Copyright 2004 IEEE. Published in the Proceedings of the 2nd International Symposium on Wireless Communication Systems 2005, 5 – 7 September 2005, Siena, Italy

Audiovisual Quality Estimation for Mobile Streaming Services

Michal Ries, Rachele Puglia**, Tommaso Tebaldi**, Olivia Nemethova*, Markus Rupp**
**Institute for Communications and Radio Frequency Engineering*
Vienna University of Technology
Gusshausstr.25, A-1040 Vienna, Austria
(mries, onemeth, mrupp)@nt.tuwien.ac.at
***Dipartimento di elettronica e informazione, Politecnico di Milano*
via Ponzio, 34/5- 20133 Milan, Italy
tommaso.tebaldi@inwind.it, rachelepuglia@hotmail.com

***Abstract*— 3G mobile terminals supporting audio and video streaming services became reality although the perceptual quality for such low bit rates is limited. To select optimal codec parameters for audio and video, it is important to consider corresponding quality requirements based on human perception. The intention of this paper is to estimate perceptual audiovisual quality for low bit rate videos. We performed subjective perceptual experiments for audio, video and audiovisual quality for music and speech video contents. These experiments were carried out for different audio and video codecs supported by today's mobile terminals. Furthermore, we study the simultaneous influence of music and speech quality on subjective perceptual video quality and mutually compensation effect that occurs in this context. Finally, we propose audiovisual metric for an automated quality estimation.**

I. Introduction

In the last years, several metrics for video quality measurements [1], [2], [3] and for audio quality measurements [4], [5] were investigated. The great majority of the publications assume only a single continuous medium, either audio, or video. Nevertheless, nowadays multimedia systems are becoming more and more important. In UMTS services, it is essential to provide required levels of customer satisfaction. Thus, the priority is to find out a multi-modal model that can be used to predict audiovisual quality. Goal of our research is to estimate the quality of mobile multimedia at the user-level (perceptual quality of service) and to find optimal codecs settings for 3G streaming scenarios. To measure video and audio quality, we have investigated several video and audio objective parameters describing the character of the sequence. In multimedia systems, video and audio modes not only interact, but there is even a synergy of component media (audio and video) [6]. Therefore, we have investigated the perceptual mutually compensation effect that appears in mobile multimedia applications.

II. Test setup for audiovisual quality evaluation

The source material, with different content as is described in Table I, for the tests is composed of music clips, cinema trailer and video call sequences. Screenshots of these sequences can be seen in Figures 1, 2, 3.

All the multimedia clips are nearly eight seconds long [7]. The resolution is QCIF (144×176 pixels). We choose a frame rate of 8 fps as it is used typically for UMTS video streaming. All sequences were encoded with H.263 and MPEG-4 (video tracks) combined with AAC and AMR (audio tracks) as shown in Table II.

Fig. 1. Screenshot of the "video clip" sequence

Fig. 2. Screenshot of the "cinema trailer"

To evaluate the subjective perceptual audiovisual quality, we worked with 20 unpaid test persons. The chosen group ranged different ages (between 17 and 30), sex, education and experience with image processing.
The tests were performed according to the ITU-T Recommendation [7], using absolute category rating (ACR). People have evaluated the audiovisual quality using a five grade MOS scale (1-bad, 2-poor, 3-fair, 4-good, 5-excellent).

According to our experiences with previous psycho-visual

Fig. 3. Screenshot of the "video call" sequence

Title	Video characteristic	Audio characteristic
Cinema trailer	fast scene changes, fast movements	movie soundtrack
Video clip	fast movements, rapid zoom in/zoom out, camera angles changes, fast scene changes	music with singer voice in foreground
Video call	fix camera, low amount of small movements	female monologue

TABLE I
CONTENT OF THE TEST SEQUENCES

experiments the subjective results are slightly different if they are displayed on UMTS handset or PC monitor. To emulate real conditions of the UMTS service, all the sequences were displayed on a UMTS handset Sony Ericsson Z1010. The viewing distance from the phone was not fixed and selected by the test person but we have noticed that all subjects were comfortable to take the phone at a distance of 20-30 cm. Moreover, since one of our intentions is to study the relation between audio quality and audiovisual quality, we have decided to take all the tests with a standard Z1010 stereo headset. With headset it was possible to obtain a higher correlation between the perceived audio quality (MOS_a) and the perceived audiovisual quality (MOS_{av}).
At the beginning of the test session, three training sequences were presented to test persons. Test sequences were presented

