Volume 25
Issue 7
August 2010
ISSN 0923-5965
ELSEVIER
SIGNAL PROCESSING
IMAGE
COMMUNICATION
Special Issue on Image and Video Quality Assessment
Guest Editors: Stefan Winkler, Frederic Dufaux, Dominique Barba, Vittorio Baroncini
Theory, Techniques & Applications
A publication of the European Association to Signal Processing (EURASIP)
Available online at www.sciencedirect.com
ScienceDirect
EURASIP

Signal Processing: *Image Communication* 25 (2010) 482–501

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

Perceptual-based quality assessment for audio–visual services: A survey[☆]

Junyong You [a,*,♣], Ulrich Reiter [a], Miska M. Hannuksela [b,♣], Moncef Gabbouj [c,♣], Andrew Perkis [a,♣]

[a] *Centre for Quantifiable Quality of Service in Communication Systems,[1] Norwegian University of Science and Technology, Elektro E, O.S. Bragstads Plass 2E, 7034 Trondheim, Norway*
[b] *Nokia Research Center, Tampere, Finland*
[c] *Department of Signal Processing, Tampere University of Technology, Finland*

ARTICLE INFO

Article history:
Received 3 May 2009
Accepted 21 February 2010

Keywords:
Objective quality metric
Subjective quality assessment
HVS
Psychophysical approach
Engineering approach
Perception
Alignment
PEAQ
Semantic importance

ABSTRACT

Accurate measurement of the perceived quality of audio–visual services at the end-user is becoming a crucial issue in digital applications due to the growing demand for compression and transmission of audio–visual services over communication networks. Content providers strive to offer the best quality of experience for customers linked to their different quality of service (QoS) solutions. Therefore, developing accurate, perceptual-based quality metrics is a key requirement in multimedia services. In this paper, we survey state-of-the-art signal-driven perceptual audio and video quality assessment methods independently, and investigate relevant issues in developing joint audio–visual quality metrics. Experiments with respect to subjective quality results have been conducted for analyzing and comparing the performance of the quality metrics. We consider emerging trends in audio–visual quality assessment, and propose feasible solutions for future work in perceptual-based audio–visual quality metrics.

1. Introduction

Multimedia services are experiencing a tremendous growth in popularity recently due to the evolution of digital communication systems. Two main media modalities, namely audio and video signals constitute the core content in most digital systems. Quality of audio–visual signals can be degraded during lossy compression and transmission through error-prone communication networks. Consequently, accurately measuring the quality of distorted audio–visual signals plays an important role in digital applications, for example, when evaluating the performance of codecs and networks, helping to improve the coding abilities or adjusting network settings based on a strategy of maximizing the perceived quality at the end-user.

Subjective assessment of audio–visual quality is considered to be the most accurate method reflecting the human perception [1]. It is, however, time-consuming and cannot be done in real time. Thus, the International Telecommunication Union (ITU) has released requirements for an objective perceptual multimedia quality model [2]. Currently, most studies regarding the understanding of human quality perception of multimedia systems have focused on individual modalities, i.e., audio and video separately. These investigations have led to a considerable progress in developing objective models based on the human perceptual system for both audio and video. A brief introduction to signal-driven audio and

[☆] This work was supported in part by the Academy of Finland (application number 213462, Finnish Program for Centres of Excellence in Research 2006–2011) while Junyong You worked at Tampere University of Technology.

[1] Centre for Quantifiable Quality of Service in Communication Systems, Centre of Excellence, appointed by The Research Council of Norway, funded by the Research Council, NTNU and UNINETT.

[*] Corresponding author. Tel.: +47 73592617.
E-mail address: junyong.you@q2s.ntnu.no (J. You).

[♣] EURASIP member.

0923-5965/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.image.2010.02.002

video quality metrics is given in the following paragraphs and the main focus of this paper is on full reference quality models for general audio and video signals.

1.1. Audio quality assessment

Perceptual audio quality assessment has been investigated for several decades. Most audio quality models are designed for handling coding distortions only. This paper will also focus on audio quality metrics for coding distortions. Traditional objective measurement methods, such as signal-to-noise ratio (SNR) or total harmonic distortion (THD), have never really been shown to relate reliably to the perceived audio quality. A number of methods for making objective perceptual assessment of audio quality have been developed as the ITU identified an urgent need to establish a standard in this area. The level-difference between the masked threshold and the noise signal is evaluated in a noise-to-masked ratio (NMR) measurement method presented by Brandenburg [3]. In the method proposed by Beerends and Stemerdink. [4], the difference in intracranial representations of the reference and distorted audio signals was transformed with a cognitive mapping to the subjective perceptual audio quality. A perceptual evaluation developed by Paillard et al. [5] first modeled the transfer characteristics of the middle and inner ear to form an internal representation inside the head of the subject, which is an estimate of the information being available to the human brain for comparison of signals, and the difference between the representations of the reference and distorted signals was taken as a perceptual quality. By comparing internal basilar representation of the reference and distorted signals, a perceptual objective measurement (POM) proposed by Colomes and Rault. [6] quantified a certain amount of degradations including the probability of detecting a distortion and a so-called basilar distance. Sporer introduced a filter bank with 241 filters to analyze and compare the reference and distorted signals in [7]. A perceptual measurement method (DIX: disturbance index) proposed by Thiede and Kabit. [8] is based on an auditory filter bank that yields a high temporal resolution and thus enables a more precise modeling of temporal effects such as pre- and post-masking. These six perceptual models [3–8] combined with some toolbox functions were integrated into the ITU recommendation BS.1387 [9]. In this recommendation, the method for objective measurement of perceived audio quality (PEAQ) is used to predict the perceived quality for wide-band audio signals with small impairment.

However, some limitations have been discovered in PEAQ. Most notably PEAQ is shown to be unreliable for signals with large impairment resulting from low bitrate coding [10]. Furthermore, PEAQ is limited to a maximum of two channels. Consequently, improvements in PEAQ have been developed. Barbedo and Lopes. [11] proposed a new cognitive model and new distortion parameters. The limitation of PEAQ up to a maximum of two channels was addressed by the development of an expert system to assist with an optimization of multichannel audio system [12]. Creusere et al. [10,13] presented an energy equalization quality metric (EEQM), which can be used in predicting the audio quality for a wide range of impairments. Furthermore, a variable called the energy equalization threshold (EET), used in EEQM, can also be appended in PEAQ as a complementary model output variable (MOV) to give a more accurate quality prediction [14].

1.2. Video quality assessment

A widely used objective video quality metric, peak signal-to-noise ratio (PSNR), has been found to correlate inaccurately with the perceived quality, since it does not take the characteristics of the human visual system (HVS) into account [15]. A number of objective methods for measuring the perceived video quality have been proposed, and many of them have been studied by the video quality experts group (VQEG) [16]. Different validation phases conducted by VQEG between 1997 and 2008 have helped the ITU in producing two recommendations for objective video quality assessment using full-reference models. ITU-T J.144 [17] recommends models for digital television pictures (i.e. coding impairments) and ITU-T J.247 [18] is intended for multimedia-type video (QCIF, CIF and VGA) transmitted over error-prone networks (i.e. coding impairments and transmission errors). The conducted validations are also reported in VQEG reports [19–21].

Objective video quality metrics are generally classified into three categories based on the availability of reference information: full-reference (FR), reduced-reference (RR), and no-reference (NR) [22]. The FR metrics have access to the reference signal. They have been studied widely and usually have the best performance in predicting the perceived quality, but the drawback is that they cannot be used for all services, for example IPTV monitoring. The following aspects are typically considered in a typical HVS-based FR quality metric: color processing, multichannel decomposition, perceived contrast and adaptation to a specific luminance or color, contrast sensitivity, spatial and temporal masking, and error pooling over various channels within the primary visual cortex [23]. The perceptual distortion metric (PDM) proposed by Winkler [24] exploited main elements of an HVS-based model and exhibited a promising performance in VQEG FR-TV Phase I evaluation. RR metrics analyze a number of quality features extracted from the reference and distorted videos and integrate them into a single predictive result. For example, Lee et al. [25] extracted a few edge pixels in each frame, and used them to compute PSNR around edge pixels. The task of NR metrics is very complex as no information about the reference medium is available. Therewith, an NR method is an absolute measurement of features and properties in the distorted video. Most NR metrics currently focused on certain distortion features, such as blockiness [26], blurring [27], and the analysis of coding parameter settings [28].

1.3. Audio–visual quality evaluation

Compared to the extensive studies on quality assessment of individual modalities, relatively little work on joint audio–visual quality assessment has been performed. Fundamental research on multi-modal perception is required to study the mutual influence between auditory and visual stimuli as well as other influence factors in audio–visual quality assessment [29]. Some experiments, reviewed below, have demonstrated that there is a significant mutual influence between the auditory and the visual domain in the perceived overall audio–visual quality. To explain the relationship between audio quality (AQ), video quality (VQ), and the overall audio–visual quality (AVQ), five combinations of stimulus types and assessment tasks, presented in Table 1, have been suggested. In the Table, AQ_V denotes the audio quality in the presence of a visual stimulus, and VQ_A denotes the video quality in the presence of an auditory stimulus. Earlier studies have shown that when a combined auditory–visual stimulus was given, the judgment of the quality in one modality was influenced by the presence of the other modality [30,31]. Some other experiments have been conducted to study how to derive AVQ based on single AQ and VQ. In these experiments, three subjective assessments corresponding to AQ, VQ, and AVQ in Table 1 have been conducted. Most studies have shown that VQ dominates AVQ in general [30,32,33], while Hands [34] suggested that AQ is more important than VQ in a teleconference setup, because the human attention is mainly focused on the auditory stimulus. Winkler's experiments [35,36] have shown that more bits should be allocated to audio to achieve a higher overall quality in very low total bitrate budgets. Moreover, the relationship between AQ, VQ, and AVQ is also influenced by some other factors, such as attention of subjects, the audio–visual content itself, usage context, and the experiment environment [37,38]. It has been proposed that the overall audio–visual quality can be derived by a linear combination and a multiplication of AQ and VQ, where the multiplication of AQ and VQ has very high correlation with the overall quality [33].

When assessing the overall audio–visual quality, synchronization between audio and video (e.g. lip sync) may be an important issue. It is known that the perception of asynchrony depends on the type of content and on the task [39]. The overall quality is degraded when audio and video do not form a single spatio-temporal coherent stream. Dixon and Spitz [40] claimed that the perceived quality degrades rapidly when asynchrony is increased. Massaro et al. [41] also reported that intelligibility is decreased when audio and video are not in sync. When the audio is about 150 ms earlier than the video, it was found that subjects might find the asynchrony annoying [42,43]. When the video is earlier than the audio, the same degradation is perceived for asynchronies that are about twice when the audio is earlier than the video. A large number of methods of audio and video synchronization have been proposed [44,45]. In this survey paper, we will introduce related issues in audio–video synchronization briefly, while we will mainly focus on objective quality metrics for spatiotemporally coherent audio–visual systems.

Another issue for developing audio–visual quality metrics, which cannot be neglected is the relationship between the semantic importance of audio–visual contents and the perceived quality. Only little work has been done on this topic. As mentioned earlier, the mutual influence of AQ and VQ is related to audio–visual contents, for example, it can be assumed that AQ is more important than VQ for typical teleconference sequences. The visual content consists principally of head and shoulders of speakers, while the audio content is semantically more important as it may convey more information. Moreover, there is a significant relation between semantic audio–visual importance and the perceived quality. For instance, when same quality degradation occurs on two audio–visual segments with different importance levels, subjects might give different quality judgments for these two segments. Although most existing quality metrics take into account the audio and video contents latently, the relationship between semantic audio–visual importance and the perceived quality has not been studied adequately. There are two challenging problems for integrating semantic importance into quality assessment. Firstly, semantics is a very subjective concept; so it is a challenging task to construct a generic semantic importance model for audio–visual contents. Existing semantic analysis methods are mainly focused on certain types of multimedia contents, such as sports video. They typically exploit audio–visual features, for example, we have proposed a semantic analysis framework for video contents based on visual perception [46]. Secondly, because semantics is a strong subjective concept, it is difficult to define an order of semantic importance among pieces of audio–visual contents. For example, a sports sequence might be important for sports fans, whereas a child may think a cartoon sequence is more important. Thus, rather than comparing different content items, the semantic importance of different temporal segments in an audio–video sequence is usually compared. Taking an example of a football sequence, goal segments are usually important for most subjects. Consequently, the quality scoring on the goal segments is potentially different from other scenes in the same sequence.

Table 1
Quality assessment for five different presentations.

