Of MOS and men: bridging the gap between objective and subjective quality measurements in mobile TV

T.C.M de Koning [a], P. Veldhoven [a], H. Knoche [*b], R.E. Kooij [a,c]
[a]TNO ICT, PO Box 5050, 2600 GB Delft, The Netherlands
[b]Dept. of Computer Science, University College London, London, WC1E 6BT UK
[c]Delft University of Technology, Dept. of Electrical Engineering, Mathematics and Computer Science, Mekelweg 4, 2628 CD Delft, The Netherlands

ABSTRACT

In this paper we explore the relation between subjective and objective measures of video quality. We computed objective MOS values from video clips using the video quality measuring tool VQM and compared it to the clips' subjective Acceptability scores. Using the ITU defined mapping (M2G) from MOS to binary Good or Better (GoB) values, we compared the M2G translated values to the clips' subjective Acceptability scores at various encoding bitrates (32-224kbps) and sizes (120x90, 168x126, 208x156 and 240x180). The results show that in the domain of mobile TV the ITU mapping M2G represents a serious overestimation of Acceptability.
The mapping M2A, between MOS and Acceptability, that we suggest provides a significant improvement of 76% in the root mean square error (RMSE) over M2G. We show that Acceptability depended on more than just the visual quality and that both content type and size are essential to provide accurate estimates of Acceptability in the field of mobile TV. We illustrate this gain in Acceptability predictions for the popular content type football (soccer). In terms of RMSE our content dependent mapping (M2Af) yielded an improvement of 39% over M2A. Future research will validate the predictive power of our suggested mapping on other video material.

Keywords: Mobile TV, Acceptability, MOS, VQM, subjective quality, objective quality

1. INTRODUCTION

The numbers of mobile TV users are expected to rise from the 3.6 million in 2006 to 120 million by 2010. So far uptake is lagging and service providers are looking into ways to improve the quality of experience of current services. Video quality is one of the defining factors of the experience of mobile TV. Service providers aim at maximizing the user experience while minimizing resource usage in the constrained environment of the mobile domain. Objective video quality measurements like PSNR and VQM can aide in configuring optimal encoding settings on a per clip basis. They algorithmically predict the quality perceived by a human observer and include mappings such that their results can be understood in terms of Mean Opinion Scores (MOS). MOS currently represent the lingua franca in subjective audio and video quality assessment in which test participants provide their ratings on standardized scales containing labels such as Excellent, Good, Fair, Poor, Bad in lab based tests. Shortcomings in MOS have sparked the development of new subjective measures. Acceptability is a binary subjective measure in which participants state whether they find a given quality acceptable or unacceptable for a given purpose or service, see [17] for more details. A number of studies has successfully employed acceptability, e.g. [15], [6] and [13]. A major advantage of the measure acceptability is that it bears a direct relation to market acceptance and it easily translates into utility curves for service providers.

In the context of voice quality the ITU defined a mapping (M2G) which relates MOS values to the binary measure Good or Better (GoB), see ITU-T G.107 [11]. To our knowledge M2G has not been validated with subjective ratings of video quality. Unfortunately, the reference provides no detailed rationale behind the mapping M2G and the relationship between Acceptability and GoB remains equally unknown. From the verbal labels one could argue that an acceptable quality would equate to a MOS quality score between the labels Fair and Good. Even if we knew the relative position of acceptable quality on the MOS scale the shape of the curves describing Acceptability and GoB along a continuum of e.g.

* h.knoche@cs.ucl.ac.uk

Multimedia on Mobile Devices 2007, edited by Reiner Creutzburg, Jarmo Takala, Jianfei Cai,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 6507, 65070P,

encoding bitrates might still differ. To better understand these two seemingly similar concepts and to validate the M2G mapping we used the video clips and their acceptability scores from a large scale mobile TV study [15]. A total of 128 participants provided acceptability scores of the video quality of audio-visual clips presented on a mobile device across a range of four content types (football, news, music, and animation), seven encoding bitrates (32-224kbps) and four sizes (120x90, 168x126, 208x156 and 240x180). We used VQM to compute the objective MOS values (MOS_{VQM}) of a total of 448 video clips each 20 seconds long. We then applied the current ITU mapping M2G to the MOS_{VQM} values to validate how well M2G would predict the video clips' acceptability scores.

Our results show that in the domain of mobile TV the current mapping M2G applied to MOS_{VQM} results in a serious overestimation of acceptability. The main contribution of this paper is a mapping from MOS in its most important range (from poor to excellent quality) to acceptability (M2A), which provides a significant improvement of 76 % in the root mean square error (RMSE) over M2G. So far, objective video quality measures are typically content independent, i.e. they do not discriminate between, e.g., sports and cartoons, and do not consider the size of the displayed content. We show that acceptability depends on more than just the visual quality and that both content type and size are essential to provide accurate estimates of acceptability in the field of mobile TV. We illustrate this gain in perceived quality predictions for the popular content type football (soccer). In terms of RMSE our content dependent mapping ($M2A_f$) yielded an improvement of 39% over M2A. Another interesting finding is that we can substantially improve the mapping if we take into account the size of the video display. At present VQM does not consider the factor size into its calculation of MOS_{VQM} values. Further research will need to validate the predictive power of our suggested mappings and extend it. As reported in [15] the acceptability of the video quality depended partly on the audio quality that accompanied the clip whereas the MOS_{VQM} values are purely based on the visual content of a clip.

