Recurrent and Dynamic Models for Predicting Streaming Video Quality of Experience

Christos G. Bampis, Zhi Li, Ioannis Katsavounidis, and Alan C. Bovik, *Fellow, IEEE*

***Abstract*— Streaming video services represent a very large fraction of global bandwidth consumption. Due to the exploding demands of mobile video streaming services, coupled with limited bandwidth availability, video streams are often transmitted through unreliable, low-bandwidth networks. This unavoidably leads to two types of major streaming-related impairments: compression artifacts and/or rebuffering events. In streaming video applications, the end-user is a human observer; hence being able to predict the subjective Quality of Experience (QoE) associated with streamed videos could lead to the creation of perceptually optimized resource allocation strategies driving higher quality video streaming services. We propose a variety of recurrent dynamic neural networks that conduct continuous-time subjective QoE prediction. By formulating the problem as one of time-series forecasting, we train a variety of recurrent neural networks and non-linear autoregressive models to predict QoE using several recently developed subjective QoE databases. These models combine multiple, diverse neural network inputs, such as predicted video quality scores, rebuffering measurements, and data related to memory and its effects on human behavioral responses, using them to predict QoE on video streams impaired by both compression artifacts and rebuffering events. Instead of finding a single time-series prediction model, we propose and evaluate ways of aggregating different models into a forecasting ensemble that delivers improved results with reduced forecasting variance. We also deploy appropriate new evaluation metrics for comparing time-series predictions in streaming applications. Our experimental results demonstrate improved prediction performance that approaches human performance. An implementation of this work can be found at https://github.com/christosbampis/NARX_QoE_release.**

***Index Terms*— Subjective and objective video quality assessment, Quality of Experience, streaming video, rebuffering event.**

I. Introduction

VIDEO data and mobile video streaming demands have skyrocketed in recent years [1]. Streaming content providers such as Netflix, Hulu and YouTube strive to offer high quality video content that is viewed by millions of subscribers under very diverse circumstances, using a plethora of devices (smartphones, tablets and larger screens), under varying viewing resolutions and network conditions. This enormous volume of video data is transmitted over wired or wireless networks that are inherently throughput limited. On the client side, the available bandwidth may be volatile, leading to video playback interruptions (rebuffering events) and/or dynamic rate changes.

These network-related video impairments adversely affect end-user quality of experience (QoE) ubiquitously; hence studying QoE has become a major priority of streaming video companies, network providers and video QoE researchers. For example, to better account for fluctuating bandwidth conditions, industry standard HTTP-based adaptive streaming protocols have been developed [2]–[8] that divide streaming video content into chunks, represented at various quality levels; whereby the quality level (or representation) to be played at any given time is selected based on the estimated network condition and/or buffer capacity. These adaptation algorithms seek to reduce the frequency and number of rebuffering events, while minimizing occurrences of low video quality and/or frequent quality switches, all of which can significantly and adversely affect viewer QoE [9]–[11]. Note that HTTP-based adaptive video streaming relies on TCP and hence frame drops [12] or packet loss [13] distortions are not an issue.

In streaming video applications, the opinion of the human viewer is the gold standard; hence integrating models of perceptual video quality and other "QoE-aware" features into resource allocation protocols is highly relevant. This requires injecting principles of visual neuroscience and human behavior modeling into the video data resource allocation strategies. Systems that can make accurate real-time predictions of subjective QoE could be used to create perceptually optimized network allocation strategies that can mediate between volatile network conditions and user satisfaction.

Here, we present a family of *continuous-time* streaming video QoE prediction models that process inputs derived from perceptual video quality algorithms, rebuffering-aware video measurements and memory-related temporal data. Our major contribution is to re-cast the continuous-time QoE prediction problem as a time-series forecasting problem. In the time-series literature, a wide variety of tools have been devised ranging from linear autoregressive-moving-average (ARMA) models [14], [15] to non-linear approaches, including artificial neural networks (ANNs). ARMA models are easier to analyze; however they are based on stationarity assumptions. However, subjective QoE is decidedly non-stationary and is affected by dynamic QoE-related inputs, such as sudden quality changes or playback interruptions. This suggests that non-stationary models implemented as ANNs are more suitable for performing QoE predictions.

Manuscript received April 21, 2017; revised December 16, 2017; accepted February 27, 2018. Date of publication March 14, 2018; date of current version April 6, 2018. This work was supported by Netflix Inc., under a research grant. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Kalpana Seshadrinathan. *(Corresponding author: Christos G. Bampis.)*

C. G. Bampis and A. C. Bovik are with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA (e-mail: bampis@utexas.edu; bovik@ece.utexas.edu).

Z. Li and I. Katsavounidis are with Netflix Inc., Los Gatos, CA 95032 USA (e-mail: zli@netflix.com; ikatsavounidis@netflix.com).

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.2018.2815842

We specifically focus on the most practical and pressing problem: predicting *continuous-time* QoE by developing QoE system models driven by a mixture of quality, rebuffering and memory inputs to ANN-based dynamic models. Building on preliminary work in [16] and [17], we advance progress towards this goal by devising efficient QoE prediction engines employing dynamic neural networks including recurrent neural networks, NARX [16], [17] and Hammerstein Wiener models [18], [19]. We thoroughly test these models on a set of challenging new subjective QoE datasets, and we conduct an in-depth experimental analysis of model and variable selection. We also study a variety of new ways of aggregating the time-series responses produced in parallel by different QoE models and initializations into a single robust continuous-time QoE estimate, and we provide demonstrations and guidance on the advantages and shortcomings of evaluation metrics that might be used to assess continuous time QoE prediction performance. We also compare the abilities of our proposed models against upper bounds on performance, i.e, human predictions.

The rest of this paper is organized as follows. Section II studies previous work on video quality assessment and QoE, while Section III discusses the design of our general QoE predictor. Next, Section IV describes the proposed predictor that we have deployed and experimented with, and the complementary continuous-time inputs that feed it. In Section V we introduce the forecasting ensemble approaches that are used to augment performance, and in Section VI a general class of QoE predictors that we designed are described. Section VII explains the experimental setup and Section VIII describes and analyzes our experimental results. Section IX concludes and discusses possible future improvements.

II. Related Work

Ultimately, video QoE research aims to create QoE prediction models that can efficiently address the resource allocation problem while ensuring the visual satisfaction of users. As such, QoE prediction models are designed and evaluated on databases of QoE-impaired videos and associated human subjective scores [13], [20]–[25]. Recently developed QoE prediction models can be conveniently divided into *retrospective* and *continuous-time* QoE predictors.

Retrospective QoE prediction models output a single number which summarizes the overall QoE of an entire viewed video. Many video quality assessment (VQA) models that only measure visual distortions from, for example, compression or packet loss fall into this category. VQA models are further classified as full-reference (FR) [26]–[32], reduced-reference (RR) [33] or no-reference (NR) [34]–[39], depending on whether all or part of a pristine reference video is used in the assessment process. Towards a similar goal, the MOAVI VQEG project [40] studies no-reference assessment of audio-visual quality, but these methods seek to detect the presence of an artifact and do not measure overall quality.

Besides video quality degradations, retrospective QoE is also affected by playback interruptions [41], [42]; hence retrospective predictive models have been proposed that compute global rebuffering-related features, such as the number or durations of rebuffering events [43], [44]. Hybrid approaches that model video quality degradations and rebuffering events have very recently been studied, resulting in models like SQI [45] and the learning-based Video ATLAS [46]. Other works [47] integrate video fidelity measurements with rebuffering information, but these approaches simply ascribe bitrate or QP values with perceptual video quality rather than deploying high-performing perceptual VQA models.

