metric and VQM vs. QP of the leftover region for different attention area QP settings for the ballet test sequence

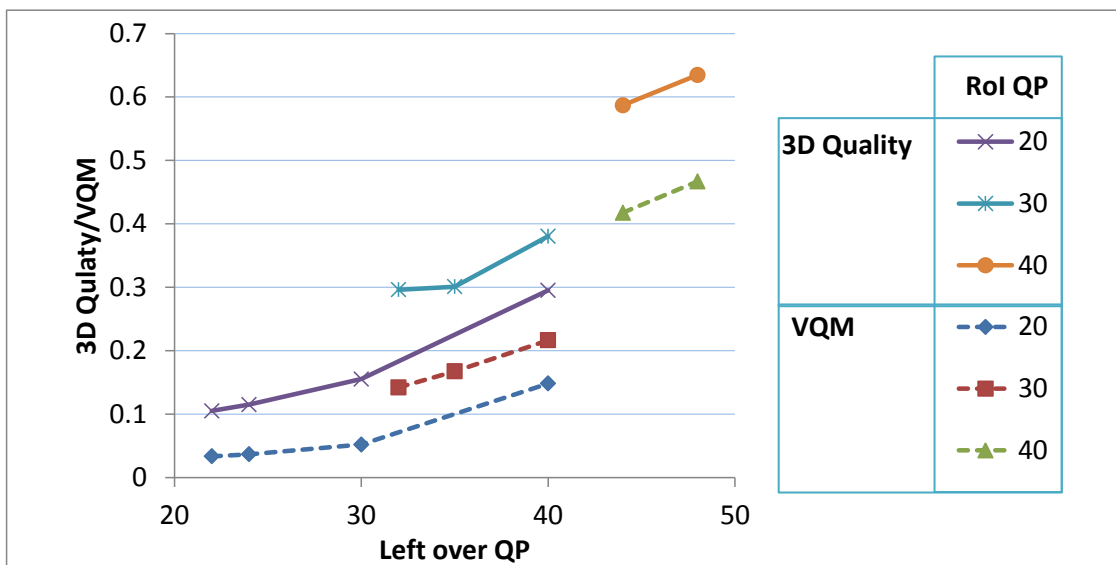


Figure 4 - Proposed 3D QoE metric and VQM vs. QP of the leftover region for different attention area QP settings for the champagne tower test sequence

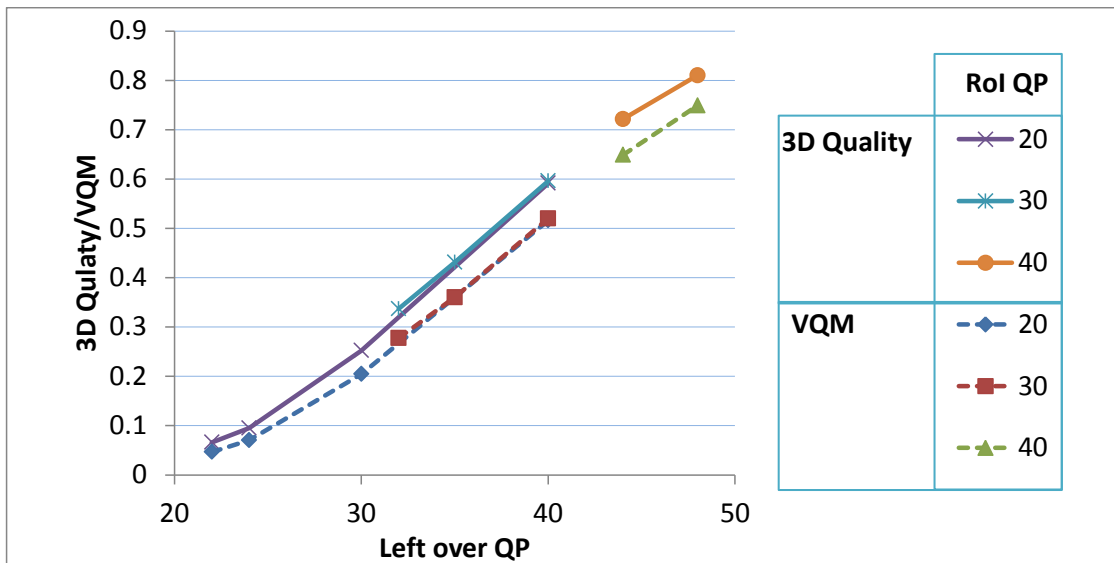


Figure 5 - Proposed 3D QoE metric and VQM vs. QP of the leftover region for different attention area QP settings for the music test sequence

2.1.3 QoE model for stereoscopic contents

Further subjective experiments have been conducted for assisting the attention area based coding for stereoscopic video. The aim of this experiment is to identify the ways of combining KPI parameters to determine the QoE for stereoscopic applications. For this experiments, video with strong attention areas were selected. They were encoded using the DIOMEDES encoder using the quantisation settings shown in Table 1. The subjective experiments were also conducted using the DSCQS method specified in ITU-T BT500 standard [4]. 15 viewers were used for this experiment. The video sequences are displayed on a 46-inch JVC passive stereoscopic display. The viewers assessed the overall 3D quality aspects of the test video sequences. The ambient illumination was set to 200 lux and the viewing distance was 3 m. Each video sequence is of 10 s long. .

Subjective results obtained from this experiment are shown in Figure 6. In order to model the visual experience recorded by the users, the feasibility of image quality KPI parameter, VQM, is examined in this study. The average VQM of the left and right views of the test sequences are shown in Figure 7. Table 3 shows the regression analysis results. Based on this regression analysis results, the following QoE metric is proposed for stereoscopic contents.

$$QoE_v = \frac{IQ_{left} + IQ_{right}}{2} \quad (10)$$

where IQ_{left} and IQ_{right} represents the image quality KPIs of left and right sequences subsequently.

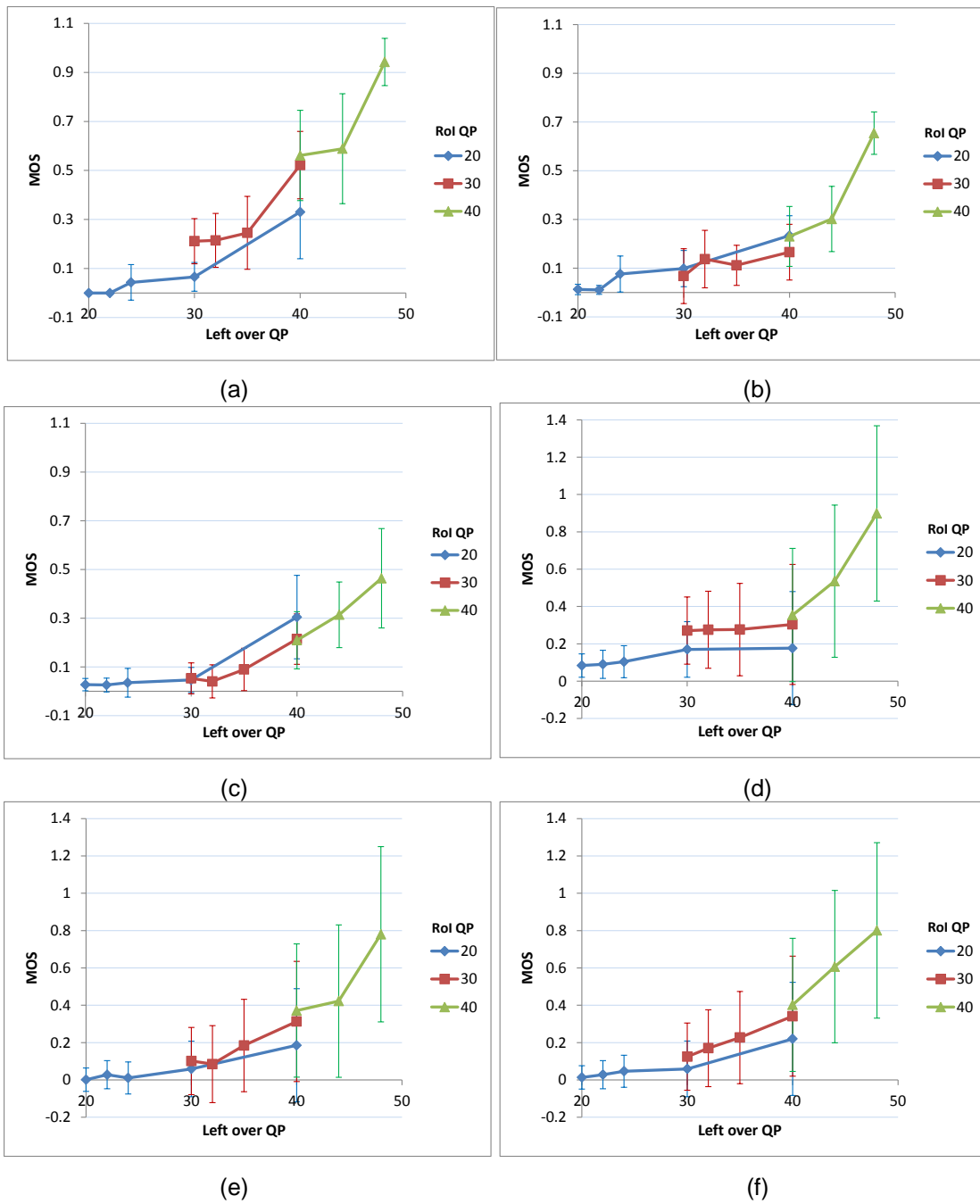


Figure 6 - MOS vs. QP of the leftover region for different attention area QP settings for (a) fencing, (b) music, (c) lecture sequences, (d) band, (e) cafe, and (f) Poznan street

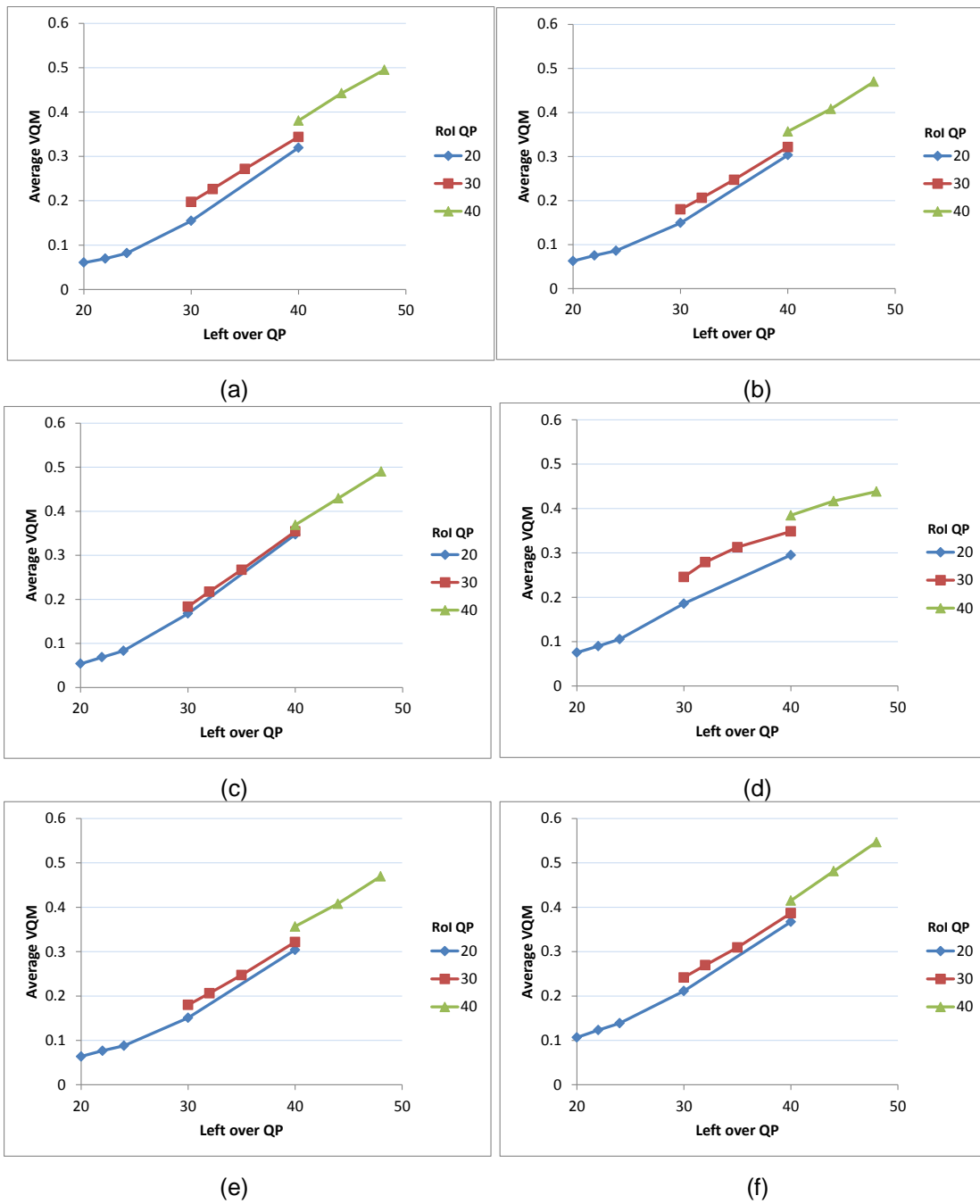


Figure 7 - Average VQM vs. QP of the leftover region for different attention area QP settings for (a) fencing, (b) music, (c) lecture sequences, (d) band, (e) cafe, and (f) Poznan street

Metric	CC	SSE	RMSE
Average VQM	0.9116	0.6043	0.0916

Table 3 - Validation of the proposed stereoscopic QoE model under attention area based coding

2.2 Perceptual Evaluation Methodologies

2.2.1 Audio-only experiments

- **Scope**

The audio-only test investigates how the decrease of bandwidth influences the users' audio experience. This helps recommending coding thresholds depending on the capacity of the available distribution channel. Audio material available by Fraunhofer IDMT was used for this test. Object oriented audio scenes are limited and impaired by the bitrates of the audio stream.

Important factors for perceived audio quality are the localization of sound sources and the spatial envelopment. Approaches like WFS enable more realistic sound including proper perception of direction, distance and elevation. An infinite number of loudspeakers and the use of a complex software model for reproducing the propagation of sound waves would create the best spatial impression. Unfortunately it is necessary to make compromises between the localization resolution and the complexity of the system, which is often combined with large costs and implementation complexity. In home applications, which are one target of DIOMEDES, the user will not be able to install a complete WFS-system. The number of loudspeakers will have to be reduced. To analyse the correlation of the systems number of loudspeakers and the achieved perceived audio quality it is necessary to conduct listening tests in terms of channel reduction. For the time being it is not possible to reduce the number of transmitted audio objects automatically before transmission whilst at the same time preserving the original intention of the sound designer. Due to this fact, the whole audio scene is transmitted and the channel reduction is done on the receiver's side.

- **Procedure**

The WFS systems at Fraunhofer IDMT were used in these tests to render test stimuli that each consisted of a full 32 channel object based audio scene. At the time of the tests, no dedicated listening test control and playback tools were available that would allow a simultaneous audio rendering and switching between a broad set of differently coded audio scenes (of 32 objects each) combined with timeline features (such as jumping to particular regions or looping marked sequences).

These limitations rule out methods such as AB-X (Rec. ITU-R BS.1116-1 [1]) or MUSHRA (Rec. ITU-R BS.1534-1 [6]), as looping and near-instantaneous switching is mandatory for these. Therefore, the well-established single stimulus method "Absolute Category Rating" (ACR) with a labelled 11-point scale, as described in [7], was used. The term "single stimulus" refers to the lack of a reference¹ to which the participants have to compare the stimuli to rate. Instead, the method asks for an "absolute rating" of each item. The stimuli are presented in sequence, one at a time, with a 10s rating period after the playback of each stimulus (see Figure 8). In this case, the actual test took approximately 25 min per participant.

