Musikpsychologie Bd. 27, 112–127, © Hogrefe Verlag, Göttingen 2017

Audio-visual quality perception in musical performance videos

David Hammerschmidt & Clemens Wöllner

Zusammenfassung

Eine zentrale Fragestellung der audiovisuellen Qualitätswahrnehmung ist, wie sich der rezeptive Gesamteindruck konstruiert. Während über alle Videogenres hinweg der Videoqualität der größte Einfluss auf die audiovisuelle Qualitätsbeurteilung zugeschrieben wird, ist der Einfluss der Audioqualität auf das qualitative Gesamturteil bei Musikvideos stärker im Vergleich zu anderen Videotypen. Die vorliegende Studie untersuchte anhand von drei professionellen Musikvideos, inwieweit die unimodalen Qualitäten den audiovisuellen Gesamteindruck sowie sich gegenseitig beeinflussen. Hierfür beurteilten Probanden die subjektiv wahrgenommenen Qualitäten des Audio- und Videosignals sowie dessen Kombination (audiovisuelle Qualität). Die Ergebnisse zeigen, dass die beiden unimodalen Qualitäten die wahrgenommene audiovisuelle Gesamtqualität bei Musikvideos gleich stark beeinflussen und nicht von der Videoqualität dominiert wird. Des Weiteren wurde die wahrgenommene Audioqualität von der Videoqualität beeinflusst jedoch nicht umgekehrt. So waren die Bewertungen der Versuchsteilnehmer für die Videoqualität nicht von der simultan präsentierten Audioqualität betroffen, die Bewertungen der Audioqualität jedoch von der Videoqualität. Dies spricht für eine stärkere Beeinflussung der wahrgenommenen Audioqualität durch die Videoqualität bei Musikvideos als umgekehrt.

Abstract

One of the main questions of audio-visual quality perception is what influences the receptive overall impression. While video quality has the highest impact on audio-visual quality assessment across all video genres, research showed that audio quality is more influential in music videos compared to other video genres. The present study investigated the extent to which unimodal qualities influence the perceived audio-visual quality as well as mutual influences of the unimodal qualities by using three professional music videos. Participants evaluated the subjectively perceived audio and video quality in addition to a combination (audio-visual quality). Results show that the quality of both modalities influenced the audio-visual quality similarly and is not dominated by the video quality. Furthermore, the perceived audio quality was affected by the video quality but not vice versa. In other words, participants' judgments were not influenced by the simultaneously presented audio quality when judging video quality, yet they were influenced by the video quality when judging audio quality. This indicates that in music videos the perceived audio quality is more strongly influenced by the video quality than vice versa.

1 Introduction

Audio-visual quality perception of music videos is a field of investigation that is influenced by a multitude of different factors. In this paper, quality is defined as a criterion of excellence for all perceivable characteristics of an audio-visual signal. Therefore, quality perception assumes a person as the final authority of quality assessments. Music videos represent a specific kind of media content that use digital signal formats and impose increased demands on compression methods as well as on perceptual characteristics, due to complex signal components. Lossy data compression is the most important factor for quality perception in practice. The compression of digital signals is necessary since uncompressed signals are too demanding for processing systems (Schulze, 2006). Therefore, compression aims to minimise the required memory space without noticeably reducing the perceptual quality of the signal's content.

The following experiment aimed at exploring how modal-specific quality impairments caused by bit rate compression have an impact on the overall perceived audio-visual quality, and to which extent unimodal qualities influence each other in music videos. To the best of our knowledge, this is the first study specifically addressing audio-visual quality perception for musical performance videos. Furthermore, we chose a design that differs from most published studies in the way that all quality assessments were only carried out in audio-visual presentation modes, since we specifically aimed at analysing mutual influence of unimodal qualities. Since the field of quality perception and its evaluation involves a wide range of factors, a short description of the most important aspects will be given first.

1.1 Perceptual and cognitive factors of quality assessment

To reduce the required memory space of a signal, compression methods use characteristics of human perception. The most relevant aspects of auditory perception are the outer and middle ear, perceptual frequency scales, excitation and detection thresholds, simultaneous, temporal, and partial masking effects as well as hearing sensitivity (Ellermeier & Hellbrück, 2008; ITU-R BS.1387-1, 2001). Most important for visual perception are the characteristics of light sensitivity, contrast sensitivity, colour perception, spatiotemporal masking effects and pattern matching (Winkler, 2005). Information below the thresholds of these properties are considered less relevant to the recipient and thus irrevocably deleted by the encoder (Schulze, 2006). However, certain processing steps of the compression produce artefacts and may add them to the original signal.

These artefacts become more perceptible with an increasing degree of compression. Artefacts can adversely affect the perceived quality. Examples of such artefacts are the visual block boundary, which generally describes discontinuities or differences at the boundaries of adjacent pixel blocks (for video artefacts, see Yuen & Wu, 1998) or the auditory "birdie" effect, which causes changes in timbre and energy variation when strong variations in frequency representations appear from one frame to the next (for audio artefacts, see Liu, Hsu & Lee, 2008).

