Learning a Continuous-Time Streaming Video QoE Model

Deepti Ghadiyaram, *Student Member, IEEE*, Janice Pan, and Alan C. Bovik, *Fellow, IEEE*

***Abstract*— Over-the-top adaptive video streaming services are frequently impacted by fluctuating network conditions that can lead to rebuffering events (stalling events) and sudden bitrate changes. These events visually impact video consumers' quality of experience (QoE) and can lead to consumer churn. The development of models that can accurately predict viewers' instantaneous subjective QoE under such volatile network conditions could potentially enable the more efficient design of quality-control protocols for media-driven services, such as YouTube, Amazon, Netflix, and so on. However, most existing models only predict a single overall QoE score on a given video and are based on simple global video features, without accounting for relevant aspects of human perception and behavior. We have created a QoE evaluator, called the time-varying QoE Indexer, that accounts for interactions between stalling events, analyzes the spatial and temporal content of a video, predicts the perceptual video quality, models the state of the client-side data buffer, and consequently predicts continuous-time quality scores that agree quite well with human opinion scores. The new QoE predictor also embeds the impact of relevant human cognitive factors, such as memory and recency, and their complex interactions with the video content being viewed. We evaluated the proposed model on three different video databases and attained standout QoE prediction performance.**

***Index Terms*— Quality of Experience, subjective video quality assessment, continuous-time QoE, stalling events, network impairments, mobile video quality.**

I. Introduction

CAPTURING, storing, sharing, and streaming of digital visual media continues to experience explosive growth in online applications of social media, entertainment, medicine, geoscience, transportation security, and e-commerce. An increasing number of video entertainment services are being offered by such major video content providers as YouTube, Netflix, HBO, and Amazon Video. On social media, video-centric mobile applications such as Facebook LIVE, Snapchat, Periscope, Google Hangouts, and Instagram LIVE are also becoming increasingly popular. Given the ubiquitous availability of portable mobile devices for video capture and access, there has been a dramatic shift towards over-the-top (OTT) video streaming and sharing of videos via social media websites and mobile applications. This growing consumption of visual media is fueling the demand for high quality video streaming on different viewing platforms having varied bandwidth capabilities and display resolutions.

As a way to account for the performance of these media-centric applications, finding ways to measure and maximize an end user's quality of experience (QoE) [2] is gaining attention among content and mobile providers. QoE in this context refers to a consumer's holistic perception and satisfaction with a given content or communication network service. Media streaming services typically employ cloud video transcoding systems, leveraging HTTP-based adaptive streaming protocols such as Dynamic Adaptive Streaming over HTTP (DASH) [3] and HTTP Live Streaming (HLS) [4] to make video delivery scalable and adaptable to the available network bandwidth. Under such protocols, videos are typically divided into *segments* (of fixed duration), where each video segment is encoded at multiple bitrates and resolutions (also called *video levels*). A stream-switching controller designed at either the server side [5] or the client side [6]–[9] *adaptively* predicts (and then requests) an "optimal" video level depending on such factors as the requester's device, the client-side data buffer occupancy, and the current network conditions. Under volatile network conditions, the controller may request video segments of varied bitrates interspersed with stalling events, thus potentially causing viewer annoyance. Some examples of video frames afflicted with combinations of compression and stalling artifacts are shown in Fig. 1.

There has been some progress made on the design of intelligent controllers that aim to reduce the number of stalls and bitrate switches; however, none of these measure an end user's continuous time-varying or overall perceived QoE. While aiming to reduce the number of stalls and bitrate switches is a reasonable approach to reduce viewer annoyance, it does not account for a viewer's time-varying QoE. A user's perceived QoE at any given instant is greatly influenced by the complex interplay of video content, the number and frequency of rebuffering events, rebuffering lengths, rebuffering locations within a video, and fluctuating bitrates, as well as by cognitive factors such as memory and recency. Being able to quickly and accurately predict the *instantaneous* viewing experience of an end user *objectively* by consolidating all of the aforementioned factors could supply crucial feedback to stream-switching

Manuscript received March 18, 2017; revised October 18, 2017 and December 12, 2017; accepted December 12, 2017. Date of publication January 5, 2018; date of current version February 12, 2018. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Amit K. Roy Chowdhury. *(Corresponding author: Deepti Ghadiyaram.)*

D. Ghadiyaram is with Facebook Inc, Menlo Park, CA 94025 USA (e-mail: deeptigp9@gmail.com).

J. Pan and A. C. Bovik are with the Laboratory for Image and Video Engineering, Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX 78712 USA (e-mail: janicespan@gmail.com; bovik@ece.utexas.edu).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author. The material presents paired sample t-tests that were conducted between SROCC values obtained from different QoE predictors (multi-learner, multi-stage, and global) on three different QoE databases. The total size of the file is 0.0802 MB. Contact deeptigp9@gmail.com for further questions about this work.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2018.2790347

Fig. 1. Sample stalled frames from videos in the LIVE Mobile Stall Video Database-II [1] encoded at different bitrates.

algorithms (at either the client or the server side). Such objective QoE predictors could also serve as an enabling step towards the design of stream-switching algorithms that can efficiently balance the tradeoffs between network operational costs and delivering videos with the highest possible quality to customers.

A. Motivation for a Continuous-Time Quality Predictor

Most existing objective models extract global stall-informative features, such as cumulative stall length and number of stalls, to train an overall QoE predictor [10]–[12]. However, the natural temporal information of video contents and stalling events, which have a crucial effect on user QoE, are not effectively captured by such global statistics. Furthermore, perceived quality also depends on a behavioral hysteresis or recency "aftereffect," whereby a user's QoE at a particular moment also depends on their viewing experiences preceding that moment [13]. For example, in the context of QoE, the memory of an early unpleasant viewing experience caused by a stalling event may negatively impact future QoE and, thus, may also negatively impact the overall QoE. A long initial delay (of length L, for example) at the beginning of a video sequence may more likely to lead to viewer abandonment, than when viewing the same video content containing multiple stalls whose total length equals L. Additionally, a stalling event occurring towards the end of a video sequence could have a more negative impact on the final overall perception of video quality, than a stall of the same length occurring at an earlier position in the same video. This dependency on previous viewing experiences is generally nonlinear and can be crucial in determining both the overall as well as the instantaneous QoE of viewers [13], but this information is not currently being exploited by contemporary QoE prediction models.

To study these effects, we have developed an objective, no-reference, continuous-time QoE predictor that we call the **T**ime-**V**arying **QoE** (TV-QoE) Indexer for processing streaming videos afflicted by stalling events and quality variations. To tackle this difficult problem, we sought to solve several sub-problems simultaneously, including how factors such as stalling event properties, the data buffer state, and video content impact an end user's QoE. Driven by these factors, we define and extract several useful distortion-informative features, and model them as continuous-time inputs. These are used to design an ensemble of Hammerstein-Wiener (HW) models [14] that serves as an integral component of our QoE model. Towards effectively modeling the joint impact of the aforementioned factors on QoE, we strategically combine this ensemble of models and accurately predict the continuous-time QoE scores of streaming videos.

B. Contributions

To thoroughly understand and model the effect of several video quality-influencing factors on QoE, we summarize our contributions below:

1) First, we describe the comprehensive set of continuous-time, stall-informative, video content-informative, and perceptual quality-informative inputs that we derive from distorted videos. These inputs contain useful evidence descriptive of the effects of stalls and quality degradations on the time-varying QoE of a streaming video (Sec. IV-A).
2) We mathematically model the dynamics of a client-side data buffer that takes into account the variations in the bitrates at which the streaming video segments are encoded, as well as the instantaneous network throughput. This dynamic model serves as a valuable indicator of the fluctuations in perceived QoE due to stalling events (Sec. IV-B).
3) We employ a Hammerstein-Wiener model that effectively captures the hysteresis effects that contribute to QoE using a *linear filter*. Further, it also accounts for the nonlinearity of human behavioral responses using nonlinear functions at the input of the linear filter (Fig. 8). Each distortion-informative input is independently used to train a HW model with memory, resulting in an ensemble of HW models (Sec. V).
4) We fuse the predictions of these individual HW models by employing and comparing two different strategies: (a) a multi-stage approach, in which the Hammerstein-Wiener models are concatenated, such that the predictions from the learners at one stage are supplied as inputs to another HW Model at a subsequent stage and (b) a multi-learner approach, in which the predictions of the individual HW models are used to train a different learner, called the *meta-learner* [15] (Sec. V). These two predictors are independently trained to predict continuous-time QoE scores.
5) To address the side problem of predicting the overall perception of the quality of experience after viewing a video, we derive useful global statistics from our comprehensive set of continuous-time inputs, and we also design a global overall QoE predictor (Sec. VI).

6) We evaluate our QoE predictors (multi-stage, multi-learner, and global) on three different video QoE databases, and analyze the performances of the proposed models when different amounts and types of information about the test video are available.

Our experiments show that the proposed global and continuous-time QoE predictors effectively capture the contributions of several QoE-influencing features and subjective effects such as memory and recency.

II. Related Work

A. VQA Databases Modeling Stalling Events

Subjective and objective video quality assessment is a very active area of research, and a number of popular public-domain subjective video quality databases [16]–[20] have been designed in the past decade. The videos in these data collections model different post-capture and in-capture spatial and temporal distortions, such as compression, transmission errors, frame freezes, artifacts due to exposure and lens limitations, focus distortions, and color aberrations. However, these databases do not model network-induced distortions, such as start-up delays and stalling events, or combinations of stalling events and compression effects. A few video quality studies have been conducted in the recent past to analyze the effects of network streaming quality on QoE. The only openly available databases are the LIVE Mobile Stall Video Database-I [21], [22] and the Waterloo Quality-of-Experience Database [23], which contain a wide-variety of video contents (174 and 200 respectively) and network-induced impairments. However, these databases only capture a single overall QoE score for each video. The newly designed LIVE Mobile Stall Video Database-II [1], on the other hand, contains per-frame mean QoE scores, in addition to overall QoE scores obtained via a subjective study using all 174 publicly available videos. The database presented in [24] also models network-induced impairments and captures continuous-time QoE scores; however, only 24 out of the 112 video contents are publicly available.