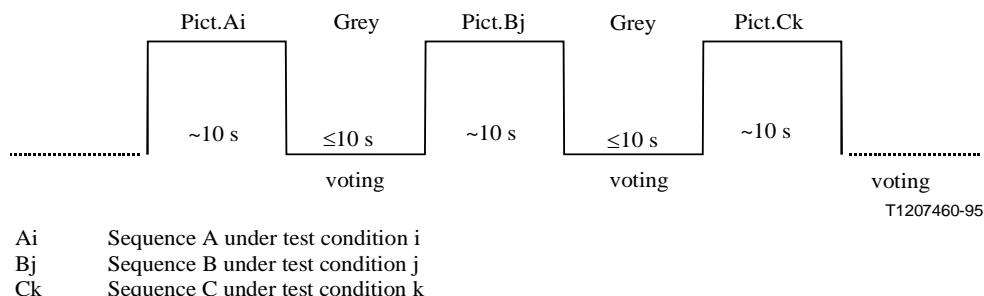


Figure 8 - Stimulus presentation in the ACR method [P.910]

¹ An uncompressed version of the stimulus.

This method has the advantage that it asks participants for a true quality judgment, rather than testing a hearing threshold. It is easy to implement and occasionally realized with pen and paper instead of a rating interface. The fact that there is no interaction of the participants (e.g., looping of stimuli) allows for a precise schedule, as each participant needs exactly the same time. However, context issues (i.e., participants are influenced in their opinion by the quality of the stimulus presented before) are a frequent problem and the method is highly demanding on the participants' expertise. Hence 14 stimuli were presented twice in this test in order to be able to screen the participants' reliability.

- **Parameters**

Test Items

Items covering a range of spatial listening impressions and the influence of different parameters were used in this test. These items were seven pieces of 20 seconds each, cut from the four following IOSONO WFS demos available at Fraunhofer IDMT:

Name	Attributes
"Jeklin"	multichannel source 16 bits per sample 2 clips Jazz music and audience music sources from fixed directions moving audience sources
"Sommerfeld (long)"	multichannel source 16 bits per sample Jazz music static sources
"Trance"	mono sources 24 bits per sample 3 clips Percussion and vocal elements moving sources
"World of Sound Dschungel (WOSD)"	24 bits per sample ambient Jungle noise and sounds moving sources

Table 4 - Attributes of the Selected Test Items

Seven pieces of 20 seconds each were cut from this material. A total of 35 impaired conditions were created out of these.

Conditions

The test items were encoded using AAC-LC (MPEG-4 Audio Object Type 2) audio compression with the MPEG-4 HE-AAC Fast Evaluation Encoder provided by Fraunhofer IIS. To create audio files compatible to the IOSONO sound system available at Fraunhofer IDMT,

the AAC files had to be decompressed again to WAV format. This was done using the corresponding decoder also provided by Fraunhofer IIS. The following bit rates were used for encoding (sample rates were set by the AAC encoder):

- 8kbps, 16kHz
- 16kbps, 16kHz
- 32kbps, 32kHz
- 64kbps, 48kHz
- 96kbps, 48kHz

The original source material featured sample rates of 48 kHz. Due to the down sampling introduced by AAC encoding, the audio material was re-sampled back to 48 kHz with the professional grade iZotope 64-bit SRC algorithm using the highest quality settings available. The demos Jeklin and Sommerfeld (long) are stored as a multichannel WAV File and had to be split into separate mono files prior to and rejoined after encoding. All clips were decoded providing the sample sizes of 16 and 24 bits per sample according to the properties of their sources.

Test System

The WFS demo system installed at Fraunhofer IDMT was used for playback of the items. It utilizes 88 speakers and 4 sub woofers for wave field synthesis.

The computational system roughly consists of 4 main processing systems: one playback PC, one control PC and two rendering PC. The playback PC stores the audio material and its according scene description information. At a command from the control PC, it will play out this data to the rendering PCs via Ethernet. The control PC serves as an operational interface. Via scripts or a graphic user interface (GUI), the user is able to select, load, and start and stop the playback of demo material. The rendering PCs combine the audio data and scene description input from the Playback PC and calculate, with 64 audio channels each, the final audio signal for each loudspeaker of the setup.

Implementation of test method and graphical user interface (GUI)

For each participant, an individual Linux shell script was created and executed on the control PC. They contained the required commands to start playback and to present the rating interface after playback. These scripts were generated automatically, randomizing the playback sequence for each participant.

The graphical user interface consisted of a simple rating slider. The application was programmed in C++ utilizing the Qt library v4.6.3. Besides of displaying the actual rating slider and a 10 second countdown, a button to immediately record the rating and skip the rest of the countdown was available as well.

The rating slider application created a score file for each participant, keeping track of the played conditions and their ratings, which then could easily be processed for evaluation. In case a participant would not have rated an item, this would have been noted in the file.

• **Statistical Analysis**

Participants & Post-screening

The initial listening test panel consisted of 28 participants, mostly employees and students of Fraunhofer IDMT. Their expertise ranged from “naïve” to “expert” [8]. Each participant received training before the actual test, in which they listened to an automated sequence: the worst condition (i.e., 8 kbps, 16 kHz), the best condition (96 kbps, 48 kHz), and the worst condition again of each of the seven items. This allowed the participants to establish a feeling of the anchor points of their personal subjective quality scale. The training took about nine minutes per participant and was conducted directly prior to each participant’s test.

As said before, the test method is highly demanding on the participants and 14 stimuli were

repeated to allow for a post-screening. The participants were allowed to have one major difference in their rating of these stimuli as well as further differences smaller than 1.5 categories on the scale. As expected, out of the 28 participants tested only 15 could be taken into analysis after the screening. These consisted of three female and twelve male participants, the youngest being 25, the oldest 38 years of age. The average age was 29.8 years.

Results

Of the participants remaining after the post-screening, “mean opinion scores” (MOS), median and confidence intervals were calculated². These results are shown in Figure 9.

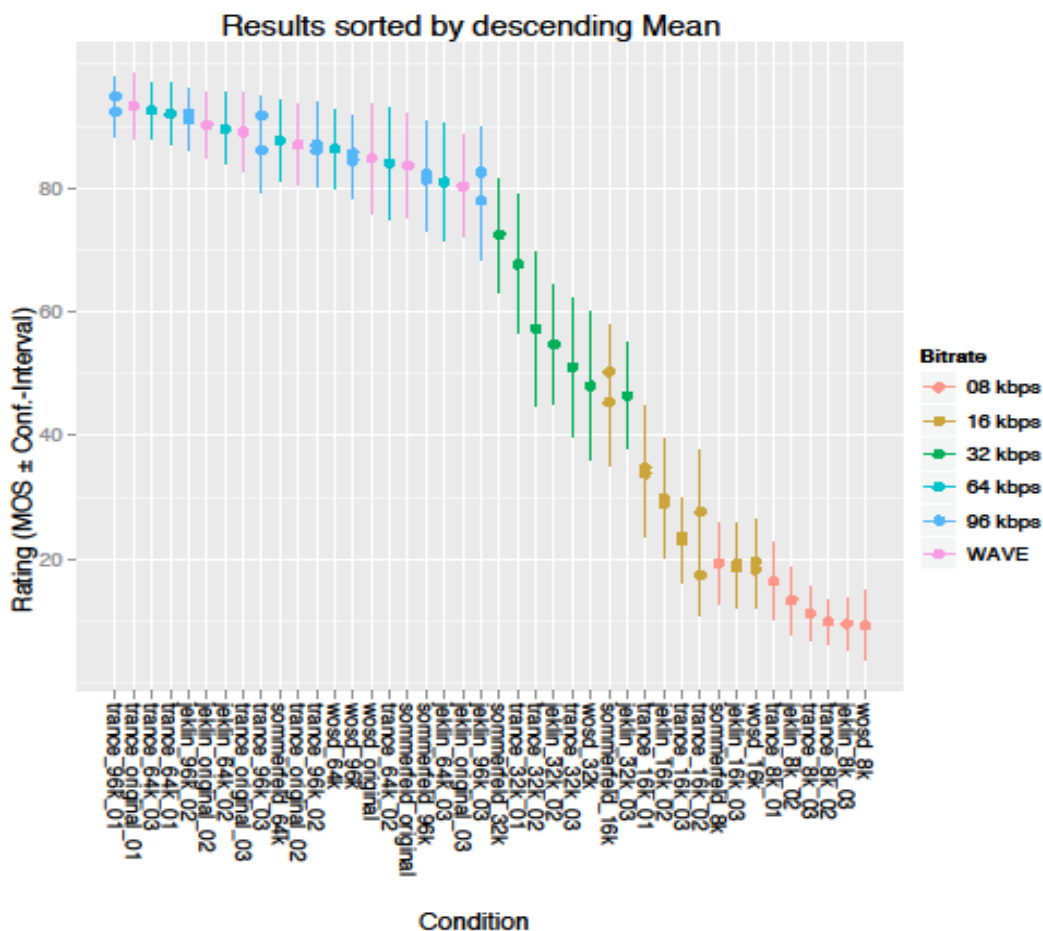


Figure 9 - Mean values for items and conditions

The items are sorted by descending mean opinion scores. The bit rates are indicated in colour³. Stimuli with two MOS-values (“dots”) are those that were presented twice during the test.

Figure 9 shows that the participants perceived the quality of bit rates equal to or higher than 64 kbps (coloured teal and blue) mostly as “excellent” (ratings higher than 80). Only one stimulus (the repetition of “Jeklin 03” at 96 kbps) has a mean opinion score of “good” (ratings

² Data of two participants was corrected as both noted that they had mistakenly given one stimulus a “0” score where they wanted to give a “100”.

³ The bit rate is denoted in the stimulus’ name as well

between 60 and 80), although the confidence interval here suggests that the true mean rating might be in the “excellent” region as well.

It is noticeable that the uncompressed stimuli (“WAVE”, coloured in magenta) were frequently rated worse than the encoded versions, with confidence intervals reaching into the “good” range, but this difference is statistically not significant due to overlapping confidence intervals. It is possible that the clear tendency towards the category “good” for the stimulus “Jeklin 03” at bit rates of 96 kbps, 64 kbps, and WAVE indicates that the majority of the participants disliked this stimulus in general⁴. Overall, these findings lead to the conclusion that the participants did not perceive significant differences between the uncompressed signals and signals encoded with 64 kbps or 96 kbps.

At 32 kbps, the participants started to rate the perceived quality as “good” or “fair”. A majority of the stimuli achieved mean opinion scores in the range of “fair”, but the rather large confidence intervals indicate that they are taken to be “good” by some participants. This, however, excludes the two lowest rated ones (“Jeklin 03” and “World of Sound Dschungel” (“WOSD”)).

The ratings and confidence intervals of the 32 kbps versions of “Sommerfeld” and “Trance 03” illustrate that for some participants these stimuli are comparable to the quality of other uncompressed stimuli. In case of “Sommerfeld” the confidence interval even overlaps with the same uncompressed stimulus.

The participants easily identified the lower bit rates, i.e. 16 kbps was mostly found to be “poor” and 8 kbps “bad”. Again, the stimuli “Sommerfeld” and “Trance 01” perform noticeably better than others, with the 16 kbps version of “Sommerfeld” clearly being rated as “fair” and the respective “Trance 01” having its confidence interval reaching into the range of “fair”. At 8 kbps, most participants rated these items “bad”, but their confidence intervals extend to “poor”. The ratings of these two stimuli indicate that they were easy to encode.

On the other hand “WOSD” and “Jeklin” are rated worst at bit rates of 8, 16, and 32 kbps. In the latter case, the confidence intervals tend to slightly reach into “poor”, while the 16 kbps versions are rated in between “poor” and “bad”. The poor performance of these two items is easily explained: “Jeklin” contains applause, which is always a major challenge for audio codecs due to its uncorrelated structure that makes prediction practically impossible and causes audio codecs to create a more or less white noise-like signal. The stimulus “WOSD” contained flaps of a bird’s wings, which showed very clear signs of degradation even at higher bit rates.

Note that the terms of the German translation of the ACR quality scale are relatively equidistant in the sense of “perceived distance”, while the terms of the original English version are not [9]. As a result, the English “poor”, for example, denotes a quality much lower than the German counterpart “mäßig”. Figure 10 shows the differences between the translations.

⁴ This is supported by the fact that the stimulus is being rated low at all bit rates.

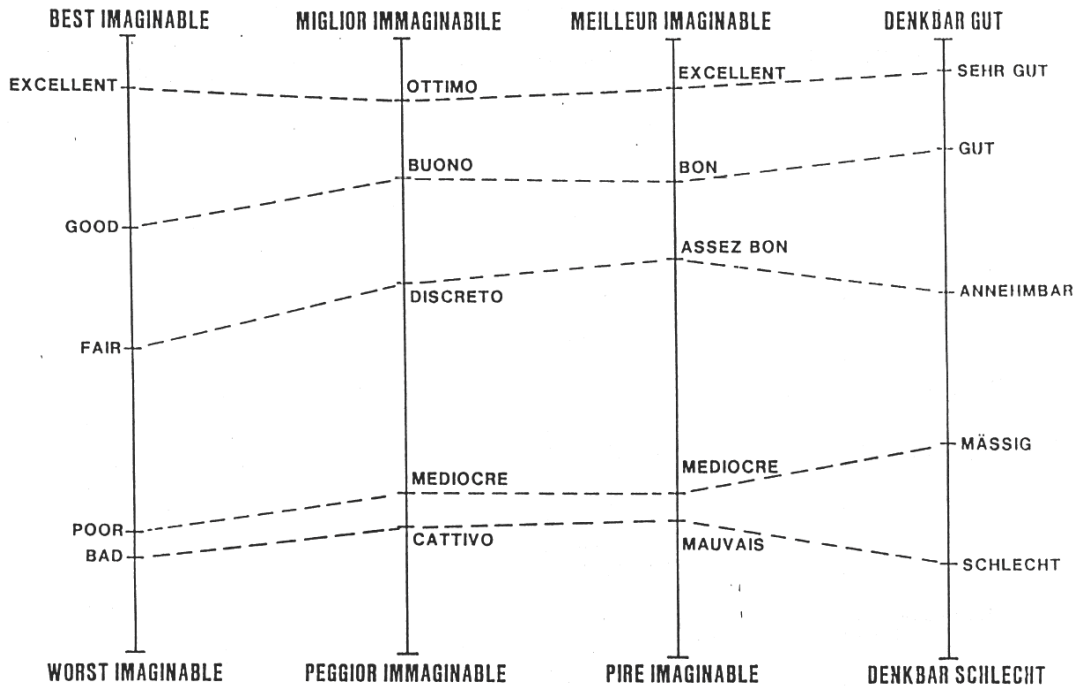


Figure 10 - Relation of quality terms

Figure 11 gives the results as boxplots. Every boxplot depicts the ratings of all stimuli of the respective bit rate. The boxplots show the median (horizontal line in the box), the 25% and 75% quantiles (boxes). The whiskers show the interquartile range times 1.5.