For audio-visual signals, additional factors such as the synchronous replay of the auditory and visual signal components are crucial for quality perception. This is due to the detection of synchronisation errors in audio-visual signals (Hollier & Rimell, 1998). If temporal asymmetry exceeds a content dependent threshold, an audio-visual signal is no longer perceived as spatiotemporal coherent, leading to impairments of intelligibility and subjective quality (Dixon & Spitz, 1980; Massaro, Cohen & Smeele, 1996). Furthermore, a number of cognitive processes are involved in the assessment of quality.

First, attention is directed towards the information that promises the most accurate solution for a current task. In the case of audio-visual quality assessment, cross-modal error masking may occur (Hollier, Rimell, Hands & Voelcker, 1999). This effect describes the situation when attention is focused on one modality so that impairments of a signal in the other modality remain unnoticed by the recipient. Thus the information of that modality is less likely to be included in the decision making process.

Secondly, long-term memory provides information that has been learned or experienced in the past. In a study by Pras, Zimmerman, Levitin and Guastavino (2009), it was shown that listening expertise is crucial for the perceptibility of auditory quality differences. In this experiment, sound engineers preferred uncompressed stimuli more often, concluding that they better recognised differences in the audio quality due to experience and practice. Specific information of a signal may also create associations and functions as a reference, and as such influence quality judgments. These references are based on any kind of listening habits such as musical preferences (Hollier et al., 1999).

Finally, working memory is responsible for the temporary storage of information as a function of the given task. The influence of working memory on quality assessment has not been extensively investigated yet. In a study by Aldridge, Davidoff, Ghanbari, Hands and Pearson (1995), participants evaluated the video quality of a signal to be lower if the poorer quality appeared at the end of a signal as compared to the beginning. Quality impairments can vary over time, whether caused by compression or by the content itself and thus affect the assessment. Studies have shown that the intensity of a past affective experience has greater impact on its evaluation than its duration, which is negligible in comparison (Fredrickson & Kahneman, 1993; Redelmeier & Kahneman, 1996). The decision-making process is therefore more likely to identify particularly striking aspects of a signal (peaks) than temporal components.

1.2 Structure of audio-visual quality perception

In addition to the factors already mentioned, there is a variety of different coding formats, various quality impairments (e.g. artefacts) and different environmental aspects (e.g. room acoustics). Generalised statements based on limited data are therefore difficult to make, since it is not possible to control all relevant internal and external factors and to systematically vary them in one experiment. For this reason, meta-analyses are particularly well-suited to draw general conclusions from previous research on audio-visual quality perception. The most common method derives the perceived audio-visual quality from the mean opinion score (MOS) of the audio and video qualities, which is intended to explain the relationship of these factors to the MOS of the overall impression (audio-visual quality) by correlating these factors. Based on eight studies using such a subjective test method, the meta-analysis of You, Reiter, Hannuksela, Gabbouj and Perkis (2010) shows that the qualities of the audio and video signals impact the overall judgment with a dominating influence of video quality on the perceived audio-visual quality. In other words, video quality is more important for audio-visual quality than audio quality. Pinson, Ingram and Webster (2011) disagree with this conclusion and suggest that many studies used different perceivable quality ranges for audio and video. In their meta-analysis, based on thirteen studies, they noted that a dominance of the video quality was only found in studies in which the evaluated range of audio and video quality impairments was unbalanced. They argue that studies with a balanced range of quality levels for both signals show that the quality in both modalities is of similar importance for judgments of the audio-visual quality. So it seems plausible that studies with an unequal ratio of perceptible quality ranges may create a dominance of video quality or at least magnify the effect. The results of You and colleagues (2010) also indicate that the mutual unimodal impact on audio-visual quality varies with the quality level. Accordingly, the influence of the audio quality increases for the overall impression when the quality of the audio and video signal is low or when the audio signal is of very poor quality. Thus, the influence on perceived audio-visual quality increases with decreasing audio quality.

The interaction of unimodal qualities, which means the mutual influence on the other perceived quality, has been less studied than the overall audio-visual quality. The mutual influence is often established by the MOS of the audio and video quality while stimuli are presented in unimodal and audio-visual modes. Results of previous studies suggest a minor mutual influence of unimodal qualities (Beerends & De Caluwe, 1999; Kitawaki, Arayama & Yamada, 2005; ITU SG 12 Contribution COM 12.61-E, 1998), such that judgments of audio quality were influenced by video quality and vice versa. Beerends and De Caluwe (1999) additionally examined whether judgments of an unimodal quality differs from audio-visual presentations but could not find any significant differences. In fact, the judgments between presentation modes differed by less than one percent in their experiment.

In a previous study, we examined whether the perceived audio quality of a musical performance video is influenced by the simultaneously presented video quality (Hammerschmidt & Wöllner, 2015). While the video quality was varied, the audio quality remained the same. Results revealed that participants recognised the different video qualities, yet also evaluated the constant audio quality to be different depending on the presented level of video quality. With increasing video quality the influence on the audio quality increased as well. The better the video quality, the better the perceived audio quality was judged based on a music video.