		MPEG-4 / H.263 total bit rate		
		56 kbps	75 kbps	105 kbps
AAC audio bit rate	8 kbps	clips 1,2,3		
	16 kbps	clips 1,2,3	clips 1,2,3	
	24 kbps	clips 1,2	clips 1,2,3	clips 1,2,3
	32 kbps		clips 1,2	clips 1,2,3
	48 kbps			clips 1,2
AMR audio bit rate	5,9 kbps	clips 1,2,3	clips 1,2,3	clips 1,2,3
	7,9 kbps	clips 1,2,3	clips 1,2,3	clips 1,2,3
	10,2 kbps	clips 1,2,3		
	12,2 kbps		clips 1,2,3	clips 1,2,3

TABLE II
CODECS AND BIT RATE COMBINATIONS; 36 COMBINATIONS FOR "CINEMA TRAILER" AND "VIDEO CLIP", 30 COMBINATIONS FOR "VIDEO CALL"; 102 TOTAL ENCODED SEQUENCES.

in an arbitrary order, with additional condition that the same sequence (even differently degraded) did not appear in succession. Two rounds of each test were taken. In the further processing of our results we have rejected the sequences which were evaluated with individual variance higher than one. Due to this, we excluded 7% of the tests results.

III. OBJECTIVE QUALITY PARAMETERS

A. Video quality estimation

First, we investigated several objective video parameters, according to [1]. The first two parameters, defined in ANSI standard $sigain$ and $siloss$, measure the gain and the loss in the amount of spatial activity respectively. If the codec operates through an edge sharpening or enhancement, a gain in the spatial activity is obtained, that is an improvement in the video quality of the image. On the other hand when a blurring effect is present in an image, it leads to a loss in the spatial activity. The other two parameters, $hvgain$ and $hvloss$, measure the changes in the orientation of the spatial activity. In particular, $hvloss$ reveals if horizontal and vertical edges suffer of more blurring than diagonal edges. The parameter $hvgain$ reveals if erroneous horizontal and vertical edges are introduced in the form of blocking or tiling distortions. These parameters are calculated over the space - time (S-T) regions. The S-T regions are described by the number of pixels horizontally, vertically and by the time duration of region. In this case one S-T region corresponds to 8x8 pixels over 5 frames.
The ANSI standard in [1] defines seven objective parameters based on spatial, temporal and chrominance properties of video streams. We have simplified this video model by rejecting three parameters, because in our study they showed only small influence on the video quality estimation. Thus, we propose the following video quality (VQ) metric based only on the four most important parameters:

$$VQ = \begin{array}{l} -0.2097 \cdot siloss + 0.5969 \cdot hvloss \\ +0.2483 \cdot hvgain - 2.3416 \cdot sigain|^{0.14}, \end{array} \quad (1)$$

where the positive parameter $sigain$ is clipped at an upper threshold of 0.14 (if $sigain > 0.14$ than set $sigain$ = 0.14) which indicates the maximum improvement of the video quality observed in the encoded sequences [1]. VQ values are between 0 and 1, where 0 indicates the best and 1 the worst quality. To calculate estimation of perceived video quality (MOS_v), VQ is mapped into the range between 1 and 5 by affine transformation:

$$MOS_v = 4(1 - VQ) + 1. \quad (2)$$

B. Audio quality estimation

In our tests we have noticed that the subjective evaluation of speech and music is unequal. Therefore it is necessary to design two independent metrics for speech and for music.

For speech quality evaluation we have adopted the auditory distance (AD) parameter, according to [5]. It measures the

dissimilarities between the original and the compressed speech signals.