B. Automatic QoE Predictors

Top-performing global and continuous-time video quality predictors [25]–[31] that have been developed in the past decade deal with post-processing distortions but not network-induced impairments, and they cannot capture the impact of stalling events interspersed with bitrate variations. A number of objective QoE predictors have been designed [10]–[12], [32]–[37]. Some of these methods derive global video statistics and are also based on the total stall length and on the number of random video stalls. However, these models only make global measurements, and therefore do not capture the time-varying levels of satisfaction experienced when viewing streaming videos.

The DQS model [38] also considers global stall statistics and a linear model to predict a continuous-time QoE score. Specifically, this model defines three events: start-up delay, first rebuffering, and multiple rebuffering (explicitly) based on empirical observations on the final QoE scores of the LIVE Mobile Stall Video Database-I [22]. The underlying assumption of the DQS model is that an end user's QoE is driven by these predefined events, and different model parameters are chosen to determine the contribution of each event to the model's quality prediction. Thus, the generalizability of the DQS model to more diverse stall patterns is questionable.

The recently proposed SQI model [23] combines perceptual video presentation quality and simple stalling event-based features to predict QoE. In a preliminary model detailed in [39], we modeled four inputs based purely on stalling events, using them to train a single Hammerstein-Wiener model to predict continuous-time perceptual quality. As we describe next, we have greatly expanded the suite of QoE-sensitive inputs to the model. We derived a total of 6 stall-based inputs based on our insights from collecting continuous-time QoE scores [1]. These include the output of a model of the client-side data buffer state, measurements of spatial and temporal video complexity, and predictions delivered by a perceptual video quality algorithm. We use these to create two separate continuous-time QoE predictors and one global QoE predictor, which we evaluate on all existing video QoE databases.

III. LIVE Mobile Stall Video Database-II

We briefly describe the key characteristics of the LIVE Mobile Stall Video Database-II [1] that is used in this work. As mentioned earlier, this new database contains 174 test video sequences modeling 26 diverse stall patterns. There are 24 reference videos of varied spatial and temporal complexities, with mobile-focused video resolutions, which were used to construct the test sequences in the dataset (Figure 1). We conducted a subjective study in a calibrated study setting gathering per-frame subjective opinion scores from about 27 unique viewers per test video. This rich data proved to be a valuable resource that allowed us to analyze and better understand various aspects of the effects of rebuffering events on end users' QoE, such as their lengths, their frequency of occurrence, and their location within videos. We refer the reader to [1] for more details regarding our comprehensive analysis of subjective behavior that occurs when viewing videos impaired by stalling events.

IV. Modeling Continuous-Time Inputs for QoE Prediction

With a goal to model the effects of stalls as well as video content, distortion, and other factors on QoE, we designed a number of stall-informative and content-informative input channels that we describe next. We list the set of measurements that our model relies on, along with some brief descriptive comments of each, in Table I.

A. Video Stall-Driven Inputs

1) Stall Length: One of our inputs ($u_1[t]$) is designed to capture the impact of stall lengths on QoE. Given a video, if $s_1[t]$ denotes the length of a stall at a discrete time instance t, then let

$$u_1[t] = e^{\alpha_1 s_1[t]} - 1, \tag{1}$$

TABLE I
DESCRIPTION OF THE PROPOSED DYNAMIC INPUTS

Dynamic Input	Brief Description
Stall Length	The duration of the stall.
Number of stalls	-
Time since previous stall	Time that has elapsed since a stall has ended (or) the time the user had to recover from an unpleasant stall experience.
Frequency of stalls	Density of stall occurring in a video.
Rebuffering Rate	The fraction of stall time in a given video.
Client-side buffer model	Models the varying dynamics of a client-side media buffer.
Scene criticality	A measure of the video content complexity [40]
Perceptual Quality	Per-frame image or video quality metrics

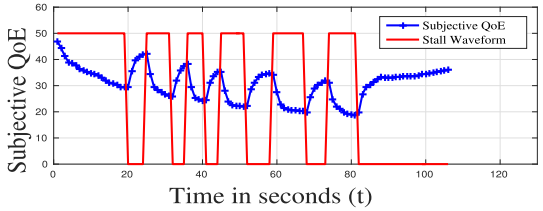


Fig. 2. An example video sequence with 6 stalling events from the LIVE Mobile Stall Video Database-II [1], where the stall waveform (in red) is overlaid on the average of the temporal subjective QoE scores from each subject (in blue). For the purpose of illustration, a value of 0 in the stall waveform indicates normal video playback, while a value of 50 in the stall waveform indicates a stalling event.

where α_1 is a scalar chosen via cross-validation (Sec. VII). The choice of a nonlinear exponential function to express the influence of stall lengths on predicted QoE is motivated by the basic observation that viewer annoyance increases with rebuffering length [1]. Using a parameterized exponential makes it possible for TV-QoE to learn the steepness of the stall-length / annoyance relationship.

2) Total Number of Stalls: We also found from our analysis of the subjective data in [1] that, as the number of stalls increases, user annoyance increases monotonically, irrespective of the video content or duration. Further, as may be observed in the example in Fig. 2, perceived QoE tends to decrease with every stall occurrence. To capture the impact of the number of stalls on QoE, we defined another dynamic input

$$u_2[t] = e^{\alpha_2 s_2[t]} - 1, \tag{2}$$

where $s_2[t]$ is the total number of stalls up to a discrete time instance t. Again, using an exponential model makes it possible to capture a viewer's annoyance against the number of stalls. The parameter α_2 is also a scalar chosen via cross-validation (Sec. VII).

3) Time Since the Previous Stall: The next continuous-time input targets recency. Viewers generally react sharply to a stall occurrence, and as the period of time following the end of a stall increases, the viewer's perceived QoE may reflect improved satisfaction with the streaming video quality. However, immediately (and for some period of time) following a stall, the viewer's perceived QoE generally reflects heightened annoyance with the streaming quality. We found clear evidence for this behavior in the continuous-time subjective data obtained on the LIVE Mobile Stall Video Database-II [1]. Figure 2 also illustrates this behavior.

Thus, the third input to our model is the *time since the preceding rebuffering event* at every discrete instant t, with values of zero representing times during stalls. If $[T_{i,end}, T_{i+1,begin}]$ denotes the discrete time interval between the stall event (s_i) ending at time $T_{i,end}$ and the next stall event (s_{i+1}) starting at time $T_{i+1,begin}$,[1] then

$$u_3[t] = \begin{cases} t - T_{i,end} & if\ [T_{i,end} \leq t < T_{i+1,begin}] \\ 0 & if\ [T_{i,begin} \leq t < T_{i,end}]. \end{cases} \tag{3}$$

4) Frequency of Stalling Events: Next, we sought to define a model input that excludes the effects of stall lengths, instead capturing the interplay between the number of stalling events and the length of *video playback time up to a given time instant* $(p[t])$. Therefore, our next input captures the effects of the density of the stalls on QoE relative to the current moment. The frequency of stalling events $(u_5[t])$ at a given time t is given by

$$u_4[t] = \frac{p[t]}{s_2[t]}, \tag{4}$$

where $s_2[t]$ is the total number of stalls up to time t.

5) Rebuffering Rate: While the frequency of stalling events captures important interactions between video length, playback time, and the number of stalls, our next input focuses exclusively on the interplay between stall lengths and the length of playback up to a given time instant $(p[t])$. To motivate the construction of this input, consider a single, very long stalling event in a video of length 90 seconds. This event may impact QoE differently than would a relatively short stalling event in a 20-second video. To effectively account for this hypothesis,

[1] If there does not exist a stall event s_{i+1}, then $T_{i+1,begin}$ denotes the end of the video.

