 DOI: 10.1109/QOMEX.2010.5518100

AUDIOVISUAL QUALITY INTEGRATION: COMPARISON OF HUMAN-HUMAN AND HUMAN-MACHINE INTERACTION SCENARIOS OF DIFFERENT INTERACTIVITY

Sebastian Möller[1], Benjamin Belmudez[1], Marie-Neige Garcia[2], Christine Kühnel[1], Alexander Raake[2], Benjamin Weiss[1]

[1] Quality and Usability Lab, Deutsche Telekom Laboratories, TU Berlin
[2] Assessment of IP-based Applications, Deutsche Telekom Laboratories, TU Berlin

ABSTRACT

The quality of audio-visual services is largely influenced by the qualities of the auditory and the visual signals. Perception of both signals is integrated and a single audio-visual rating is formed. The integration function depends, however, on the type of application. In this paper, we compare different integration functions which have been determined from empirical data collected in human-human and human-machine scenarios with different degrees of interactivity. The results show that the impact of each modality depends on both, the type of scenario and its degree of interactivity.

Index Terms **—** Audio-visual quality, multimodality, integration function, interactivity

1. INTRODUCTION

Quality has been defined as the results of a perception and a judgment process, in which the perceiving human compares the perceptual event happening inside her brain to some expected or desired reference [1]. Thus, quality is a relative entity, and depends on the particular application it is linked to. Nevertheless, valid and reliable quality measurements can be obtained by asking test participants for an Absolute Category Rating (ACR) on the perceived quality. This rating obviously reflects, amongst others, the test context and the application it has been obtained for. The most common ACR rating scale is the one on "overall quality" defined in ITU-T Rec. P.800 [2] for speech and in ITU-T Rec. P.910 [3] for video applications; the mean rating on this scale averaged over all test participants and stimuli belonging to one test condition is called a mean opinion score, MOS.

For audio-visual services like IP-based television (IPTV), the overall quality will depend on both the quality of the visual signal and the one of the auditory signal. Thus, separate MOS values can be obtained for the audio-only (MOS_A), the video-only (MOS_V) and the audio-visual (MOS_{AV}) situation. The first two can be further subdivided into situations where the audio signal is judged upon in a listening-only context ($MOS_{A|A}$) vs. a listening-and-viewing context ($MOS_{A|AV}$), or where the video signal is judged upon in a viewing-only context ($MOS_{V|V}$) vs. a listening-and-viewing context ($MOS_{V|AV}$)[1]. Corresponding MOS values can also be collected for interactive audio-visual services, like video-telephony or video-conferencing. The paradigm can further be transferred to human-machine interaction with human-like characters, so-called Embodied Conversational Agents (ECAs). For the development of ECAs, it is useful to quantify the quality of the (mostly synthesized) speech signal and the quality of the graphical animation alone, and to combine it to form an overall quality judgment of the entire ECA.

The combination of audio (or speech) and video quality judgments to an overall quality judgment for an audio-visual service is important for service planning and monitoring: Developers need to know where to invest (e.g. in terms of bandwidth, signal processing capability, routing priority, optimization effort) in order to optimize the overall quality for the user. For audio-visual human-to-human[2] communication scenarios, a number of studies have been carried out in the past, and linear as well as nonlinear integration functions are summarized in ITU-T Rec. P.911 [4]. For human-machine interaction with ECAs, such studies are far less frequent.

In this article, we begin with presenting some modality properties which seem to be relevant for audio-visual quality integration (Chapter 2) and review studies reported in the literature (Chapter 3). Then we describe five studies which have been carried out in our lab in the past 18 months (Chapter 4), and analyze them with respect to the quality integration function (Chapter 5). We put an emphasis on the effect of the type of interaction scenario (human-human vs. human-machine) and the degree of interactivity (passive vs. semi-interactive vs. fully-interactive) on the weightings associated with the two modalities. We conclude by interpreting the obtained results in the light of the different application scenarios.

---

[1] In the reminder of the paper, we note $MOS_{A|A} = MOS_A$ and $MOS_{V|V} = MOS_V$.

[2] We consider IPTV as a human-to-human scenario, as the transmitted source signal belongs to the natural domain.