Similar efforts have been recently initiated as part of the AVHD-AS/P.NATS Phase 2 project [48], a joint collaboration between VQEG and ITU Study Group 12 Question 14, which includes numerous industry and university proponents. These research efforts have the same broad goal as our work, which is to design objective QoE prediction models for HTTP-based adaptive streaming [49], [50]. The P.NATS models combine information descriptive of rebuffering and video quality as determined by bitstream or pixel-based measurements. These approaches operate on a temporal block basis (e.g. on GOPs). Our work has two fundamental differences. First, we deploy continuous-time predictors that measure QoE with finer, per-frame granularity and these QoE responses can be further aggregated over any desired time interval when designing adaptive rate allocation strategies. Furthermore, we train neural network models that exploit long-term memory properties of subjective QoE, which is a distinctive feature of our work.

Continuous-time QoE prediction using perceptual VQA models has received much less attention and is a more challenging problem. In [18], a Hammerstein-Wiener dynamic model was used to make continuous-time QoE predictions on videos afflicted only by dynamic rate changes. In [17], it was shown that combining video quality scores from several VQA models as inputs to a non-linear autoregressive model, or simply averaging the individual forecasts derived from each can deliver improved results. In [51], a simple model called DQS was developed using cosine functions of rebuffering-aware inputs, which was later improved using a learned Hammerstein-Wiener system in [19]. The system only processed rebuffering-related inputs, using a simple model selection strategy. Furthermore, only the final values of the predicted time-series were used to assess performance. As we will explain later, time-series evaluation metrics need to take into account the temporal structure of the data. To the best of our knowledge, the only approach to date that combines perceptual VQA model responses with rebuffering measurements is described in [16], where a simple non-linear autoregressive with exogenous variables (NARX) model was deployed to predict continuous QoE.

A limitation of previous QoE prediction studies has been that experimental analysis was carried out only on a single dynamic model and on a single subjective database. Since predictive models designed or learned and tested on a specific dataset run the risk of inadvertent "tailoring" or overtraining, deploying more general frameworks and evaluating them on a variety of different datasets is a difficult, but much more valuable proposition. We also believe that insufficient attention has been directed towards how to properly apply evaluation metrics to time-series QoE prediction models. Optimal model parameters can significantly vary across different test videos; hence carefully designed cross-validation strategies for model selection are advisable. In addition, it is possible to better generalize and improve QoE prediction performance by using forecast ensembles that filter out spurious forecasts. Finally,

previous studies of continuous QoE have not investigated the limits of QoE prediction performance against human performance; calculating the upper bounds of QoE model execution is an exciting and deep question for QoE researchers.

To sum up, previous research studies on the QoE problem have suffered from at least one, and usually several, of the following limitations:

1) including either quality or rebuffering aware inputs
2) relying on a single type of dynamic model
3) limited justification of model selection
4) using evaluation metrics poorly suited for time-series comparisons
5) limited evaluation on a single video QoE database
6) do not exploit time-series ensemble forecasts
7) do not consider *continuous-time* human performance

Our goal here is to surmount 1-7 and to further advance efforts to create efficient, accurate and real-time QoE prediction models that can be readily deployed to perceptually optimize streaming video network parameters.

III. Designing General Continuous-Time QoE Predictors

In our search for a general and accurate continuous-time QoE predictor, we realized that subjective QoE is affected by the following:

1) Visual quality: low video quality (e.g. at low bitrates) or bandwidth-induced fluctuations in quality [11], [21] may cause annoying visual artifacts [13], [20] thereby affecting QoE.
2) Playback interruption: frequent or long rebuffering events adversely affect subjective QoE [41], [43]. Compared to degradations on visual quality, rebuffering events have remarkably different effects on subjective QoE [10], [21].
3) Memory (or hysteresis) effects: Recency [11], [21], [52] is a phenomenon whereby current QoE is more affected by recent events. Primacy occurs when QoE events that happen early in a viewing session are retained in memory, thereby also affecting the current sense of QoE [53].

Broadly, subjective QoE "is a non-linear aggregate of video quality, rebuffering information and memory" [9]–[11], [16]. Recently, the learning-driven Video ATLAS model [46] proposed to combine these different sources of information to predict QoE in general streaming environments where rebuffering events and video quality changes are commingled. Nevertheless, that model is only able to deliver overall (end) QoE scores. Towards solving the more difficult continuous-time QoE prediction problem, the following points should be considered:

1) At least three types of "QoE-aware inputs" must be fused: VQA model responses, rebuffering measurements and memory effects.
2) These inputs should have high descriptive power. For example, high-performance, perceptually-motivated VQA models should be preferred over less accurate indicators such as QP values [47] or PSNR. QoE-rich information can reduce the number of necessary inputs and boost the general capabilities of the QoE predictor.
3) Dynamic models with memory are able to capture recency (or memory) which is an inherent property of QoE.
4) These dynamic models should have an adaptive structure allowing for variable numbers of inputs. For example, applications where videos are afflicted by rebuffering events are not always relevant.
5) Multiple forecasts may be combined to obtain robust forecasts when monitoring QoE in difficult, dynamically changing real-world video streaming environments.

An outcome of our work is a promising tool we call the General NARX (GN) QoE predictor. Table I summarizes the notation that we will be using throughout the paper. In the following sections, we motivate and explain the unique features of this new method.

IV. The GN-QoE Predictor

Our proposed GN-QoE prediction model is characterized by two main properties: the number and type of continuous-time features used as input and the prediction engine that it relies on. In this section, we discuss in greater detail the QoE-aware inputs of our system and the neural network engine that we have deployed for continuous-time QoE prediction.

A. QoE-Aware Inputs

The proposed GN-QoE Predictor relies on a non-linear dynamic approach which integrates the following *continuous-time QoE-aware* inputs:

1) The high-performing ST-RRED metric as the VQA model. Previous studies [16], [21], [46], [54], have shown that ST-RRED is an exceptionally robust predictor of subjective video quality. As was done in [17], it is straightforward to augment the GN-QoE Predictor by introducing additional QoE-aware inputs, if they verifiably contribute QoE prediction power. For example, the MOAVI key indicators [40] of bluriness or blur loss distortion could be applied in order to complement the current VQA input. At the same time, we recognize that simple and efficient models are desireable in practical settings, especially ones that can be adapted to different types of available video side-information.

   Quality switching [11], [55] also has a distinct effect on subjective video QoE. While we do not explicitly model quality switching, the memory component of the NARX engine allows it to exploit ST-RRED values over longer periods of time as a proxy for video segments having different qualities.
2) We define a boolean continuous-time variable R_1 which describes the playback status at time t which takes value $R_1 = 1$ during a rebuffering event and $R_1 = 0$ at all other times. This input captures playback-related information. We also define the integer measure R_2 to be the number of rebuffering events that have occurred until time t.
3) M: the time elapsed since the latest network-induced impairment such as a rebuffering event or a bitrate change occurred. M is normalized to (divided by) the overall video duration. This input targets recency/ memory effects on QoE. Figure 1 shows a few examples

TABLE I
DESCRIPTION OF THE ACRONYMS AND VARIABLES USED THROUGHOUT THE PAPER

Acronym	Description	Acronym	Description
VQA	video quality assessment	r	# training data splits for cross-validation
QoE	quality of experience	N_T	# training QoE time-series
QP	quantization parameter	S	# shuffles for performance bounds
NARX	non-linear autoregressive neural network	N_f	# frames for a given video
RNN	recurrent neural network	OL	open-loop configuration
HW	Hammerstein-Wiener	CL	closed-loop configuration
V-N/R/H	VQA-driven QoE with NARX/RNN/HW	ANN	artificial neural network
R-N/R/H	rebuffering-driven QoE with NARX/RNN/HW	FR	full-reference
G-N/R/H	general QoE-aware with NARX/RNN/HW	RR	reduced-reference
R_1	playback status indicator at time t	NR	no-reference
R_2	# rebuffering events until time t	RMSE	root-mean-squared error
M	time elapsed since last distortion (memory)	OR	outage ratio
D_1	LIVE HTTP Streaming Video Database [18]	DTW	dynamic time warping
D_2	LIVE Mobile Stall Video Database-II [19]	$\mathbf{D}$	pairwise DTW distance matrix
D_3	LIVE-Netflix Video QoE Database [21]	CI	confidence interval
d_u	# external variable lags	SROCC	Spearman's rank order correlation coefficient
d_y	# input lags	PLCC	Pearson's linear correlation coefficient
H	# hidden nodes	LD	number of layer delays in RNN
N	# videos in a subjective QoE database	α	significance level for hypothesis testing
T	# training initializations	m	# comparisons in Bonferroni correction