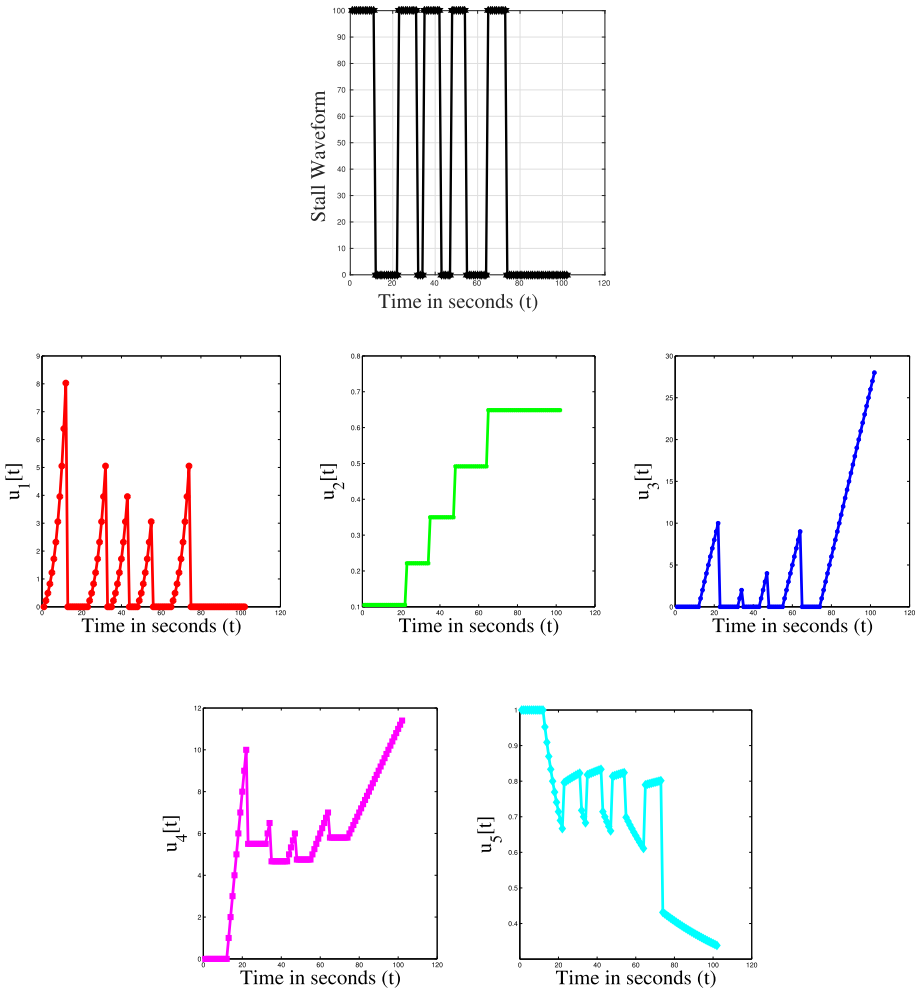


Fig. 3. (Top row) A sample test video impaired by intermittent stalling events. (Bottom two rows) Stall-descriptive continuous input waveforms computed from a video sequence as described in Sec. IV-A. The vertical axis labels the type of input. Best viewed in color.

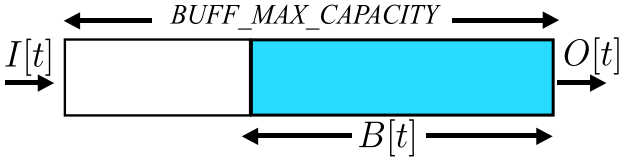


Fig. 4. Illustration of a possible client-side data buffer state.

we define the rate of rebuffering events as:

$$u_5[t] = \frac{r[t]}{r[t] + p[t]}, \tag{5}$$

where $r[t]$ is the total sum of stall lengths up to time t, and $p[t]$ is the playback time up to time instant t.

We illustrate the aforementioned inputs for a sample stall pattern in Fig. 3.

B. Modeling the Dynamics of the Client-Side Data Buffer

As previously mentioned, OTT services employ adaptive bitrate streaming algorithms, wherein the end-to-end network conditions are constantly monitored, and the bitrates of future video segments are chosen based on the current data buffer status of the client's media player, with a goal to minimize the occurrences of stalling events. In OTT streaming under constrained network conditions, the state of the data buffer varies dynamically, and a stream-switching controller constantly chooses either to request a lower bitrate video segment or to risk the possibility of stall occurrence. Thus, the dynamics of a data buffer have a direct impact on streaming video quality but are not being modeled in any existing QoE models [23], [38], [39].

Existing publicly available databases do not provide dynamic data buffer capacity information to accompany their spatially- or temporally-distorted videos, because the videos were constructed by manually inserting stalls. Hence we designed and used a simple model of a client-side data buffer, which we describe next. Note that, in the event that a media-streaming service or a QoE database can make available the actual buffer capacity trace, then it could be directly plugged into our QoE learner without needing to explicitly model the client's data buffer as we do below.

Assumptions and Notation: We make the following assumptions about the client-side data buffer, which are reflected in our mathematical model. A possible client-side data buffer is illustrated in Fig. 4.

- That a client-side data buffer is of a fixed size with a capacity measured in seconds. $BUFF\_MAX\_CAPACITY$ is

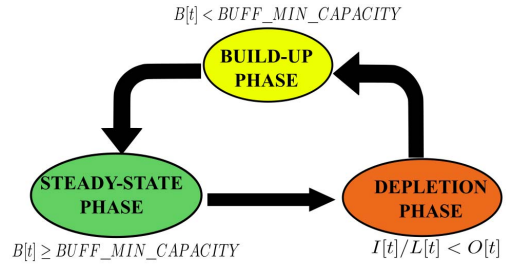


Fig. 5. Possible states of the client-side data buffer model.

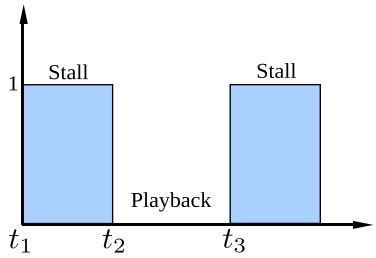


Fig. 6. Example video sequence with one stall between t_1 and t_2 and another stall at t_3. A value of 0 in this waveform indicates successful video playback, while a value of 1 indicates a stall event.

defined as the maximum amount of video content that can be stored in a buffer.

- Without loss of generality, that each video segment that is being adaptively transmitted from the media server is 1 second long.
- That the buffer occupancy builds and depletes exponentially. However, a different function can be easily applied in place of the exponential function.
- That $BUFF\_MIN\_CAPACITY = 1$, which is the minimum amount of video (in seconds) that should be present in the data buffer for it to be able to handle input bitrate variations. This quantity can also be understood as requiring the data buffer to contain at least one second's worth of video content in order to continue playback on the client's media player. If the buffer state does not satisfy this minimum requirement, the result on the client side would be the occurrence of a rebuffering event.
- $O[t]$ is the rate at which the video content leaves the buffer at time t. It can take one of the two possible values

$$O[t] = \begin{cases} 1, & \text{during playback} \\ 0, & \text{during stall} \end{cases} \tag{6}$$

  i.e., one second of video content leaves the data buffer during each second of playback, and no video content leaves during a playback interruption.
- Let $B[t]$ be the amount of buffer that is occupied with video content and $\Delta B[t]$ be the rate of change of the buffer occupancy at a given discrete time instant t.
- Let $L[t]$ denote the bitrate at which the incoming video segment is encoded, and $I[t]$ be the network throughput at time t.

At a given discrete time instant t, the rate of change of the buffer occupancy can be defined as follows [5]:

$$\Delta B[t] = \frac{I[t]}{L[t]} - O[t], \tag{7}$$

i.e., $\Delta B[t]$ is the difference between the amount of video (in seconds) that is entering the buffer and the amount of video (in seconds) that is leaving the buffer. Thus, variations in the buffer occupancy can be introduced due to changes in $I[t]$ or $L[t]$. When videos are encoded under a constant bitrate (CBR) regime, then $L[t]$ is fixed over the entire duration of the video sequence being streamed, which would not be the case for videos encoded under a variable bitrate (VBR) regime.

In the proposed model, the client-side data buffer can exist in one of the following three phases (illustrated in Fig. 5):

1) A **steady-state phase** where there is sufficient content in the data buffer to be transmitted to the client's media player at time t, i.e., $B[t] \geq BUFF\_MIN\_CAPACITY$.
2) A **build-up phase**, where the buffer builds up until it reaches $BUFF\_MIN\_CAPACITY$, i.e., until $B[t] < BUFF\_MIN\_CAPACITY$. In this phase, set $O[t] = 0$ until $I[t]/L[t] \approx BUFF\_MIN\_CAPACITY$. In other words, until the buffer contains at least one second of content, no amount of video content can leave the buffer, and thus, a rebuffering event occurs.
3) A **depletion phase** that occurs when $I[t]/L[t] < O[t]$. In this phase, the amount of data leaving the buffer is greater than the amount of data entering the buffer, which causes the buffer to slowly deplete until there is no more data to transmit.

We will now illustrate how a data buffer might transition from one state to another through a general scenario. Consider a video sequence $v[t]$ (illustrated in Fig. 6) such that

1) there is a stall event between times t_1 and t_2;
2) there is smooth continuous playback between times t_2 and t_3;
3) and there is another stall event that occurs after t_3.

A possible data buffer scenario can be described as follows (Fig. 7):

1) At a discrete time instant $t = t_1$, the buffer starts off empty, but it *must* enter the **build-up phase** before $t = t_2$ for playback to begin.
2) Next, the buffer *must* be in the **steady-state phase** for some time (t_s) between t_2 and t_3.
3) The buffer should next enter the **depletion phase** and *must* be completely empty at $t = t_3$, for a stall to occur at $t = t_3$.

We effectively model the different phases of the buffer as follows:

1) **Modeling the buildup phase (between t_1 and t_2):**
   - We *randomly* sample a discrete time instant t_b between times t_1 and t_2.
   - We fit an exponential function between data points (t_b, 0) and (t_2, $BUFF\_MIN\_CAPACITY$).
2) **Modeling the depletion phase (between t_2 and t_3):**
   - We *randomly* sample a discrete time instance t_d between times t_2 and t_3.

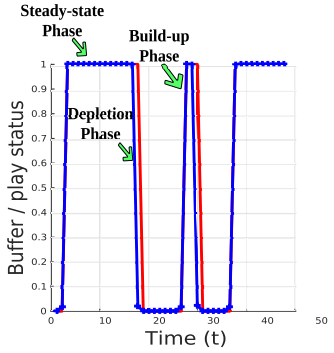


Fig. 7. Illustrating a possible client-side buffer state (in blue) for a given playback state (in red). Best viewed in color.

- We fit an exponential function between data points $(t_d, BUFF\_MIN\_CAPACITY)$ and $(t_3, 0)$.

During smooth playback (steady-state phase), the buffer capacity is not necessarily always at $BUFF\_MIN\_CAPACITY$, but instead can fall anywhere in the range $[BUFF\_MIN\_CAPACITY, BUFF\_MAX\_CAPACITY]$. However, we chose not to model the steady-state phase of the data buffer, because it does not cause any quality degradations in the streaming video, and therefore does not influence viewer QoE.

