Audio and Video Channel Impact on Perceived Audio-visual Quality in Different Interactive Contexts

Benjamin Belmudez[1], Sebastian Moeller [2], Blazej Lewcio[3], Alexander Raake[4], Amir Mehmood[5]

Quality and Usability Lab, Deutsche Telekom
Ernst Reuter Platz 7, 10583 Berlin, Germany
[1] benjamin.belmudez@telekom.de
[2] sebastian.moeller@telekom.de
[3] blazej.lewcio@telekom.de
[4] alexander.raake@telekom.de
[5] amir@net.t-labs.tu-berlin.de

***Abstract*—With the advent of audio-visual IP clients, video telephony becomes a realistic option in many application scenarios. In order to guarantee an adequate quality to its users, providers of audio-visual telephony services need to know the impact of the audio and video transmission channel characteristics on perceived Quality of Experience (QoE) in a realistic interactive setting. For this aim, a conversational video telephony experiment was conducted where the audio and video channel settings were adjusted in a controlled way, and participants were asked about the perceived audio, video and overall quality after carrying out a conversation over the audio-visual channel. We analyze the results with respect to the impact the two modalities have, as well as with respect to the impact of the conversation scenario.**

I. Introduction

The interaction between audio quality, video quality and audiovisual quality in a context of multimedia communication has been previously studied in the literature [1] [2] [3] [4] [5] [6]. Several models have been proposed for predicting audio and video impact on the overall quality, one of which was retained for the ITU-T Rec. P.911 [7]. The procedure used to study this interaction was to evaluate separately the effects of video resp. audio with or without the presence of audio resp. video, and then to relate this to the overall audiovisual quality. Most of the experiences used for establishing the models have been carried out in a passive context of listening or viewing only. A study showed that it was possible to map the results in a passive context to an interactive context [8]. However, model coefficients usually depend on the type of degradations applied to the signals, on the experimental conditions, and on the content of the material being evaluated, which is one of the most important point [9]. We designed an experiment in a realistic fashion by using an IP-based client, an interactive context with different conversation scenarios, and with realistic types of degradations. Through this experiment, we investigate on one side the impact of the audio and video channels on the overall quality. We then compare these results obtained in a realistic context to the literature. On the other side, we aimed at analyzing the influence of the scenario and tasks on both media quality evaluation.

MMSP'09, October 5-7, 2009, Rio de Janeiro, Brazil.
978-1-4244-4464-9/09/$25.00

II. Subjective Quality Assessment Experiment

A. Videoconference IP-based client

We used an open-source VoIP client called PJSIP [10] to which a video functionality was added (cf. Fig. 1). The video stream was generated by a camera producing raw video frames, complying with the format YUV422 [11], with a precisely adjustable framerate. For this experiment, we used the VGA format with a framerate of 25 frames per second. We integrated to the client the standard implementation of the video codec MPEG2 included in the open-source library FFmpeg [12]. The video bitstream produced by the decoder was then sliced accordingly to the RFC 2250 [13]. The slices were then encapsulated in RTP packets following the RTP protocol RFC 3550 [14]. The jitter buffer implemented is an adaptative ring buffer capable of receipting fragmented video content. Before the jitter buffer, we implemented a mechanism of synchronization according to the RTCP protocol, so that the interstream delay remains quite low (between 0 and 45 ms), at least under the threshold of detectability which is around 80ms [15]. Once the video frames were decoded, they were directly copied to a video hardware overlay where the conversion YUV to RGB is computed in order to reduce the end-to-end delay. Concerning the audio we used the codec already included in the client, logarithmic PCM according to ITU-T Rec. G.711 [16] for the narrow-band (300-3400 Hz) case, and an open-source wideband codec known as Speex [17], both with the packet loss concealment algorithm deactivated.

