Quality-of-Experience Beyond MOS: Experiences with a Holistic User Test Methodology for Interactive Video Services

Sebastian Egger*, Michal Ries†, Peter Reichl*
*Telecommunications Research Center Vienna (ftw.)
Donau-City-Strasse 1, A-1220 Vienna, Austria
(egger, reichl)@ftw.at
†Institute of Communications and Radio-Frequency Engineering
Vienna University of Technology
Gusshausstrasse 25, A-1040 Vienna, Austria
(mries)@nt.tuwien.ac.at

***Abstract*—Interactive video services, like video telephony, social TV or on-line gaming, are about to become a significant part of the service portfolio for telecommunication service providers. As the future commercial success of these services will depend essentially on their end-to-end quality as perceived by the end user, appropriate Quality-of-Experience (QoE) measurement methods are of paramount importance. The aim of this paper is to provide an application-oriented re-evaluation of actual QoE metrics for these services and to design a test methodology for evaluating user perceived experience which goes beyond standard Mean Opinion Score (MOS) metrics. To this end, we describe a specific test scenario designed for assessing the currently recommended question sets from two ITU recommendations. We determine the goodness of fit of these question sets to the user perceived quality dimension. Altogether, the resulting reduced set of items (questions) provides a significant step towards a more realistic methodology for assessing audio-visual quality perception.**

***Index Terms*—Quality of Experience, Quality of Service, audio-Visual quality, perceptual quality, social presence, conversational interactivity.**

I. Introduction

Despite of their obvious potential to supplement and enrich standard voice-only communication, for long years audio-visual communication services have failed to make a valid business case, probably due to a lacking awareness of its added value on the customer side. Nevertheless, there is a growing belief that these services will become a major part of the service portfolio of telecommunication service providers in the foreseeable future, while the focus of service design and development is targeted much more specifically to the user and her real needs.

In this ongoing evolution, the notion of service quality will play a paramount role. The fact that service quality evaluation in general becomes weak without a strong user-centric component has been recognized by ITU-T already 15 years ago, e.g. their Rec. E.800 [10] which provides a general overview of sub-dimensions contributing to the user perceived Quality of Service (QoS) of telecommunication services. The contribution of these sub-dimensions is quantitatively correlated with respective technical QoS parameters and accordingly the overall QoS results from a summation of those parameters. However, subsequent research work has been primarily focused on the technical aspects of this approach, thus more and more lacking sufficient evidence of the user's quality perception (for a detailed review of this development please refer to [19]).

In order to bridge this gap, the ITU has repeatedly proposed subjective quality assessment methods for multimedia applications, for instance [13] and [12]. More recently, these approaches have been superseded by the fully user-centred approach of Quality-of-Experience (QoE), which is referring to a higher abstraction layer compared to QoS. Essentially, QoE can be considered as a perceptual layer and an extension to the application layer defined in the OSI model [1]. To further illustrate the kaleidoscope of QoE dimensions, Figure 1 depicts a comprehensive selection of various influence factors playing a crucial role for the resulting overall QoE from a user perspective. Nevertheless, this selection is still only a subset of dimensions contributing to QoE, as has been argued e.g. in [19] and [16]. Note that, among the depicted sub-dimensions, the notion of user expectations is of particular interest as its influence is basically neglected in all currently available QoE assessment methodologies. User expectations do influence the quality level necessary for satisfactory mediated communication, hence they vitally contribute to the overall QoE.

Whereas originally QoE has been mainly investigated in the context of Voice-over-IP (cf. [2], [8], there is an increasing need to introduce appropriate test methodologies also for the QoE of interpersonal interactive video services as a straightforward extension of current voice quality assessment methodologies. Such a QoE measure should reflect the users' perceived quality experience of the content presented by the application layer. As such the application layer acts as a user interface reflecting the overall result of the individual QoS from the underlying layers [24].