C. Video Content-Driven Inputs

As mentioned earlier, in addition to stalling events, a viewer's QoE can further be affected by the interplay of other factors such as video quality (due to the presence of distortions), and the spatial and temporal complexities of the video. During our subjective study in [1], we instructed subjects to not judge a video based on their interest in the content, but we did not provide instructions regarding the audio or the video presentation quality. To deepen our understanding in these regards, we sought to study the contributions of these aspects on QoE.

1) Perceptual Video Quality: Perceptual video quality can be defined as the quality of a digital video as *perceived* by human observers, as a reaction to the presence of different forms of spatial and temporal distortions. Rebuffering events, while a form of distortion, do not fall in this category. Bitrate variations and rebuffering events co-occur in streaming videos, and although rebuffering events are more likely to dominate a viewer's QoE, rapid bitrate variations can also significantly impact an end user's dynamic viewing experience and must be accounted for when designing a QoE predictor.

Towards this end, we incorporate either a full-reference, a reduced-reference, or a no-reference video quality assessment (VQA) algorithm [28], [41], [42] in our model, depending on the application scenario. Given the information provided by an objective VQA algorithm, we compute a perceptual VQA score at every second, which provides a continuous-time waveform of perceptual quality. This serves as another continuous-time input to our QoE predictor.

2) Video Space-Time Perceptual Measurement: Videos contain highly diverse spatial and temporal complexities, and different video contents may be retained differently in memory [43], [44]. These may both interact with past memories of

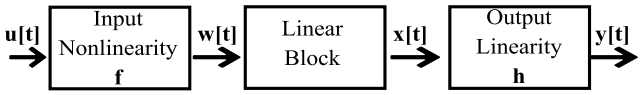


Fig. 8. Block diagram representing the structure of a Hammerstein-Wiener model.

unsatisfactory viewing experiences, e.g., caused by rebuffering events or bitrate drops. The perceptual video quality input designed in Sec. IV-C1 is not sufficient to capture this aspect, so we chose to use a variant of a spatial-temporal metric called *scene criticality* [40]. Let F_n denote the luminance component of a video frame at instant n, and (i, j) denote spatial coordinates within the frame. A frame filtered with the spatial Sobel operator [45] is denoted as $Sobel(F_n)$. Also define the frame difference operation $M_n(i, j) = F_n(i, j) - F_{n-1}(i, j)$. As formulated in [46], spatial perceptual information (SI) and temporal perceptual information (TI) measurements are computed as

$$SI[n] = STD_{space}\Big[Sobel(F_n(i, j))\Big], \tag{8}$$

$$TI[n] = STD_{space}\Big[M_n(i, j)\Big], \tag{9}$$

where STD_{space} denotes the standard deviation computed over all the pixels of a given image (F_n or M_n). These are simple, widely used measurements of video activity [46].

By combining these quantities, a continuous-time scene criticality input at every n is arrived at:

$$Criticality[n] = log_{10}\Big[SI[n] + TI[n]\Big]. \tag{10}$$

We study the efficacies of each of these content-driven inputs on continuous-time QoE prediction in Sec. VII. Note that in a real implementation setting, all of the continuous-time inputs can be easily computed on the fly in real time.

V. Training a Continuous-Time QoE Predictor

A. Hammerstein-Wiener Model

When designing a dynamic model that can accurately predict perceived QoE, structural simplicity and computational efficiency are highly desirable. Moreover, the dynamic model should also crucially account for the affects of subjective hysteresis and memory on viewers' QoE [13], [47]. While a simple linear system model would be desirable, human visual responses contain numerous nonlinearities [48]–[50], which should also be modeled. However, there are no existing behavioral models that can be used to directly and explicitly model the combined effects of the diverse considered inputs, nor of how the inputs relate. Therefore, towards simultaneously capturing the nonlinearities in human visual responses and the hysteresis effect, we employed a simple but powerful classical nonlinear system identification approach called the Hammerstein Wiener (HW) model [51], which can accept multiple dynamic inputs and use them to produce dynamic output predictions.

The core of the HW model is a linear filter with memory [51] to capture the hysteresis effect, with an input point nonlinearity of a very general form to allow the model to learn nonlinearities. This simple design makes it possible to capture both linear and nonlinear aspects of human behavioral

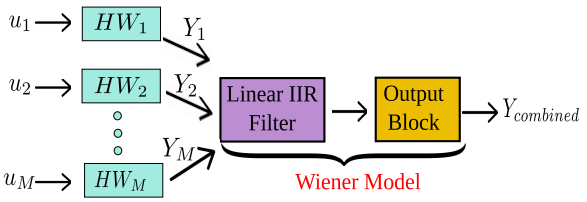


Fig. 9. The multi-stage framework for predicting dynamic QoE.

responses. The output linear scaling block simply scales the output of the linear filter to continuous-time quality scores. Figure 8 shows a block diagram of the single-input single-output (SISO) Hammerstein-Wiener model that we use.

We denote any given continuous-time input (described in Section IV) as $u[t]$ and the resulting continuous-time output as $y[t]$ (see Fig. 8). The input non-linear and output static functions are denoted by f and h respectively. The linear filter block of our model has the following form:

$$x[t] = \sum_{d=0}^{n_b} b_d w[t-d] + \sum_{d=1}^{n_f} f_d x[t-d] = \mathbf{b}^T(w)_{t-n_b:t} + \mathbf{f}^T(x)_{t-n_f:t-1}, \quad (11)$$

where $w[t]$ is the output of the nonlinear input block at time t. The parameters n_b and n_f define the model order, while the coefficients $\boldsymbol{b} = (b_1, \ldots, b_{n_b})^T$ and $\boldsymbol{f} = (f_1, \ldots, f_{n_f})^T$ are learned.

At the input, we process the signal with a generalized sigmoid function of the form

$$w[t] = \beta_3 + \beta_4 \frac{1}{1 + \exp(-(\beta_1 u[t] + \beta_2))}. \quad (12)$$

The output block, which scales the output of the linear IIR filter to a continuous-time QoE prediction, is a simple linear function of the form

$$y[t] = \gamma_1 x[t] + \gamma_2, \quad (13)$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_4)^T$ and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)^T$ are also learned.

B. An Ensemble of Hammerstein-Wiener Models

The HW model is the building block of our continuous-time QoE predictor. Each of the distortion-informative continuous-time inputs (detailed in Sec. IV) is independently used to train a HW model, thereby leading to an ensemble of M HW models. Our next task is to accurately combine them to jointly model the interactions between these factors. Formally, if $Y_i \ \forall i = 1, 2, ..M$ are the continuous-time outputs predicted from each HW model (HW_i), then

$$Y_{combined} = \boldsymbol{\Phi}(Y_1, Y_2, Y_3, ..Y_M), \quad (14)$$

where $\boldsymbol{\Phi}$ is a function that maps the individual outputs to a combined desired output $Y_{combined}$. In our case, we have a total of 9 inputs (6 stall-derived and 2 content-derived) to design an ensemble of $M = 9$ SISO HW models. We chose to strategically combine these models by learning $\boldsymbol{\Phi}$ via the following two alternative approaches:

1) Multi-Stage Approach: In this approach, we utilize the predictions from each HW model (HW_i) to train another Wiener model[2] (Fig. 9). Specifically, each individual prediction serves as input to another linear filter (in the second stage), followed by an output linearity block. Thus, the model in the second stage is a MISO (Multiple Input, Single Output) Wiener model. Given a test video's distortion-informative inputs, we use the trained multi-stage framework to directly derive the final continuous-time QoE prediction $Y_{combined}$.

[2]A Hammerstein-Wiener model without an input non-linearity block is a Wiener model [52].

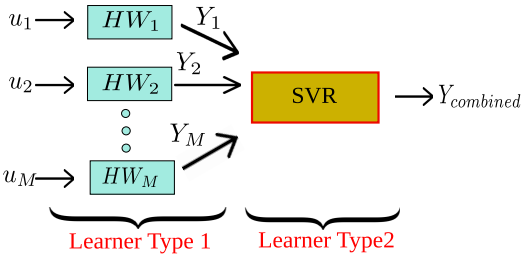


Fig. 10. The multi-learner framework for predicting dynamic QoE.

Although we utilized a two-stage model here, we note that this can be easily extended to more stages if desired. For example, by training separate MISO Wiener models for stall-derived and content-derived inputs, then fusing them using another MISO model, a third stage could be added to the framework, and so on.

2) Multi-Learner Approach: As mentioned earlier, the LIVE Mobile Stall Video Database-II [1] supplies per-instant ground truth QoE scores for each test video sequence. We denote the continuous-time output of a HW_i model for a given video content of length V seconds as $Y_i = [y_{i1}, y_{i2}, y_{i3}, \ldots y_{iV}]$. In this approach, we first construct a set of instance-label pairs for each video content, $(\bar{\boldsymbol{y}}_{\boldsymbol{n}}^{\boldsymbol{I}}, y_n^L) \ \forall \ n = 1, 2, ..V$, where y_n^L is the ground truth subjective QoE at time instant n and $\bar{\boldsymbol{y}}_{\boldsymbol{n}}^{\boldsymbol{I}} = [y_{1n}, y_{2n}, \ldots y_{Mn}]$ are the predictions from each of the M HW models at time instant n. Using the instant-label pairs of all the video contents in the training set, we train a support vector regressor (SVR) to learn the mapping function $\boldsymbol{\Phi}$. We illustrate this learning framework in Fig. 10. Thus, in this approach, we use multiple learners: Hammerstein-Wiener models that predict the continuous-time outputs Y_i, and an SVR that learns $\boldsymbol{\Phi}$. Given a test video's distortion-informative inputs, we use the pre-trained HW models and $\boldsymbol{\Phi}$ to directly derive $Y_{combined}$ using an SVR. Since the SVR is trained on the predictions of other learners, it is a *meta-learner* [15]. Other learners (random forests, multilayer perceptron, etc…) could also be used in place of the SVR.

