IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 25, 2017, accepted August 31, 2017, date of publication September 11, 2017, date of current version October 25, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2750918

Audio-Visual Multimedia Quality Assessment: A Comprehensive Survey

ZAHID AKHTAR AND TIAGO H. FALK, (Senior Member, IEEE)
INRS-EMT, University of Quebec, Montreal H5A 1K6, Canada

Corresponding author: Zahid Akhtar (zahid.eltc@gmail.com)

ABSTRACT Measuring perceived quality of audio-visual signals at the end-user has become an important parameter in many multimedia networks and applications. It plays a crucial role in shaping audio-visual processing, compression, transmission and systems, along with their implementation, optimization, and testing. Service providers are enacting different quality of service (QoS) solutions to issue the best quality of experience (QoE) to their customers. Thus, devising precise perception-based quality metrics will greatly help improving multimedia services over wired and wireless networks. In this paper, we provide a comprehensive survey of the works that have been carried out over recent decades in perceptual audio, video, and joint audio-visual quality assessments, describing existing methodologies in terms of requirement of a reference signal, feature extraction, feature mapping, and classification schemes. In this context, an overview of quality formation and perception, QoS, QoE as well as quality of perception is also presented. Finally, open issues and challenges in audio-visual quality assessment are highlighted and potential future research directions are discussed.

INDEX TERMS Subjective quality assessment, objective quality metric, multimedia quality, signal-driven model, audiovisual perception, quality of service, data-driven analysis.

I. INTRODUCTION

The recent evolution of digital communication systems (e.g., 3G and 4G) has led to an explosion of multimedia services and applications, such as IPTV, mobile multimedia on smartphones, social networking (e.g., Facebook), immersive multimedia and virtual reality based games, video conferencing, and educational multimedia presentations, to name a few. These multimedia applications now have become an integral (if not indispensable) part of daily lives, and expected to grow further exponentially. Multimedia service providers are formulating various techniques to provide better quality of experience (QoE), which is increasingly being demanded by end-users. Thus, human's opinion about quality is critical in the design and deployment of any current and future multimedia networks and services [1].

Audio and video are two core modalities in most multimedia applications. Despite recent advances, audio-visual signals suffer from impairments through both lossy source encoding and transmission over error prone channels, leading thereby to degraded quality of the multimedia signal [2]. For instance, as shown in Fig. 1, a video sample received by the end user may posses a wide range of quality due to different transmission or rendering errors. Accurately estimated quality of the transmitted audio-visual signals may contribute hugely in multimedia services and communication networks. In fact, quality assessment for digital signals is one of the basic and challenging problems in the field of multimedia processing and its practical situations, such as process evaluation, implementation, optimization of encoding and decoding, testing and monitoring (e.g., in transmission and manufacturing sites). Moreover, how to evaluate audio and video quality plays a central role in shaping most (if not all) multimedia services, algorithms and systems [3]. Few other examples of technological dependence upon audio-visual quality assessment are signal acquisition, synthesis, enhancement, compression, watermarking, storage, retrieval, reconstruction, rendering, and presentation (e.g., display on mobile device).

Quality assessment (QA) of an audio, video, or audio-visual signal measures it's degradation during acquisition, compression, transmission, processing, and reproduction. In today's highly interconnected digital societies, reliable quality assessment decidedly helps not only in meeting the promised QoS (quality of service) but also in improving the end user's QoE [4]. QA methods can be categorized into two broad classes: subjective and objective. Subjective (perceptual) QA methods are based on groups of trained (or naive) users viewing multimedia content, and

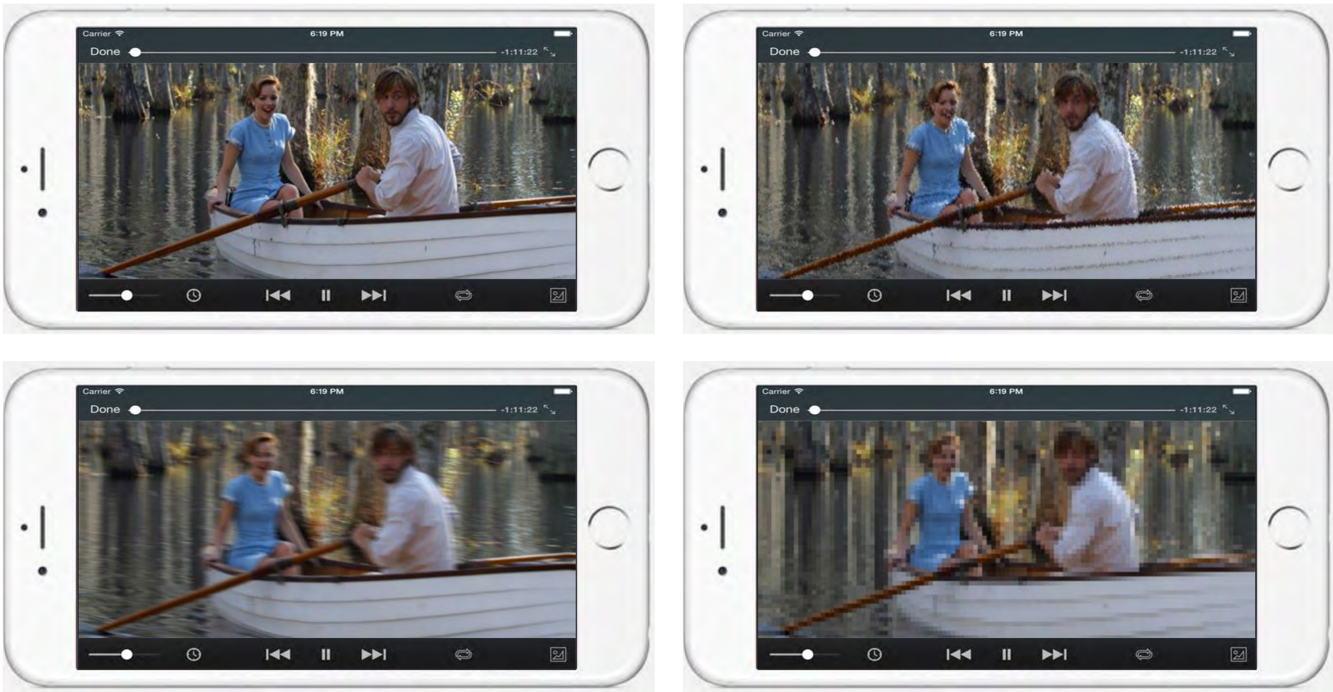

FIGURE 1. **Examples of video frames with different quality received on user's mobile device.**

providing ratings for quality [9]. However, subjective methods are time-consuming, laborious and not applicable in real-time. It is thus imperative to devise computational models that are able to predict the evaluation of an average observer. To this end, objective methods have been proposed, which are based on signal fidelity measures (e.g., signal-to-noise-ratio) or network parameters (e.g., packet loss rates). Despite objective audio-visual QA algorithms being computationally simple and well defined with clear physical meanings, they have been shown to be poor predictor of perceived quality because they usually disregard the viewing conditions, the characteristics of human audio-visual perception, and not every change in a multimedia content is noticeable, not each fragment receives the same attention level, and not every change yields the same extent of perceptual effect with the same magnitude of change [6]. Therefore, the International Telecommunication Union (ITU) has outlined basic requirements for objective perceptual multimedia quality modeling [1].

Multimedia quality assessment can be of first (the multimedia content maker), second (the subject(s) of a multimedia sample), or third party (neither the maker nor the subject(s)) level [12]. The main focus of this survey is the perception of third-party observers, since it represents the most practical and meaningful situation in applications as well as in modeling. Though recent progress in developing objective quality assessment models in line with the human perceptual system for multimedia services, it is still a long and challenging odyssey [1], [6], [7] owing to the multi-disciplinary complex nature of the problem (related to psychology, physiology, vision and audio research and computer science), the limited understanding of the human vision and auditory mechanisms, and the diversified scope of available applications and requirements. Moreover, it is easy to notice in the literature that most published works on quality assessments have been focused on individual modalities only, i.e., audio and video independently.

Over the years, several survey papers [6], [10], [11] and books [4], [9], [12] on quality assessment have been published, but with limited scope. For instance, [10], [12] discussed only video quality assessment methods, while [9], [11] gives details about audio quality evaluation techniques mostly focusing on objective quality models. You *et al.* [6] provided a review on audio-visual perceptual quality assessment methods. However, they focused mainly on so-called full-reference quality models (i.e., that require a reference signal) and coding distortions, thus ignoring several issues such as quality degradation by packet losses during transmission and so on. Further, You *et al.* [6] did not detail QoS, QoE and QoP (quality of perception), which have newly emerged and raising great research interest. This paper significantly differs from the previous articles as it provides comprehensive overview of the evolution of multimedia perceptual quality assessment methods including quality formation and perception, datasets and current challenges and future research directions. Among the significant contributions of this survey article, we can cite:

- A description of quality formation and perception including various quality influential factors,

- A general overview of QoS, QoE, and QoP in context of audiovisual quality assessment,
- A survey of a wide range of audio, video, and audio-visual quality assessment methodologies following a systematic categorization with use of reference signals, and feature extraction and mapping schemes,
- A synopsis of publicly available databases for audio, video and audio-visual perceptual quality assessment, and
- A discussion of open issues and future research directions for uni- and multi-modal quality assessment and QoE.

The rest of the paper is structured as follows. Section II discusses quality assessment, perception and formation. Different quality influential factors are discussed in Section III. Section IV summarizes the existing QoS, QoP and QoE methods. Section V presents a survey of existing quality assessment pertaining to audio, video, and audio-visual channels. In Section VI, publicly available databases for quality evaluation purposes are enlisted. Future research directions, and conclusions are described in Sections VII and VIII, respectively.

II. AUDIO-VISUAL MULTIMEDIA QUALITY ASSESSMENT

This section introduces the key notions related to concept of multimedia quality and its formation and evaluation.

A. DEFINITION OF QUALITY

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 [13]. While according to QUALINET white paper [15], *"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 "qualitas" (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) [15].

Fundamentally speaking, quality is the outcome of a human judgment based on various criteria. Some of them can be based on measurable intrinsic information of the signal, while others are based on cognitive processes thereby usually unmeasurable. Namely, quality can be conceived of as an umbrella term, since several variables contribute to form a cognizance of quality. For instance, for audio quality, covariates such as listening effort, loudness, pleasantness of tone and intelligibility are vital. For visual and audiovisual quality, in turn, factors such as image size, frame rate, and packet loss, degree of audio-visual synchronization, respectively, play a crucial role.

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.

B. 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 [4]. There are various theories and studies attempting to describe how humans perceive physical events via their sensory system [16], [17]. 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 [16]. 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, 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 [19], [20]. 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 [21], 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 [19].

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.

C. 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.

1) 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 Radiocommunication 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. Generally, the absolute category rating (ACR) is employed asking users to make a single rating for the test sample using ITU recommended 5-point category scale ranging from 'bad' to 'excellent' as depicted in Fig. 2. The final quality score is obtained by averaging the rating values registered by multiple subjects, which is referred as mean opinion score (MOS) and difference mean opinion score (DMOS) for single- and double-stimulus methodologies, respectively [6].

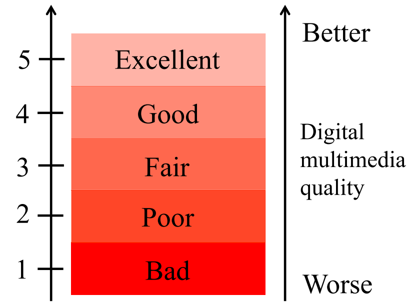

FIGURE 2. The ITU recommended ACR quality measurement scale. Human observers are usually asked to rate the digital multimedia sample in terms of annoyance, where annoyance is a measure of how 'bad' the observer believe impairment is; as annoyance value is correlated with strength of the impairment.

To study the impact of environmental or contextual factors on MOS, an international experimental study using 10 datasets from different laboratories was conducted in [23]. 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.

2) 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 [9], [10], [24]. 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 [13].

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 [13], [25]:

i *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.
ii *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.
iii *Parametric planning models*—These models employ encoding and networks parameters to predict quality. Thus, they demand a priori knowledge about the system in question.
iv *Bitstream-layer models*—These models predict the quality using encoded bitstream and packet-layer information that is utilized in parametric packet-layer models.
v *Hybrid models*—The models in this class usually integrate two or more of the above-mentioned models.

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.

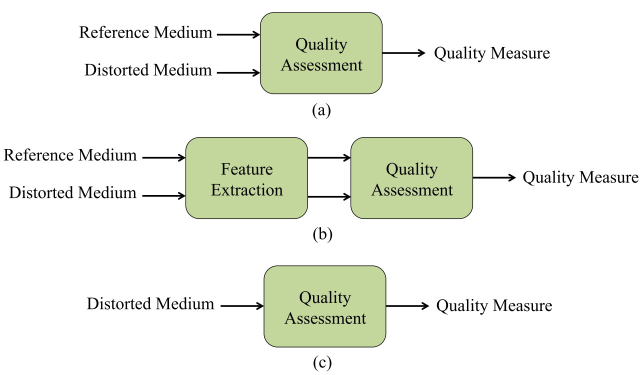

FIGURE 3. Overview of (a) Full-reference method, (b) Reduced-reference, (c) No-reference method.

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 (as shown in Fig. 3a), 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.

3) 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 [13], besides other factors (e.g., audio-visual content itself) that are detailed in Section III-A. 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 assess-

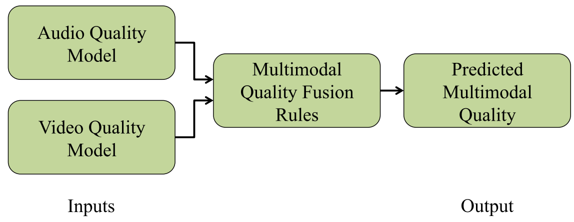

FIGURE 4. Basic multimedia quality estimation model.

TABLE 1. Quality estimation for seven different presentations.

Stimuli	Assessment
Audio	Audio quality
Audio	Audiovisual quality
Video	Video quality
Video	Audiovisual quality
Audio + Video	Audio quality
Audio + Video	Video quality
Audio + Video	Audiovisual quality

ment 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:

$$Q_{AV} = a_0 + a_1 Q_A + a_2 Q_V + a_3 Q_A Q_V \tag{1}$$

where Q_{AV}, Q_A, Q_V and $\{a_0, a_1, a_2, a_3\}$ are predicted audio-visual quality, audio quality, video quality and weights, respectively. Though a_0 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 A_Q and V_Q has high correlation with the overall predicted quality [6].

III. 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.

A. 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 smartphone), 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). Table 2 presents some possible causes of each of the aforementioned quality factors.

B. 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

TABLE 2. Summary and examples of potential quality influential factors.

Factors	Examples	Explanation
Human Influential Factors		
Low-level:		
physical	Gender, age, visual or auditory acuity	
emotional	Mood	Each user has their own perception
mental	Attention level	towards a multimedia quality based
High-level:		on individual expectation and attitude.
understanding	Socio-cultural background, socio-economic state	
interpretation	Goal, motivation	
evaluation	Previous experiences and knowledge, skills	
Technological Influential Factors		Attributes, properties and
Content-related	Audio bandwidth, dynamic range, video motion and detail	characteristics which dictate the
Media-related	Encoding, resolution, sampling rate, frame rate, synchronization	technically produced quality of
Network-related	Bandwidth, delay, jitter, loss, error rate, throughput	a service or application.
Contextual Influential Factors		
Physical context	Location, space, environmental characterstics, motion	Describes situational ambient properties
Temporal context	Time, periodic cycle of use	to indicate how a user may perceive the
Social context	Inter-personal relations	multimedia content, since perceived quality
Economic context	Costs, subscription type, brand	varies according to when, where, and with
Task context	Nature of experience, task type, interruptions, parallelism	whom the media is exploited.
Technical or informational context	Compatibility, interoperability	

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 audio-visual 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]. Hollier and Rimell [27] and Peltoketo [28]conducted several experimental studies on temporal asymmetry with different stimuli samples considering audio-visual communications systems. They pointed out that audio cannot lead the visual stimuli/percept owing to the difference in sound and light travelling rates. The findings in [27] and [28] have hugely influenced the synchronization thresholds recommendation in ITU-T J.100 [29], which are 40 ms for video lead and 20 ms for audio lag.

IV. QUALITY OF SERVICE, QUALITY OF EXPERIENCE AND QUALITY OF PERCEPTION

Recently, research and industry have been shifting towards encompassing the end user as the most prominent factor in the multimedia quality assessment to attain broader aspects, such as Quality of Experience (QoE) or Quality of Perception (QoP) rather than only Quality of Service (QoS). This section discusses the underlying concepts of QoS, QoP, and QoE.

A. QUALITY OF SERVICE (QoS)

QoS is often used to express the performance level of multimedia applications and networks. The QoS de-facto definition generally used in the literature based on physical and measurable performance factors of networks including delivery platforms is "a collection of networking technologies and measurement tools that allow for the network to guarantee delivering predictable results [13]". The term QoS is usually utilized with two different meanings. First, it refers to the concepts and measures of network performance (e.g., jitter, delay). Second, it refers to mechanisms such as Integrated Services. Several characteristics, such as performance, responsiveness, availability, adaptivity, dependability, security and application aspects are involved to form the QoS. Due to heterogeneities of the applications, QoS has been explained diversely in independent publications. In this section, we aim to systematically present the QoS taxonomy, influencing factors and performance aspects. Considering multimedia end-to-end architecture, QoS can be divided into three layers: user, application, and resource.