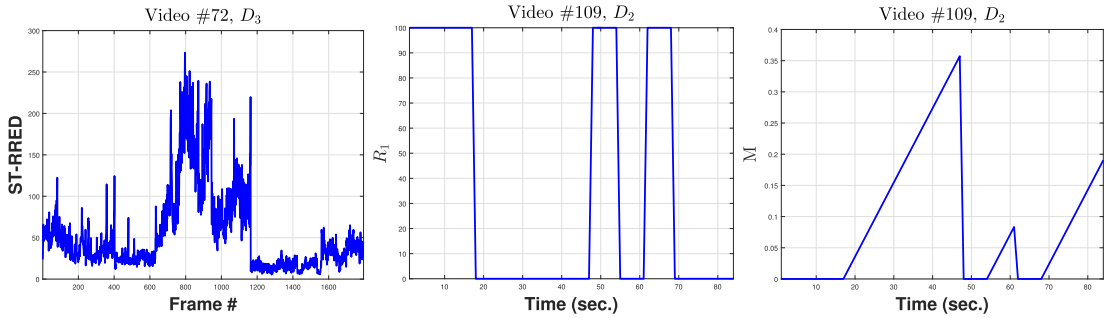

Fig. 1. Examples of the proposed continuous time QoE variables. Left to right: ST-RRED computed on video #72 of the LIVE-NFLX Video QoE Database (D_3), and R_1 and M on the LIVE Mobile Stall Video Database-II (D_2).

of these continuous-time inputs measured on videos from various subjective databases.

B. NARX Component

The GN-QoE Predictor relies on the non-linear autoregressive with exogenous variables (NARX) model [16], [56], [57]. The NARX model explicitly produces an output y_t that is the result of a non-linear operation on multiple past inputs $(y_{t-1}, y_{t-2}, \dots)$ and external variables $(\mathbf{u}_t)$:

$$y_t = f(y_{t-1}, y_{t-2}, \dots, y_{t-d_y}, \mathbf{u}_t, \mathbf{u}_{t-1}, \mathbf{u}_{t-2}, \dots, \mathbf{u}_{t-d_u}) \quad (1)$$

where $f(\cdot)$ is a non-linear function of previous inputs $\{y_{t-1}, y_{t-2}, \dots, y_{t-d_y}\}$, and previous (and current) external variables $\{\mathbf{u}_t, \mathbf{u}_{t-1}, \mathbf{u}_{t-2}, \dots, \mathbf{u}_{t-d_u}\}$, where d_y is the number of lags in the input and d_u is the number of lags in the external variables. To capture the recency effects of subjective QoE, the memory lags d_y and d_u need to be large enough. In practice, we determine these parameters using cross-validation (see Section VII-B). In Appendix II-D we show that GN-QoE is able to capture recency effects when predicting QoE.

In a NARX model, there are two types of inputs: past outputs that are fed back as future inputs to the dynamic model, and external (or "exogenous") variables (see Fig. 1). The former are scalar past outputs of the NARX model, while the latter are past and current values of QoE-related information, e.g. the video quality model responses, and can be vector valued. To illustrate this, Fig. 2 shows an example of the NARX architecture: there are three exogenous inputs $\mathbf{u}(t)$, each containing a zero lag component and five past values. By contrast, past outputs cannot contain the zero lag component.

The function $f(\cdot)$ is often approximated by a feed-forward multi-layer neural network [58] possibly having variable number of nodes per hidden layer. Here we focus on single-hidden layer architectures having H hidden nodes. There are two approaches to training a NARX model. The first approach is to train the NARX without the feedback loop, also known as an open-loop (OL) configuration, by using the ground truth values of $\mathbf{y}_t$ when computing the right hand side of (1). An example of the ground truth scores is shown in Fig. 3. The second approach uses previous estimates of y_t, also known as a closed-loop (CL) configuration [17]. Both approaches can be used while training; however, application of the NARX must be carried out in CL mode, since ground truth subjective data is not available to define a new time-series. The advantages of the OL approach are two-fold: the actual subjective scores are used when training, and the neural network to be trained is feed-forward; hence static backpropagation can be used [59].

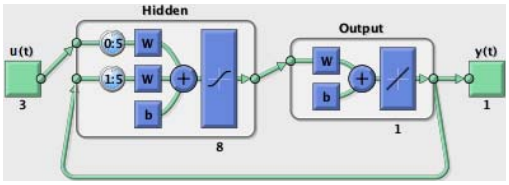

Fig. 2. The dynamic CL NARX system with 3 inputs, 8 neurons in the hidden layer and 5 feedback delays. The recurrency of the NARX occurs in the output layer [59].

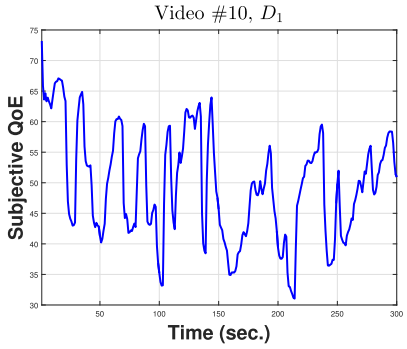

Fig. 3. Exemplar subjective QoE scores on video #10 from the LIVE HTTP Streaming Video Database (denoted by D_1).

It has been shown [17] that, in practice, the CL configuration requires longer training times and yields worse predictive performance; hence we use the OL configuration when training and the CL configuration only when testing. An example of the CL configuration of the NARX model is shown in Fig. 2. For simplicity, we used a tangent sigmoid activation function and a linear function in the output layer. The role of the linear function is to scale the outputs in the range of the subjective scores, while the sigmoid activation function combines past inputs and external variables in a non-linear fashion. Given that the problem is of medium size, we chose the Levenberg-Marquardt [60], [61] algorithm to train the model [62]. To reduce the chances of overfitting in the OL training step, we used an early stopping approach [63]: the first 80% of the samples were used to train the OL NARX, while the remaining 20% were used to validate it. In Appendix I, we discuss these implementation details of the NARX predictor, including the choice of the training algorithm, the activation function and data imputation strategies.

The GN-QoE Predictor follows a learning-driven approach which requires careful cross-validation and design. Still, preliminary experiments led us to the conclusion that a single time-series prediction may be insufficient for the challenging problem of continuous-time QoE prediction. Next, we describe another unique feature of the GN-QoE Predictor: the use of forecasting ensembles.

V. Forecasting Ensembles

A. Motivation

Ensemble learning is a long-standing concept that has been widely applied in such diverse research fields as forecasting [64], [65] and neural network ensembles [66], [67]. We are specifically interested in time-series forecasting ensembles, where two or more continuous QoE predictions are aggregated. In our application, we utilize a variety of dynamic approaches that have various parameters, such as the number of input delays. The results of these models may also depend on the neural network initialization. Generally, relying on a single model may lead to drawbacks such as:

1) Uncertain model selection. For example, in the stationary time-series and ARMA literature [14], [15], model order selection typically relies on measurements of sample autocorrelations or on the Akaike Information Criterion. However, in neural network approaches, this problem is not as well-defined.
2) Using cross-validation for model selection may not always be the best choice. Different choices of the evaluation metric against which the QoE predictor is optimized may yield different results. Furthermore, an optimal model for a particular data split may not be suitable for a different test set. While much larger QoE databases could contribute towards ameliorating this issue, the barriers to creating these are quite formidable, suggesting multi-modal approaches as an alternative way to devise effective and practical solutions.
3) The QoE dynamics within a given test video may vary widely, reducing the effectiveness of a single model order.