 DOI: 10.1109/QOMEX.2010.5518100

2. MODALITY PROPERTIES

The aim of most audio-visual services is to convey information. For human-machine interactions, characteristics of modalities have been defined [5] describing

- how the modalities relate to each other in fulfilling this task: *complementary* vs. *redundant*
- how the modalities relate in their temporal order: *parallel* vs. *sequential*

These properties are also valid for a human-to-human interaction scenario involving an audio-visual transmission channel. For example, during an audio-visual communication parts of the information might be conveyed via speech and others via the video (complementary); the information might only be complete if both modalities are used in a specific order (sequential, e.g. talking and then showing something), or if they are used in parallel. Depending on the interaction scenario, one or the other characteristic may be dominant: For example, in a video-telephony situation most information may be conveyed via speech, and the visual signal just supports identifying the intentions and emotions of the speaker. In contrast to this, the verbal reference to a physical set-up in the same situation may force the attention towards the visual signal, and the speech signal becomes secondary only.

Parallel usage of different modalities may distribute the cognitive load in a better way, and thus lead to a higher quality. According to Wickens' model [6], the resources can be classified as to the input modality, the stage of processing (perception, cognition, response), and the type of code or response (spatial/manual vs. verbal). The better the resources are distributed along these dimensions, the lower the cognitive load. Although some results contradictory to this have been reported as well, e.g. by Schomaker et al. [7], it can be assumed that the possibility to distribute the resources will have a positive impact, in particular in highly-demanding situations where users have been observed to act more multimodally [8].

3. INTEGRATION FUNCTIONS

The integration of audio and video qualities to an overall audio-visual quality metric can be performed on different levels. Because of its convenience, most commonly MOS values are combined in a linear or non-linear way, like in the following approaches:

$$\mathrm{MOS}_{\mathrm{AV}} = c_1 \cdot \mathrm{MOS}_{\mathrm{A}} + c_2 \cdot \mathrm{MOS}_{\mathrm{V}} + c_4 \quad (1)$$

$$\mathrm{MOS}_{\mathrm{AV}} = c_3 \cdot \mathrm{MOS}_{\mathrm{A}} \cdot \mathrm{MOS}_{\mathrm{V}} + c_4 \quad (2)$$

$$\mathrm{MOS}_{\mathrm{AV}} = c_1 \mathrm{MOS}_{\mathrm{A}} + c_2 \mathrm{MOS}_{\mathrm{V}} + c_3 \mathrm{MOS}_{\mathrm{A}} \mathrm{MOS}_{\mathrm{V}} + c_4 \quad (3)$$

For media streaming and passive listening-and viewing-only situations, Eq. 2 has proven to provide reliable estimations. ITU-T Rec. P.911 [1] summarizes the results of four studies where this integration function led to Pearson correlations between predicted and measured $\mathrm{MOS}_{\mathrm{AV}}$ scores in the range $r = 0.93\ldots0.99$, with constants $c_3 = 0.107\ldots0.121$ and $c_4 = 1.1\ldots1.5$. It is stated that these values have been calculated for a 9-point quality scale, for synchronized audio-visual signals, and for signals where audio and video have a similar impact on quality. Further results for Eq.s 2 and 3 are summarized in [9], showing correlations $r > 0.9$ in most cases. For interactive (video-telephony or video-conferencing) situations, no integration functions are given.

Apart from the impact of the auditory and the visual signal alone, also their synchronization may play a role, in particular when persons are speaking and lip synchrony is an issue. Asynchrony may impact the overall quality, and it can be analyzed on a local level (e.g. for individual phonemes and corresponding visemes), or on a global level (e.g. when audio and video delay are unequal). While we disregarded the global asynchrony to avoid overloading the experiments with too many factors and test conditions, local synchrony is inherently an issue for a talking head where the articulatory movements of the animated head have to be aligned with the speech signal through some type of articulatory or morphing model. We therefore included the aspect of local asynchrony in the investigations of the ECA scenarios, and admit that global asynchrony is an interesting topic for further study.