B. Test bed

We performed the videoconference tests in two separate rooms where the sound and light properties were adjusted according to ITU-T Rec. P.910 [11]. Both terminals were controlled remotely and any degradations (ex: packet loss)

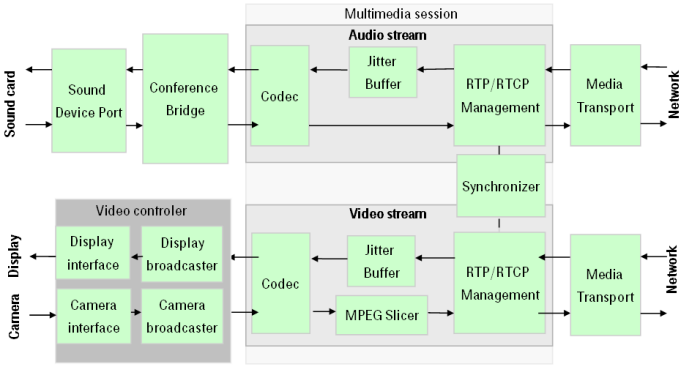

Fig. 1. PJSIP client : modules structure

introduced on the video or audio streams have been realized externally. To do so, we used an external server equipped with a Freebsd tool called Dummynet [18] in order to emulate network-like conditions, such as packet loss, between the sender and receiver clients. The packet loss applied to the streams was controlled separately for both streams, and a network monitoring showed that the effective packet-loss rate applied to the stream was very close to the intended one.

C. Test plan

Three quality levels for both audio and video were predefined by our test set-up, corresponding to three settings of the audio and video coding and transmission characteristics, respectively. For video, the best quality was expected with a bitrate of 1024 kbit/s and with no packet loss or jitter; a medium quality was expected with a bitrate of 512 kbit/s and a packet loss rate of 0.5%, and finally a low quality was expected with a bitrate of 512 kbit/s and 5% of packet loss. All conditions were symmetric so that the test participants experienced the same quality on both ends of the connection. For the audio, the best quality was expected by using the wideband Speex codec with no packet loss, the medium quality with the narrowband G.711 log PCM codec, and finally the low quality with the G.711 and 6% of packet loss. Basically, the degradations applied the original signals were produced either by changing the codec and the packet loss rate for audio, or by changing the bit rate and the packet loss rate for video. Among all the possible combinations of these parameters, we chose 13 different conditions, cf. table I.

Conversations between participants were stimulated by providing two different types of scenarios, one emphasizing more the video aspect and one focusing more on the audio part. The first type of scenario was the "building block scenario", described in to ITU-T Rec. P.920 [19], where one of the participants had an object already made and explains to the other, who has the jumbled pieces, how to build this object (so-called "Lego" scenario). The objects were composed of approximately ten small colored pieces. The second scenario was the "short conversation test scenarios" (SCT) [20] frequently used for conversation tests on speech quality: one of the participants enquires various types of information from the other one who possesses the answers. The duration of each conversation was around 3-4 minutes and the conversation mode was free conversation as in a real videoconference call.

Twenty-four participants carried out the experiment. They were all inexperienced in evaluating audiovisual quality in such a context, but the majority already experienced a videoconference call. They were balanced in gender and aged between 18 and 30 years. Each participant rated every condition once and for each modality. We randomized the order of the conditions, used 5 different SCT scenarios and eight different objects for the lego scenario so that each subject experienced each condition with a different conversation scenario or lego object. The participants used three discrete 11-point scales to rate the audio, video and audiovisual qualities at the end of each session as specified in ITU-T Rec. P.920 [19].

TABLE I
ALLOCATION OF TEST CONDITIONS AND SCENARIOS

	Audio codec Packet loss		
Video Codec Bitrate Packet loss	Speex 0 %	G711 0 %	G711 6 %
MPEG2 1024 kbs 0 %	Lego SCT	Lego SCT	Lego SCT
MPEG2 512 kbs 0.5 %	Lego		Lego
MPEG2 512 kbs 5 %	Lego SCT	Lego	Lego
No Video	SCT		

III. ANALYSIS OF RESULTS

A. Impact of the channel settings

In this section, we make use of the multivariate ANOVA with repeated measures in order to analyze the impact of the different factors on the quality perception. Performing an ANOVA on the audio MOS revealed a statistically significant influence of the audio transmission settings on the perceived audio quality (MOS_A; F=3.5, p=0.032). However, a post hoc test (Scheffe) showed that there was only a difference between the best and the worst audio channel settings. We did not find any interaction between the audio channel settings and the perceived video quality (MOS_V (F=0.183, p=0.673), a result also found in [8]. As expected, we found a statistically significant impact of the video channel settings on MOS_V (F=62.47, $p<0.001$) but no significant interaction between the video channel settings on the MOS_A (F=0.3, p=0.59). A post hoc test (Scheffe) showed that MOS_V differed for all video channel settings.