1) USER-LAYER

A user-layer QoS specification is required, so that at the start a user can specify, at abstract level, the QoS requirements, e.g., frame and sampling rates, resolution, cost, and security criteria, perhaps using a GUI. At the end, he/she can provide perceived QoS parameters, such as multimedia content detail, resolution, etc.

2) APPLICATION-LAYER

Once the users have specified their requirements, the next stage is to translate and map those requested QoS to lower layer parameters. This layer is known as application-layer that normally makes no assumption regarding operating systems and network conditions, thus is hardware and platform independent. There are two types of features that are used at the application-layer, i.e., performance-specific (quantitative parameters, e.g., resolution) and behavior-specific (qualitative parameters, e.g., how to manage the service in case of any network bandwidth issue). A specification language is used to provide definite notions to system designer to avoid misconception and time consumption.

3) RESOURCE-LAYER

QoS requirements are specified in a high-level abstract manner, which are then translated into more concrete resource demands, i.e., description of physical resources needed for the application including their allocation, mechanisms and transport protocols. The resource-layer specifications can be classified into coarse granularity and fine granularity categories. The coarse granularity expect a meta-level specification, where generally resource-layer QoS specifications only specify resource requirements without allocation time or detailing resource instances. The fine granularity expects concrete descriptions of required resources, which include explicit narration of quantitative and qualitative QoS requirements, allocation time and adaptation rules.

As also discussed in Section III, QoS is influenced by system as well as user factors. Thus, QoS performances can be evaluated at the system and the user side during the quality formation process. At the system side, the performance can be quantified in terms of input performance (i.e., accuracy of biometrics/emotion/behavior recognizers), input modality appropriateness (i.e., theoretical knowledge of modality properties and its aptness according to environment), interpretation performance (i.e., accuracy of underlying semantic concepts), dialogue management performance (i.e., counting of dialogue success rate), contextual appropriateness (i.e., quantification of Grice's Cooperativity Principle), output modality appropriateness (i.e., interrelations between modalities) and form appropriateness (i.e., the output provided to the user which can be measured via its intelligibility, comprehensibility, etc.). At the user side, the interaction performance can be quantified by efforts (i.e., perceptual, cognitive and physical) required from the user and freedom of interaction.

B. QUALITY OF PERCEPTION (QoP)

QoS describes technical quality of system but neglects the fidelity and utility aspect from users. Thus, to address this limitation, Ghinea and Thomas [181] introduced the notion of Quality of Perception (QoP) and defined it as ''QoP is a term which encompasses not only a user's satisfaction with the quality of multimedia presentations, but also his/her ability to analyze, synthesise and assimilate the informational content of multimedia displays''. Defining multimedia quality using only either subjective or objective factors is insufficient because of multidimensional nature of multimedia, therefore QoP combines both subjective evaluation based on first part of the definition, i.e., user's satisfaction with the quality of multimedia presentations (denoted as QoP-S), and objective one based on second part of the definition, i.e., user's ability to analyze, synthesize and assimilate the informational content of multimedia (denoted by QoP–IA). QoP-S is made of two components, i.e., QoP–LOE (user's level of enjoyment while experiencing multimedia content) and QoP–LOQ (user's judgement concerning the objective level of quality assigned to the multimedia content being experienced). Specifically, QoP–IA usually expressed as a percentage measure to reflect a user's level of information assimilated from experienced multimedia content. While, QoP–LOE and QoE–LOQ are obtained by users' traditional rating methods. Authors in [182] investigated effect of varying multimedia presentation frame rates on user's QoP and eye paths. The presented results show that higher frame rates normally do not lead to higher QoP or level of participant information assimilation, besides not influencing median coordinate value of eye path either. But, it does enhance overall user enjoyment and quality perception. Apteker *et al.* [183] studied video at varying bandwidths and frame rates to determine user QoP termed as 'user watchability'. In this work, it was explicitly found that content of video and fidelity remarkably impact QoP.

C. QUALITY OF EXPERIENCE (QoE)

User satisfaction and perception are shaped by various other aspects, which may/may not necessarily be regulated by the performance of specific service components. Therefore, recently the term Quality of Experience (QoE) has been introduced to describe how a user perceives the usability, acceptability and satisfaction of the service [4]. QoE goes beyond conventional end-to-end QoS integrity parameters to cover a multitude of different aspects (e.g., user's mental state) to improve the experienced quality by the user. Namely, QoE is the perceptual QoS from perspective of the users. In [13], QoE is stated as ''the degree of delight or annoyance of a person whose experiencing involves an application, service, or system. It results from the person's evaluation of the fulfilment of his or her expectations and needs with respect to the utility and/or enjoyment in the light of the person's context, personality and current state''.

QoE is determined by psychological as well as cognitive determinants, e.g., habits, feelings, requirements and expectations. It is paramount to obtain quantified QoE by translating system's performance together with users' perception in the form of statistical and interpretable values. The quantified QoE can be obtained employing either 'direct QoE measurements' (i.e., rating done by real subjects; also called subjective QoE) or 'indirect QoE measurements' (i.e., logging user behavior and relating it to perceived QoE; also called objective QoE). In the latter category, use of physiological measures have been recently investigated in several studies [167].

1) SUBJECTIVE QoE ASSESSMENT METHODS

Since human consumers are the ultimate judges for any multimedia content/service, for optimization and analysis subjective QoE assessment methods are usually carried out by surveying, interviewing, and statistical sampling of users/customers for their perceptions, requirements, and quality. Broadly speaking, subjective QoE studies can be labeled as qualitative or quantitative techniques. The qualitative techniques capture human perceptions, feelings and opinions through verbal behaviors, e.g., comments on blogs. The quantitative techniques capture human perceptions, feelings and intentions through numbers and statistics.

2) OBJECTIVE QoE ASSESSMENT METHODS

Objective QoE assessment methods are grouped into QoS (technology) centric or human cognitive (physiological) centric techniques. As a former group's approach, most of the time perceptual-based (objective) quality assessment methods for audio, video and audiovisual signals discussed below in Sections V-A, V-B, and V-C are applied to quantify the QoE. Anyway, Skowronek and Raake [168] specifically investigated the relationship between number of interlocutors, cognitive effort and perceived QoE of multimedia conferencing and telemeetings. They found that better technical solutions causes less cognitive efforts and better QoE. Adaptive video streaming protocols have been proposed in [169] to achieve better QoE over multimedia wireless networks that schedules video chunks and their qualities at given time. Wang and Dey [170] devised a mobile gaming user experience (MGUE) model to quantify user's QoE using cloud mobile gaming (CMG). Other studies attempted to identify the relationship between QoE and QoS. For instance, an expression to capture the exponential relation between the QoE and QoS parameters was proposed in [171]. Particularly, QoE is expressed as a function of loss and reordering ratio caused by jitter, and considered that the change of QoE is based on the current level of QoE such that same amount of change in QoS value happens with different sign, as shown in (Eq. 2):

$$\frac{\partial QoE}{\partial QoS} \sim (QoE - \gamma). \tag{2}$$

Similarly, Shaikh *et al.* [172] defined a linear relationship between the QoE and multiple QoS parameters such as bandwidth, throughput and delay on the QoE as:

$$\log(QoE) = a_0 + a_1 QoS_1 + a_2 QoS_2 + \ldots + a_n QoS_n. \tag{3}$$

Finally, the QoE/QoS exponential correlation was modelled by applying an exponential transformation on (Eq. 3) as:

$$QoE = e^{a_0} + e^{a_1 QoS_1 + a_2 QoS_2 + \ldots + a_n QoS_n}, \tag{4}$$

where constants a_i were estimated by the least squares method. Alberti *et al.* [173], in turn, defined the relationship between QoE and QoS parameters as non-linear by the following expression:

$$QoE = \sum_{i=0}^{N-1} a_i QoS_i^{k_i}, \tag{5}$$

where a_i are the constants and k_i are the exponents for N parameters.

TABLE 3. Comparison of different neuroimaging technologies.

Method	Neuronal Activity	Hemoglobin Dynamics	Time Resolution	Spatial Resolution	Subjects Mobility
EEG	Yes	No	ms	cm	Yes
MEG	Yes	No	ms	cm	Limited
fMRI	Yes	Yes	s	1mm	Limited
NIRS	Yes	Yes	ms	10mm	Yes

The approaches in the cognitive centric group try to use neurophysiological insight for perceived QoE through human body area sensors and networks using techniques, such as electroencephalography (EEG), magnetoencephalography (MEG), functional magnetic resonance imaging (fMRI), and near-infrared spectroscopy (NIRS). The EEG and MEG data provide high time resolution, while fMRI and NIRS provide good spatial resolution but poor temporal resolution [174], [175], as presented in Table 3. Although each physiological/cognitive centric technique has its strengths and weaknesses, they provide precise quantitative information about human behavior and perceived QoE. Thus, studies on fusion of cognitive centric and existing quality assessment methods have received a recent spurt by the research community. For instance, it was reported in [177] that NIRS and physiological biosignal sensors may be used to characterize subjective image preferences with up to that 72% accuracy. Arndt *et al.* [178] and Moldovan *et al.* [179] utilized EEG to correlate perceived quality of videos with varying properties. Likewise, [167], [175], [176], [180] investigated user's EEG signals to characterize speech/audio QoE. Particularly, the study in [176] showed that measuring human affective states is important for objective measurement of perceived QoE. The studies [167], [175] concluded that speech quality is inversely proportional to EEG feature (perceived QoE).

Immersive 360-degree virtual reality (VR360) applications are burgeoning and users interact with virtual elements in 3D environments created by VR techniques. Particular devices, e.g., head-mounted displays, stimulate 3D sight, hearing and touch. Usually, in VR360 the simulated environment is built by real-time dynamic 3D stereo/Binocular and binaural rendering. Up to some extent, VR QoE may be defined as a compelling and immersive experience, which does not drive the user sick. Traditional objective QoE methods can not be applied directly for VR QoE. Recently, few works have focused on VR360 QoE assessment. For instance, Zhou *et al.* [110] devised a stereoscopic images quality assessment method based on disparity map, which can be used not only for three dimensional multimedia systems but also for 3D image/video broadcasting. In turn, Rozenn *et al.* [198] studied how to evaluate QoE of 3D

TABLE 4. A representative list of audio quality assessment algorithms. CC: Correlation Coefficient; RMS: Root Mean Squared Error; AES: Absolute Error Score.

Category	Method	Feature description	Database	Figure of merit	Year
Intrusive methods	PSQM [52]	Time synchronized spectral power densities on frames	CCITT LD-CELP	CC	1994
	PEAQ [36]	Fast Fourier transform and filter bank-based models for masking	MPEG90, MPEG91, ITU92DI, ITU92CO, ITU93, MPEG95, EIA95, CRC97	CC, AES	2001
	PESQ [57]	Perceptual frequencies and compressive loudness scaling	ITU-T P-series	CC	2001
	POLQA [59]	Disturbance density (additive distortions and subtracted distortions)	NB_BT_P862_BGN_ENG, WB_GIPS_EXP3, SWB_48kHz303_OPTICOM	CC	2011
	ViSQOL [46]	Spectro-temporal short-term fourier transform spectrogram	IEEE Harvard Speech Corpus	CC	2012
	AutoMOS [60]	Recurrent long short-term memory cells	Corpus of Google's TTS engines	CC, RMS	2016
Nonintrusive methods	PLP [66]	Perceptual linear prediction coefficients and vector quantization	Bell Lab	CC	1995
	ANIQUE+ [200]	Frame, mute and non-speech distortions	Private datasets	CC	2005
	POSQE [69]	Vector quantization and self-organizing map	Nortel Networks	CC	2010
	HASQI [74]	Linear and nonlinear measurements of envelope and temporal fine-structure modifications	Private dataset	CC	2010
	SRMR [70]	Auditory-inspired modulation filterbank analysis	IEEE sentence corpu	CC	2014
	PREQUEL [77]	Acoustic output and binaural recordings with a head and torso simulator	Private dataset	CC	2016

audio binaural rendering. Perrin *et al.* [199] predicted sense of presence as a variant of QoE in immersive audiovisual communications by using also physiological signals (i.e., EEG, ECG (electrocardiography), and respiration). Likewise, [200] evaluated heart rate and electrodermal activity as an objective QoE parameter for immersive VR environments. Besides VR360, in the past few years, high dynamic range (HDR) and high frame rate (HFR) applications have also emerged and their QoE assessment has turned into emerging research topics. Representative examples of works on HDR/HFR quality assessment include those reported in [13], [145], and [191], where authors investigated subjective quality assessment experiment on videos compressed at different frame rates, quantization levels and spatial resolutions. The progress on perceptual QoE of VR, HDR and HFR remains limited, however, thereby making it difficult to assess the exact gain by switching from 2D to 3D or from low to high frame rates.

V. AUDIO, VIDEO AND AUDIOVISUAL MULTIMEDIA QUALITY: EXISTING ASSESSMENT METHODS AND METRICS

In this section, a comprehensive overview of perceptual (objective) quality assessment methods for audio, video and audiovisual multimedia signals are presented. Each signal type (i.e., audio, video, and audiovisual) is addressed in a different subsection (i.e., Sections V-A, V-B, and V-C). This way, the reader can gain a more clear perspective of the current panorama in the field of multimedia perceptual quality assessment.

A. STATE-OF-THE-ART IN AUDIO QUALITY ASSESSMENT

Sound can generally be categorized into two groups as high-fidelity audio (i.e., all kinds of sound) and speech (i.e., language content). It is of fundamental importance to measure sound quality[1] in several applications to meet human user's quality expectations and feelings. Aside from the widely used ACR scale, another popular subjective method is the double blind Multi Stimulus with Hidden Anchor (MUSHRA) [31], adopted as a ITU-recommendation (ITU-R BS.1534) [9]. In turn, audio objective quality assessment algorithms can be broadly classified into three classes: *intrusive* (also known as full-reference, comparison-based, or input-to-output), *nonintrusive* (also known as no-reference, output-based or single-ended) and *parametric* (also known as planning or glass box) methods. A brief description of representative audio quality assessment methods is presented in Table 4.

1) INTRUSIVE METHODS

Intrusive models compare an original signal with a degraded signal under test. The published works on intrusive methods can be further sub-classified as psychoacoustic and cognitive/perceptual models.

a: PSYCHOACOUSTIC MODELS

According to the domain transformation utilized, psychoacoustic models are grouped into two clusters: time domain and spectral domain measures.

a.1: TIME DOMAIN MEASURES

Time domain analysis is useful mostly for analog or waveform coding systems where target is to reproduce the waveform. The signal-to-noise ratio (SNR) and total harmonic distortion (THD) [11] are well-known examples of time domain measures in which signals are time aligned to

[1]In this article, we use the terms audio quality and sound quality interchangeably, unless explicitly stated otherwise.

compute the noise and corresponding quality. Different variants of SNR measures have been presented in the literature, e.g., segmental SNR (SNR measurement over short periods), frequency weighted segmental SNR (different weights for different frequency bands), granular segmental SNR (for granular noise), noise-to-masked ratio (i.e., level difference between masked threshold and noise signal), signal-to-interference ratio (SIR), signal-to-distortion ratio (SDR), and signal-to-artifact ratio (SAR) [9]. Though SNR measures are good estimators for waveform codecs audio quality, they are poor estimator of subjective audio quality especially under a larger range of distortions [14].

a.2: SPECTRAL DOMAIN MEASURES

Spectral domain measures are more practical since they are less sensitive to time misalignments and phase shifts in the signals. In recent years, several spectral domain based audio quality evaluation schemes have been proposed, e.g., psychoacoustic model of PEAQ (perceptual evaluation of audio quality; ITU standard for audio quality (BS.1387) [32]. Specifically, PEAQ transforms the time domain signals into a frequency basilar membrane representation via Fast Fourier Transform (FFT) to model outer and inner ear, and/or filter bank-based models to model human ear with backward masking to obtain perceived quality estimation. A novel method that models sound pressure levels and tracks temporal maskers frame to frame with boundary detection was presented in [33]. Huber and Kollmeier [34] proposed a technique named PEMO-Q that maps the internal ear representation via psychoacoustically validated model of auditory processing for internal ear representation. The reported results showed better accuracy than PEAQ for a wide range of distortions except linearly distorted signals. Other notable spectral domain audio quality assessments works are LLR (log likelihood ratio) based on speech production models [35], Itakura-Saito distortion measure (i.e., a variant of LLR) [14], cepstral distance based on linear prediction coefficients [36], DIX (disturbance index) based on temporal resolution analysis using filter bank [37], NCM (Normalized Covariance Metric based on covariance between auditory-inspired envelopes of the clean and processed signals) [38], STOI (Short-Time Objective Intelligibility like NCM but over short time frames including both signals are time-aligned) [39], MSSIM (mean structural similarity of spectrogram of frequency) [40], NSIM (Neurogram Similarity Index Measure based on responses from auditory nerves) [41], ViSQOL (Virtual Speech Quality Objective Listener based on spectro-temporal–Short-term Fourier Transform (STFT) spectrogram–measure to account for human sensitivity to degradations in speech quality) [42], and VISQOLAudio (an extension of ViSQOL to increase the hearing frequency bandwidth) [43]. Generally speaking, spectral domain measures are mostly related to speech codecs design and speech production models, thus their performance is limited by the constraints of the speech production models as well as models' failure.

b: COGNITIVE/PERCEPTUAL MODELS

The work in [44] is one of the very first attempts to devise a perceptual-based audio quality assessment. The proposed method is based on auditory spectrum distance (ASD) model that compares time frequency and loudness representation of both reference and test signals. Numerous novel cognitive quality assessment methods inspired by the work in [44] have been proposed in the literature. For instance, bark spectral distortion (BSD) technique in [45], which models frequency scale warping using bark transformation, besides other features of audio perceptual processing, e.g., ear sensitivity, loudness level and band integration in the cochlea. The method in [45] has been extended in [46], named as Modified BSD measure, which incorporates noise-masking threshold and difference and normalization of loudness. Beerends and Stemerdink [47] devised a scheme named perceptual speech quality measurement (PSQM) that analyzes temporal and continuous distortions and spectral power densities in both signals. Latter, PSQM was approved and recommended by ITU-T P.861. However, PSQM did not account for temporal masking effects and impacts caused by packet loss or other time clipping effects. PSQM was extended in [48] to address its limitations; the extended method was named PSQM+. It was empirically concluded in [49] that listeners adapt and respond differently to spectral deviations spanning different time and frequency scale, which was adopted in [50] to annex PSQM. The annexed method, also called measuring normalizing blocks (MNB) model, measures the perceptual distance between the signals across multiple time and frequency scales. A logistic regression is employed to compute the final perceived audio quality using time- and frequency-measuring blocks. Rix and Hollier [51] proposed a algorithm called perceptual analysis measurement system (PAMS) that estimate the perceived audio clarity of an output signal as compared with the input signal. Though PMS is quite similar to PSQM, it utilizes different signal processing techniques as well as different perceptual model. The training process of PAMS is computationally expensive, since it is not easy to optimize the model parameters and mapping function.