Designing an audio quality metric we have noticed a difference in the subjective audio evaluation when the sequences are encoded with the codec AMR or AAC. The maximal correlation between our quality prediction and the measured MOS_a (mean opinion score for audio) is obtained with a translation of the speech audio metric in the two cases of AMR (equation (3)) and AAC coding (equation (4)).

$$MOS_a^{AMR} = -6.996AD^2 + 10.95AD + 1.165 \quad (3)$$

$$MOS_a^{AAC1} = -6.996AD^2 + 10.95AD + 0.370 \quad (4)$$

AD is here normalized between 0 and 1. The reason for this translation is due to the operation of the codecs. The AAC codec utilizes a wider range of frequencies; thus it degrades objectively the signal less than AMR. However, the subjective audio evaluation is higher for AMR. Indeed AMR is a codec designed for speech. It degrades the signal in a way that human's ears do not perceive it. Therefore, although the objective degradation is stronger for AMR, the subjective speech evaluation is higher.

To evaluate the quality of the fit of different metrics for our data, we used a correlation factor defined as follows:

$$r = \frac{\mathbf{x}^T\mathbf{y}}{\sqrt{(\mathbf{x}^T\mathbf{x})(\mathbf{y}^T\mathbf{y})}}. \quad (5)$$

In our case vector $\mathbf{x}$ corresponds to MOS and vector $\mathbf{y}$ to the prediction metric. We obtained 98% correlation for the AMR and 84% correlation for the AAC metric.

For music quality evaluation we have used a more suitable audio metric according to [4]. We have split the original and the encoded streams into 32 ms frames with 50% overlapping [4]. Successively each frame of the two signals has been transformed in the perceptual Bark frequency scale [4]. In this way we obtain both temporal and frequency information of the original and the encoded signals. According to the internal representation of audio signals in the human auditory system, the signals are elaborated through the Zwicker's law [9] that takes into account how the human's ears perceive sound loudness. The first parameter [4], integrated frequency distance (IFD), measures how much the powers of the original and of the encoded signals diverge. The other two parameters, denoted as D_ind and A_ind (disturbance indicators), consider how much the presence of noise and the loss of time-frequency components influence the audio quality.

The resulting music audio metric [4] is a linear combination of the parameters IFD, D_ind and A_ind:

$$MOS_a^{AAC2} = 3.1717 + \frac{4.8809}{IFD} + 0.3562 \cdot A\_ind + 0.0786 \cdot D\_ind. \quad (6)$$

This metric exhibits 91% for AAC codec correlation with the subjective evaluation.

Coefficients	MOS_{av}^{I}	MOS_{av}^{II}	MOS_{av}^{III}	MOS_{av}^{IV}
K	$-1,5025$	$0,9135$	$-0,9222$	$-1,1895$
A	$0,7411$	0	$0,5691$	$0,7126$
V	$0,7380$	0	$0,5064$	$0,5947$
AV	0	$0,2329$	$0,1697$	$0,0677$
A'	0	0	0	$-0,0395$
V'	0	0	0	$-0,0031$
r	$0,8879$	$0,8415$	$0,9106$	$0,9117$

TABLE III
COEFFICIENTS AND CORRELATION OF "FAST MOVEMENT" (CINEMA TRAILER/VIDEO CLIP) MODEL.

Coefficients	MOS_{av}^{I}	MOS_{av}^{II}	MOS_{av}^{III}	MOS_{av}^{IV}
K	$-0,4934$	$0,9987$	$-0,6313$	$0,5723$
A	$0,4327$	0	$0,2144$	$0,2686$
V	$0,5420$	0	$0,0124$	$9,6508$
AV	0	$0,1536$	$0,1184$	$0,2244$
A'	0	0	0	$-0,0940$
V'	0	0	0	$-0,0171$
r	$0,8614$	$0,8915$	$0,9023$	$0,9086$

TABLE IV
COEFFICIENTS AND CORRELATION OF VIDEOCALL MODEL.

IV. METRIC DESIGN AND TEST RESULTS

The obtained MOS depends on the sequence character [8]. For instance, video calls contain lower amount of spatial and temporal information leading to loss of critical ability of the subjective judgment when the media has small video information contents. Therefore, we need one audiovisual model with different coefficients for "video call" and "cinema trailer" or "music clip". We also have to take into account the mutual compensation property and synergy of component media. We investigated the following model:

$$\begin{aligned} MOS_{av} = \quad & K + A \cdot MOS_a + V \cdot MOS_v + AV \cdot MOS_a \cdot \\ & MOS_v + A' \cdot MOS_a^2 + V' \cdot MOS_v^2, \end{aligned} \quad (7)$$

where K is a constant, and A, V, AV, A', V' are weights of MOS_a and/or MOS_v. Inputs of this model are above described audio and video metrics. We were searching for the best trade-off between complexity and correlation with the subjective audiovisual quality. The best result was obtained for a parabolic model MOS_{av}^{III} (Table: III, IV) if we neglect the last two model coefficients (equation: (7)). Figures 4 and 5 show the performance of the proposed model MOS_{av}^{III}.