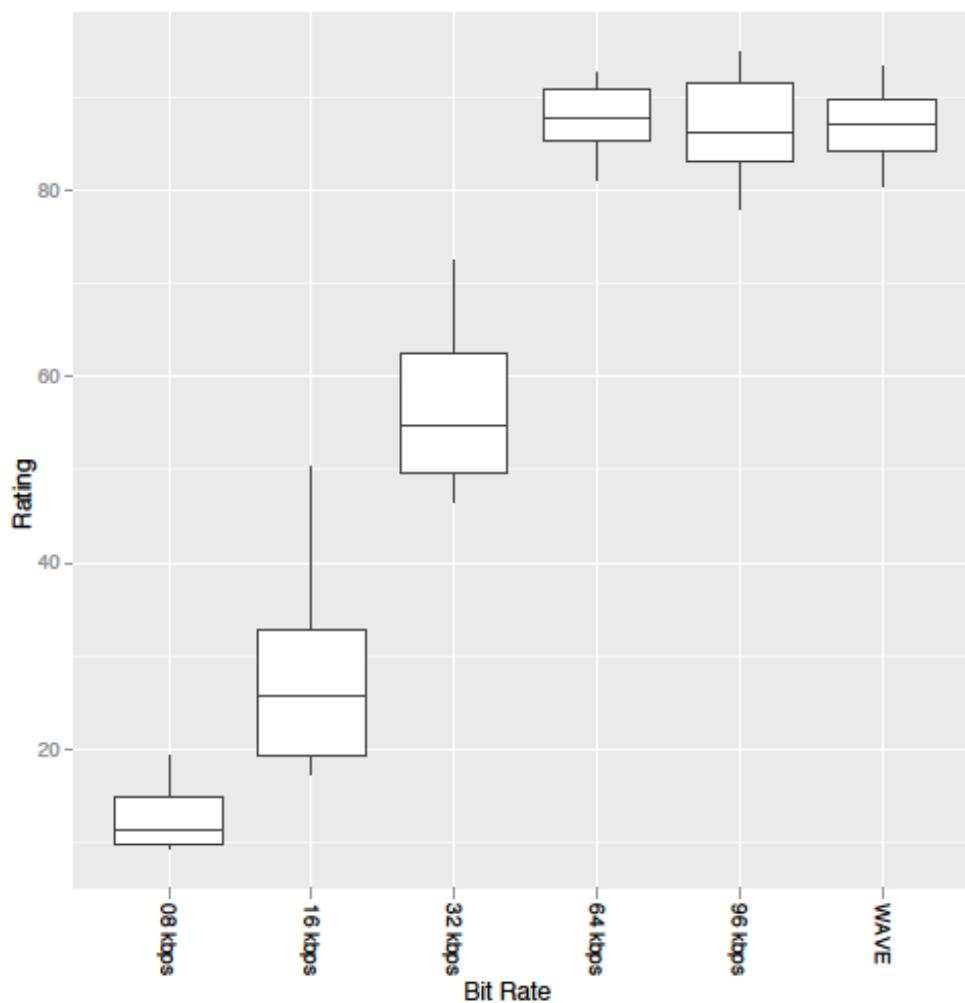


Figure 11 - Boxplots for ratings of all stimuli of the respective bit rate

The Boxplots clearly confirm the previous findings: In total, the participants could not discern the uncompressed signals from the 96 and 64 kbps at all, while they were perfectly able to distinguish between 32, 16, and 8 kbps.

The Boxplots also show a rather large gap between the higher qualities and 32 kbps. On the one hand, it is easily conceivable that a bit rate between 32 and 48 kbps might provide a perceived quality that would fill this gap. On the other hand, some stimuli encoded with 32 kbps were already indiscernible with higher qualities. Another explanation for the gap could simply be that the scale is not equidistant at that point: note that Figure 10 shows this gap as well.

Despite the issues of the method, the test leads to the conclusion that bit rates of 64 kbps or higher are perceived as “excellent” and as good as the original “wave” file. Without a reference – which is the case in real life situations – even the most experienced listeners were unable to discern the higher bitrates, frequently giving 64 kbps stimuli scores of up to 100, despite that they were presented the uncompressed version minutes ago in the training. The fact that with some content even 32 kbps proved to be hard to distinguish from higher bit rates indicates that for some parts of the audience this bit rate might be unnoticeable or at least sufficient.

The results of this assessment, especially the knowledge concerning perceptually noticeable quality differences, will be brought into the DIOMEDES project. Based on this evaluation,

coding thresholds depending on the capacity of the available distribution channel are recommended. The results of this investigation also can be used for further work towards the improvement of the QoE models.

2.2.2 Audio-visual experiments

- **Scope**

In the audio-visual test the horizontal spatial congruency of auditory and visual objects in a 3D sound and 3D video setup is evaluated. The goal of the experiments is to find out the acceptance threshold of horizontal angular displacement of auditory and visual objects in a 3D audio-visual scene. Research on the congruency of audio and video objects has already been done for 2D video systems ([10],[11]). The audio-visual effects of angular displacement within 3D video and audio reproduction systems have not been researched yet.

For the audio-visual test the audio objects are placed in different angles to the corresponding video object. Stereoscopic video material and corresponding audio recordings from UNIS will be used for the test. The localization of the audio objects will be an important factor in this test. Thus, a WFS (wave field synthesis) system will be used for audio reproduction. This enables precise audio object positioning and localization. As home applications are in the scope of DIOMEDES project, a system with reduced number of loudspeakers will also be tested. With the results it will be possible to make statements about the displacement thresholds, for which an unimpaired perception is possible.

- **Procedures**

The well-established single stimulus method “Absolute Category Rating” (ACR), as described in [7], was used for the test. The term “single stimulus” refers to the lack of a reference to which the participants have to compare the stimuli to rate. Instead, the method asks for an “absolute rating” of each item. The stimuli are presented in sequence, one at a time, with a 7s rating period after the playback of each stimulus (see Figure 8). In this case, the actual test took approximately 40 min per participant.

All items were presented twice and rated on two different scales. On the first scale participants were asked for the overall quality of the video on the “Quality Scale” [12]. On the second scale participants were asked to rate, if they could perceive a difference in the position of the presented audio and video on the “Impairment Scale” [12].

“Quality Scale” [12]

5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

“Impairment Scale” [12]

5	Inaudible
4	Audible but not annoying
3	Slightly annoying
2	Annoying
1	Very annoying

The Test started with a training phase, in which the participants watched and rated 10 items. These items represented the whole quality scale of the items under test. The test participants were able to train the rating of the items on the two scales. After 7 seconds of rating time the next item was presented automatically. In the actual audio-visual test the participants watched all 48 items twice and rated them.

- **Parameters**

Test Items

The items were produced from DIOMEDES Demo content provided by UNIS. The stereoscopic videos “music scenario” and “lecture scenario” were used for the tests. In the “music scenario” three musicians play their instruments. Advantage of this video is that all

three instruments are recorded separately and so they can be used as separate audio objects. With the WFS system these audio objects can be placed independently at any position in the room. In the “lecture scenario” there is one audio source, the male speaker. This audio object can also be placed at any position in the room. Table 5 shows the screenshots of the two contents and describes the stereoscopic video sequences in more detail.



Screenshot	Description
	<p>“music scenario”</p> <p>Length: ~ 10 sec</p> <p>Audio: 3 audio objects, one for each instrument</p> <p>Video: medium spatial details, low temporal motion, medium amount of depth, low depth dynamism, medium depth complexity, no scene cuts</p>
	<p>“lecture scenario”</p> <p>Length: ~ 10 sec</p> <p>Audio: one audio object for the male speaker</p> <p>Video: medium spatial details, low temporal motion, medium amount of depth, low depth dynamism, medium depth complexity, no scene cuts</p>

Table 5 - Screenshots and descriptions of the used test items

Conditions

In total 48 items were produced with different conditions. The above mentioned contents “music scenario” and “lecture scenario” were used. Furthermore, the audio was presented on two different systems. One system was the WFS system with 88 loudspeakers; the other system was a system with a reduced number of loudspeakers, in which only every fourth loudspeaker was used for audio reproduction.

The test items were encoded using AAC-LC (MPEG-4 Audio Object Type 2) audio compression with the MPEG-4 HE-AAC Fast Evaluation Encoder provided by Fraunhofer IIS. To create audio files compatible to the IOSONO sound system available at Fraunhofer IDMT, the AAC files had to be decompressed again to WAV format. This was done using the corresponding decoder also provided by Fraunhofer IIS. The following bit rates were used for encoding (sample rates were set by the AAC encoder):

- 32 kbps
- 48 kbps
- 64 kbps

According to the results from the audio-only experiments (see section 2.2.1) these bitrates were used as this seems to be an interesting range between excellent and poor quality perception. Lower bitrates are not tested, as they were perceived as having a poor or bad quality, which is not acceptable.

The audio objects were positioned at the position of the video objects on the screen. These positions were taken as the initial position at 0°. For the angular displacement the whole auditory scene was rotated about the viewing point by 5°, 10° and 20° degrees.

Test System

As one system the WFS demo system installed at Fraunhofer IDMT was used for playback of the audio objects. It utilizes 88 speakers and 4 sub woofers for wave field synthesis. As a second system the same WFS system is used, but only every fourth loudspeaker (22 loudspeakers) is used for audio playback.

The computational system for the tests consists of the following components. One audio processing system conducts the decoding and rendering of the audio scene to the loudspeaker setup. One PC processes the incoming stream and conducts video decoding and rendering of the video scene to the LG monitor. Another PC is the streaming and control PC that transmits the test sequences (audio and video) to the audio and video rendering PCs. This streaming PC also stores the video and audio material and its according scene description information.

The video is played back on a LG 47LD950 TV (47") that uses a passive polarized screen with polarized glasses for the 3D effect. The videos are presented via PC HDMI connection in side-by-side mode. The LG TV uses a passive polarization technique to display the stereoscopic video in 3D mode. Figure 12 shows a schematic representation of the test set-up in the listening room at Fraunhofer IDMT. Positions of loudspeakers, TV screen and listener/viewer are depicted.

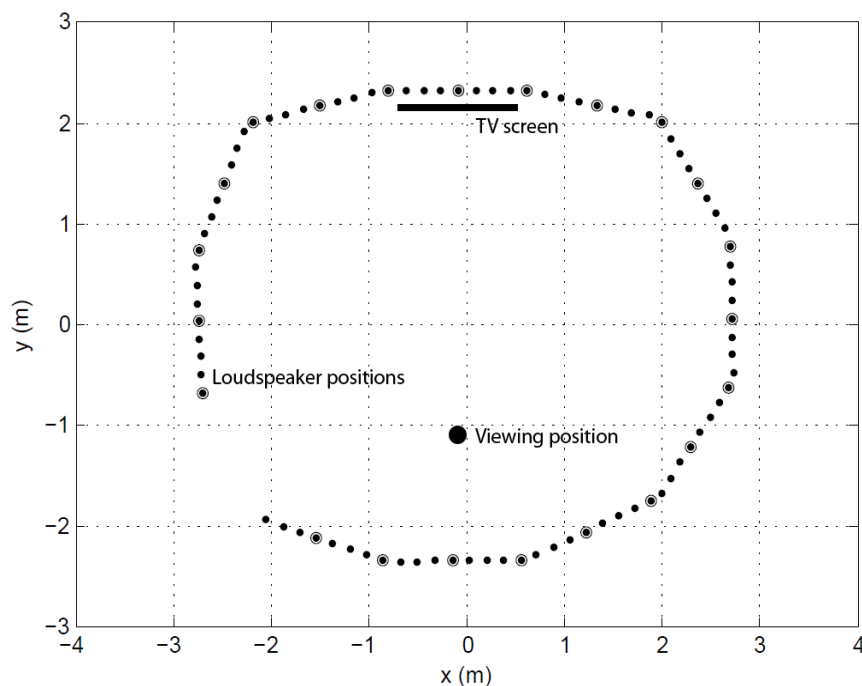


Figure 12 - Schematic representation of test set-up, as system with reduced loudspeakers only the ones with one dot and a circle are used

As two systems are used for audio and video playback, the synchronization is rather challenging. The audio and video streams are multiplexed and played back over network connections from the corresponding audio and video playback devices. A script directs the output and runs the playback automatically.

Implementation of test method

For each participant an individual script, each with a different randomized order of the test items was produced. This script was started on the control PC and played back all items. Test participants rated each item one after another on provided answer sheets with the two rating scales. Paper answer sheets were used to not affect the perceived stereoscopic video quality with another 2D monitor for a rating GUI.

- **Statistical Analysis**

Participants & Screening

Participants were screened for stereoscopic vision using the Randot-stereo test. Students and employees from Fraunhofer IDMT took part in the test. A total of 14 test participants (9 male and 5 female) took part in the test.

Analysis of data

The data was analysed using Excel 2003 and SPSS 17. No outliers were detected and all data from all participants was used for the statistical analysis. Using the Kolmogorov-Smirnov Test, all data was checked for normal distribution. This analysis showed that the ratings of the individual items were not normal distributed (all $p > 0.05$). Therefore, non-parametric tests were used for further analysis. Friedman and Wilcoxon Signed Rank Test were used to analyse the relation between independent and dependent variables for the ratings of overall quality and perception of position changes. Wilcoxon Signed Rank Test was especially used to analyse significant differences between two related samples.

Results

Overall quality rating

Friedman test revealed a significant influence of the parameter combination when the ratings were averaged over both contents (Friedman: $p < 0.05$). Further Wilcoxon tests were done to analyse the influence of each tested parameter. The reproduction system had no influence on the perceived overall quality. Within all comparisons, no significant difference could be found (Wilcoxon: all comparisons $p > 0.05$).

For some parameter combination the content has an influence on the quality rating. That being the case for parameter combinations 32 kbps and the angle of the audio object of 0° , 10° and 20° (Wilcoxon: all comparisons $p < 0.05$). For all other parameter combination, no influence of the content on the quality rating could be found (Wilcoxon: all comparisons $p > 0.05$). It seems that the influence of the presented content only matters if the quality of the reproduced audio is low (32 kbps). Figure 13 shows the results of the overall quality rating for the two contents and the other parameters (bitrate and angle). Furthermore, it can be seen that the quality rating, very much depends on the angular displacement of the audio objects.

The angle of the audio objects has a significant influence on the perceived overall quality. No significant influence could be found between parameter combination 32 kbps and angles of 0° and 5° (Wilcoxon: $Z = -0.156$, $p > 0.05$), for all other parameter combination the angle of the audio object position displacement had an significant influence on the quality rating (Wilcoxon: all comparisons $p < 0.05$).

The influence of the bitrate on the quality ratings is analysed for each angular displacement of the audio object separately. For the original audio object placement (0°) the quality of the 32 kbps items is rated worse than all other items with higher bitrates. A significant difference in the ratings could be found between these signals (Wilcoxon: 32 kbps compared to 48/64/192 kbps: $p < 0.01$). Comparisons between the higher bitrates showed no significant differences (Wilcoxon: comparisons: 48/64, 48/192, 64/192 kbps: $p > 0.05$). For an audio object displacement of 5° a significant difference could only be found for the bitrates 32 and 192 kbps (Wilcoxon: $Z = -2.888$, $p < 0.01$). For all other bitrate combination no significant influence on the overall quality rating could be found (Wilcoxon: all comparison $p > 0.05$). For an angular displacement of 10° the bitrate had only a low effect between bitrates 48 and 64 kbps

(Wilcoxon: $Z = -1.922$, $p < 0.05$). For all other comparisons, no influence of the bitrate could be found (Wilcoxon: all comparisons: $p > 0.05$). For the angular displacement of 20° a significant influence of the bitrate could be found between bitrates 32/48, 32/64, 32/192 and 64/192 kbps (Wilcoxon: all comparisons $p < 0.05$). For bitrates 48/64 and 48/192 no significant effect could be found for the quality rating (Wilcoxon: all comparisons $p > 0.05$).

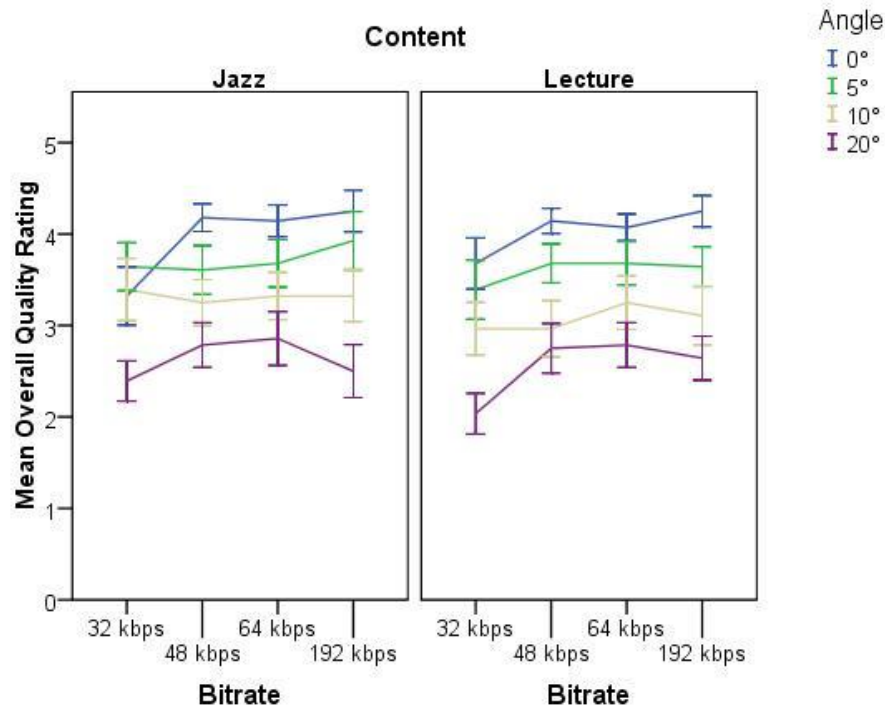


Figure 13 - Mean Overall Quality Rating for the two contents, Error bars show the 95% confidence intervals

The analysis of the data shows that the participants were able to perceive the angular displacement of the audio object and that the perceived overall quality depends on the angular displacement respectively on the congruence of audio and video objects. Figure 14 shows the mean overall quality ratings averaged over the content and reproduction system. As shown before, the bitrate and the angular displacement of the audio object mainly influence the perceived overall quality of the presented items. The overall quality ranges from poor (Item: 30 kbps, 20° angular displacement, Mean = 2.21, $sd = 0.594$) to good (Item: 192 kbps, 0° angular displacement, Mean = 4.25, $sd = 0.513$), with the most quality ratings between fair and good.