Beerends *et al.* [47] improved the traditional PSQM to better correlate the subjective MOS. The improved version was dubbed PSQM99. A new measure that integrates the robust time-alignment techniques of PAMS and the accurate perceptual modelling of PSQM99 was approved by ITU-T under recommendation P.862 as perceptual evaluation of speech quality (PESQ) [52]. PESQ was originally conceptualized to approximate the listening audio quality in wireless, VoIP and fixed networks, and has been widely adopted by many vendors as a standard method. However, it was empirically found in [53] that PESQ performs better mainly for signals processed by modern vocoders compared to the signals with distortions generated by the transmission channel limited to 8 KHz with P.862.3 for 16 KHz. Thus, it is better to use PESQ in conjunction with other methods that consider different parameters as well (e.g., frequency response, loudness ratings) [5]. Thus, POLQA (Perceptual Objective Listening

Quality Assessment) [54] was introduced by ITU-T to predict overall speech quality in narrowband (300–3400 Hz), wideband (50–7000 Hz) and super-wide band (50–14000 Hz) and their speech processing components. More recently, assessing perceived quality of synthesized audio (speech) [55], multichannel and automotive audio [56] and blind source separation [57] has become areas of growing interest. All in all, several studies found that cognitive models are very data-dependent and perform poorly under strong time-wrapping distortions and Enhanced Variable Rate Codecs (EVRC).

2) NONINTRUSIVE METHODS

Although intrusive methods are more accurate, they normally are unsuitable for real-time applications, besides requiring difficult synchronization between the reference and processed signals. Objective audio quality assessment methods that estimate the audio quality using only the test (or degraded) signal are known as nonintrusive methods. Nonintrusive techniques can be divided into two classes: a priori-based and source-based approaches.

a: A PRIORI BASED APPROACHES

A priori based approaches first learn a set of well-characterized distortions and then establish a statistical relationship between this set and subjective opinions. For instance, the technique in [58] measured output-based speech quality for wireless communication systems by analyzing visual features of the spectrogram of audio signal. The method computes variance and dynamic range in a block-wise manner, and then averages all the blocks to yield final quality score. Gray *et al.* [60] proposed a novel use of the vocal-tract modelling technique that can be employed for nonintrusive quality assessment of speech stream over networks. The reported results showed efficacy of the technique, but also the sensitivity to speaker gender. In turn, an auditory non-intrusive quality estimation (ANIQUE) model [195] was formulated using temporal envelope representation of speech motivated by functional roles of human auditory system both at peripheral and central levels to be later mapped to a final quality score by artificial neural network (ANN). Since ANIQUE's accuracy is inversely proportional to speech naturalness, ANIQUE+ has been devised to overcome the limitation.

b: SOURCE BASED APPROACHES

The source-based approaches can be considered as more universal methods, since they make a priori assumptions about expected clean signal properties rather than the distortions that may occur; this way they can deal with ample range of distortion types. One of the initial attempts to develop source-based audio quality assessment algorithms is [61], where the model compared the variety of clean with distorted audio signals using perceptual-linear prediction (PLP). However, the method is computationally expensive because is based on Vector Quantizers (VQ) technique, and its generalized capability is inferior. To overcome some of these drawbacks, Falk and Chan [62] replaced VQs by Gaussian mixture models (GMMs) and proposed a consistency measure to estimate quality. Improved results were later achieved once clean and degraded GMMs were utilized [62]. A perception-based quality evaluation is presented in [63] that computes objective distances between perceptually-based parametric vectors representing degraded speech signal to appropriately matching reference vectors extracted from a pre-formulated clean reference codebook. Similarly, POSQE (Perceptual Output-based Speech Quality Evaluation) based on vector quantization and self-organizing map was devised in [64]. Falk *et al.* [65] devised SRMR (speech-to-reverberation-modulation energy ratio) normalized metric based on auditory-inspired modulation filterbank analysis of temporal envelopes of the speech and pitch signals. Also, few models have been developed to predict the audio quality ratings by hearing impaired listeners [66]–[68]. For instance, the Hearing-Aid Speech Quality Index (HASQI) in [69] takes into account the effect of noise, nonlinear distortion and linear filtering for the perceived speech quality; however it is very senstive to loudness pattern distortion. In turn, Beerends *et al.* [72] have presented the PREQUEL (Perceptual Reproduction Quality Evaluation for Loudspeakers) that simulates the binaural recordings of the reference signals using head and torso simulator to quantify the overall loudspeakers' perceived sound quality by assessing their acoustic output. In recent years, developing hybrid methods, e.g., [70], [71], [73], that combine properties of both a priori- and source-based techniques is also gaining momentum.

3) PARAMETRIC METHODS

Parametric models estimate the quality using specifications of network design process and/or parameters, such as echo, delay, frequency-weighted insertion loss (so-called ''loudness rating'') and packet loss. Most of these specifications can be accurately modelled by a small number of statistical measures. A well-known example of parametric approach is ITU recommendation P.563 that utilizes in-service, nonintrusive measurement devices (INMD) [59]. An INMD evaluates objective parameters of voice channels on live call traffic without hindering the call, and with knowledge of network and human auditory system produces quality values. Such quality estimates are only applicable for transmission planning purposes but not for actual customer opinion prediction. To address this drawback, there exist one more ITU-T recommended computational model known as the E-model [22] that can be used in conjunction with INMD by transmission planners to estimate the quality and users' satisfaction. Though proven to be proficient for network related perceptual effects, E-model becomes less precise with modern terminal equipments (e.g., handsets involving noise reduction) because of several simplifying assumptions (e.g., linearity and order independence).

B. STATE-OF-THE-ART IN VIDEO QUALITY ASSESSMENT

Over the years, a large number of video objective quality models has been proposed and different international

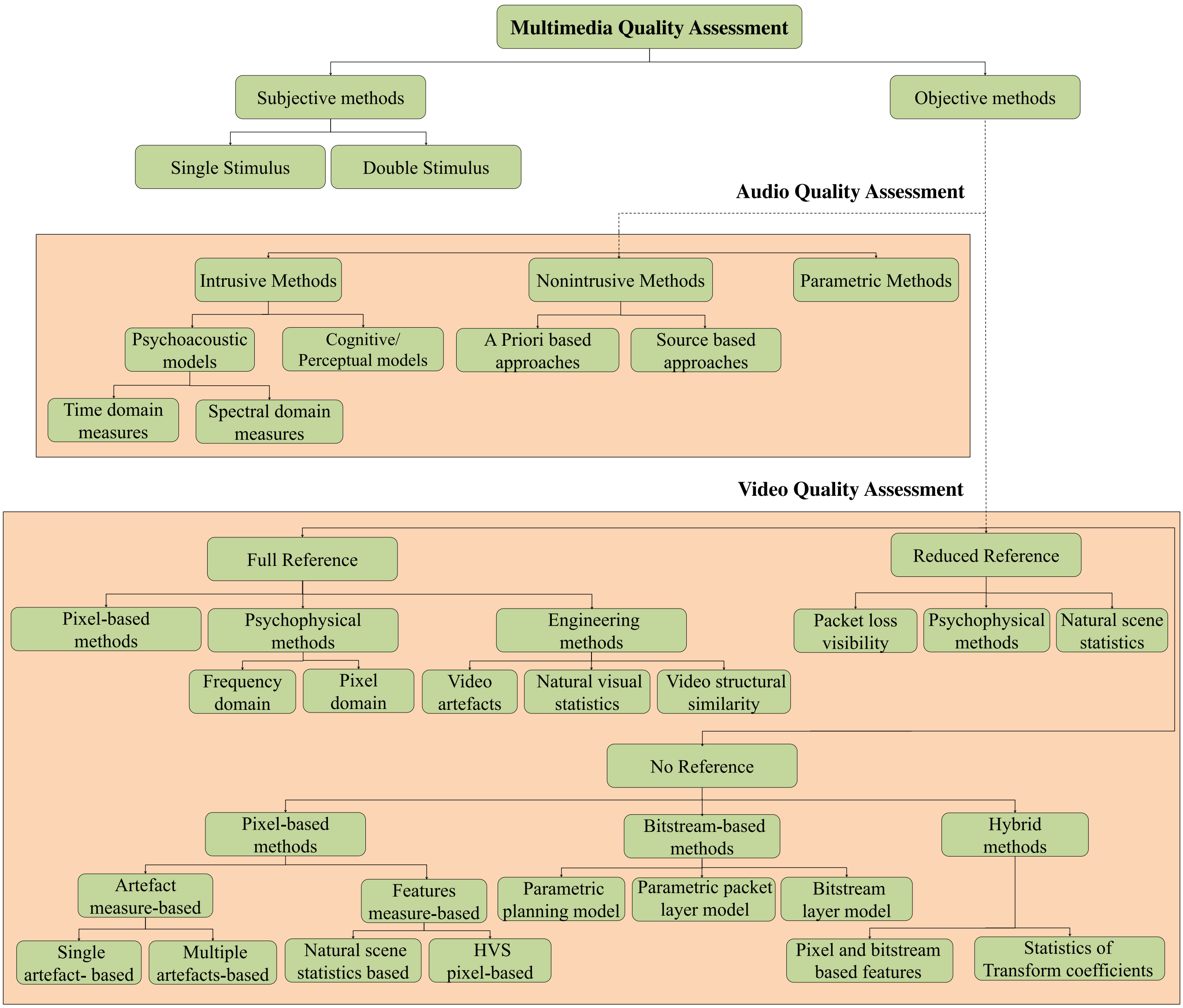

FIGURE 5. Classification of multimedia quality assessment methods.

organizations have tried to standardize video quality evaluation metrics. The objective video quality assessment (OVQA) methods may be considered as a two-stage process composed of feature detection and feature pooling into a final quality score. OVQA metrics are typically rooted on either vision based modelling or signal-driven approaches. The former exploits relevant psychophysical properties and physiological knowledge (thus also known as 'psychophysical approaches'), while the latter uses signal extraction and analysis (thus also referred as 'engineering approach' or 'natural visual characteristics based approach'). From Fig. 5, it can be seen that objective methods can be FR, RR, and NR. Usually FR is based on psychophysical approaches, while RR and NR belong to engineering approach. In the following subsections, these three main categories (and subcategories) are described and a brief summary is presented in Table 5.

1) FULL-REFERENCE METRICS

These video quality metrics/algorithms can be coarsely classified as *pixel-based*, *psychophysical*, and *engineering* methods.

a: PIXEL-BASED METHODS

Methods in this category are also referred to as 'data metrics'. Two widely used pixel-based techniques are Mean Squared Error (MSE) and Peak-Signal-to-Noise Ratio (PSNR). The former measures the video (frame/image) difference to denote the power of the distortion, while the latter measures fidelity to denote the resemblance between two samples. Though pixel-based metrics are simple and computationally inexpensive, they correlate poorly with perceived quality, as neither features of HVS nor video content or viewing

TABLE 5. A representative list of video quality assessment algorithms. CC: Correlation Coefficient; RMS: Root Mean Squared Error; SRCC: Spearman Rank-order Correlation Coefficient; PSNR: Peak signal-to-noise ratio; CART: Classification And Regression Tree; Generalized Linear Model (GLM).

Category	Method	Feature description	Database	Figure of merit	Year
Full-Reference	PVQM [86]	Luminance edginess, color error and temporal decorrelation	Database with digital codec, analog PAL, VHS and Betacam distortions	CC	2002
	Multi-scale SSIM [99]	Luminance, contrast, structure, image details at different resolutions (Multi-scale)	LIVE JPEG/JPEG2000	CC, RMS, OR, SRCC	2003
	Video-VIF [96]	Natural scene statistics and video motion	VQEG Phase I FR-TV	CC, SRCC	2005
	MOVIE [93]	Sapatial, temporal and spatio-temporal distortions	VQEG Phase I FR-TV	CC, OR, SRCC	2010
	AFViQ [85]	Contrast sensitivity, foveated vision, visual attention	LIVE and VQEG HDTV	CC, RMS	2013
	STME [94]	Correlation between spatio-temporal motion energies	LIVE VQA	CC	2016
Reduced-Reference	PQSM [113]	Visual saliency, attention, and eye movement	Videos with framerate at 25/30Hz	PSNR	2003
	Packet loss visibility in MPEG-2 [104]	Content-independent and content-dependent factors	Private database of packet losses based MPEG-2 videos	CART GLM	2006
	Packet loss visibility in H.264 [108]	SSIM, camera motion and proximity to a scene change	H.264 videos with 352×240 resolutions	PRIM	2007
	RR-GGD [119]	Discrete cosine coefficients and mutual information	LIVE, MICT, CSIQ	CC, SRCC, RMS	2013
	RR-VQA [116]	Motion in stereo videos and binocular perception characteristics	NAMA3DS1-COSPAD	CC, SRCC, RMS	2016
No-Reference	Hybrid [154]	Packet lengths, motion intensity and luminance discontinuity	SDTV	CC, RMS	2007
	LBM [129]	Blocking artefacts and properties of HSV	LIVE	CC, SRCC	2008
	T-V-model [146]	Coding bit-rate and packet loss percentage	Private dataset	CC	2008
	ANFIS [144]	Video content based encoding and transmission parameters	Private encoded dataset	CC, RMS	2009
	Video quality without decoding[148]	Size of frames and motion in video	H.264 videos of size 1440×1080	CC, RMS CC, RMS	2010
	SACONVA [142]	3D shearlet transform and convolutional neural network (CNN)	LIVE, IVPL, CSIQ	CC, SRCC	2016

conditions are taken into account [8]. Engelke *et al.* [201] designed a temporal trajectory aware video quality measure (TetraVQM) by combining PSNR and a simple saliency model.

b: PSYCHOPHYSICAL METHODS

FR psychophysical methods are modelled based on HVS characteristics related to visual perception, such as contrast sensitivity, colors perception, masking effects, spatial and temporal features and frequency selectivity [74]. Most psychophysical methods construct a sensitivity or response computational model of the HVS as a function of stimulus. In other words, in these approaches perceptual attributes motivated from computational models of low-level vision are computed to produce a reduced description of the video to be used latter to quantify effects of distortions and content on perceived quality. Psychophysical approaches can further be divided into frequency domain and pixel domain.

