AUDIO-VISUAL MULTIMEDIA QUALITY ASSESSMENT FACTORS

Dr. T. Amitha
Professor /CSE Dept, SVS GROUPS OF INSTITUTION
Dr. B.Raghu ,Principal & Professor/CSE, SVS GROUPS OF INSTITUTION , Waraganal, Telangana.

ABSTRACT:
The notion of *quality* is an abstract concept and contemplated as a construct of the mind, which is easy to understand but difficult to define. In multimedia field, quality is typically used with an engineering goal in mind due to the fact that quality is a key criterion to evaluate systems, services or applications during both design and operation phases While according to QUALINET white paper *"quality is the outcome of an individual's comparison and judgment process, which includes perception, reflection about the perception, and the description of the outcome"*. Contrary to definitions/concepts in which quality is seen as "qualities" (i.e., a set of inherent characteristics), QUALINET considers quality in terms of the evaluated excellence or goodness, of the degree of need fulfillment, and in terms of a "quality event", where event is an observable occurrence and determined in space (i.e., where it occurs), time (i.e., when it occurs), and character (i.e., what can be observed) Quality can be gauged both at the service provider or user sides. QoS and QoE describe aspects related to the acceptability of a service and degree of sentiment of a person experiencing an application, system, or service, respectively. Understanding human (quality) perception processes would help to apprehend how the quality impression is created in the mind of the user. Therefore, in the following subsection we discuss the human perception process.

Keywords: Qos, QoE, Subjective Quality Assessment, Objective Quality Assessment

Quality Formation Process

A critical design goal for an audio-visual multimedia coding/transmission/decoding/display system is to produce audio and video signals of quality to be acceptable and pleasant to the human observer. It is well known that the formation of quality hugely depends on the human perception process. There are various theories and studies attempting to describe how humans perceive physical events via their sensory system]. Understanding how human observers view/hear, interpret and respond to visual/audio stimuli would help to formulate not only design principles for audio/video encoding, decoding and display but also methods for their perceived quality evaluation. Human quality perception may be defined as a conscious sensory experience and process made of low-level sensorial and high-level cognitive processing levels The physical stimulus or signals (e.g., a sound wave for an auditory signal) are converted into electric signals for the nervous system by the low-level sensorial processing level. In turn, the conscious processing (i.e., interpretation and understanding) of the neural signals are carried out by high-level cognitive processing to form a perceived quality judgment. Though, quality judgment originates from the neuronal processing of a physical stimulus, it is also influenced by contextual information (i.e., physical environment), other modalities, mental states (e.g., mood, emotions, attitude, goals, and intentions) and previous knowledge or experiences. *Visual perception* is the ability to interpret the surrounding environment through what we see. Due to great complexity, many theories regarding the relationships among visual psychological phenomena are in the hypothesis stage. However, several studies have shown that luminance non-linearity, contrast sensitivity, masking effects, multi-channel parallel and visual attention are necessary building blocks of visual perception . Visual attention refers to a cognitive operation that selects relevant and filters out irrelevant visual information. Existing visual attention

theories can be grouped into space-based (i.e., attention is directed to discrete regions of space within the visual field) and object-based (i.e., attention is directed to the object, rather than its location per se). From a Psychology point of view, visual attention can be either bottom-up saliency (i.e., influenced by low-level features of the environment/target) or top-down saliency (i.e., influenced by person's cognitive processing).

Auditory perception is regulated by two prominent elements, i.e., masking and binaural hearing, besides attention. Auditory masking is a perceptual event in which subject cannot respond in the presence of one perceived auditory stimulus to another one (i.e., generally lower level signal). While, the perception of the direction of a sound source in the space including blur of a sound is feasible due to *binaural hearing*. It has been experimentally proved that the differences in the intensity and timing of sounds perceived by both ears are exploited as cues for directional perception.

On the whole, like many functions of the nervous system, there exist several unproven audio and video perception theories. However, there are two main processing schemes (which are commonly adopted in the literature as well as in the practice): bottom-up and top-down. It is believed by the bottom-up and top-down processing theorists that low-level sensory information and higher-level cognitive processes, respectively, are the most vital determinants of what humans perceive; while some scientists state that the truth may be lying somewhere in between.

Quality Assessment

There are basically two categories of quality assessment (QA) methods, namely the subjective methods that involve human observers to assess the quality of multimedia contents, and objective methods that compute the quality automatically using mathematical models.