1.3 Content dependency

Another important aspect of quality perception is the broad spectrum of signal content types whose specific characteristics influence the perceived quality. Regarding the dependence of content on audio-visual quality perception we are aware of only one study that specifically investigated this factor. Garcia, Schleicher and Raake (2011) compared five content types (Film Trailer, Football, Interview, Film and Music Video) for differences in the perceived audio-visual quality. The stimuli differed visually in terms of their level of detail, the complexity of structure and movement. The auditory conditions differed with regard to the content type (language and/or music) and the musical genre (Classical Music, Pop Music). Results showed that video quality was more important than audio quality for the perceived audio-visual quality in all content types. However, results indicate a different impact of audio quality on audio-visual quality ratings depending on content type. The impact of both unimodal qualities on the audio-visual quality was overall similar for all content types except the music video. It could be shown that the audio quality of a music video is significantly more important. The representation of music in form of a musical performance video thus seems to place the auditory information more in the focus of attention of the recipient, since other stimuli that may include background music did not differ from non-musical stimuli. The results suggest an increased influence of the audio quality on the perception of audio-visual quality for music videos.

As Garcia et al. (2011) indicate, a multitude of different content types may include music, so it is necessary to clarify how a music video is characterised. For the current study, music videos are defined as concert recordings or performance videos. This is the dominant form of visual presentation of music and a foundation of the typology of musical videos (Jost, Klug, Schmidt, Reautschnig & Neumann-Braun, 2013). A music video is the synchronous visual realisation of auditory events in the form of a musical performance in which place, time and actions are homogeneous. The visual content appears as a complete source of the sound. The sound generation determines the visual actions and thus the greatest possible focus of the recipient is placed on the auditory signal. The videos are classified as complex by the level of information and include a high degree of movement, image sections, camera movements and often vigorous changes in light conditions (ibid.).

Furthermore, Pras and colleagues (2009) investigated the influence of musical genres on the auditory quality perception of compressed signals. They argue that the perceptibility of quality differences is less related to individual genres rather than to differences between electronically amplified (Pop, Rock/Metal) and purely acoustic music (Contemporary Music, Orchestral Music, Opera). Participants preferred more frequently the uncompressed compared to the compressed versions for electronically amplified music.

In a study by Ruzanski (2006), musical pieces representing different genres (Notturno, Capriccio, Soft Rock, Heavy Rock, Rap) were also examined for effects of lossy compression. No dependence of the auditory quality perception on genre was found in this experiment as well. The results rather suggest that quality differences in compressed music are less perceptible when the music is of low dynamic range. However, both studies did not use a representation of various genres, since each one included only one piece of music.

The subjective quality of video signals is particularly dependent on image complexity. It can be assumed that the complexity of a compressed video increases with its entropy (measure of the amount of information) and therefore requires an increased data memory (Strutz, 2009). Thus, visual artefacts can occur which in turn have a negative effect on the perception of video quality. A distinction is made between spatial information such as the number of edges and corners in the image frame and the temporal information like the nature and direction of movement (Ries, Crespi, Nemethova & Rupp, 2007). Movements within the image as well as movements of the camera and the speed of movement increase the level of information and so require more data space for coding (Strutz, 2009). For example, Ries and colleagues (2007) classified signals for their model of subjective video quality according to the quantity of movement, their speed as well as the colour information.

Taken together, music videos seem to emphasise the role of audio quality for audio-visual quality perception compared to other content types, yet video quality is still more important than audio quality. Furthermore, it is assumed that perceived quality of music videos is dependent on signal characteristics (e.g. auditory dynamic range and visual image complexity) rather than genre.

1.4 Hypotheses

For this study, we were (a) interested in the effect of unimodal qualities on the perceived audio-visual quality. What unimodal quality is more important for the overall quality impression in music videos? In agreement with the studies reported above, we hypothesised that video quality is of greater importance. Furthermore, we were specifically interested in (b) the effect of quality level in one modality on the perception of quality in the other modality. As already reported, these effects are assumed to be mutual and of equal magnitude.

2 Method

2.1 Participants

A total of 28 individuals took part in our study. Data from three participants were not included in the analysis due to outliers in age (64 and 59), and in one case due to uncorrected visual acuity. In the end, judgments of 25 participants were obtained from 13 male and 12 female participants (age: $M=24.6$, $SD=3.8$). Participants had a mean experience of 9.9 years of playing an instrument ($SD=6.4$ years) and a varying level of musical expertise from amateur to semi-professional. Although participants were musically experienced they are not considered experts for the perceptibility of quality impairments (Pras et al., 2009), since none of them was involved in sound engineering or mastering. Before the beginning of the test we asked each participant in a questionnaire which modal-specific quality would be most important to them while watching music videos. 64 percent of the participants stated that audio and video qualities are of equal importance. Audio quality was more important for 32 percent, and four percent stated that the video quality is the most important factor.