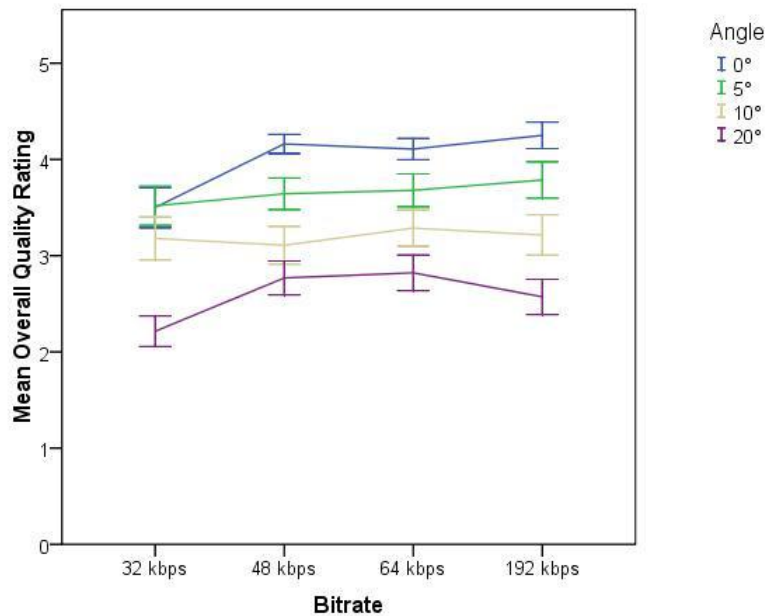


Figure 14 - Mean Overall Quality Rating, averaged over content and reproduction system, Error bars represent 95% confidence intervals

Angular displacement of audio and video objects

Wilcoxon tests were done to analyse the influence of each tested parameter. For all parameter combinations the influence of the content was analysed using Wilcoxon test. No significant influence of the content on the impairment ratings could be found (Wilcoxon: all comparisons $p > 0.05$). A significant influence of the audio reproduction system could only be found between the parameter combination 48 kbps and 10° audio object displacement (Wilcoxon: $Z = -2.449$, $p < 0.05$). However, Figure 15 depicts the mean impairment ratings averaged over the two contents for the parameters bitrate and angle. Small differences between the mean ratings are visible for some parameter combinations, but they are not significant.

Furthermore, one can see differences in the mean impairment ratings for each angular displacement of the audio object. The original audio objects position (0°) and the displacement of 5° are rated between excellent and good. A difference between these two audio object positions might have not been perceivable for all participants. However, a significant effect could still be found between both angular positions (Wilcoxon: 0°/5° for all bitrates $p < 0.05$). So participants were able to detect an angular deviation of 5° but were not annoyed by it. A significant effect of the audio object position on the impairment rating could further be found for the angles 10° and 20° for all bitrates (Wilcoxon: all comparisons $p < 0.001$). Participants rated a 10° deviation of audio and video object between perceivable, but not annoying and slightly annoying. A deviation of 20° was rated between slightly annoying and annoying.

Effects of the bitrate on the impairment rating could only be found for some parameter combinations. The bitrate had a significant influence on the impairment ratings for an angle of 0° between bitrates 48/192 kbps and 64/192 kbps (Wilcoxon: $p < 0.05$; all other combinations Wilcoxon: $p > 0.05$). For an angular deviation of 5° a significant influence of the bitrate could be found for bitrate 48 kbps (Wilcoxon: all comparisons for 48 kbps $p < 0.05$). In Figure 16 it can be seen that the ratings for angle 5° and bitrate 48 kbps are a bit lower than the ratings for the other bitrates for the same deviation. For an angular deviation of 10° no significant effect of the bitrate on the impairment ratings could be found. Items with a bitrate of 32 kbps and a deviation of 20° were rated slightly lower than the items with higher bitrates (Wilcoxon:

comparisons with 32 kbps $p < 0.05$). The influence of the bitrate on the impairment ratings is only visible for some parameter combinations and seems to depend on the angular deviation. One reason might be that different artefacts occur for different angles and bitrates. Items with a deviation of 20° were perceived worse than the other items. Especially the items with 32 kbps and 20° deviation were rated worse (Mean = 2.12, sd = 0.634). Items with the original position of the audio object (at the place of the corresponding video object) were rated best. Especially the item with 192 kbps and 0° deviation was rated best (Mean = 4.66, sd = 0.478), meaning that participants didn't perceived any deviation of audio and video objects.

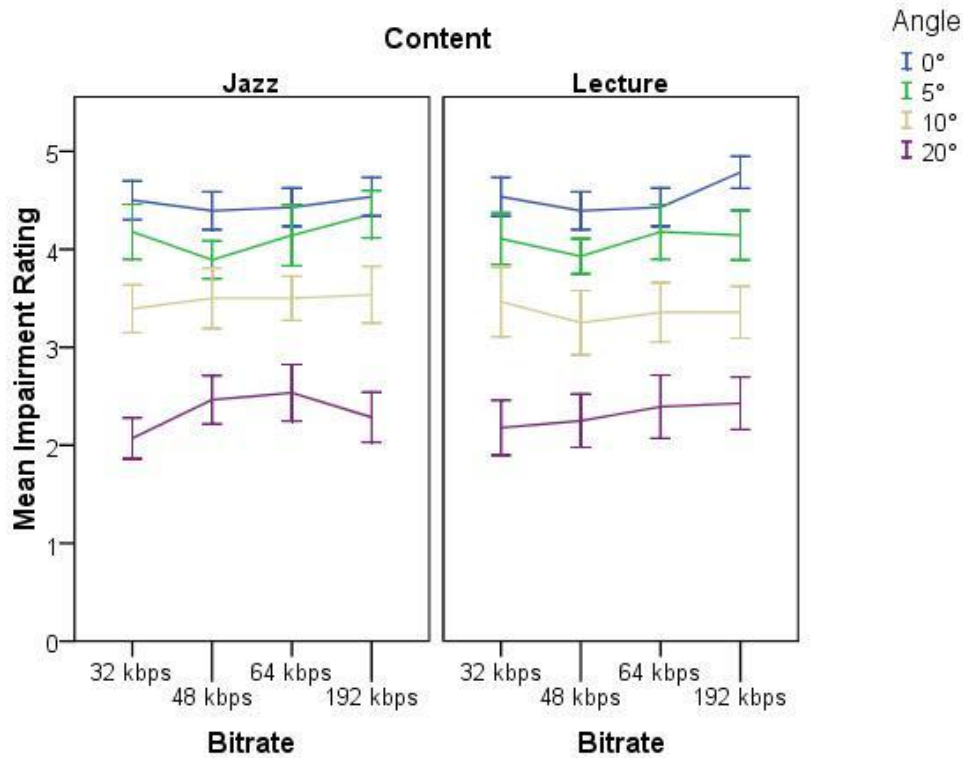


Figure 15 - Mean Impairment Ratings of the Angular displacement of audio and video objects

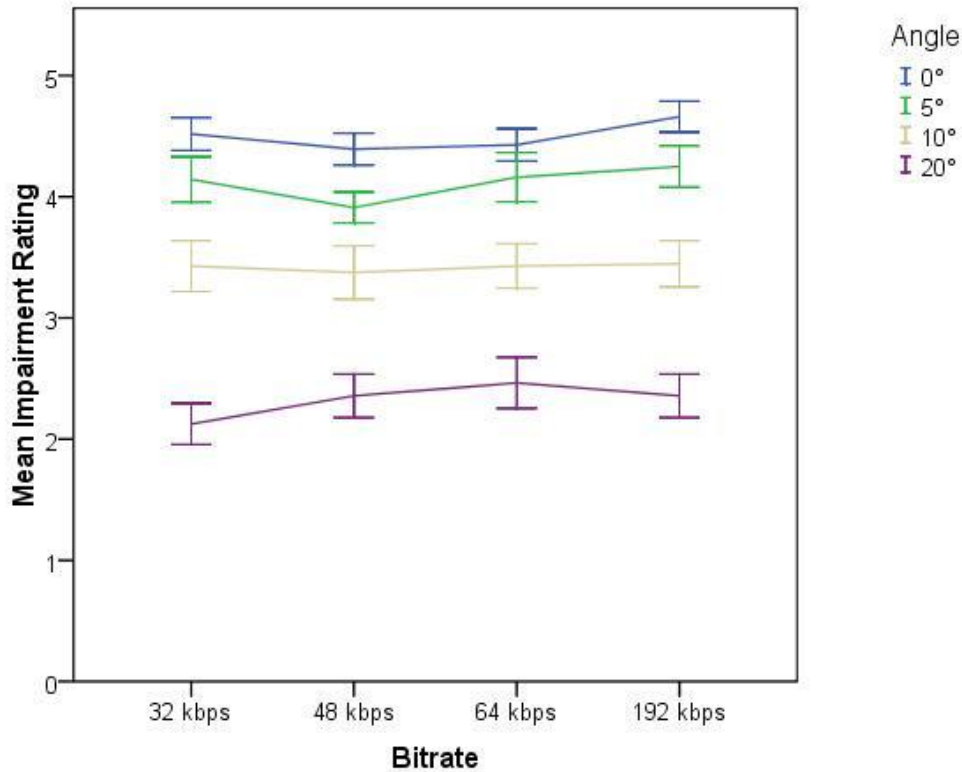


Figure 16 - Mean Impairment Ratings for parameters bitrate and angle

In summary, it can be said that the test participants were able to detect the angular deviation of the audio object respectively to the video object. A deviation of more than 5° was perceived as slightly annoying or even annoying.

Summary

The audio-visual quality evaluation showed that participants perceive an angular deviation or displacement of the audio object respectively to the video object as annoying, when the deviation is more than 5°. Furthermore, only a small influence of the bitrate could be found in the quality ratings. Based on the results of the previous audio-only test, only bitrates from 32 kbps and higher were used in the test, as lower bitrates were poor or even bad (compare 2.2.1). For these chosen bitrates and the presence of a 3D video the effect of the bitrate on the perceived overall quality was small or not even detectable. Further tests with different bitrates would be needed to evaluate the influence of the audio bitrate on the perceived overall quality in more detail.

The main goal of the tests, evaluating the influence of a displacement of audio and video objects, was achieved and an influence of the angular deviation of the audio objects on the perceived overall quality could be shown. The actual threshold between not annoying and slightly annoying perceived displacement might be between 5° and 10°. Further tests could evaluate the threshold in more detail. However, these first results show that an accurate positioning of audio objects respectively to the video objects on the screen is important for good quality perception as well as an enjoyable and not annoying media perception.

3 ATTENTION MODELLING

In the following sections the developed Audio and Visual Attention models are presented.

3.1 Visual Attention Modelling

This chapter is an extension to the description of the visual attention modelling given in Deliverable D3.2 (or D3.3) chapter 2.3. Please find the detailed description of the basic concepts of the model in those deliverables.

In the following the further development of the attention model is shown. This includes a revised and more detailed categorisation of the model components, as well as a more in-depth description of the details and usage of the developed framework. Additionally the results of the conducted user tests are presented in section 3.1.3.

3.1.1 Model Concept

The developed models are based on a multilevel system using a graph based approach to flexible combine different analysing and processing algorithms. The following Figure 17 gives a general overview.

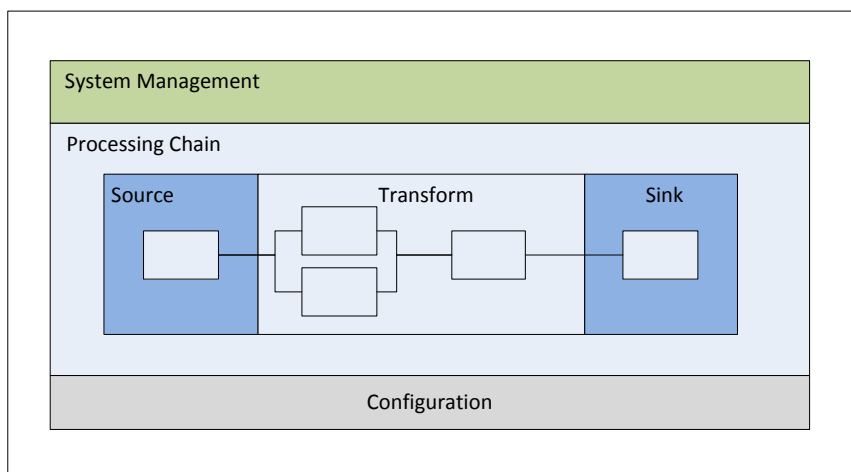


Figure 17 - Visual Attention Framework

The actual models are implemented within the processing chain and make use of the configuration. The elements that form a processing chain are called filters in the following. Source and sink filters form the data interfaces of the system. The transform filters are used to define the attention models.

The structure of graphs built from transform filters is basically derived from the type of the filters. According to this they can be categorized as depicted within the following Table 6.

Category	Description
Feature Extraction	Image / Video processing algorithms to generate feature maps
Feature Map Processing	Merging / Combining of feature maps
Feature Map Manipulation	Prioritization of regions within feature maps and stabilization of generated feature maps
Object Extraction	Extraction of geometrical representation of objects from feature maps
Object Rating	Rating / Filtering of extracted objects
Object Clustering	Merging / Combining of objects according to matching rules
Image Pre-/Post-Processing	Image manipulation

Table 6 - Visual Attention Filter Category

The feature extraction algorithms are used for the basic detection of visual saliency. This is done by analysing different image properties e.g. structure, colour and motion.

During feature map processing multiple extracted feature maps are combined into a single one which is used for further processing. This new map is also called saliency map in the following. It represents an overall saliency value from stimuli of different features.

Feature map manipulation algorithms are used for adapting feature maps to different conditions, like raising the importance of specific areas or creating more constant results over time by eliminating or down-rating outlier cues.

Object extraction produces geometrical representations of objects from feature maps, e.g. bounding boxes or polygons.

Object rating applies some weighting values to objects according to object properties.

Object clustering combines objects based on their properties.

Image pre-/post-processing can be applied to images as well as feature maps. The algorithms performed here are used to enhance image properties or improve feature maps.

For creating attention models only feature extraction and feature map processing are mandatory. All algorithms from other categories are useful additions to enhance the model.

3.1.2 Functionality

Operation Principle

The basic element of the visual attention system is feature extraction. The model uses different computer-vision algorithms in order to adapt for different types of video content. All feature extraction algorithms produce feature maps. The number of feature extractors is not limited. Normally two to four are used.

Those maps are represented by grayscale images whereas zero values indicated non-salient regions and non-zero values show salient regions with higher values represent stronger saliency. An example is given below within the section "Example feature extraction algorithm".

The idea of distractor detectors (described in D3.2/D3.3) was deferred as it was found to be too specific for this content and not widely applicable for much other content.

After the feature extraction, the generated feature maps are merged. The general saliency

map depends on the combination of different feature extraction algorithms. Doing this, the importance of regions found in different feature maps in the same position is increased while regions only present in one feature map are suppressed.