Subjective Quality Assessment

In order to reliably measure the perceived quality by human auditory and/or visual systems, subjective tests are performed where groups of trained or naive human observers provide quality ratings [1]. This evaluation procedure is known as subjective quality assessment that seeks to quantify range of opinions that users express when they see/hear the digital content. Subjective quality assessment is carried out generally in a well-controlled environment using standardized recommendations (e.g., International Telecommunication Union Radio communication Sector [ITU-T] guidelines). Subjective quality assessment can be categorized as double stimulus or single stimulus methods. In double stimulus methodology, subject is presented with the source and test samples to evaluate their qualities. In single-stimulus methodology, the subject is presented with the test only without the source as reference to evaluate quality. The single-stimulus methodology is more useful in realistic test environment, such as conversational tests in which two subjects interactively listen and talk through transmission system under evaluation to provide quality. The scale for rating can be either numerical or categorical, and either continuous or discrete. The rating can be obtained after or during stimulus presentation to acquire overall quality or temporal quality variations, respectively.

To study the impact of environmental or contextual factors on MOS, an international experimental study using 10 datasets from different laboratories was conducted in . The study concluded that the performance obtained from 24 users under a controlled environment was analogous to the one obtained from approximately 35 users under a public environment. Though subjective quality assessment techniques can reliably determine the perceived quality, they are time consuming, expensive, laborious, not instantaneous, and could not be incorporated in adaptive systems that adjust their operating parameters automatically based on measured quality feedback. Moreover, subjective ratings usually have high variance between subjects possibly due to different expectations/experiences

of technology, viewing/hearing distance, digital media player, subject's mood and vision/hearing ability.

Objective Quality Assessment

Although subjective quality assessment provides reliable human perception quality cues, it cannot be applied in real-time in-service quality evaluation. Thus, objective quality assessment methods have been developed to replace the human panel by a computational model for predicting results of a subjective test. Namely, the goal of objective quality assessment is to automatically estimate MOS values, which are as close as possible to quality scores obtained from subjective quality assessment .The numerical measures of quality obtained from the objective method (also referred to as objective or predicted MOS) are expected to better correlate with human subjectivity. There are various metrics to measure the relationship between subjective MOS and predicted MOS. Two most common statistical metrics used to report the performance of objective quality assessment methods are 'Root Mean Square Error (RMSE)' and 'Pearson Correlation'. An objective quality assessment algorithm having a high correlation (usually greater than 0.8) is apprised as efficacious

Two main advantages of objective quality assessment usage are defining meaning of MOS for a given application (i.e., people know what a MOS of 3 means in terms of quality), and reproducible MOS prediction (i.e., different people utilizing the tool for the same test samples obtain the same results). Objective quality measurement techniques can be classified into five groups, as per the ITU recommendation, based on the type of input data being utilized by the metrics

1. *Media-layer models*—the models in this category do not require any information about the system in question. Particularly, these models utilize only audio or video samples to estimate the quality, and can be applied to applications such as codec optimization and codec comparison.
2. *Parametric packet-layer models*—the solutions to predict quality in this group are lightweight since parametric packet-layer models have to only process the packet-header information without dealing with the media signals.
3. *Parametric planning models*—these models employ encoding and networks parameters to predict quality. Thus, they demand a priori knowledge about the system in question.
4. *Bit stream-layer models*—these models predict the quality using encoded bit stream and packet-layer information that is utilized in parametric packet-layer models.
5. *Hybrid models*—the models in this class usually integrate two or more of the above-mentioned models.
6. On the other hand, objective quality assessment techniques can also be classified into three categories: full-reference (FR), reduced-reference (RF) and no-reference (NR) according to the availability of the reference (original/ideal), partial information about the reference, or no reference for evaluating quality, respectively.

FR methods measure the impairment in the test signal with respect to a reference signal, thereby requires availability of entire original signal. Though it provides a highly accurate objective quality assessment owing to the use of original signal this is considered expensive and often not applicable for all services and applications, e.g., IPTV monitoring. RR methods evaluate the quality by comparing a small amount of respective features extracted from reference and test samples. Since the RR methods utilize information from source signal, they are fairly precise but less than FR methods. Both FR and RR are vital for non-real-time quality monitoring. NR methods predict the quality using only the test signal without the requirement of an explicit reference signal. Since these methods do not need the reference signal and make assumptions about the multimedia content and types of distortions, they are less accurate. With respect to reference requirements, FR and RR are also termed as double-ended, while NR as single-ended metrics. In addition, depending on usability, objective methods can also be categorized as out-of service and in-service methods. In the former, no time

constraints are placed and the original sequence can be available. In the latter, time constraints are placed and quality is evaluated during streaming.

