MAXIMIZING AUDIOVISUAL QUALITY AT LOW BITRATES

Stefan Winkler

Genista Corporation
Rue du Theâtre 5
1820 Montreux, Switzerland

stefan.winkler@genista.com

Christof Faller

Audiovisual Communications Lab
Ecole Polytechnique Fédérale de Lausanne
1015 Lausanne, Switzerland

christof.faller@epfl.ch

ABSTRACT

We carried out a number of subjective experiments for audiovisual, audio-only, and video-only quality assessment. We selected content and encoding parameters at very low bitrates that are typical of mobile applications. Using these data, we explore the influence of video codecs and frame rate as well as audio channels and sampling rate on quality. Finally, the optimal trade-off between bits allocated to audio and video inside a bitstream is investigated.

1. INTRODUCTION

Video quality (VQ) assessment [15, 16] has become rather well established by now, as evidenced by the number of research publications and products available, as well as the collaborative efforts of the Video Quality Experts Group (VQEG) and recent standards for TV [8]. Speech and audio quality (AQ) assessment techniques have an even longer history. Speech and audio quality metrics have been standardized as PESQ [11] and PEAQ [7], respectively.

Audiovisual quality (AVQ), however, is an entirely different matter. There have been a few studies in the past [2, 13], but little work has been done at low bitrates. Therefore, we designed a number of subjective experiments using content, codecs, and bitrates typical of emerging mobile applications. Based on the results of these tests, an analysis of AV coding parameters is presented in this paper. More details on the experiments, the interactions between audio and video quality, and an evaluation of quality metric predictions on the data can be found in [17].

The paper is organized as follows. Section 2 describes the experimental setup in terms of source material, test conditions, and subjective assessment. The influence of video codecs and frame rate on video quality is discussed in Section 3. The effect of the number of audio channels (mono or two-channel stereo) and sampling rate on audio quality is discussed in Section 4. Finally, the optimal bit budget allocation trade-off between audio and video is investigated in Section 5.

2. EXPERIMENTAL SETUP

2.1. AV Source Clips

The content of the source clips and the range of coding complexity was chosen to be representative of a typical scenario for watching video on a mobile device. The source material comprises 6 short clips of about 8 seconds each. The video and audio content of these scenes is summarized in Table 1. The video source material was originally in TV format; for our tests we de-interlaced and downsampled it to QCIF frame size (176x144). The audio source material was 16-bit PCM stereo sampled at 48 kHz.

2.2. Test Material

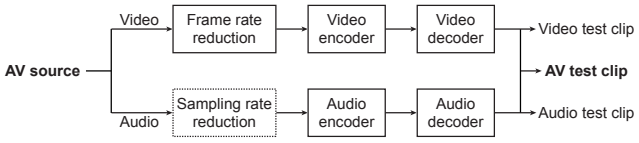

Fig. 1. Encoding Setup.

For the video track, we chose the MPEG-4 AVC/H.264 [4, 10] coding standard (baseline profile) as well as traditional MPEG-4 part 2 [5] and H.263 [9]. The JM reference software* version 8.5 was used for H.264 encoding; QuickTime Pro version 6.5 was used for H.263 and MPEG-4 encoding. Before encoding, the frame rate of the source clips was reduced to 8 fps or 15 fps using VirtualDub.†

For the audio track, we chose the MPEG-4 AAC-LC coding standard [6]. QuickTime Pro was again used for encoding, with the "recommended" sampling rate for each target bitrate (i.e. the sampling rate reduction was carried out internally by the encoder).

* The JM reference software is available at http://bs.hhi.de/~suehring/tml/

† VirtualDub is available at http://www.virtualdub.org/

Table 1. Video and audio content of test scenes.

Scene	Name	Video	Audio	Duration
A	Buildings	slow horizontal pan across a city skyline, followed by a vertical pan up a building facade	orchestral background music	7.48 sec.
B	Conversation	camera switching between head-and-shoulders shots of a woman and a man talking	male and female voices	8.36 sec.
C	Football	American football scene from VQEG [14]; high motion	crowd cheering and chanting; female commentator	7.60 sec.
D	Music video	music video clip; high motion	rock music with vocals	8.08 sec.
E	Trailer 1	action movie trailer; scene cuts and high motion	theme music and voice-over	8.84 sec.
F	Trailer 2	romance movie trailer with credits; scene cuts	theme music and voice-over	8.08 sec.