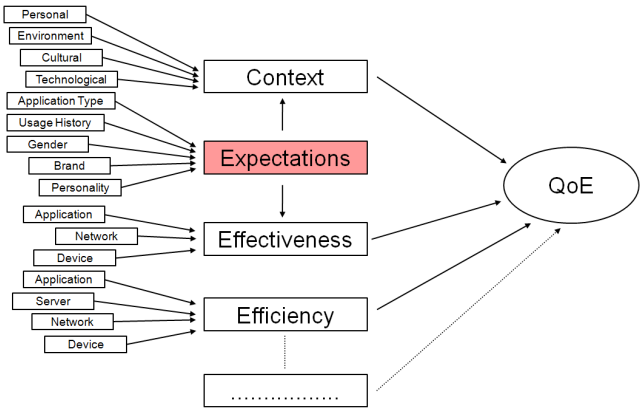

Fig. 1: QoE dimensions (extending [3])

The goal of our research is to derive a suitable methodology for the QoE assessment of interactive video services based on classical experience assessment methods [13], [12] enriched with sociopsychological measures [18]. Additionally, we have used the methodology proposed in [4] to assess the interactivity level of audio-visual conversations, thus including the user's differing interaction behavior and as such incorporating interaction properties into the quality metric. In this way, we explicitly address the QoE dimension of expectations (cf. Figure 1) which is influenced by the interactivity level. As shown in [6], highly interactive conversations in general demand for better quality, due to specific user expectations. To gather data for the verification of the used metrics (i.e. items), we have conducted user tests on video call scenarios with varying QoS parameters. The scores obtained in these trials have been utilized to evaluate the fit of the used items on measuring the users' QoE by applying multivariate statistical methods.

The remainder of the paper is organized as follows: In Section II we describe the test design for the evaluation of user perceived quality in an interactive video communication system. Quantitative test results and results from qualitative data gathering methods are discussed in Section III. Section IV summarizes our results and provides some conclusions and recommendations for related QoE testing methods.

II. Test Design

As a starting point, we describe a fundamental study which has been carried out as paired user test. In the context of evaluating human social interaction, this paired approach has the following three advantages: increased realism, less need for facilitator intervention and higher quality of results [23].

The study has been conceived as a series of controlled user experiments in order to observe the impact of the factor variables (video resolution and allocated bandwidth) on the subject's quality perception.

In each session, lasting about 45 minutes, two participants were exposed to three different scenarios as described below. We randomized their sequence in order to provide a counterbalanced study design and to avoid biased results due to e.g. learning effects. This approach enabled us to gather detailed comparison results of user perception for the different scenarios we exposed them to.

After each scenario, participants were asked to fill in a questionnaire. The questionnaires, which are described in detail in section II-C, include a consistent set of items addressing the user experience. In addition, items have been included that relate to specific aspects of a scenario. Among such specifics are the overall impression of the test situation and the need of adjustments to the video-conferencing system in use. In addition to the quantitative data collected through these questionnaires we have analyzed video recordings of the subjects in a qualitative manner. The goal of this qualitative analysis was to identify interaction cues exchanged and to detect their occurrence. The interaction cues are discussed in depth in section II-D.

A. Conversational Tasks

In order to increase the degree of conversation activity we asked the participants to accomplish a shared task. As we wanted to include the visual channel, standard tasks for interactive audio-only conversations could not be used (cf. [7]) because they lack visual interactivity. Therefore, the test pairs have been asked to jointly complete a Lego construction set as suggested in [14]. To this end, one of the subjects held the instructions while her partner had the parts in front of him. The test subjects were allowed to show either the instructions or the Lego parts in front of the camera. By permitting this, we wanted to enable the possibility of providing semantical information via inclusion of external references.

The Lego sets provided are originally targeted toward an age group of eight to twelve year old children. In conjunction with the division of instruction and parts, this task is supposed to be challenging enough to create the need of extensive information exchange between the participants.