4. EXPERIMENTS

Five experiments have been carried out at TU Berlin in the past 18 months which address audio-visual quality in different contexts and interactivity settings, see Table 1. The most relevant details of these experiments are summarized in the following paragraphs; for an in-depth description of each experiment, the reader is referred to the cited references.

Table 1: List of experiments

	Type of interaction	
Interactivity	Human-human	Human-machine
Low	Test 1: IPTV	Test 3: ECA passive
Medium	--	Test 4: ECA simulated interaction
High	Test 2: Video-telephony	Test 5: ECA interactive

Test 1 is a series of three viewing-only, listening-only and viewing-and-listening experiments carried out in the frame of the T-V-Model project of Deutsche Telekom, see [9]. It included 5 HD audio-visual contents of 16 s duration each which are representative of different TV-programs, differing in their spatial and temporal complexity (movie trailer/speech on music, interview/ speech, soccer/ speech on noise, movie/ classical music and music video/ pop music with singer). The sequences were degraded with different audio and video codecs, and partially with packet loss.

 DOI: 10.1109/QOMEX.2010.5518100

The applied Packet Loss Concealment introduced freezing and slicing. Listening and viewing conditions were compliant to ITU-T Rec. P.800 for the auditory and ITU-R Rec. BT.500-11 for the visual part. 24 subjects participated in each test and rated each condition on a continuous 11-point quality scale as recommended in ITU-T Rec. P.910 [3].

Test 2 is an interactive experiment for testing the impact of audio and video degradations in video-telephony, see [11]. 24 test participants carried out interactions over an impaired audio-visual channel following two scenarios: A building block or "Lego" scenario where one participant had an object made of Lego blocks which the other participant had to build on her own, following the instructions and visual examples of the partner; and a "short conversation test" (SCT) scenario which is recommended for speech-only conversations in ITU-T Rec. P.805 [12] and where the participants have to exchange information to perform some transaction, e.g. travel booking or an appointment with the doctor. Transmission degradations included speech and video codecs of different bitrate, and packet loss. After each conversation, test participants were asked to rate the audio, video, and audio-visual quality on a continuous 11-point rating scale, which have been averaged to corresponding $MOS_{A|AV}$, $MOS_{V|AV}$ and MOS_{AV} values.

Tests 3 and 4 addressed different ECA versions consisting of three animated heads combined each with two male Text-To-Speech (TTS) systems, see [13]. In Test 3, sentences relating to the smart-home domain of approx. 2 sec. length were generated with six differen ECA versions. They were rated by 14 participants in a pure listening-and-viewing context with respect to their speech quality, visual quality, and overall quality, on a 5-point ACR scale. In Test 4, the four best-rated of the six ECA versions were presented to 24 test participants as metaphors of a smart-home system. The participants interacted each with each ECA to perform a series of simple tasks, the results of which were however only simulated and presented on the screen. After each interaction, ratings of speech quality, visual quality, overall quality, the goodness of fit between the animated head and the voice, as well as on the synchronization were collected on 5-point ACR scales.

Test 5 was a replica of Test 4, however in a fully-operational smart-home environment where the test subjects could experience the effects of their interaction via feedback from home devices installed in a living room [14]. The four ECA versions of Test 4 were amended by a speech-only version, and each test participant interacted with two ECAs and the speech-only version. After each interaction, ratings of speech quality, visual quality and overall quality were solicited from all 49 test participants on 5-point ACR scales.

5. ANALYSIS

For each of the experiments, we calculated different types of integration functions and calculated the Pearson correlation r and the root mean squared error RMSE. In order to reach comparable results, we first transformed ratings obtained on a scale in the range [1;5] to the range [1;9], 9 corresponding to the label "excellent" and 1 to "bad".

5.1. Test 1: IPTV

In the IPTV tests, the audio-visual quality seems to be predominantly influenced by the video quality. The correlation between MOS_V and MOS_{AV} is much higher (0.83) than the one between MOS_A and MOS_{AV} (0.49), and the weights for MOS_V are always higher than the ones for MOS_A. Averaging the ratings over all test participants and five content conditions, and then calculating all three models of Eq.s 1-3, we obtain the results in Table 2.