Table 2. Video test conditions.

Condition	Codec	Frame rate	Bitrate
1	H.264	8 fps	24 kb/s
2	H.264	8 fps	32 kb/s
3	H.264	8 fps	40 kb/s
4	H.264	8 fps	48 kb/s
5	H.263	8 fps	48 kb/s
6	MPEG-4	8 fps	48 kb/s
7	H.264	15 fps	24 kb/s
8	H.264	15 fps	32 kb/s
9	H.264	15 fps	40 kb/s
10	H.264	15 fps	48 kb/s

Table 3. Audio test conditions.

Condition	Channels	Sampling rate	Bitrate
1	mono	8 kHz	8 kb/s
2	mono	16 kHz	16 kb/s
3	mono	22 kHz	24 kb/s
4	mono	32 kHz	32 kb/s
5	mono	22 kHz	32 kb/s
6	stereo	22 kHz	32 kb/s
7	stereo	16 kHz	32 kb/s

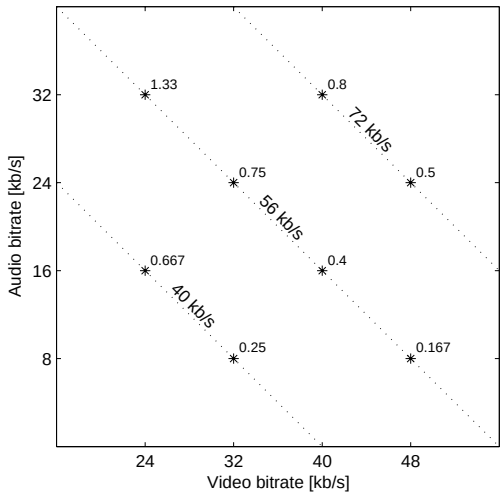

Fig. 2. Audiovisual test conditions. The stars denote the video and audio bitrate combinations used in the test. The diagonal dotted lines connect points with the same total data rate. Every point is labeled with its A/V bitrate ratio.

Video conditions 1–4 from Table 2 were then combined with audio conditions 1–4 from Table 3 for a total of 8 audiovisual test conditions as illustrated in Figure 2. Of particular interest is a total data rate of 56 kb/s, which can be transmitted over a typical 64 kb/s wireless link (leaving room for packetization overhead).

2.3. Subjective Assessment

The laboratory set-up follows ITU-T Rec. P.910 [12]. We use the Absolute Category Rating (ACR) methodology from this recommendation. With ACR, the test clips are viewed one at a time and rated independently on a discrete 11-level scale from "bad" (0) to "excellent" (10).

6 female and 18 male subjects aged 25–36 years participated in the subjective test. The test consisted of one session of about 40 minutes, including a short training session, which preceded the actual test. The training comprised three audiovisual clips demonstrating the extremes of the expected audiovisual, audio and video quality ranges. The subjects were allowed to adjust the volume to a comfortable level during the training session.

The actual test consisted of three parts for audiovisual, audio-only (blank screen), and video-only (no accompanying audio) evaluation using the material described above. Subjects were asked to rate the quality of the presentation in each case. The order of the clips was randomized individually for each subject.

The monitor used in the subjective assessments was a Dell 1703FP 17" LCD screen. For our test material, we found subjects to be comfortable at a viewing distance of around 8 times the height of the video picture, which corresponds to about 30-40 cm in our setup. For audio playback, an external D/A converter (Emagic EMI A26) was

connected to the PC. High quality headphones (Sennheiser HD 600) were directly connected to the D/A converter. The test was conducted in a sound insulated room.

Genista's *QualiView* software was used for the playback of the sequences. It reads the decoded test clips stored in uncompressed AVI format and plays them on the PC. After each clip, the voting dialog is presented on the screen, and the rating entered by the subject is recorded.

3. VIDEO CODECS AND FRAME RATE

3.1. Codecs

Codec selection was principally determined by the 3GPP* file format as defined in [1]. It is of particular interest for packet-switched video streaming in 3G networks. However, version 6.5 of QuickTime Pro only supports H.263 and MPEG-4 part 2 for the video track. Therefore, we were forced to use the JM reference encoder for H.264, even though its output is not necessarily 3GPP-compliant.