Stimuli	Assessment	Quality abbreviation
Audio only	Audio quality	AQ
Audio+Video	Audio quality	AQ_V
Video only	Video quality	VQ
Audio+Video	Video quality	VQ_A
Audio+Video	Audio–visual quality	AVQ

1.4. *Goals and organization of the paper*

As a basis of developing an audio–visual quality model, we first study the popular objective quality metrics for single auditory and visual modality, and analyze the characteristics and performance of these metrics. Some issues related to audio–visual quality assessment are then studied, including audio–video synchronization, the interactive process of multi-modalities (i.e. audio and video in this paper), semantic multimedia analysis and its relation with quality assessment, temporal averaging methods for long-term audio–visual sequences, etc. In addition, we briefly review subjective quality assessment methodologies for audio–visual signals. To make this survey paper more specific, we will focus on the general audio signals and FR models.

The rest of this paper is organized as follows: we introduce alignment methods for audio signals, video signals, and audio–video synchronization in Section 2. The audio quality metrics, especially PEAQ, are introduced in Section 3. Section 4 introduces some well-known FR and RR video quality metrics and presents the algorithm analysis and experimental results. Subjective quality assessment methodologies for audio–visual signals are reviewed in Section 5. The mutual influence of audio , video and audio–visual qualities, as well as some relevant issues and trends of audio–visual quality assessment, are investigated in Section 6. Finally, some conclusions are drawn in Section 7.

2. Alignment for audio–video signals

Alignment between distorted audio–visual signals and original signals has a significant influence on quality assessment. Slight misalignment may not affect the subjective quality evaluation by human, while it will greatly reduce the accuracy of an objective quality metric. In addition, the audio–video synchronization is another important issue for audio–visual quality assessment. This section will investigate briefly the related issues in the alignment of audio–video signals, and approaches for audio–video alignment used in our experiments will be introduced.

2.1. Audio alignment

Alignment between distorted and original audio signals is an important issue for audio quality evaluation. The misalignment is mostly perceptually irrelevant, but can affect an objective quality measure considerably. Both coding and transmission errors cause a time delay between the distorted and original audio signals. For example, in advanced audio coding (AAC), the time delay is one frame because the psychoacoustic model needs to estimate what block/window switch will be performed. For parametric stereo (PS) and spectral band replication (SBR) in the high-efficiency AAC (HE-AAC), a filter delay is required for quadrature mirror filter (QMF) and hybrid analysis [47]. So, once the mode of PS is determined, the time delay will be constant and hence it only needs to be estimated once. However, transmission errors may cause different delays for different packets, and a finer time alignment block by block is required in this case. In addition, level alignment can be done by scaling a test signal with a constant factor that is chosen to result in equal overall rms values. This way of level alignment will work if the overall gain of the system has no long-term drifts. Although time alignment is not an integral part of PEAQ, it can be achieved in other ways.

A commonly used approach is to compute the cross-correlation function of the temporal envelopes between the distorted and original signals. The delay of one signal relative to the other is given by the position of the maximum of the correlation function. Alternatively, audio signals can be aligned visually using specialized software. In our experiments, we used the cross-correlation function to estimate a coarse delay. Subsequently, a software named Sonic Visualiser with MATCH Vamp plugin based on an efficient dynamic time warping algorithm [48] was employed for validating the coarse delay estimation, and for performing a more accurate visual alignment.

2.2. Video alignment

Analog to digital video conversion, video coding, and transmission errors may introduce spatial, temporal, and gain/level shift misalignment between the distorted and original videos. Video alignment has to be considered in FR and RR objective video quality assessment, because the quality is based on a comparison between the distorted and original videos. Normally, misalignments are assumed to be constant for short video sequences and hence need to be calculated once [49], if no temporal impairment is taken into account. However, if transmission error and therefore temporal impairment are introduced, the situation becomes complicated, as the misalignment will vary temporally. Any metrics should include a video alignment step, which can be performed either independently of the quality estimation algorithm or using some of the extracted features in the metrics. ITU-T J.247 [18] also includes corresponding alignment methods for different objective quality metrics. For example, the NTT model moves a video frame of the distorted video to find the pixel and frame shifts that give the minimum difference in terms of luminance between the distorted and reference videos. The Yonsei model in J.247 uses extracted edge pixels in aligning the distorted and reference videos by minimizing MSE over all edge pixels in a sliding window. Generally, most alignment methods try to search for suitable spatial, temporal and gain/level shifts to maximize the correlation between the distorted and the aligned reference videos [50,51].

2.3. Audio–video synchronization

In systems reproducing audio–visual content, play back of synchronized auditory and visual stimuli is considered mandatory. Interestingly, the detection thresholds for unsynchrony are not temporally symmetric. Hollier and Rimell. [29,52] have performed a number of experiments focusing on audio–visual communications systems to examine this temporal asymmetry with different types of stimuli. They compared a talking head audio–visual scene with a bouncing pen scene and an audio–visual stimulus in which an ax hits an object a single time. They concluded that the general trend in error detection asymmetry is apparent for all stimulus types. Furthermore, the distinctness of the ax stimulus results in a greater probability of detection than for the pen stimulus. For the talking head stimulus, the error detection rate is consistent with the other stimuli when the audio lags behind the video, but greater than either the ax or the pen stimuli when the audio leads the video. Apparently, test subjects compared the artificial stimuli presented in the lab with real life experiences. In real life, due to the physical nature of different traveling speeds of sound and light, audio can never lead the visual percept.

These findings on synchronization error detection asymmetry are also reflected in the recommended synchronization thresholds given in ITU-T J.100 [53], which are 20 ms for audio lead and 40 ms for audio lag. The recommendation suggests these fixed values for all types of television content and is intended to ensure that synchronization errors remain imperceptible for all possible varieties of content. This relatively small threshold means that the human perceptual system is generally quite sensitive to errors in synchrony.

In the field of perceptual psychology, interaction between the aural and visual modalities is well documented. A rather large number of singular effects have been scientifically researched. Because in these experiments often extremely simple auditory and visual stimuli are used, it is often hard to extrapolate the results for more complex media applications. Therefore, it is interesting to look at the field of film music investigations. A number of models for film music perception have been proposed. Lipscomb e.g. [54] suggested that there are at least two implicit judgments made during the perceptual processing of the movie experience: an association judgment and a mapping of accent structures. According to their model, the mapping of accent structure is determined by the consistency with which important events in the musical score coincide with important events in the visual scene – the synchrony between the two modalities. When test subjects were presented with extremely simple auditory and visual stimuli, accent alignment played an important role in the determination of ratings of effectiveness. As the stimuli became more complex, the importance of accent structure alignment appeared to diminish and the association judgment assumed a dominant role. This means that apparently synchronization loses its importance with growing complexity of the audio–visual stimuli.

3. Perceptual evaluation of audio quality

Objective audio quality models that incorporate properties of the human auditory system have existed since the 1970s and were mainly applied to speech codecs. As mentioned earlier, a number of psychoacoustic models have been proposed to measure the perceived audio quality, and six of them [3–8] were extracted and integrated by the ITU into the standard method, PEAQ [9]. Although there are many other metrics that have good performance, we concentrate on PEAQ and the improved methods. The first reason is that we mainly focus on the wide-band audio rather than the narrow-band speech; so we do not consider the numerous speech quality metrics even though they correlate well with the subjective speech quality assessment. Second, it is proven that PEAQ is good enough for evaluating the performance of audio codecs and transmission networks, and the ITU has no plans for a replacement to this model at the present time. In this section, we will introduce the main principle of PEAQ and the experimental results on a conformance test and two sample rates: 48 and 44.1 kHz. In addition, the experimental results of improved PEAQ and some other audio quality metrics, with respect to the results of a subjective quality evaluation conducted by MP3-tech [55], will also be presented.

3.1. PEAQ and experimental results

PEAQ was standardized by ITU-R BS.1387 during 1998–2001 [9], which specified a method for objective assessment of the perceived audio quality of a device under test, e.g. a high bitrate codec. PEAQ consists of two versions: one is called the basic version, which is intended for applications that require high processing speed. The other is the advanced version, which is intended for applications requiring the highest achievable accuracy [56]. The basic version uses only an FFT-based ear model, and employs both the concept of comparing internal representations and a masked threshold. The masked threshold is a threshold above which human can perceive the difference between the distorted and reference audio signals. The advanced version of PEAQ makes use of the FFT-based ear model as well as a filter bank-based ear model. The masked threshold concept is applied using the FFT-based ear model, whereas the concept of comparing internal representations is applied using the filter bank-based ear model. The basic version uses eleven model output variables (MOV), which are derived from the ear model to measure loudness of distortions, amount of linear distortions, changes in the temporal envelope, a noise-to-mask ratio, noise detection probability, and harmonic structure in the error signal. In the advanced version, five MOVs derived from the filter bank measure the loudness of non-linear distortions, the amount of linear distortions, and disruptions of the temporal envelope. Furthermore, the MOVs based on the FFT include a noise-to-mask ratio and a cepstrum-like measure of harmonic structure in the error signal. During the computation of MOVs, spectral averaging over frequency bands and temporal averaging over temporal frames are applied. Frame selection strategy is also performed on the signals and the MOVs. These MOVs extracted either from the basic or the advanced versions are mapped to a single quality index objective difference grade (ODG) by a multi-layer neural network with either 3 (basic version) or 5 (advanced version) units in a single hidden layer. The neural network is trained based on all available data, such as MPEG90, ITU93, etc. [9].

ITU provided 16 testing items including the original audio signals and their distortions for conformance test. These items were sampled at 48 kHz with 16-bit PCM. We implemented the basic versions of PEAQ in MATLAB according to the instruction of ITU BS.1387 and an examination written by Kabal [57]. Subsequently, a conformance test on the standard audio signals using our implementation and two other public software implementations of the PEAQ basic version (implementation in MATLAB by McGill University and EAQUAL software in C language), was performed. Mean absolute difference (MAD) and root mean squared error (RMSE) are calculated between the metric results and the standard ODG values provided by ITU. Table 2 presents the experimental results. ITU suggested that an acceptable tolerance interval is ±0.02 of the ODG for all test items. According to the experimental results, although our implementation of the basic version cannot produce results within this tolerance exactly for all items, it is comparable to the results of the public software implementations. So, we think that our implementation of the basic version is valid and can be used in evaluating the perceptual audio quality.

The standard PEAQ only supports audio signals at 48 kHz sample rate. However, most popular audio signals are sampled at other rates, such as 44.1 and 24 kHz. A subjective audio quality test conducted by MP3-tech used audio signals at 44.1 kHz sample rate [55]. Therefore, we tried to find the best way to use PEAQ with audio signals sampled at 44.1 kHz. We still used the standard audio signals in the experiment based on the following assumption: the down-sampling and up-sampling will not influence audio quality. Although this assumption is not very convincing, it can compare two approaches for using PEAQ on 44.1 kHz samples. The experiment is designed as follows: first, the standard 48 kHz signals are down-sampled to 44.1 kHz, and a modified PEAQ, where two parameters related to sample rate in grouping into critical bands and excitation patterns are changed accordingly, is performed on these down-sampled signals. Second, these 44.1 kHz signals are up-sampled to 48 kHz again, and the standard PEAQ implementation is then performed. So, we can compute two groups of quality results, and they are listed in the right columns in Table 2. The MAD and RMSE are still used for comparison. According to the results, the performance of the second approach is better than the first. So, we believe that up-sampling should be performed first and then the standard PEAQ algorithm is used when evaluating the quality of the audio signals at 44.1 kHz sample rate.

3.2. *Audio quality metrics for parametric coding scheme*

Although PEAQ has a credible performance for assessing the perceived audio quality, it is important to note that PEAQ is designed to evaluate reconstructed audio signals that have relatively high quality. This is made clear that ODG is designed to approximate subjective difference grade (SDG), which is determined using the testing framework described in ITU Recommendation ITU-R BS.1116 for small impairments [58]. With the wide application of parametric audio codecs, such as high-efficiency advanced audio coding (HE-AAC) versions1 and 2, developing a metric that can predict audio quality with large impairment is becoming an urgent issue.

HE-AAC is a lossy data compression for digital audio. It is an extension of low complexity AAC (AAC LC) optimized for low-bitrate applications such as streaming audio. HE-AAC v1 used spectral bank replication (SBR) to enhance the compression efficiency in the frequency domain. HE-AAC v2 couples SBR with parametric stereo (PR) to enhance the compression efficiency of stereo signals. To the best of our knowledge, there is no audio quality metric designed to handle HE-AAC v1 or v2. In this paper, we concentrate on the energy equalization approach (EEA) proposed by Creusere et al. [10,13,14,59], PEMO-Q proposed by Huber and Kollmeier. [60], as well as two other simple metrics: measuring normalizing block (MNB) [61] and mean structural similarity

Table 2
Experimental results of PEAQ (basic version).