Table 2: IPTV models

Eq.	c_1	c_2	c_3	c_4	r	RMSE
(1)	0.301	0.530	--	0.585	0.94	0.52
(2)	--	--	0.089	2.425	0.94	0.51
(3)	0.014	0.202	0.067	1.957	0.97	0.36

Although all models perform very well, the prediction accuracy can be slightly improved by including both linear and non-linear terms in the model. This highlights an interaction effect between audio and video quality in this scenario.

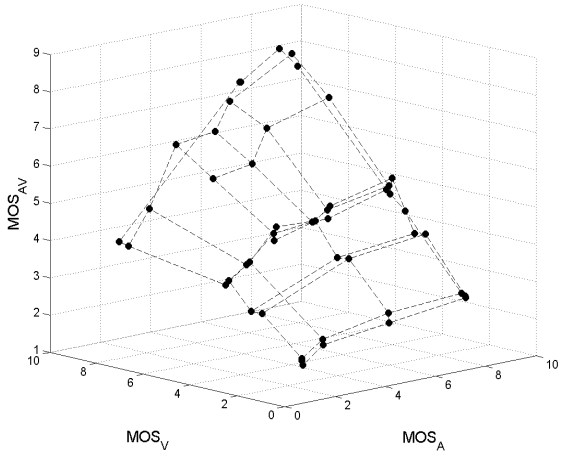

Figure 1: Impact of audio and visual quality on MOS_{AV} in Test 1 [9].

Figure 1 shows that the impact of audio quality is much higher when the video quality is high. To a lesser extent, impact of video quality is also higher with high audio quality. Apparently, video degradations are the basic quality-limiting factor in this scenario; only if the video quality is high, the audio can have a noticeable effect.

5.2. Test 2: Video-telephony

As in the IPTV situation, the correlation between $MOS_{V|AV}$ and MOS_{AV} is much higher (0.82) than the one between

 DOI: 10.1109/QOMEX.2010.5518100

$MOS_{A|AV}$ and MOS_{AV} (0.59). However, there is no clear dominance of the visual quality over the auditory one. Calculating the three models of Eq.s 1-3 for the entire data set, the impact of audio is slightly higher that the one of video, see Table 3. The difference is however not as clearly pronounced as in the IPTV scenario. All models obtain a very high correlation between subjectively measured and predicted MOS_{AV} scores, which is probably due to the fact that only few data points (MOS values for 13 test conditions) are modeled.

Table 3: Video-telephony models

Eq.	c_1	c_2	c_3	c_4	r	RMSE
(1)	0.64	0.54	--	-0.89	0.99	0.06
(2)	--	--	0.09	2.90	0.99	0.07
(3)	0.50	0.40	-0.02	-0.07	0.99	0.06

As Figure 2 shows, the impact of audio on the audio-visual quality is much higher when the video quality is high. With low video quality, video degradations mask the auditory ones. However, the range of covered audio qualities is much more restricted than the one of video qualities.

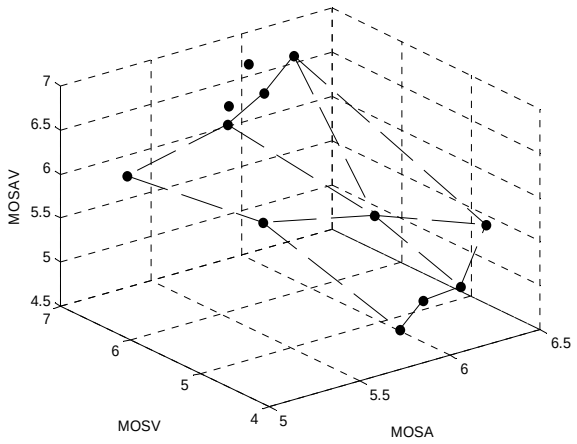

Figure 2: Impact of audio and visual quality on MOS_{AV} in Test 2 [11]

Two distinct scenarios have been used in Test 2. In the "Lego" scenario, the video channel provides indispensable and complementary information, and it is used both sequentially and in parallel with the audio channel (e.g. to indicate where to put a specific block). In contrast to this, in the SCT scenarios the video channel is mostly redundant, and it is probably used in parallel to the audio one. Computing separate models for both scenarios, it appears that in the SCT scenario the impact of audio is higher than in the "Lego" scenario, see Table 4. Because only four conditions were tested with the SCT scenario, these results are tentative only, and only simple linear models of Eq. 1 have been calculated.