B. Scenarios

The two subjects were exposed to three different scenarios. In the first two scenarios ("HiFi" and "LoFi") the participants were connected with a video-conferencing system. The scenarios differed in visual resolution and allocated bandwidth (see Table I). The used video-conference system was based on a customized VLC [25]. The standard VLC settings were tweaked to achieve sufficient performance regarding the one-way delay (for further details cf. [4]). For presenting the video to the participants we used common 19-inch LCD screens. The video window was the only opened window and the desktop background was set to grey as recommended in [14]. Detailed technical settings are provided in Table I.

Frame Rate	15 fps
Video Codec	H.264/AVC baseline profile [15].
Audio Codec	AAC [9] at 16 kbps
HiFi	VGA resolution (640 x 480 pixels) at a video bitrate of 512 kbps
LoFi	QVGA resolution (320 x 240 pixels) at a video bitrate of 256 kbps

TABLE I: A/V Settings

The average one-way delay achieved with this setting was about 250ms. Acoustically, headsets were used to capture and play the voice signals.
The third scenario was a face-to-face setting where the participants were seated opposite to each other. To avoid cheating, a 35cm high optical barrier was set up between them. In this scenario the participant holding the instruction was not allowed to show it, while the other one was allowed to show the parts for identification purposes. The face-to-face scenario has been chosen to work as reference condition and allows to measure the user perceptions differentially on an intrapersonal level via the semantic differential method (discussed in more detail in the following section). By comparing intrapersonal differences we can rule out variations based on contrastive sensing of scenarios by different subjects.

C. Evaluation Questionnaires

We have developed two different questionnaires targeted towards user perception and demographic data, respectively.

The *User Perception Questionnaire* has been designed in order to quantify the user perception with respect to the system settings. To this end, we have used two types of items, i.e. semantic differential scores (SD, items (1a) to (1o), cf. Table II) and Mean Opinion Scores (MOS, items (2) to (12), cf. Table III). The *semantic differential score* as used within this user study objectively measures the perceived social presence, in contrast to other studies which utilized it for measuring the aesthetic and activity dimension of user perception (cf. [26]). It has been included in order to investigate to which extent this method can be used for assessing the user perception of the system quality in an objective manner through the social presence sensed. Basically, this method uses contrary adjective pairs to measure the social presence perceived by interactants [18]. Note that the collection of items (adjective pairs) in Table II is based on the results of [22] where it is demonstrated that this method is viable to differentiate the amount of social presence perceived by interactants communicating over different interaction media. Another motivation for including this method refers to the question of potential interrelation between the SD and conversational interactivity measured through qualitative analysis (cf. II-D). This assumption is based on the fact that higher interactivity contributes vitally to the presence perception of the opposite interactant (cf. [4]). The final SD values are calculated as average values over all adjective pairs after removing their changing polarities (which are necessary for bias prevention). Within the PCA analysis discussed below, we further recoded the results into a five grade scale to simplify comparison to the other items gathered.

The second type of items, i.e. the *Mean Opinion Score items*, have been used to obtain a standard-compliant system quality rating. In general, MOS items are used to assess the perceived quality of transmission systems and are included in several ITU recommendations on assessment of communication systems. For instance, MOS tests for telephone communication are specified in ITU-T Rec. P.800 [11] and are adapted

(1a)	impersonal	1	2	3	4	5	6	7	personal
(1b)	cold	1	2	3	4	5	6	7	warm
(1c)	ugly	1	2	3	4	5	6	7	beautiful
(1d)	small	1	2	3	4	5	6	7	big
(1e)	sensitive	1	2	3	4	5	6	7	insensitive
(1f)	colourless	1	2	3	4	5	6	7	colourful
(1g)	asocial	1	2	3	4	5	6	7	social
(1h)	passive	1	2	3	4	5	6	7	active
(1i)	convenient	1	2	3	4	5	6	7	inconvenient
(1k)	humorous	1	2	3	4	5	6	7	humorless
(1l)	binding	1	2	3	4	5	6	7	non-binding
(1m)	confident	1	2	3	4	5	6	7	unconfident
(1n)	movable	1	2	3	4	5	6	7	unmovable
(1o)	happy	1	2	3	4	5	6	7	sad