After creation of the saliency map there are two non-exclusive options for further processing: Feature Map Processing and Object Processing (Extraction, Rating and Clustering).

Using the first option the saliency map is manipulated to achieve better quality, incorporate additional information or improving the stability of the results (over time).

The second option can be applied with or without the first option. Here the saliency information is extracted from the saliency map and further processed using a geometrical representation.

Both options may provide some equivalent operations which are performed on the different data representation forms.

Depending on the application scenario for the attention model different data interfaces for input and output are usable. The current implementation is file based.

For the video input Microsoft DirectShow [14] is used.

For the output of the attention data a custom file format as shown in the following Table 7 is used. This file is then processed by the video encoder.

Item	Number of bits
Picture width in macro blocks	8
Picture height in macro blocks	8
Frames per second	8
Number of frames	16
For all frames repeat { For y=0 to picture height { For x=0 to picture width { Value of visual saliency [frame number][y][x] } } }	8

Table 7 - Visual Attention Output Format

Illustrative Example

Figure 18 gives an expression on how such saliency map looks like. This was taken from the content captured within this project. Both actors performing the fencing scene are detected and highlighted. Over time adaption is applied to get more consistent results.

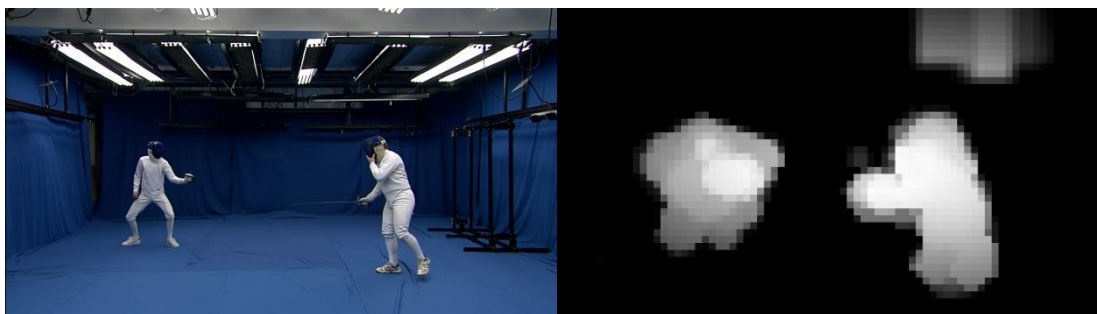


Figure 18 - Fencing Scene Attention Map

Example feature extraction algorithm

Since the feature extraction step is a key component of the model, one of the used algorithms is briefly presented in the following. It can be described as Global Motion Detection and Compensation, which can be used for 2D and with a small adaption for 3D content as well.

The proposed algorithm is used when camera motion is involved within a scene, e.g. pan or zoom. For successive video frames the global motion that occurs between them is estimated. This is done as follows:

- Prominent edges or corners in the two successive images are searched for.
- Found points are matched in both frames. The points in the first frames represent the initial position and the points in the second frame represent the new position.
- This yields an over-determined equation.
- By finding the best matching transformation matrix the homography between the images is calculated. In the current implementation LucasKanade algorithm [13] in combination with RANSAC is used.
- The camera movement is compensated by transforming one image using the found transformation matrix.

Using this, it is possible to detect moving objects within the scene, which produce saliency.

The same algorithm can also be applied for stereo or multi-view 3D content. Instead of successive frames, different views are used. The saliency detected here does not relate to motion, but instead produces foreground – background differences based on stereo depth properties.

It is possible to use temporal motion detection and foreground-background separation in parallel. The use of one variant of the algorithm does not exclude the other one.

The role of metadata

Since the large variety of content with different characteristics the developed approach also includes the usage of available metadata. Those play an important role for optimizing the attention model for specific content or content types.

Two different categories of metadata have unique impact on the design of the models. Those

categories are: technical metadata and descriptive metadata.

Technical metadata provide information about the production equipment, e.g. camera data like lenses, focal length, etc.

Descriptive metadata provide abstract information about the content itself, e.g. genre type.

Both types are of interest for creating the visual attention model. Information about camera movement for example can be used to select a proper feature extraction algorithm for motion detection. Genre type information can be utilized to select special feature extractors for specific scenes like face tracking for an interview situation.

The creation of attention models

As described before the system is highly flexible without a fixed structure. A scripting interface based on Lua scripting language [15] is used for creating the visual attention processing chain. It includes the script interpreter as well as the necessary library functions. The library uses the API provided by the management layer of the framework. This allows the definition of the static structure of the model as well as controlling the flow of execution. Complete processing chains can be defined in script files and executed without any additional effort for configuration or management.

Figure 19 shows a short example how an attention model can be created. In the example a short processing chain is created. It has a single feature extraction algorithm, called Spectral Residual. This filter is fed from the DirectShow source and the output feature map is displayed on screen.

The described approach was chosen because the scripting interface provides many options for interaction. Apart from the file based option shown in the example there are additional options like interactive command line or even interaction with other applications.

```
-- include convenience functions
require("visionInterface")

local inputFile = "C:/path/to/video.avi"

io.write("Initializing System...\n")
-- initialize system
vision.init()

io.write("Loading Filters...\n")
-- load File Source Filter
v1 = vision.loadFilterAndReport("DsSource")
-- load Screen Renderer Filter
v2 = vision.loadFilterAndReport("ScreenRenderer")
-- load Spectral Residual Filter
fSr = vision.loadFilterAndReport("SpectralResidual")
-- load Image Preprocessor Filter

io.write("Loading Filter Settings...\n")
-- load Settings File for Spectral Residual Filter
vision.loadFilterSettingsAndReport(fSr, "SpectralResidual.settings")

io.write("Connecting Filters...\n")
-- connect File Source and Spectral Residual
vision.connectFilterAndReport(v1, fSr)
--connect Spectral Residual and Screen Renderer
vision.connectFilterAndReport(fSr, v2)

io.write("Timer Setup...\n")
-- create predefined clock (25fps)
-- connect clock to renderer
vision.createAndSetClockAndReport(v2)

io.write("Opening Filters...\n")
-- open input file
vision.openFilterAndReport(v1, inputFile)
-- open screen renderer
vision.openFilterAndReport(v2, "")

io.write("Running Graph...\n")
-- execute graph and wait until completed
vision.startProcessingWaitUntilComplete()
io.write("Run complete!\n")

io.write("Cleanup timers...\n")
-- disconnect clock from renderer
vision.removeAndDeleteClockAndReport(v2)

io.write("Closing filters...\n")
-- close source file
vision.closeFilterAndReport(v1)
-- close renderer
vision.closeFilterAndReport(v2)

io.write("Unloading Filters...\n")
-- unload File Source Filter
vision.unloadFilterAndReport(v1)
-- unload Spectral Residual
vision.unloadFilterAndReport(fSr)
-- unload Screen Renderer Filter
vision.unloadFilterAndReport(v2)

io.write("Cleanup complete!\n")
-- cleanup system
vision.cleanup()

io.write("Exiting!\n")
```

Figure 19 - Attention Model Script

Dissemination

This work was presented at the IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB) 2011. It was submitted under the title: “AVISION – Audio and visual attention models applied to 2D and 3D audio-visual content”.

3.1.3 Evaluation

For evaluation purposes a subjective viewing experiment was carried out using an eye tracking system. The aim of this test was to compare the results of developed models with the human visual attention process. This does not primarily aim to provide a rating on the developed models, but demonstrate the influence of content type and why the proposed framework concept was chosen.

The viewing experiment was set up as follows: Nine different clips were presented to 18 viewers in 2D as well as 3D stereo. All clips were encoded equally without any prioritisation of specific image areas. They had different length, ranging from 10 to 30 seconds. The used sequences were kindly provided by the DIOMEDES project, Fraunhofer Institut, tpc AG and KUK Film.

A Tobii X60 eye tracker system was used to track the viewer’s gaze position while watching the clips.

For every user session the eye tracker was calibrated using a special calibration pattern. The calibration has been updated in a post processing step, as it was found that the accuracy was degrading towards the sides of the screen. This was done by adjusting the gaze positions, which were tracked for the calibration pattern to match the calibration pattern more exactly. This correction was finally applied to all results from the specific viewer.

A JVC 3D display with polarization technology was chosen to perform the viewing. Preliminary tests showed that the eye tracker works well with this display technology. In contrast to this, it was not possible to use shutter glasses of a Panasonic 3D Display because the glasses as well as the eye tracker use infrared light. This causes a loss of synchronisation of the shutter glasses.

A one-to-one comparison between modelled attention and human attention is not possible here because the model does not implement all steps relevant for the human attention (see also D3.2 or D3.3 chapter 2.3), e.g. steps like cognitive processing is very limited within the model. So this cannot be an exact correlation measurement. Additionally the method of gaze position tracking introduces additional inaccuracy of the measurement result. It also does not work equally well for different people. As a direct consequence, results from two viewers were rejected completely because of the enormous mismatch even for the calibration pattern. However these tests give a good impression on the models behaviour and how close it is to the human attention process.

The following tables show the results of the evaluation based on the gaze data of all viewers for the specific video tracks. They show the percentage of gaze positions that lie within the attention area marked by the model for 2D and 3D.

Clip Name	% within attention area
24h	37
Badminton	86
Budo	46
Fencing	26
Pablo	70
Lecture	80
Mercedes	26
Music	76
Fraunhofer Image Film	60

Table 8 - Visual Attention evaluation results for 2D

Clip Name	% within attention area
24h	37
Badminton	75
Budo	41
Pablo	55
Lecture	72
Mercedes	25
Music	63
Fraunhofer Image Film	55

Table 9 - Visual Attention evaluation results for 3D

As it can be seen the overall results are quite differing depending on the clips contents. Also the 3D performance is lower as for 2D. This is depending of different factors as quality of the stereo content and different viewing conditions. Especially the unusual viewing experience related to focus and convergence while viewing 3D content contributes to this.

There are different characteristics of the content, which influence the behaviour and the results of the models:

- Image structure, e.g. sharpness, contrasts
- Proportions of foreground and background
- Overall image resolution/size

Six of the nine videos scored high results: Badminton, Budo, Pablo, Lecture, Music and Fraunhofer Image Film. Low results of the other clips can be related to different causes: The 24h clip has two main properties that degrade results of the models. There are many shots within the clips that cause the model to readapt for new scenes. During this process results are not exact. Also most algorithms rely on good contours within the image but the scenes in the 24h clip suffer from low sharpness because of high percentage of motion blur and drizzling rain in the video. Mercedes contains many close shots with high zoom whereas the model tries to find objects within the scene that are significantly smaller than the overall action area.

Therefore it is not able to find objects which fill the whole scene. This does not match the targeted use case of the model to detect contextual most important image regions because the whole image contains only one object. This behaviour is exactly as expected beforehand. For the fencing scene the model primarily detects the actors mainly because of the scene size to object size relation, given within the shot. Gaze points of the viewer however focus mainly on the épée. This can be corrected to apply additional parameters considering this very special case.

As already mentioned, this evaluation is not strictly aiming on giving a rating to the models, but more to show the influence of the types of content and why the proposed modular framework with adaptable parameter sets was developed. This also includes the use of the described metadata as well as the open interfaces to extend the models as required by different use cases.

3.2 Audio Attention Modelling

In this section, the findings from the subjective tests investigating the audio attention are firstly described in detail, leading to principles to enable efficient transmission of auditory information. Then a new subjective test is introduced, in which the findings are validated.

3.2.1 Introduction and subjective test procedure

An auditory event always contains one or more sound sources or objects. In the presence of multiple objects, depending on the context, the listener often pays more attention to specific objects which he/she believes contain more important auditory information. Audio attention modelling attempts to mimic this mechanism of human auditory system. In other words, audio attention modelling is a process to determine which attribute of the sound objects would contain the more important information and thus would need to draw more attention, as a human listener would do. More detailed background information was provided in D3.2 and D3.3.

A subjective test was designed and conducted to investigate the relative difference in the perceived salience of changes in acoustical properties. Multiple audio objects were presented together to the listeners through various combinations of processing. The variations were introduced in terms of amplitude, timbre (by means of bandwidth limitations), and direction (using stereo loudspeaker setup). The subjects were asked to rate the overall perceived audio quality compared to the reference. Table 10 briefly summarises the attributes varied in combinations for the listening test.

Content	Loudness	Quality (from bandwidth variation)	Direction
Speech with Music	"Normal" / "Loud" (2x amplitude than the other)	"Low" / "Medium" / "High" (reference)	Centre-wide / 60° Right

Table 10 - Acoustic attributes varied in combination for the subjective listening test

3.2.2 Results and analysis

Figure 20 to Figure 25 are drawn again as in D3.2 and D3.3 to show the results and the implications in more detail. Firstly, Figure 20 implies that when all the other conditions remain the same, the quality degradation in speech would be perceived more salient than music (as indicated by the circle on the plot). Although this tendency does not seem statistically significant, it is supported further by comparing Figure 21 and Figure 22. It is clearly seen that the perceived quality of all the quality degradations was graded lower in general when music was louder, compared to when speech was louder. This means that the subjects perceived the stimuli as poorer when speech was quieter. Comparison of Figure 24 and Figure 25 shows the same tendency, where louder music in the stimuli did not receive as high grade as louder speech in the stimuli. This seems due to the fact that speech tends to convey clearer information than music and thus might draw more attention.

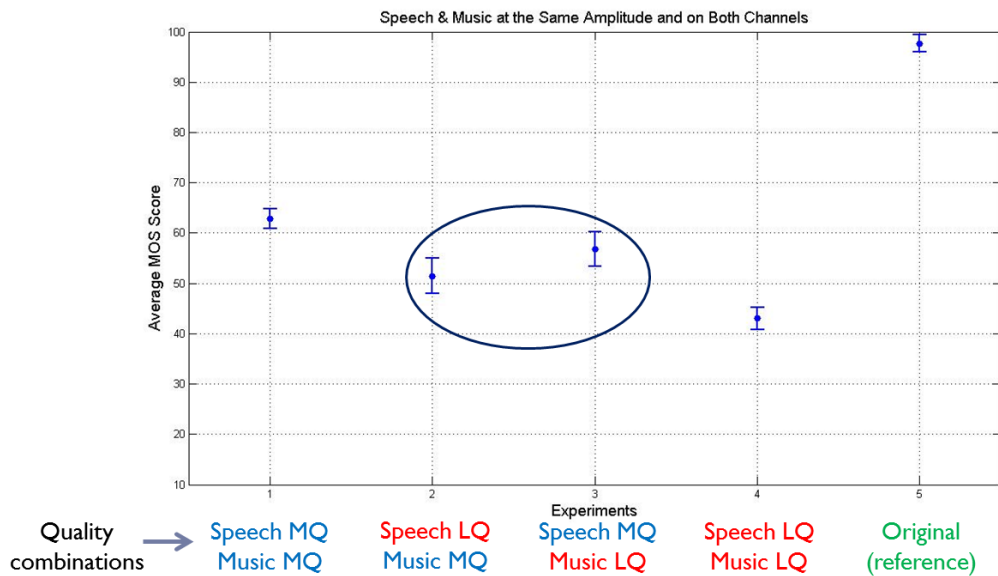


Figure 20 - Mean Opinion Scores for combinations of quality degradations in stimuli - music and speech at normal loudness and wide at the centre. The quality degradation in speech seems to be perceived more salient than music (as indicated by the circle on the plot)

Another finding from Figure 21 and Figure 22 is that the louder content was dominant in the quality degradation attention. Figure 21 shows that when speech was louder in the stimuli, the quality degradation in speech was more salient. Figure 22 shows the opposite – when music was louder, the quality degradation in music was more salient.