Test items	Basic version				Sample rates	
	ITU ODG	Authors	McGill	EAQUAL	44.1 kHz	48 kHz
acodsna.wav	*−0.676*	−0.675	−0.679	−0.69	−0.531	−0.649
bcodtri.wav	*−0.304*	−0.291	−0.292	−0.27	−0.388	−0.487
ccodsax.wav	*−1.829*	−1.793	−1.797	−1.81	−1.323	−1.742
ecodsmg.wav	*−0.412*	−0.369	−0.368	−0.37	−0.280	−0.407
fcodsb1.wav	*−1.195*	−1.181	−1.168	−1.13	−0.886	−1.120
fcodtr1.wav	*−0.598*	−0.552	−0.561	−0.55	−0.540	−0.751
fcodtr2.wav	*−1.927*	−1.790	−1.788	−1.77	−1.599	−1.871
fcodtr3.wav	*−2.601*	−2.317	−2.457	−2.43	−2.137	−2.370
gcodcla.wav	*−0.386*	−0.380	−0.374	−0.38	−0.259	−0.387
icodsna.wav	*−3.786*	−3.786	−3.772	−3.77	−3.706	−3.789
kcodsme.wav	*0.038*	0.043	0.045	0.06	0.115	0.002
lcodhrp.wav	*−0.876*	−0.844	−0.834	−0.83	−0.616	−0.834
lcodpip.wav	*−0.293*	−0.064	−0.035	−0.28	−0.035	−0.111
mcodcla.wav	*−2.331*	−2.290	−2.267	−2.18	−1.633	−2.023
ncodsfe.wav	*0.045*	0.048	0.048	0.05	0.043	0.034
scodclv.wav	*−0.435*	−0.385	−0.413	−0.34	−0.369	−0.413
MAD	**–**	***0.058***	***0.054***	***0.057***	***0.225***	***0.089***
RMSE	**–**	***0.100***	***0.086***	***0.079***	***0.293***	***0.127***

(MSSIM) [62], and test their performance on audio quality assessment for the parametric coding scheme.

In EEA, a truncation threshold *T* is set on the spectrogram of an original audio signal and it is varied until the truncated spectrogram has the same energy as the spectrogram of the distorted signal. Ultimately, *T* is adjusted until the truncated version of the original spectrum has the same island-like character as the distorted signal, and thus *T* serves as a measure of how island-like the spectrum of the distorted signal is. *T* can be used as an independent measure of perceived audio quality, and it can also be integrated into PEAQ as an additional MOV. The authors of EEA believe that EEA is an effective measure of the quality in highly impaired audio [14]. Another advantage is that the time alignment is not required in EEA because misalignment between the original and distorted signals has no influence on finding the threshold *T*. However, EEA is designed and tested using music signals, and the applicability to speech and other general acoustic signals is unknown. In addition, when adding this threshold *T* as an additional MOV in PEAQ, the authors concluded that the three-layer neural network in PEAQ is unnecessary. Therefore, a least-square procedure was used in finding an optimal linear weighting for each MOV to get a scalar value estimate of audio quality, and such simple linear weighting is less susceptible to over-training and thus likely to provide a more robust quality metric [10]. Furthermore, a minimax optimal MOV selection was proposed for finding the most suitable subsets within all the MOVs (11/5 original MOVs in PEAQ basic/advanced version+threshold *T*) to minimize a cost function: the maximum absolute error between the objective metric output and the subjective quality measurement. Because the metric was proposed for a wide range of audio distortions, the subjective experiments for intermediate and large impairments measured by ITU BS.1534 (MUSHRA) [63] were also included in the optimization procedure.

Based on a psychoacoustically validated, quantitative model of the effective peripheral auditory processing, Huber and Kollmeier [60] proposed an objective assessment of the perceived audio quality by expanding a speech quality measure in [64]. To evaluate the quality of a given distorted audio signal relative to a corresponding high quality reference signal, an auditory model was employed to compute the internal representations of the signals, which were partly assimilated in order to account for assumed cognitive aspects. The linear cross-correlation coefficient of the assimilated internal representations represents a perceptual similarity measure (PSM). PSM shows a good correlation with subjective quality assessments if different types of audio signals are considered separately. A second quality measure PSMt is represented by the fifth percentile of the sequence of instantaneous PSM, which has better accuracy of signal independent quality prediction than the original PSM. Finally, PSMt can be mapped to an ODG scale by a regression function composed of a hyperbola and a linear function derived from a numerical fitting procedure.

In addition, we also tested some other simpler metrics for audio quality assessment in the parametric coding scheme, and MNB and MSSIM methods were chosen because of their simple computation. MNB [61] uses a simple but effective perceptual transformation obtained using a Bark frequency scale, and a distance measure, which consists of a hierarchy of measuring normalizing blocks. Each measuring normalizing block integrates two perceptually transformed signals over a time or frequency interval to determine the average difference across that interval. This difference is then normalized out of one signal, and is further processed to generate one or more measurements. The linear combination of these measurements gives an auditory distance. SSIM was initially proposed by Wang et al. [65] for image quality assessment, and Kandadai et al. [62] extended this idea to audio structure similarity in two ways. First, it is assumed that the structure depends on each time sample and its position with respect to a small temporal neighborhood of samples around it. Then the audio sequence can be split into temporal frames with 50% overlap. The SSIM on each frame is applied separately, and the mean SSIM over all frames is calculated as the audio quality. Second, a time–frequency transform, such as a 256-point MDCT with a 50% overlapping window, is applied to the audio sequences. The structural similarity in both the temporal and frequency domains can be evaluated by applying SSIM to the 2-dimensional blocks of the time–frequency representation.

In our experiment, the subjective results from a public audio test conducted by MP3-tech were employed [55]. Although the purpose of this test is to evaluate the quality provided by state-of-the-art MPEG audio encoders, we believe that it can also be used in evaluating the performance of audio quality metrics on the parametric coding schemes. Seven encoders in total using 18 audio signals at 44.1 kHz sample rate were tested, including: 3GPP reference encoder HE-ACC v1, Coding Technologies HE-AAC v1 and v2, Nero Digital HE-AAC v1 and v2, L.A.M.E. MP3 as a high anchor, and Apple Quick Time/iTunes LC-AAC as a low anchor. Six metrics were evaluated in our experiment, including PEAQ basic version, EEA, and the metric combining the threshold *T* in EEA as an additional MOV in PEAQ (EET_PEAQ), PEMO-Q, MNB, and the first approach of MSSIM. Two methods of EET_PEAQ were included: one used the minimax-optimal MOV selection, while the other did not. All the metrics were implemented in MATLAB. PEMO-Q [66] is a demo version in MATLAB, which only supports signals that are no longer than 4 sec; so an audio sequence is first divided into 4s segments and the overall quality is an average of model results over all segments. In addition, EEA only supports single channel signals, so the quality of stereo signals was taken as the average of model results over the two channels. The audio signals were aligned in advance using the method described in Section 2 A, and re-sampling may be required for suiting to the demand of these metrics. The subjective measure used a 5-point impairment scale (1 – very annoying, 2 – annoying, 3 – slightly annoying, 4 – perceptible but not annoying, and 5 – imperceptible). The evaluation criteria of the metrics were chosen to relate to the prediction accuracy, monotonicity, and consistency. Therefore, RMSE, the Pearson linear correlation coefficient, and the Spearman rank order correlation coefficient between the subjective results and ODG computed by the metrics were used. Averaged computation

Table 3
Evaluation results of objective audio quality metrics for parametric audio coding.

Criteria	PEAQ	EEA	EET_PEAQ (with optimization)	EET_PEAQ (without optimization)	PEMO-Q	MNB	MSSIM
RMSE	1.76	1.46	1.07	0.96	1.54	2.02	2.35
Pearson	0.478	0.577	0.659	0.683	0.614	0.439	0.401
Spearman	0.582	0.603	0.712	0.707	0.663	0.536	0.552
Time (s)	248.7	196.3	262.2	398.5	223.5	199.2	162.0

time for one sequence was taken as the criterion of metric complexity. Table 3 presents the experimental results. According to the evaluation results, EET_PEAQ achieves the best performance, while other methods are not promising for parametric coding schemes. Thus, although these metrics have been proved to be good for measuring the audio quality with small impairments. We are still a long way from developing more robust methods for a wide range of audio distortions. In addition, re-sampling may also influence the performance of these metrics, so developing robust metrics to different sample rates is still an open issue.

4. Objective video quality metrics

The goal of objective video quality metrics is to give quality predictions, which are in accordance with subjective assessment results. Thus, a good objective metric should take the psychophysical process of the human vision and perception system into account. Main characteristics of the HVS include modeling of contrast and orientation sensitivity, frequency selection, spatial and temporal pattern masking and color perception [24]. Due to their generality, HVS-based metrics can in principle be used for a wide variety of video degradations. These metrics retrospect to the 1970s and 1980s, when Mannos and Sakrison [67] and Lukas and Budrikis. [68] developed the first image and video quality metrics based on the visual model. Later, well-known HVS-based metrics include a visual difference predictor (VDP) proposed by Daly [69], Sarnoff just noticeable differences (JND) method by Lubin and Fibush. [70], PDM [24], etc. Psychophysically driven metrics usually have good performance because they mimic the formation process of the human vision and perception. However, the computation is accordingly complex due to the complexity of such a process. In addition, although the human perception for video quality is formed according to the perceptual process above, the subjective assessment may be determined by certain important aspects of the HVS. Thus, simplified metrics based on the extraction and analysis of certain features or artifacts in video have been proposed. These metrics, categorized as the engineering approach, assess how pronounced the detected artifacts are in order to estimate the overall quality. Metrics following the engineering approach do not necessarily disregard the attributes of the HVS, as they often consider psychophysical effects as well, but image content and distortion analysis rather than fundamental vision modeling is the conceptual basis for their design. In this section, we introduce main existing FR and RR metrics and present a comprehensive analysis of their computational principles, results, and complexity. Table 4 presents an overview of the introduced video quality metrics in this paper, including brief explanations, approach categories, processing units or domains, and reference numbers.

Metrics derived from the psychophysical approach usually follow the psychophysical process of the human vision and perception, where some of them consider most characteristics of this process while others just take a certain characteristic of the HVS. PDM proposed by Winkler is a representative method of the former, and it contains four steps performed on both the reference and distorted videos synchronously: color space conversion, perceptual decomposition, contrast and gain control, and error detection and pooling.

Psychophysical metrics may be simplified in two ways. First, all processes of the HVS do not need to be considered in quality computation, as in [73] and [80]. Second, only selected regions of the images may be enough to calculate the quality, because the HVS is more sensitive to severer distortions or larger degraded regions [81]. The distortions occurring in certain regions are invisible because of the masking effect. Thus, a reduction in computing cycles can be achieved based on an the analysis of image contents. Lee and Kwon [80] proposed to use a spatiotemporal wavelet transform to decompose the input video into different frequencies, and then a quality index is based on pooling the difference in the decomposed results between the reference and distorted videos. Similarly, Guo et al. [73] used a Gabor filter bank to perform the decomposition into different frequencies and orientations. As the discrete cosine transform (DCT) is widely used in video coding, a digital video quality (DVQ) metric has been proposed by Watson et al. [74] based on the human vision. MSU method [75] used DCT as the channel decomposition tool, whilst other perceptual processes, such as color transform, contrast filtering, were also integrated.

As psychophysical metrics are usually computationally complex, a number of metrics in the engineering approach category have been proposed. These kinds of metrics compare the quality features extracted from the reference and distorted videos, in which certain characteristics of the HVS are taken into account. Coding of digital video causes certain types of visual artifacts and distortions, such as blockiness, blurring, color bleeding, and ringing, and these types of distortions can be classified into the

Table 4
Overview of perceptual-based video quality metrics.

Availability of reference	Brief explanation of metrics (quality features in NR metrics)	Approaches (psychophysical/ Engineering)	Processing unit/domain	Reference
Full reference	PDM	Psychophysical	Whole image	[24]
	JND methods			[70–72]
	PEVQ	Engineering	Image blocks	[18]
	NTT model			
	Gabor difference based		Whole image	[73]
	Psytechnics model		Resized to QCIF	[18]
	DVQ	Psychophysical	DCT domain	[74]
	MSU model	Engineering		[75]
Reduced reference	PSNR based	Engineering	Whole image	[76]
	Attention based		Visual attention regions	[77–79]
	PSNR around edge areas		Whole image	[18,25]
	Wavelet transform based			[80]
	NTIA model and its simplification		Image blocks	[49,81,82]
	SSIM		Whole image/blocks	[65,83]

category of spatial distortion. In addition, packet loss and other transmission errors cause temporal distortion in the form of freezing of the latest correct frame or temporal propagation of the distortion caused by error concealment of image slices is affected by the transmission errors. Most engineering metrics are designed to detect artifacts and distortions caused by coding and transmission errors.