We found a global effect of the audio and video channel settings on the audiovisual MOS_{AV} (audio: F=8.09, p=0.009; video: F=42.82, $p<0.001$). We notice that the audiovisual

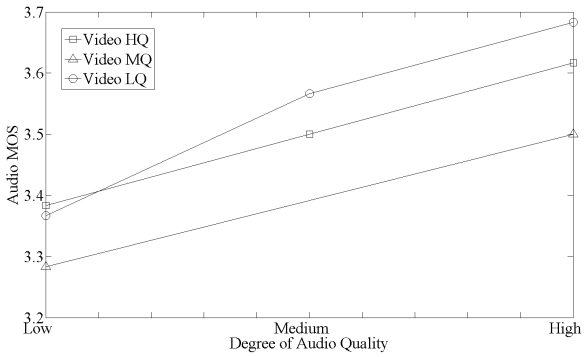

Fig. 2. Audio Quality

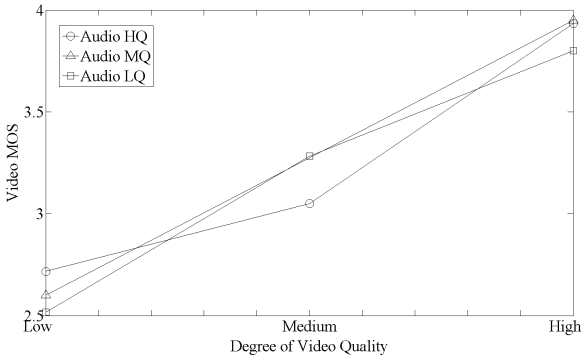

Fig. 3. Video Quality

MOS_{AV} is more significantly influenced by the video than by the audio channel settings as previously found in the literature.

B. Audiovisual mapping

A correlation analysis showed that the audiovisual MOS_{AV} is more correlated to the video MOS_V (Pearson correlation r = 0.82) than to MOS_A (Pearson correlation r = 0.59). The audio and video MOS are poorly correlated (Pearson correlation r = 0.27). As found in the literature, the audiovisual quality is more impacted by the video than by the audio quality.

After performing a non linear fit to our data, we found the following model to relate the audio and video qualities to the audiovisual quality:

$$Q_{AV} = 2.9 + 0.9 \times (Q_A \times Q_V)$$

with a correlation of $R = 0.9502$, and a covered variance of $R^2 = 0.9029$ as well as a root means square error of $RMSE = 11.3627$. This model is similar to the one reported in ITU-T Rec. P.911 [7], however, the value specified in this recommendation were of 1.3 for the constant and of 0.11 for the interaction coefficient. We assume that this model was based on experiments realized mainly in a passive (listening- and viewing-only) context with other types of degradations applied to the audio and video signals than we did in our test. We presume that the variability in the coefficients comes from the interactive aspect of our experiment.

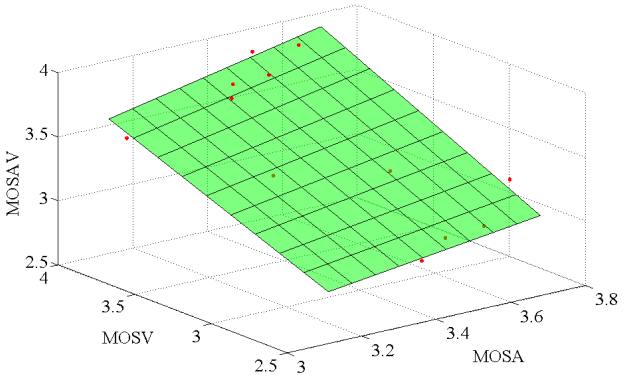

Fig. 4. Mapping MOS_{AV}=f(MOS_A,MOS_V)

We can notice from Fig. 4 that when the video quality is low, the audiovisual quality remains low and is very little affected by the audio quality, which is a consequence of the fact that the video quality is more correlated with the audiovisual quality.