C. Advantages of the Proposed Dynamic Frameworks

- **Structural Flexibility:** The proposed ensemble framework is extremely flexible, since it can be further supplemented with any number of additional inputs (or by eliminating any ineffective ones), without changing the general structure of the model.
- **Computational Efficiency:** Each of the SISO HW models are extremely fast (the average training time on a video of average length 86 seconds was 0.54 seconds).[3]

[3]These runtimes were obtained using MATLAB's implementation of the Hammerstein Wiener model [52] when executed on Ubuntu 14.04 OS with an Intel i7 CPU (single processor) and 32 GB of RAM.

TABLE II
DESCRIPTION OF THE PROPOSED GLOBAL VIDEO QOE FEATURES

Type of the feature	Definition
Number of stalls	-
Sum of the lengths of all stalls	-
Rebuffering Rate	$\frac{TotalPlaytime}{TotalVideoLength}$
Frequency of stalling events	$\frac{TotalPlaytime}{NumberofStalls}$
Time since the end of last quality impairment	-
Perceptual Quality Score	-

The models can also be trained in parallel to improve the overall computation time on a test video.

- **Modeling the effects of Memory:** The long-term and short-term effects of memory on viewing experience could be easily modeled by adding more SISO HW models to the ensemble using the same kinds of distortion-informative dynamic inputs, but with varied values of the memory parameters (n_b and n_f defined in (11)).

Note that a single MISO HW model could potentially be designed instead of constructing an ensemble of SISO HW models. However, this approach has several disadvantages: 1) the train and test times would be very high when training on multiple nonlinearly transformed inputs. 2) a MISO HW model cannot be trained on different inputs in parallel, and 3) jointly learning multiple input non-linear functions requires a very large amount of training data which is not available in any of the existing QoE databases.

VI. AN OVERALL QOE PREDICTOR WITH GLOBAL VIDEO FEATURES

Although continuous-time QoE predictors are valuable, there is also a need for accurate, computationally efficient overall (end-of-video) QoE predictors that could be used when the resources of the stream-switching controllers are limited or when a different analysis is desired. Thus, we also trained an overall QoE predictor by designing comprehensive global features (listed in Table II) which are derived by effectively encapsulating the aforementioned continuous-time inputs. With regard to the perceptual quality score feature, as we describe in Sec. VII-E, we tested different pooling strategies and different objective VQA algorithms in regard to their ability to derive a single, effective, representative quality score to be used as a feature to feed our global model.

A non-linear mapping was learned between these global features and the corresponding real-valued overall QoE scores of the training videos, using an SVR with a radial basis kernel function. Given any test video's features as input to the trained SVR, a final QoE score may be predicted. The optimal model parameters of the learner were found via cross-validation. Our choice of the model parameters was driven by the obvious aim of minimizing the learner's fitting error to the validation data (details in Sec. VII).

VII. EXPERIMENTS

We evaluated the proposed TV-QoE model and all other currently known continuous-time QoE and global QoE predictors on three different databases: the LIVE Mobile Stall Video Database-II [1], the Waterloo QoE Database [23], and the recent LIVE-Netflix Video QoE Database [24]. Every distorted video in the LIVE Mobile Stall Video Database-II is afflicted by at least one stalling event. However, 60 of the 180 distorted videos in the Waterloo QoE Database, and 56 of the 112 videos of the LIVE-Netflix Video QoE Database are afflicted only by compression artifacts. Since stall-based inputs are not applicable to videos having only compression artifacts, we constructed two disjoint video sets: V_s and V_c, comprising videos afflicted with only compression artifacts and videos afflicted with combinations of stalling events and compression artifacts (if any), respectively.[4] In each of the experiments we describe below, we evaluated the performance of the various predictors on both of these disjoint video collections, wherever applicable.[5]

For every experiment, each database (and video set) was partitioned into training and testing data (80/20 split) with non-overlapping content. To mitigate any bias due to the division of data, the process of randomly splitting each dataset was repeated 50 times. Since global TV-QoE and one of the compared models (V-ATLAS [53]) are learning-based, in each iteration, a model was trained from scratch on the 80% of the data that was set aside for training, then evaluated on the remaining 20% of the test data. FTW [54] and the Streaming QoE Index (SQI) [23] are training-free algorithms, but for a fair comparison with the learning-based models, we report their performance on the test data alone. All of the existing (global and continuous-time QoE) predictors such as SQI [23], FTW [54], and V-ATLAS [53] capture different aspects of stall patterns and distortions. V-ATLAS [53] also computes the normalized (relative to the video duration) time per video over which a bitrate drop took place. The authors follow the simple notion that the relative amount of time that a video is more heavily distorted is directly related to the overall QoE.

For each test split, depending on the type of predictor being evaluated (continuous-time or global), we computed three different metrics as described below:

1) **Continuous-time performance** was evaluated by computing the median of the per-frame correlation and root mean square error (RMSE) between the subjective and the estimated continuous QoE for each distorted test video. The median of these per-video correlations and errors was computed as a performance indicator of the given split.
2) **Overall QoE performance** of a global QoE predictor for a given split was evaluated by computing the correlation and RMSE between the predicted overall QoE and the ground truth overall QoE of the test videos.

For continuous-time as well as global predictors, we report the median Pearson Linear Correlation Coefficient (PLCC), median Spearman Rank-Order Correlation Coefficient (SROCC), and the median RMSE across the 50 test splits. Higher median correlation values indicate better performance

[4]Skipping stall-based inputs is the same as setting all stall-based inputs to zero, provided that these instances are carefully handled in the feature normalization step.

[5]Note that V_c is the empty set $\emptyset$ for LIVE Mobile Stall Video Database-II.

TABLE III

PERFORMANCE OF CONTINUOUS-TIME QOE PREDICTORS ON THE LIVE MOBILE STALL VIDEO DATABASE-II. NOTE THAT THE PER-FRAME QOE VALUES LIE IN THE RANGE [0, 100]. THE BEST PERFORMING MODEL IS INDICATED IN BOLD FONT

Learner Type		PLCC	SROCC	RMSE
Multi-learner (TV-QoE-1)	Stall only Inputs	0.9599	**0.9474**	4.6305
Multi-learner	Stall only Inputs + Scene Criticality	**0.9601**	0.9444	**4.4241**
Multi-learner	Stall only Inputs + Scene Criticality + NIQE [41]	0.9297	0.9262	5.5052
Multi-stage (TV-QoE-2)	Stall only Inputs	0.9394	0.9378	5.3155
Multi-stage	Stall only Inputs + Scene Criticality	0.9429	0.9330	5.0244
Multi-stage	Stall only Inputs + Scene Criticality + NIQE [41]	0.9348	0.9162	5.2517
	SQI + NIQE [23]	0.8348	0.6988	4.4901

of a QoE prediction model with better monotonicity and linear accuracy, while lower median RMSE values indicate better accuracy of the model. Since SQI and FTW are training-free algorithms, their predictions were passed through a logistic non-linearity [55] mapping them to the ground truth QoE scores before computing PLCC. Furthermore, since the Waterloo QoE Database [23] does not contain ground truth continuous-time subjective scores, we were only able to evaluate global QoE models on that database. The continuous-time TV-QoE predictors were superior to all the compared models on all databases with statistical significance. Due to space constraints, we report the results of the statistical significance tests that we conducted on the results of every experiment described below in the supplementary material.

Parameter Selection: To find the optimal parameters for each individual Hammerstein-Wiener QoE prediction model in the ensemble, we determined the model order parameters (n_b, n_f, $\mathbf{b}$, $\mathbf{f}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$), and the input nonlinearities via cross-validation on the LIVE Mobile Stall Video Database-II [1]. Specifically, we divided the entire dataset into 70% training, 10% validation, and 20% test sets. We conducted a simple grid-search of the parameter values to train each model on the training dataset, then evaluated its performance on the validation dataset, which is disjoint from the test data. We found that the values $n_b = 4$ and $n_f = 3$ served as the final model parameters for each of the SISO Hammerstein-Wiener models in the ensemble. We also determined the values of the weights α_1 in (1) and α_2 in (2) using cross-validation. Specifically, we performed a grid search varying both scalars between 0.1 and 0.7 in steps of 0.1, trained a series of models using the training data, and evaluated the performance of each on the validation data. We found that the models with $\alpha_1 \approx 0.2$ and $\alpha_2 \approx 0.1$ yielded maximum correlation scores, and thus, these values were used in all the experiments on all the databases.

Henceforth, the multi-learner will be referred to as TV-QoE-1, while the multi-stage learner will be referred to as TV-QoE-2. The parameters of both the Wiener model in the TV-QoE-2 framework (Sec. V-B1) and the SVR in the TV-QoE-1 framework (Sec. V-B2) were also determined via cross-validation. For the Wiener model in the TV-QoE-2, we found that the value of $n_b = 1$ and a simple linear output block yielded maximum correlation scores on all the databases. In the TV-QoE-1 framework, we employed an SVR using a non-linear radial basis kernel function.