The rest of this paper is organized as follows. In section 2 we briefly discuss the subjective and objective video quality measures and we also give the acceptability results reported in [15]. In section 3 we discuss the data set of 448 video clips and how we resized the major part of the data set such that we can determine MOS values of all video clips. In section 4 we report the obtained MOS values. In section 5 we compare the ITU-mapping M2G with our own content-independent mapping M2A. We also show in the same section that M2A can be further improved by taking the content type into account. In section 6 we show that yet another improvement can be realized by also taking the size of the video into account. Finally, concluding remarks and suggestions for future research can be found in section 7.

2. BACKGROUND

Quality assessment represents an important measure to simulate, calibrate and operate multimedia delivery systems. Objective quality assessment approaches are computer based and are therefore automated, low cost techniques to obtain measures that quantify the fidelity/quality of content without the need for human judgment. These make them particularly suitable for system planning and optimization. Some of these objective measures employ human perceptual models to derive results that better match the way humans would rate the quality.

Mean opinion scores (MOS) represent an established and widely used method for subjective quality assessment of audio and as of lately video content. The International Telecommunication Union (ITU) standardized a five-grade scale to rate transmission quality (of audio quality) [8]. Participants rate the quality of short audio on a scale with the labels Excellent, Good, Fair, Poor and Bad, which are mapped to values from five to one. For service providers the range from Excellent to Fair is most important. The resulting averaged scores are called Mean Opinion Scores (MOS). This approach has been extended to the quality assessment of video clips in [10]. Many of the objective video quality measures like PSNR, VQM, and SSIM have existent mappings to MOS values.

Some research has shown that the intervals on the MOS scale are not conceptually equal in size [1] which poses problems for the computation of means. Because of cultural differences in using the scale and the vocabulary used for the labels the ITU scale it has been argued that it does not even represent an internationally ordinal scale [18]. Criticism has led to the introduction of other measures for audio-visual quality assessment. Binary measures like watchability [3] and acceptability [6], [13] in which participants discriminate between only two states, have been successfully used in a number of studies.

Acceptability as introduced in [17] is based on just noticeable differences a concept that dates back to the work of Fechner in psychophysics done in the eighteen hundreds [5]. The "Method of Limits" approach was geared towards the detection of thresholds in human perception. The intensity of a stimulus was increased in discrete steps until it was just detectable to a human subject. The subject would indicate the detection of a stimulus by a binary YES/NO response. The

method of limits has been adopted for acceptability ratings. Participants are asked to state when they think that the quality of a video content is unacceptable or acceptable for an application or service in a context under study [17]. During these tests participants may switch back and forth between these two opinions continuously as a stimulus is being presented that exhibits different quality levels. Many of the objective video quality measures like PSNR, VQM, SSIM include or have been extended by a mapping to MOS.

In the context of voice quality the ITU defined a mapping (M2G) which relates MOS values to the binary measure Good or Better (GoB), see ITU-T G.107 [11]. To our knowledge this mapping M2G has not been validated with subjective ratings of video quality. We only found some research that compared subjective ratings to PSNR scores e.g. [7],[14].

Previous research has shown that the acceptability of video quality depends besides the video encoding bit rate on the displayed content and size [15]. We present the summary of the acceptability scores averaged across the four content types of that study (News, Football, Music and Animation) in Figure 1. Most objective quality measurements do neither consider size nor content types into their calculations of video quality.

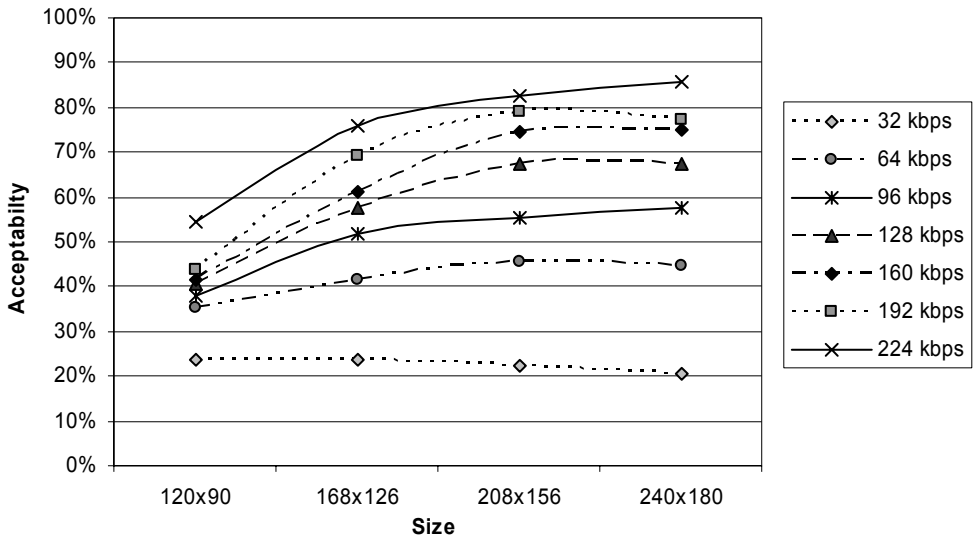

Figure 1: Acceptability of the encoding bitrates for different sizes

3. PREPARING THE VIDEO CLIPS

Our study focuses on video quality assessment with respect to three different dimensions: (1) content types, (2) sizes and (3) video encoding bit rates. Within the dataset, each content type, i.e. Animation (A), Football (F), Music (M) and News (N), was represented by four clips of 2:20 minutes. Every clip was chunked into seven 20 second segments, each of which we encoded at a different bit rate. This resulted in a total of 448 video clips. The structure of the video clips for one of the four sizes (i.e. 120x90) is visualized in Table 1. For later reference, the light grey marked blocks refer to all video clips encoded at 224 kbps. The dark grey marked blocks refer to the video clips of content type football.