Since a single time-series predictor might yield subpar prediction results, we have developed ensemble prediction models that deliver more robust prediction performance by deemphasizing unreliable forecasts. These ensemble techniques were applied to each of the forecasts generated. For example, testing GN-QoE using κ different combinations of model orders d_u and d_y, λ different neural network initializations and μ possible values for the neurons in the hidden layer, produces $\kappa\lambda\mu$ forecasts which are then combined together yielding a single forecast. In the next section, we discuss these ensemble methods in greater detail.

B. Proposed Ensemble Methods

We have developed two methods of combining different QoE predictors. The first determines the best performer from a set of candidate solutions. We relied on the dynamic time warping (DTW) distance [68] which measures the similarity between two time-series that have been time-warped to optimally match structure over time: a larger DTW distance between two time-series signifies they are not very similar. The benefit of DTW is that it accounts for the temporal structure of each time-series and that it makes it possible to compare signals that are similar but for rebuffering-induced delays. We computed pairwise DTW distances between all predictors, thereby producing a symmetric matrix of distances $\mathbf{D} = [d_{ij}]$, where $d_{ij} = d_{ji}$ is the DTW distance between the ith and jth time-series predictions. Similar to the subject rejection method proposed in [21], we hypothesize that $v_i = \sum_j \mathbf{D}_{ij}$, i.e., the sum across rows (or columns) of $\mathbf{D}$ is an effective measure of the reliability of the ith predictor. A natural choice is

$$i_o = \arg\min_i \; v_i, \tag{2}$$

where i_o denotes the single best predictor. Note that i_o may not necessarily coincide with the time-series prediction resulting from the best model parameters (as derived in the

TABLE II

SUMMARY OF THE VARIOUS COMPARED QoE PREDICTORS. X DENOTES THAT THE PREDICTOR IN THE ROW POSSESSES THE PROPERTY DESCRIBED IN THE COLUMN. WE HAVE FOUND THAT INCLUDING R_2 IN THE G-PREDICTORS PRODUCES NO ADDITIONAL BENEFIT (SEE APPENDIX II)

QoE Predictor	Learner	VQA	R_1	R_2	M	ensemble
VN	NARX	X				X
RN	NARX		X	X		X
RMN	NARX		X	X	X	X
GN	NARX	X	X		X	X
VR	RNN	X				X
RR	RNN		X	X		X
RMR	RNN		X	X	X	X
GR	RNN	X	X		X	X
VH	HW	X				X
RH	HW		X	X		X
RMH	HW		X	X	X	X
GH	HW	X	X		X	X

cross-validation step). The second approach is to assign a probabilistic weight to each of the C candidate predictors:

$$\tilde{y}_t = \sum_{c=1}^{C} w_c \hat{y}_{ct}, \quad w_c = \frac{1/v_c}{\sum_c 1/v_c}, \tag{3}$$

where $w_c \in [0, 1]$ determines (weights) the contribution of the cth predictor to the ensemble estimate $\tilde{y}_t$. Along with these two ensemble methods, we also evaluated several other commonly used ensemble methods, including mean, median and mode ensembles. Mean ensembles have proven useful in many forecasting applications [69], while median and mode ensembles are more robust against outliers [70].

VI. THE G-FAMILY OF QoE PREDICTORS

The GN-QoE Predictor is versatile and can exploit other VQA inputs than the high performance ST-RRED model [54]. Indeed, it allows the use of any VQA model (FR, RR or NR), depending on the available reference information. As in [16] and [46], this enables the deployment of these models in a wide range of QoE predictions applications.

Taking this a step forward, we have developed a wider family of predictors based on the ST-RRED, R_1 and M inputs, that also deploy other dynamic model approaches. For example, Layer-Recurrent Neural Networks (denoted here as RNNs) [71] or the Hammerstein-Wiener (HW) dynamic model [18], [19] can be used instead of NARX, yielding models called GR-QoE and GH-QoE, respectively. This general formulation also allows us to consider model subsets that relate and generalize previous work. For example, the GH-QoE model, when using only ST-RRED as input (denoted by VH in Table II) may be considered as a special case of [18]. We summarize the proposed family of G-predictors and other predictors that use subset of these inputs, and their characteristics in Table II. Since the same QoE features are shared across GN-, GR- and GH-QoE, we next discuss the learning models underlying GR-QoE and GH-QoE.

A. GR-QoE Models

Recurrent Neural Networks (RNNs) [71] have recently gained popularity due to their successful applications to various tasks such as handwriting recognition [72] and speech recognition [73]. The main difference between the NARX and RNN architectures, is that while the former uses a feedback connection from the output to the input, RNNs are feedforward neural networks that have recurrent connections in the hidden layer. Therefore, the structure of an RNN allows it to dynamically respond to time-series input data. The recurrency property of RNNs allows them to model the recency properties of subjective QoE. An example of such a neural network is shown in Fig. 4.

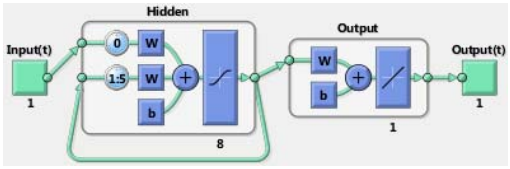

Fig. 4. The dynamic RNN approach with 1 input, 8 neurons in the hidden layer and 5 layer delays: the recurrency occurs in the hidden layer rather than in the output layer [59].

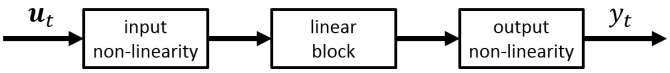

Fig. 5. The HW dynamic approach.

Given that the amount of available subjective data is insufficient to train a deep neural network, we decided to train relatively simple RNN models, i.e., neural networks having only one hidden layer and up to 5 layer delays. As in NARX, we used a tangent sigmoid activation function and a linear function at the output layer.

B. GH-QoE Models

Unlike the NARX and RNN models, the HW model, which is block-based (see Fig. 5), has only been deployed for QoE prediction on videos afflicted by rate drops [18] or rebuffering events [19]. The HW structure is relatively simple: a dynamic linear block having a transfer function with n_f poles and n_b zeros, preceded and followed by two non-linearities. The poles and zeros in the transfer function allow the HW model to capture the recency effects in subjective QoE, while the non-linear blocks account for the non-linear relationship between the input features and QoE.

The family of G-QoE predictors (see Table II) can be applied to any subjective database containing videos afflicted by quality changes, rebuffering events or both, by simply choosing the model (QoE feature) subset that is applicable to each case. Following our G-notation, we also define predictors V- (which use only VQA model responses), R- (only rebuffering features) and RM- (rebuffering and memory). We next describe the various subjective datasets we used to evaluate the various approaches.

VII. SUBJECTIVE DATA AND EXPERIMENTAL SETUP

We now discuss the experimental aspects behind our QoE prediction systems. We first describe the three different subjective QoE databases that we used and our parameter selection strategy. Next, we discuss the advantages and caveats of various continuous-time performance metrics and their differences. We conclude this section with a discussion on performance bounds of continuous-time QoE predictors.