C. Comparison to other models

We have estimated the degradations introduced to the audio and video channels by means of instrumental models for each of the pre-defined quality levels. For the audio quality, we used a wideband extension of the E-model [21] [22] to estimate the R-value indicating an assumed transmission rating:

- High quality Audio R = 113; MOS = 4.27
- Medium quality Audio R = 93; MOS = 3.69
- Low quality Audio R = 38; MOS = 1.58

Instrumental estimations of the video quality were derived using the Structure Similarity index [23]:

- High quality video: MOS = 4.5
- Medium quality video: MOS = 4.05
- Low quality video: MOS = 3.1

When we feed these values to the interaction model trained on our observed data, we get a correlation coefficient between the predicted values and the observed ones of 0.58 with an RMSE of 1.52. This low correlation can be explained by the fact that the test participants were less sensitive to the degradations on the audio channel compared to the the video channel (see Fig 2): they rated the audio on a smaller scale range, and the difference between wideband and narrowband in the case of a clear audio channel was nearly not noticeable. As pointed out in [8], test participants might have rated the audio quality rather on an acceptability criteria.

Different models relating the audiovisual quality to the audio and video quality are found in the literature. The model which has been retained for ITU-T Rec. P.911 [7] is the interaction model. We fed our observed values of MOS_A and MOS_V to these models and computed the correlation between the observed MOS_{AV} and the ones predicted by the models:

- Linear model:
  ITS: $Q_{AV} = -0.677 + 0.217 \times Q_A + 0.888 \times Q_V$
  R = 0.9645, RMSE = 0.5286

- Interaction model:
  P.911: $Q_{AV} = 1.3 + 1.1 \times (Q_A \times Q_V)$
  R = 0.9925, RMSE = 0.9501
- Complete model:
  ITS: $Q_{AV} = 0.517 - 0.0058 \times Q_A + 0.654 \times Q_V + 0.042 \times (Q_A \times Q_V)$
  R = 0.9660, RMSE = 0.5653
  KPN: $Q_{AV} = 1.12 + 0.007 \times Q_A + 0.24 \times Q_V + 0.088 \times (Q_A \times Q_V)$
  R = 0.9857, RMSE = 0.5303

We have to point out that the good correlation comes partly from the fact that the range of our observed audiovisual MOS values is rather centered on the MOS scale.

D. Influence of the scenario

We expect that the type of task (more focused on audio or on video) used during a videoconference call influences the quality perceived by our test participants.

The "Lego" task used in this experiment mainly focuses the participant's attention on the video part because the visual aspect provides more help to build the object than the oral explanation provided by the interlocutor. The short conversation task focuses the attention mostly on the audio because the video channel brings few useful information to perform the task.

In order to study this effect, we have considered the case where the video channel setting is constant (high quality) and let the audio channel setting vary (corresponding to the first line of the Tab I). However, because in the first part of the analysis we only found a difference between the high and low audio levels, only these two cases were considered. We computed an ANOVA on these data to check the scenario influence on the audio MOS_A. We found a small but significant effect of the scenario for the bad audio channel setting (F=6.24, p=0.02), however no significant effect for the good audio channel setting (F=1.2, p=0.29). This result confirms our hypothesis in the sense that in case of major audio degradations, the audio quality will get better ratings if the subject's attention is more focused on the video. However, in case of no audio degradations, the hypothesis is not confirmed.

If we now considered the opposite case where the audio has a good quality and the video a bad quality, we should normally expect to get a better video rating as long as the subject's attention is more focused on the audio part. However, an ANOVA shows no impact of the scenario on the video MOS_V (F=1.3, p=0.27).

Finally we wanted to know whether, in the case were the audio channel setting is good, the addition of video improves the audio quality compared to a pure audio call. A t-test showed that this hypothesis had to be rejected (F=2.3, p=1.4).

IV. Conclusion

The results of this experiment showed that in an interactive context, where the people have to focus on a task and interact with unknown interlocutors, the audio quality is rated in a less differentiated way: this might be an indication of audio quality being mainly judged on an overall acceptability criteria. The test participants were more sensitive to the video degradations and were able to distinguish the differences between the predefined levels of quality. Finally, the type of scenario showed to be a factor which influences significantly the perceived audio quality depending on how much the subjects dedicate their attention to either the audio or the video channel. In a further work, it would be interesting to quantify the effect of the attention focus on the subjective assessment.