Audiovisual Quality Assessment (AVQ)

The psychophysical processes responsible for the perception of uni-modal stimuli (i.e., audio or video) have been extensively studied and well accepted. However, little research on audiovisual quality perception (i.e., a multimodal process involving both human visual and auditory systems) has been conducted leading to the lack of theoretical and practical understandings of perceived multimodal quality. In other words, from a engineering point of view, it is still unknown how to most efficiently model the perception of audiovisual quality. Likewise, from a neurophysiological point of view, there is a long way to go to answer the question 'for multimodal quality processing, at what stage is the information originated from various brain's functional areas and how are they aggregated?'

Although detailed understanding of low-level multimodal quality perception is yet available, some experimental analyses have observed that there is a noteworthy mutual influence between auditory and visual stimuli in the overall perceived quality , According to the well-adopted 'late fusion' theory, the audio and visual modalities are internally processed to yield individual auditory and visual qualities, which are then integrated towards the end stages of the overall perceived quality estimation procedure. It seems rational to utilize relatively matured audio and video perceptual quality measures as primary inputs to the AVQ models. As depicted in Fig. 4, the elementary inputs to perception-based multimodal quality assessment model are derived from independent psychophysical based audio and video quality assessment modules. The multimodal fusion schemes are then applied to individual base information from elementary inputs (modalities) to produce perceived multimodal quality. As such, the choice of fusion rule(s) is a very decisive and vital for design and performance of AVQ methods. A fully functional AVQ model is expected to account for different quality attributes (e.g., spatial-temporal properties), other influential factors and missing data issue (i.e., when any (or more) of the unimodal input is missing). There can be seven combinations of stimulus types and quality assessment tasks, as presented in Table 1. For instance, Stimuli-Assessment: Audio-Audiovisual quality pair indicates the audiovisual quality when information from video modality is missing and only audio stimulus is present. Since audio and visual information play most dominant roles in perceived audiovisual quality, therefore the multimodal quality is commonly derived by a linear combination and a multiplication using audio and video qualities as:

$$QAV=a0+a1QA+a2QV+a3QAQV \quad (1)$$

where QAV,QA,QV and {a0,a1,a2,a3} are predicted audiovisual quality, audio quality, video quality and weights, respectively. Though a0 is irrelevant to the correlation between the predicted and perceived qualities, it improves the fit in terms of the residual between them. It is also worth noticing that the multiplication of AQ and VQ has high correlation with the overall predicted quality [6].

Audiovisual Multimedia Quality: Factors and Degradation

This section describes the factors that may influence quality of audio or/and visual samples. Further, audio and visual features that are commonly utilized in objective quality assessment are studied.

Factors Influencing Audiovisual Multimedia Quality

For better assessment algorithms, it is appreciated to understand complex and strongly interrelated factors that impact user interaction behaviors as well as perceived quality. Some factors are inevitable, while some are due to inherent limitations of the multimedia signal itself. These factors can be grouped into three categories: human, technological and contextual influential factors.

- *Human Influential Factors:* encompass variant or invariant characteristics of the human user that may impact quality judgment, which includes physical/mental constitution/emotional state, demographic, and socio-economic background. These attributes are either static (e.g., gender, age) or dynamic (mental states, motivation). The user factors may take part in sensory or/and cognitive quality processes. The early sensory (i.e., low-level) quality process is affected by user's physical, emotional and mental states, e.g., user's auditory acuity, user's mood, and attention. The cognitive (i.e., higher-level/top-down) quality process relates to the interpretation of stimuli based on user's knowledge and background that include individual's need, motivation, preference, and so on.
- *Technological Influential Factors:* encompass agent (an interaction partner) and functional factors of the system. The examples of agent factors are technical attributes (e.g., speech recognition). The examples of functional factors are functional capabilities (e.g., number of tasks) and domain characteristics (e.g., entertainment system). The system factors may be further divide into four classes as network-related (i.e., associated to data transmission over a network, e.g., bandwidth), device-
- related (i.e., associated to communication end system/device, e.g., high resolution smart phone), media-related (i.e., associated to media configuration, e.g., frame rates) and content-related (i.e., associated to amount of media information, e.g., voice/spoken vs. musical contents).

- *Contextual Influential Factors:* encompass physical environment (e.g., office) and service factors (i.e., non-physical system attributes, e.g., system access restrictions). The context factors can also be broken down as physical context (i.e., location and space characteristics, e.g., peaceful/noisy place), temporal context (i.e., experience's temporal aspect, e.g., month June or spring season), social context (i.e., interrelationship among users, e.g., hierarchical dependencies like boss and employee), economic context (i.e., business perspective, e.g., cost per usage), task context (i.e., experience of user for perceived quality, e.g., effect of multitasking while quality rating), and technical and information context (i.e., relationship between the involved or optional systems and devices, e.g., interconnectivity of devices over Bluetooth).