A. Subjective Video QoE Databases

In [18], a subjective video QoE database (denoted by D_1 for brevity) was created containing 15 long video sequences afflicted by quality fluctuations relevant to HTTP rate-adaptive video streaming. This database consists of 8 different video contents of 720p spatial resolution encoded at various H.264 bitrate levels, with associated time-varying subjective scores. Rebuffering events were studied in [74] using a different database (denoted by D_2), where diverse rebuffering patterns were inserted into 24 different video contents of various spatial resolutions. Unlike [18], this subjective QoE database allows the study of rebuffering-related characteristics (such as the number, locations and durations of the rebuffering events) and their effects on time-varying and overall QoE. A total of 174 distorted videos are part of this database.

A deficiency of these early studies is that they were not driven by any bandwidth usage models and did not contain videos containing both rebuffering events and quality variations. In realistic streaming applications, dynamic rate adaptations and rebuffering events occur, often in temporal proximity depending on the client device's resource allocation strategy [5]–[7]. Towards bridging this gap, we built the new LIVE-NFLX Video QoE Database [21] (D_3). This database contains about 5000 continuous and retrospective subjective QoE scores, collected from 56 subjects on a mobile device. It was designed based on a bandwidth usage model, by applying 8 distortion patterns on 14 spatio-temporally diverse video contents from the Netflix catalog and other publicly available video sources. These impairments consist of constant and/or dynamic rate drops commingled with rebuffering events.

We used these three subjective databases to extensively study the performance of the continuous-time GN-, GR- and GH-QoE predictors. Next, we describe the cross-validation strategy that we used to determine the best parameter setting for each of these prediction engines.

B. Cross-Validation Framework for Parameter Selection

We now introduce our cross-validation scheme for continuous-time QoE prediction. Notably, the proposed recurrent models are highly non-linear; hence the traditional time-series model estimation techniques used in ARMA models [14] are not possible. Further, subjective QoE prediction is highly non-stationary; therefore the most suitable model order may vary within a given QoE time-series or across different test time-series. As a result, determining the best model parameters, e.g., the input and feedback delays in the GN-QoE model (d_u and d_y), the number of poles (n_f) and zeros (n_b) in the transfer function of a GH model, or the number of layer delays (LD) in a GR model, must be carefully validated (see Table III).

Here we propose a novel cross-validation framework that is suitable for *streaming video* QoE predictors. This idea builds on a simpler approach that was introduced in [17]. In data-driven quality assessment applications, the available data is first split into content-independent training and testing subsets, then the training data is further split into smaller "validation" subsets for determining the best parameters. Content independence ensures that subjective biases towards different contents is alleviated when training and testing. In the case of data-driven continuous-time QoE predictors, it is more realistic to split the data in terms of their distortion patterns, since the testing network conditions (which have a direct effect on the playout patterns) are not known *a priori*.

TABLE III

PARAMETERS USED IN OUR EXPERIMENTS. ON ALL THREE DATABASES WE FIXED $r = 3$ AND $T = 5$. K CAN BE ANY OF THE FOLLOWING THREE: G, V OR RM DEPENDING ON THE SUBJECTIVE DATABASE THAT THE PREDICTORS WERE APPLIED

Model	KN			KR		KH		
parameter	d_u	d_y	H	LD	H	n_b	n_f	H
D_1	[10,12,14]	[10,12,14]	[5,8]	[3,4,5]	[5,8]	[10,12,14]	[10,12,14]	10
D_2	[4,5,6]	[4,5,6]	[5,8]	[3,4,5]	[5,8]	4	4	10
D_3	[8,10,15]	[8,10,15]	[5,8]	[3,4,5]	[5,8]	[8,10,15]	[8,10,15]	10

The non-deterministic nature of these time-series predictions adds another layer of complexity. As an example, given a set of QoE time-series used for training, we have found that different initial weights produce different results for GN- and GR-QoE Predictors; hence their performance should be estimated across initializations. By comparison, previous continuous-time QoE prediction models [16], [18], [19] have used a single model order. To sum up, training a successful continuous-time QoE predictor requires:

1) Determining the best set of parameters using cross-validation on the available continuous-time subjective data.
2) Ensuring content-independent train and test splits.
3) Distorted videos corresponding to the same network or playout pattern should belong only in the train or the test set.
4) To account for different neural network initializations, multiple iterations need to be performed on per training set.

Based on these properties, we now discuss our cross-validation strategy in detail. Let $i = 1 \ldots N$ index the video in a database containing N videos. First, randomly select the ith video as the test time-series. To avoid content and other learning biases, remove from the training set all videos having similar properties as the test video, such as segments that belong to the same video content. Depending on which subjective database is used, we applied the following steps. For D_3, we removed all videos having either the same content or the same distortion pattern [16]. For D_1 and D_2, we removed all videos having the same content. This process yielded a set of N_T training QoE time-series for each test video, where $N_T = 10$, 129 and 91 for D_1, D_2 and D_3 respectively.

Next, we divided the training set further into a training subset and a validation subset. This step was repeated r times to ensure sufficient coverage of the data splitting. We also found that the HW component of the GH-QoE model was sensitive to the order of the training data in a given training set. To account for this variation, we also randomized the order of the time-series in this second training set. Then, we evaluated each model configuration on every validation set in terms of root-mean-square error (RMSE), and averaged the RMSE scores, yielding a single number per model configuration. The model parameters that yielded the minimum RMSE were selected to be the ones used during the testing stage. When testing, we used all of the training data and the

optimized model parameters that were selected in the cross-validation step. To account for different weight initializations, we repeated the training process T times; then averaged the performances across initializations.

During cross-validation, we used the RMSE evaluation metric to select the best performing model configuration. Nevertheless, other evaluation metrics may also contribute important information when comparing continuous-time QoE prediction engines. In the following section, we investigate these metrics in greater detail.

C. Evaluation Metrics

After performing the time-series predictions, it is necessary to select suitable evaluation metrics to compare the output p with the ground truth time g. In traditional VQA, e.g, in [27] and in hybrid models of retrospective QoE [45], [46], the Spearman rank order correlation coefficient (SROCC) is used to measure monotonicity, while Pearson's Linear Correlation Coefficient (PLCC) is used to evaluate the linear accuracy between the ground truth subjective scores and the VQA/QoE predicted scores. These evaluation metrics have also been used in studies of continuous-time QoE prediction [17]–[19].

Yet, it is worth asking the question: "Is there a single evaluation metric suitable for comparing subjective continuous-time QoE scores?" We have found that each evaluation metric has its own merits; hence they have to be considered collectively.

We now discuss the advantages and shortcomings of the various evaluation metrics that can be used to compare a ground truth QoE time-series $g = [g_i]$ and a predicted QoE waveform $w = [w_i]$ where i denotes the frame index. Continuous-time subjective QoE is inherently a dynamic system with memory; hence we have developed continuous-time autoregressive QoE models. However, SROCC and PLCC are only valid under the assumption that the samples from each set of measurements were independently drawn from within each set; whereas subjective QoE contains strong time dependencies and inherent non-stationarities.

There are other evaluation metrics that are more suitable for time-series comparisons, i.e.,

1) The root-mean-squared error (RMSE), which captures the overall signal fidelity: $\sqrt{(\sum_{i=1}^{N_f}(w_i - g_i)^2)/N_f}$, where N_f is the number of frames.
2) The outage rate (OR) [18], which measures the frequency of times when the prediction w_i falls outside twice the confidence interval of g_i: $\frac{1}{N_f}\sum_{i=1}^{N_f}\mathbb{1}|w_i - g_i| > 2\mathrm{CI}_{g_i}$, where CI_{g_i} is the 95% confidence interval of the ground truth g at frame i across all subjects.
3) The dynamic time warping (DTW) distance can also be employed [16], [21], [68] to capture the temporal misalignment between w and g.

Each of these metrics has shortcomings:

1) The RMSE is able to capture the scale of the predicted output, but cannot account for the temporal structure.
2) The OR is intuitive and suitable for continuous-time QoE monitoring, but does not give information on how the predicted time-series behaves within the confidence bounds.