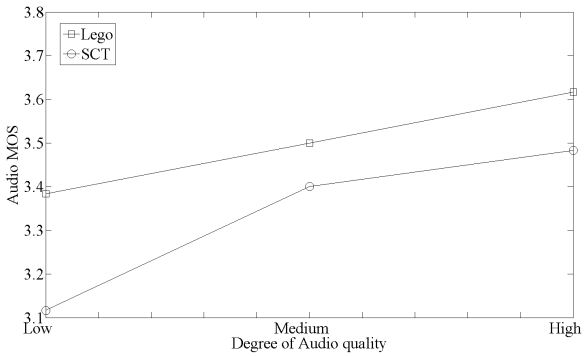

Fig. 5. Influence of the scenario on the MOS Audio for high Video Quality

References

[1] ITU-T Contribution COM 12-19-E, "Relations between audio, video and audiovisual quality," Source: KPN Research, Netherlands, ITU-T Study Group 12 Meeting, Geneva, Feb. 1998.

[2] ITU-T Contribution COM 12-20, "Experimental combined audio/video subjective test method," Source: Bellcore, USA, ITU-T Study Group 12 Meeting, Geneva, Dec. 1993.

[3] ITU-T Contribution COM 12-37, "Extension of combined audio/video quality model," Source: Bellcore, USA, ITU-T Study Group 12 Meeting, Geneva, Sep. 1994.

[4] ITU-T Contribution COM 12-64-E, "Results of an audiovisual desktop video teleconferencing subjective experiment," Source: ITS, USA, ITU-T Study Group 12 Meeting, Geneva, Nov. 1998.

[5] D. S. Hands, "A basic multimedia quality model," *IEEE Trans. Multimedia*, vol. 6, no. 6, pp. 806–816, Dec. 2004.

[6] C. Jones and D. J. Atkinson, "Development of opinion-based audiovisual quality models for desktop video-teleconferencing," in *IEEE International Workshop on Quality of Service (IWQoS'98)*, May 1998, pp. 196–203.

[7] ITU-T Recommendation P.911, "Subjective audiovisual quality assessment methods for multimedia applications," Dec. 1998.

[8] ITU-T Contribution COM 12-61-E, "Study of the influence of experimental context on the relationship between audio, video and audiovisual subjective quality," Source: France Telecom/CNET, France, ITU-T Study Group 12 Meeting, Geneva, Nov. 1998.

[9] A. Watson and M. A. Sasse, "Measuring perceived quality of speech and video in multimedia conferencing applications," in *Proceedings of the sixth ACM international conference on Multimedia*, Sep. 1998.

[10] (2009) Open source SIP stack and media stack for presence, im/instant messaging, and multimedia communication. [Online]. Available: http://www.pjsip.org/

[11] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," Sep. 1998.

[12] (2009) The FFmpeg website. [Online]. Available: http://www.ffmpeg.org/

[13] (1998) RFC 2250 - RTP Payload Format for MPEG1/MPEG2 Video. [Online]. Available: http://www.faqs.org/rfcs/rfc2250.html

[14] (2003) RFC 3550 - RTP: A Transport Protocol for Real-Time Applications. [Online]. Available: http://www.faqs.org/rfcs/rfc3550.html

[15] R. Steinmetz, "Human perception of jitter and media synchronization," *IEEE J. Sel. Areas Commun.*, vol. 14, pp. 61–72, Jan. 1996.

[16] ITU-T Recommendation G.711, “Pulse code modulation (pcm) of voice frequencies,” Nov. 1988.
[17] Speex : A Free Codec for Free Speech. [Online]. Available: http://speex.org/
[18] The FreeBSD Website. [Online]. Available: http://www.freebsd.org/
[19] ITU-T Recommendation P.920, “Interactive test methods for audiovisual communication,” May 2000.
[20] S. Mller, *Assessment and prediction of speech quality in telecommunications*. US-Boston MA: Kluwer Academic Publishers, 2000.
[21] ITU-T Recommendation G.107, “The e-model, a computationel model for use in transmission planning,” May 2000.
[22] ITU-T Recommendation P.834.1, “Extension of the methodology for the derivation of equipment impairment factors from instrumental models for wideband speech codecs,” ITU-T, Geneva, 2009 - To be published.
[23] Z. Wang, H. R. S. A. C. Bovik, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.