Degradations of Audio and Visual Signals

In order to better understand audiovisual quality assessment it might be helpful to closely inspect the different artifacts that commonly manifest in audio and video signals. The audio/visual degradations are manifested by the properties of the signal capture device, encoding, decoding, compression or transmission mechanism, or end device being used by the human subjects. The typical examples of visual degradations are blurring (i.e., loss of spatial information or edge sharpness due to incorrect focus, motion or context factors), edginess (i.e., the distortions happened at the edges), motion jerkiness due to jitter (i.e., time-discrete intermission of the original continuous, smooth scene), blockiness (i.e., discontinuity at the boundaries of two adjacent blocks owing to video coding schemes), jerkiness (i.e., non-fluent and non-smooth presentation of frames), flickering (i.e., noticeable discontinuity between consecutive frames), color bleeding (i.e., smearing of colors between areas of differing chrominance), ringing (i.e., shimmering effect around high contrast edges) illumination, and color naturalness (affected by color rendering). The typical examples of audio degradations are loudness (i.e., a psycho-physiological attribute correlating of physical strength), reverberation, naturalness, pitch fluctuations, distortion, and delay. Spatial or temporal misalignment or unsynchronization, in turn, is most vital degradation in audiovisual multimedia content. *Alignment* between degraded and original audio-visual signals, and *synchronization* of audio and video channels more considerably affect objective quality assessment than subjectively [6].

References:

1) Barkowsky M, Eskofier B, Bialkowski J, Bitto R, Kaup A (2009) Temporal trajectory aware video quality measure. IEEE J Sel Top Sig Proc 3(2):266–279, Issue on Visual Media Quality AssessmentCrossRefGoogle Scholar
2) Barland R, Saadane A (2006) A reference free quality metric for compressed images. In: Proc of 2nd int workshop on video processing and quality metrics for consumer electronicsGoogle Scholar
3) Beerends JG, De Caluwe FE (1999) The influence of video quality on perceived audio quality and vice versa. J Audio Eng Soc 47(5):355–362Google Scholar
4) Campisi P, Carli M, Giunta G, Neri A (2003) Blind quality assessment system for multimedia communications using tracing watermarking. IEEE Trans Signal Process 51(4):996–1002CrossRefMathSciNetGoogle Scholar
5) Campisi P, Le Callet P, Marini E (2007) Stereoscopic images quality assessment. In: Proceedings of 15th European signal processing conference (EUSIPCO07)Google Scholar
6) Cermak GW (2009) Consumer opinions about frequency of artifacts in digital video. IEEE J Sel Top Sig Proc 3(2):336–343CrossRefGoogle Scholar
7) Chandler DM, Hemami SS (2007) VSNR: a wavelet-based visual signal-to-noise ratio for natural images. IEEE Trans Image Process 16(9):2284–2298CrossRefMathSciNetGoogle Scholar
8) Channappayya SS, Bovik AC, Caramanis C, Heath RW (2008) Design of linear equalizers optimized for the structural similarity index. IEEE Trans Image Process 17(6):857–872CrossRefMathSciNetGoogle Scholar
9) Channappayya SS, Bovik AC, Heath RW (2006) A linear estimator optimized for the structural similarity index and its application to image denoising. In: IEEE international conference on image processing, pp 2637–2640Google Scholar
10) Chen GH, Yang CL, Xie SL (2006) Gradient-based structural similarity for image quality assessment. In: IEEE international conference on image processing, pp 2929–2932Google Scholar
11) Chen MJ, Bovik AC (2009) No reference image blur assessment using multiscale gradient. In: 1st international workshop on quality of multimedia experience (QoMEX)Google Scholar
12) Chen MJ, Bovik AC (2010) Fast structural similarity index. In: IEEE international conferene on acoustics, speech and signal processing (ICASSP)Google Scholar
13) Datta R, Joshi D, Li J, Wang JZ (2006) Studying aesthetics in photographic images using a computational approach. Lect Notes Comput Sci 3953:288CrossRefGoogle Scholar
14) Datta R, Li J, Wang JZ (2008) Algorithmic inferencing of aesthetics and emotion in natural images: an exposition. In: IEEE intl conf image proc, pp 105–108Google Scholar
15) Dignan L (2010) Tiered mobile data plans accelerate: verizon wireless to follow AT&T's lead. ZDnet.com