Table 4: Video-telephony models for different conversation scenarios, following Eq. 1.

Scenario	c_1	c_2	c_4	r	RMSE
Lego	0.59	0.53	-0.63	0.99	0.07
SCT	0.79	0.54	-1.74	0.99	0.01

5.3. Test 3: ECA Passive

In the passive listening-and-viewing situation, test participants managed quite well to separate the speech and the video quality of the presented ECAs. The rating of the visual quality ($MOS_{V|AV}$) depends only on the animated head, and the rating of speech quality ($MOS_{A|AV}$) depends far more on the TTS used than on the talking head. For this reason, a simple model without interaction term already leads to a very high correlation, cf. Table 5. Both models (1) and (3) are slightly but significantly (F-test) better than model (2).

Table 5: ECA Passive models

Eq.	c_1	c_2	c_3	c_4	r	RMSE
(1)	0.495	0.371	--	0.863	0.96	0.21
(2)	--	--	0.073	3.428	0.95	0.26
(3)	0.503	0.379	-0.001	0.813	0.96	0.21

Interestingly, in this situation the speech quality is far more dominant than the visual quality of the ECA. The prediction accuracy is very similar to what can be achieved in the human-to-human interaction scenarios. It does not improve when including an interaction term into the model, showing that speech and visual quality are mainly independent in this situation.

5.4. Test 4: ECA Simulated Interaction

In order to investigate additional influences which might occur in interactive situations, we further included the scales on goodness of fit between animated head and voice as well as the synchronization between speech and lip movements in the analysis of the Test 4 results. Assuming a linear relationship, the synchronization proved to have a significant impact on overall audio-visual quality, whereas the goodness of fit between animated head and voice did not. However, the performance of the linear and the non-linear models did not benefit from the additional variable, and the mixed linear-nonlinear model contains too many parameters compared to the sparse training data (3 heads · 2 voices · 2 tasks = 12 data points). We therefore calculated only the models of Eq.s 1-3 with their original variables $MOS_{A|AV}$ and $MOS_{V|AV}$, see Table 6.

The models' fit is significantly worse than in the passive situation of Test 3. Apparently, the overall quality seems to be impacted more and more by other factors than the speech and the visual quality alone. From the latter, the visual

 DOI: 10.1109/QOMEX.2010.5518100

quality of the animated agent is now the dominating factor over the speech quality. No significant effect of the interaction term is observed.

Table 6: ECA Simulated Interaction models

Eq.	c_1	c_2	c_3	c_4	r	RMSE
(1)	0.212	0.434	--	2.385	0.86	0.12
(2)	--	--	0.049	4.413	0.85	0.11
(3)	0.319	0.556	-0.020	1.731	0.86	0.13

5.5. Test 5: ECA Interactive

Compared to Test 4, the ECA was used as part of a fully interactive system in Test 5. Thus, it can be expected that the overall quality rating of the ECA is also influenced by other factors of the system. From the four ratings collected in this experiment, only speech quality, visual quality and synchronization contributed significantly to the overall quality and were included in the models. We calculated here only the linear and the purely nonlinear models corresponding to Eq.s 1 and 2, as the full model contains too many parameters compared to the training data, and used the individual judgments instead of the mean as there were four system configurations only included in the test:

$$MOS_{AV} = 0.225 \cdot MOS_{A} + 0.430 \cdot MOS_{V} + 0.288 \cdot MOS_{SYN} + 1.220 \quad (4)$$

$$MOS_{AV} = 0.006 \cdot MOS_{A} \cdot MOS_{V} \cdot MOS_{SYN} + 4.870 \quad (5)$$

In contrast to Test 3 and in line with Test 4, video quality is again the dominating factor. Speech quality is less important, even less than the synchronization between speech and lip movements. The linear model of Eq. 4 is able to predict audio-visual quality with a correlation r = 0.64 (RMSE = 1.05), and the non-linear model of Eq. 5 with r = 0.56 (RMSE = 1.11); apparently, the models' prediction performance is very poor compared to the passive and non-interactive settings (even a full model with linear and non-linear terms only reaches r = 0.69 and RMSE =1.00). A separate question addressing the overall system quality showed that this is significantly influenced by MOS_{AV} (Wilcoxon ranked sign test).