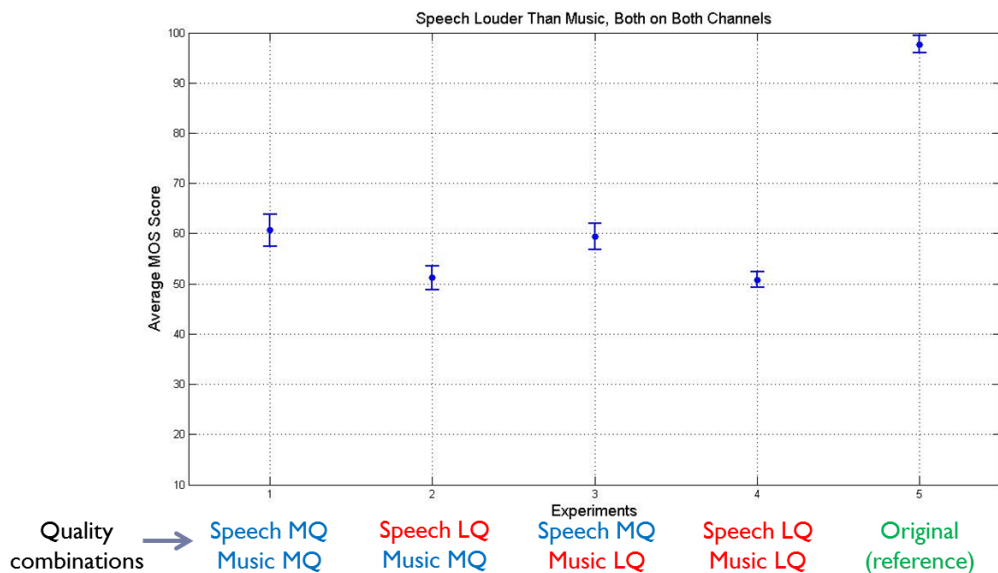


Figure 21 - Mean Opinion Scores for combinations of quality degradations in stimuli - speech louder than music, both wide at the centre.

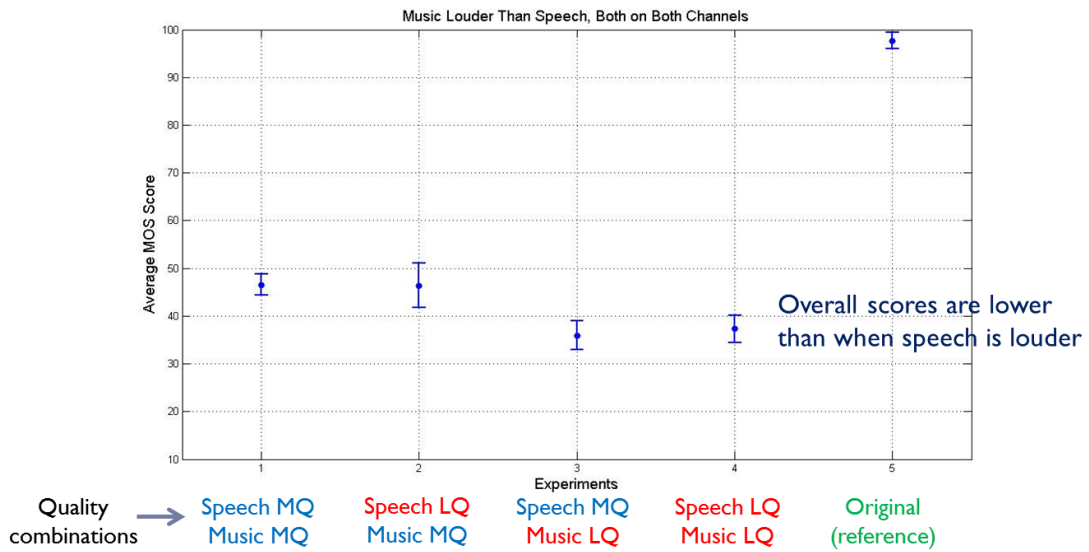


Figure 22 - Mean Opinion Scores for combinations of quality degradations in stimuli - music louder than speech, both wide at the centre.

From Figure 23 to Figure 25, it can additionally be inferred that the direction change of content has resulted in notable decline of the MOS scores, and that amplitude could compensate the difference in perception (Figure 24).

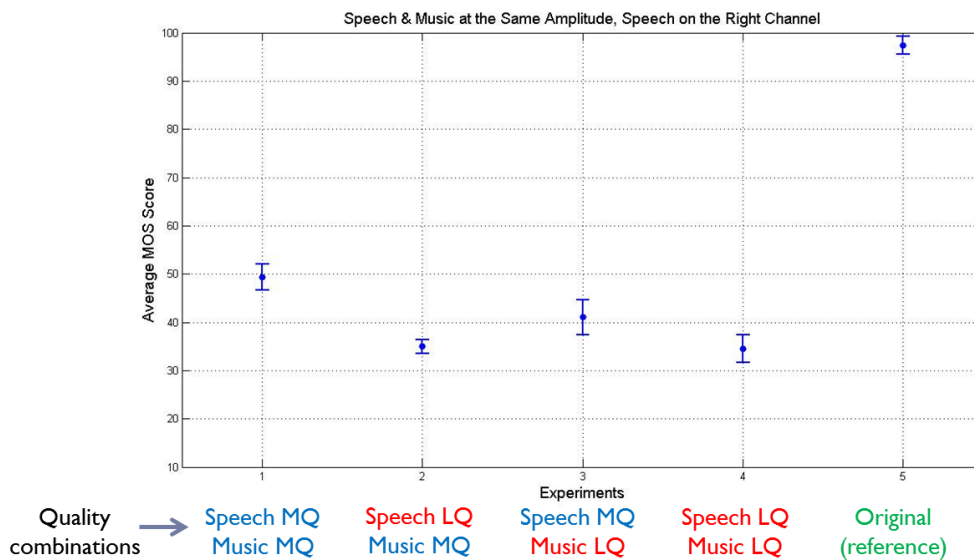


Figure 23 - Mean Opinion Scores for combinations of quality degradations in stimuli - both at the same amplitude, speech from the right channel with music wide at the centre.

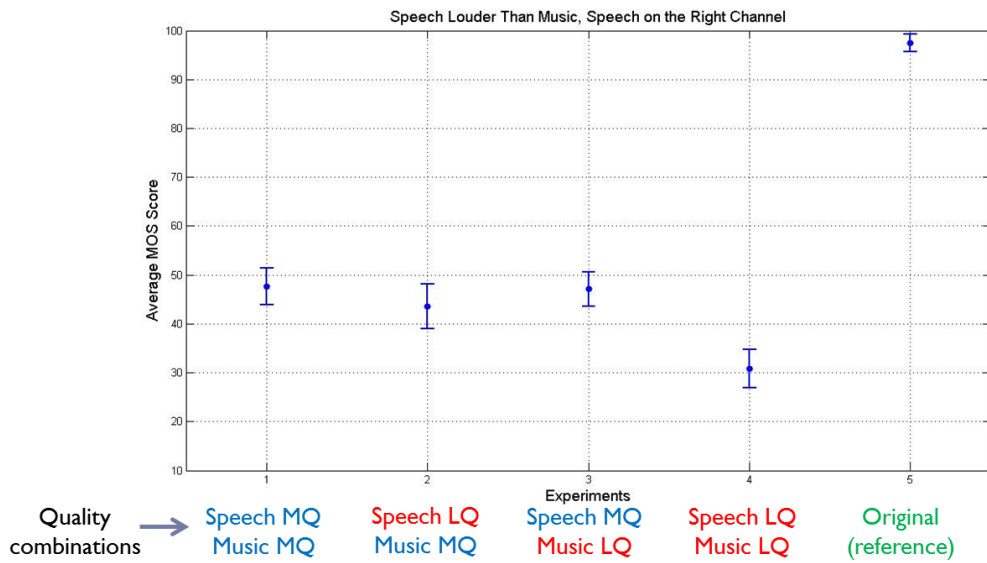


Figure 24 - Mean Opinion Scores for combinations of quality degradations in stimuli - speech louder than music, speech from the right channel with music wide at the centre

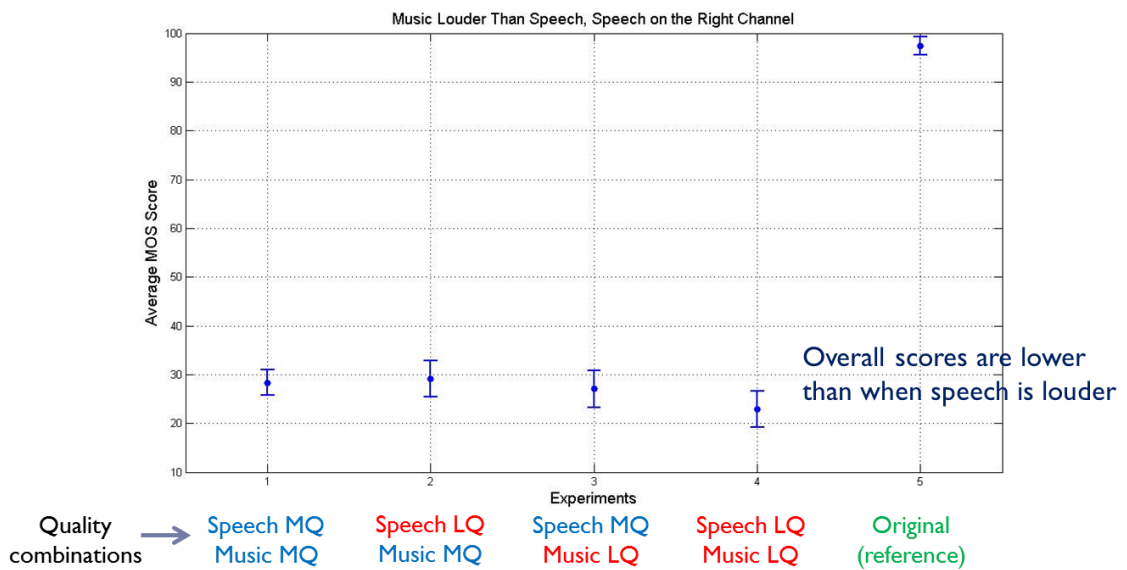


Figure 25 - Mean Opinion Scores for combinations of quality degradations in stimuli - music louder than speech, speech from the right channel with music wide at the centre

The findings can be summarised as follows: Firstly, degradation or change in more informative content is more salient. Secondly, amplitude changes are found more salient than timbral changes from bandwidth variation. Thirdly, direction changes are found more salient than timbral changes, but can be compensated to a certain amount by adjusting the loudness

3.2.3 Principles for application

Based on the findings, some principles can be established for audio attention modelling.

- Scene information from descriptive metadata

The scene information needs to be obtained first to establish context-based encoding strategy. For example, the names of the sound objects in the captured scene, their positions and amplitudes can be firstly retrieved.

- Classification of objects

The current findings suggest that the objects be classified in terms of 1) how informative they are – for example speaking person or a wind noise, 2) loudness to the listener, and 3) width or directivity. The hierarchy for bandwidth saving will be determined based on this classification.

- Hierarchy for bandwidth saving

Firstly, it is worth noting that whether the bandwidth limitation would be effective on the audio needs to be known in a given scenario where audio-visual contents are encoded for transmission. In other words, since the amount of audio data in general is smaller than that of video, it needs to be known whether the amount of saved bandwidth from limiting audio quality would be meaningful in terms of overall bandwidth efficiency. Once it is determined that the audio bandwidth needs to be limited further, the hierarchy can be made such that once the objects are classified according to the informative importance, their amplitudes will be checked before the direction, and the quality of louder objects will be preserved with higher priority.

3.2.4 Validation experiment

In order to find out the validity of the above findings regarding the audio bandwidth limiting hierarchy, another subjective test was conducted in which the degradation salience of various sound objects could be compared.

An auditory scene was synthesised in binaural stereo listening setup, consisting of 6 sound objects – drums, bass guitar, acoustic guitar and strings which played an arranged music piece, and a female voice and a male voice whose contents were not related to each other. The sound objects were arranged to be reproduced in various signal amplitudes and in various directions. The drums were placed at the centre with an average amplitude (RMS power) of -43dB. The bass guitar was placed at the centre with an average amplitude of -25dB. The acoustic guitar was placed at the left side (reproduced at the left channel only) with an average amplitude of -25dB. The strings were placed at the right side (at the right channel only) at -37dB. The female voice was panned approximately 45 degrees to the left and was reproduced at an average amplitude of -25dB. Lastly, the male voice was placed approximately 72 degrees to the right at an amplitude of -46dB. The initial bitrate for reference was 2822.4kbps (44.1kHz sampling frequency, 32 bits, 2 channels). For the experiment, the bitrate of each object was reduced to 352.8kbps (11.025kHz, 16 bits, 2 channels) in turn within the mixture of all objects. Table 11 summarises the objects used for the auditory scene creation and the attributes controlled for the subjective test.

Object	Average Amplitude (RMS power)	Panning (direction)	Bitrate
Drums	-43dB	Centre	2822.4kbps → 352.8kbps
Bass guitar	-25dB	Centre	2822.4kbps → 352.8kbps
Acoustic guitar	-25dB	90° Left	2822.4kbps → 352.8kbps
Strings	-37dB	90° Right	2822.4kbps → 352.8kbps
Female voice	-25dB	45° Left	2822.4kbps → 352.8kbps
Male voice	-46dB	72° Right	2822.4kbps → 352.8kbps

Table 11 - Sound objects and attributes controlled for the validation experiment

Six stimuli were created in this way. The bitrate of one object was reduced per a stimulus. They were presented to 8 listeners, who were asked to compare them to the reference (with no bitrate degradation). MUSHRA [5] was used as the test method, in which the listeners graded the perceived audio quality of each stimulus, compared to the reference, using a scale of scores from 0 to 100. The reference was provided hidden as one of the stimuli. Figure 26 shows the collected results. The crossed marks indicate the average scores, and the bars indicate the standard deviations.

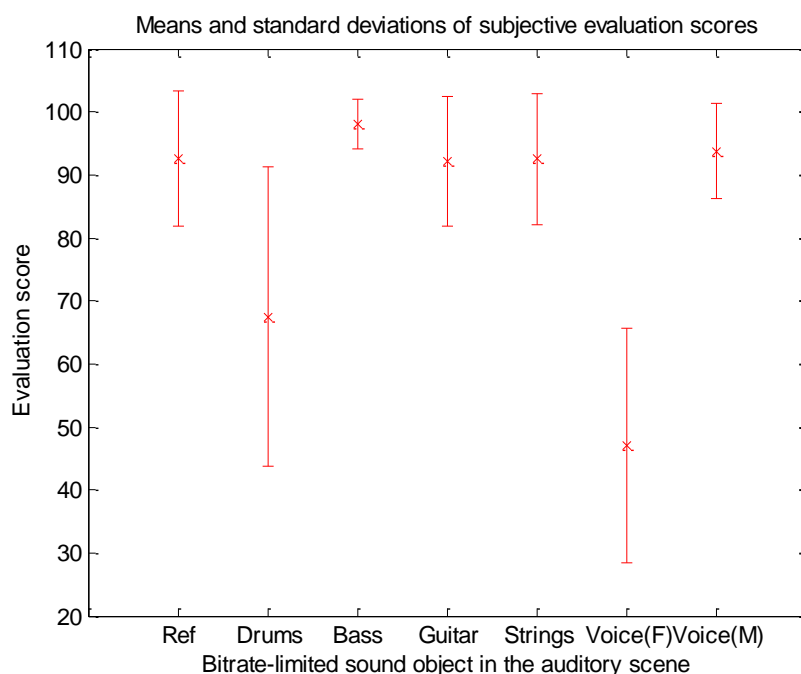


Figure 26 - Means and standard deviations of subjective evaluation scores of the provided auditory scene where various objects were individually degraded in terms of bitrate.

It is seen that the degradation in the female voice, despite the same amplitude as the bass and the guitar, was the most noticeable. The next noticeable degradation is found in the drums sound, although its amplitude was smaller than the other musical instruments. The degradations in the other objects are not noticeable compared to the reference. This validates the findings from the previous experiment and provides further implications as follows:

- a) The degradation of the more informative objects is more salient.
- b) Whether an object is more “informative” or not could be inferred from the frequency range it occupies – the female voice and the drum sound covers wider frequency area than the other sound objects.

Within the group of objects with similar significance of information, the hierarchy can be determined from the amplitude.