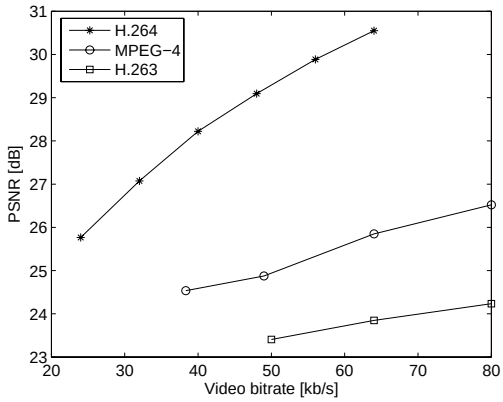

Fig. 3. Peak signal-to-noise ratio (PSNR) of the *music* clip as a function of bitrate for H.264 (stars), MPEG-4 (circles) and H.263 (squares).

Unfortunately, the QuickTime encoders for MPEG-4 and especially H.263 did not produce substantial quality variations within the bitrate range of interest. Furthermore, they did not achieve the target bitrates at the low end of the range. This is demonstrated in Figure 3. Viewers would have been unable to discern the quality of the different test clips. The H.264 JM reference encoder does not have these problems.† We therefore decided to use H.264 for almost all test conditions.

To compare the performance of the three codecs in terms of perceived quality, we now look at video test conditions 4–6 from Table 2. The VQ mean opinion scores (MOS) shown in Figure 4 are further evidence that H.264 clearly outperforms the two other codecs. The only exception is perhaps trailer 2, in which H.264 has a hard time coping with the scene cuts. No clear winner can be determined between H.263 and MPEG-4.

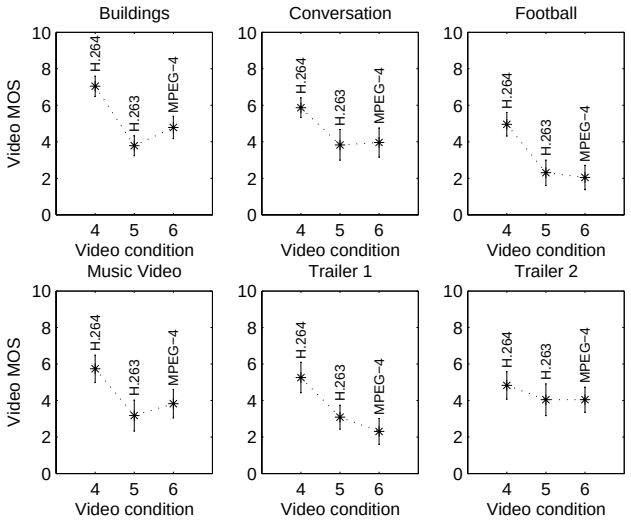

Fig. 4. Video MOS comparison for different codecs. The error bars indicate the 95%-confidence intervals.

We also carried out paired t-tests of the null hypothesis that the three possible codec pairs come from equal means. The resulting *p*-values, shown in Table 4, confirm that the QuickTime H.263 and MPEG-4 codecs are not significantly different in visual quality, while H.264 is significantly better than both.

Table 4. *p*-values of t-test ($\alpha = 0.05$) comparing pairs of samples from different codecs.

Codecs	*p*-value
H.263 vs. MPEG-4	0.442
H.263 vs. H.264	0
H.264 vs. MPEG-4	0

3.2. Frame Rate

Video test conditions 1–4 and 7–10 from Table 2 differ only in frame rate (8 fps and 15 fps, respectively). The VQ MOS and 95%-confidence intervals for these conditions are shown in Figure 5. In most cases, the perceived video quality is markedly better for 8 fps than for 15 fps at the same bitrate. The difference is least pronounced for the low-motion "conversation" scene, but interestingly also for the two high-motion trailers, which contain the most scene cuts.

* 3rd Generation Partnership Project, see http://www.3gpp.org.

† The absolute quality gain using H.264 is also evident from Figure 3; however, it is worth noting that the H.264 reference encoder implementation is almost 100 times slower than the two QuickTime encoders.