Table 1: Structure of video clips at size 120x90 (S1)

		Bitrate						
		224	192	160	128	96	64	32
120 x 90 (S1)	A	A1S1P1	A1S1P2	A1S1P3	A1S1P4	A1S1P5	A1S1P6	A1S1P7
		A2S1P1	A2S1P2	A2S1P3	A2S1P4	A2S1P5	A2S1P6	A2S1P7
		A3S1P1	A3S1P2	A3S1P3	A3S1P4	A3S1P5	A3S1P6	A3S1P7
		A4S1P1	A4S1P2	A4S1P3	A4S1P4	A4S1P5	A4S1P6	A4S1P7
	F	F1S1P1	F1S1P2	F1S1P3	F1S1P4	F1S1P5	F1S1P6	F1S1P7
		F2S1P1	F2S1P2	F2S1P3	F2S1P4	F2S1P5	F2S1P6	F2S1P7
		F3S1P1	F3S1P2	F3S1P3	F3S1P4	F3S1P5	F3S1P6	F3S1P7
		F4S1P1	F4S1P2	F4S1P3	F4S1P4	F4S1P5	F4S1P6	F4S1P7
	M	M1S1P1	M1S1P2	M1S1P3	M1S1P4	M1S1P5	M1S1P6	M1S1P7
		M2S1P1	M2S1P2	M2S1P3	M2S1P4	M2S1P5	M2S1P6	M2S1P7
		M3S1P1	M3S1P2	M3S1P3	M3S1P4	M3S1P5	M3S1P6	M3S1P7
		M4S1P1	M4S1P2	M4S1P3	M4S1P4	M4S1P5	M4S1P6	M4S1P7
	N	N1S1P1	N1S1P2	N1S1P3	N1S1P4	N1S1P5	N1S1P6	N1S1P7
		N2S1P1	N2S1P2	N2S1P3	N2S1P4	N2S1P5	N2S1P6	N2S1P7
		N3S1P1	N3S1P2	N3S1P3	N3S1P4	N3S1P5	N3S1P6	N3S1P7
		N4S1P1	N4S1P2	N4S1P3	N4S1P4	N4S1P5	N4S1P6	N4S1P7

We will use VQM to compute MOS_{VQM}, the video clips' objective video quality in terms of MOS. As explained in the previous section VQM is a full reference method, i.e. VQM computes the quality of a video clip by comparing it to its corresponding original. This means that we need the corresponding, high quality, originals as a reference to compare them with the 448 degraded video clips used by [15]. All of the original video clips were encoded at 3.2 Mbps or higher. Each video clip, both original and degraded, has a frame rate of 12.5 fps.

Now, if we would determine the objective measure MOS for all degraded clips, at their original sizes, it would not be possible to compare MOS values between different sizes. The reason for this is that VQM intrinsically only determines the visual quality of a degraded clip, irrespective of its size. Therefore, we need to set a baseline for the size for the sake of comparison. We have chosen to use the largest size, 240x180, as the baseline. Therefore we have up-sampled all clips smaller than 240x180 to this size by using the Precise Bicubic resize method in VirtualDub [16].

This resulted in 448 degraded video clips and 448 original video clips, all at a size of 240x180. This means that we are now in a position to determine the objective measure MOS for all 448 video clips by using VQM, see Figure 2.