6. DISCUSSION

For each of the experiments, we could derive audio-visual quality integration models which fit the experimental data more or less well, but which still have to be validated against unknown data. These models vary with respect to the impact of the different constituents (audio quality, video quality, plus partially synchronicity) depending on the type of scenario (human-to-human vs. human-to-machine) and the degree of interactivity involved. We can make the following observations:

- In the passive human-to-human scenario (IPTV), the visual quality clearly dominates over the auditory quality. This is shown both by a higher weighting of the visual quality compared to the auditory one, as well as by a higher correlation between visual and audio-visual quality.
- In the interactive human-to-human scenario (video-telephony), the impact of the audio is slightly higher, and although the correlation between video and overall quality is higher, it seems that the audio plays a more dominant role. In particular in the short conversation test scenarios – where the video channel is mainly redundant and used in parallel to the audio channel – impact of audio is dominant, whereas in the "Lego" scenario – where the video is complementary and used both sequentially and in parallel to the audio channel – the situation between the two modalities is more equilibrated. Apparently, the usage patterns of video and audio influence the weighting of the two modalities in the integration model.
- In both human-to-human scenarios, there is an interaction between audio and visual quality. As a result, the impact of the auditory quality on the audio-visual quality is more pronounced with high visual quality than with low visual quality. Or, low visual quality seems to mask the effect of the audio.
- In the non-interactive human-machine interaction scenarios with the ECA, speech quality (in an audio-visual context) is more important than visual quality. This finding was not expected, in particular because in the passive listening-and-viewing situation no particular task was given to the test participants the resolution of which would have required speech information. In the simulated and the fully interactive settings, the dominance of speech over visual quality is inverted.
- For the semi-interactive and the fully interactive human-machine interaction scenario, the synchronicity between animated head and voice seem to be an important factor; the goodness of fit between head and voice, in turn, was not a significant factor. It should be noted however that these factors have not been assessed in the passive setting.
- The correlation between observed MOS_{AV} and the predicted values using the cited models is very high in the two human-to-human scenarios, as well as in the passive human-machine interaction scenario; it is lower in the (semi and fully) interactive human-machine scenarios. Apparently, the link of the audio-visual signal with a background system (in our case a smart-home system) – be it simulated or fully implemented – incites the test participants to not only consider the quality of the audio-visual signal alone, but to link it to the quality of the entire system. In fact, a significant impact of the audio-visual quality on system quality could be observed in the interactive setting.

These results are briefly summarized following the categorization of scenarios of Table 1, see Table 7.

 DOI: 10.1109/QOMEX.2010.5518100

The results have some implications for planning high-quality audio-visual services. Whereas most effort might be needed for ensuring good visual quality in IPTV situations, at least parts of the effort should be dedicated to the speech channel in case of video-telephony and interactions with an ECA. In a realistic interaction with an ECA, the audio-visual quality will not only depend on the auditory (TTS) and the visual (animated agent) signal, but will also depend on how these two are synchronized in order to create realistic lip movements.

Table 7: Implications of the experiments for the quality integration functions. >>: strongly dominates over; > dominates over; ↔: interaction.

	Type of interaction	
Interactivity	Human-human	Human-machine
Low	▪ $MOS_A << MOS_V$ ▪ $MOS_V \leftrightarrow MOS_A$ ▪ excellent model fit	▪ $MOS_A >> MOS_V$ ▪ MOS_V and MOS_A independent ▪ excellent model fit
Medium	[not tested]	▪ $MOS_A << MOS_V$ ▪ MOS_V and MOS_A independent ▪ moderate model fit
High	▪ $MOS_V < MOS_A$ ▪ $MOS_V \leftrightarrow MOS_A$ ▪ excellent model fit	▪ $MOS_A << MOS_V$ ▪ MOS_V and MOS_A independent ▪ poor model fit

7. CONCLUSIONS

We presented a series of experiments addressing the quality of audio-visual signals in human-to-human and human-machine interaction scenarios. We showed that the integration function relating auditory and visual quality to overall (audio-visual) quality depends on the type of scenario as well as on its interactivity. Depending on these factors, we created a matrix which shows dominance and interaction of these two constituents, as well as the goodness of fit of the derived prediction models.