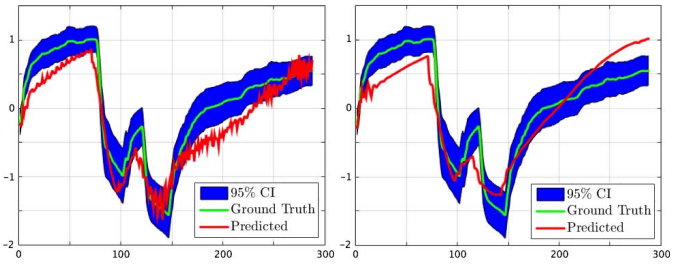

Fig. 6. Vertical axis: QoE; horizontal axis: time (in samples). OR does not describe the prediction's behavior within the CI. (a) OR = 5.90. (b) OR = 13.19.

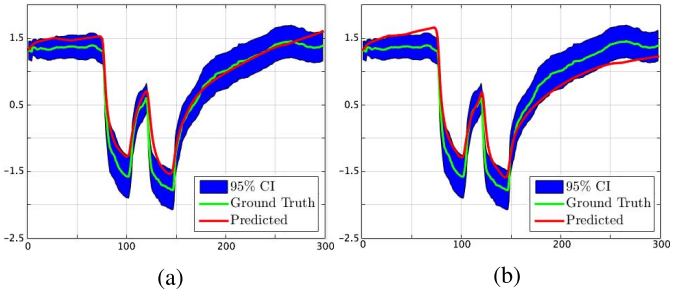

Fig. 7. Vertical axis: QoE; horizontal axis: time (in samples). DTW better reflects the temporal trends of the prediction error although it is harder to interpret. (a) DTW = 2.96. (b) DTW = 19.56.

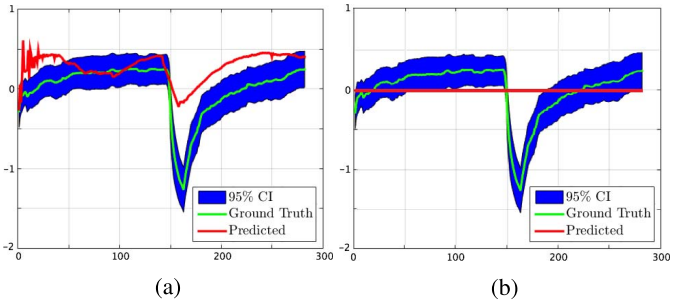

Fig. 8. Vertical axis: QoE; horizontal axis: time (in samples). RMSE does not effectively account for the local temporal structure of the prediction error. (a) RMSE = 0.36. (b) RMSE = 0.33.

3) DTW captures temporal trends, but the DTW distance is hard to interpret, e.g., a smaller distance is always better but a specific value is hard to interpret.

We demonstrate these deficiencies in Figs. 6, 7 and 8. Figure 6 shows that the outage rate on the left is lower; however the predicted QoE is noisy. By contrast, while the predicted QoE on the right has a larger OR, it is more stable and it appears to track the subjective QoE more accurately. Figure 7 shows that, while the DTW distance between the two time-series predictions is very different, both predictions nicely capture the QoE trend. Lastly, while RMSE captures the correct QoE range, an artificially generated time-series containing a zero value performs better than the temporal prediction but misses all of the trends (see Fig. 8). Clearly, any single evaluation metric is likely to be insufficiently descriptive of performance; hence we report all three of these metrics, along with the SROCC, to draw a clearer picture of relative performance.

D. Continuous-Time Performance Bounds

While the previously discussed evaluation metrics can be used to compare QoE predictors, they do not yield an absolute ranking against the putative upper bound of human performance. As stated in [13]: "The performance of an objective model can be, and is expected to be, only as good as the performance of humans in evaluating the quality of a given video." We measured the "null" (human) level of performance as follows. We divided the subjective scores of each test video into two groups of the same size, one considered as the training set and the other as the test set. Let A_i and B_i be the two sets, i.e., A_i is the train set for the ith test video and B_i the corresponding test set. For a given evaluation metric, we averaged the subjective scores in A_i and B_i and compared them. To account for variations across different splits, this process was repeated S times per test video, yielding subsets A_{is} and B_{is} at each iteration s. We fixed $S = 10$. Then, we computed the median value over s, yielding the median prediction performance of the ith test video. Finally, to obtain a single performance measure on a given database, we calculated the median value over all test videos.

VIII. Experimental Results

In this section, we thoroughly evaluate and compare the different approaches in terms of their qualitative and quantitative performance. Recall that only database D_3 contains both quality changes and playback interruptions; hence we applied the V-predictors on D_1, the RM-predictors on D_2 and the G-predictors on D_3.

To examine statistical significance, we used the non-parametric Wilcoxon significance test [75] using a significance level of $\alpha = 0.05$. To account for multiple comparisons, we applied Bonferroni correction which adjusts α to $\frac{\alpha}{m}$, where m is the number of comparisons. In all of the reported statistical test results, a value of '1' indicates that the row is statistically better than the column, while a value of '0' indicates that the row is statistically worse than the column; a value of '-' indicates that the row and column are statistically equivalent.

A. Qualitative Experiments

We begin by visually evaluating the different models on a few videos from all three QoE databases. Figure 9 shows the performance of the VN-QoE Predictor on video #8 of database D_1; the continuous time predictions of the best cross-validated model closely follow the subjective QoE, and all individual models yielded similar outputs. In such cases, it may be that forecasting ensembles yield little benefit.

By contrast, Fig. 10 shows QoE prediction on video #16 of database D_2. All three dynamic approaches suffered either from under- or over-shoot. The RMR-QoE Predictor produced some spurious forecasts. In this instance, an ensemble method could increase the prediction reliability, but, in this example, the RMH-QoE Predictor performed well.

The example in Fig. 11 proved challenging for both the GN- and GR-QoE Predictors: the best cross-validated GN model was unable to capture the subjective QoE trend, while the GR model produced an output that did not capture the first part of the QoE drop. These examples highlight some of the challenges of the problem at hand: finding the best neural network model can be difficult. By contrast, the GH model was able to produce a much better result. Notably, all three dynamic approaches suffered from spurious forecasts, again suggesting that forecasting ensembles could be of great use.

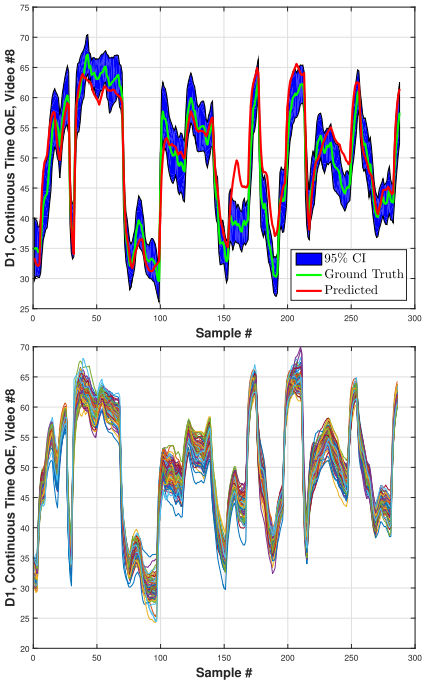

Fig. 9. The VN-QoE Predictor on video #8 of database D_1. Top: prediction using the best cross-validated model; bottom: predictions from all the models.

TABLE IV

OR Significance Testing ($m = 3$) on the Class of V-Predictors (Without Ensembles) on D_1 Using ST-RRED

Model Type	VN	VR	VH
VN	-	1	1
VR	0	-	-
VH	0	-	-

B. Quantitative Experiments - D_1

We begin our quantitative analysis by discussing the prediction performances of the compared QoE prediction models (class V-) on the LIVE HTTP Streaming Video Database (D_1). We first statistically compared the VN, VR and VH predictors in terms of OR when using ST-RRED (see Table IV). Among the three compared dynamic approaches, the VN-QoE Predictor consistently outperformed the VR and VH models. It has been previously demonstrated [56] that the NARX architecture is less sensitive than RNN models when learning long-term dependencies.