The NTIA model proposed by Pinson and Wolf [81] divides video sequences into spatiotemporal (S-T) blocks, and a number of features measuring the spatial gradient activity, chrominance information, contrast information and absolute temporal information in each of these blocks are computed. The features extracted from the reference and distorted videos are then compared using functions that model visual masking of the spatial and temporal impairments. Several spatial, temporal collapsing and clipping functions are applied in the masked results because the HVS is usually more sensitive to the worst distortions. We presented a simplified NTIA model, which was based on quality features extracted from spatial gradients [82]. These features can express the coding artifacts of blockiness, blurring and added noise, whilst the influence caused by frame freezing is also taken into account. Furthermore, the computation speed can be increased evidently by adaptively adjusting the sizes of the S-T blocks according to spatial and temporal perceptual information.

There are some other similar metrics to extract the distortion features from the reference and distorted videos. Perceptual evaluation of video quality (PEVQ), standardized in ITU-T J.247 [18], computes four indicators for spatial distortion analysis: luminance indicator based on luminance difference of edge images between the reference and distorted videos, chrominance indicator by a similar approach as for the luminance, and two temporal variability indicators for omitted and introduced components. The temporal distortion is measured by a frame repeat indicator. Finally, the perceived video quality is estimated by mapping all indicators to a single number using a sigmoid approach. Also standardized in ITU-T J.247, the psytechnics model analyzes spatial frequency, edge distortion, blurring, and blockiness on resized input videos, and the overall quality prediction is a linear combination of the above analysis results, temporal distortion caused by dropped/frozen frames, and a spatial complexity analysis on the distorted video.

The human visual system cannot perceive any changes between adjacent pixels below a just noticeable distortion (JND) threshold due to underlying spatial/temporal sensitivity and masking properties. This characteristic can be widely used in video coding and quality assessment. Obviously, any unnoticeable signal difference does not need be coded and reflected in a distortion measure. Several methods for finding JND thresholds have been proposed operating in a sub-band (DCT or wavelet) domain or a pixel domain [71].

Several metrics either using PSNR as a component in their computation or integrating PSNR into HVS have been proposed, because of the simple computation and clear physical meanings of PSNR. In the method proposed by Yang et al. [72], a JND threshold is estimated, and a modified PSNR is computed by subtracting the JND threshold from the difference between the reference and distorted videos. NTT's FR model uses PSNR as one of the five features to compute an objective quality [18]. Lee et al. [76] proposed to use an optimization procedure to compute the weighted PSNR for three color components, and the experimental results showed that this method provides improved performance over the conventional PSNR. In addition, NTIA scales PNSR as presented in Eq. (1), results into a measure, which perform better than the original PSNR [49].

$$\mathrm{PSNR\_}M = \frac{1}{1+e^{0.1701(PSNR-25.6675)}}, \quad 10 \le \mathrm{PSNR} \le 55 \qquad (1)$$

Moreover, there are some metrics that take into account the image characteristics and combine them with HVS. Under an assumption that the HVS is highly adapted to extract structural information from field of vision, Wang et al. [65] proposed that a measure of structural information change can provide a good approximation to perceive image distortion. The structural similarity (SSIM) index is used in measuring the image distortion based on a comparison of luminance, contrast and structure between the original and distorted images. This SSIM metric was then extended to measure the video distortion by integrating chrominance and motion information [83]. Lee et al. [25] found that subjects tend to give low quality scores to a video sequence whose edges are noticeably degraded, even though the overall MSE is not large. So, they proposed to use PSNR in the edge areas instead of the whole images, which was proven to be more correlative with subjective evaluation. This method, combined with two other features, which reflect the blockiness and blurriness degradation, was then standardized in ITU-T J.247. Although these kinds of metrics do not take into account the characteristics of the HVS in detail, it is relatively simple to compute and shows an acceptable performance because image characteristic is modeled in a perceptual manner.

In addition, visual attention is another important attribute of the human visual and perceptual system, while it is ignored in most existing quality metrics. Most of the current metrics consider the distortion on all subregions or pixels equally. Actually, many physiological and psychological experiments have demonstrated that human attention is not allocated equally to all regions in field of vision, but focused on certain regions known as salient regions [84]. Some tentative works have been done on integrating the human attention analysis into quality assessment. Lu et al. [77] proposed significance map estimation for visual quality assessment, and evaluated its application in a JND model. Based on the saliency attention model in [84], Feng et al. [78] investigated some different weighting methods at the pixels in salient regions for MSE, MAD, and SSIM metrics. However, no appropriate metrics have been proposed, which can exploit the characteristics of the human attention adequately. In [79], we analyzed the capability of visual attention in visual quality assessment, and proposed an effective video quality metric by taking into consideration the human attention analysis.

We surveyed main approaches of objective video quality above, including the psychophysical and engineering methods. These methods cover a wide range of existing metrics, and some of them have been standardized by ITU. To evaluate their performance and complexity further, we implemented some of them: PDM [24], a method using wavelet transform [80], Gabor difference metric [73], Yonsei method [25], JND metric [72], SSIM model [83], a simplified NTIA model [82], and a visual attention based metric [79] proposed by the authors. The public NTIA general model [85] and the MSU software [86] were also tested. The conventional PSNR was taken as the benchmark.

In our experiments, a total of 392 video clips and the corresponding subjective quality results were employed, which included 320 VQEG FR-TV Phase I test clips, 60 temporal scalability test clips, and 12 mobile test clips. The temporal scalability test was a single stimulus assessment to compare the performance of different temporal scalability parameters in an H.264/AVC codec. Four different video contents were employed with two resolutions: VGA and QVGA. The mobile test was a double stimulus continuous quality scale assessment, which employed three different content types with QCIF resolution to test the performance of H.264/AVC codec at four bit rates: 24, 32, 40, and 48 kbps. Because different subjective experiments usually have different rating scales, different test conditions, and many other test variables that change from one laboratory to another, it is difficult to compare or combine the results of two or more subjective experiments directly. Pinson and Wolf, [87] proposed an objective method for combining multiple subjective data sets, which can map the multiple subjective data sets onto a single common scale using an iterated nested least squares algorithm (INLSA). In our experiments, we used INLSA to map all subjective results into the range [0,1]. Then, a non-linear regression suggested in [20] was fitted to the data set of the metric results VQ and the mapped MOS values, and restricted to be monotonic over the range of the MOS values. The following function fitted to the data [MOS_P, MOS] was used in the regression

$$\mathrm{MOS}_P = \frac{b1}{1+e^{-b2\times(\mathrm{VQ}-b3)}} \tag{2}$$

where $b1$, $b2$, and $b3$ denote the regression parameters. The non-linear regression was used to transform the set of metrics results VQ to a set of predicted MOS values, MOS_P, which were then compared with the mapped MOS values.

Four evaluation criteria were chosen to assess the prediction accuracy, monotonicity, and consistency of the metrics; hence RMSE, the Pearson linear correlation coefficient, the Spearman rank order correlation coefficient between MOS_P and MOS, and outlier ratio were used.

Table 5 gives the evaluation results of the tested metrics. According to the evaluation, the NTIA model and the visual attention based metric, as well as PDM, achieve the best performance, while PSNR is the worst. For the simplified NTIA model and the visual attention based metric proposed by the authors, we used half of the original undistorted video scenarios and their corresponding distorted clips in the training of the coefficients and thresholds. Then, the metrics' performance was evaluated with respect to the remaining sequences. Although the training clips were different from the test clips, the performance comparison between these two models and other metrics might be unfair, because we used data coming from the same set to train our models. Further analysis of the proposed models will be performed using other video data sets in our future work.

Statistically speaking, there are currently no objective metrics available that can replace subjective quality assessments. We suspect that a more accurate metric must take into account not only the internal characteristics of video clips and the HVS, but also external factors such as testing environment and a priori knowledge of subjects. In some special applications, e.g. when a coarse video quality estimation rather than an exact prediction is required, we believe that a simpler metric, e.g. based on PSNR combined with semantic video analysis, can be used.

Table 5
Evaluation results of objective video quality metrics.

Criteria	PDM	Wavelet	Gabor	Yonsei	JND	SSIM	MSU	NTIA	Simplified NTIA	Attention based metric	PSNR
RMSE	0.089	0.127	0.119	0.113	0.099	0.117	0.128	0.086	0.096	0.081	0.139
Pearson	0.874	0.786	0.823	0.786	0.745	0.775	0.746	0.886	0.843	0.891	0.683
Spearman	0.821	0.815	0.796	0.804	0.773	0.698	0.755	0.865	0.821	0.841	0.558
Outlier ratio	0.71	0.72	0.61	0.72	0.59	0.66	0.59	0.71	0.67	0.63	0.77

5. Subjective methodology for audio–visual quality assessment

As introduced in the preceding sections, ITU has come up with a number of normative recommendations of how to perform subjective assessments of perceived quality in general. These suggestions are internationally recognized and allow a comparison of results of assessments carried out in different laboratories. They define the test conditions as well as the form of presentation. Suggestions on rating scales to be used as well as on the classification of the test material are given. Several important recommendations will be discussed in the sequel.

Unfortunately, all these recommendations are mostly related to assessments in one single modality. Most recommendations either focus on audio or video quality alone, without taking into account the possible cross-modal effects. There are only very few recommendations that relate directly to quality assessments of audio–visual program material.

ITU-R BT.500-11 [88] provides details on methodologies for the evaluation of television picture quality. Most important in the audiovisual context, ITU-R BT.500-11 contains recommendations for viewing distances, illumination levels, screen sizes for different resolution and aspect ratio displays, etc. ITU-R BS.775-1 [89] makes suggestions on the reproduction setup for multichannel audio with accompanying picture. The relation between screen size, aspect ratio and loudspeaker positions is discussed. ITU-R BS.1286 [90] guides the testing of audio systems in the presence of an accompanying image. It should be applied in conjunction with recommendations ITU-R BS.1116-1, BS.1284, or BS.1285. Appendix 1 of ITU-R BS.1286 contains a list of viewing distances and conditions for different image sizes, aspect ratios, and image definitions.

ITU-R BT.1359 [91] provides recommendations for the relative timing of television sound and vision. These recommendations are related to inter-modal synchrony issues discussed in Section 2.C. Finally, Question ITU-R 102/6 [92] requests suggestions for further standardizing the methodologies used for the subjective assessment of audio and video quality, based on the fact that existing standards do not cover all aspects of audiovisual perception.

The telecommunication standardization sector of ITU has also come up with a number of recommendations related to audio–visual quality assessments. ITU-R P.910 [93] describes non-interactive subjective assessment methods that can be used to evaluate one-way overall video quality for multimedia applications such as videoconferencing. It also suggests characteristics of source sequences to be used, such as duration, content, number of sequences etc., ITU-R P.911 [1] is similar to P.910, but applies to audio–visual (instead of visual only) subjective assessment. Both are valid for non-interactive audio–visual program material. Aspects of interactive applications are not considered here.

Today's assessments rely on a combination of subjective and introspective methods, e.g. with questionnaires and rating scales. These methods imply that the test subject needs to be questioned during or immediately after the presentation of an item. Because it is necessary for the test subject to reflect upon the perceived impressions before rating an experience, he/she is induced to consume the content in a much more conscious way, and the effects to be assessed might not occur any longer. New approaches try to assess the degree of agreement of test subjects as a measure of perceived quality by using physical, physiological or behavioral investigations, but these are only starting to give relevant data [94]. It is therefore still necessary to ask the test subjects directly.

The most common test methods of subjective assessments can be categorized into single stimulus (also called absolute category rating, ACR) methods, pair comparison methods, and multi-stimulus methods. Alternatively, degradation category rating (DCR) is widely used. Subjective testing always requires the choice of a methodology and a rating scale amongst many possibilities, an issue that cannot be discussed in more detail here.

Audio–visual quality assessment can either give quantitative or qualitative data. Coolican [95] gives a definition of the terms *quantitative data* and *qualitative data*. Qualitative data are data "left in its original form of meaning (e.g. speech, text) and not quantified", whereas quantitative data are data "in numerical form, the results of measurement". Hence, qualitative assessments are assessments that allow test subjects to use their own words for describing the percepts, whereas quantitative assessments work with predetermined attributes that are quantified by subjects using a given rating scale.

Usually, quantitative assessment methods are preferred, because the resulting data are easier to analyze. Unfortunately, these methods often require an increased degree of preparation of the test subjects. This can either consist of a lengthy and repeated vocabulary development for all kinds of descriptive analysis processes (see e.g. [96]), or in a training that familiarizes test subjects with the semantic identifiers of the attributes to be rated. Still, quantitative methods are far more often used in assessments of audio and visual quality than qualitative methods. A successful mixed method approach, combining quantitative and qualitative data, has been described in [97,98].

6. Perceptual-based audio–visual quality metrics

In the preceding sections, we have investigated and analyzed a number of audio quality and video quality metrics. These methods assume only a single modality, either video or audio. Nevertheless, it has been shown by subjective tests (e.g. in [30]) that there is a strong mutual influence between audio and video on the experienced overall quality. At present, there is no reliable metric available for measuring the audio–visual quality automatically. In order to produce a reliable and accurate audio–visual metric, one must develop a firm understanding of the human perceptual processes first [29]. A number of methods have been proposed for performing user studies of the audio–visual quality, but there still remains a gap in defining the perceptual and cognitive basis of audio–visual quality assessment [98,99]. For an understanding of how subjects perceive audio–visual quality, it seems fundamental to research how auditory and visual stimuli are perceived, and at what stage in the human perceptual process they are fused to form a single overall quality experience.