b.1: FREQUENCY DOMAIN

The quality is determined by measuring impairments in different frequency regions using transforms such as wavelets, Gabor filters, Fourier, DCT (Discrete Cosine Transform), etc. One of the pioneering video quality metrics based on HVS was devised by Lukas and Budrikis [75]. The developed model is composed of two stages: a nonlinear spatio-temporal model of a visual filter, and a masking function. The masking function describes the spatial and temporal activity by point-by-point weighting of the filtered error for non-uniform backgrounds, while the former stage describes the threshold attributes on uniform backgrounds. The error averaged over the video frames is finally used as a perceived (predicted) quality. Lambrecht *et al.* [76] developed MPQM (Moving Picture Quality Metric) to simulate spatio-temporal model of HVS with a filter bank technique. The MPQM is particularly based on two characteristics of human perception, i.e., contrast sensitivity and masking effect, since eye's sensitivity varies as a function of spatial frequency, orientation and temporal frequency, while perception of a stimulus is a function of its background. The authors in [77] proposed Digital Video Quality (DVQ) model to calculate visual difference between reference and distorted videos using DCT. The model incorporates contrast masking, spatial and temporal filtering, aspects of luminance and chromatic channels, probability summation, and spatial frequency channels to assess quality. After pre-processing, the video sequences are then processed with block DCT of size (8 × 8 pixels) to estimate local contrast and just-noticeable differences for visual quality of the sequence. The reported experiments concluded that proposed metric was not a good fit for low bit rate videos. Xiao [78] extended the DVQ model to incorporate another human eye's characteristic, i.e., spatio-temporal patterns sensitivity to eyes is inversely proportional to spatial and temporal frequencies. A wavelet transform based method was devised in [79] that employs multi-level and 3D wavelet transform to compute spatial and temporal degradations. A novel metric to model advanced contrast sensitivity of the HVS based on the mechanisms of vision foveation[2] and visual attention named Attention-driven Foveated Video Quality metric (AFViQ) was proposed in [80]. In particular, AFViQ simulates dynamic foveation by estimating video fixation using eye movement leading to a wavelet-based distortion visibility quality measure. In order to provide empirical efficacy of AFViQ, authors evaluated it with different attention/saliency maps obtained from the graph-based visual saliency (GBVS), video spatial-temporal saliency, and a video attention models.

b.2: PIXEL DOMAIN

An objective video quality model exploiting the HVS feature of sensitivity to edges and local changes in luminance was developed in [81]. The model is known as Perceptual Video Quality Metric (PVQM), which is also called the Swisscom/KPN metric. The perceptual quality is predicted by a linear combination of three distortion indicators (i.e., edginess, temporal decorrelation, and color error). The edginess, temporal decorrelation, and color error account for loss or introduction of sharpness, perceived spatial distortion, and temporal variability causing error, respectively. Another video perpetual quality metric was proposed in [82] and [83] that uses distortion-invisibility, blockiness, and content fidelity factor. The method was modified in [84] to use a Sobel filter to approximate the gradient of local luminance to attain improved performance. Chandler and Hemami [85] devised visual signal-to-noise ratio (VSNR) metric by detecting perceptual distortions via visual masking and visual summation. Opticom introduced a video quality metric called Perceptual Evaluation of Video Quality (PEVQ) [26] based on PVQM model. Specifically, PEVQ utilizes gradient filter, and computes spatial distortion measures (i.e., edginess in luminance, edginess in chrominance, temporal variability indicators) and a temporal distortion measure (i.e., absolute difference between current and previous frame).

c: ENGINEERING METHODS

Methods in engineering approach are based on visual statistical features (e.g., covariance of certain distortion patterns) and visual features (e.g., blockiness), thus also called as 'natural visual characteristics' based methods. Published FR video quality engineering approaches can be broadly subgrouped into three categories: video artefacts, natural visual statistics, and video structural similarity.

[2]The HVS discerns different volume of detail/resolution across the area of view, with highest resolution at the point of fixation. The point of fixation is projected onto the center of the eye's retina called fovea [18].

c.1: VIDEO ARTEFACTS

Since FR pixel-based quality metrics are unsuited for videos encoded at a low bitrate, the Low-bitrate Video Quality Model (LVQM) [86] was developed. LVQM incorporates three aspects, namely distortion-invisibility (based on luminance, spatial-textural and temporal masking), block fidelity (since low bit compression introduces block boundaries distortion) and content richness fidelity (based on luminance occurrences). Lee and Sim [87] developed a metric to indicate the visual degradation in digital mobile videos, which is calculated as the weighted sum of three factors: block edginess, blockiness and blurriness. In [88] a video quality metric called MOVIE (MOtion-based Video Integrity Evaluation) index was formulated using Gabor filter banks to emulate the middle temporal (MT) visual area of the visual cortex in the human brain, since the MT visual area is known to be critical for the perception of video quality. The MOVIE index evaluates distortions both individually in space and time domains as well as in the space-time domain to specify the motion quality and trajectories. Likewise, perceiving motion based on spatiotemporal energy is exploited in [89] for video quality prediction.

c.2: NATURAL VISUAL STATISTICS

Videos are natural scenes having different statistical information than random signals. Nonetheless, video compression artefacts precipitate unnaturalness in the samples. The statistical information differences between original and compressed videos can be quantified by combining Natural Scene Statistics [90] and distortion models. Towards this aim, the well-known model called Video Visual Information Fidelity (V-VIF) [91] was designed. The VVIF basically combines visual statistics with HVS modelling using Gaussian Scale Mixtures and mutual information.

c.3: VIDEO STRUCTURAL SIMILARITY

The methods in this genre aim to estimate the similarity (fidelity) between original and distorted videos by top-down techniques to model functionality of the overall HSV. The Video Structural Similarity (VSSIM) index [92], [93] exploits the fact that HSV is distinctly developed to capture the structure of the video and thereby utilizes structural distortions as a source to estimate perceptual distortions. In particular, SSIM (Structural Similarity Index Metric) computes the 'difference of structure' between the original and the distorted videos via analysis of luminance, contrast and structure at the local region, frame, and sequence levels. Several versions of SSIM, such as Multi-scale SSIM [94], Spatial weighted SSIM [95], Speed Weighted SSIM [96], Visual fixation weighted SSIM [97], quality weighted SSIM [97], have been also proposed in the literature to incorporate sampling density, viewer's distance, fidelity of spatial information, motion speed, etc. in the process. Tao [98] employed matrix singular value decomposition (M-SVD) to compute the underlying video structure and consequent quality measure.

2) REDUCED-REFERENCE (RR) METRICS

RR video quality methods extract the most characteristic features from the reference video, and perceived quality is then estimated by comparing those features in video under test. RR video metrics can be coarsely classified as *packet loss visibility*, *psychophysical* and *natural scene statistics* based techniques.

a: PACKET LOSS VISIBILITY BASED METHODS

In [99] and [100], tree-structured data analysis based on Classification And Regression Tree (CART), and Generalized Linear Model (GLM), respectively, is conducted to classify whether packet loss is visible or invisible. In [102] and [101], multiple packet loss and H.264 considering the frames in which packet loss occurs, the magnitude and angle of the motion were studied, while in [103] and [100], the visibility of packet loss via SSIM, and Patient Rule Induction Method (PRIM) and Group-of-Picture (GoP) are adopted for packet loss classification. Aabed and AlRegib [104] exploited optical flow to evaluate the quality degradations in video streaming service due to coding and network errors.

b: RR PSYCHOPHYSICAL METHODS

The approaches in this group are developed on modelling HVS. For instance, [105] utilized several HSV related features, such as blurriness and blockiness that are distinguished by harmonic amplitude analysis and local harmonic strength values for quality estimation. Similarly, [106] modeled RR quality estimation using contrast sensitivity function of HVS by contourlet transform. The method in [107] combined color perception, psychophysical subband decomposition and masking effect with structural similarity to attain RR metric. Lu *et al.* [108] developed a saliency-weighted RR metric to simulate the quality perception called perceptual quality significance map (PQSM) to be used in estimating the visual distortion. The PQSM is an array and its elements represent relative perceptual-quality significance levels for the corresponding regions for images/video. Particularly, the method in [108] utilizes visual attention, eye fixation/movement, and the path of vision/retina. Since, the selectivity characteristic of HVS (Human Visual System) pays more attention to certain area/regions of visual signal due to certain combination of salient features in video, cues from domain knowledge, and association of other media (e.g., audio). Karacali and Krishnakumar [109] devised a real time RR metric known as Simplified Perceptual Quality Region (SPQR) for video conferencing application that detect face and its discrepancies among frames. A RR quality metric for stereo videos was proposed in [110] and [111], respectively, using view together with disparity zero-watermarks based on gradient vectors, and temporal characteristics of video and binocular perception in HVS.

c: NATURAL SCENE STATISTICS

These algorithms assume that real-world videos are made of natural scenes, thus their statistical features would be deranged by any kind of distortion, which can be utilized to quantify the perceived quality. The standard natural scene statistics (NSS) based RR model called wavelet-domain natural image statistic metric (WNISM) was proposed in [112]. The divisive normalization transform (DNT) was used to overcome the limitations of wavelet transformation in [113]. While, in [114] Tetrolet transform was employed to compute statistical dependencies and quality. Ma *et al.* [115] argued and empirically showed that generalized Gaussian density (GGD) can depict the coefficient distribution in reorganized DCT (RDCT) domain for better RR video quality prediction. The Kullback-Leibler divergence, weighted entropy difference in DCT bands and discrete wavelet transform (DWT) of locally weighted gradient magnitudes were successfully used to estimate the high level perceived quality in [116]–[118], respectively.

3) NO-REFERENCE (NR) METRICS

NR metrics can meet the requirement real-time quality and QoE assessments. But, NR methods are difficult to design since no reference/original video is available during test. Many efforts recently have been placed on development of NR methods. Existing NR techniques can be roughly divided into three groups: *pixel*, *bitstream* and *hybrid* methods.

a: PIXEL-BASED METHODS

Pixel-based NR (P-NR) methods analyze certain artifacts related to a particular type of degradation in video quality. They can be further divided into two subgroups: artefact measure- and features measures-based.

a.1: ARTEFACT MEASURE-BASED METHODS

Artefact measure-based metrics quantify common visual artefacts (e.g., blur, noise) and impairments for perceived video quality. Artefact measure-based can further be classified into two clusters: single artefact and multiple artefacts based P-NR methods.

a.1.1: SINGLE ARTEFACT BASED METHODS

As the name suggests the methods in this category are developed by considering a given model of a single degradation factor, such as blurring, blocking, ringing, noise and frame freeze [119]. The work in [120] quantifies quality in terms of global blur relying on histograms of discrete cosine transform (DCT) coefficients present in MPEG and JPEG encoded data. However, it performs well only for out-of-focus blur but not for uniform background or over-illuminated samples. Contrary to edge blur detection methods, the framework in [121] is to evaluate blur at macroblock boundaries and averaging the block level measure to yield overall quality. In addition, the framework also uses content-sensitive masking. This method is widely used for videos encoded following the H.264/AVC standard. The technique proposed in [122] claimed to be working for any type of blurriness without being sensitive to the source of blur. A gradient image and a Markov model is used to attain the quality prediction. Chen *et al.* [123] claim that their proposed method can be used for any video format. The method is a frequency domain pixel-based bi-directional (horizontal and vertical) measure. Liu *et al.* [124] developed an HVS-based blocking method to gauge quality via a grid detector that discovers blocking locations. The method is computationally inexpensive and the use of visual masking makes it easy to locate blockiness visible only to human perception. Moreover, [125] integrated HVS masking with human visibility index to estimate ringing nuisance and perceived quality to attain performance level comparable to FR methods. Since noise is usually introduced during video acquisition, processing, recording or transmission, the work in [126] uses high-pass directional operators to compute an estimate of average noise variance to be exploited for quality assessment. Moreover, to measure jerkiness (both frame jitter and frame freeze) as a measure to quality assessment of videos with varying resolution from QCIF to HD, a technique using mean square difference (MSD) of frames is devised in [127]. Pastrana-Vidal and Gicquel [128] proposed a generalized model for different fluidity break situations, such as regular, irregular, isolated, sporadic, and several discontinuity durations including various distributions and densities.

a.1.2: MULTIPLE ARTEFACTS BASED METHODS

Single artefact based techniques may not lead to satisfactory quality perceived assessment in presence of other artefacts. Thus, estimations of different artifacts are fused to yield a single quality score. For instance, Oelbaum *et al.* [129] formulated a rule-based video quality assessment technique that integrates the information from blockiness, blurriness, spatial activity, temporal predictability, edge continuity, motion continuity, and color continuity using multivariate data analysis method. Romaniak *et al.* [130] created a composite method to correlate well with subjective quality assessment via blocking and flickering measure of H.264/AVC encoded videos. The metric proposed in [131] employs a multiple regression for weighted integration of three artifacts (i.e., blurring, blockiness, and jitter/jerkiness) both in the luminance and chrominance planes for perceived quality estimation of standard-definition television (SDTV) sequences. A modular method to account for frame freeze/jerkiness and clearness/sharpness in MPEG-4 encoded videos has been studied in [132], which combines artifacts both from spatial and temporal domain to achieve a final perceived quality score. Culibrk *et al.* [202] explored the effect of bottom-up motion saliency features for the problem of MPEG-2 coded VQA and proposed a no-reference video quality estimator by analyzing video coding artifacts separately for salient motion and other regions of the frames.

a.2: FEATURES MEASURE-BASED METHODS

The methods in this group decompose a video signal into various features to represent specific aspects of visual information and their relation to the corresponding perceptual quality. Based on their particular functions, methods in this class are partitioned into two sets: natural scene statistics and HVS pixel-based features.

a.2.1: NATURAL SCENE STATISTICS BASED METHODS

The natural scene statistics (NSS) for corresponding quality values was studied in [133] to engineer a NR quality scheme that utilizes curvelet, wavelet, and cosine transform to analyze distortions, such as noise, blur, and artifacts introduced by compression. Likewise, in [134], a model based on temporal statistics of videos (i.e., natural motion statistics obtained from independent component analysis) was presented. The idea of 2D- and 3D-based statistical features based quality estimator has been investigated in [135] for stereoscopic visual information.

a.2.2: HVS PIXEL-BASED METHODS

These methods estimate perceived quality relying on certain HSV statistics derived from pixels of a video. An NR HVS based quality estimation for color video has been derived in [136], where different channels of the HVS have been processed with 3D multispectral wavelet decomposition considering pixel's contrast and luminance values. A perceptual mask weighted flow tensor between successive frames is employed to yield a final quality score. A general-purpose framework that is based on 3D shearlet transform and convolutional neural network (CNN) was proposed in [137]. Ries *et al.* in turn, showed that content of a video can help much in perceived quality assessment [138]. Their developed method predicts the video quality by classification of feature vector made of statistics on pixel motion (e.g., uniformity of the pixel movement) together with bitrate and frame rate information. Similarly, Khan *et al.* [139] exploited content of the video for quality estimation by combining encoding and transmission level parameters. The technique concluded, using an adaptive network-based fuzzy inference system (ANFIS) and a regression model for score computation, that transmission parameters (e.g., packet error rate) have more impact on the perceived quality than the compression parameters (e.g., frame rate). In turn, the mean square error distortion (i.e., pattern of lost macroblocks) caused by network impairments for a H.264/AVC encoded video was studied in [140] for perceived quality evaluation.

b: BITSTREAM-BASED METHODS

These methods adopt usage of bitstream data for quality estimation. As such, they do not need to process the full video data, since information from the bitstream (e.g., coding modes, motion vectors) are readily available. Nonetheless, bitstream-based methods are natively coding standard specific as different encoders have independent formats. According to the level of information used for processing, bitstream-based methods can further be divided into three categories: *parametric planning model*, *parametric packet-layer model*, and *bitstream layer model*.

b.1: PARAMETRIC PLANNING MODEL

The parametric planning techniques use codec type, packet loss rate, and bitrate for a crude quality evaluation.

The well-know example of this category is Opinion model for video-telephony applications described in ITU-T (G.1070). A quality prediction model for H.264/AVC videos in IPTV is presented in [141], which translates the encoding, packet and client information into overall perceived quality. The MSE of patterns of packet loss may also give some insight of the perceived quality, which is the base of the work in [142], where a model to establish the relationship between MSE and average motion vector length resulted in reliable quality estimates.

b.2: PARAMETRIC PACKET-LAYER MODEL

The visual quality estimation work in [143] does not require decoding the video at any level and uses the relationship between error concealment, motion in the video, importance of the frame regions and size of frames to compute the quality score. The work in [144] uses a nonlinear relationship between an objective quality metric and two quality-related parameters (i.e., value of the interval between intra-frames and packet loss rate). Different effects of packet loss over visible degradation was probed in [145] for H.264/AVC and HD videos. It was found that more than 75% human users perceived an artifact when packet loss was visible.

b.3: BITSTREAM LAYER MODEL

Bitstream layer models can do any type of analysis of the bitstream except usage of pixel data, thus are comparatively more complex, but offer better performance. Yang *et al.* [146] argued that their framework can be used for real time quality estimation, where the quality score is achieved by pooling QoS parameters, such as, packet loss rate, spatial and temporal complexities from the bitstream information. The investigation in [147] concluded that fusion of DCT coefficients data, a packet loss model identical to the one presented in ITU-T.G.1070, and a frame type- and error pattern-dependent model yields best visual quality prediction. Nonetheless, it was shown in [148] that the Cauchy distribution is more suitable for quality estimation than DCT coefficients.

c: HYBRID METHODS

Techniques that combine the coded bitstream and decoded media statistics are termed no-reference hybrid quality estimation methods. Hybrid methods can be divided into two categories: pixel and bitstream-based features or artifacts, and statistics of transform coefficients.

c.1: PIXEL AND BITSTREAM-BASED FEATURES

Yamada *et al.* [149] proposed a hybrid bitstream (i.e., packet lengths) and pixel domain (i.e., motion intensity and luminance discontinuity) quality estimator. Another hybrid non-reference quality framework, which fuses information from the packet layer (packet loss rate, packet size), bitstream layer (frame error, frame duration), and media layer (blurring, blocking), has been presented in [150] for videos transmitted over long term evolution (LTE) networks. Likewise, in [3] and [151] the Application Performance Metrics (APM) that characterizes the impact of rebuffering events user-viewing activities on the QoE for HTTP video streaming service and Universal Mobile Telecommunication System (UMTS) quality metric that characterizes video transmission including video content over wireless networks have been studied.

c.2: STATISTICS OF TRANSFORM COEFFICIENTS

The perceived quality can be assessed as a fusion of transform coefficients, bitstream features and pixel domain, e.g., [152], [153] in which PSNR was obtained for MPEG-2 coded videos via DCT coefficients as a Laplacian distribution. Nonetheless, its accuracy for quality evaluation for B type frames is low. Therefore, authors later integrated picture energy with DCT coefficients to attain improved accuracy even for SDTV and HDTV sequences.