3.2.5 Summary and conclusion

Investigations have been made into the perception of quality degradations of various acoustical attributes of sound objects, towards attention modelling for bandwidth-efficient audio coding and transmission. Two sets of subjective tests have confirmed that a hierarchy of audio objects could be made in a captured auditory scene based on the scene description, in terms of the salience of their quality degradation. It has been found that the more informative objects with wider frequency spectrum would draw more attention, and that amongst the objects covering similar frequency ranges the ones with larger amplitudes would draw more attention. The findings suggest the audio attention modelling processes as described in the previous subsections and summarised again in Figure 27.

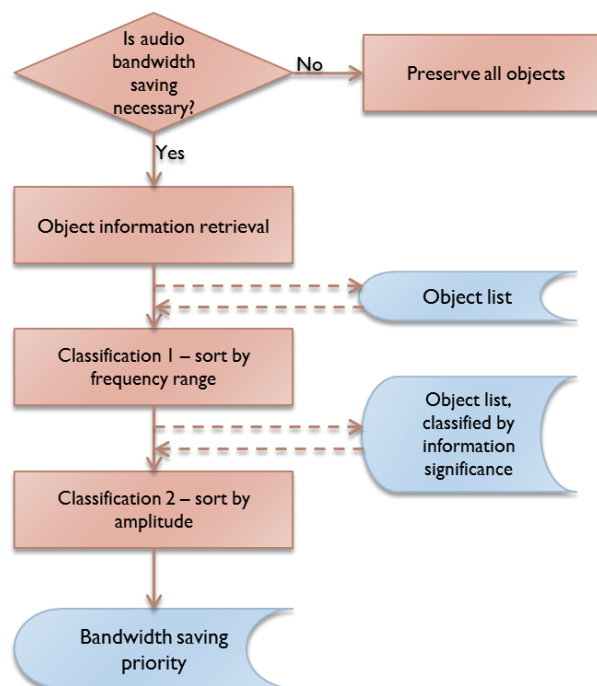


Figure 27 - Suggested processes for audio attention modelling

3.2.6 Loudness Analysis Model

- **Introduction**

In D3.2 the loudness of an audio signal was identified as one of the most important factors affecting the attention of a listener. It was shown by subjective testing, that a louder signal tends to dominate in terms of quality, while other signals played at the same time go almost unnoticed.

Another aspect is the easy measurability of loudness values in contrast to other properties which interact with the human audio perceptual system e.g. timbre, location and isolation. This predestines a loudness measurement to be the core of the practical implementation of a grouping process, which classifies audio objects based on the impact they arouse at a listeners perception. Such a process could be used to improve the audio coding compression while maintaining the QoE, as seen in Figure 28.

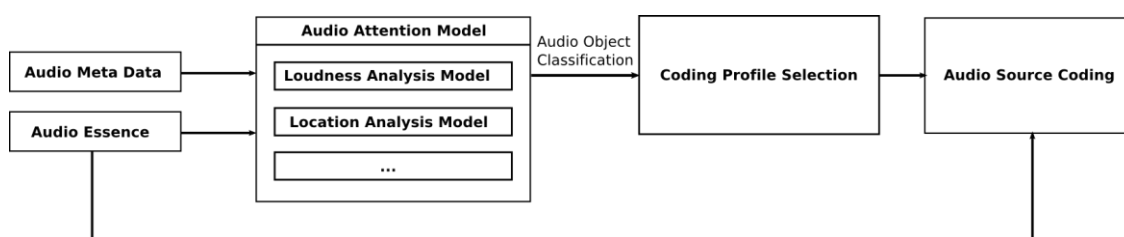


Figure 28 - Integration of the Loudness Analysis Model

Due to the fact that there was no approach of automatically measuring the loudness of an object based audio scene available by now, an analysis model had to be developed for the DIOMEDES specific needs. This model should be capable of analysing the PCM-based audio information and combine it with the metadata of each audio object. Therefore the loudness analysis model simulates the loudness of a whole audio scene, comparable to the listener's impression in front of a WFS loudspeaker array. As independence from a renderer is one of the basic ideas behind object based audio formats, it was tried to model the approach of loudness analysis as generic and physically authentic as possible. For that reason a large set of parameters is provided to enable the alignment of the loudness calculation to the features of the proposed rendering engine.

The term “loudness” is defined as a psychoacoustic perception quantity and describes the perceived intensity of the stimulus sound pressure level. It varies significantly from a measurable physical stimulus such as the sound pressure level itself. Within the DIOMEDES loudness analysis model two independent ways of loudness measurement are provided, which adopt generally different approaches:

- Loudness modelling according to Zwicker (DIN 45631 and ISO 532 B)
- K-weighted RMS loudness modelling referring to the EBU recommendation R128

The Zwicker model calculates the loudness by simulating the physical processes of the human auditory system in multiple stages. Thus, it is a real psychoacoustic approach of predicting the loudness perception caused by a certain stimulus. As depicted in Figure 29, and due to the reason that the sound pressure level isn't usually measured at the eardrum, initially the transfer function of the head and the tympanum have to be taken into consideration. This is practically implemented by an analogue filtering process. In a second stage the frequency dependence of the human hearing is simulated by calculating the excitation level I_g for each of the first 24 critical frequency bands. This leads to a neuronal pattern, which represents the excitation of the basilar membrane. From this, the specific loudness N' of every frequency group can be determined. The overall loudness finally results from the integration of all bands.[18]

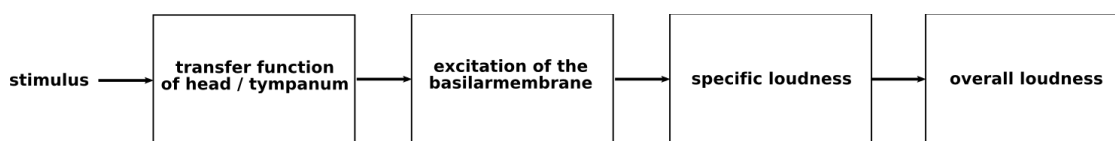


Figure 29 - Functional model of loudness measurement according to Zwicker

For the loudness calculation of time-variant sound sources, as they are designated to be used in DIOMEDES, the basic Zwicker model has to be expanded by the measurement method specified in DIN45631/A1. In so doing, the duration of a stimulus is additionally incorporated into the calculation rule. This is important for the reason that the human auditory system integrates during its loudness assessment until the perception stabilizes at around 200ms.[16]

K-weighted RMS is deduced from the EBU recommendation R128 and describes an entirely different method of loudness measurement, which doesn't date from psychoacoustic modelling. It was basically designed as an easy and comprehensible way of measuring the loudness of broadcast related audio signals and corresponds to the listeners' perception in certain situations. On behalf of simplicity it is not intended to be an exact model of the human hearing. The K-weighted RMS model can be equally divided into multiple consecutive processing steps, as shown in Figure 30.

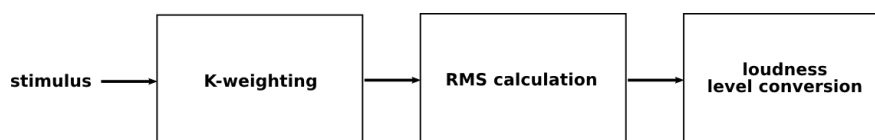


Figure 30 - Loudness measurement according to EBU R128

First, the audio signal is weighted in the frequency domain for the purpose of considering the varying sensitivity of the human hearing at different bands of frequencies. In a second step the RMS (root mean square) value is calculated to reproduce the integration aspect and finally, within a third step, the result presented as a level value.[17]

Both models were implemented due to their importance within different scopes of application. In the context of this investigation it could not be determined which way of loudness measurement works best, in order to enhance the audio coding process. This mainly refers to the current state of implementation of the DIOMEDES audio attention model as well as the high expenditure of time needed for such a sophisticated study. All the more, further testing of this issue would be highly desirable within the scope of future work.

- **Loudness Analysis Model in DIOMEDES**

The DIOMEDES loudness analysis model consists of six different types of sub modules, the input module, the scene model, the config module, the interface module, the binaural summation and the loudness analysis module, as can be seen in Figure 31.

The input module is responsible for parsing and structuring the metadata of the audio source objects, the room, as well as the listener in order to process them further within the scene module. There is a source struct available for every single audio object within a certain scene, while the room and the listener struct are existing only once. In the source struct, information such as the position of the source within a virtual room, the gain factor and the directivity are stored. In order to correctly process time-variant audio scenes, the metadata structs are designed in an adaptive way. The information is read out from the metadata stream every 1024 PCM samples and thereby renewed block wise. In DIOMEDES just a very simple set of metadata is provided to describe the audio scene parameters and for that reason only a

specific choice of metadata is used for the analysis process. E.g. neither the room struct nor the listener struct is used. The model however, was developed generic in order to be easily adaptable to more powerful audio description languages and other types of renderers. The following table shows the subset of description parameters actually used for DIOMEDES:

Struct	Parameter	Description
source	.name	the filename of the audio source
	.pcm	the PCM data of the source
	.fs	the sampling frequency the material has been recorded with
	.FrameNum	the frame number of the actual sample block
	.xPos .yPos .zPos	cartesian coordinates
	.Gain	amplification factor between 0 and 1
	.start	time delay from the start of the scene to the playback start of the referring audio object
	.stop	time delay from the start of the scene to the playback end of the referring audio object

Table 12 - Description parameters used in the DIOMEDES audio metadata

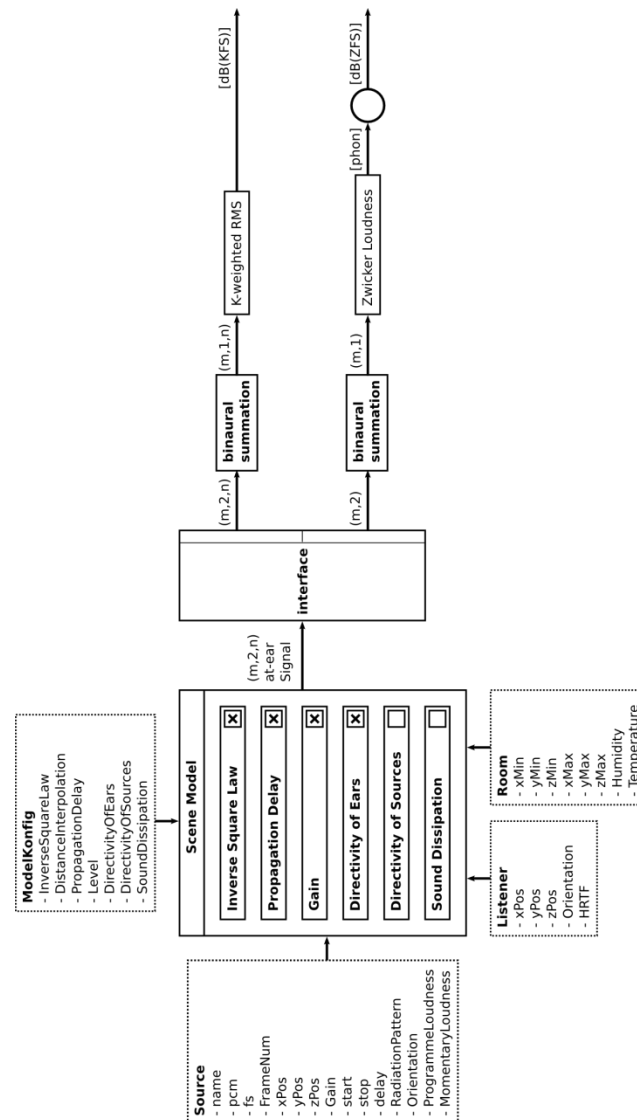


Figure 31 - Schematic representation of the Loudness Analysis Model

The scene model is the centrepiece of IRT's loudness analysis and contains the calculation procedures capable of shaping the PCM data of the audio objects according to their given metadata into an audible binaural audio scene. Therefore several physical related processes could be applied to the audio sources:

Process	Function
Inverse Square Law	For the model no realistic sound field is assumed, like it would establish in a room with natural acoustics. This is mainly for the reason that no room information is provided for the renderer within DIOMEDES. Later, a simple shoe box room model could be a reasonable replacement for this process. In the suppositional anechoic conditions however, the inverse-square law is a valid phenomenon and describes the decrease of sound intensity over the distance. This process calculates the resultant attenuation of the signal.
Propagation Delay	Calculates the propagation delay between the source and the listener position. It is also used to model the Doppler effect for moving sources and thereby affecting the pitch of the audio material.
Gain	Calculates the amplification of a source based on the gain factor.
Directivity of Ears	Calculates the directivity of the human ear by applying a head related transfer function to the PCM data. This process creates a binaural two-channel signal, which can be processed further.
Directivity of Sources	Calculates directivity to the audio source objects, by employing algorithmic models referring to the Huygens principle, in order to approximate the directivity pattern of real audio sources. (Is not used in DIOMEDES)
Sound Dissipation	Calculates the treble attenuation caused by the air in dependence of the humidity and temperature of the room, as well as the distance of the sound source. (Is not used in DIOMEDES)

Table 13 - Physical based audio processes implemented into the scene model

The config module is used to configure the scene model by turning on/off the physical calculation procedures. This feature can be used to adapt the loudness analysis model to different types of audio description languages and rendering engines.

The interface module is used to convert the binaural signal matrices for every audio object, given out by the scene model. This is an essential process, since the Zwicker and the K-weighted RMS determination methods need different audio input formats.

The binaural summation takes account of the Zwicker model is exclusively defined for frontally incident sound sources, thus for sound events which cause the same auditory perception at the left and the right ear of a listener. For an object based audio scene however, also sound sources from other angles of incidence had to be considered. Therefore the loudness measurement method is enhanced by an approach of Silvon and Ellermeier [19][20], which describes the psycho-acoustically correct approximation of binaural loudness to a single loudness value. In addition to that, the binaural summation module is also responsible of remodelling the binaural signals into channel-based information, needed as an input for the K-weighted RMS system.

The loudness analysis model finally implements the two algorithms of loudness measurement provided. The results are given out as logarithmic level values in dB(KFS) or dB(ZFS) according to the related measurement process. For this purpose, the output value of the Zwicker model first had to be normalized to a reference sound pressure, then converted to the phone scale and finally expressed as a level.

- **Implementation**

A Matlab programme, implementing the specified model, had been set up as a proof of concept. As such, it is capable of analysing file based audio objects and parsing IDMT's dedicated metadata format for DIOMEDES. As seen in Figure 32, an optional user interface was created for the ease of configuration and illustration of results.

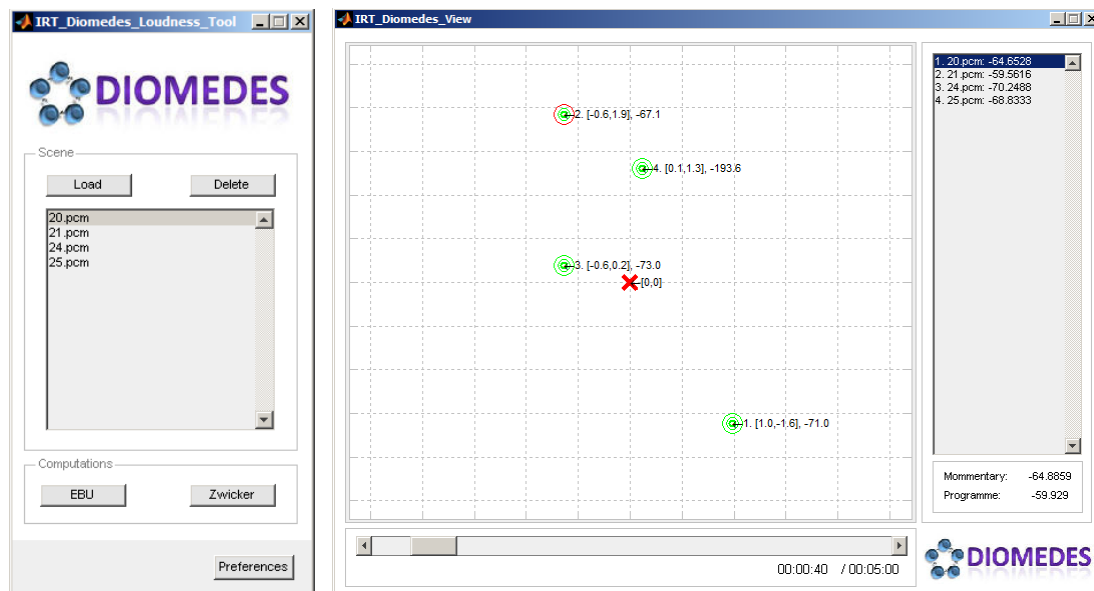


Figure 32 - User interface of the loudness analysis Matlab programme

It shows the position and loudness level of every audio object and furthermore sorts and classifies them in dependency on this attribute. A slider can be used to shift the time and watch the scene changing within a certain interval. Besides this, the GUI also displays the overall loudness value of the measured audio scene.