6.1. Mutual influence between AQ, VQ, and AVQ

A series of subjective experiments have been conducted to study the mutual influence between AQ, VQ, and AVQ since the 1990s. As mentioned before, the qualities of five different presentations in Table 1 could be evaluated for analyzing the mutual influence between AQ, VQ, and AVQ. In the experiment conducted by Beerends and De Caluwe. [30], it was shown that AQ_V and VQ_A did not immediately improve the prediction of the audio–visual quality. Therefore, most experiments have been conducted for assessing AQ, VQ, and AVQ, and exploring their relationship. Such experiments try to derive AVQ according to AQ and VQ, but the influence of one modality on the other modality is not investigated. Consequently, most of the research has been concentrated on deriving a model to conclude AVQ from AQ and VQ. A commonly used fusion model [30–33] is

$$AVQ = a_0 + a_1 AQ + a_2 VQ + a_3 AQVQ \tag{3}$$

where the parameters $\{a_1, a_2, a_3\}$ denote different weights of audio and video quality, as well as the multiplication factor for the overall quality. The parameter a_0 is irrelevant to the correlation between the predictive quality and the perceived quality, but it improves the fit in terms of the residual between them. The overall audio-visual quality is influenced by a couple of factors, in which the individual AQ and VQ are the most important. Synchronicity, i.e. the offset between the audio and video stimuli, is another key element.

Hayashi et al. [100] proposed to take audio–video synchronization into account in the fusion model. Quality degradation (DQ) due to audio–visual delay for videophone services is derived, and the perceived multimedia quality is computed from AVQ and DQ using a similar fusion method as presented in Eq. (3).

The results of a quality assessment can also be affected by some other factors, such as the goal of the assessment, the testing environment, and the test methodology [101,102]. However, it is hard to model these external factors in a computable approach. Moreover, as stated earlier, we assumed perfectly lip-synchronized stimuli in this study. Hence, we concentrate on investigating the mutual influence of AQ and VQ on the AVQ for spatiotemporally coherent audio–visual systems in the following paragraphs.

Although the fusion method in Eq. (3) has been recognized by many researchers, there are no commonly agreed values or derivations for the four fusion parameters. Instead, the optimal values of the fusion parameters are different in various studies for different audio–visual content items. Table 6 summarizes the fusion parameters that we are able to find in the literature.

Early studies proposed AVQ models based on the multiplication of AQ and VQ [30,33], and the additive item, as in Eq. (4). These methods achieved promising results with most of the variation seen in the additive parameter a_0.

$$\mathrm{AVQ} = a_0 + a_3 \mathrm{AQVQ} \tag{4}$$

This type of fusion is in accordance with some studies on the human cognitive understanding, indicating that aural and visual information might be combined in an early phase of the human perception formation [105]. Nevertheless, fusion using single multiplication as in Eq. (4) does not reflect the differences in the influence of auditory-only and visual-only stimuli on the overall quality. Actually, a number of subjective assessments indicate that the influence of audio and video on the overall quality is unbalanced [30–34]. In these assessments, video quality dominates the overall quality in general. Audio seems to be more important than video under certain special circumstances only, such as in video conference situations, and in music TV.

In order to investigate the fusion parameters for the derivation of AVQ, we collected all subjective scores of AQ, VQ, and AVQ from the publications listed in Table 6. Because most literature on this issue reports the subjective quality scores in figures rather than score values explicitly, we employed a software named GetData Graph Digitizer [106] to extract the score values from the figures. After that, we performed the same analysis as in the original references, e.g. computation of correlation coefficients. Whenever we obtained the same as or very close results to the original references, i.e. the difference between our results and the results in the original references was below a very small threshold (e.g. 0.02). The extracted score values were used in our evaluation. In contrast, if the extracted scores could not exhibit attributes similar to those in the original references, they were excluded. In addition, the original experimental conditions, such as audio–visual content category and bit rates, were also gathered when the literature described them explicitly. The analysis of our investigation is presented in the following paragraphs.

Table 6
Overview of fusion parameters for audio–visual quality assessment.

Laboratory	a_0	a_1	a_2	a_3	Correlation	Methodology	Rating scale	Reference
KPN	1.12	0.007	0.24	0.088	0.98	ACR	[1,9]	[30]
	1.45	0	0	0.11	0.97			
Bellcore	1.07	0	0	0.111	0.99			[31]
	1.295	0	0	0.107	0.99			
ITS	−0.677	0.217	0.888	0	0.978		[1,5]	[32]
	1.514	0	0	0.121	0.927			
	0.517	−0.0058	0.654	0.042	0.98			
NTT	1.17	−0.144	0.186	0.154	0.96			[100]
	0.908	−0.192	0.258	0.193	0.96			
ICRFE	−0.9222	0.5691	0.5064	0.1697	0.911			[103]
	−0.6313	0.2144	0.0124	0.1184	0.902			
BT	1.15	0	0	0.17	0.85	SSQS		[34]
	0.95	0	0.25	0.15	0.83			
	4.26	0.59	0.49	0	0.97	DSCQS	[0,100]	
	−3.34	0.85	0.76	−0.01	0.99			
EPFL	1.98	0	0	0.103	0.9	ACR	[0,10]	[36]
	−1.51	0.456	0.77	0	0.94			
ICU	0	0.38	0.44	0.18	0.95	DCR		[104]
	0	0.43	0.32	0.26	0.95			
	0	0.35	0.58	0.07	0.95			

The test conditions, materials, and methodologies of these experiments differed from each other. Thus, the obtained fusion parameters were different in different experiments. However, according to the obtained correlation coefficients in Table 6, the fusion model in Eq. (3) works well in all experiments. Therefore, we conclude that it is feasible to combine all experimental results and then perform a correlation analysis. In addition, because the ranges of the rating scales were different, all subjective scores were rescaled into 9-point scales, i.e., the quality scale interval was set to [1,9]. The following function was used to rescale the subjective scores.

$$S = 8\frac{S_0 - S_{\text{best}}}{S_{\text{best}} - S_{\text{worst}}} + 9 \tag{5}$$

where S and S_0 denote the rescaled score and original score, respectively. S_{best} is the limit corresponding to the best quality of the original scale, and S_{worst} denotes the limit corresponding to the worst quality.

After rescaling the quality values, the Pearson linear correlation coefficient (R) and RMSE were used for the correlation analysis. A multivariable regression algorithm was applied to find the best fusion parameters in Eq. (3), in which the correlation was maximized whilst the RMSE was minimized. Moreover, to make the analysis results easier to compare, we limited the weights of the fusion parameters $\{a_1, a_2, a_3\}$ to the interval [0,1], whereas the additive parameter a_0 was allowed to be out of this range. It should be noted that such limitations might introduce a bias for the correlation analysis. Therefore, we have also performed an analysis without such constraints. The experimental results obtained were very similar to the results in this paper. Thus, we report the results with such constraints on the fusion parameters, because then we can clearly see the difference between different classes.

An analysis on all test clips was performed. The correlation R and RMSE were 0.55 and 1.85 between AQ and AVQ, respectively, and 0.83 and 1.2 between VQ and AVQ, respectively. According to the correlation analysis, both AQ and VQ have a significant effect on the overall quality. A clear dominance of the video quality is observed in general. Furthermore, the correlation between the multiplication AQ × VQ and AVQ is 0.93, which indicates that the multiplication has the most significant influence on AVQ. This is in accordance with the experimental results in [30,33]. Even after the quality fusion in Eq. (3) with optimal parameters, the correlation increased only marginally, from 0.93 to 0.95. The results therefore indicate that for a general audio–visual quality metric, the multiplication of AQ and VQ is adequate for predicting the tendency of the overall quality, whereas the weight a_3 and the additive constant a_0 can improve the prediction accuracy.

In order to study the mutual influence under different conditions of compression settings and degradation strengths, we analyzed the correlation and derived the optimal fusion parameters for groups of test cases. Two different classification criteria were used: bitrate and subjective quality level. For dividing the test clips into different groups according to the bitrates, two empirical thresholds, 24 and 64 kbps, were used for the audio and video bitrates, respectively. These empirical thresholds stem from an informal subjective test in which five subjects were asked to give a threshold of bit rate at which they could clearly perceive the difference when the audio or video was encoded below and above the bit rate. For the grouping according to quality levels, we divided the quality scale interval [1,9] into three equal parts. The analysis results are presented in Tables 7 and 8 for bitrate classes and quality level classes, respectively.

Based on our analysis on the mutual influence between AQ, VQ, and AVQ, some general conclusions can be drawn as follows:

- Both audio quality and video quality contribute to the overall audio–visual quality, and their multiplication has the highest correlation with the overall quality. So,

Table 7
Correlation analysis for AQ, VQ, and AVQ for different bitrate classes.

Audio bitrate	Video bitrate	Number of clips	R/RMSE (AQ, AVQ)	R/RMSE (VQ, AVQ)	R(AQ × VQ, AVQ)	Best fusion parameters $\{a_0, a_1, a_2, a_3\}$	R/RMSE (AVQ_P, AVQ)
Low ≤ 24 kbps	All	44	0.22/1.60	0.82/1.10	0.90	−0.03, 0.14, 0.45, 0.07	0.94/0.42
	≤ 64 kbps	*24*	*0.67/0.98*	*0.47/1.07*	*0.92*	*−0.2, 0.2, 0.43, 0.07*	*0.94/0.3*
	> 64 kbps	20	−0.23/2.1	0.93/1.1	0.87	1.1, 0, 0.38, 0.05	0.95/0.4
High > 24 kbps	All	100	0.42/1.98	0.88/1.03	0.91	0.84, 0.09, 0.43, 0.05	0.94/0.6
	≤ 64 kbps	24	−0.27/1.8	0.91/0.59	0.81	3.62, 0, 0.35, 0	0.91/0.49
	> 64 kbps	76	0.43/2.03	0.88/1.13	0.91	0.8, 0.1, 0.43, 0.05	0.94/0.66
All	≤ 64 kbps	48	0.67/1.45	0.55/0.86	0.90	−0.23, 0.21, 0.52, 0.05	0.93/0.35
	> 64 kbps	96	0.43/2.05	0.88/1.13	0.92	0.94, 0.05, 0.39, 0.06	0.95/0.63

(R and RMSE denote the Pearson correlation coefficient and the root-mean-square error, respectively, between the indicated quality measures. AQ, VQ, AVQ denote the subjective measure of the audio, video, and audio–visual qualities, respectively, and AVQ_P is the quality predicted from AQ and VQ with the fusion method in which the best fusion parameters are derived by the multivariable regression analysis. The use of *italics* denotes that AQ dominates AVQ.)

Table 8
Correlation analysis for AQ, VQ, and AVQ for different quality level classes.

AQ level	VQ level	Number of clips	R/RMSE (AQ, AVQ)	R/RMSE (VQ, AVQ)	R(AQ × VQ, AVQ)	Best fusion parameters $\{a_0, a_1, a_2, a_3\}$	R/RMSE (AVQ_P, AVQ)
Low [1∼3.3]	All	44	0.33/1.27	0.86/1.91	0.88	−0.02, 0.49, 0.42, 0	0.89/0.48
	Low	14	0.35/0.86	0.9/0.36	0.74	0.68, 0.07, 0.58, 0	0.91/0.28
	Middle	21	0.31/1.15	0.6/1.42	0.57	0.82, 0.13, −2.15, 0	0.67/0.63
	High	*9*	*0.84/1.9*	*−0.03/3.6*	*0.84*	*−0.13, 0.99, 0.25, 0*	*0.87/0.37*
Mid. [3.3∼6.6]	All	99	0.24/1.74	0.92/0.94	0.94	1.04, 0.05, 0.31, 0.08	0.94/0.56
	Low	32	0.13/2.44	0.82/0.62	0.80	2.15, 0.03, 0.18, 0	0.84/0.61
	Middle	40	0.52/0.92	0.77/0.61	0.82	0.25, 0, 0.39, 0.11	0.84/0.53
	High	*27*	*0.63/1.67*	*0.02/1.48*	*0.57*	*1.16, 0.85, 0, 0.03*	*0.64/0.66*
High [6.6∼9]	All	65	0.31/2.29	0.94/0.93	0.95	−1.96, 0.43, 0.87, 0	0.95/0.56
	Low	13	0.20/3.87	0.74/1.33	0.78	2.42, 0, 0, 0.07	0.78/0.47
	Middle	33	0.47/2.01	0.61/0.89	0.73	−0.61, 0.4, 0.57, 0.01	0.74/0.54
	High	19	0.49/0.81	0.74/0.63	0.82	4.99, 0.15, 0.23, 0	0.85/0.47
All	Low	59	0.6/2.59	0.7/0.8	0.86	−0.13, 0.23, 0.71, 0.01	0.89/0.4
	Middle	94	0.73/1.44	0.5/0.94	0.85	0.62, 0.17, 0.42, 0.05	0.87/0.53
	High	*55*	*0.9/1.48*	*0.11/1.83*	*0.90*	*−1.32, 0.67, 0, 0.09*	*0.91/1.65*

when constructing a general audio–visual quality metric, the overall quality can be predicted by the weighted multiplication and an additive shift. The weight and addition are irrelevant to the correlation, but they can improve the prediction accuracy.