C. STATE-OF-THE-ART IN AUDIOVISUAL QUALITY ASSESSMENT

There exists ample studies on quality assessment of individual modalities, including the psychophysical processes involved in their quality perception. However, audiovisual quality assessment, which is a multi-modal information process, is a relatively under explored field. Though various details of neurophysiological processing of audiovisual data remain unknown, empirical studies have demonstrated that the auditory and the visual domain have mutual influence on the perceived overall audiovisual quality [13]. One set of studies, e.g., [154], [155], indicated that the video channel is more important in perceived audiovisual quality. Other, however, e.g., [2], have suggested that the audio channel is more vital than video one, specially in teleconference scenario where humans pay more attention to audio information. Similarly, the experiments in [156] found that more bits allocated to audio may lead to attaining a higher perceived audiovisual quality at very low bitrates applications, e.g., VoD on mobile devices. All in all, audio or video channel's importance depends on application and/or context.

Since audiovisual quality perception involves interactions between two sensory modalities, one modality can modify the perceptual experience formed by the other. For instance, speech intelligibility can be improved by attaching a visual channel that shows the lip movements. Preliminary empirical analysis conducted by Rimmel *et al.* [157] observed the mutual compensation between modalities, i.e., increased quality of one modality remarkably improved the perceived quality of another one in video telephony. Detailed studies about the effects of various types of interaction between audio and video modalities on perceptual quality can be seen, e.g. in [2], [5], and [6]. Majority of previous works indicate that video quality has more influence on perceptual audio quality than vice versa [154]. However, contrary results have been reported in the literature, e.g., the study in [2] showed audio quality to be more vital than video quality in 'talking head' scenarios. While, homogenous work in [158] noticed no influence of audio on video quality but only very weak influence of video on audio quality. Mki *et al.* [194]

investigated correlations of the audio quality, video quality and interaction of these with audiovisual quality. The study found that video quality has higher correlation with audiovisual quality than audio quality. Moreover, it was reported that interaction of audio and video quality has higher correlation with audiovisual quality than either of the individual ones. Overall, it is a widely accepted fact that audiovisual quality, besides individual audio and video qualities, is influenced by other factors as well, such as different context (passive viewing and listening or interactive setting), content, attention of the user and task [6].

Another key factor that contributes to perceived audiovisual quality and intelligibility is *synchronization* between audio and video stimuli [30]. It is known that audiovisual quality is inversely proportional to asynchrony [13]. Improper synchronization can distract and annoy the viewer, which may reduce the clarity of the intended message and quality [159]. As per the ITU-R BT.1359, the threshold of acceptability for audio leading video is about 90 ms and in reverse situation about 185 ms, on average. There exists several audio and video synchronization methods, as also discussed in Section III-B.

Though there is a significant relationship between the perceived audiovisual quality and the audio-visual contents [5], limited research has been conducted on the topic. The majority of the existing methods excogitate the audio and video contents latently due to semantics/content being very subjective (e.g., news may be interesting for adults but children may think cartoon is important) thus it is very challenging to devise universal semantics importance model. Some researchers believe that the overall audiovisual quality can be attained by a weighted linear combination of perceptual quality and semantic quality; the former is the satisfaction of a user perceiving the multimedia signals and the latter is the perceived amount of information conveyed by the signals [6]. While, some suggests that the objective audiovisual perceptual quality model that takes into account also the content of the multimedia may be modelled at two different levels: cognitive and affective levels. The cognitive level can be used to model how a subject perceives the content and the affective level can be used to define the affective characteristics of the content [5]. Few recently proposed quality assessment frameworks that take the content type into account are [160], [161]. Specifically, Song *et al.* [161] attempted to identify the relationship between the audiovisual quality, content and QoE. The audiovisual content was materialized in terms of user interest that was defined as a physically expressed state of concentration which can be visually recognized when a user is involved in the audiovisual content/story by his/her eyes. Moller and Raake [13], [199] investigated the influence of audio-visual Focus of Attention, namely saliency, in the perceived quality of standard definition multimedia audio-visual content. The study found that higher spatial resolutions on the sound-emitting regions in image sequences leads to the same quality when compared to the case where all moving objects receive high priorities for the spatial resolution and even to the cases without blurring, unless the blur effect is too strong.

The perceptual and cognitive basis of audiovisual quality assessment, i.e., at which stage of the perceptual processing chain the modalities are actually fused, is yet not fully determined. However, the majority of the researcher have adopted the late fusion theory, in which auditory and visual channels are processed internally to yield respective quality values that are integrated at a late stage to form a single overall perceived quality [2]. Audiovisual quality is therefore usually described as a fusion of two dimensions (i.e., audio and video qualities), as shown in Fig. 4. The most common fusion model used and adopted in several studies [2], [5], [6] is the one reported in (Eq. 1). However, it is worth mentioning that there are no commonly agreed values or derivations for the four fusion parameters (a_0–a_3) in (Eq. 1); values reported in the literature range from $a_0 = -3.34$–4.26, $a_1 = -0.19$–0.85, $a_2 = 0$–0.89, to $a_3 = -0.01$–0.26. Few studies on human cognitive understanding suggest that audio and video channel might be integrated in an early phase of human perception formation [162]. Based on this, several researchers [2], [154] proposed audiovisual quality models as a multiplication of audio and video quality with equal importance, as shown in (Eq. 6):

$$Q_{AV} = a_0 + a_1 Q_A Q_V. \tag{6}$$

Similarly, Martineza *et al.* [163] proposed three audiovisual perceived quality metrics. The first model is simple linear model as given by (Eq. 7):

$$Q_{AV} = a_0 + a_1 Q_A + a_2 Q_V. \tag{7}$$

The second metric is based on the weighted Minkowski model as:

$$Q_{AV} = (a_1 {Q_A}^P + a_2 {Q_V}^P)^{\frac{1}{P}}, \tag{8}$$

where the exponent P is obtained from the fit for Minkowski model. The third metric is a power model as given by (Eq. 9):

$$Q_{AV} = (a_1 + a_2 {Q_A}^{P1} {Q_V}^{P2}) \tag{9}$$

As some studies [154], [155] suggested that visual modality can be more dominant than audio in perceived audiovisual quality formation, specially for videos with high motion data, thus authors in [2] presented the following model:

$$Q_{AV} = a_0 + a_1 Q_V + a_2 Q_A Q_V \tag{10}$$

Although models in equations 1–10 attained fairly accurate predicted audiovisual quality in some studies when audio and video quality spans are the same, it does not reflect the differences in the influence of audio only and video only stimuli on the overall quality. Moreover, they also can not fully capture some other influential factors, e.g., goal of assessment, testing environment, and impact of impairments, synchronicity. Thus, Saidi *et al.* [164] proposed a audio-video synchronization based quality prediction model as follows:

$$Q_{AV} = a_0 + a_1 Q_A + a_2 Q_V + a_3 Q_A Q_V + a_4 D_{synch}, \tag{11}$$

where the desynchronization term is set to $D_{synch} = 5 - MOS_{synch}$. In the experiments, authors obtained MOS_{synch} from a specific 5 point impairment scale: 1-very annoying, 2-annoying, 3-slightly annoying, 4-perceptible but not annoying, and 5- imperceptible due to desynchronization. Another synchronization based multimedia perceived quality model is presented by Heyashi *et al.* [165] as:

$$Q_{MM} = a_0 + a_1 Q_{AV} + a_2 Q_D + a_3 Q_{AV} Q_D, \quad (12)$$

where Q_{MM}, Q_{AV} and, Q_D are predicted multimedia quality, audiovisual quality and quality degradation calculated based on audiovisual delay, respectively.

Most multimedia quality models in the literature are proposed for short-duration sequences with a single content/ scene. Thus, when used with long-duration sequences, they are usually applied at short temporal segments and later averaged for overall quality with equal weights for each segment. Simple averaging, may not be appropriate for long sequences with multiple contents and scenes of varying complexity, which should be assigned larger weights. Towards this, You *et al.* [6] presented a weighted temporal averaging method for long-term sequences as:

$$Q_{AV} = \sum_i W_i S_i (a_{0,i} + a_{1,i} Q_A + a_{2,i} Q_V + a_{3,i} Q_A Q_V), \quad (13)$$

where i, W_i and S_i denote different segments whose duration might be different from each other because of different multimedia contents, the weight of this segment that is affected by some external factors and quality level (as different quality levels make different contributions to the overall quality), and the semantic/affective importance of a segment obtained from a content analysis model, respectively. It is worth noticing that the fusion parameters a_0, a_1, a_2, a_3 might be different for different segments. The search for an optimal temporal quality pooling method, however, is still an open issue.

The ITU-T has proposed some standardized audiovisual quality prediction models, e.g., ITU-T P.1201, ITU-T G.1070 and ITU-T G.1071. The ITU-T P.1201 model was proposed to compute the audiovisual quality of streaming services. It is suitable both for lower resolution (e.g., mobile TV) and higher resolution applications (e.g., IPTV). The model is non-intrusive and utilizes packet-header information to provide individual predictions of audio, video, and audiovisual quality via the five-point MOS scale. The ITU-T G.1070 model was recommended for video telephony. The overall multimedia quality is estimated by network, application and terminal equipment parameters. It is more useful for quality of experience and quality of service planners. The model can be applied to compute independent speech quality (using speech codec type, packet loss rate, bit rate and talker echo loudness rating), video quality (using video format, display size and codec type, packet loss rate, bit rate, key frame interval and frame rate), and multimedia quality (using individual speech and video quality, audiovisual asynchrony and end-to-end delay). While, ITU-T G.1071 model was proposed for network planning of audio and video streaming services, it is applicable for lower- as well as higher-resolution services. It is worth noticing that this model is limited to QoS/QoE planning, and cannot be used for quality benchmarking and monitoring. The network-planning assumptions (e.g., video resolution, audio, and video codec types and profiles, audio and video bitrates, packet-loss rate and distribution) are employed to attain the separate predictions of audio, video and audiovisual quality. More details of standardization activities regarding audiovisual quality assessment and the related standards can be seen in [4], [5], and [13]. Though, these standardized methods reach high prediction accuracy, they intrinsically have limited applications. Thus, researchers are trying to improve these models as well as proposing new techniques.

Few recent studies, e.g. [5], [13], [166], have attempted to estimate the audiovisual perceived quality with machine learning algorithms, such as neural networks and random forest ensemble. Machine learning based approaches do not require intermediate predictions for audio and video quality, and still successfully capture the complex relationships between influence factors, thereby achieving high accuracy and generalization capability. Recently, a novel trend to assess user perception of audiovisual quality using electroencephalography (EEG) and other physiological measurement devices have emerged [4]. The empirical results depict high correlation between perceived multimedia and physiological data [13].

VI. DATABASES FOR AUDIO-VISUAL QUALITY ASSESSMENT

Databases of audio, video or audiovisual signals annotated with subjective ratings constitute essential ground truth for training, testing, and benchmarking methods for perceptual-based quality assessment. Over the years, several data sets have been released in the public domain. In this section, we present an overview of a few *representative* databases of uni- and multi-modal signals, including physiological, that have been used in the literature, which are also summarized in Table 6.

A. AUDIO

1) ITU93 [37]

It is based on seven audio stereo sequences (i.e., Asa Jinder, bagpipe, bass clarinet, castanets, harpsichord, German male speech and violin) that were processed by different tandem code configurations of MPEG layer 2 at 192, 256 and 360 kbit/s/channel. There are total of 42 listening test signals whose quality values were rated by 33 subjects.

2) MPEG95 [22]

It is based on six mono sequences (i.e., bag pipe, castanets, glockenspiel, harpsichord, pitch pipe and English female speech) processed by 22 encoding variations of six audio codecs. There are 132 listening test signals available with subjective quality ratings given by 63 subjects.

TABLE 6. Publicly available audio, video and audiovisual quality assessment datasets.

Modality	Dataset	Description of characteristics and distortions	Subjective ratings (e.g., MOS)			Year
			Audio	Video	Audiovisual	
Audio	ITU93 [41]	7 audio stereo types: Asa Jinder, bagpipe, bass clarinet, castanets, harpsichord, German male speech and violin with MPEG1 Layer 2 tandem codec at 92, 256, 360 kbit/s	Yes	No	No	1993
	MPEG95 [26]	22 encoding variations of six audio codecs	Yes	No	No	1995
	REVERB challenge [189]	3 subsets of both clean and reverberant speech signals with 1-ch, 2-ch, and 8-ch recordings at a sampling frequency of 16 kHz	No	No	No	2016
	Live Music [190]	4 genres (i.e., rock, pop, electronic, and country) of real and synthetically altered live music recordings; Kind of noises: amplitude compression and amplification, butterworth filtering, white noise and crowd noise additions	Yes	No	No	2013
	Blizzard Challenge [191]	50 children's text book and audiobooks spoken by a British female speaker with 44.1 kHz sampling rate, 2 channels, 16 bit encoding	Yes	No	No	2016
Video	Poly NYU VQ [192]	Quantization error; video format and resolution: CIF (352x288) QCIF (176x144), 30 frame-rates	No	Yes	No	2008
	LIVE VQ [193]	Compression and transmission error; video format and resolution: YUV+264/M2V (768x432); 25/50 frame-rates	No	Yes	No	2010
	VQEG HDTV [194]	Compression and transmission error; 1920x1080 resolution; 59 frame-rates; 1x (0.7-4.2) PLR%	No	Yes	No	2010
	MMD [195]	Compression error for mobile TV; low-high motion; 1x per Seq bitrates; 480p resolution; 25 frame-rates	No	Yes	No	2012
	CVD2014 [196]	Compression and video acquisition related distortions, e.g., flickering, jerky; Videos captured from 73 cameras; different video format and resolution, e.g., QCIF (176x144), QVGA (352x240), HD (1280x720), FHD (1920x1080)	No	Yes	No	2014
Audiovisual	PLYM [197]	Compression and transmission error; low motion; 1x per Seq bitrates; 144p resolution; 8, 15 frame-rates; 5x (0.01-0.20) PLR%	Yes	Yes	Yes	2010
	TUM [198]	Compression error; low-high motion; 4x per Seq bitrates; 1080p resolution; 50 frame-rates	No	No	Yes	2012
	VQEG [27]	Compression and transmission error; low-high motion; 3x per Seq bitrates; 480p resolution; 30 frame-rates	No	No	Yes	2012
	VTT [199]	Compression and transmission error; low-high motion; 1x per Seq bitrates; 480p, 720p, 1080p resolution; 20-30 frame-rates; 5x (0.3-4.8) PLR%	Yes	Yes	Yes	2013
	INRS [171]	Compression and transmission error; low motion; 4x per Seq bitrates; 720p resolution; 4x(10-25) frame-rates; 5x (0-5) PLR%	No	No	Yes	2016

3) REVERB CHALLENGE [184]

It was used in 2014 REVERB challenge [184], and consists of three subsets: a training, a development, and an evaluation set. Both clean and reverberant speech signals recorded as 1-ch, 2-ch, and 8-ch recordings at a sampling frequency of 16 kHz are available publicly.

4) LIVE MUSIC DATASET [185]

The database is comprised of two subsets of live music recordings of four music genres (i.e., rock, pop, electronic, and country) for perceptual audio quality assessment, which were annotated by 60 subjects with normal hearing. The first subset contains 500 live music recordings with human annotations obtained via a web-based interface; while the second one contains 2,400 synthetically altered live music recordings in 8 different quality conditions.

5) BLIZZARD CHALLENGE 2016 [186]

This dataset was used for text to speech synthesis Blizzard Challenge 2016, and consists of speech and text data of professional audiobooks. In particular, about 5 hours of British English speech data (44.1 kHz sampling rate, 2 channels, 16 bit encoding) from a single female speaker is provided.

B. VIDEO

1) POLY NYU VQ [187]

This database contains three individual but related test using videos with different frame rates and quantization parameters. Specifically, distorted videos were generated by different temporal, spatial, and SNR resolutions. A total of 31 viewers participated in the test, while 20 ratings for each processed video sequence is available.

2) LIVE VQ [188]

The LIVE Video Quality database includes 15 video sequences with recent and advanced codecs such as MPEG-2 and H.264 compressions, simulated transmission of H.264 packetized bitstreams through error-prone IP networks and wireless networks. Each video was assessed by 38 human subjects. The videos in this dataset span a much wider range of quality, e.g., the low quality to those found in found in online video streaming application, such as YouTube.