We can observe (Figure 8) that in the "video call" scenario the MOS_{av} depends more on the audio quality than on the video quality; first because in this case the audio information is more important than the video. Therefore, the obtained MOS_{av} results show that the most suitable combination for the "video call" scenario is a combination of H.263 video codec with AMR audio codec. It makes no sense to use an AAC codec, because AAC needs higher throughput for the same perceptual quality performance of human speech.

For "film trailer" and "video clip" scenarios we obtain better MOS results with MPEG-4 than with H.263 (Figures 6, 7) because these sequences contain a lot of spatial and temporal

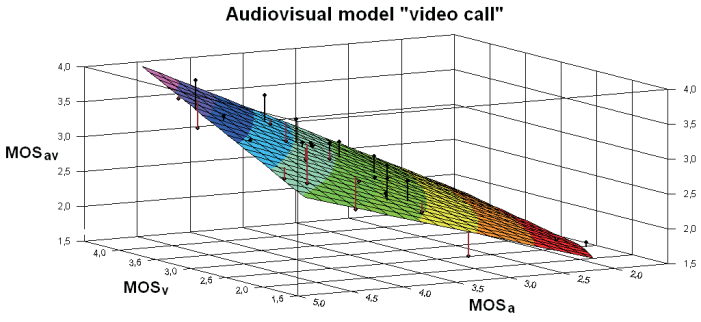

Fig. 4. MOS_{av}^{III} model (surface) and MOS_{av} (measured points) for "video call"

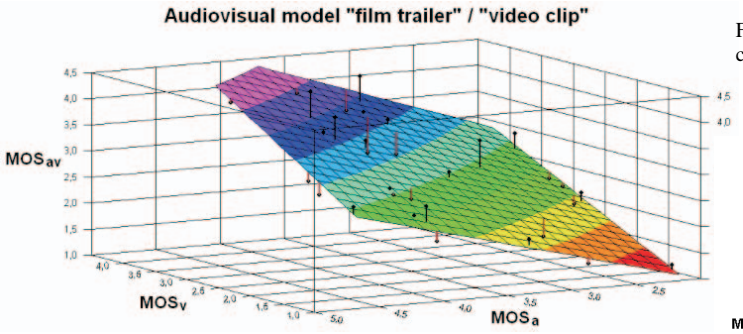

Fig. 5. MOS_{av}^{III} model (surface) and subjective MOS_{av} (measured points) for "cinema trailer"/"video clip"

changes (fast camera movements, scene cuts, zoom out/in). It was evident that the AMR codec cannot achieve sufficient results for music content. For fast movement sequences we can clearly observe the mutually compensation effect. The MOS_{av} is significantly more influenced by audio quality for the lowest bit rate (56 kbps). It is caused by loss of spatial information due to the compression, where higher audio quality can compensate the lower video quality. On the other hand MOS_{av} is not strongly influenced by audio quality for the higher bit rates (75, 105 kbps).

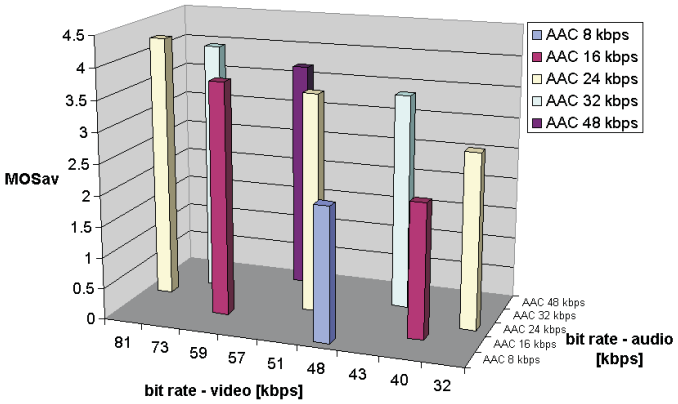

Fig. 6. Subjective MOS_{av} for "cinema trailer" (encoded with MPEG-4 video codec, AAC audio codec)