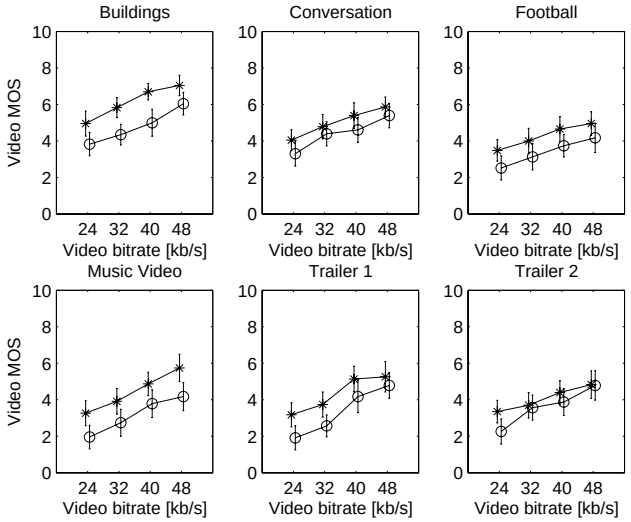

Fig. 5. Video MOS as function of bitrate at 8 fps (stars) and 15 fps (circles). The error bars indicate the 95%-confidence intervals.

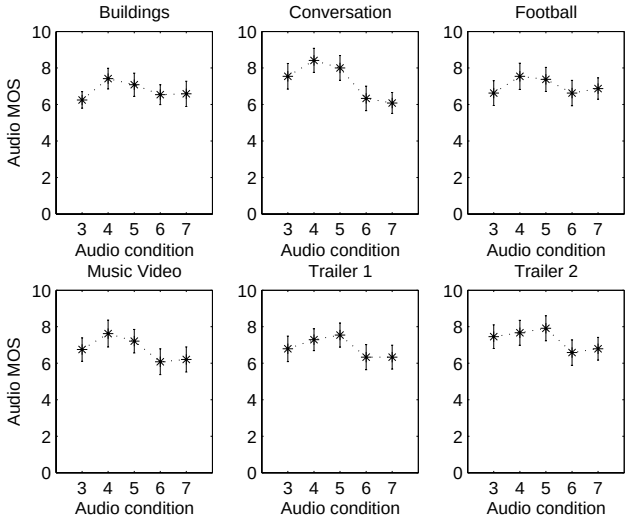

Fig. 6. Audio MOS comparison for mono/stereo, different sampling rates, and two bitrates (see Table 3 for details). The error bars indicate the 95%-confidence intervals.

Again we carried out paired t-tests of the null hypothesis that 8 fps and 15 fps come from equal means at each bitrate. The resulting *p*-values, shown in Table 5, lead to the rejection of the null hypotheses, thus indicating that a frame rate of 8 fps results in significantly higher video quality than 15 fps at a given bitrate.

Table 5. *p*-values of t-test ($\alpha = 0.05$) comparing pairs of samples with frame rates of 8 fps and 15 fps.

Bitrate	***p*-value**
24 kb/s	$2.66 \cdot 10^{-14}$
32 kb/s	$1.69 \cdot 10^{-10}$
40 kb/s	$9.72 \cdot 10^{-13}$
48 kb/s	$1.84 \cdot 10^{-6}$

4. AUDIO CHANNELS AND SAMPLING RATE

We now study the impact of various audio coding parameters on the perceived audio quality. For this purpose we had included four audio test conditions with the same bitrate (32 kb/s) but varying parameters in the test (conditions 4–7 in Table 3). We also include condition 3 in this analysis, as it only differs from condition 5 in bitrate. The question is how the audio bandwidth (directly related to audio coder sampling rate) and the number of audio channels (mono or two-channel stereo) affect the audio quality.

Figure 6 shows the AQ MOS for the relevant audio test conditions. For all six clips, the perceived audio quality is higher when mono audio coding is used (conditions 4&5) than when stereo audio coding is used (conditions 6&7).

t-tests were carried out on all 10 possible condition pairs. The resulting *p*-values are shown in Table 6. The only condition pairs that are not significantly different are 4&5 and 6&7. This implies that changing the audio sampling rate has no significant effect on quality, regardless of whether mono or stereo is used. However, mono encoding is significantly better than stereo encoding in all four cases. In fact, even 24 kb/s mono is better than 32 kb/s stereo (while 32 kb/s mono is always better than 24 kb/s mono).