TABLE IV

CONTRIBUTION OF THE PROPOSED STALL AND VIDEO CONTENT-BASED DYNAMIC INPUTS TOWARDS CONTINUOUS-TIME QOE ON THE 50 TEST SPLITS OF THE LIVE MOBILE STALL VIDEO DATABASE-II [1]. THE VIDEO CONTENT-BASED INPUTS ARE ITALICIZED

Dynamic Inputs	PLCC	SROCC
Stall position	0.6946	0.6962
Number of stalls	0.4399	0.4744
Time since previous stall	0.9109	0.8919
Stall density	0.6264	0.6945
Buffer model	0.7893	0.7829
Frequency	0.6812	0.7640
Rebuffering rate	0.5554	0.5495
Scene Criticality	*0.5701*	*0.4399*
NIQE [41]	*0.0758*	*0.0811*

A. Performance of Continuous-Time Predictors on LIVE Mobile Stall Video Database-II

First, we evaluated the performance of continuous-time QoE models on the distorted videos of the LIVE Mobile Stall Video Database-II. The results are reported in Table III. Since we proposed two different ways of combining the ensemble of Hammerstein-Wiener models, we report the performance of both of these models. SQI, which is the only other existing continuous-time QoE predictor, uses a per-frame quality metric to compute the spatial quality on each frame. Specifically, the instantaneous QoE (Q_n) at each frame n is computed as the sum of a video presentation quality (P_n), i.e., spatial quality, and a stall-dependent experience quality (S_n). Since the LIVE Mobile Stall Video Database-II does not have reference videos, we relied on a popular per-frame NR-IQA metric, NIQE [41], to compute the continuous-time SQI scores.

It may be observed from Table III that the proposed set of dynamic inputs and learners significantly outperform SQI. It may also be observed that the proposed multi-learner approach (Sec. V-B2) performs better than the multi-stage approach (Sec. V-B1), for every given input combination. The likely reason for this is that the SVM is better able to account for nonlinear correlations between the first stage outputs. We will show next that NIQE scores [41] are poor indicators of instantaneous QoE, so including NIQE as an input slightly impairs the performance of TV-QoE. Figure 11 illustrates a few examples of the ground truth and the predicted continuous-time QoE waveforms of a few test videos from the proposed

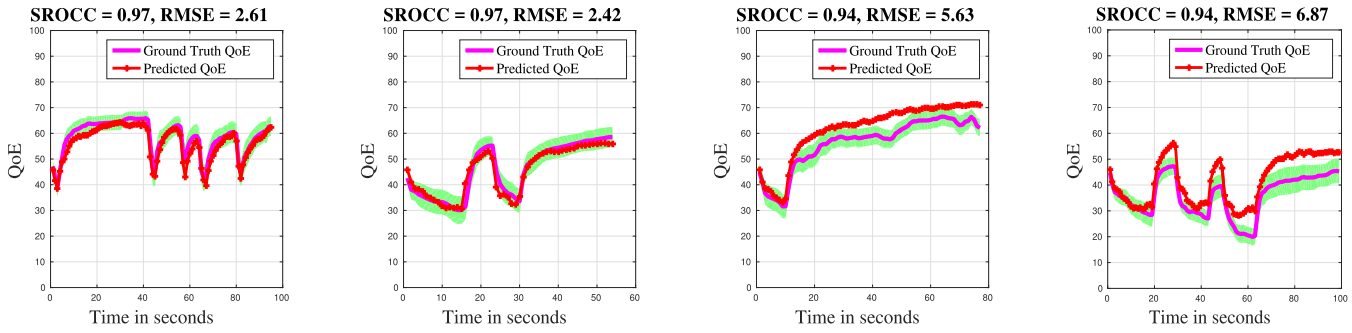


Fig. 11. Some examples of the continuous-time predictions obtained from the proposed algorithm (indicated in red) on different test video sequences of the LIVE Mobile Stall Video Database-II. The ground truth dynamic QoE response is indicated in magenta and the associated 95% confidence interval derived from the responses from individual subjects is indicated in green. Spearman Rank Ordered Correlation (SROCC) and Root Mean Squared Error (RMSE) between the instantaneous predicted and ground truth QoE is also reported in each plot.

TABLE V

PERFORMANCE OF CONTINUOUS QOE PREDICTORS ON THE VIDEO SET V_s OF THE LIVE-NETFLIX VIDEO QOE DATABASE. NOTE THAT THE PER-FRAME QOE VALUES LIE IN THE RANGE $[-2.26, 1.52]$. THE BEST PERFORMING MODEL IS INDICATED IN BOLD FONT

Learner Type		Quality Predictor	PLCC	SROCC	RMSE
Multi-learner (TV-QoE-1)	**Stall only Inputs**	-	**0.9131**	**0.8579**	0.3536
Multi-learner	Stall only Inputs + Scene Criticality	NIQE [41]	0.9059	0.8306	0.3672
Multi-learner	Stall only Inputs + Scene Criticality	SSIM [42]	0.8694	0.7820	0.3151
Multi-learner	Stall only Inputs + Scene Criticality	STRRED [28]	0.8905	0.8061	**0.3004**
Multi-stage (TV-QoE-2)	Stall only Inputs		0.8800	0.7970	0.3851
Multi-stage	Stall only Inputs + Scene Criticality	NIQE [41]	0.8738	0.7775	0.4026
Multi-stage	Stall only Inputs + Scene Criticality	SSIM [42]	0.8345	0.7248	0.3584
Multi-stage	Stall only Inputs + Scene Criticality	STRRED [28]	0.8496	0.7471	0.3777
	SQI	NIQE [41]	0.6821	0.4281	0.3433
	SQI	SSIM [42]	0.6892	0.3793	0.3450
	SQI	STRRED [28]	0.6705	0.3275	0.3581

approach (using the multi-learner approach and the stall-based inputs in isolation). It may be observed that the proposed model does not overfit to the existing dataset, but instead attempts to accurately predict the varying trends in each dynamic QoE prediction. In some of the examples, it may be observed that the QoE predictions occasionally fall outside of the 95% confidence interval, despite maintaining a strong monotonic relationship with the ground truth dynamic QoE.

B. Intrinsic Analysis of the Individual Dynamic Inputs

To better understand the relationship between our input set and the dynamic QoE, we trained separate Hammerstein-Wiener Models on each input on the same 50 random, non-overlapping train and test splits of the LIVE Mobile Stall Video Database-II, as were used in Sec. VII-A. We report the median SROCC and PLCC scores over these 50 iterations in Table IV. These results illustrate the degree to which each of these inputs accurately predict perceived QoE, while also justifying the choice of our inputs. Nevertheless, this analysis does not account for any relationships between the various inputs. It may also be observed that the per-second NIQE scores [41] performed rather poorly at predicting QoE scores when videos were afflicted by stalling events. Thus including this input when conducting continuous-time QoE prediction degrades performance (Sec. VII-A). Of course, the NIQE model utilizes only spatial information and does not benefit from any reference signal or training process.

C. Performance of Continuous-Time Predictors on the LIVE-Netflix Video QoE Database

As mentioned, we divided the entire collection of 112 videos in the LIVE-Netflix Video QoE Database into two disjoint video sets: V_c and V_s. Videos belonging to V_s contain both compression and stalling artifacts, while those in V_c contain only compression artifacts. Hence, on the video set V_c, we did not use any stall-based inputs,[6] relying instead only on the content-driven inputs. We report the performance of TV-QoE-1, TV-QoE-2, and SQI in Table VI. For videos in V_s, however, we used both stall-based as well as content-based inputs, and report the performance in Table V. Furthermore, we also considered scenarios where either a FR, RR, or NR VQA model would be incorporated into the QoE predictor. It may be observed from these results that TV-QoE significantly outperforms SQI on both video sets, especially when videos were afflicted by both stalls and compression artifacts. Further, the multi-learner approach (TV-QoE-1) yielded better

[6]This is same as setting stall-based inputs to zero.

TABLE VI

PERFORMANCE OF CONTINUOUS QOE PREDICTORS ON THE VIDEO SET V_c OF THE LIVE-NETFLIX VIDEO QOE DATABASE. NOTE THAT THE PER-FRAME QOE VALUES LIE IN THE RANGE $[-2.26, 1.52]$. THE BEST PERFORMING MODEL IS INDICATED IN BOLD FONT

Learner Type		Quality Predictor	PLCC	SROCC	RMSE
Multi-learner (TV-QoE-1)	Scene Criticality	NIQE [41]	0.2412	0.1711	0.5462
Multi-learner	Scene Criticality	SSIM [42]	0.6314	0.3998	0.4723
Multi-learner	**Scene Criticality**	**STRRED [28]**	**0.6733**	**0.5776**	0.3965
Multi-stage (TV-QoE-2)	Scene Criticality	NIQE [41]	0.2512	0.1594	0.5609
Multi-stage	Scene Criticality	SSIM [42]	0.6387	0.3786	0.4732
Multi-stage	Scene Criticality	STRRED [28]	0.6728	0.5715	0.3969
	SQI	NIQE [41]	0.2123	0.1408	**0.3102**
	SQI	SSIM [42]	0.2392	0.0934	0.3136
	SQI	STRRED [28]	0.1984	0.1917	0.3214

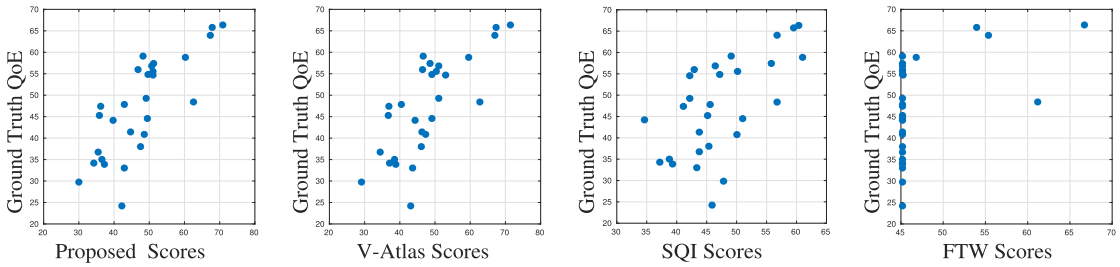


Fig. 12. Scatter plots of the ground truth overall QoE scores and the predicted overall QoE scores obtained on a single test split from four different global QoE predictors on the LIVE Mobile Stall Video Database-II [1]. The proposed Global TV-QoE model (left most), is statistically significant than all other global QoE predictors.

performance than the multi-stage approach (TV-QoE-2) on both video sets of the LIVE-Netflix Video QoE Database.