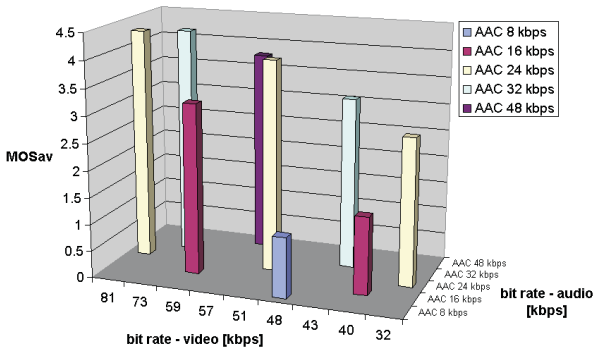

Fig. 7. Subjective MOS_{av} for "video clip" (encoded with MPEG-4 video codec, AAC audio codec)

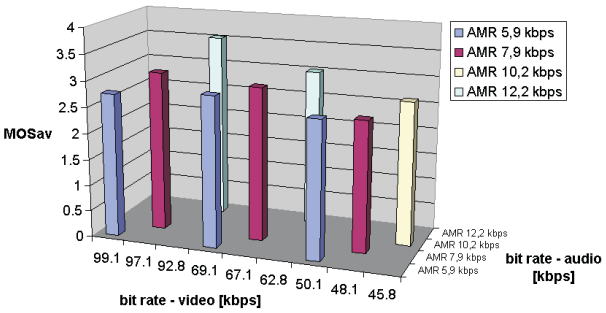

Fig. 8. Subjective MOS_{av} for "video call" (encoded with H.263 video codec, AMR audio codec)

V. Conclusion

The scope of the paper is to estimate audiovisual quality and to propose a metric for mobile applications. We observe that audio quality, video quality and sequence character are important factors to determine the overall subjective perceived quality. A mutual compensation property of audio and video can also be clearly seen from our results. We propose an audiovisual metric for automated prediction of audiovisual quality, which is suitable for mobile streaming services. Its correlation with tests results is 91% in the fast movement scenario, and 90% in the "video call" scenario. Our future work will be to determine the relation between information entropy and subjective judgment.

VI. Acknowledgment

The authors would like to thank mobilkom austria AG&Co KG 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&Co KG.

REFERENCES

[1] ANSI T1.801.03, "American National Standard for Telecommunications - Digital transport of one-way video signals. Parameters for objective performance assessment," American National Standars Institute, 2003.

[2] S. Winkler, Ch. Faller: "Maximizing audiovisual quality at low bitrates," in Proc. Workshop on Video Processing and Quality Metrics for Consumer Electronics, Scottsdale, AZ, 2005.

[3] A.A. Webster, C.T. Jones, M.H. Pinson, S.D. Voran, S. Wolf, "An objective video quality assessment system based on human perception," in Proc. SPIE Human Vision, Processing and digital display, vol. 1913, pp. 15-26, San Jose, CA, 1993.

[4] ITU-T Recommendation P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow band telephone networks and speech codecs," International Telecommunication Union, 2001.

[5] S. Voran, "Objective Estimation of Perceived Speech Quality. Part II: Evaluation of the Measuring Normalizing Block Technique," in Proc. IEEE Transactions on speech and audio, vol. 7, no. 4, pp. 385-390, 1999.

[6] S. Tasaka, Y. Ishibashi, "Mutually Compensatory Property of Multimedia QoS," IEEE Transactions, Nagoya Institute of Technology, Nagoya, Japan, 2002.

[7] ITU-T Recommendation P.911, "Subjective audiovisual quality assessment methods for multimedia application," International Telecommunication Union, 1998.

[8] O. Nemethova , M. Ries, E. Siffel, M. Rupp, "Quality Assessment for H.264 Coded Low-Rate and low-Resolution Video Sequences," Proc. of Conf. on Internet and Inf. Technologies (CIIT), St. Thomas, US Virgin Islands, pp. 136-140, 2004.

[9] E. Zwicker, R. Feldtkeller, "Das Ohr als Nachrichtenempfänger," S. Hirzel Verlag, Stuttgart, 1967.