Table 6. *p*-values of t-test ($\alpha = 0.05$) comparing pairs of samples with different coding parameters.

Conditions	***p*-value**	**Conditions**	***p*-value**
4 vs. 5	0.233	3 vs. 4	$4.39 \cdot 10^{-8}$
4 vs. 6	$1.48 \cdot 10^{-12}$	3 vs. 5	$3.17 \cdot 10^{-3}$
4 vs. 7	$1.25 \cdot 10^{-9}$	3 vs. 6	$2.48 \cdot 10^{-5}$
5 vs. 6	$8.24 \cdot 10^{-11}$	3 vs. 7	$2.18 \cdot 10^{-2}$
5 vs. 7	$8.04 \cdot 10^{-8}$		
6 vs. 7	0.700		

It is not surprising that the non-parametric transform coder AAC-LC yields better quality for mono considering the low audio bitrates in our test. The audio bandwidth available for two stereo channels is much lower than for a single mono channel when both are coded at the same bitrate. Therefore, the stereo audio appears more distorted, and subjects prefer mono audio with less degradation. The recently standardized High Efficiency AAC (HE-AAC) [3] avoids the issue that at low bitrates only a low audio bandwidth can be afforded for stereo. Using HE-AAC, stereo may be preferred even at these bitrates.

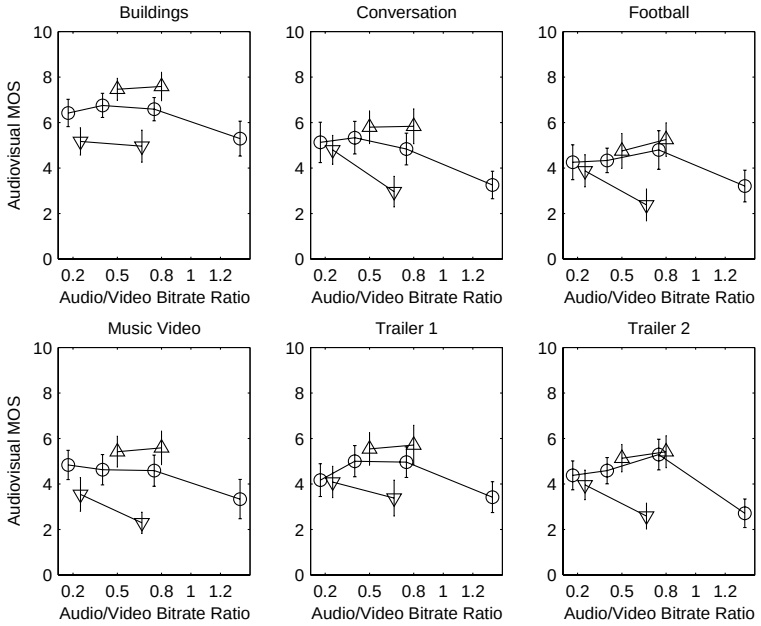

Fig. 7. Audiovisual quality as a function of audio/video bitrate ratio at total bitrates of 56 kb/s (circles), 40 kb/s (downward-pointing triangles) and 72 kb/s (upward-pointing triangles). Refer to Figure 2 for the exact A/V bitrate ratios of each data point. The error bars indicate the 95%-confidence intervals.

5. AUDIO-VIDEO BIT BUDGET ALLOCATION

The AVQ MOS values for the six clips are shown as a function of the audio/video bitrate ratio (cf. Figure 2) in Figure 7. Focusing first on 56 kb/s (circles), where we have the most sample points, we note the following. The audio/video bitrate ratio with the highest AVQ depends to a large extent on the specific clip. For five out of the six clips the optimum ratio is in the center range around 16/40–24/32.

In the visually most complex scenes, e.g. "football" and the two trailers, a high relative audio bitrate seems to produce the best overall quality, whereas the less demanding scenes ("buildings" and "conversation") benefit from a high video bitrate. This seems counter-intuitive, since one would expect that complex scenes need more bits for the video. On the other hand, a bitrate increase may result only in a negligible improvement in video quality for such a scene, while an increase by the same amount can significantly improve the audio. This could explain why the bits may in fact be better spent on the audio when the video is very complex.