- Generally, video quality dominates the overall quality, whereas audio quality is more important for cases in which the bit rates of coded audio and video are both low, or the video quality is above a certain quality threshold. With decrease in audio quality, its influence on the overall quality is increasing. In addition, for certain audio–visual contents or applications in which audio is evidently more important than video, such as teleconference, news, and eventually music video, audio quality dominates the overall quality. In such applications, a greater weight should be assigned to audio quality when constructing an audio–visual metric. In other cases, video quality should be given a greater weight. Then, the audio quality component in the fusion equation can even be discarded, i.e., the VQ term and the multiplication AQ × VQ are adequate for predicting the overall quality.

In addition, there are three different papers that consider the quality change in one modality in the presence of the other modality [30,31,102]. Thirty six test cases were studied in total, and the correlation results for the quality of one modality in the presence of the other are: R(AQ_V, AQ) = 0.991, R(AQ_V, VQ) = 0.365, and R(VQ_A, AQ) = 0.374, R(VQ_A, VQ) = 0.992. Therefore, when measuring individual audio or video quality in audio–visual stimuli, the influence of the other modality might be small, but cannot be neglected totally. The magnitude of the impact of the presence of the other modality seems to be similar for both audio and video, i.e., R(AQ_V, VQ) ≈ R(VQ_A, AQ).

Finally, in our analysis of previous experiments, we also found that the mutual influence between AQ and VQ may depend on the video characteristics. The higher the motion and the more complex the picture content is, the more dominance of the video quality is observed. So, motion information and picture complexity should be taken into account in developing an audio–visual quality metric.

In summary, we may classify the influence factors of audio–visual quality into three levels. Influence factors of level 1 are those related to the generation of audio–visual stimuli, such as the loudspeaker and visual reproduction setup, as well as the content itself. Influence factors of level 2 affect sensory perception, which involves the physiology of the user (acuity of hearing and vision, masking effect caused by limited resolution of the human sensors, etc.) as well as all other factors directly related to the physical perception of stimuli. Influence factors of level 3 are related to the processing and interpretation of the perceived stimuli, which span the widest range of factors: experience, expectations, and background of subjects; task; degree of interactivity (if any); type of application; etc. [29,107].

6.2. Relevant issues and trends in audio–visual quality metrics

The first issue in developing an audio–visual quality metric is the interactive process of multiple modalities. In order to understand the multi-modal perceptual fusion, it is worth-looking at modern theories of attention. These can be basically classified into two approaches: early selection and late selection [105]. In the early selection theory, all stimuli that reach the sensory system are processed until individual physical attributes are explicitly represented. In the study of attention, this is called “Broadbent’s Filter theory” [108]. If we transfer this concept to the realm of quality perception, we may deduct that an overall perceived quality impression must then be a function of the individual attributes. A fusion or binding of individual quality attributes takes place at the end of the processing chain hence this theory is called late fusion. Some other experiments support that the auditory and visual stimuli can be integrated at the quality level. Opposed to this, late selection theory in attention argues that the recognition of familiar objects proceeds unselectively, as one cannot voluntarily choose to (or refuse to) recognize or identify something. All available input is processed to result in the representation of a perceptual object generated from a fused set of attributes. As a consequence, an overall quality impression must then be the result of analyzing the perceptual object as a whole, i.e. audio–visual fusion takes place at an earlier stage than quality level, which is called early fusion. Many experiments support this conclusion.

However, in the latter case, i.e. early fusion, the construction of audio–visual quality models seems to be very complicated. It means that we need to consider not only the quality distortion between distorted and reference signals of an individual stimulus, but also the influence between different modalities. This suggests that an objective audio–visual quality metric should be represented as a joint function, Q (audio, video), rather than a separate function [Q (audio), Q (video)]. Hence, an appropriate cognition model should be constructed so that it can process audio and video signals simultaneously, and take into account the cross-modal influences. However, such a cognitive model may be very complex and cannot be used in real-time applications.

Here, we assume that the human brain perceives the auditory and visual stimuli independently at a certain point in time or during a very short duration. This means that fusion is assumed to occur at the quality level at this point in time or during this duration, if we cite the concept of limit in mathematics. In this approach, we do not need to consider the influence of a modality on the individual quality of the other modality. Currently, most studies on audio–visual quality assessment adopt the late fusion theory, presumably because it is easier to handle. Whether or not this approach is correct is a topic of ongoing research.

Therefore, applying the late fusion theory, we can construct audio–visual quality models in an engineering approach in two steps: find an appropriate fusion model or appropriate fusion parameters using Eq. (3) to compute the overall audio–visual quality based on audio quality and video quality in a short temporal segment; find an appropriate time averaging method to pool the quality values over all segments into an overall quality of a whole audio–visual sequence. The first step can be figured out according to the analysis on the mutual influence. At present, there is no definite solution for solving the second step. For example, PEAQ uses linear, squared and windowed averages for the temporal averaging. Moreover, most existing video metrics are proposed for measuring short sequences with a single scene, in which methods such as Minkowski summation or directive averaging are used. These methods, however, may be inappropriate for long-term sequences with multiple complex scenes. Therefore, temporal averaging is still an open issue, and it will lead to the second issue relevant for the development of audio–visual metrics – semantic analysis of content.

Semantic analysis for audio–visual content has been studied for several decades. It is mainly applied in audio–visual retrieval, management, summarization, etc. [109,110]. Actually, it is suspected that semantic audio–visual importance is strongly related to perceived quality. Yet, this topic is relatively unexplored. Most audio and video quality metrics are based on the perceptual perspective, for example, the existing FR metrics usually predict the difference between the reference and distorted signals from the viewpoint of the human perception. However, even when perceptual quality is reduced (e.g. lower frame rate), subjects may understand the content and the perceived quality is thought to be unchanged [111]. There has been some work focused partially on this topic. Cucchiara et al. [112] proposed to encode video objects with different perceptual fidelities based on the importance of objects, and the overall quality was evaluated using an object-weighted MSE measure. Thang et al. thought that the overall quality consists of perceptual quality (PQ) and semantic quality (SQ), and the overall quality is regarded as a weighted linear combination of PQ and SQ. They applied graph theory in modeling the relationship between the overall quality, PQ and SQ in [113,114]. The semantic quality mainly models the perceived amount of information conveyed by audio–visual signals, regardless of how the signals are presented, whereas the perceptual quality is defined as the satisfaction of a subject perceiving the signals, regardless of what information is conveyed. Further, Hanjalic and Xu, [115] thought that video content should be modeled at two different levels: cognitive and affective levels, in which the former is to model how a subject perceives video content, and the latter is to define the affective characteristics of video content. We believe that different affections caused by audio–visual content may lead to different impacts on the perceived quality for different subjects. For example, an optimistic subject may like more a comedy segment, such that he/she will give a more positive quality score on the comedy segment than on tragedy segments. Of course, an optimistic subject may also give a more negative quality score for the comedy segment because he/she has stricter demand for this kind of segments. In addition, semantic importance analysis can also be applied in the temporal averaging over all time segments, as mentioned in the previous paragraph. A time segment with more important semantic content should be assigned a bigger weight when performing time averaging. Synthetically, we can assume that an overall quality consists of a combination of perceived quality and semantic or affective importance, in which the former is evaluated by audio–visual quality metrics, whereas the latter can be derived from content analysis. Based on the analysis of the mutual influence between audio and video quality, of the relationship between perceived quality and semantic information, as well as the time averaging, we suggest that a perceptual quality model can be constructed in the following paradigmatic form:

$$\mathrm{OQ} = \sum_i W_i S_i (a_0 + a_1 \mathrm{AQ} + a_2 \mathrm{VQ} + a_3 \mathrm{AQVQ})_i \tag{6}$$

where OQ denotes the overall audio–visual quality of an audio–visual sequence; i denotes different segments whose duration might be different from each other because of different audio–visual contents; W_i represents a weight of this segment, which is affected by some external factors and quality level, e.g. different quality levels make different contributions to the overall quality [81,82]; S_i denotes the semantic importance of a segment derived from a content analysis model; and Σ represents time averaging, which might be a direct average. It is noticed that the fusion parameters $\{a_0, a_1, a_2, a_3\}$ might be different in different segments. Fig. 1 illustrates an audio–visual quality model, taking into account the semantic information and other influence factors. The grayed-out blocks denote a potential quality drop caused by unsynchronized audio and video signals.

A reasonable simplification of Eq. (6), which can be applied in practical systems, is a combination of semantic analysis and a simple quality metric such as PSNR. Although PSNR is not always correlating well with results obtained from subjective quality evaluations, we believe that a suitable integration of semantic analysis into PSNR could improve the predictive accuracy of this type of simple metrics. For example, we have found that the accuracy of predicting the video quality can be increased by integrating the information of spatial and temporal activities into PSNR [116]. Furthermore, as an important neurophysiologic concept, human attention is an indispensable factor in constructing the semantic analysis model as well as the quality metric [79,84]. In addition, we have mentioned above that AQ and VQ have different influence on the overall AVQ in certain situations, in which audio and video have different importance levels, for example in music video and teleconference sequences. Therefore, constructing a semantic analysis model for comparing the importance between the audio and video is also important. This topic on the relationship analysis between semantic audio–visual analysis and quality assessment will be studied further in our future work.

For the temporal averaging, another issue should be taken into account, i.e. different contributions of temporal segments with different quality levels to the overall quality. The human perceptual system is usually more sensitive to large impairments. Our experiment in [82] demonstrated that bigger weights should be assigned to the segments suffering from severe impairment. We have also investigated different spatial and temporal averaging methods by using some image quality metrics to predict the quality of packet loss video streams [117]. Our experimental results indicated that the human perception on video quality is mainly influenced by those regions and frames with the most severe distortions. The NTIA model [49,81] uses spatial and temporal collapsing functions to express such characteristic. However, this issue is not studied adequately, and more

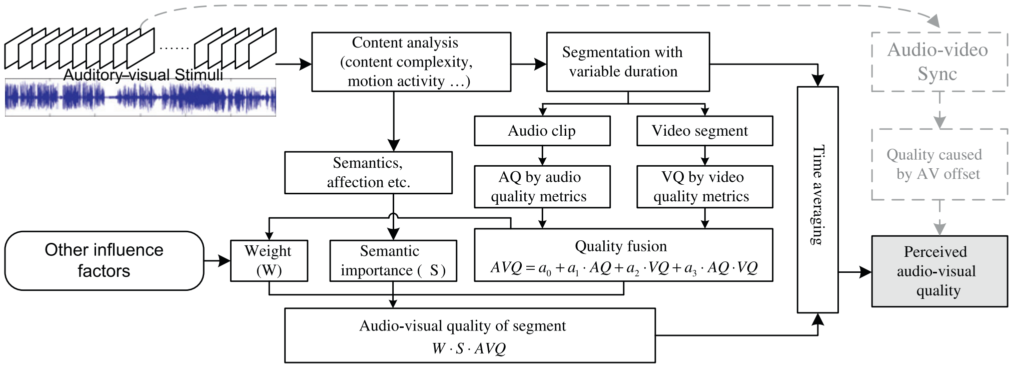
Auditory-visual Stimuli
Content analysis (content complexity, motion activity …)
Segmentation with variable duration
Audio-video Sync
Audio clip
Video segment
Semantics, affection etc.
AQ by audio quality metrics
VQ by video quality metrics
Time averaging
Quality caused by AV offset
Other influence factors
Weight (W)
Semantic importance (S)
Quality fusion
$AVQ = a_0 + a_1 \cdot AQ + a_2 \cdot VQ + a_3 \cdot AQ \cdot VQ$
Perceived audio-visual quality
Audio-visual quality of segment
$W \cdot S \cdot AVQ$

Fig. 1. Flowchart of an audio–visual quality metric taking into account the semantic analysis and other influence factors in quality assessment.

subjective measurements analyzing the influence of different quality levels on the overall quality perception are required, as well as a thorough investigation of the corresponding psychophysical knowledge.