2.2 Stimuli

The music videos consisted of commercially unavailable concert recordings of the "Reeperbahn Festival 2014" from the broadcasting cooperation Norddeutscher Rundfunk. Three 15 sec. long purely instrumental clips of three different artist groups were used. The first clip contained the Blues-Rock song "If you don't love me" by The Wild Feathers (Clip_1), the second clip was the Dance-Pop song "Move" by Mausi (Clip_2) and the third clip was "Hurts to be loved by you" by Kill It Kid (Clip_3) which is associated with the genre Alternative. Each of the clips contained electronically amplified music with an instrumentation of electric guitars, electric bass guitar, drums and synthesizers. Referring to Ruzanski (2006), the psycho-acoustical parameters of mean intensity and dynamic range (*SD* of intensity) are listed in Table 1. Additionally, spectral centroids for an approximate characterisation of the frequency spectrums are listed as well. The calculation of the individual parameters was done via Adobe Audition CC 2015.0 and Praat 5.4.09.

The visual complexity of the three clips, which is a main factor for video quality perception, has been calculated for the visual signal components. Using the software VideoAnalysis 5.1.3, quantity of motion (QoM) was calculated and averaged over time (see Table 1). The QoM determines the mean change of all pixels from frame to frame and divides it by the absolute number of all pixels. The results thus provide values between 0 and 1, value 1 representing a change in all pixels from one frame to the next. The QoM can therefore be regarded as a measure of image complexity (Adde, Helbostad, Jensenius, Taraldsen & Støen, 2009). Since camera cuts usually cause such a drastic change their number is also given in the table for each clip.

Regarding cognitive factors, participants rated their familiarity of the songs on a discrete 7-point scale (1= not known at all, 7 = very well known). Analyses show that Clip_1 (*M* = 1.35, *SD* = 0.75), Clip_2 (*M* = 1.52, *SD* = 1.19) and Clip_3 (*M* = 1.6, *SD* = 1.15) were mostly unknown to participants, so there was little to no direct reference regarding the songs and therefore the recording as well.

Tab. 1:
Parameters of audio and visual signal components

Audio Signal	Spectral Centroid *M* (*SD*) in Hz	Intensity *M* (*SD*) in dB
Clip_1	2040.73 (155.97)	67.58 (13.43)
Clip_2	4650.04 (258.35)	71.73 (24.23)
Clip_3	3891.08 (330.55)	70.88 (9.67)
Video Signal	**QoM** *M* (*SD*)	**Camera Cuts** *N*
Clip_1	.46 (.19)	4
Clip_2	.83 (.15)	6
Clip_3	.67 (.14)	6

Note: Means (*M*) and standard deviations (*SD*) of the audio signal for the spectral centroid and intensity (averaged over time). The *SD* of intensity equals the dynamic range. Means (*M*) and standard deviations (*SD*) of the visual signal for quantity of motion (QoM) averaged over time and number (*N*) of camera cuts for each clip.

In the study, three audio and three video impairments were used and combined in every possible way (see Table 2). The clips were encoded according to the MPEG-2 (H.262) standard. This standard is mainly known for its use in DVD's and digital television broadcasting. The audio signal was encoded by the MPEG-1 – Audio Layer II. The compression based on this type of encoding is lossy. A description of the compression methods for the visual component can be found in Mitchell (1997) and for the auditory component in Noll (1997). The quality impairments were implemented in advance with Adobe Premiere Pro CC 2015.0 for both modalities. The default encoding settings remained unchanged and only bit rates for the audio and video signal were manipulated (for a detailed description of bit rate reduction, see Lerch, 2008). Table 2 lists the relevant technical properties and the applied bit rates. The bit rate distribution for the signals was set to a constant distribution (CBR) over time. The quality levels of both signal components therefore differed only by bit rate. In this experiment, we named the applied quality levels "low" for the smallest bit rate, "medium" for the mid-range and "high" for the highest bit rate. The three quality levels per modality and the three clips resulted in 27 stimuli (3 x 3 x 3).

Tab. 2:
Technical properties of the stimuli and the applied bit rate impairments

	Video	Audio
Codec	H.262 (MPEG-2)	MPEG 1 – Audio Layer II
Specifications	720 x 576 Pixel, PAL	48 kHz Sample Rate
	16:9 Widescreen	16 Bit Depth
	Profile: MO Level: ML	Stereo
	25 Frames per Second	
Method	Constant Bit Rate (CBR)	Constant Bit Rate (CBR)
Bit Rates (Quality Levels)	Low: 2 mbit/s	Low: 96 kbit/s
	Medium: 5 mbit/s	Medium: 128 kbit/s
	High: 8 mbit/s	High: 384 kbit/s

2.3 Experimental design

In order to test the hypotheses as specified above, a subjective rating test was carried out in which the participants evaluated the audio-visual quality (AVQ), audio quality (AQ) and video quality (VQ) using a repeated-measure design. The goal was to generate the MOS (mean opinion score) for the three modal-specific quality levels (MOS_AVQ, MOS_VQ, MOS_AQ). The experiment was divided into three blocks in which the audio-visual quality was evaluated in the first block and then randomly either the audio quality block or video quality block was presented first. Each of the three blocks included all 27 stimuli and consistently presented them in audio-visual condition. All stimuli within one block were randomised. The 27 stimuli and the three-time presentation (once per block) resulted in 81 trials total. The test duration was approximately 45 minutes. The continuous audio-visual presentation of the overall audio-visual as well as the unimodal quality assessments differs from the design mostly used in the reported studies above. In these studies, influences of audio and video quality on audio-visual quality perception were derived from unimodal presentations for judgments of audio and video qualities. For the current experiment, we chose continuous audio-visual presentation for all blocks since it allows for deriving the impact of the unimodal qualities (AQ, VQ) on the AVQ as well as the mutual influences of the audio and video quality directly with the same method. Furthermore, the design in which the AVQ had to be evaluated first was meant to ensure a more heuristic judgment of the audio-visual quality as there was no influence by previous evaluations of unimodal qualities.