Depending on the variety of the interaction scenario, we expect that the goodness of fit of the models derived for human-to-human interaction may also decrease, e.g. when additional information (text, still pictures, vibration) is transmitted in parallel to the audio-visual signal. In such situations, the impact of the audio-visual signal on overall quality might be reduced, as it was the case with the interactive ECA. In addition, we did not yet address the issue of global synchrony between audio and video for the human-to-human interaction scenario, and we will further analyze the impact of the audio-visual content.

8. REFERENCES

[1] Jekosch, U., *Voice and Speech Quality Perception. Assessment and Evaluation*, Springer, Berlin, 2005.

[2] ITU-T Rec. P.800, *Methods for Subjective Determination of Transmission Quality*, Int. Telecomm. Union, Geneva, 1996.

[3] ITU-T Rec. P.910, *Subjective Video Quality Assessment Methods for Multimedia Applications*, Int. Telecomm. Union, Geneva, 2008.

[4] ITU-T Rec. P.911, *Subjective Audiovisual Quality Assessment Methods for Multimedia Applications*, Int. Telecomm. Union, Geneva, 1998.

[5] Coutaz, J., Nigay, L., Salber, D., Blandford, A.E., May, J., Young, R.M., "Four easy pieces for assessing the usability of multimodal interaction: The CARE properties", in: *Human-Computer Interaction, Proc. Interact'95* (Nordby, K., Helmersen, P.H., Gilmore, D.J., Arnesen, S.A., eds.), Chapman & Hall, London, pp. 115–120, 1995.

[6] Wickens, C. D., "Multiple resources and mental workload", *Human Factors: The Journal of the Human Factors and Ergonomics Society* 50, pp. 449–455, 2008.

[7] Schomaker, L., Nijtmans, J., Camurri, A., Lavagetto, F., Morasso, P., Benoît, C., Guiard-Marigny, T., Le Goff, B., Robert-Ribes, J., Adjoudani, A., Defée, I., Münch, S., Hartung, K., Blauert, J., "A taxonomy of multimodal interaction in the human information processing system", *Report of the ESPRIT Basic Research Action 8579 MIAMI*, NICI, Nijmegen, 1995.

[8] Oviatt, S., "Multimodal interfaces", in: *The Human Computer Interaction Handbook* (Sears, A., Jacko, J.A., eds.), Lawrence Erlbaum, New York, 2. Ed., Chapter 21, p.413–432, 2008.

[9] Garcia, M., Raake, A., "Impairment-factor-based audio-visual quality model for IPTV", in: *Proc. First International Workshop on Quality of Multimedia Experience (QoMEX 2009)*, 29–31 July, San Diego, 2009.

[10] Möller, S., *Assessment and Prediction of Speech Quality in Telecommunications*, Kluwer Academic Publ., Boston, 2000.

[11] Belmudez, B., Möller, S., Lewcio, B., Raake, A., Mehmood, A., "Audio and video channel impact on perceived audio-visual quality in different interactive contexts", in: *2009 IEEE Int. Workshop on Multimedia Signal Processing (MMSP'09)*, 5–7 Oct., Rio de Janeiro, 2009.

[12] ITU-T Rec. P.805, *Subjective Evaluation of Conversational Quality*, Int. Telecomm. Union, Geneva, 2007.

[13] Weiss, B., Kühnel, C., Wechsung, I., Fagel, S., Möller, S., "Quality of talking heads in different interaction and media contexts", accepted for *Speech Communication*, 2010.

[14] Kühnel, C., Weiss, B., Möller, S., "Talking heads for interacting with spoken dialog smart-home systems", in: *Proc. 10th Ann. Conf. of the Int. Speech Communication Assoc. (Interspeech 2009)*, ISCA, pp. 304–307, 2009.