TABLE II: semantic differential items as suggested by [18], [22]

(2)	Did you have the impression that the cooperation with your partner was working well?	(Scale 1 - 5)
(3)	How well could you understand your partner from an acoustic point of view?	(Scale 1 - 5)
(4)	How well could you understand your partner from a visual point of view (gestures, mimics)?	(Scale 1 - 5)
(5)	How did the achievement of the task solution work?	(Scale 1 - 5)
(8)	Did you have the impression that your partner was spontaneous reacting on your linguistic questions / instructions?	(Scale 1 - 5)
(9)	Did you have the impression that your partner was spontaneous reacting on your gestures / visual instructions?	(Scale 1 - 5)
(10)	How satisfied have you been with the overall quality (acoustic and visual judged together) of the preceding situation?	(Scale 1 - 5)
(11)	How satisfied have you been with the acoustic quality of the preceding situation?	(Scale 1 - 5)
(12)	How satisfied have you been with the visual quality of the preceding situation?	(Scale 1 - 5)

TABLE III: Questions targeted towards user perceived quality

for audio-visual communications in Rec. P.920 [14]. We have used both these recommendations as a source for compiling an adequate list of MOS items addressing the user perceived quality of interactive audio-visual scenarios, see Table III[1]. As test method we have used Absolute Category Rating (ACR) as it generally reflects real world usage scenarios better [20], [21]. Additionally, ACR simplifies the comparison between MOS values and SD items with respect to the user perceived quality. The participants answered the questions after each scenario using a five grade scale (1 - bad, 5 - excellent). Note that questions (10), (11), (12) cannot be applied to the face-to-face scenario and therefore have been omitted there.

As mentioned above, the user perception questionnaire has been supplemented by an additional *Demographic Questionnaire* in order to examine the possible influence of latent variables. To this end, we have gathered standard demographic data like age, education, sex and employment status. Overall, we recruited 10 pairs of test subjects via public announcements (14 male and 6 female), aged between 19 and 35 years (mean

[1]numbering starts with (2) as the SD items start with (1a) to (1o) and question (6) and (7) are not included as they where targeted towards the test setup itself and are not important to the QoE evaluation.

= 27.0 years, median = 28.0 years).

D. Qualitative Conversational Interactivity Assessment

To sustain our aim of enriching standard MOS items with further non-standard metrics, we focus on investigating the suspected correlation between conversational interactivity and user perceived quality. Here, the analysis of human interaction in [4], leads us to the following hypothesis: In the case of insufficient visual resolution as in the LoFi scenario (cf. Table I), a shift from visual interaction cues to vocal interaction cues has to be expected. As a consequence, such a decreased visual interactivity level could therefore be used as an indicator for lower quality in the visual channel.

To check this hypothesis, we have used the following interaction cues for assessing conversational interactivity levels in the visual and acoustic channel:

- Speaker alternation rate
- Head nods, head movements
- Shifts of gaze
- Mimic
- Body orientation
- Posture
- Approval gestures
- Graphical gestures
- Pointing towards external references

We have used video recordings of every session to code the respective rates of the interaction cues. Due to the lack of automatized cue identification procedures, we have manually counted the overall number of cues taking place and used the resulting rate for expressing the level of conversational interactivity related to each session.

III. Results and Findings

The given task to assemble a Lego construction set was in general well accomplished by the participants. All participants were familiar with its execution. We did, however, not expect that the subjects behave quite differently in the way they use the video-conferencing system. Our original assumption was that they use the visual channel to acknowledge if they e.g. were holding the right brick in their hands. Also hinting on the instruction manual in front of the camera was expected. Moreover the resolution of the HiFi scenario has been chosen such that the participants could identify at least the type of brick connection to be realized. We mentioned these options to the participants while explaining the tasks, but did not emphasize them in order to avoid biasing the test subjects. It was interesting to observe that, while some of the subjects extensively used such visual aid, the others did not make use of it at all. Particularly in the LoFi scenario some of the participants preferred to give instructions mainly verbally and, by doing that, they rarely looked up from the instruction manual. As a result of this observed behavior, the visual channel appeared to be relatively useless for this type of users.