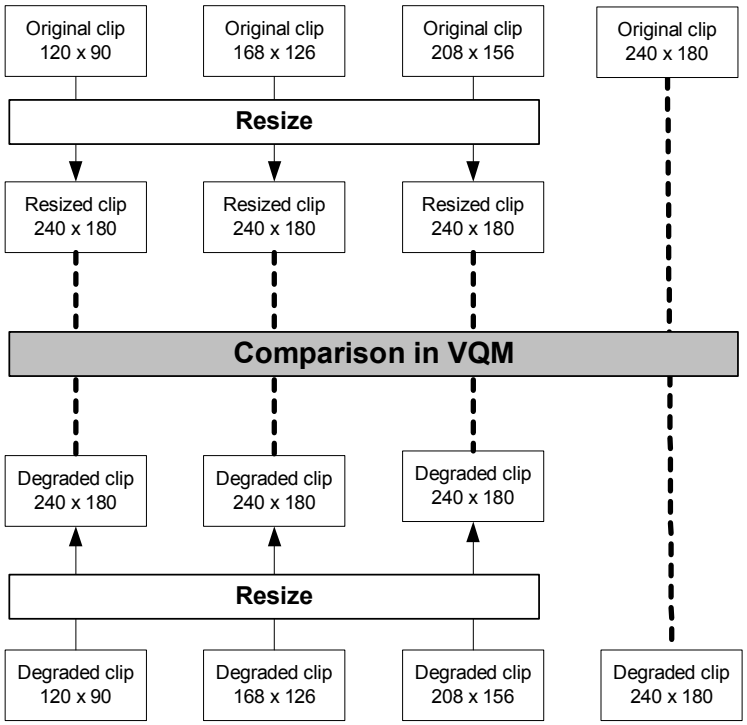

Figure 2: Resizing part of the video clips

4. OBJECTIVE QUALITY OF THE VIDEO CLIPS: MOS_{VQM}

In the previous section we described how every clip was related to its corresponding non-downgraded original, i.e. the video clip of the same content, at equal testing size and original bit rate. Next the proper program and parameter settings in VQM were determined. The actual video quality model that we used is the General Model. The NTIA General Model has been standardized by ANSI (T1.801 [2]) and is included in two ITU Recommendations (ITU-R BT.1683 [9] and ITU-T J.144 [12]). The NTIA General Model was selected for submission to the VQEG full reference phase-2 test since it provides the most robust, general purpose metric that can be applied to the widest range of video systems. We chose to use the General Model because of all six proponents in the VQEG phase-2 test, it had achieved the highest overall correlation with subjective data.

We used the VQM implementation for PC, Version 2.2. Frame sizes tested for this release include: NTSC (720x486), CIF (352x288), QCIF (176x144), SIF (360x240), 720x480, 320x240, 240x180, PAL (720x576). Frame Rates tested for this release include: 30, 29.97, 25, and 15 fps. We processed all clips with VQM with a frame rate setting of 15 fps.

Table 2 contains the resulting MOS_{VQM} values for each unique combination of content type, size and bit rate. Note that every MOS_{VQM} value in Table 2 is actually obtained by averaging across four clips' individual MOS_{VQM} values (cf. Table 1).

Table 2: MOS_{VQM} values for every combination of content, size and bit rate

		Bitrate						
		224	192	160	128	96	64	32
S1	A	3.96	3.96	4.05	3.97	3.85	3.81	3.52
	F	4.08	4.10	4.07	3.87	3.65	3.30	2.65
	M	4.13	3.97	3.98	3.78	3.49	3.28	2.68
	N	3.81	4.04	3.90	3.89	3.56	3.42	2.85
S2	A	4.29	4.29	4.37	4.27	4.14	4.03	3.48
	F	4.23	4.16	4.15	3.90	3.68	3.28	2.52
	M	4.34	4.16	4.10	3.94	3.56	3.36	2.36
	N	4.21	4.25	4.18	4.05	3.67	3.51	2.81
S3	A	4.41	4.44	4.51	4.41	4.26	4.06	3.49
	F	4.24	4.15	4.17	3.89	3.65	3.26	2.27
	M	4.39	4.22	4.12	4.01	3.53	3.35	2.35
	N	4.38	4.33	4.25	4.09	3.72	3.46	2.74
S4	A	4.52	4.53	4.57	4.45	4.33	4.07	3.48
	F	4.25	4.14	4.07	3.78	3.60	3.16	2.11
	M	4.36	4.22	3.87	3.66	3.35	3.00	2.13
	N	4.48	4.38	4.29	4.11	3.67	3.48	2.56

The results given in Table 2 can be visualised in many ways. Knoche et al [15] visualize the content-independent relation between acceptability and video size for different video bit rates, see Figure 1. Analogously, we present the content-independent relation between MOS_{VQM} values and video size for different bit rates in Figure 3.

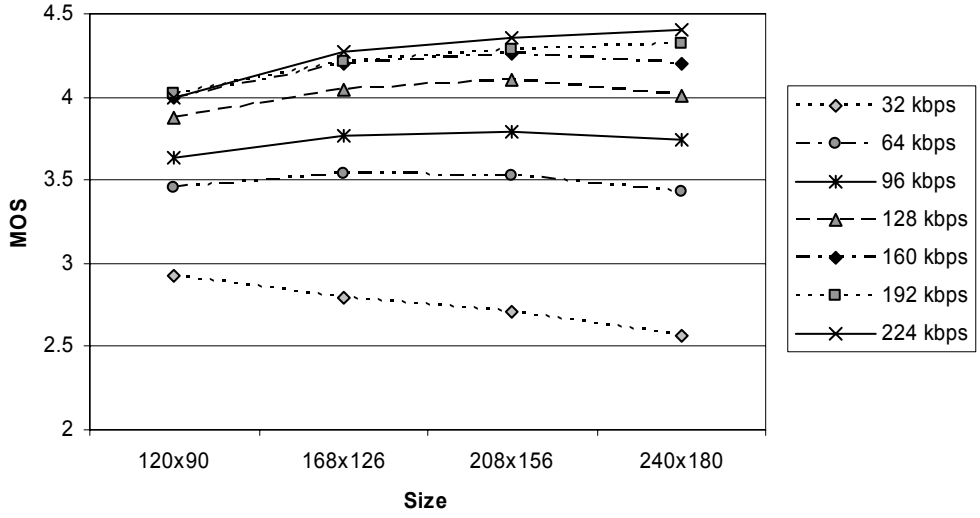

Figure 3: MOS_{VQM} values for different sizes and bit rates

Note that by visual inspection we can already confirm that the MOS_{VQM} curves in Figure 3 have the same tendency as the acceptability curves in Figure 1. For instance, both MOS_{VQM} and acceptability curves increase with increasing size at the coding bit rate 224 kbps, while both curves decrease towards the largest size for the coding bit rate 32 kbps. In the next section we will study in depth the relation between the acceptability values obtained by Knoche et al [15] and the MOS_{VQM} values given in Table 2.