The source code makes a number of interfaces available, which could be easily accessed by a superordinate audio attention wrapper class. Beyond that, the code is transferable to a more efficient programming language in order to enhance the real-time capabilities.

- **Verification measurement**

For validation of the operability of the physical model and implementation, the results of a real loudness measurement setup in the anechoic chamber had been compared with the output of the model.

Therefore a moving loudspeaker arrangement was mounted on a rail in front of an artificial head. As shown in Figure 33, the orientation of the driving direction had been orthogonal to the 0°-azimuth axis of the head. The loudspeaker was driven by a daring deed connected through a cable control system, which allowed detailed regulation of the acceleration process. White noise served as a test signal and was played back while the loudspeaker was moving. During the measurement, the microphone signals from the artificial head were recorded together with camera shots of the scene.

For evaluation purpose the movie was analysed with editing software in order to extract the relevant position information of the loudspeaker in dependence to the time. With the time – position information on the other hand, a respective audio scene could be simulated by the loudness analysis model. Finally the deviation of the 1/3-octave-band spectra of the measurement and the simulation was calculated. This measured value can serve as an

indicator of coherence between the model and the loudness conditions within a real sound field. The results can be obtained from Figure 34 to Figure 36.

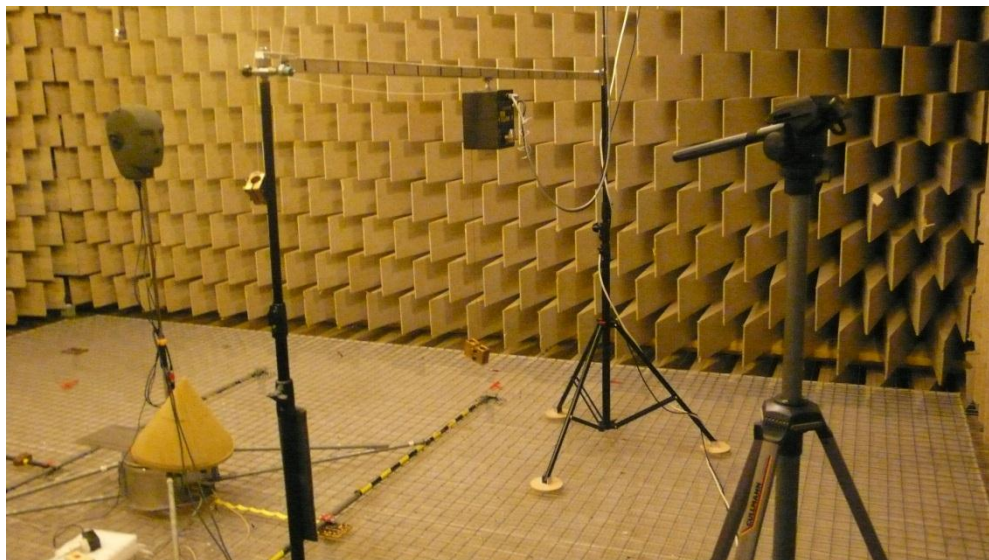


Figure 33 - Measurement setup for determining the loudness of a time variant object based audio scene in the anechoic chamber

In a second experiment an additional static loudspeaker was placed sideways of the artificial head to simulate a second sound object. The result is shown in Figure 37.

It can be seen, that the deviation with a calculation procedure turned off, tends to be higher than with all processes running. For the simulation the directivity of sources, the directivity of the ears, as well as the inverse square law were considered. The Doppler Effect and the sound dissipation however had been of no consequence, because of the short distance between the loudspeaker and the measuring point. It is also evident, that with a second sound source the deviation is around 1dB higher. With a further increasing amount of sources, the deviation converges exponentially to a certain value.

The model in general may be regarded as sufficiently precise for its intended purpose. The average deviation shown in Figure 34 is just as low as 1,2dB and therewith very close to the human perception threshold.

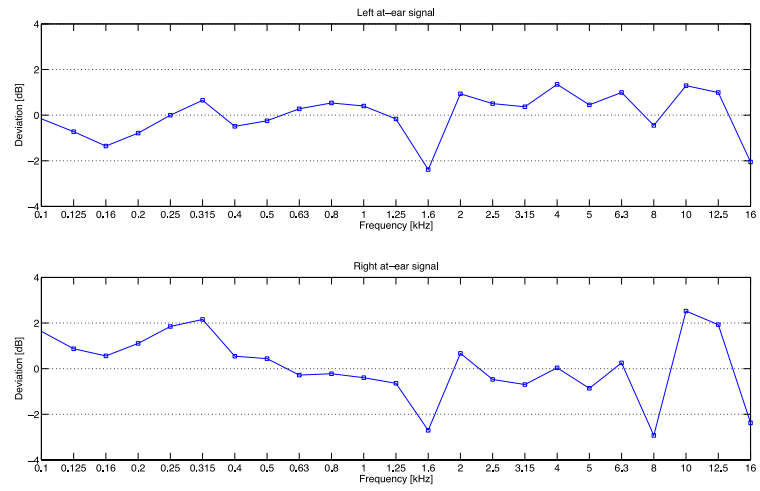


Figure 34 - Deviation with all calculation processes running

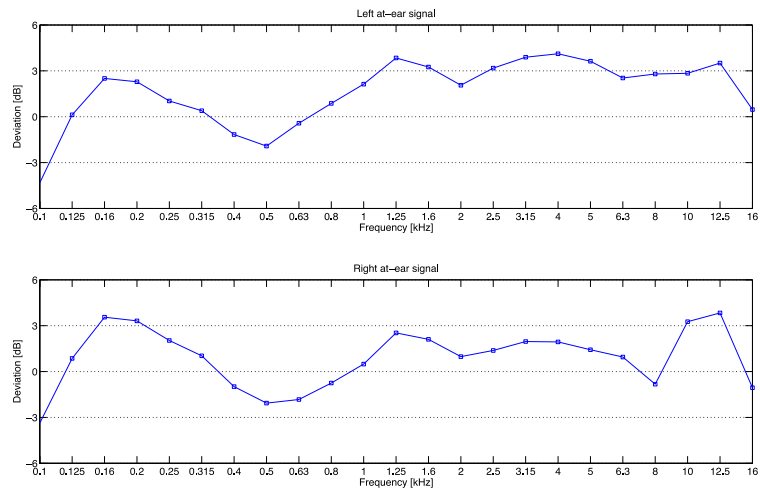


Figure 35 - Deviation with directivity of sources turned off

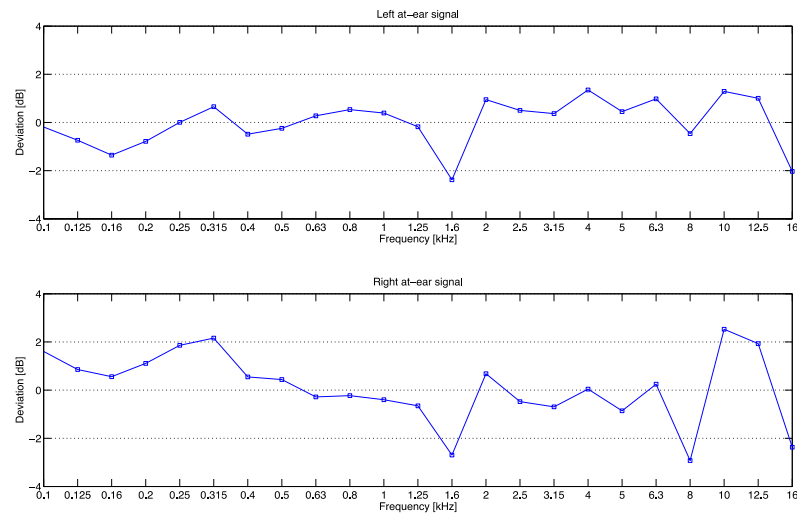


Figure 36 - Deviation with sound dissipation turned off

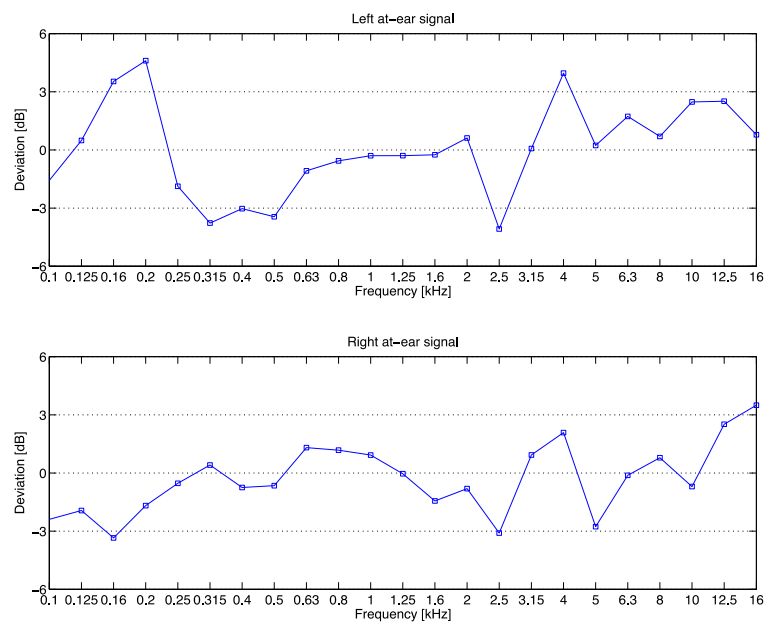


Figure 37 - Deviation with two source objects

- **Conclusion and Future Work**

The created model seems to be promising in terms of using the gained loudness information to improve the audio coding process of object based scenes. The applicability in practice however finally depends on how much the operation principle of the loudness analysis model and renderer coincide. Therefore the adaptability of the model is one of the most important factors and should be further improved.

Important milestones of a future work could be the implementation of a basic room model and the optimization towards a better real-time capability.

4 CONCLUSIONS

This document has described the development of the Quality of Experience model as well as the audio and visual attention modelling and is an extension to deliverables 3.2 and 3.3.

The QoE model presented aims at identifying the relevant KPI for the DIOMEDES application scenario and computing the Quality of Experience on the receiver end. The relevant algorithms were presented and also the evaluation methodologies and results were shown.

The attention models are based on human attention concepts. The implementation concepts are described and also the subjective evaluation and its results. The aim of the attention models is to identify the most important and most relevant elements of the content. This information can be used for region based coding, including the use of different quality layers.

This deliverable concludes the tasks 3.2 “QoE Modelling for Compression” and 3.3 “3D Visual and Audio Attention Modelling”.

REFERENCES

- [1] C.T.E.R. Hewage, S.T. Worrall, S. Dogan, S. Villette, and A.M. Kondoz, "Quality Evaluation of Color Plus Depth Map-Based Stereoscopic Video," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, Apr. 2009.
- [2] Yasakethu, S.L.P.; De Silva, D.V.S.X.; Fernando, W.A.C.; Kondoz, A.; "Predicting sensation of depth in 3D video," *Electronics Letters*, vol.: 46, no. 12, pp. 837 – 839, 2010.
- [3] S.L.P. Yasakethu, S.T. Worrall, D.V.S.X. De Silva, W.A.C. Fernando, and A.M. Kondoz, "A Compound Depth And Image Quality Metric For Measuring The Effects Of Packet Loss On 3D Video", *Int. Conf. Digital Signal Processing (DSP 2011)*, Corfu, Greece, July 2011.
- [4] ITU-T Recommendation BT.500–11, "Methodology for the Subjective Assessment of the Quality of Television Pictures," Jun. 2002.
- [5] Recommendation ITU-R BS.1116-1, "Methods for the subjective assessment of small impairments in audio systems including Multichannel Sound Systems", International Telecommunication Union, Radiocommunication Sector, October 1997
- [6] Recommendation ITU-R BS.1534-1, "Method for the subjective assessment of intermediate quality level of coding systems", International Telecommunication Union, Radiocommunication Assembly, January 2003
- [7] Recommendation ITU-T P.910, "Subjective video quality assessment methods for multimedia applications". International Telecommunication Union, Telecommunication standardization sector, April 2008
- [8] Bech, Søren; Zacharov, Nick; "Perceptual audio evaluation. Theory, method and application", Chichester, England, Hoboken, NJ: John Wiley & Sons, 2006
- [9] Report ITU-T BT.1082-1, "Studies toward the unification of picture assessment methodology", 1990
- [10] Komiyama, Setsu; "Subjective Evaluation of Angular Displacement between Picture and Sound Directions for HDTV Sound Systems", *Journal Audio Eng. Soc.*, Vol. 37, No. 4, April 1989
- [11] De Bruijn, Werner P. J.; Boone, Marinus M.; "Subjective experiments on the effects of combining spatialized audio and 2D video projection in audio-visual systems", 112th AES Convention, Paper 5582, Munich, Germany, May 2002
- [12] Recommendation ITU-R BS.1284-1, "General methods for the subjective assessment of sound quality". International Telecommunication Union, Radiocommunication Assembly, December 2003
- [13] Lucas, B.; and Kanade, T; "An Iterative Image Registration Technique with an Application to Stereo Vision", *Proc. of 7th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 674-679
- [14] <http://msdn.microsoft.com/en-us/library/ms783323.aspx> (visited on 2011-09-28)
- [15] <http://www.lua.org/> (visited on 2011-09-28)
- [16] EDIN45631/A1; "Berechnung des Lautstärkepegels und der Lautheit aus dem Geräusch- spektrum — Verfahren nach E. Zwicker — Änderung 1: Berechnung der Lautheit zeitvarianter Geräusche - Normentwurf. Beuth Verlag, Berlin, 2007
- [17] Recommendation EBU Technical R 128, "Loudness normalisation and permitted

maximum level of audio signals”, European Broadcasting Union, Geneva, 2010

- [18] Eberhard Zwicker, H. F.; Schroeder, M. (Hrsg.); “Psychacoustics”. Springer-Verlag, Berlin, 1990
- [19] Ville Pekka Sivonen, W. E.; “Directional loudness in an anechoic soundfield, head related transfer functions, and binaural summation.”, In: Journal of Acoustical Society of America 119, S. 2965–2980, 2006
- [20] Wolfgang Ellermeier, V. P. S.; “Directional dependence of binaural loudness.” In: Fortschritte der Akustik – Daga ’06, 2006

APPENDIX A: GLOSSARY OF ABBREVIATIONS

A	
AAC-LC	Advanced Audio Coding Low Complexity
ACR	Absolute Category Rating
D	
dB(KFS)	DB k-weighted in reference to full scale
dB(ZFS)	DB processed according to Zwicker in reference to full scale
DDM	Disparity Distortion Metric
DIN	Deutsches Institut für Normung
DP	Depth Perception
DSCQS	Double Stimulus Continuous Quality Scale
E	
EBU	European Broadcasting Union
G	
GUI	Graphical User Interface
H	
HE-AAC	High Efficiency Advanced Audio Coding
I	
IQ	Image Quality
ITU	International Telecommunication Union
K	
KPI	Key Performance Indicator
M	
MPEG	Motion Pictures Experts Group
MOS	Mean Opinion Score
P	
PCM	Pulse-Code-Modulation
Q	
QoE	Quality of Experience
QP	Quantisation Parameter
R	
RMS	Root Mean Square
RoI	Region of Interest
V	
VQM	Video Quality Model
W	
WAV	Waveform Audio File Format
WFS	Wave Field Synthesis
Z	
ZMA	Z-direction (depth direction) Motion Activity