The rest of this section deals with a detailed analysis of the results, starting with an analysis of the SD items and the MOS questions, focusing on their joint relevance for the user's perceptive dimension, before we finally analyze the conversational interactivity levels of the scenarios with qualitative methods.

A. Quantitative Results

The first significant outcome of our case study refers to the orthogonality and redundancy in the MOS related question set compiled in Table III which, as mentioned above, reflects the current state-of-the-art MOS methodology. To this end, we have decided to resort to Principal Component Analysis (PCA) [17] as a standard statistical method for mapping a high-dimensional data set in considerably less dimensions. In our case, the first two PCA components turned out to be sufficient for an adequate modeling of the data variance, as they jointly account for at least 62% of variability in all scenarios. We applied the biplot technique [5] to depict the correlation between the two principal components and the different raw items (Table III) plus the average SD value. Note that in Fig. 2,3,4 the x- and y-axis refer to the first two principal components, whereas the individual vectors depict the projections of the MOS items and the average SD value with respect to the principal components. Thus, biplotting allows to illustrate inter-item distances and to indicate clustering of items as well as to display variances and correlations of the individual items.

Fig. 2,3,4 contain the results of our PCA analysis. Summarizing, we observe that the parameter vectors of question pairs (2) + (5) and (8) + (9) are located rather close to each other, and therefore we may conclude that questions (2) + (5) and (8) + (9) are semantically related, respectively. As a consequence, we can in each case exclude one of the two related questions. Finally, we propose to prefer question (2) to question (5) as the latter one is targeting the specific task to be undertaken by the test subjects, and similarly to prefer question (8) to question (9) as the audio channel examined in (8) has been utilized much more frequently for information exchange compared to the visual channel targeted by (9).

Beyond recognizing potential redundancy, the PCA analysis allows additionally to identify the strong interdependency between the audio and overall quality in the LoFi scenario, as can be deduced from the fact that in Fig.3 the endpoints of items (10) and (11) are located extremely close to each other, thus demonstrating the dominating importance of the audio channel in this scenario. In the HiFi scenario, things look quite different, as here the PCA results show a closer distance between overall quality (10) and visual quality (12) (observe the respective vectors in Fig. 4). Summarizing, we may conclude that the HiFi scenario is influenced stronger by the visual quality perception than the audio one, whereas the LoFi scenario behaves the other way round.

Our final quantitative results refer to the SD data collection. Here, it was first of all interesting to observe that the subjects needed quite a substantial amount of time to answer the SD items of the questionnaire at all. As far as the informative value of this item type in terms of perceived social presence is concerned, we observe that on an *interpersonal level* the

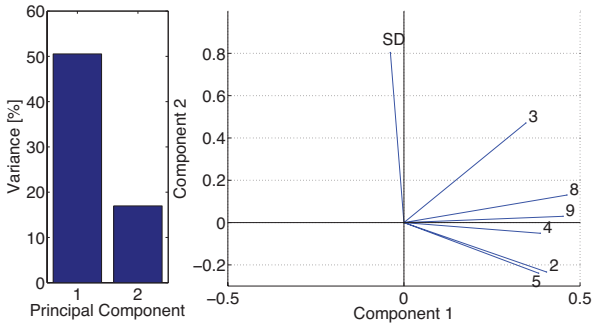

Fig. 2: Visualization of PCA results for Face-to-face scenario.