Before each block, participants completed a baseline training in which they were familiarised with the relevant quality range in a given block. For example, the training for the block in which the AQ had to be evaluated, the applied low and high quality level was presented alternately and three times each (3 training units including 6 trials each). The trainings were performed on similar but not identical music videos (other clips of the same recordings). The quality ratings were given on a discrete 7-point scale with a minimum value 1 representing 'very poor' and a maximum value 7 representing 'very good' quality. Participants were not informed about the applied number of quality levels. The test design, presentation and quality judgments were realised in E-Prime 2 Pro. The test took place under constant and controlled conditions in the project room of the institute. The audio signal was produced by an IDT Definition Audio Codec soundcard and Beyerdynamik DT-880 Pro headphones. The video signal was presented by a 24-inch Dell U2412M monitor with black background (RGB = 0, 0, 0) and generated by an Intel HD Graphic 4000 graphics card. The distance of the participants to the monitor was between 60 – 80 cm and the sound level of the audio signals was kept the same across participants.

3 Statistical Analysis and Results

3.1 Judgments of audio-visual quality

The first question aimed at finding out whether the audio-visual quality impression of music videos is more influenced by the video quality or audio quality. For this purpose, a repeated-measure ANOVA was carried out with two factors (VQ, AQ) and three levels each (low, medium, high quality). Beforehand, the judgments of the audio-visual quality were averaged across the three clips in

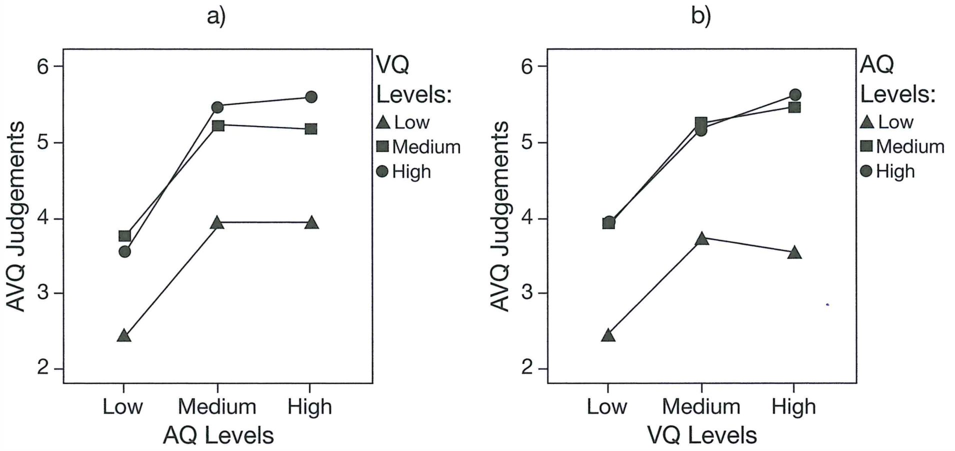

Fig. 1:
Estimated marginal means for judgments of audio-visual quality (MOS_AVQ)
Note: Figure 1a shows the means of the AVQ judgements for the factor VQ dependent on AQ (x-axis). Figure 1b shows the same judgements for AQ dependent on VQ (x-axis).

order to obtain MOS_AVQ. The main effects of the ANOVA (Greenhouse-Geisser-corrected) show that the factor VQ (*F* [1.41, 33.84] = 59.34, $p < .001$, $\eta^2 = .71$) as well as the factor AQ (*F* [1.37, 32.79] = 54.93, $p < .001$, $\eta^2 = .70$) influenced AVQ. The interaction between these two factors showed no significant result ($p = .089$). Bonferroni post-hoc tests were carried out in order to assess whether the individual quality levels were distinguishable. The pairwise comparison of the means for the factor VQ showed that the low and medium quality levels clearly differed in their ratings ($p < .001$). In contrast, the means of the medium and high level did not differ ($p > .05$) (see Figure 1a). The results for the levels of factor AQ look similar. The means between the low and medium quality levels differed ($p < .001$), yet no difference between the means of medium and high quality level was found ($p > .05$) (see Figure 1b).

3.2 Judgments of unimodal qualities

The second hypothesis was whether the unimodal qualities influence each other and if so, how strong the effect of one modality would be on the other. To answer this question, two repeated-measure ANOVAs for the judgments of the video and audio quality were carried out as before, including the same factors and levels. First, the results for judgments of the video quality will be reported followed by the results for the audio quality.