If the total bitrate budget is reduced to 40 kb/s, the optimum audio/video bitrate ratio decreases, i.e. relatively more bits should be allocated to the video. The opposite trend can be observed when the total bitrate increases to 72 kb/s. In this case, the optimum appears to shifts to the right, i.e. a higher relative bitrate for the audio seems favorable. Unfortunately, our test does not include enough data points to draw firm conclusions on this matter.

6. CONCLUSIONS

We investigated the influence of various encoding parameters on audio, video and audiovisual quality for a number of test scenes encoded at very low bitrates (24-48 kb/s for video and 8-32 kb/s for audio). The main findings can be summarized as follows:

- The QuickTime Pro encoders for H.263 and MPEG-4 have very similar quality. H.264 (JM reference software) clearly outperforms both of them.
- Encoding at 8 fps produces higher-quality video than 15 fps at the same bitrate.
- Choosing mono instead of stereo produces higher-quality audio. Changing the sampling rate or even the bitrate has much less effect on the resulting audio quality.
- The optimum audio/video bitrate allocation depends on scene complexity. The more complex the scene and the higher the total bitrate budget, the more bits should be allocated to audio. At a total bitrate of 56 kb/s, the optimum is roughly 32-40 kb/s for video and 16-24 kb/s for audio.

However, we would like caution against extrapolating all of these conclusions to much higher bitrates or other codecs without further testing.

7. ACKNOWLEDGMENTS

We acknowledge the support of Prof. Sabine Süsstrunk at EPFL's Audiovisual Communications Lab, who provided some of the testing facilities for the subjective experiments. We also thank the people who participated in our tests as subjects.

8. REFERENCES

[1] 3GPP Technical Specification 26.244: "Transparent end-to-end packet switched streaming service (PSS); 3GPP file format (3GP)(Release 6)." 3rd Generation Partnership Project, 2004.

[2] J. G. Beerends, F. E. de Caluwe: "The influence of video quality on perceived audio quality and vice versa." *J. Audio Eng. Soc.* **47**(5):355–362, 1999.

[3] M. Dietz, L. Liljeryd, K. Kjörling, O. Kunz: "Spectral band replication – a novel approach in audio coding." in *Proc. AES Convention*, Munich, Germany, 2002.

[4] ISO/IEC 14496-10: "Coding of audio-visual objects – Part 10: Advanced video coding." International Organization for Standardization, Geneva, Switzerland, 2004.

[5] ISO/IEC 14496-2: "Coding of audio-visual objects – Part 2: Visual." International Organization for Standardization, Geneva, Switzerland, 2004.

[6] ISO/IEC 14496-3: "Coding of audio-visual objects – Part 3: Audio." International Organization for Standardization, Geneva, Switzerland, 2001.

[7] ITU-R Recommendation BS.1387-1: "Method for objective measurements of perceived audio quality (PEAQ)." International Telecommunication Union, Geneva, Switzerland, 2001.

[8] ITU-R Recommendation BT.1683: "Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference." International Telecommunication Union, Geneva, Switzerland, 2004.

[9] ITU-T Recommendation H.263: "Video coding for low bit rate communication." International Telecommunication Union, Geneva, Switzerland, 1998.

[10] ITU-T Recommendation H.264: "Advanced video coding for generic audiovisual services." International Telecommunication Union, Geneva, Switzerland, 2003.

[11] 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, Geneva, Switzerland, 2001.

[12] ITU-T Recommendation P.910: "Subjective video quality assessment methods for multimedia applications." International Telecommunication Union, Geneva, Switzerland, 1996.

[13] A. Kohlrausch, S. van der Par: "Auditory-visual interaction: From fundamental research in cognitive psychology to (possible) applications." in *Proc. SPIE*, vol. 3644, pp. 34–44, San Jose, CA, 1999.

[14] VQEG: "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment." 2000, available at http://www.vqeg.org/.

[15] S. Winkler: *Digital Video Quality – Vision Models and Metrics*. John Wiley & Sons, 2005.

[16] S. Winkler, F. Dufaux: "Video quality evaluation for mobile applications." in *Proc. SPIE*, vol. 5150, pp. 593–603, Lugano, Switzerland, 2003.

[17] S. Winkler, C. Faller: "Audiovisual quality evaluation of low-bitrate video." in *Proc. SPIE*, vol. 5666, San Jose, CA, 2005.