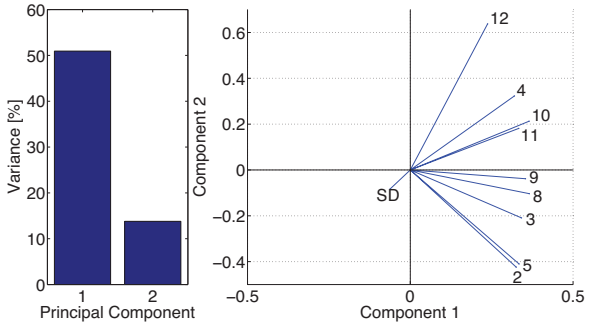

Fig. 3: Visualization of PCA results for LoFi scenario.

averaged SD scores for individual test subjects did not provide a significantly different evaluation of the perceived social presence in each of the three scenarios (the ”once cool, always cool” phenomenon). On an intrapersonal level, however, each test subject exhibits her own origin of the internal evaluation scale per SD item (cf. [18])[2].

Therefore, a differential analysis of the SD scores on an interpersonal level might lead to wrong conclusions, and we provide only results for the intrapersonal level. Here, most notably, the differences between the face-to-face and the LoFi scenario are significant (Z=-2.31, p=0.021)[3], hence we may state that in the face-to-face scenario a higher extent of social presence is sensed by the interactants. For both other cases, however, we could not detect a significant difference (Z=-1.68, p=0.094 for face-to-face vs. HiFi and Z=-1.28, p=0.205 for LoFi vs. HiFi).

Therefore, based on the result for the LoFi vs. HiFi case, we come to the conclusion that the SD method is not suited to detect a quality difference due to pure visual quality impairments as tested in our study. Furthermore, we observe that the parameter vector of SD is the only one which is projected to the negative part of the biplot diagram, thus indicating that it does not correlate with the other parameter vectors. In terms of perceived quality, we therefore conclude that there is only a weak relation between the SD and the

[2]for instance test subject A might perceive the social presence in the LoFi scenario as ”cold”, in the HiFi scenario as ”lukewarm” and in the face-to-face scenario as ”warm”, while subject B might judge the same scenarios as ”warm”, ”hot” and ”boiling” (note that the intermediate steps are identical for A and B, whereas the basement level differs)

[3]For identification of statistical differences, Wilcoxon signed ranks tests were calculated

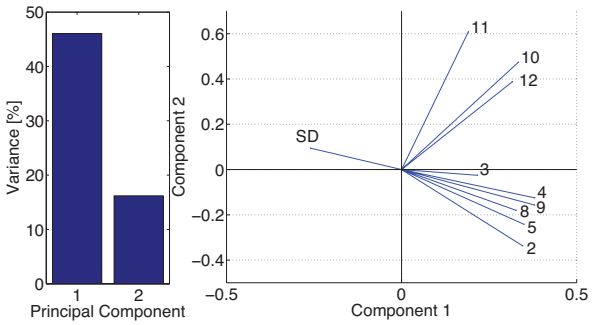

Fig. 4: Visualization of PCA results for HiFi scenario.

MOS items (see Figures 3, 2, 4).

B. Qualitative Results

Each pairwise analysis of a scenario was conducted separately for the interaction cues enumerated in II-D and for each video channel[4]. For assessing the speaker alternation rate we mixed the audio files for both directions into one stereo file which was then further used for the analysis.

The computation of the interaction cues shows that there is no difference on the nonverbal interactivity level for the LoFi and the HiFi scenario (Z=-0.66, p=0.51) nor exists a difference in the speaker alternation rate (Z=-0.15, p=0.89) for these two scenarios. This finding is surprising as we would have expected higher visual interactivity levels for the HiFi scenario. Moreover, we have noted also a trend for general higher interactivity levels (acoustical and visual) for the scenario being presented first to the subjects, indicating a gain of efficiency due to adaption to the task over time. From the observations we conclude the predominance of the acoustical channel which was used as main source of information for all subjects. This might be based on the well established utilization of acoustic only interaction while most of our subjects were not familiar with audio-visual interaction systems. We could witness that the subjects use rather elaborate strategies for the substitution of visual interaction cues with acoustic information.