In D_1, there is no rebuffering in the distorted videos and hence it is straightforward to study the performance between various leading VQA models: PSNR, NIQE [35], VMAF (version 0.3.1) [32], MS-SSIM [76], SSIM [77] and ST-RRED [33] (see Table V).

Unsurprisingly, NIQE performed the worst across all dynamic approaches; after all, it is a no-reference frame-based video quality metric. PSNR delivered the second worst performance, but it does not capture any perceptual quality

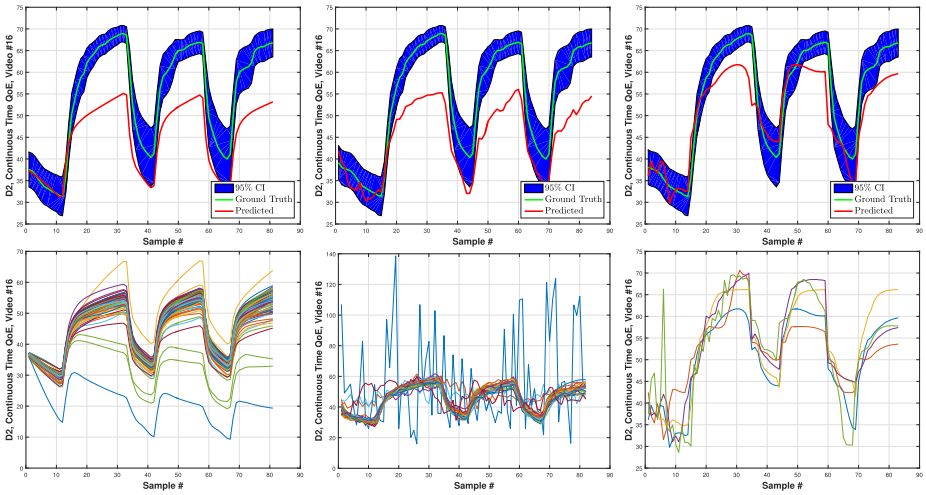

Fig. 10. Columns 1 to 3: The RMN-, RMR- and RMH-QoE Predictors applied to video #16 of database D_2. First row: prediction using the best cross-validated model; second row: predictions from all models.

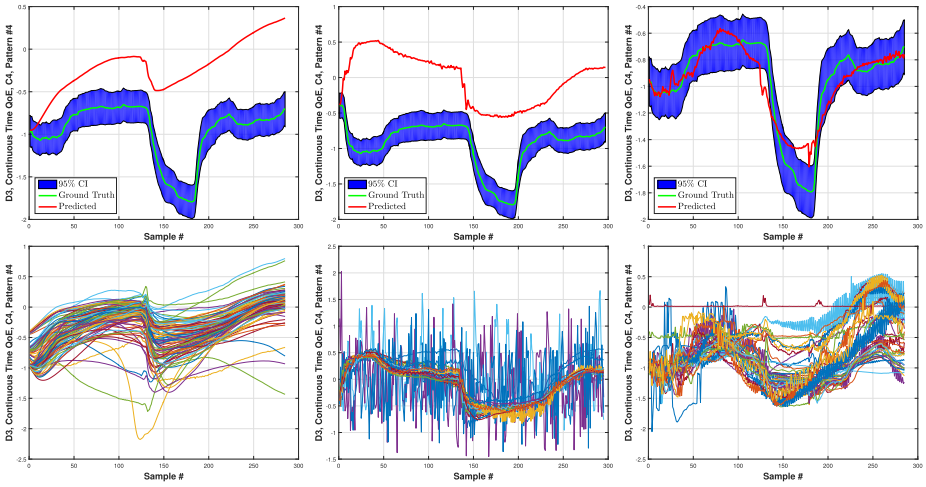

Fig. 11. Columns 1 to 3: The GN-, GR- and GH-QoE Predictors applied on pattern #4 of database D_3. First row: prediction using the best cross-validated model; second row: predictions from all models.

TABLE V

MEDIAN OR PERFORMANCE FOR THE CLASS OF V-QOE PREDICTORS ON DATABASE D_1 (SEE ALSO TABLE XVI)

Model Type	VN	VR	VH
NIQE [35]	34.79	42.84	42.78
PSNR	25.07	36.16	29.51
VMAF [32]	12.38	24.05	23.04
MS-SSIM [76]	5.73	17.64	31.82
SSIM [77]	5.46	17.43	30.69
ST-RRED [33]	5.90	20.81	15.31

information. MS-SSIM, SSIM and ST-RRED all performed well when deployed in the VN-QoE Predictor; but when it was inserted into the HW model, ST-RRED delivered the best performance. As shown in Table VI, the OR performance differences between VMAF 0.3.1, MS-SSIM, SSIM and ST-RRED were not statistically significant for the VN model; but all three of them performed significantly better than PSNR and NIQE. It should be noted that these statistical comparisons were performed at a very strict confidence level of $\frac{\alpha}{m} = \frac{0.05}{15}$ (due to Bonferroni correction with $m = 15$), hence these comparisons are conservative.

Our results show that perceptual VQA models, when combined with dynamic models that learn to conduct continuous-time QoE prediction, do not perform equally well; hence deploying high performance VQA models can contribute to improved QoE prediction. Deciding upon the choice of the VQA feature is application-dependent; yet we believe injecting perceptual VQA models into these models is much more beneficial than using QP or bitrate information.

We now study the efficacy of ensemble forecasting approaches. The naming convention of the ensemble methods is as follows: "best": pick best (from cross-validation) model parameters when testing, "avg": averaging of all forecasts, "med": taking the median of all forecasts, "mod": estimating the mode, "DTW-single": determining i_o in (2), "DTW-prob": probabilistic weighting of forecasts in (3).

Table VII shows that NARX again performed better than the other models across all ensemble methods. Using an ensemble method different than the mean yielded results similar to the mean. This suggests that the VN-QoE predictions were stable across different initializations and configurations (see also Fig. 9), given that more robust estimators such as the non-parametric mode produced results similar to the

TABLE VI

OR SIGNIFICANCE TESTING ($m = 15$) WHEN THE VN-QOE PREDICTOR WAS APPLIED ON D_1 ACROSS VARIOUS VQA MODELS. SIMILAR RESULTS WERE PRODUCED BY THE OTHER EVALUATION METRICS

Model	NIQE	PSNR	VMAF	MS-SSIM	SSIM	ST-RRED
NIQE	-	0	0	0	0	0
PSNR	1	-	0	0	0	0
VMAF	1	1	-	-	-	-
MS-SSIM	1	1	-	-	-	-
SSIM	1	1	-	-	-	-
ST-RRED	1	1	-	-	-	-

TABLE VII

MEDIAN OR PERFORMANCE FOR VARIOUS TIME-SERIES ENSEMBLE METHODS APPLIED ON THE CLASS OF V-PREDICTORS ON DATABASE D_1 USING ST-RRED (SEE ALSO TABLE XVII)

Model Type	VN	VR	VH
best	5.90	20.81	15.31
avg	5.25	15.59	14.69
med	5.25	9.15	13.99
mod	4.55	8.48	13.99
DTW-single	5.59	8.81	15.04
DTW-prob	5.25	10.51	14.69

mean ensemble which can be sensitive to outliers. Unlike VN and VH, using better ensemble estimators improved the OR performance of VR predictions by 5-10%. This may be explained by the larger uncertainty involved in the VR predictions, which is alleviated by our forecasting ensembles. Notably, determining the single best predictor using DTW in (2) performed better than the predictions based on the "best" model parameters during cross-validation. This verifies our earlier observation: the optimal model may vary over different data splits. The probabilistic weighting scheme in (3) delivered performance that was competitive with other ensemble methods, such as the median. Given that this scheme is also non-parametric and data-driven, these results are encouraging.