5. MAPPING MOS_{VQM} TO ACCEPTABILITY

5.1 Content independent mapping; data

The aim of this section is to establish a mapping between the objective measure MOS_{VQM} and the subjective quality measure acceptability. First we consider the case where both subjective and objective measures are averaged across all content types. The subjective and objective measurement data needed to construct the mapping is extracted from Knoche et al. [15] and the results from the previous section. On a high level, the mapping is visualized in Figure 4.

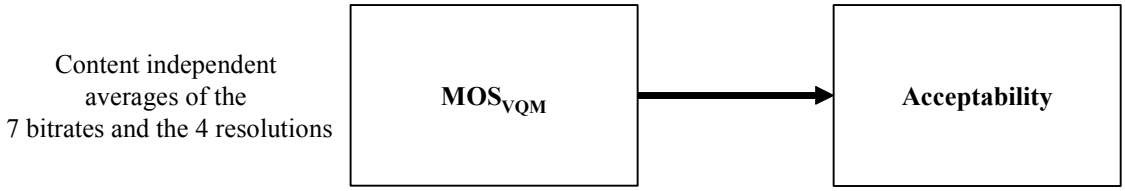

Figure 4: Mapping on content independent level from MOS_{VQM} to Acceptability

5.2 Content independent mapping; results

For each of the four sizes, MOS_{VQM} values for the seven bitrates plotted in Figure 3, will be related to their corresponding acceptability values as presented in Figure 1. This leads to the series of 28 data points, which are depicted in Figure 5. along with the mapping M2G between MOS and the binary measure GoB, as defined in ITU-T G.107 [ITU-T g.107].

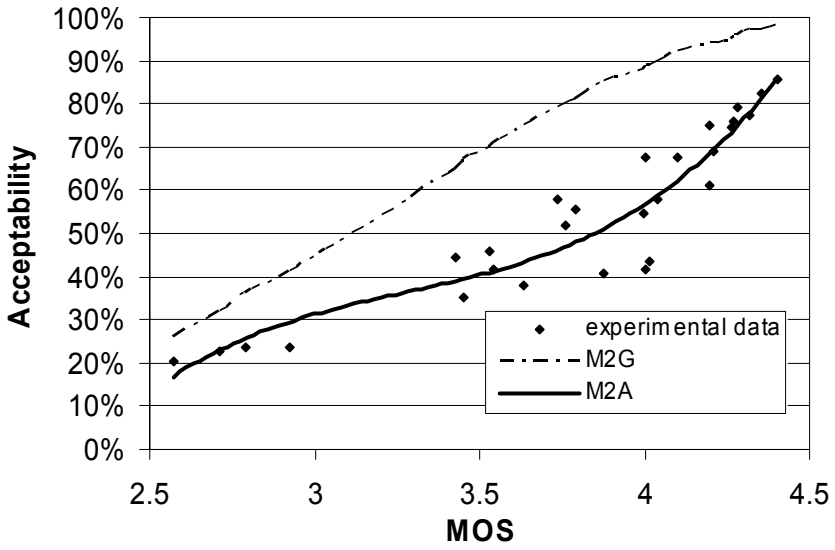

Figure 5: Generic mapping on content independent level

Apparently, M2G resulted in a serious overestimation of Acceptability scores. We therefore suggest an improved mapping - M2A (MOS to Acceptability)- based on our experimental data. Assuming a third order polynomial fit we find the following equation through regression analysis:

$$\text{M2A: } Acceptability = 0.22\ MOS_{VQM}^3 - 2.13\ MOS_{VQM}^2 + 7.14\ MOS_{VQM} - 7.81. \quad (1)$$

A visual inspection clearly shows the improvement of M2A over M2G, see Figure 5. The measure that we use to express how well a model explains subjective data is the linear correlation coefficient R, also known as the Pearson correlation. The correlation coefficient is related to the coefficient of determination R^2, which is a measure for how well the model explains the variation in the subjective data and can be 1 for maximum correlation. The correlation coefficient R for M2G and the new mapping M2A is 0.87 and 0.94, respectively. Thus, it follows that the new mapping M2A outperforms M2G for the specific MOS domain (2.40 – 4.40). Another measure to illustrate the superiority of M2A is the Root Mean Square Error (RMSE). For aggregated content these are, 1.42 for M2G and 0.34 for M2A respectively - an improvement of M2A on M2G of 76%.

As the goal is to create a generic mapping between MOS_{VQM} values and acceptability, it should be applicable on the entire domain of both acceptability (0 - 1) and MOS_{VQM} values (1 - 5). Since the generic mapping M2A given in (1) does not meet this requirement, the mapping should be extended, although currently no data is available about the entire domain. As both acceptability and MOS_{VQM} values have the same semantic direction and are both limited to a specific domain, we can assume that both minimum and maximum scores on both scales are equal, resulting in the conclusion that an acceptability score of 0 equals a MOS_{VQM} value of 1 and an acceptability score of 1 equals a MOS_{VQM} value of 5. Taking this into account, for three different ranges, the following mapping function for M2A can be formulated:

$$Acceptability = 0.11\ MOS_{VQM} - 0.11 \qquad if\ 1.00 < MOS_{VQM} \leq 2.57 \quad (2)$$

$$Acceptability = 0.22\ MOS_{VQM}^3 - 2.13\ MOS_{VQM}^2 + 7.14\ MOS_{VQM} - 7.81 \qquad if\ 2.57 < MOS_{VQM} \leq 4.40 \quad (3)$$

$$Acceptability = 0.23\ MOS_{VQM} - 0.15 \qquad if\ 4.40 < MOS_{VQM} \leq 5.00 \quad (4)$$