These results suggest that for our setting the interactivity levels of the audio-visual conversation can not be used solely for further analysis regarding the system QoE. As a remedy we suggest a combinatory interactivity measure utilizing both the visual and verbal interactivity. Nevertheless, further research on interactivity measures and their value for QoE assessment is needed.

IV. Conclusions and Recommendations

This paper is devoted to a case study investigating new ways to derive a reliable quality measurement methodology for interactive audio-visual conversations which exceeds the standard MOS metrics. The proposed generalization concerns two different aspects, i.e. the inclusion of conversational interactivity as a characteristic of the expectation dimension

[4]From the pairwise setting we got one video and one audio channel for each direction

in QoE, together with the usage of the method of semantic differentials as social presence indicator.

In a first step, we have focused on deriving a sufficiently orthogonal set of MOS-related questions, starting from a state-of-the-art compilation based on related ITU-T recommendations. By using Principal Component Analysis we have been able to identify redundancies between questions related to inter-partner cooperation. Hence, it is recommendable to omit two of the propsed questions without significant loss of information yielding in a more compact and less time consuming question set.

From the obtained data we furthermore conclude that the SD method exhibits several limitations concerning its suitability to prove differences in the user perceived system quality if the system is impaired by decreased visual resolution only. Therefore, we recommend to further assess and adapt the SD measure to ascertain its abilities to detect visual and acoustical impairments on a sociopsychological level before incorporating it into a quality metric.

Multivariate statistical results show that mutual dependencies between audio and video quality are significantly influenced by the video resolution used. According to our results, the audio quality can be considered more important for the LoFi scenario, which is confirmed by our empirical observations. In case low resolution video is used we suggest therefore to additionally include detailed questions regarding the experienced audio quality, based on the relative importance of the audio connection in such scenarios.

Finally, the conversational interactivity level of the communication turned out to be difficult to assess in this setting, mainly due to deeply rooted preference for voice-only communications by the test subjects, probably as a result of life-long adaptation. However, with increasing dissemination of interactive video services and the subsequent rise of acquaintance within our target group, this attitude will change in the mid-term future, thus increasing the prominence of the novel approach sketched in this paper. Therefore, future work will concentrate on further adapting and fine-tuning the presented methodology in order to enable reliable QoE prediction beyond plain MOS.

ACKNOWLEDGMENT

The authors would like to thank mobilkom austria AG for supporting their research. The views expressed in this paper are those of the authors and do not necessarily reflect the views within mobilkom austria AG. Parts of this research have been performed within the project U0 SUPRA++ at the Telecommunications Research Center Vienna (ftw.) and have been funded by the Austrian Government and the City of Vienna within the competence center program COMET.

REFERENCES

[1] B. Bauer and A. Patrick, "A Human Factors Extension to the Seven-Layer OSI Reference Model," 2004.

[2] A. P. C. da Silva, M. Varela, E. de Souza e Silva, R. M. M. Le ao, and G. Rubino, "Quality assessment of interactive voice applications," *Comput. Netw.*, vol. 52, no. 6, pp. 1179–1192, 2008.

[3] K. De Moor and L. De Marez, *The Challenge of User- and QoE-centric research and product development in today's ICT-environment.*, ser. Innovating for and by users. Office for Official Publications of the European Communities, 2008, pp. 77–90.

[4] S. Egger, "MAVIA - Mediated Audio-Visual Interaction Analysis," Master's thesis, Insitute of Sociology, University of Graz, Graz, Austria, September 2008.

[5] K. R. Gabriel, "The biplot graphic display of matrices with application to principal component analysis," *Biometrika*, no. 58, pp. 453–467, 1971.

[6] F. Hammer, P. Reichl, and A. Raake, "The well-tempered Conversation: Interactivity, Delay and perceptual VoIP Quality," *Communications, 2005. ICC 2005. 2005 IEEE International Conference on*, vol. 1, pp. 244–249 Vol. 1, May 2005.

[7] F. Hammer, "Quality Aspects of Packet-Based Interactive Speech Communication," Ph.D. dissertation, SPSC Lab, Faculty of Electrical and Information Engineering, University of Technology Graz, Graz, Austria, June 2006.