To obtain the MOS_VQ, judgments of the video quality were averaged over the three clips as well. The results show that the factor VQ showed the strongest effect with *F* [1.43, 34.34] = 461.12, $p < .001$, $\eta^2 = .95$. The result for the factor AQ showed no significant effect on judgments of the video quality ($p > .05$). Thus, no effect of the audio quality on perceived video quality was found and no interaction between these factors ($p > .05$). The post-hoc test for the pairwise comparisons of the means of VQ levels shows that the judgments of low and medium ($p < .001$) as well as the medium and high quality ($p < .001$) clearly differ (see Figure 2a). This result is in contrast to the judgments of MOS_AVQ, which differed at the medium and high VQ level.

To analyse the effect of video quality on the audio quality, judgments of audio quality were also averaged over all three clips to obtain the MOS_AQ. The results of the ANOVA for the factor AQ showed a significant main effect (*F* [1.45, 34.86] = 155.27, $p < .001$, $\eta^2 = .87$). The analysis for factor VQ showed that the judgment of audio quality was affected by the video quality (*F* [1.89, 45.29] = 4.45, $p < .05$, $\eta^2 = .16$). The effect is small but still evident. The interaction between AQ and VQ was not significant ($p > .05$). The post-hoc test for the AQ levels showed that the means of low and medium qualities differ ($p < .001$) (see Figure 2b). The means for the medium and high qualities were not significantly different ($p > .05$). Since the ANOVA resulted in an effect of the VQ factor on the judgments of the audio quality, it is of interest which quality levels were distinguished by participants. The pairwise comparisons for VQ levels show that only the means for the low and high levels of VQ differ ($p < .05$), indicating that these two levels influenced the perceived audio quality the most.

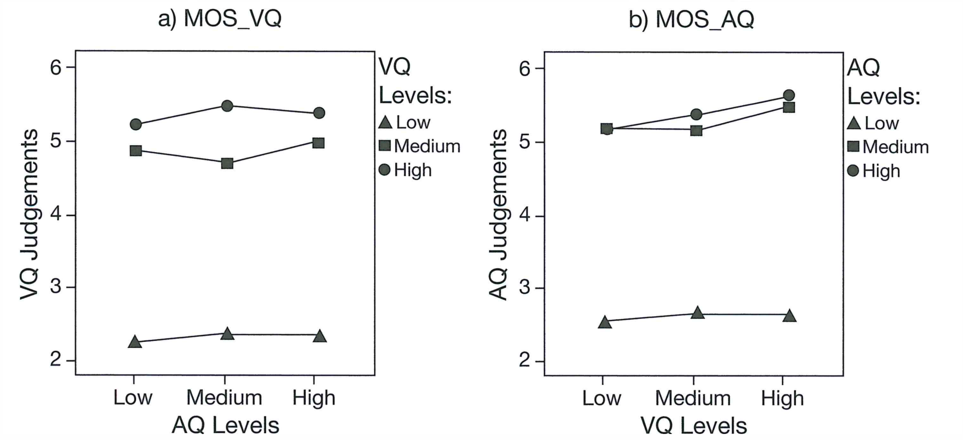

Fig. 2:
Estimated marginal means for judgment of audio and video qualities

3.3 Further Results

According to Pinson et al. (2011) the range between the highest and poorest perceived quality is an important aspect since a range in favour of one modality could affect the assessment of audio-visual quality. Therefore, the ratio of the subjective ranges for the perceived audio and video quality (MOS_AQ, MOS_VQ) has been calculated according to their method. The result for the calculation max(AQ)–min(AQ) to max(VQ)–min(VQ), in which (max) stands for the applied high qualities and (min) for low qualities, shows a ratio of 92.05 percent in favour of video quality. This result is more acceptable compared to the ratios of the studies cited in their paper (47 %–106 %). So the effect of uneven perceptual quality ranges was relatively low in the current experiment.

Regarding individual video clips, we were interested in possible differences in participants' overall assessments. Therefore, audio quality judgments were averaged over all audio quality levels and video quality judgments over all video quality levels to gain the mean AQ and VQ for each clip. Results showed that Clip_2 (*M*=4.68, *SD*=0.53) had the highest AQ rating, Clip_1 a similar rating (*M*=4.63, *SD*=0.83), and Clip_3 the lowest AQ rating (*M*=3.93, *SD*=0.711). Comparing the results with the intensity parameters (see Table 1), it becomes evident that the clip with the lowest AQ rating (Clip_3) also had the lowest dynamic range (*SD* of intensity) of all three clips. Results for the ratings of VQ for each clip indicate that Clip_3 (*M*=4.42, *SD*=0.52) and Clip_1 (*M*=4.41, *SD*=0.76) had similar ratings, whereas Clip_2 showed the lowest VQ rating over all quality levels (*M*=3.71, *SD*=0.53). Comparing the ratings with the QoM parameters in Table 1, Clip_2 had the highest QoM value, hence the highest image complexity.

4 Discussion

In contrast to previous studies, the results obtained for the MOS_AVQ do not support assumptions that video quality is the dominant factor in the perception of audio-visual quality. Our results rather suggest a similar influence of the video and audio quality on audio-visual quality perception. Thus, we could show that a possible unimodal difference in influence of audio and video quality on the overall quality impression of music videos is rather small. However, all results of this experiment should be seen in the context of the relatively low number of participants, which is a limitation of this experiment.