No uniform fusion parameters for Eq. (3) have been found so far, and the fusion parameters are strongly correlating with audio–visual content and other factors. As mentioned before, semantic content analysis might be helpful to find appropriate weights of audio and video, in evaluating the overall audio–visual quality. For example, the concept of entropy can express the information conveyed in audio or video signals. Thus, we may assume that the fusion model or its parameters can be obtained automatically, based on an analysis of audio–visual content. In this case, the regression operation between the model results and the subjective quality results will not be required to find the fusion parameters, since it is supposed that subjective quality results are not available in developing objective quality metrics. This topic will also be studied in our future work.

For the audio-only and video-only quality metrics, there are some issues that have not been studied adequately. At the same bit rates, one video sequence with a larger size is usually allocated fewer bits for every frame than a sequence with a smaller size, such that the computed quality of the former is lower than the latter. However, subjects usually prefer a large image size because it can potentially provide more details. Our earlier subjective experiment [118] demonstrated that the perceived quality of a video sequence with CIF resolution coded at 384 kbps and 12.5 fps is rated higher than the same video with QCIF resolution coded at 384 kbps and 25 fps. However, the quality as computed by the NTIA model is lower for the former than for the latter. A similar phenomenon is found in audio quality measurements. VQEG has conducted the validation of quality metrics for multimedia assessment with three different resolutions, including VGA, CIF, QCIF, whereas the performance was analyzed for video sequences with the same resolution [18,21]. So, determining how to integrate the factors such as image size, frame rate, number of audio channels, sample rate, etc. into an overall quality metric is another important but unresolved issue.

Last but not the least, the contribution of quality metrics on the performance improvement of audio–visual services should be studied further. Although a quality metric would originally be proposed to evaluate the quality of the distorted signals, it may also be applied for improving coding schemes, transmission abilities, etc. For example, Yang et al. [72] proposed to use a JND-adaptive motion estimation scheme and residue filtering for rate control. All of these can improve the overall performance of a video coding scheme. In addition, a quality metric has wide applications in mode decision, as well as in error concealment in audio and video encoding and decoding. Although several methods have been proposed in this area, more work needs to be performed in the future.

7. Conclusions

We surveyed perceptual-based audio and video quality assessment methods in this paper. The main existing methods were introduced and analyzed, and the experimental results with respect to subjective assessment were presented. Alignment of audio and video signals and audio–video synchronization were reviewed. For audio quality metrics, we mainly concentrated on PEAQ and some improvement methods as well as two simple metrics. The basic version of PEAQ was implemented and validated by the conformance test using standard audio sequences, and the performance of PEAQ in different sample rates was also tested. Furthermore, we tested the performance of some metrics on parametric audio coding schemes, while no reliable metrics for parametric coding have been found. For objective video quality metrics, we investigated FR and RR models in psychophysical and engineering approaches. Experiments with respect to the available subjective assessment results were performed to analyze and evaluate the performance of the existing FR and RR models and our proposed metrics. The experimental results demonstrated that the current objective quality metrics still cannot replace subjective quality assessment, although their performance is comparatively promising. Possible improvements for future development were also discussed. Subjective audio–visual quality methodologies and ITU recommendations were also investigated. Finally, the mutual influence among audio quality, video quality, and overall audio–visual quality was studied, and some general conclusions were drawn. The relevant issues in developing perceptual-based audio-visual metrics were investigated and some trends were presented. Although this paper tried to provide a comprehensive survey of the perceptual-based audio–visual quality metrics, some issues have not been included, such as quality degradation caused by packet losses in transmission. These topics will be studied in future work.

The quality metrics for individual audio and video modalities have been studied for several years, while the audio–visual quality is a relatively unexplored issue. Furthermore, the studies on quality assessment of multimodality, and multiparty for wideband applications have just been started [119]. These issues are closely correlated with the psychophysical and psychoacoustic knowledge, and there still remains a big gap between these two scientific disciplines and quality assessment. For example, it is still unknown whether early fusion or late fusion between audio and video dominate the human perception in audio–visual quality assessment. As we have highlighted in this survey paper, many interesting quality measurement approaches and improvements have been proposed, and several standards have been made or are in the making. Nonetheless, we are still a long way from audio–visual quality metrics that are widely applicable and universally recognized.

Acknowledgement

The authors would like to thank three anonymous reviewers and the editor for their valuable and constructive comments on this paper.

References

[1] Recommendation ITU P.911, Subjective audiovisual quality assessment methods for multimedia application, ITU Telecommunication Standardization Sector, December 1998.

[2] ITU Recommendation J.148, Requirements for an objective perceptual multimedia quality model, ITU Telecommunication Standardization Sector, May 2003.

[3] K. Brandenburg, Evaluation of quality for audio encoding at low bit rates, in: Proceedings of the Contribution to the 82nd AES Convention, preprint 2433, London, United Kingdom, 1987.

[4] J.G. Beerends, J.A.J.A. Stemerdink, A perceptual audio quality measure based on a psychoacoustics sound representation, J. Audio Eng. Soc. 40 (1992) 963–978 December.

[5] B. Paillard, P. Mabilleau, J. Soumagne, Perceval: perceptual evaluation of the quality of audio signals, J. Audio Eng. Soc. 40 (1992) 21–32.

[6] C. Colomes, M. Rault, A perceptual model applied to audio bit-rate reduction, J. Audio Eng. Soc. 43 (1995) 233–240 April.

[7] T. Sporer, Objective audio signal evaluation – applied psychoacoustics for modeling the perceived quality of digital audio, in: Proceedings of the 103rd AES-Convention, preprint 4512, New York, United States of America, October 1997.

[8] T. Thiede, E. Kabit, A new perceptual quality measure for bit rate reduced audio, in: Proceedings of the Contribution to the 100th AES Convention, preprint 4280, Copenhagen, Denmark, 1996.

[9] ITU-R Recommendation 1387-1, Method for objective measurement of perceived audio quality, ITU Telecommunication Standardization Sector, 1998–2001.

[10] C.D. Creusere, K.D. Kallakuri, R. Vanam, An objective metric for human subjective audio quality optimized for a wide range of audio fidelities, IEEE Trans. Audio Speech Lang. Process. 16 (1) (2008) 129–136 January.

[11] J. Barbedo, A. Lopes, A new cognitive model for objective assessment of audio quality, J. Audio Eng. Soc. 53 (1/2) (2005) 22–31.

[12] S. Zielinski, F. Rumsey, R. Kassier, S. Bech, Development and initial validation of a multichannel audio quality expert system, J. Audio Eng. Soc. 53 (1/2) (2005) 4–21.

[13] R. Vanam, C.D. Creusere, Scalable perceptual metric for evaluating audio quality, in: Proceedings of the Conference Record of the 39th Asilomar Conference Signals, Systems and Computers, pp. 319–323, 2005.

[14] C.D. Creuser,e, R. Vanam, Understanding perceptual distortion in MPEG scalable audio coding, IEEE Trans. Speech Audio Process. 13 (13) (2005) 422–431 May.

[15] B. Girod, What's wrong with mean-square error, in: A.B. Watson (Ed.), Digital Images and Human Vision, MIT Press, Cambridge, MA, 1993, pp. 207–220.

[16] Video Quality Experts Group, 〈http://www.vqeg.org〉.

[17] ITU-T Recommendation J.144, Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference, ITU Telecommunication Standardization Sector, March 2001.

[18] ITU-T Recommendation J.247, Objective perceptual multimedia video quality measurement in the presence of a full reference, ITU Telecommunication Standardization Sector, August 2008.

[19] VQEG, Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment, VQEG, March . 2000.

[20] VQEG, Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment, phase II (FR-TV 2), VQEG, August 2003.

[21] VQEG, Final report from the Video Quality Experts Group on the validation of objective models of multimedia quality assessment, phase I, VQEG, September 2008.

[22] S. Winkler, Video quality and beyond, *Proc. Eur. Signal Process. Conf.* (September 2007) 150–154.

[23] S. Winkler, Perceptual video quality metrics – A review, in: H.R. Wu, K.R. Rao (Eds.), Digital Video Image Quality and Perceptual Coding, CRC Press, 2006, pp. 155–180 Chapter 5.

[24] S. Winkler, in: Digital Video Quality: Vision Models and Metrics, John Wiley & Sons, 2005.

[25] C. Lee, S. Cho, J. Chow, J. Choe, T. Jeong, W. Ahn, E. Lee, Objective video quality assessment, Opt. Eng. 45 (1) (2006) 017004-1–017004-11 January.

[26] S. Liu, A.C. Bovik, Efficient DCT-domain blind measurement and reduction of blocking artifacts, IEEE Trans. Circuits Syst. Video Technol. 12 (12) (2002) 1139–1149 December.

[27] P. Marziliano, F. Dufaux, S. Winker, T. Ebrahimi, A no-reference perceptual blur metric, Proc. IEEE Int. Conf. Image Process. 3 (September 2002) 57–60.

[28] M. Ries, O. Nemethova, M. Rupp, Reference-free video quality metric for mobile streaming applications, in: Proceedings of the 8th International Symposium on DSP and Communication Systems, December 2005, pp. 98–103.

[29] M.P. Hollier, A.N. Rimell, D.S. Hand, R.M. Voelcker, Multi-modal perception, J. BT Technol. 17 (1999) 35–46 January.

[30] 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) (1999) 355–362 May.

[31] N. Kitawaki, Y. Arayama, T. Yamada, Multimedia opinion model based on media interaction of audio–visual communications, in: Proceedings of the 4th International Conference on Measurement of Speech and Audio Quality in Networks (MESAQIN '05), Prague, Czech Republic, June 2005, pp. 5–10.

[32] C. Jones, and D.J. Atkinson, Development of opinion-based audiovisual quality models for desktop video-teleconferencing, in: Proceedings of the 6th International Workshop on Quality of Service (IWQoS '98), Napa Valley, CA, 18–20 May 1998.

[33] ANSI-Accredited Committee T1 Contribution, T1A1.5/94-124, Combined A/V model with multiple audio and video impairments, Bellcore, USA, April 1995.

[34] D.H. Hands, A basic multimedia quality model, IEEE Trans. Multimedia 6 (6) (2004) 806–816 December.

[35] S. Winkler, and C. Faller, Maximizing audiovisual quality at low bitrates, in: Proceedings of the Workshop on Video Processing and Quality Metrics for Consumer Electronics, Scottsdale, AZ, January 2005.

[36] S. Winkler, C. Faller, Perceived audiovisual quality of low-bitrate multimedia content, IEEE Trans. Multimedia 8 (5) (2006) 973–980 October.

[37] M.R. Frater, J.F. Arnold, A. Vahedian, Impact of audio on subjective assessment of video quality in videoconference application, IEEE Trans. Circuits Syst.Video Technol. 11 (9) (2001) September.

[38] ITU-T Contribution COM 12-61-E, Study of the influence of experimental context on the relationship between audio, video and audiovisual subjective qualities, France Telecom/CNET, September 1998.

[39] R. van Eijk, A. Kohlrausch, J.F. Juola, S. van de Par, Audiovisual synchrony and temporal order judgments: effects of experimental method and stimulus type, Percept. Psychophys. 70 (6) (2008) 955–968.

[40] N.F. Dixon, L. Spitz, The diction of auditory visual desynchrony, Perception 9 (1980) 719–721.

[41] D.W. Massaro, M.M. Cohen, P.M.T. Smeele, Perception of asynchronous and conflicting visual and auditory speech, J. Acoust. Soc. Am. 1000 (1980) 1777–1786 September.

[42] ITU-R SG11 11A/55, Evaluation of the subjective effects of timing errors between sound and vision signals in television, November 1995.

[43] S. Rihs, The influence of audio on perceived picture quality and subjective audio–video delay tolerance in RACE MOSAIC deliverable no: R211180CESR007.B1, chapter 13, June 1995.

[44] G. Blakowski, R. Steinmetz, A media synchronization survey: Reference model, specification, and case studies, IEEE J. Selected Areas Commun. 14 (1) (1996) 5–35 January.

[45] R. Steinmetz, Human perception of jitter and media synchronization, IEEE J. Selected Areas Commun. 14 (1) (1996) 61–72 January.

[46] J. You, G. Liu, L. Sun, H. Li, A multiple visual models based perceptive analysis framework for multilevel video summarization, IEEE Trans. Circuits Systems Video Technol. 17 (3) (2007) 273–285 March.

[47] Y. Gao, Audio coding standard overview: MPEG4-AAC, HE-AAC, and HE-AAC v2, in: F.-L. Luo (Ed.), Mobile Multimedia Broadcasting Standards: Technology and Practice, Springer Press, 2009, pp. 607–627 Chapter 21.

[48] Sonic Visualiser Software, [online] available: 〈http://www.sonicvisualiser.org/〉, MATCH Vamp plugin, [online] available: 〈http://www.vamp-plugins.org/〉.

[49] S. Wolf, and M. Pinson, Video quality measurement techniques, NTIA Report 02-392, June 2002.