[8] T. Hossfeld, P. Tran-Gia, and M. Fiedler, "Quantification of quality of experience for edge-based applications," in *20th International Teletraffic Congress (ITC20)*, Ottawa, Canada, jun 2007.

[9] International Organization for Standardization, *ISO/IEC 14496-3:1999: Information technology — Coding of audio-visual objects — Part 3: Audio*. Geneva, Switzerland: International Organization for Standardization, 1999, available in English only. [Online]. Available: http://www.iso.ch/cate/d25035.html

[10] International Telecommunication Union, "Terms and definitions related to quality of service and network performance including dependibility," *ITU-T Recommendation E.800*, Aug. 1994.

[11] ——, "Methods for subjective determination of transmission quality," *ITU-T Recommendation P.800*, Aug. 1996.

[12] ——, "Subjective audiovisual quality assessment methods for multimedia applications," *ITU-T Recommendation P.911*, December 1998.

[13] ——, "Subjective video quality assessment methods for multimedia applications," *ITU-T Recommendation P.910*, September 1999.

[14] ——, "Interactive test methods for audiovisual communications," *ITU-T Recommendation P.920*, May 2000.

[15] ——, "H.264: Advanced video coding for generic audiovisual services," *ITU-T Recommendation H.264*, March 2005. [Online]. Available: {http://www.itu.int/rec/recommendation.asp?type=folders\&\#38;lang=e\&\#38;parent=T-REC-H.264}

[16] K. Kilkki, "Quality of experience in communications ecosystem," *Journal of Universal Computer Science*, vol. 14, no. 5, pp. 615–624, 2008, http://www.jucs.org/jucs_14_5/quality_of_experience_in.

[17] W. J. Krzanowski, *Principles of multivariate analysis: a user's perspective*. New York, NY, USA: Oxford University Press, Inc., 1988.

[18] C. E. Osgood, G. J. Suci, and P. H. Tannenbaum, *The Measurement of Meaning*. Urbana: University of Illinois Press, 1957.

[19] P. Reichl, J. Fabini, and C. Happenhofer, M.and Egger, "From QoS to QoX: A Charging Perspective," in *Proceedings of the 18th ITC Specialist Seminar on Quality of Experience*. Blekinge: Blekinge Insitute of Technology, May 2008, pp. 35 –44.

[20] M. Ries, O. Nemethova, and M. Rupp, "Video Quality Estimation for Mobile H.264/AVC Video Streaming," *Journal of Communications*, vol. 3, no. 1, pp. 41–50, 2008.

[21] M. Ries, R. Puglia, T. Tebaldi, O. Nemethova, and M. Rupp, "Audivisual Quality Estimation for Mobile Streaming Services," in *Proceedings of the 2nd International Symposium on Wireless Communications (ISWCS 2005)*, Siena, Italy, September 2005.

[22] J. Short, E. Williams, and B. Christie, *The Social Psychology of Telecommunications*. New York: John Wiley and Sons, 1976.

[23] T. Shrimpton-Smith, B. Zaman, and D. Geerts, "Coupling the Users: The Benefits of paired User testing for IDTV," in *Euro ITV 2006: Proceedings of the 4th Euro iTV Conference*. Athens: Springer - Lecture Notes in Computer Science, May 2006, pp. 214–221.

[24] M. Siller and J. Woods, "Improving quality of experience for multimedia services by qos arbitration on a qoe framework," in *QoE Framework, International Conference on Packet Video*, 2003, pp. 28–29.

[25] VideoLan team, "VLC media player," 2008. [Online]. Available: {http://www.videolan.org/vlc/}

[26] K. Yamagishi and T. Hayashi, "Analysis of psychological factors for quality assessment of interactive multimodal service," in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, ser. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, B. E. Rogowitz, T. N. Pappas, & S. J. Daly, Ed., vol. 5666, Mar. 2005, pp. 130–138.