With regard to the mutual influence of the unimodal qualities, results indicate a cross-modal influence of the video quality on perceived audio quality. Previous studies suggest a small but noticeable influence for both modalities on each other (Beerends & De Caluwe, 1999; Kitawaki et al., 2005; ITU SG 12 Contribution COM 12.61-E, 1998). The results suggest that the perceived video quality was not influenced by the audio quality, which should not exclude other possibilities. Rather, it indicates that the perceived audio quality is more dependent on video quality. Since more attention is given to the audio quality of music videos, this result supports even more notions of unequal mutual influences of unimodal qualities. A possible impact could be that participants seemed to have difficulties to distinguish between the medium and high audio quality level as the results of the mean comparison for MOS_AVQ and MOS_AQ suggest. Yet this would raise an interesting question for further research. The quality perception of audio signals could be more easily influenced by a simultaneously presented video signal than vice versa. To verify such hypotheses, further experiments are necessary. For music videos, one possible approach would be to control stimuli according to the presented parameters in a study.

Regarding these parameters, differences in averaged judgments for AQ and VQ of each individual clip suggest partially similar results to studies reported above. Ruzanski (2006) suggested that quality differences of audio signals with low dynamic range are less perceivable. In our study AQ of the clip with the lowest dynamic range showed the lowest rating, which would be contrary to the results of Ruzanski. The averaged VQ ratings showed that the clip with the highest image complexity was rated the lowest, which is in consensus with reported studies stating that image complexity is crucial for compressed video signals and the perceived quality. However, no further conclusions can be drawn from our results since more video clips are needed to statistically analyse the impact of these parameters as factors of quality perception. Besides that, more quality levels are necessary to specify the effect size. It would also be of interest to vary stimuli for different musical genres, since not many experiments addressed these aspects in an audio-visual context.

Despite the limitations, the test design showed to be well-suited to address these questions. Whether or not it is useful to completely randomise the presentation blocks is a topic open for further discussion. We decided to gather the audio-visual quality judgments first since the decision making process involves a shared focus on both modalities. Consequently, serial order effects cannot be fully excluded; however we think that the mechanism of evaluating audio-visual quality itself rules out a task-specific emphasis on one modality. We assumed that this order leads to more heuristic judgments of the overall quality impression and thus a unimodal bias caused by this order is unlikely.

This study is a contribution to the broad field of quality perception and highlights the particular role of music videos. It especially raises the need to study the mutual influence of perceived audio and video qualities on each other for music videos as well as other content types. Audio-visual quality perception is a major aspect for possible improvements in user experiences of media content. In recent years, usage of audio-visual media increased also suggesting further investigation of this research field.

References

Adde, L., Helbostad, J.L., Jensenius, A.R., Taraldsen, G. & Støen, R. (2009). Using computer-based video analysis in the study of fidgety movements. *Early human development, 85* (9), 541–547. http://doi.org/10.1016/j.earlhumdev.2009.05.003

Aldridge, R., Davidoff, J., Ghanbari, M., Hands, D. & Pearson, D. (1995). Measurement of scene-dependent quality variations in digitally coded television pictures. *IEE Proceedings-Vision, Image and Signal Processing, 142* (3), 149–154. http://doi.org/10.1049/ip-vis:19951937

Beerends, J.G. & De Caluwe, F.E. (1999). The influence of video quality on perceived audio quality and vice versa. *Journal of the Audio Engineering Society, 47* (5), 355–362.

Dixon, N.F. & Spitz, L. (1980). The detection of auditory visual desynchrony. *Perception, 9* (6), 719–721. http://doi.org/10.1068/p090719

Ellermeier, W. & Hellbrück, J. (2008). Hören – Psychoakustik – Audiologie. In S. Weinzierl (Hrsg.), *Handbuch der Audiotechnik* (S. 41–86). Berlin: Springer Science & Business Media.

Fredrickson, B.L. & Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of personality and social psychology, 65* (1), 45. http://doi.org/10.1037/0022-3514.65.1.45

Garcia, M.N., Schleicher, R. & Raake, A. (2011). Impairment-factor-based audiovisual quality model for IPTV: influence of video resolution, degradation type and content type. *EURASIP Journal on Image and Video Processing*. http://doi.org/10.1155/2011/629284

Hammerschmidt, D. & Wöllner, C. (2015). The influence of image compression rate on perceived audio quality in music video-clips. In R. Timmers, N. Dibben, Z. Eitan, R. Granot, T. Metcalfe, A. Schiavio & V. Williamson (Eds.), *Proceedings of ICMEM 2015. International Conference on the Multimodal Experience of Music*. Sheffield: HRI Online Publications. Retrieved from https://www.hrionline.ac.uk/openbook/chapter/ICMEM2015-Hammerschmidt

Hollier, M.P. & Rimell, A.N. (1998, September). *An experimental investigation into multi-modal synchronization sensitivity for perceptual model development*. Paper presented at the 105th Audio Engineering Society Convention, San Francisco, CA.