3) VQEG HDTV [189]

The dataset is composed of 6 subsets but only 5 subsets are publicly available. The test conditions are MPEG-2 and H.264 compression with two types of network impairments, i.e., slicing error and freeze error caused by burst packet loss.

4) MADE FOR MOBILE DATASET (MMD) [190]

It consists of 19 pairs of extracted video sequences from 22 professionally produced clips with 18 observers for subjective test. The aim of the database is to assess content production rules as well as video quality between mobile devices and TV.

5) CVD2014 [191]

The CVD (Camera Video Database) utilizes real cameras instead of introducing distortions via post-processing that leads to a complex distortion space (e.g., sharpness, jerkiness) in regard to the video acquisition process. The dataset is comprised of 234 videos that are recorded using 78 different cameras. The subjective ratings are also included.

C. AUDIOVISUAL

1) PLYM [192]

The PLYM dataset was created to study the audiovisual quality predictions for video calls over wireless applications. The subjective tests for 60 audio, 60 video and 60 audio-visual sample with 16 observers are available. The videos were encoded with the H.263 and G.711 law codecs using 6 motion, 2 video frame rates and 5 packet loss rates.

2) TUM [193]

This data is targeted for high definition videos audiovisual quality assessment with 1080p50 format. The video sequences were encoded with the H.264/AVC video codec including different bitrates and encoding impairments, e.g., blurring and flicker. The subjective scores were obtained from 21 users.

3) VQEG [23]

There are 10 audiovisual subsets in this database produced by six different international laboratories in a study to determine the most appropriate way to perform audiovisual quality testing. The audiovisual sequences were coded to attain three coding qualities, i.e., high, medium, and low. Particularly, the H.264/AVC video codec and Advanced Audio Coding (AAC) with 6 and 3 bitrate levels, respectively, were used for encoding. While, the subjective scores were obtained from 35 observers.

4) VTT [194]

It consists of 125 audiovisual sequences from streaming services with subjective quality values provided by 125 users. The H.264 video and AAC audio streams were adopted for the test with varying impact of resolution, movement quantity, packet loss rate, and mean loss burst size.

5) INRS [166]

It contains 160 unique configurations for audiovisual content with different media compression and network distortion parameters, e.g., video frame rate, packet loss rate, and quantization and noise reduction parameters. The H.264 video codec and AMR-WB audio codec were employed to encode video and audio streams; while 30 subjects rated the overall audiovisual quality.

D. PHYSIOLOGICAL

1) PHYSIOLOGICAL EVALUATION OF SYNTHESIZED SPEECH QoE (PhySyQX) [196]

It is an EEG dataset using a Biosemi ActiveTwo system. The quality ratings were obtained from 21 healthy participants by presenting 44 synthesized speech stimuli (approximately 20 s long), generated from 7 commercially available TTS systems along with 4 natural voices.

2) DATABASE FOR EMOTION ANALYSIS USING PHYSIOLOGICAL SIGNALS (DEAP) [197]

It is composed of EEG and peripheral physiological signals of 32 participants, when they watched 40 one-minute long music videos with varying emotional content. The participants provided the quality rating for each video in terms of the levels of arousal, valence, like/dislike, dominance, and familiarity.

VII. OPEN ISSUES AND FUTURE RESEARCH DIRECTIONS

Despite great progress in the audio, video and audiovisual quality (QoE) assessment, a range of issues remains to be addressed. In the following, some of the open issues and research directions are described.

A. GENERALIZATION CAPABILITY

It is easy to see in the literature that a given multimedia perceptual-based (objective) quality assessment model will typically perform well for some content, context or degradation types, but not so well for others on which either the model was not tuned or proposed for, thereby leading to low generalization capability. For instance, an audiovisual quality assessment model developed for videoconferencing will most likely not perform well on video streaming applications. While, multimedia is transmitted over broad set of network infrastructure (e.g., jitter, packet loss, and bandwidth), with varying characteristics (e.g., codec, spatial and temporal information, bitrate), contents (e.g., sports, news), contexts (e.g., office, street), different capture devices (e.g., PC, smartphone) and setups (e.g., conversational, multimedia streaming), development of generalized quality assessment models will greatly advance the state-of-the-art in quality assessment & QoE field.

B. ADVANCED MACHINE LEARNING (AML) BASED ASSESSMENT

Traditional quality assessment methods are often based on explicit modeling of the highly non-linear behavior of human perception. As a result, many traditional models are prone to overfitting or have questionable overall reliability. Conversely, AML based methods try to mimic quality perception instead of designing an explicit model of the human auditory or visual system. There exist few preliminary studies on use of AML for unimodal (audio or video) quality assessment, but audiovisual objective quality models based on AML, such as dictionary learning and deep learning, have seldom been explored. AML paradigms can be utilized for robust segmentation, representation learning, feature extraction/selection, classification and finding temporal correlations within and between different modalities to attain higher interoperability and generalization capability of the models. Future QoE/QoP audiovisual models should explore AML paradigms.

C. MULTIMODAL QUALITY PERCEPTION

The audiovisual quality perception is a multimodal process, which integrates visual and auditory sensory channels.

There are two well-known theories for multimodal fusion: early and late fusion. Most of the works in the literature have adopted the early fusion theory. However, multimodal quality perception yet suffers from advanced theoretical understanding from a neurophysiological point of view. There is huge demand for understanding many complexities (e.g., spatial and temporal proximity and resolution between modalities and stimuli) involving audiovisual quality perception in both subjective and objective domain. Given the neuroimaging advances seen to date, more neurophysiological QoE studies should be conducted to shed light on this matter.

D. IMMERSIVE QoE ASSESSMENT

The inclusion of a third dimension brings more challenges in quality assessment models. The depth impression by stereoscopic displays and multichannel audio signals are another potential source of either quality improvement or distortions. Some studies tried to apply existing 2D models for 3D unimodal and multimodal QoE, but such an approach does not account for specific distortions, such as stereoscopic crosstalk. Thus, while some works have specifically targeted 3D perceived multimedia quality assessment, this is a research topic still in its infancy stage. The 3D QoE is multifaceted with distortion, display and discomfort issues, and their impact and relation to overall 3D quality is poorly understood. Existing methods only consider two factors, i.e., depth and display. There are no prediction models for 3D naturalness and why some users feel dizzy or nauseous. The latter case can be better understood by devising methods for 'simulator of sickness' in 3D QoE, which may later be useful in designing 3D QoE assessment metrics.

E. LARGE-SCALE ANNOTATED MULTIMODAL DATASET

Progress in multimedia QoE deeply depends on the existence of comprehensive large-scale databases that contain different coding, transmission, and decoding inaccuracies, and various potential content and contexts. Though several disparate databases are available, they are very limited in size and broadness. Large-scale public multimedia databases (including corresponding subjective ratings, and if possible recorded physiological signals) will help to compare various QoE models, discover inter and intra relationship between different factors and phenomena, and to make strong conclusions in terms of statistical significance. Crowdsourcing techniques may help obtaining annotated large-scale databases [13].

F. CONVERSATIONAL QUALITY ASSESSMENT

Conversational quality assessment, where multiple subjects talk using unimodal (only speech) or multimodal channels over a test connection, is important for telecommunication devices, networks, and algorithms. Conversational QoE can probe various dimensions including handsets/devices combination, side tone, echo, level and delay impairment, and the effect of relationship between interacting subjects, which are usually not assessable via listening-only or talking-and-listening tests. Conversational QoE tests are generally considered more expensive, thus are relatively rare in the literature. There are few human-human interaction QoE assessment methods, but the researchers have mainly ignored human-machine and sizeable-group conversational QoE [6].

G. NO-REFERENCE/NONINTRUSIVE QoE METRICS

Usually, full- or reduced-reference based quality assessment methods attain higher accuracy, but they are not usable in all applications owing to their need of reference signal. No-reference/nonintrusive QoE metrics are gaining momentum [5]. Particularly, nonintrusive audio (speech) quality metrics with high predictive power are highly coveted [24].

H. PSYCHOPHYSIOLOGY-BASED QoE ASSESSMENT

The quintessential psychophysical techniques quantitatively evaluate the relationship between physical stimuli and the conscious perceptions, while psychophysiology looks into the physiological bases of perceptual and cognitive processes. Namely, psychophysiology evaluates implicit responses to physical stimuli rather than explicit ones, which may avoid potentially misleading subjective ratings. Recent studies have shown that use of psychophysiological measures in quality assessment algorithms (e.g., a method based on analysis of neuronal activity) can lead to better QoE assessment. There is a need for designing better non-learning or learning-based fusion schemes to combine psychophysiological and psychophysical assessment [175]. Because individually they have limited capability; their integration can improve overall insight into QoE. The lack of standards for physiological methodologies for QoE is hampering the progress. Moreover, the lack of public databases containing ground truths has further stymied research on this topic. Current trend of physiological measurements being integrated into personal computing devices also provide an opportunity to devise techniques for continuous QoE monitoring in a minimally invasive way.

I. MULTIMEDIA NETWORKS MANAGEMENT VIA QoE

The management of multimedia services over access networks is another challenging issue of QoE due to the larger heterogeneity of the devices, user's requirements and communication channels. Current multimedia access networks management depends mainly on time-consuming and costly manual and reactive process, especially when anomalies occur. To overcome this limitation, autonomic management framework can be developed to maximize user's QoE. In other words, perceptual quality measures can be used to systematically steer, in real-time, management algorithm parameters (e.g., video rate adaptation, admission control, and traffic flow adaptation) for optimized QoP/E [3].

VIII. CONCLUSION

A recent spurt in multimedia services over wired and wireless networks has also triggered perceptual quality assessment research. In particular, there is a huge demand for methods that are capable of estimating and quantifying the coding, transmission and decoding (reception) quality, services, experience and satisfaction as perceived by the end-user. Though the perceptual multimedia quality assessment proved to be a difficult task, a plentiful of research and development efforts have been devoted to it and its applications, thereby leading to significant progress in the field. This article provided a survey of existing multimedia quality assessment methods with a focus on perceptual-based audio, video and audiovisual quality measurement techniques. The paper also presented a classification of audio and video quality metrics based on their underlying methodologies. Moreover, influential factors, quality of services, quality of experience, quality of perception, quality assessment using physiological signals, and representative public audio, video, audiovisual and physiological databases have been discussed. Still, there are various issues remaining to be addressed to attain increased understanding of the many complexities of human perception for both individual and multimodal qualities. Thus, the paper discussed some of the open issues and challenges in the filed. We are still a long way from any dependable multimodal quality/experience/perception assessment method, which will require interdisciplinary research efforts of different domains, such as human vision, physiology, and psychophysiology, etc.

REFERENCES

[1] *Requirements for an Objective Perceptual Multimedia Quality Model*, document ITU Rec.J.148, ITU Telecommunication Standardization Sector, 2010.

[2] D. S. Hands, ''A basic multimedia quality model,'' *IEEE Trans. Multimedia*, vol. 6, no. 6, pp. 806–816, Dec. 2004.

[3] A. Khan, L. Sun, and E. Ifeachor, ''QoE prediction model and its application in video quality adaptation over UMTS networks,'' *IEEE Trans. Multimedia*, vol. 14, no. 2, pp. 431–442, Apr. 2012.

[4] T. Dagiuklas, *Multimedia Quality of Experience (QoE): Current Status and Future Requirements*. New York, NY, USA: Wiley, 2015.

[5] B. Belmudez, *Audiovisual Quality Assessment and Prediction for Videotelephony*. Berlin, Germany: Springer, 2014.

[6] J. You, U. Reiter, M. M. Hannuksela, M. Gabbouj, and A. Perkis, ''Perceptual-based quality assessment for audio—Visual services: A survey,'' *Signal Process., Image Commun.*, vol. 25, no. 7, pp. 482–501, 2010.

[7] S. Winkler and P. Mohandas, ''The evolution of video quality measurement: From PSNR to hybrid metrics,'' *IEEE Trans. Broadcast.*, vol. 54, no. 3, pp. 660–668, Sep. 2008.

[8] Z. Wang and A. C. Bovik, ''Mean squared error: Love it or leave it? A new look at signal fidelity measures,'' *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 98–117, Jan. 2009.

[9] S. Bech and N. Zacharov, *Perceptual Audio Evaluation—Theory, Method and Application* . New York, NY, USA: Wiley, 2006.

[10] W. Lin and C.-C. Jay Kuo, ''Perceptual visual quality metrics: A survey,'' *J. Visual Commun. Image Represent.*, vol. 22, no. 4, pp. 297–312, 2011.

[11] D. Campbell, E. Jones, and M. Glavin, ''Audio quality assessment techniques—A review, and recent developments,'' *Signal Process.*, vol. 89, no. 8, pp. 1489–1500, 2009.

[12] B. W. Keelan, *Handbook of Image Quality*. New York, NY, USA: Marcel Dekker, 2002.

[13] S. Möller and A. Raake, *Quality of Experience: Advanced Concepts, Applications and Methods*. Berlin, Germany: Springer, 2014.

[14] U. Jekosch, *Voice and Speech Quality Perception: Assessment and Evaluation* (Signals and Communication Technology). New York, NY, USA: Springer-Verlag, 2005.

[15] K. Brunnstrom *et al.*, ''Qualinet white paper on definitions of quality of experience,'' in *Proc. 5th Qualinet Meet.*, 2013, pp. 1–24.

[16] E. B. Goldstein, *Sensation and Perception*. Boston, MA, USA: Cengage Learn. 2009.

[17] R. W. Fleming, ''Visual perception of materials and their properties,'' *Vis. Res.*, vol. 94, pp. 62–75, Jan. 2014.

[18] H. Davson, ''Human Eye,'' in *Encyclopdia Britannica*. Chicago, IL, USA: Encyclopædia Britannica, Inc., 2010.

[19] R. J. Zatorre and P. Belin, ''Spectral and temporal processing in human auditory cortex,'' *Cerebral Cortex*, vol. 11, no. 10, pp. 946–953, 2002.

[20] F. W. Campbell and J. G. Robson, ''Application of Fourier analysis to the visibility of gratings,'' *J. Physiol.*, vol. 197, no. 3, pp. 551–566, 1968.

[21] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA, USA: MIT Press, 1999.

[22] *The E-model, a Computational Model for use in Transmission Planning*, document ITU-T Rec. G.107, 2005.

[23] M. H. Pinson *et al.*, ''The influence of subjects and environment on audiovisual subjective tests: An international study,'' *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 640–651, Oct. 2012.

[24] U. Engelke and H.-J. Zepernick, ''Perceptual-based quality metrics for image and video services: A survey,'' in *Proc. EuroNGI Conf. Next Gen. Int. Netw.*, May 2007, pp. 190–197.

[25] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, ''Objective video quality assessment methods: A classification, review, and performance comparison,'' *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 165–182, Jun. 2011.

[26] *Objective Perceptual Multimedia Video Quality Measurement in the Presence of a full Reference*, document ITU-T Rec. J.247, ITU Telecom. Standardization Sector, 2008.

[27] M. P. Hollier and A. N. Rimell, ''An experimental investigation into multimodal synchronization sensitivity for perceptual model development,'' in *Proc. Audio Eng. Soc. Conv.*, 1998, pp. 1–6.

[28] V.-T. Peltoketo, ''Objective verification of audio-video synchronization,'' Sofica Ltd, Seinajoki, Finland, Tech. Rep. 20-2012, 2012.

[29] *Tolerances for Transmission Time Differences Between the Vision and Sound Components of a Television Signal*, document Rec. ITU-T J.100, Int. Telecommun. Union, Geneva, Switzerland, 1990.

[30] D. Patterson and L. Evans, ''Synchronization of audio-visual elements in Web applications,'' in *Proc. 3rd Austral. Web Conf. (AWC)*, vol. 166. 2015, pp. 3–10.

[31] *Method for the Subjective Assessment of Intermediate Quality Levels of Coding Systems*, document Rec. IRU-R BS.1534-1, ITU, 2010.

[32] *Method for Objective Measurements of Perceived Audio Quality*, document ITU-R Rec. BS.1387-1, 2001.

[33] C. Cave, ''Perceptual modelling for low-rate audio coding,'' M.S. thesis, Dept. Elect. Comput. Eng., McGill Univ., Montreal, QC, Canada, 2002.

[34] R. Huber and B. Kollmeier, ''PEMO-Q—A new method for objective audio quality assessment using a model of auditory perception,'' *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 6, pp. 1902–1911, Nov. 2006.

[35] F. Itakura, ''Minimum prediction residual principle applied to speech recognition,'' *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-23, no. 1, pp. 67–72, Feb. 1975.

[36] N. Kitawaki, H. Nagabuchi, and K. Itoh, ''Objective quality evaluation for low-bit-rate speech coding systems,'' *IEEE J. Sel. Areas Commun.*, vol. SAC-6, no. 2, pp. 242–248, Feb. 1988.

[37] T. Thiede, ''Perceptual audio quality assessment using a non-linear filter bank,'' Ph.D. dissertation, Dept. Elect. Eng. Inf. Technol., Tech. Univ. Berlin, Berlin, Germany, 1999.