5.3 Content specific mapping; data

The usefulness of the mapping M2A depends on its applicability. The mapping function M2A provides a strong and evident relation between acceptability and MOS_{VQM} scores. Due to the fact that there appears to be significant difference for the relation between acceptability and MOS_{VQM} values among content types, it might also prove necessary to differentiate among content types. In order to test to what degree the generic mapping function M2A can be applied to the level of a specific content type we have examined the content type football. We base this analysis on the same data set [15] which consists of 28 different scores on the content type football. To calculate the corresponding objective MOS_{VQM} scores, sizes and bit rates are related and averaged across the four values from the blocks that contain football content. The concept is visualized in Figure 6.

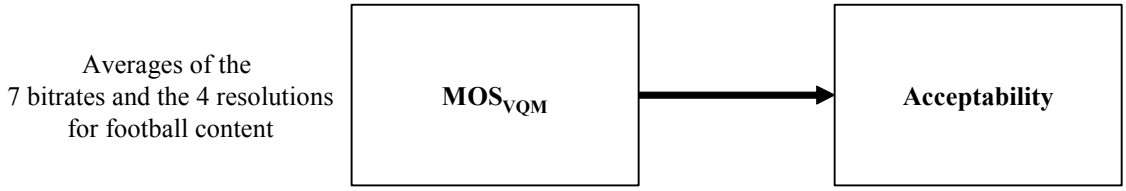

Figure 6: Mapping on content specific level

5.4 Content specific mapping; data

Following the same procedure as in the previous subsections, we visualize the 28 data points related to football, and the content independent mapping M2A in Figure 7.

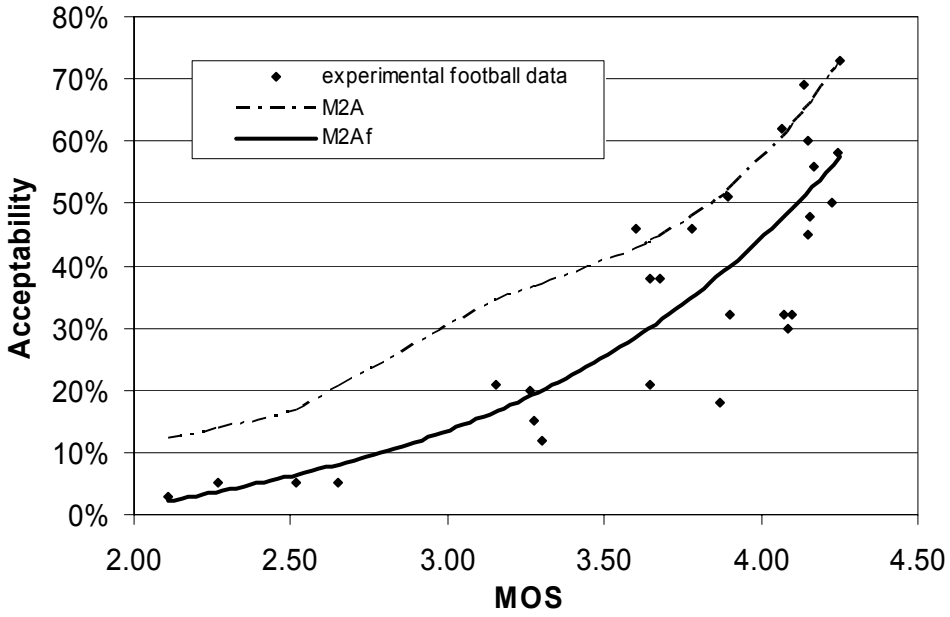

Figure 7: Mapping of content type football

Now clearly M2A overestimates the acceptability for football content. Next we introduce a mapping $M2A_f$, specifically for the content football:

$$M2A_f: \quad Acceptability = 0.03\ MOS_{VQM}^3 - 0.20\ MOS_{VQM}^2 + 0.51\ MOS_{VQM} - 0.47 \tag{5}$$

From Figure 7 it is obvious that the football specific mapping $M2A_f$ performs better than the generic content independent mapping M2A. This difference is hardly apparent from the correlation coefficient R (R = 0.84 for M2A, R = 0.85 for $M2A_f$), but is more clearly illustrated if the Root Mean Square Error (RMSE) is calculated. The RMSE for M2A and $M2A_f$ are 2.17 and 0.56, respectively. This, the content dependent mapping $M2A_f$ yields a reduction of 1.61 in terms of RMSE, which amounts to an improvement of 74% over M2A.

Just like M2A, also the mapping for the content type football should be extended in order to be applicable to the entire domain of acceptability scores and MOS_{VQM} values. This resulted in three different mappings for each domain:

$$Acceptability = 0.02\ MOS_{VQM} - 0.02 \qquad if\ 1.00 < MOS_{VQM} \leq 2.11 \tag{6}$$

$$M2A_f: \quad Acceptability = 0.03\ MOS_{VQM}^3 - 0.20\ MOS_{VQM}^2 + 0.51\ MOS_{VQM} - 0.47 \qquad if\ 2.11 < MOS_{VQM} \leq 4.25 \tag{7}$$

$$Acceptability = 0.57\ MOS_{VQM} - 1.85 \qquad if\ 4.25 < MOS_{VQM} \leq 5.00 \tag{8}$$

From Figure 5 and Figure 7 we conclude that the generic mapping M2A cannot be applied with the same results to the experimental data for a specific content type. Therefore we can state that although the mapping M2A on content independent level provides a high correlation on the generic level, it can be improved on content specific level by calculating mappings for each content type. We have illustrated for one content type, namely football.