Hollier, M.P., Rimell, A.N., Hands, D.S. & Voelcker, R.M. (1999). Multi-modal perception. *BT Technology Journal, 17* (1), 35–46. http://doi.org/10.1023/A:1009666623193

ITU-R BS. 1387–1. (2001). *Method for objective measurements of perceived audio quality*. Geneva: ITU Telecommunication Standardization Sector. Retrieved from http://www.itu.int/rec/R-REC-BS.1387/en

ITU SG 12 Contribution 61 COM 12–61-E. (1998). *Study of the influence of experimental context on the relationship between audio, video and audiovisual subjective qualities*. Geneva: ITU Telecommunication Standardization Sector.

Jost, C., Klug, D., Schmidt, A., Reautschnig, A. & Neumann-Braun, K. (2013). Einleitung. Zur historischen, ästhetischen und systematischen Verortung des Musikvideos als paradigmatischen Fall der Audiovision. In C. Jost, D. Klug, A. Schmidt, A. Reautschnig & K. Neumann-Braun (Hrsg.), *Computergestützte Analyse von audiovisuellen Medienprodukten* (Qualitative Sozialforschung, Vol. 22, S. 7–17). Wiesbaden: Springer.

Kitawaki, N., Arayama, Y. & Yamada, T. (2005). Multimedia opinion model based on media interaction of audio-visual communications. *Proceedings of the 4th International Conference on Measurement of speech and Measurement of Speech and Audio Quality in Networks* (MESAQIN 2005), 5–10.

Lerch, A. (2008). Bitratenreduktion. In S. Weinzierl (Hrsg.), *Handbuch der Audiotechnik* (S. 849–884). Berlin: Springer Science & Business Media.

Liu, C.M., Hsu, H.W. & Lee, W.C. (2008). Compression artifacts in perceptual audio coding. Audio, Speech, and Language Processing. *IEEE Transactions on audio, speech and language processing, 16* (4), 681–695. http://doi.org/10.1109/TASL.2008.918979

Massaro, D.W., Cohen, M.M. & Smeele, P.M. (1996). Perception of asynchronous and conflicting visual and auditory speech. *Journal of the Acoustical Society of America, 100* (3), 1777–1786. http://doi.org/10.1121/1.417342

Mitchell, J.L. (1997). *MPEG video compression standard*. Berlin: Springer Science & Business Media. http://doi.org/10.1007/978-1-4899-4587-7

Noll, P. (1997). MPEG digital audio coding. *IEEE signal processing magazine, 14* (5), 59–81. http://doi.org/10.1109/79.618009

Pinson, M., Ingram, W. & Webster, A. (2011). Audiovisual quality components. *IEEE Signal Processing Magazine, 6* (28), 60–67. http://doi.org/10.1109/MSP.2011.942470

Pras, A., Zimmerman, R., Levitin, D. & Guastavino, C. (2009, October). *Subjective evaluation of mp3 compression for different musical genres*. Paper presented at the 127th Audio Engineering Society Convention, New York. Retrieved from https://www.researchgate.net/publication/257068576_Subjective_Evaluation_of_MP3_Compression_for_Different_Musical_Genres

Redelmeier, D.A. & Kahneman, D. (1996). Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *Pain, 66* (1), 3–8. http://doi.org/10.1016/0304-3959(96)02994-6

Ries, M., Crespi, C., Nemethova, O. & Rupp, M. (2007). Content based video quality estimation for H. 264/AVC video streaming. *2007 IEEE Wireless Communications and Networking Conference,* 2668–2673. http://doi.org/10.1109/WCNC.2007.496

Ruzanski, E.P. (2006). Effects of MP3 encoding on the sounds of music. *IEEE Potentials, 25* (2), 43–45. http://doi.org/10.1109/MP.2006.1649011

Schulze, J. (2006). Kompressionsverfahren für Video und Audio. In R. Schmitz, R. Kiefer, J. Maucher, J. Schulze & T. Suchy (Hrsg.), *Kompendium Medieninformatik: Mediennetze* (Vol. 1, S. 1–82). Berlin: Springer Science & Business Media.

Strutz, T. (2009). *Bilddatenkompression: Grundlagen, Codierung, Wavelet, JPEG, MPEG, H.264* (S. 1–17). Berlin: Springer. http://doi.org/10.1007/978-3-8348-9986-6

Winkler, S. (2005). Vision. In S. Winkler (Hrsg.), *Digital video quality: vision models and metrics* (S. 5–34). Hoboken, NJ: John Wiley & Sons.

You, J., Reiter, U., Hannuksela, M. M., Gabbouj, M. & Perkis, A. (2010). Perceptual-based quality assessment for audio-visual services: A survey. *Signal Processing: Image Communication, 25* (7), 482–501. http://doi.org/10.1016/j.image.2010.02.002

Yuen, M. & Wu, H. R. (1998). A survey of hybrid MC/DPCM/DCT video coding distortions. *Signal Processing, 70* (3), 247–278. http://doi.org/10.1016/S0165-1684(98)00128-5