D. Cross-Dataset Evaluation of TV-QoE

As an example of cross-database evaluation, we trained the better performing multi-learner model, TV-QoE-1, on the LIVE Mobile Stall Video Database-II, then tested it on 20% of the video set V_s of the LIVE-Netflix Video QoE Database. This is the only meaningful cross-database comparison we can make, since otherwise the overlap of features that can be extracted from the pair of databases under consideration becomes too small. In this experiment, we only considered stall-driven, scene criticality, and NIQE as the continuous-time inputs, since only these input features that are shared by the two databases. We found that TV-QoE-1 achieved a median PLCC of 0.7823 and a median SROCC of 0.6355 over the same set of 50 test splits that were used in the Sec. VII-C. These are outstanding numbers, given that (Table V) the cross-dataset performance of TV-QoE-1 was superior to that of SQI on the LIVE-Netflix Video QoE Database.

E. Performance of Global QoE Predictors

Next, we evaluated the performance of the proposed global features (Table II) and other global QoE predictors under identical train/test settings on all three databases and report the results in Tables VII, VIII, and IX. We computed the perceptual quality scores using various quality predictors and tested several pooling strategies to derive a single quality score from the per-frame perceptual quality score, to be used as a global feature. However, due to space constraints, we only report the results obtained from the pooling strategy that yielded the best performance. Note that the LIVE Mobile Stall Video Database-II does not contain pristine videos, so we relied on the no-reference (NR) picture quality model NIQE [41] to supply VQA scores on this database. For the other two databases, the reference videos are available, so we report the performance using full-reference (SSIM [42]), reduced-reference (ST-RRED [28]), and no-reference VQA models (NIQE [41]). Note that when evaluating the proposed global QoE predictor on the video sets V_c of different databases, we utilized only the video content-based inputs, since the stall-informative global features do not capture any information. When evaluating the proposed global QoE predictor and V-ATLAS, we trained an SVR with a radial basis kernel function, by separately finding the optimal SVR parameters via cross-validation on all three databases.

TABLE VII

PERFORMANCE OF GLOBAL QOE MODELS ON THE LIVE MOBILE STALL VIDEO DATABASE-II [1]. NOTE THAT THE FINAL QOE VALUES LIE IN THE RANGE $[0, 100]$. THE BEST PERFORMING MODEL IS INDICATED IN BOLD FONT

	PLCC	SROCC	RMSE
TV-QoE Global Stall Features	0.7099	0.6836	8.7609
TV-QoE Global Stall Features + Max-Pooled NIQE	**0.7757**	**0.7797**	**7.7914**
SQI + NIQE	0.4828	0.4565	9.8512
V-ATLAS + Max-Pooled NIQE	0.7541	0.7572	8.1541
FTW	0.4411	0.6689	10.5074

TABLE VIII

PERFORMANCE OF GLOBAL QOE PREDICTORS ON THE WATERLOO QOE DATABASE [23]. NOTE THAT THE FINAL QOE VALUES LIE IN THE RANGE [0, 100]. THE BEST PERFORMING MODEL IS INDICATED IN BOLD FONT

		V_s			V_c		
	Quality Predictor (Pool type = mean)	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
Global TV-QoE Features	NIQE [41]	0.3200	0.3216	14.8681	0.2983	0.3356	19.9181
Global TV-QoE Features	SSIM [42]	**0.8660**	0.8604	8.5568	**0.8956**	**0.8531**	**12.2217**
SQI [23]	NIQE [41]	0.3046	0.4134	14.1765	0.2393	0.3357	18.6965
SQI [23]	SSIM [42]	0.8582	**0.8623**	**7.5603**	0.8910	0.8462	12.2412
V-ATLAS [53]	NIQE [41]	0.3200	0.3216	14.8681	0.2983	0.3356	19.9181
V-ATLAS [53]	SSIM [42]	**0.8660**	0.8604	8.5568	**0.8956**	**0.8531**	**12.2217**
FTW [54]		NaN	NaN	-	-NA-	-NA-	-NA-

TABLE IX

PERFORMANCE OF GLOBAL QOE PREDICTORS ON THE LIVE-NETFLIX VIDEO QOE DATABASE [24]. NOTE THAT THE FINAL QOE VALUES LIE IN THE RANGE $[-1.6, 1.6]$. THE BEST PERFORMING MODEL IS INDICATED IN BOLD FONT

		V_s			V_c		
	Quality Predictor (Pool type = mean)	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
Global TV-QoE Features	NIQE [41]	0.6719	0.3318	0.4126	0.7676	0.4909	0.6616
Global TV-QoE Features	SSIM [42]	0.7821	0.7045	0.3475	0.9030	0.8000	0.7452
Global TV-QoE Features	STRRED [28]	0.8564	0.7591	0.3196	0.9246	0.8091	0.5663
SQI	NIQE [41]	0.1977	0.0272	0.4051	0.4895	0.3773	0.6630
SQI	SSIM [42]	0.6132	0.5500	0.3185	0.8262	0.8000	0.4596
SQI	STRRED [28]	**0.8597**	0.7500	**0.2151**	0.8061	0.8000	0.3820
V-ATLAS	NIQE [41]	0.8165	0.6091	0.3245	0.8170	0.6045	0.6250
V-ATLAS	SSIM [42]	0.7951	0.6591	0.3346	0.9406	0.8545	0.3902
V-ATLAS	STRRED [28]	0.8547	**0.7636**	0.3095	**0.9462**	**0.8591**	**0.3586**
FTW		0.2797	0.2778	0.3984	-NA-	-NA-	-NA-

We found that the proposed global QoE predictor outperforms all existing QoE predictors on the LIVE Mobile Stall Video Database-II (Table VII). It is also clear from these results that including NIQE as a global perceptual quality metric benefits the QoE prediction. The scatter plots of the predicted and the ground truth QoE scores for one test split are illustrated in Fig. 12. With regard to the Waterloo QoE Database, there are a couple of oddities in the results arising from the database design. Each of the 120 videos in the Waterloo QoE Database belonging to V_s are of the same length and each contains one stalling event of duration 5 seconds. In this peculiar scenario, the otherwise different global TV-QoE and V-ATLAS features capture exactly the same information, thereby yielding identical performances (Table VIII). Moreover, since the FTW model [54] is based on only two features (the number and the summed length of stalls), it predicts the same quality score on all video contents in the Waterloo QoE Database. On the LIVE-Netflix QoE Database, the global TV-QoE model competes very well with the performances of V-ATLAS and SQI (Table IX).

VIII. CONCLUSIONS AND FUTURE WORK

We presented a continuous-time video QoE predictor that effectively captures the effects of a variety of QoE-influencing factors, and that models the client-side data buffer model, subjective hysteresis, and that is able to accurately predict viewers' instantaneous QoE. We have also designed a global QoE predictor that achieves top performance on all existing QoE databases. The success of the proposed models encourages us to design quality-aware stream-switching algorithms which could control the position, location, and length of stalls, given a network bandwidth budget and the end user's device information, such that the end user's QoE is maximized. Such a model could have a direct and immediate impact on existing adaptive stream-switching algorithms that are used in the client players of global content providers such as YouTube and Netflix and could also propel user-centric mobile network planning and management.

ACKNOWLEDGEMENTS

The authors thank Zhengfang Duanmu for providing us the source code of the SQI algorithm and Christos Bampis for sharing the video content of the LIVE-Netflix QoE Database.

REFERENCES

[1] D. Ghadiyaram, J. Pan, and A. Bovik, "A subjective and objective study of stalling events in mobile streaming videos," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.

[2] K. Brunnström *et al.*, *Qualinet White Paper on Definitions of Quality of Experience*, 5th Qualinet Meeting, Novi Sad, Serbia, 2013. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00977812

[3] *Part 6: Dynamics Adaptive Streaming Over HTTP (DASH)*, document ISO/IEC FCD 23001-6, MPEG Requirements Group, 2011.

[4] R. Pantos and W. May, "HTTP live streaming," IETF Draft, Jun. 2010. [Online]. Available: https://tools.ietf.org/html/draft-pantos-http-live-streaming-04

[5] L. De Cicco, S. Mascolo, and V. Palmisano, "Feedback control for adaptive live video streaming," in *Proc. 2nd Annu. Conf. Multimedia Syst.*, 2011, pp. 145–156.

[6] L. De Cicco, V. Caldaralo, V. Palmisano, and S. Mascolo, "ELASTIC: A client-side controller for dynamic adaptive streaming over HTTP (DASH)," in *Proc. Int. Packet Video Workshop (PV)*, Dec. 2013, pp. 1–8.

[7] Z. Li *et al.*, "Probe and adapt: Rate adaptation for HTTP video streaming at scale," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 4, pp. 719–733, Apr. 2014.

[8] J. Jiang, V. Sekar, and H. Zhang, "Improving fairness, efficiency, and stability in HTTP-based adaptive video streaming with FESTIVE," in *Proc. Int. Conf. Emerg. Netw. Experim. Technol.*, 2012, pp. 97–108.

[9] T.-Y. Huang, R. Johari, N. McKeown, M. Trunnell, and M. Watson, "A buffer-based approach to rate adaptation: Evidence from a large video streaming service," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 4, pp. 187–198, 2014.