6. SIZE MATTERS

It is apparent from the results in the previous section that even though the content specific mapping $M2A_f$ is an improvement over the content independent M2A, the overall fit between $M2A_f$ and the experimental football data is not very good. This can be seen in Figure 7 and it also follows from the correlation coefficient which is only R = 0.85. The main explanation for the poor fit between the mapping and the experimental data can be found when size is taken into account. To illustrate this, in Figure 8, we have grouped all data points that belong to the same size, and as such, every size is visualized as a line of data points.

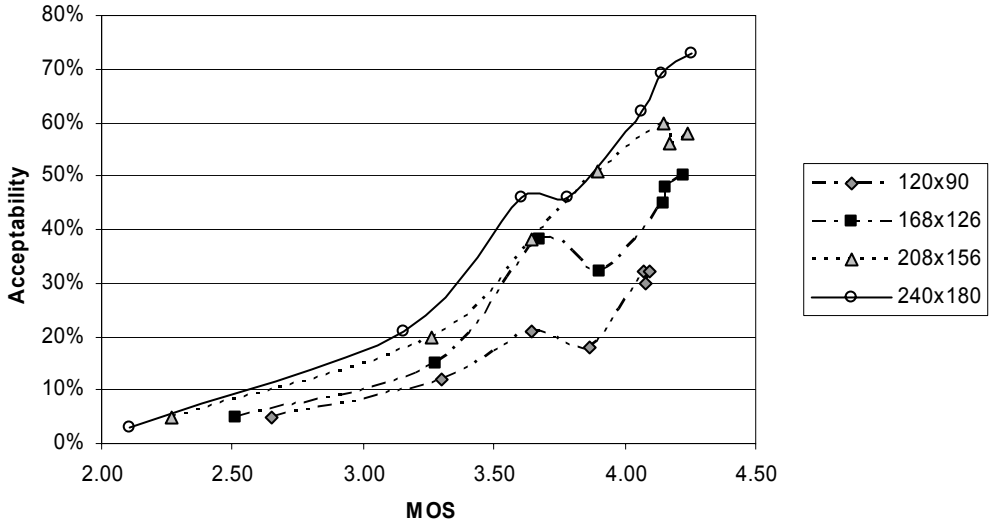

Figure 8: Acceptability for different sizes: content football

It is clear that the individual lines in Figure 8 do not show the strong variability for acceptability as the data points in Figure 7. This indicates that we could construct even more accurate mappings between acceptability and MOS_{VQM} if we took size into account. As an example we have constructed the mapping $M2A_{fS3}$, which describes the relation between MOS_{VQM} and acceptability for football content at the size S3 (208x156). For this data set, both for $M2A_f$ and $M2A_{fS3}$, the correlation coefficient satisfies R > 0.95 while the RMSE is 0.03 and 0.007 for $M2A_f$ and $M2A_{fS3}$, respectively. The importance of size was not directly apparent from the mapping M2A which was obtained by averaging across content, see Figure 5. This, however, does not compel that size does not matter on an aggregated level. To illustrate the impact of size, Figure 5 was recreated with the same data points but with the data grouped by size, see Figure 9. The data points in Figure 9 are connected on the basis of the size they represent, providing a clear difference between the sizes. Therefore, by taking size into account, we could improve the performance of the content independent mapping M2A.

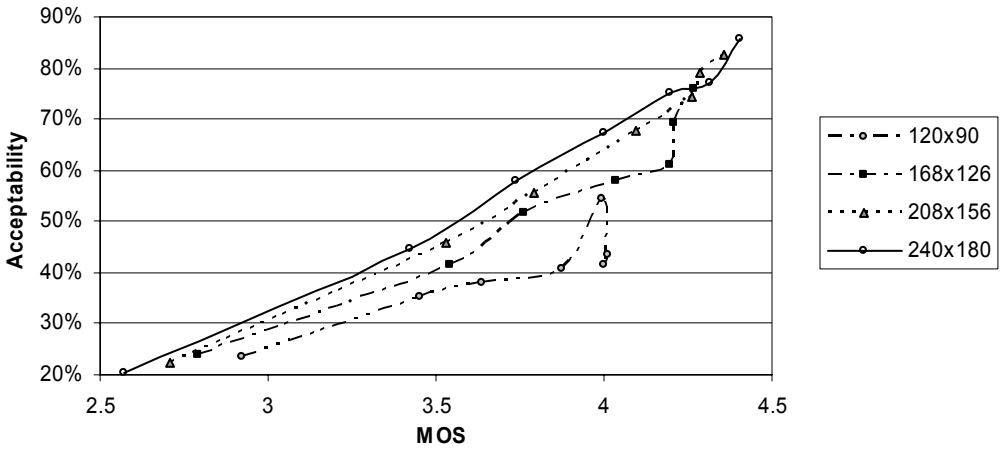

Figure 9: Acceptability for different sizes: content independent

7. CONCLUSIONS

In this paper we have studied the relation between subjective and objective video quality scores that we obtained through the subjective video quality measure Acceptability and the objective video quality metric VQM. We focused on the domain of Mobile TV with typical sizes and encoding bitrates. We have shown that mapping MOS values to acceptability according to M2G, a mapping suggested by the ITU which relates MOS to binary Good or Better values, results in a serious overestimation of acceptability.