[50] ATIS Technical Report T1.TR.73, 2001, Video normalization methods applicable to objective video quality metrics utilizing a full reference technique, Alliance for Telecommunications Industry Solutions, 1200 G Street, NW, Suite 500, Washington DC, October 2001.

[51] ITU-T Recommendation P.931, Multimedia communications delay, synchronization and frame rate measurement, ITU Telecommunication Standardization Sector, December 1998.

[52] M.P. Hollier, A.N. Rimell, An experimental investigation into multimodal synchronization sensitivity for perceptual model development, in: Proceedings of the AES 105th Convention, September 1998, San Francisco, California USA, Preprint 4790.

[53] Recommendation ITU-T J.100, Tolerances for transmission time differences between vision and sound components of a television signal, International Telecommunication Union, Geneva, Switzerland, 1990.

[54] S.D. Lipscomb, Cross-modal integration: synchronization of auditory and visual components in simple and complex media, Proc. Forum Acusticum (1999) Berlin, Germany.

[55] MP3-tech A.A.C. public test, [online] available: 〈http://www.mp3-tech.org/ content/?48kbps%20AAC%20public%20test〉.

[56] T. Thiede, W.C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J.G. Beerends, C. Colomes, PEAQ – the ITU standard for objective measurement of perceived audio quality, J. Audio Eng. Soc. 48 (2000) 3–29 February.

[57] P. Kabal, An examination and interpretation of ITU-R BS.1387: perceptual evaluation of audio quality, TSP Lab Technical Report, Department of Electrical & Computer Engineering, McGill University, December 2003.

[58] Recommendation ITU BS.1116-1, Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems, ITU Telecommunication Standardization Sector, 1994–1997.

[59] R. Vanam, and C.D. Creusere, Evaluating low bitrate scalable audio quality using advanced version of PEAQ and energy equalization approach, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2005, vol. 3, Philadelphia, PA, March 2005, pp.189–192.

[60] R. Huber, B. Kollmeier, PEMO-Q: a new method for objective audio quality assessment using a model of auditory perception, IEEE Trans. Audio. Speech. Lang. Process. 14 (6) (2006) 1902–1911 November.

[61] S. Voran, Objective estimation of perceived speech quality – Part I: development of the measuring normalizing block technique, IEEE Trans. Speech Audio Process. 7 (4) (1999) 371–382 July.

[62] S. Kandadai, J. Hardin, and C.D. Creusere, Audio quality assessment using the mean structural similarity measure, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2008, Las Vegas, USA, March–April 2008, pp.221–224.

[63] ITU Recommendation BS.1534-1, Method for the subjective assessment of intermediate quality level of coding systems, ITU Telecommunication Standardization Sector, 2001–2003.

[64] M. Hansen, B. Kollmeier, Objective modeling of speech quality with a psychoacoustically validated auditory model, J. Audio Eng. Soc. 48 (2000) 395–409.

[65] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612 April.

[66] PEMO-Q Software, [online] available: 〈http://www.hoertech.de/cgi-bin/ wPermission.cgi?file=/web_en/produkte/tools.shtml〉.

[67] J.L. Mannos, D.J. Sakrison, The effects of a visual fidelity criterion on the encoding of images, IEEE Trans. Inf. Theory 20 (4) (1974) 525–536 July.

[68] F.X.J. Lukas, Z.L. Budrikis, Picture quality prediction based on a visual model, IEEE Trans. Commun. 30 (7) (1982) 1679–1692 July.

[69] S. Daly, The visible difference predictor: An algorithm for the assessment of image fidelity, in: A.B. Watson (Ed.), Digital Images and Human Vision, MIT Press, 1993, pp. 179–206.

[70] J. Lubin, and D. Fibush, Sarnoff JND vision model, T1A1.5 Working Group Document #97-612, ANSI T1 Standards Committee, 1997.

[71] W. Lin, Computational models for just-noticeable difference, in: H.R. Wu, K.R. Rao (Eds.), Digital Video Image Quality and Perceptual Coding, CRC Press, 2006, pp. 281–304 Chapter 9.

[72] X.K. Yang, W.S. Ling, Z.K. Lu, E.P. Ong, S.S. Yao, Just noticeable distortion model and its applications in video coding, Signal Process.: Image Commun. 20 (2005) 662–680.

[73] J. Guo, M.V. Dyke-Lewis, H.R. Myler, Gabor difference analysis of digital video quality, IEEE Trans. Broadcast. 50 (3) (2004) 302–311 September.

[74] A.B. Watson, J. Hu, J.F. McGowan III, DVQ: a digital video quality metric based on human vision, J. Electron. Imaging 10 (1) (2001) 20–29.

[75] F. Xiao, DCT-based video quality evaluation, [online] Available: 〈http://compression.ru/video/quality_measure/vqm.pdf〉.

[76] C. Lee, O. Kwon, T. Jung, S. Cho, H. Kim, Weighted PSNR for objective measurement of video quality, Proc. Visualization ImagingImage Process. (2002) 130–133.

[77] Z. Lu, W. Lin, X. Yang, et al., Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation, IEEE Trans. Image Process. 14 (11) (2005) 1928–1942 November.

[78] X. Feng, T. Liu, D. Yang, and Y. Wang, Saliency based objective quality assessment of decoded video affected by packet losses, in: Proceedings of the IEEE International Conference on Image Processing, California, USA, October 12–15, 2008, pp. 2560–2563.

[79] J. You, A. Perkis, M. Hannuksela, and M. Gabbouj, Visual quality assessment based on human attention analysis, in: Proceedings of the ACM Multimedia 09, Beijing, China, October 2009, pp. 561–564.

[80] C. Lee, O. Kwon, Objective measurements of video quality using the wavelet transform, Opt. Eng. 42 (1) (2003) 265–272 January.

[81] M.H. Pinson, S. Wolf, A new standardized method for objectively measuring video quality, IEEE Trans. Broadcast. 50 (3) (2004) 312–322 September.

[82] J. You, M. Hannuksela, and M. Gabbouj, An objective video quality metric based on spatiotemporal distortion, in: Proceedings of the IEEE International Conference on Image Processing, Cairo, Egypt, November 2009, pp. 2229–2232.

[83] Z. Wang, L. Lu, A.C. Bovik, Video quality assessment based on structural distortion measurement, Signal Process.: Image Commun. 19 (2) (2004) 121–132 February.

[84] L. Itti, C. Koch, Computational modeling of visual attention, Nat. Rev. Neurosci. 2 (3) (2001) 194–203 March.

[85] Video Quality Metric (VQM) Software, [online] available: 〈http://www.its.bldrdoc.gov/n3/video/VQM_software.php〉.

[86] MSU Video Quality Measurement Tool, [online] available: 〈http://compression.ru/video/quality_measure/video_measurement_too l_en.html〉.

[87] M. Pinson, and S. Wolf, An Objective method for combining multiple subjective data sets, in: Proceedings of the SPIE Video Communication and Image Processing Conference, Lugano, Switzerland, July 2003.

[88] Recommendation ITU-R BT.500-11, Methodology for the subjective assessment of the quality of television pictures, International Telecommunication Union, Geneva, Switzerland, 2002.

[89] Recommendation ITU-R BS.775-1, Multichannel stereophonic sound system with and without accompanying picture, International Telecommunication Union, Geneva, Switzerland, 1994.

[90] Recommendation ITU-R BS.1286, Methods for the subjective assessment of audio systems with accompanying picture, International Telecommunication Union, Geneva, Switzerland, 1997.

[91] Recommendation ITU-R BT.1359, Relative timing of sound and vision for broadcasting, International Telecommunication Union, Geneva, Switzerland, 1998.

[92] Question ITU-R 102/6, Methodologies for subjective assessment of audio and video quality, International Telecommunication Union, Geneva, Switzerland, 1999.

[93] Recommendation ITU-T P.910, Subjective video quality assessment methods for multimedia applications, International Telecommunication Union, Geneva, Switzerland, 1999.

[94] M. Meehan, B. Insko, M. Whitton, and F.P. Brooks, Physiological measures of presence in stressful virtual environments, in: Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques, San Antonio, Texas, 2002, pp 645–652.

[95] H. Coolican, in: Research Methods and Statistics in Psychology, 4th Edition, Hodder Arnold, London, UK, 2004 (ISBN 0-240-81258-3).

[96] S. Bech, N. Zacharov, in: Perceptual Audio Evaluation – Theory, Method And Application, John Wiley & Sons Ltd., Chichester, West Sussex, England, ISBN 0-470-86923-2, 2006.

[97] U. Reiter and S. Jumisko-Pyykkö, Watch, press and catch – impact of divided attention on requirements of audiovisual quality, in: Proceedings of the 12th International Conference on Human–Computer Interaction, HCI2007, Beijing, PR China, July 22–27, 2007.

[98] S. Jumisko-Pyykkö, U. Reiter, and C. Weigel, Produced quality is not perceived quality – a qualitative approach to overall audiovisual quality, in: Proceedings of the 3DTV Conference, Kos Island, Greece, May, 2007.

[99] Y. Yahya, J. Donaldson, J. Jenkins, Assessing multimedia quality from the user's perceptive, Malays. J. Comput. Sci. 12 (1999) 9–18.

[100] T. Hayashi, K. Yamagishi, T. Tominaga, A. Takahashi, Multimedia quality integration function for videophone services, Proc. IEEE Int. Conf. Global Telecommun.(GLOBECOM '07) (2007) 2735–2739 November.

[101] Y. Zhong, I. Richardson, A. Sahraie, and P. Mcgeorge, Influence of task and scene content on subjective video quality, in: Proceedings of the International Conference on Image Analysis and

Recognition, LNCS Part I, ICIAR ‘04, Porto, Portugal, September–October 2004, pp. 295–301.

[102] ITU SG 12 Contribution 61 COM 12-61-E, Study of the influence of experimental context on the relationship between audio, video and audiovisual subjective qualities, ITU Telecommunication Standardization Sector, September 1998.

[103] M. Ries, R. Puglia, T. Tebaldi, O. Nemethova, and M. Rupp, Audiovisual quality estimation for mobile streaming services, in: Proceedings of the second International Symposium on Wireless Communications Systems (ISWCS ’05), Siena, Italy, September 2005, pp. 173–177.

[104] T.C. Thang, J.W. Kang, and Y.M. Ro, Graph-based perceptual quality model for audiovisual contents, in: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME’07), Beijing, China, July 2007, pp. 312–315.

[105] H.E. Pashler, in: The Psychology of Attention, MIT Press, Cambridge, MA, 1998.

[106] Digitizing Software: GetData Graph Digitizer, [online] available: 〈http://getdata-graph-digitizer.com〉/.

[107] U. Reiter, Dissertation: bimodal audiovisual perception in interactive application systems of moderate complexity, [online] available: 〈http://www.db-thueringen.de/servlets/DocumentServlet?id=14098&lang=en〉.

[108] D.E. Broadbent, in: Perception and Communication, Pergamon Press, London, UK, 1958.

[109] M.R. Naphade, T.S. Huang, Extracting semantics from audiovisual contents: The final frontier in multimedia retrieval, IEEE Trans.-Neural Networks 13 (4) (2002) 793–810 July.

[110] J. You, M. Hannuksela, and M. Gabbouj, Semantic audiovisual analysis for video summarization, in: Proceedings of the IEEE International Conference on Region 8 Eurocon ‘09, Saint-Petersburg, Russia, May 2009, pp. 1358–1363.

[111] G. Ghinea, and J.P. Thomas, QoS impact on user perception and understanding of multimedia clips, in: Proceedings of the ACM Multimedia 98, Bristol, 1998, pp. 49–54.

[112] R. Cucchiara, C. Grana, A. Prati, Semantic video transcoding using classes of relevance, Int. J. Image Graphics 3 (1) (2003) 145–169 January.

[113] T.C. Thang, Y.M. Ro, Multimedia quality evaluation across different modalities, Proc. SPIE Electron. Imaging 5668 (2005) 270–279 January.

[114] T.C. Thang, Y.S. Kim, C.S. Kim, Y.M. Ro, Quality models for audiovisual streaming, Proc. SPIE Electron. Imaging 6059 (2006) 1–10 January.

[115] A. Hanjalic, L.-Q. Xu, Affective video content representation and modeling, IEEE Trans. Multimedia 7 (1) (2005) 143–154 February.

[116] J. Korhonen, and J. You, Improving objective video quality assessment with content analysis, in: Proceedings of the fifth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM) Scottsdale, USA, January *2010*.

[117] J. You, J. Korhonen, and A. Perkis, Spatial and temporal pooling of image quality metrics for perceptual video quality assessment on packet loss streams, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Dallas, USA, March 2010.

[118] D. Isherwood, Unpublished report: Audio–visual subjective testing of DVB-H IPDC coding parameters, Nokia Research Center, March 2005.

[119] T. Hayashi, and K. Yamagishi, State of the art of multimedia quality assessment methods, ITU-T Workshop on Video and Image Coding and Application (VICA), July 2005.