[38] J. Ma, Y. Hu, and P. C. Loizou, ''Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions,'' *J. Acoust. Soc. Amer.*, vol. 125, no. 5, pp. 3387–3405, 2009.

[39] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, ''A short-time objective intelligibility measure for time-frequency weighted noisy speech,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 4214–4217.

[40] S. Kandadai, J. Hardin, and C. D. Creusere, ''Audio quality assessment using the mean structural similarity measure,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar./Apr. 2008, pp. 221–224.

[41] A. Hines and N. Harte, ''Speech intelligibility prediction using a neurogram similarity index measure,'' *Speech Commun.*, vol. 54, no. 2, pp. 306–320, 2012.

[42] A. Hines, J. Skoglund, A. Kokaram, and N. Harte, ''ViSQOL: The virtual speech quality objective listener,'' in *Proc. Int. Workshop Acoust. Signal Enhancement*, Sep. 2012, pp. 1–4.

[43] A. Hines, E. Gillen, D. Kelly, J. Skoglund, A. Kokaram, and N. Harte, ''ViSQOLAudio: An objective audio quality metric for low bitrate codecs,'' *J. Acoust. Soc. Amer.*, vol. 137, no. 6, p. EL449, 2015.

[44] M. Karjalainen, "A new auditory model for the evaluation of sound quality of audio systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1985, pp. 608–611.
[45] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE J. Sel. Areas Commun.*, vol. 5, no. 5, pp. 819–829, Jun. 1992.
[46] W. Yang, M. Benbouchta, and R. Yantorno, "Performance of the modified Bark spectral distortion as an objective speech quality measure," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 1998, pp. 541–544.
[47] J. G. Beerends and J. A. Stemerdink, "A perceptual speech-quality measure based on a psychoacoustic sound representation," *J. Audio Eng. Soc.*, vol. 42, no. 3, pp. 115–123, 1994.
[48] J. G. Beerends, E. J. Meijer, and A. P. Hekstra, *Improvement of the P.861 Perceptual Speech Quality Measures*, document COM 12-20, ITU-T Study Group 12, 1999.
[49] D. J. Atkinson, *Proposed Annex A*, document Rec. p.861, ITU-T Study Group 12 Contribution 24 (Com 12-24 E), ITU, 1997.
[50] S. Voran, "Objective estimation of perceived speech quality. I. Development of the measuring normalizing block technique," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 4, pp. 371–382, Jul. 1999.
[51] A. W. Rix and M. P. Hollier, "The perceptual analysis measurement system for robust end-to-end speech quality assessment," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Jun. 2000, pp. 1515–1518.
[52] *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for end-to-end Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*, document ITU-T Rec.P.862, 2001.
[53] D. Picovici and A. E. Mahdi, "New output-based perceptual measure for predicting subjective quality of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2004, p. 633.
[54] *Perceptual Objective Listening Quality Assessment*, document ITU-T Rec. P.863, ITU, 2011.
[55] B. Patton, Y. Agiomyrgiannakis, M. Terry, K. Wilson, R. A. Saurous, and D. Sculley. (Nov. 2016). *AutoMOS: Learning a non-Intrusive Assessor of Naturalness-of-Speech*. [Online]. Available: https://arxiv.org/abs/1611.09207
[56] J.-H. Flesner, S. D. Ewert, B. Kollmeier, and R. Huber, "Quality assessment of multi-channel audio processing schemes based on a binaural auditory model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 1340–1344.
[57] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2046–2057, Sep. 2011.
[58] O. C. Au and K. H. Lam, "A novel output-based objective speech quality measure for wireless communication," in *Proc. 4th Int. Conf. Signal Process.*, Oct. 1998, pp. 666–669.
[59] *Analysis and Interpretation of INMD Voice-Service Measurements*, document ITU-T Rec.P.562, ITU, 2000.
[60] P. Gray, M. P. Hollier, and R. E. Massara, "Non-intrusive speech-quality assessment using vocal-tract models," *IEEE Proc.-Vis., Image Signal Process.*, vol. 147, no. 6, pp. 493–501, Dec. 2000.
[61] C. Jin and R. Kubichek, "Output-based objective speech quality using vector quantization techniques," in *Proc. 29th Asilomar Conf. Signals, Syst. Comput.*, Oct./Nov. 1995, pp. 1291–1294.
[62] T. H. Falk and W.-Y. Chan, "Nonintrusive speech quality estimation using Gaussian mixture models," *IEEE Signal Process. Lett.*, vol. 3, no. 2, pp. 108–111, Feb. 2006.
[63] A. E. Mahdi, "Perceptual non-intrusive speech quality assessment using a self-organizing map," *J. Enterprise Inf. Manage.*, vol. 19, no. 2, pp. 148–164, 2006.
[64] E. A. Mahdi and D. Picovici, "New single-ended objective measure for non-intrusive speech quality evaluation," *Signal, Image Video Process.*, vol. 4, no. 1, pp. 23–38, 2010.
[65] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1766–1774, Sep. 2010.
[66] H. Salehi and V. Parsa, "On nonintrusive speech quality estimation for hearing aids," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2015, pp. 1–5.
[67] D. Suelzle, V. Parsa, and T. H. Falk, "On a reference-free speech quality estimator for hearing aids," *J. Acoust. Soc. Amer.*, vol. 133, no. 5, p. EL412, 2013.
[68] H. Salehi and V. Parsa, "Nonintrusive speech quality estimation based on perceptual linear prediction," in *Proc. IEEE Can. Conf. Elect. Comput. Eng.*, May 2016, pp. 1–4.
[69] J. M. Kates and K. H. Arehart, "The hearing-aid speech quality index (HASQI)," *J. Audio Eng. Soc.*, vol. 58, no. 5, pp. 363–381, 2010.
[70] F. Rahdari, M. Eftekhari, and R. Mousavi, "A two-level multi-gene genetic programming model for speech quality prediction in voice over Internet protocol systems," *Comput. Elect. Eng.*, vol. 49, pp. 9–24, Jan. 2016.
[71] T. H. Falk and W. Y. Chan, "Hybrid signal-and-link-parametric speech quality measurement for VoIP communications," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 8, pp. 1579–1589, Nov. 2008.
[72] J. G. Beerends, K. Nieuwenhuizen, and E. L. van den Broek, "Quantifying sound quality in loudspeaker reproduction," *J. Audio Eng. Soc.*, vol. 64, no. 10, pp. 784–799, 2016.
[73] N. Mamun, W. A. Jassim, and M. S. A. Zilany, "Prediction of speech intelligibility using a neurogram orthogonal polynomial measure (NOPM)," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 4, pp. 760–773, Apr. 2015.
[74] Y. Chen, K. Wu, and Q. Zhang, "From QoS to QoE: A tutorial on video quality assessment," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 1126–1165, 2nd Quart., 2015.
[75] F. Lukas and Z. Budrikis, "Picture quality prediction based on a visual model," *IEEE Trans. Commun.*, vol. COM-30, no. 7, pp. 1679–1692, Jul. 1982.
[76] O. Verscheure and C. Van den Branden Lambrecht, "Perceptual quality measure using a spatiotemporal model of the human visual system," *Proc. SPIE*, vol. 2668, pp. 450–461, Mar. 1996.
[77] A. B. Watson, J. Hu, and J. McGowan, "Digital video quality metric based on human vision," *J. Electron. Imag.*, vol. 10, no. 1, pp. 20–29, 2001.
[78] F. Xiao, "DCT-based video quality evaluation," Stanford Univ., Stanford, CA, USA, Tech. Rep. EE392J, 2000.
[79] C. Lee and O. Kwon, "Objective measurements of video quality using the wavelet transform," *Opt. Eng.*, vol. 42, no. 1, pp. 265–272, 2003.
[80] J. You, T. Ebrahimi, and A. Perkis, "Attention driven foveated video quality assessment," *IEEE Trans. Image Process.*, vol. 23, no. 1, pp. 200–213, Jan. 2014.
[81] A. P. Hekstra *et al.*, "PVQM: A perceptual video quality measure," *Signal Process., Image Commun.*, vol. 17, no. 10, pp. 781–798, 2002.
[82] E. Ong, X. Yang, W. Lin, Z. Lu, and S. Yao, "Video quality metric for low bitrate compressed videos," in *Proc. Int. Conf. Image Process.*, Oct. 2004, pp. 3531–3534.
[83] E. Ong, W. Lin, Z. Lu, and S. Yao, "Colour perceptual video quality metric," in *Proc. IEEE ICIP*, Sep. 2005, pp. 1–4.
[84] P. Ndjiki-Nya, M. Barrado, and T. Wiegand, "Efficient full-reference assessment of image and video quality," in *Proc. IEEE Int. Conf. Image Process.*, Sep./Oct. 2007, pp. II-125–II-128.
[85] D. M. Chandler and S. S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2284–2298, Sep. 2007.
[86] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 312–322, Sep. 2004.
[87] S.-O. Lee and D.-G. Sim, "New full-reference visual quality assessment based on human visual perception," in *Proc. Int. Conf. Consumer Electron.*, Jan. 2008, pp. 1–2.
[88] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 335–350, Feb. 2010.
[89] P. Yan and X. Mou, "Video quality assessment based on correlation between spatiotemporal motion energies," *Proc. SPIE*, vol. 9971, p. 9971, 2016.
[90] A. Srivastava, A. B. Lee, E. P. Simoncelli, and S.-C. Zhu, "On advances in statistical modeling of natural images," *J. Math. Imag. Vis.*, vol. 18, no. 1, pp. 17–33, 2003.
[91] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.
[92] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.
[93] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[94] Z. Wang, E. P. Simoncelli, and A. C. Bovik, ‘‘Multiscale structural similarity for image quality assessment,’’ in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2003, pp. 1398–1402.
[95] Z. Wang and X. Shang, ‘‘Spatial pooling strategies for perceptual image quality assessment,’’ in *Proc. Int. Conf. Image Process.*, Oct. 2006, pp. 2945–2948.
[96] Z. Wang and Q. Li, ‘‘Video quality assessment using a statistical model of human visual speed perception,’’ *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 24, no. 12, pp. B61–B69, 2007.
[97] A. K. Moorthy and A. C. Bovik, ‘‘Visual importance pooling for image quality assessment,’’ *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 193–201, Apr. 2009.
[98] P. Tao and A. Eskicioglu, ‘‘Video quality assesment using M-SVD,’’ *Proc. SPIE*, vol. 6494, p. 6494, 2007.
[99] S. Kanumuri, P. C. Cosman, A. R. Reibman, and V. A. Vaishampayan, ‘‘Modeling packet-loss visibility in MPEG-2 video,’’ *IEEE Trans. Multimedia*, vol. 8, no. 2, pp. 341–355, Apr. 2006.
[100] T.-L. Lin, S. Kanumuri, Y. Zhi, D. Poole, P. C. Cosman, and A. R. Reibman, ‘‘A versatile model for packet loss visibility and its application to packet prioritization,’’ *IEEE Trans. Image Process.*, vol. 19, no. 3, pp. 722–735, Mar. 2010.
[101] S. Paluri, K. K. R. Kambhatla, B. A. Bailey, P. C. Cosman, J. D. Matyjas, and S. Kumar, ‘‘A low complexity model for predicting slice loss distortion for prioritizing H.264/AVC video,’’ *Multimedia Tools Appl.*, vol. 75, no. 2, pp. 961–985, 2016.
[102] F. Tommasi, V. De Luca, and C. Melle, ‘‘Packet losses and objective video quality metrics in H.264 video streaming,’’ *J. Vis. Commun. Image Represent.*, vol. 27, pp. 7–27, Feb. 2015.
[103] A. R. Reibman and D. Poole, ‘‘Predicting packet-loss visibility using scene characteristics,’’ in *Proc. Packet Video*, Nov. 2007, pp. 308–317.
[104] M. A. Aabed and G. AlRegib, ‘‘Reduced-reference perceptual quality assessment for video streaming,’’ in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 2394–2398.
[105] I. P. Gunawan and M. Ghanbari, ‘‘Reduced-reference picture quality estimation by using local harmonic amplitude information,’’ in *Proc. London Commun. Symp.*, 2003, pp. 353–358.
[106] D. Tao, X. Li, W. Lu, and X. Gao, ‘‘Reduced-reference IQA in contourlet domain,’’ *IEEE Trans. Syst. Man, Cybern. B, Cybern.*, vol. 39, no. 6, pp. 1623–1627, Dec. 2009.
[107] M. Carnec, P. Le Callet, and D. Barba, ‘‘Objective quality assessment of color images based on a generic perceptual reduced reference,’’ *Signal Process., Image Commun.*, vol. 23, no. 4, pp. 239–256, Apr. 2008.
[108] Z. Lu, W. Lin, E. Ong, X. Yang, and S. Yao, ‘‘PSQM-based RR and NR video quality metrics,’’ *Vis. Commun. Image Process.*, vol. 5150, pp. 633–640, Jul. 2003.
[109] B. Karacali and A. S. Krishnakumar, ‘‘Measuring video quality degradation using face detection,’’ in *Proc. 35th IEEE Sarnoff Symp. (SARNOFF)*, May 2012, pp. 1–5.
[110] W. Zhou, G. Jiang, M. Yu, F. Shao, and Z. Peng, ‘‘Reduced-reference stereoscopic image quality assessment based on view and disparity zero-watermarks,’’ *Signal Process., Image Commun.*, vol. 29, no. 1, pp. 167–176, 2014.
[111] M. Yu and K. Zheng, ‘‘Binocular perception based reduced-reference stereo video quality assessment method,’’ *J. Vis. Commun. Image Represent.*, vol. 38, pp. 246–255, Jul. 2016.
[112] Z. Wang and E. P. Simoncelli, ‘‘Reduced-reference image quality assessment using a wavelet-domain natural image statistic model,’’ *Electron. Imag.*, vol. 56, pp. 149–159, Jan. 2005.
[113] R. Soundararajan and A. C. Bovik, ‘‘RRED indices: Reduced reference entropic differencing framework for image quality assessment,’’ in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 1149–1152.
[114] A. Abdelouahad, M. El Hassouni, H. Cherifi, and D. Aboutajdine, ‘‘Image quality assessment measure based on natural image statistics in the tetrolet domain,’’ in *Proc. 5th Int. Conf. Image Signal Process.*, 2012, pp. 451–458.
[115] L. Ma, S. Li, and K. N. Ngan, ‘‘Reduced-reference image quality assessment in reorganized DCT domain,’’ *Signal Process, Image Commun.*, vol. 28, no. 8, pp. 884–902, 2013.
[116] M. Liu, K. Gu, G. Zhai, P. Le Callet, and W. Zhang, ‘‘Perceptual reduced-reference visual quality assessment for contrast alteration,’’ *IEEE Trans. on Broadcasting*, vol. 63, no. 1, pp. 71–81, Mar. 2017.
[117] Y. Zhang, J. Wu, G. Shi, and X. Xie, ‘‘Reduced-reference image quality assessment based on entropy differences in DCT domain,’’ in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2015, pp. 2796–2799.
[118] S. Golestaneh and L. J. Karam, ‘‘Reduced-reference quality assessment based on the entropy of DWT coefficients of locally weighted gradient magnitudes,’’ *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5293–5303, Nov. 2016.
[119] C. Chen, M. Izadi, and A. Kokaram, ‘‘A no-reference perceptual quality metric for videos distorted by spatially correlated noise,’’ *ACM Multimedia*, vol. 45405, pp. 1277–1285, Oct. 2016.
[120] X. Marichal, W. Y. Ma, and H. Zhang, ‘‘Blur determination in the compressed domain using DCT information,’’ in *Proc. Int. Conf. Image Process.*, 1999, pp. 386–390.
[121] D. Liu, Z. Chen, H. Ma, F. Xu, and X. Gu, ‘‘No reference block based blur detection,’’ in *Proc. Int. Workshop Quality Multimedia Exper.*, 2009, pp. 75–80.
[122] C. Chen, W. Chen, J. A. Bloom, ‘‘A universal reference-free blurriness measure,’’ *Proc. SPIE*, vol. 7867, p. 78670B, Jan. 2011.
[123] C. Chen and J. Bloom, ‘‘A blind reference-free blockiness measure,’’ in *Proc. 11th Pacific Rim Conf. Adv. Multimedia Inf. Process., I*, 2010, pp. 112–123.
[124] H. Liu and I. Heynderickx, ‘‘A no-reference perceptual blockiness metric,’’ in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2008, pp. 865–868.
[125] H. Tao, N. Klomp, and I. Heynderickx, ‘‘A no-reference metric for perceived ringing artifacts in images,’’ *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 4, pp. 529–539, Apr. 2010.
[126] A. Amer and E. Dubois, ‘‘Fast and reliable structure-oriented video noise estimation,’’ *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 1, pp. 113–118, Jan. 2005.
[127] S. Borer, ‘‘A model of jerkiness for temporal impairments in video transmission,’’ in *Proc. Int. Workshop Quality Multimedia Exper. (QoMEX)*, 2010, pp. 218–223.
[128] R. R. Pastrana-Vidal and J. C. Gicquel, ‘‘Automatic quality assessment of video fluidity impairments using a no-reference metric,’’ in *Proc. Int. Workshop Video Process. QualityMetrics Consum. Electron.*, 2006, pp. 1–6.
[129] T. Oelbaum, C. Keimel, and K. Diepold, ‘‘Rule-based no-reference video quality evaluation using additionally coded videos,’’ *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 294–303, Apr. 2009.
[130] P. Romaniak, L. Janowski, M. Leszczuk, and Z. Papir, ‘‘Perceptual quality assessment for H.264/AVC compression,’’ in *Proc. IEEE Consumer Commun. Netw. Conf. (CCNC)*, Jan. 2012, pp. 597–602.
[131] X. Liu, M. Chen, T. Wan, and C. Yu, ‘‘Hybrid no-reference video quality assessment focusing on codec effects,’’ *Trans. Internet Inf. Syst.*, vol. 5, no. 3, pp. 592–606, 2011.
[132] R. R. Pastrana-Vidal and J. C. Gicquel, ‘‘A no-reference video quality metric based on a human assessment model,’’ in *Proc. Int. Workshop Video Process. Quality Metrics Consumer Electron.*, 2007, pp. 1–8.
[133] J. Shen, Q. Li, and G. Erlebacher, ‘‘Hybrid no-reference natural image quality assessment of noisy, blurry, JPEG2000, and JPEG images,’’ *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2089–2098, Aug. 2011.
[134] M. A. Saad and A. C. Bovik, ‘‘Natural motion statistics for no-reference video quality assessment,’’ in *Proc. Int. Workshop Quality Multimedia Exper.*, 2009, pp. 163–167.
[135] M.-J. Chen, L. K. Cormack, and A. C. Bovik, ‘‘No-reference quality assessment of natural stereopairs,’’ *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3379–3391, Sep. 2013.
[136] A. Maalouf and M. C. Larabi, ‘‘A no-reference color video quality metric based on a 3D multispectral wavelet transform,’’ in *Proc. Int. Workshop Quality Multimedia Exper. (QoMEX)*, 2010, pp. 11–16.
[137] Y. Li *et al.*, ‘‘No-reference video quality assessment with 3D shearlet transform and convolutional neural networks,’’ *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 6, pp. 1044–1057, Jun. 2016.
[138] M. Ries, C. Crespi, O. Nemethova, and M. Rupp, ‘‘Content based video quality estimation for H.264/AVC video streaming,’’ in *Proc. IEEE Wireless Commun. Netw. Conf.*, Mar. 2007, pp. 2668–2673.
[139] A. Khan, L. Sun, and E. Ifeachor, ‘‘Content-based video quality prediction for MPEG4 video streaming over wireless networks,’’ *J. Multimedia*, vol. 4, no. 4, pp. 228–239, 2009.
[140] G. Valenzise, S. Magni, M. Tagliasacchi, and S. Tubaro, ‘‘No-reference pixel video quality monitoring of channel-induced distortion,’’ *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 4, pp. 605–618, Apr. 2012.