A new, content independent mapping M2A has been suggested which relates MOS values to acceptability. M2A clearly outperforms M2G, as can be seen from correlation coefficients, which are $R = 0{,}94$ and $R = 0{,}87$ for M2A and M2G, respectively. The superiority of M2A over M2G is further illustrated by the Root Mean Square Error (RMSE) which is 1.42 and 0.34 for M2G and M2A, respectively. We can further improve on M2A if we take content type and size into account.

Applying the content independent mapping M2A to a specific content type might lead to serious discrepancy between the mapping and experimental data. For the content type football we have constructed a content specific mapping $M2A_f$. For M2A and $M2A_f$ the correlation coefficients are comparable, namely $R = 0.84$ and $R = 0.85$, respectively. In terms of the RMSE, $M2A_f$ (RMSE = 0.56) clearly outperforms M2A (RMSE = 2.17).

If we take size into account, then the performance of the content specific mappings can be further improved. We have illustrated this by constructing the mapping $M2A_{fS3}$ for football content at the size S3 (208x156). We found that for the relevant data set both for $M2A_f$ and $M2A_{fS3}$, the correlation coefficient satisfies $R > 0.95$ while the RMSE is 0.03 and 0.007 for $M2A_f$ and $M2A_{fS3}$, respectively.

Future research will validate the predictive power of our suggested mappings on other video material and produce a mapping that includes video size as a parameter.

ACKNOWLEDGEMENT

This work is partially supported by the IST-507295 Multi-Service Access Everywhere (MUSE) project, see www.ist-muse.org. The overall objective of MUSE is the research and development of a future low cost, full service access and edge network, which enables the ubiquitous delivery of broadband services. The work reported in the paper is conducted within the MUSE lab trials and demonstration Taskforce (TF4) where extensive effort is put in defining a complete test suite for full-service end-to-end testing. Furthermore, the work was supported by the IST-2005-27034 Universal Satellite Home Connection (UNIC) project. The scope of UNIC is to provide TV-centric interactive multimedia services via satellite and set top boxes. Portable devices will connect to the STB wirelessly and enable further interaction with and consumption of content around the home and conjunction with other users. The presented work is part of the service assessment effort.

REFERENCES

1. Aldridge, R. P., Hands, D. S., Pearson, D. E., Lodge, N. K. Continuous assessment of digitally-coded television pictures. *IEE Proceedings - Vision, Image and Signal Processing*, 145 ((2)), (1995), 116-123
2. American National Standards Institute (ANSI) *Digital transport of one-way video signals – Parameters for objective performance assessment.* (T1.801.03 - 2003) ANSI, (2003)
3. Apteker, R. T., Fisher, A. A., Kisimov, V. S., Neishlos, H. Distributed multimedia: user perception and dynamic QoS. In *Proceedings of SPIE*, 1994, 226-234
4. Bouch, A. *A User-centered Approach to Network Quality of Service and Charging.* PhD Thesis University College London, 2001
5. Fechner, G. T. *Elemente der Psychophysik* Leipzig: Breitkopf und Härtel, 1860
6. Hands, D. S. & Wilkins, M. A Study of the Impact of Network Loss and Burst Size on Video Streaming Quality and Acceptability. In *Proceedings of the 6th International Workshop on Interactive Distributed Multimedia Systems and Telecommunication Services*, London, UK: Springer-Verlag, 1999, 45-57

7. Hauske, G., Stockhammer, T., Hofmaier, R. Subjective Image Quality of Low-Rate and Low-Resolution Video Sequences. In *Proceedings of the 8th International Workshop on Mobile Multimedia Communications*, 2003
8. International Telecommunications Union (ITU) *P.800.Methods for subjective determination of transmission quality.* (ITU-T P.800) , (2004)
9. ITU-R *Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference.* (ITU-R BT.1683) Geneva, Switzerland: (2004)
10. ITU-T *Subjective Video Quality Assessment Methods for Multimedia Applications - in Multimedia Services.* (P.910) , (1999)
11. ITU-T *The E-model, a computational model for use in transmission planning.* (ITU-T G.107) , (2003)
12. ITU-T *Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference.* (ITU-T J.144) , (2004)
13. Jumisko-Pyykkö, S., Kumar, V. M. V., Korhonen, J. Unacceptability of instantaneous errors in mobile television: from annoying audio to video. In *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services*, 2006, 1-8
14. Knoche, H., Carthy, J., Sasse, M. A. How low can you go? The effect of low resolutions on shot types. *Personalized and Mobile Digital TV Applications*, (2007)
15. Knoche, H., McCarthy, J., Sasse, M. A. Can Small Be Beautiful? Assessing Image Resolution Requirements for Mobile TV. In *Proc.of ACM Multimedia 2005*, ACM, 2005, 829-838
16. Lee, A., *VirtualDub*., www.virtualdub.org. 2007
17. McCarthy, J., Sasse, M. A., Miras, D. Sharp or smooth? Comparing the effects of quantization vs. frame rate for streamed video. In *Proc.CHI*, 2004, 535-542
18. Watson, A. & Sasse, M. A. Measuring Perceived Quality of Speech and Video in Multimedia Conferencing Applications. In *Proceedings of ACM Multimedia '98*, 1998