[10] K. Watanabe, J. Okamoto, and T. Kurita, "Objective video quality assessment method for evaluating effects of freeze distortion in arbitrary video scenes," *Proc. SPIE*, vol. 6494, p. 64940P, Jan. 2007.

[11] M. A. Masry and S. S. Hemami, "A metric for continuous quality evaluation of compressed video with severe distortions," *Signal Process., Image Commun.*, vol. 19, no. 2, pp. 133–146, 2004.

[12] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz, "Quantification of YouTube QoE via crowdsourcing," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2011, pp. 494–499.

[13] K. Seshadrinathan and A. C. Bovik, "Temporal hysteresis model of time varying subjective video quality," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process. (ICASSP)* May 2011, pp. 1153–1156.

[14] L. Ljung, *System Identification: Theory for the User*. Upper Saddle River, NJ, USA: Prentice-Hall, 1986.

[15] D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259, 1992.

[16] 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.

[17] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. de Veciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 652–671, Oct. 2012.

[18] VQEG HDTV Group. *VQEG HDTV Database. Video Quality Experts Group (VQEG)*. [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/hdtv/hdtv.aspx

[19] 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.

[20] D. Ghadiyaram, J. Pan, A. C. Bovik, A. K. Moorthy, P. Panda, and K.-C. Yang, "Subjective and objective quality assessment of mobile videos with in-capture distortions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 1393–1397.

[21] D. Ghadiyaram, A. Bovik, H. Yeganeh, R. Kordasiewicz, and M. Gallant. (2016). *LIVE Mobile Stall Video Database—I*. [Online]. Available: http://live.ece.utexas.edu/research/LIVEStallStudy/index.html

[22] D. Ghadiyaram, A. C. Bovik, H. Yeganeh, R. Kordasiewicz, and M. Gallant, "Study of the effects of stalling events on the quality of experience of mobile streaming videos," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Dec. 2014, pp. 989–993.

[23] Z. Duanmu, A. Rehman, K. Zeng, and Z. Wang, "Quality-of-experience prediction for streaming video," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2016, pp. 1–6.

[24] C. G. Bampis, Z. Li, A. K. Moorthy, I. Katsavounidis, A. Aaron, and A. C. Bovik, "Study of temporal effects on subjective video quality of experience," *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5217–5231, Nov. 2017.

[25] Y. Tian and M. Zhu, "Analysis and modelling of no-reference video quality assessment," in *Proc. Int. Conf. Comput. Autom. Eng. (ICCAE)*, Mar. 2009, pp. 108–112.

[26] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1352–1365, Mar. 2014.

[27] L. Ma, S. Li, and K. N. Ngan, "Reduced-reference video quality assessment of compressed video sequences," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 10, pp. 1441–1456, Oct. 2012.

[28] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, pp. 684–694, Apr. 2012.

[29] 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.

[30] M. A. Masry and S. S. Hemami, "A metric for continuous quality evaluation of compressed video with severe distortions," *Signal process., Image commun.*, vol. 19, no. 2, pp. 133–146, 2004.

[31] 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.

[32] W. Zou, F. Yang, J. Song, S. Wan, W. Zhang, and H. R. Wu, "Event-based perceptual quality assessment for HTTP-based video streaming with playback interruption," *IEEE Trans. Multimedia*, to be published.

[33] A. Raake, M.-N. Garcia, W. Robitza, P. List, and S. Göring, and B. Feiten, "A bitstream-based, scalable video-quality model for HTTP adaptive streaming: ITU-T P.1203.1," in *Proc. 9th Int. Conf. Quality Multimedia Exper. (QoMEX)*, May/Jun. 2017, pp. 1–6.

[34] A. Balachandran, V. Sekar, A. Akella, S. Seshan, I. Stoica, and H. Zhang, "Developing a predictive model of quality of experience for Internet video," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 43, no. 4, pp. 339–350, 2013.

[35] C. Alberti *et al.*, "Automated QoE evaluation of dynamic adaptive streaming over HTTP," in *Proc. 5th Int. Workshop Quality Multimedia Exper. (QoMEX)*, Jul. 2013, pp. 58–63.

[36] M. T. Vega, D. C. Mocanu, and A. Liotta, "A regression method for real-time video quality evaluation," in *Proc. 14th Int. Conf. Adv. Mobile Comput. Multimedia*, 2016, pp. 217–224.

[37] D. Z. Rodríguez, R. L. Rosa, E. C. Alfaia, J. I. Abrahão, and G. Bressan, "Video quality metric for streaming service using DASH standard," *IEEE Trans. Broadcast.*, vol. 62, no. 3, pp. 628–639, Sep. 2016.

[38] H. Yeganeh, R. Kordasiewicz, M. Gallant, D. Ghadiyaram, and A. C. Bovik, "Delivery quality score model for Internet video," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 2007–2011.

[39] D. Ghadiyaram, J. Pan, and A. C. Bovik, "A time-varying subjective quality model for mobile streaming videos with stalling events," *Proc. SPIE*, vol. 9599, pp. 911–959, Sep. 2015.

[40] C. Fenimore, J. Libert, and S. Wolf, "Perceptual effects of noise in digital video compression," *SMPTE J.*, vol. 109, no. 3, pp. 178–187, 2000.

[41] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[42] 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.

[43] A. Lang, K. Dhillon, and Q. Dong, "The effects of emotional arousal and valence on television viewers' cognitive capacity and memory," *J. Broadcast. Electron. Media*, vol. 39, no. 3, pp. 313–327, 1995.

[44] T. Sharot and E. A. Phelps, "How arousal modulates memory: Disentangling the effects of attention and retention," *Cognit., Affective, Behavioral Neurosci.*, vol. 4, no. 3, pp. 294–306, 2004.

[45] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Reading, MA, USA: Addison-Wesley, 1992.

[46] *Subjective Video Quality Assessment Methods for Multimedia Applications*, document ITU-T Rec. P.910, International Telecommunication Union, 2008.

[47] D. E. Pearson, "Viewer response to time-varying video quality," *Proc. SPIE*, vol. 3299, pp. 16–25, Jul. 1998.

[48] A. B. Watson and J. A. Solomon, "Model of visual contrast gain control and pattern masking," *J. Opt. Soc. Amer. A*, vol. 14, no. 9, pp. 2379–2391, Sep. 1997.

[49] P. C. Teo and D. J. Heeger, "Perceptual image distortion," in *Proc. IEEE ICIP*, vol. 2. Nov. 1994, pp. 982–986.

[50] S. Daly, "Visible differences predictor: An algorithm for the assessment of image fidelity," *Proc. SPIE*, vol. 1666, pp. 179–206, Aug. 1992.

[51] J. A. Nelder, "The fitting of a generalization of the logistic curve," *Biometrics*, vol. 17, no. 1, pp. 89–110, 1961.

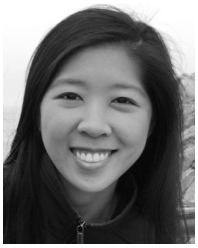
[52] *Identifying Hammerstein Weiner Models*. [Online]. Available: https://www.mathworks.com/help/ident/ug/identifying-hammerstein-wiener-models.html

[53] C. G. Bampis and A. C. Bovik. (2017). "Learning to predict streaming video QoE: Distortions, rebuffering and memory." [Online]. Available: https://arxiv.org/abs/1703.00633

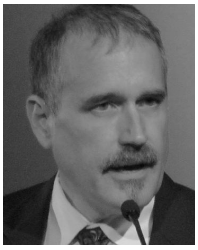
[54] T. Hoßfeld, R. Schatz, E. Biersack, and L. Plissonneau, "Internet video delivery in YouTube: From traffic measurements to quality of experience," in *Data Traffic Monitoring and Analysis*, vol. 7754. 2013, pp. 264–301.

[55] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

Deepti Ghadiyaram received the B.Tech. degree in computer science from the International Institute of Information Technology, Hyderabad, in 2009, and the M.S. and Ph.D. degrees from The University of Texas at Austin in 2013 and 2017, respectively. She is currently a Research Scientist with Facebook. She was a recipient of the Microelectronics and Computer Development Fellowship from 2013 to 2014 and a recipient of the Graduate Student Fellowship to the top 1% of the students by the Department of Computer Science for the academic years from 2013 to 2016. Her research interests include image and video processing, particularly perceptual image and video quality assessment, computer vision, and machine learning.



Janice Pan received the B.S. and M.S. degrees in electrical engineering from The University of Texas at Austin in 2013 and 2016, respectively. She is currently pursuing the Ph.D. degree in electrical engineering with the Laboratory for Image and Video Engineering. She was a recipient of the Virginia and Ernest Cockrell, Jr. Fellowship in Engineering from 2013 to 2017. Her research interests include computer vision, machine learning, and perceptual video quality assessment.



Alan C. Bovik (F'95) received the B.S., M.S., and Ph.D. degrees in electrical and computer engineering from the University of Illinois, Champaign, IL, USA, in 1980, 1982, and 1984, respectively. He is currently the Cockrell Family Regents Endowed Chair Professor with The University of Texas at Austin. He has authored or co-authored the books, including *The Handbook of Image and Video Processing*, the *Modern Image Quality Assessment*, and *The Essential Guides to Image and Video Processing*. He received the 2017 Edwin H. Land Medal from the Optical Society of America, the 2015 Primetime Emmy Award for Outstanding Achievement in Engineering Development, the 2013 IEEE Signal Processing Society Society Award, and about ten journal best paper awards, including the 2016 IEEE Signal Processing Society Sustained Impact Award. He created the IEEE International Conference on Image Processing, Austin, TX, USA, in 1994. He co-founded and was a longest-serving Editor-in-Chief of the IEEE TRANSACTIONS ON IMAGE PROCESSING.