[141] A. Raake *et al.*, "T-V-model: Parameter-based prediction of IPTV quality," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar./Apr. 2008, pp. 1149–1152.
[142] J. Han, Y.-H. Kim, J. Jeong, and J. Shin, "Video quality estimation for packet loss based on no-reference method," in *Proc. Int. Conf. Adv. Commun. Technol.*, 2010, pp. 418–421.
[143] T. Yamada, S. Yachida, Y. Senda, and M. Serizawa, "Accurate video-quality estimation without video decoding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 2426–2429.
[144] I. Sedano, K. Brunnström, M. Kihl, and A. Aurelius, "Full-reference video quality metric assisted the development of no-reference bitstream video quality metrics for real-time network monitoring," *EURASIP J. Image Video Process.*, vol. 2014, p. 4, Dec. 2014.
[145] N. Staelens *et al.*, "No-reference bitstream-based visual quality impairment detection for high definition H.264/AVC encoded video sequences," *IEEE Trans. Broadcast.*, vol. 58, no. 2, pp. 187–199, Jun. 2012.
[146] F. Yang, S. Wan, Q. Xie, and H. R. Wu, "No-reference quality assessment for networked video via primary analysis of bit stream," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 11, pp. 1544–1554, Nov. 2010.
[147] M. Chin, T. Brandão, and M. P. Queluz, "Bitstream-based quality metric for packetized transmission of H.264 encoded video," in *Proc. Int. Conf. Syst., Signals Image Process. (IWSSIP)*, 2012, pp. 312–315.
[148] S.-Y. Shim, J.-H. Moon, and J. K. Han, "PSNR estimation scheme using coefficient distribution of frequency domain in H.264 decoder," *Electron. Lett.*, vol. 44, no. 2, pp. 108–109, Jan. 2008.
[149] T. Yamada, Y. Miyamoto, and M. Serizawa, "No-reference video quality estimation based on error-concealment effectiveness," in *Proc. Packet Video*, 2007, pp. 288–293.
[150] S. Zhao, H. Jiang, Q. Cai, S. Sherif, and A. Tarraf, "Hybrid framework for no-reference video quality indication over LTE networks," in *Proc. 23rd Wireless Opt. Commun. Conf. (WOCC)*, 2014, pp. 1–5.
[151] R. K. P. Mok, E. W. W. Chan, X. Luo, and R. K. C. Chang, "Inferring the QoE of HTTP video streaming from user-viewing activities," in *Proc. ACM SIGCOMM Workshop Meas. Up Stack*, 2011, pp. 31–36.
[152] A. Ichigaya, Y. Nishida, and E. Nakasu, "Nonreference method for estimating PSNR of MPEG-2 coded video by using DCT coefficients and picture energy," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 6, pp. 817–826, Jun. 2008.
[153] A. Ichigaya, M. Kurozumi, N. Hara, Y. Nishida, and E. Nakasu, "A method of estimating coding PSNR using quantized DCT coefficients," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 2, pp. 251–259, Feb. 2006.
[154] J. Beerends and F. De Caluwe, "The influence of video quality on perceived audio quality and vice versa," *J. Audio Eng. Soc*, vol. 47, no. 5, pp. 355–362, 1999.
[155] B. Benjamin, S. Moeller, B. Lewcio, A. Raake, and A. Mehmood, "Audio and video channel impact on perceived audio-visual quality in different interactive contexts," in *Proc. IEEE Int. Workshop Multimedia Signal Process.*, Oct. 2009, pp. 1–5.
[156] S. Winkler and C. Faller, "Perceived audiovisual quality of low-bitrate multimedia content," *IEEE Trans. Multimedia*, vol. 8, no. 5, pp. 973–980, Oct. 2006.
[157] A. N. Rimell, M. P. Hollier, and R. M. Voelcker, "The influence of cross-modal interaction on audio-visual speech quality perception," in *Proc. Audio Eng. Soc. Conv.*, 1998, p. 4791.
[158] *Study of the Influence of Experimental Context on the Relationship Between Audio, Video and Audiovisual Subjective Qualities*, document ITU-T Contribution COM 12-61-E, 1998.
[159] R. Eg, C. Griwodz, and P. Halvorsen, "Audiovisual robustness: Exploring perceptual tolerance to asynchrony and quality distortion," *Multimedia Tools Appl.*, vol. 74, no. 2, pp. 345–365, 2015.
[160] M. N. Garcia, R. Schleicher, and A. Raake, "Impairment-factor-based audiovisual quality model for IPTV: Influence of video resolution, degradation type, and content type," *EURASIP J. Image Video Process.*, vol. 2011, Dec. 2011, Art. no. 629284.
[161] J. Song, F. Yang, Y. Zhou, S. Wan, and H. R. Wu, "QoE evaluation of multimedia services based on audiovisual quality and user interest," *IEEE Trans. Multimedia*, vol. 18, no. 3, pp. 444–457, Mar. 2016.
[162] H. E. Pashler, *The Psychology of Attention*. Cambridge, MA, USA: MIT Press, 1998.
[163] H. Martinez and M. Farias, "Full-reference audio-visual video quality metric," *J. Electron. Imag.*, vol. 23, no. 6, p. 061108, 2014.
[164] I. Saidi, L. Zhang, V. Barriac, and O. Deforges, "Audiovisual quality study for videotelephony on IP networks," in *Proc. IEEE Workshop Multimedia Signal Process.*, Sep. 2016, pp. 1–6.
[165] T. Hayashi, K. Yamagishi, T. Tominaga, and A. Takahashi, "Multimedia quality integration function for videophone services," in *Proc. IEEE Global Telecommun. Conf.*, Nov. 2007, pp. 2735–2739.
[166] E. Demirbilek and J. C. Grégoire, "Towards reduced reference parametric models for estimating audiovisual quality in multimedia services," in *Proc. IEEE Int. Conf. Commun.*, May 2016, pp. 1–6.
[167] K. R. Laghari *et al.*, "Objective characterization of human behavioural characteristics for QoE assessment: A pilot study on the use of electroencephalography features," in *Proc. IEEE Globecom Workshops*, Dec. 2013, pp. 1168–1173.
[168] J. Skowronek and A. Raake, "Investigating the effect of number of interlocutors on the quality of experience for multi-party audio conferencing," in *Proc. 12th Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, 2011, pp. 829–832.
[169] A. Bokani, M. Hassan, and S. Kanhere, "HTTP-based adaptive streaming for mobile clients using Markov decision process," in *Proc. 20th Int. Packet Video Workshop*, 2013, pp. 1–8.
[170] S. Wang and S. Dey, "Cloud mobile gaming: Modeling and measuring user experience in mobile wireless networks," *SIGMOBILE Mobile Comp. Commun. Rev.*, vol. 16, no. 1, pp. 10–21, 2012.
[171] M. Fiedler, T. Hossfeld, and P. Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service," *IEEE Netw.*, vol. 24, no. 2, pp. 36–41, Mar./Apr. 2010.
[172] J. Shaikh, M. Fiedler, and D. Collange, "Quality of Experience from user and network perspectives," *Ann. Telecommun.-Ann. Télécommun.*, vol. 65, no. 1, pp. 47–57, 2010.
[173] C. Alberti *et al.*, "Automated QoE evaluation of dynamic adaptive streaming over HTTP," in *Proc. 5th Int. Workshop Quality Multimedia Exper. (QoMEX)*, 2013, pp. 58–63.
[174] K. ur R. Laghari, O. Issa, F. Speranza, and T. H. Falk, "Quality-of-experience perception for video streaming services: Preliminary subjective and objective results," in *Proc. Asia Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2012, pp. 1–9.
[175] K. ur R. Laghari *et al.*, "Neurophysiological experimental facility for quality of experience (QoE) assessment," in *Proc. IFIP/IEEE Int. Symp. Integr. Netw. Manage.*, May 2013, pp. 1300–1305.
[176] R. Gupta, K. Laghari, H. Banville, and T. H. Falk, "Using affective brain-computer interfaces to characterize human influential factors for speech quality-of-experience perception modelling," *Hum.-Centric Comput. Inf. Sci.*, vol. 6, no. 1, pp. 1–19, 2016.
[177] T. H. Falk, Y. Pomerantz, K. Laghari, S. Möller, and T. Chau, "Preliminary findings on image preference characterization based on neurophysiological signal analysis: Towards objective QoE modeling," in *Proc. 4th Int. Workshop Quality Multimedia Exper.*, 2012, pp. 146–147.
[178] S. Arndt, J. N. Antons, R. Schleicher, S. Möller, and G. Curio, "Perception of low-quality videos analyzed by means of electroencephalography," in *Proc. 4th Int. Workshop Quality Multimedia Exper.*, 2012, pp. 284–289.
[179] A.-N. Moldovan, I. Ghergulescu, S. Weibelzahl, and C. H. Muntean, "User-centered EEG-based multimedia quality assessment," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, Jun. 2013, pp. 1–8.
[180] J. Antons, B. Blankertz, G. Curio, S. Möller, A. K. Porbadnigk, and R. Schleicher, "Subjective listening tests and neural correlates of speech degradation in case of signal-correlated noise," in *Proc. Audio Eng. Soc. Conv.*, 2010, p. 8206.
[181] G. Ghinea and J. P. Thomas, "Quality of perception: User quality of service in multimedia presentations," *IEEE Trans. Multimedia*, vol. 7, no. 4, pp. 786–789, Aug. 2005.
[182] S. R. Gulliver and G. Ghinea, "Stars in their eyes: What eye-tracking reveals about multimedia perceptual quality," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 34, no. 4, pp. 472–482, Jul. 2004.
[183] R. T. Apteker, J. A. Fisher, V. S. Kisimov, and H. Neishlos, "Video acceptability and frame rate," *IEEE Multimedia Mag.*, vol. 2, no. 3, pp. 32–40, Sep./Nov. 1995.
[184] K. Kinoshita *et al.*, "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP J. Adv. Signal Process.*, vol. 2016, no. 1, p. 7, 2016.
[185] Z. Li, J. Wang, J. Cai, Z. Duan, H. Wang, and Y. Wang, "Non-reference audio quality assessment for online live music recordings," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 63–72.

[186] *Blizzard Challenge*. Accessed: Sep. 28, 2017. [Online]. Available: https://synsig.org/index.php/Blizzard_Challenge_2016
[187] Y. F. Ou, Y. Zhou, and Y. Wang, "Perceptual quality of video with frame rate variation: A subjective study," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 2446–2449.
[188] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, Jun. 2010.
[189] VQEG. (2010). *Report on the Validation of Video Quality Models for High Definition Video Content*. [Online]. Available: http://www.vqeg.org/
[190] W. Robitza, Y. Pitrey, M. Nezveda, S. Buchinger, and H. Hlavacs, "Made for mobile: A video database designed for mobile television," in *Proc. 6th Int. Workshop Video Process. Quality Metrics Consum. Electron. (VPQM)*, 2012, pp. 1–6.
[191] M. Nuutinen, T. Virtanen, M. Vaahteranoksa, T. Vuori, P. Oittinen, and J. Häkkinen, "CVD2014—A database for evaluating no-reference video quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3073–3086, Jul. 2016.
[192] M. Goudarzi, L. Sun, and E. Ifeachor, "Audiovisual quality estimation for video calls in wireless applications," in *Proc. IEEE Global Telecommun. Conf.*, Dec. 2010, pp. 1–5.
[193] C. Keimel, A. Redl, and K. Diepold, "The TUM high definition video datasets," in *Proc. 4th Int. Workshop Quality Multimedia Exper.*, 2012, pp. 97–102.
[194] T. Mäki, D. Kukolj, D. Đordević, and M. Varela, "A reduced-reference parametric model for audiovisual quality of IPTV services," in *Proc. 5th Int. Workshop Quality Multimedia Exper. (QoMEX)*, 2013, pp. 6–11.
[195] D.-S. Kim, "ANIQUE: An auditory model for single-ended speech quality estimation," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 821–831, Sep. 2005.
[196] R. Gupta, H. J. Banville, and T. H. Falk, "Multimodal physiological quality-of-experience assessment of text-to-speech systems," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 22–36, Feb. 2017.
[197] S. Koelstra *et al.*, " DEAP: A database for emotion analysis using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan./Mar. 2012.
[198] N. Rozenn *et al.*, "A roadmap for assessing the quality of experience of 3D audio binaural rendering," in *Proc. EAA Joint Symp. Auralization Ambisonics*, 2014, pp. 1–7.
[199] A.-F. Perrin, M. Řeřábek, and T. Ebrahimi, "Towards prediction of Sense of Presence in immersive audiovisual communications," *Electron. Imag.*, vol. 16, pp. 1–8, Feb. 2016.
[200] D. Egan, S. Brennan, J. Barrett, Y. Qiao, C. Timmerer, and N. Murray, "An evaluation of Heart Rate and ElectroDermal Activity as an objective QoE evaluation method for immersive virtual reality environments," in *Proc. Int. Conf. Quality Multimedia Exper.*, 2016, pp. 1–6.
[201] U. Engelke, M. Barkowsky, P. Le Callet, and H.-J. Zepernick, "Modelling saliency awareness for objective video quality assessment," in *Proc. Int. Workshop Quality Multimedia Exper. (QoMEX)*, 2010, pp. 212–217.
[202] D. Ćulibrk, M. Mirkovic, V. Zlokolica, M. Pokric, V. Crnojevic, and D. Kukolj, "Salient motion features for video quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 4, pp. 948–958, Apr. 2010.

ZAHID AKHTAR received the Ph.D. degree in electronic and computer engineering from the University of Cagliari, Italy. He is currently a Post-Doctoral Researcher with the INRS-EMT Center, University of Quebec, Montreal, Canada. His research interests include computer vision, pattern recognition, and image processing with applications in biometrics, affective computing, security systems, and multimedia quality assessment. He is a member of the IEEE Signal Processing Society.

TIAGO H. FALK (SM'14) received the B.Sc. degree from the Federal University of Pernambuco, Brazil, in 2002, and the M.Sc. and Ph.D. degrees from Queen's University, Canada, in 2005 and 2008, respectively, all in electrical engineering. In 2010, he joined INRS, Montreal, Canada, in 2010, where he is currently an Associate Professor and heads the Multimedia/Multimodal Signal Analysis and Enhancement Laboratory. His research interests include multimedia/biomedical signal analysis and enhancement, pattern recognition, and their interplay in the development of biologically inspired technologies.