C. Quantitative Experiments - D_2

Next, we discuss our results on LIVE Mobile Stall Video Database-II (D_2) (see Table VIII). Overall, the RMN-QoE Predictor outperformed both the RMR and RMH-QoE Predictors, by achieving an excellent outage rate. We found these improvements to be statistically significant. Notably, using ensemble methods greatly improved OR (by more than 10% for both the RMR and RMH models) across all dynamic models. Using an ensemble method other than the mean led to a drop of OR by almost 15% in the case of the RMR-QoE Predictor. This again demonstrates the merits of using a forecasting ensemble for QoE prediction. Note that an outage rate of 0 does not mean that the prediction is perfect; it only indicates that the ensemble predictions were within two times the confidence interval.

We also compared the performance of the proposed continuous-time QoE predictors with a subset of the subjective predictions as an upper bound, as described in Section VII-D. We found that ensemble forecasts can improve on the prediction performance, but that there is still room for performance improvements (see Appendix II).

When tested on databases D_1 and D_2, the prediction performance of the proposed dynamic approaches was promising; especially when the predictions were combined in an ensemble. However, neither of these databases models both rebuffering events and video quality changes. In the next subsection, we explore the prediction performance of the studied QoE prediction models on the more challenging database D_3.

TABLE VIII

MEDIAN OR PERFORMANCE FOR VARIOUS TIME-SERIES ENSEMBLE METHODS APPLIED ON THE CLASS OF RM-PREDICTORS ON DATABASE D_2 (SEE ALSO TABLE XVIII)

Model Type	RMN	RMR	RMH
best	6.84	21.08	16.22
avg	0.00	11.48	3.71
med	0.00	6.62	4.29
mod	0.00	7.60	4.03
DTW-single	0.00	7.25	3.88
DTW-prob	0.00	7.25	3.38

TABLE IX

RMSE SIGNIFICANCE TESTING ($m = 3$) ON THE CLASS OF G-PREDICTORS (WITHOUT ENSEMBLES) ON D_3 USING ST-RRED

Model Type	GN	GR	GH
GN	-	1	0
GR	0	-	0
GH	1	1	-

TABLE X

MEDIAN RMSE PERFORMANCE FOR VARIOUS TIME-SERIES ENSEMBLE METHODS APPLIED ON THE CLASS OF G-PREDICTORS ON DATABASE D_3 USING ST-RRED (SEE ALSO TABLE XIX)

Model Type	GN	GR	GH
best	0.28	0.37	0.22
avg	0.24	0.29	0.16
med	0.29	0.29	0.11
mod	0.24	0.28	0.10
DTW-single	0.25	0.30	0.13
DTW-prob	0.24	0.29	0.12

D. Quantitative Experiments - D_3

We investigated the performance of the class of G-predictors applied to the more complex problem of QoE prediction when both rate drops and rebuffering occur by using database D_3. Due to rebuffering, computing VQA models is not possible without first removing the stalled frames from each distorted video. Using the publicly available metadata [78], we identified stalled frames and removed them from the distorted YUV video, then calculated the VQA feature, e.g. ST-RRED, on the luminance channels of the distorted and reference videos. As shown in see Table IX, the GH-QoE Predictor performed statistically better than the GN-QoE Predictor, while the GR-QoE Predictor lagged in performance. It is likely that more hidden neurons would enable the GN and GR models to perform better.

We also investigated the performance improvements of forecasting ensembles (see Table X). Overall, all forecasting ensembles greatly improved the performance of all dynamic models.

As with D_2, we also compared the performance of these QoE predictors with their upper bound (see Appendix II). Interestingly, we found that the ensemble predictions sometimes delivered better performance than the subjective upper bound; an observation that we revisit in Appendix II.

TABLE XI

OR COMPARISON BETWEEN DIFFERENT ACTIVATION FUNCTIONS WHEN TRAINING THE NARX COMPONENT ON D_1 (VN) AND ON D_2 (RMN). ROWS AND COLUMNS CORRESPOND TO THE ACTIVATION FUNCTION USED IN THE HIDDEN AND THE OUTPUT LAYER RESPECTIVELY

Database	D_1 (VN)			D_2 (RMN)		
Activation	tansig	logsig	purelin	tansig	logsig	purelin
tansig	10.38	20.28	5.90	10.59	31.38	7.68
logsig	8.97	22.55	5.10	10.34	33.26	7.92
purelin	9.28	31.28	11.10	26.04	50.55	5.48

TABLE XII

COMPARISON BETWEEN DIFFERENT TRAINING ALGORITHMS USING THE NARX COMPONENT ON DATABASES D_1 (VN) AND D_2 (RMN). THE NUMBER OF ITERATIONS WAS SET TO 1000

Database	D_1				D_2			
Metric	RMSE	OR	DTW	SROCC	RMSE	OR	DTW	SROCC
trainlm	4.00	5.72	15.39	0.91	4.43	7.47	4.15	0.93
trainbfg	3.90	4.86	14.73	0.90	6.33	17.53	6.65	0.81
trainrp	4.26	7.79	17.59	0.89	9.21	29.25	9.96	0.71
trainscg	4.20	6.04	16.53	0.89	6.59	21.09	7.21	0.79
traincgb	4.01	5.35	15.93	0.89	6.14	18.36	6.34	0.82
traincgf	4.27	6.86	16.54	0.88	6.11	18.57	6.59	0.82
traincgp	4.07	5.83	15.59	0.89	6.46	21.09	6.57	0.80
trainoss	4.49	6.97	18.14	0.87	7.19	24.27	7.30	0.80
traingdx	6.33	17.72	22.26	0.80	11.87	38.49	10.04	0.66

IX. CONCLUSIONS AND FUTURE WORK

In this work, we designed simple, yet efficient continuous-time streaming video QoE predictors by feeding QoE-aware inputs such as VQA measurements, rebuffering and memory information into dynamic neural networks. We explored three different dynamic model approaches: non-linear autoregressive models, recurrent neural networks and a block-based Hammerstein-Wiener model. To reduce forecasting errors, we also proposed ensemble forecasting approaches and evaluated our algorithms on three subjective video QoE databases. We hope that this work will be useful to video QoE researchers as they address the challenging aspects of continuous-time video QoE monitoring.

We now ask a more fundamental question: moving forward, which design aspect of these predictors is most important? Is it the choice of the dynamic model e.g. HW vs. NARX or selecting more sophisticated continuous-time features? The results in Tables VI and XIII, XVI (see Appendix II) demonstrate that a better VQA model (e.g. ST-RRED vs. MS-SSIM) or adding more rebuffering-related continuous-time inputs may not always yield statistically significant performance improvements. Tables V, VII, VIII, IX (and Tables XVII, XVIII and XIX in Appendix II) demonstrate that, among the three dynamic models, the RNN were consistently poorly performing while the performance differences between the NARX and HW components were not conclusive: on D_1 and D_2 the NARX-based predictors were better than HW, while for D_3 the HW component improved upon NARX. Meanwhile, using ensemble prediction methods yielded performance improvements in most cases by producing reliable and more robust forecasts. However, these improvements may not be significant if the individual forecasts are similar to each other.

In our preliminary experiments, we also discovered that when our proposed QoE prediction engines were trained on one publicly available database, then tested on another, they delivered poor performance likely due to their different design, e.g., only D_3 studies both rebuffering events and quality changes. This highlights an issue that is at the core of data-driven, continuous-time QoE prediction: lack of publicly-available and diverse subjective data. Existing databases, including D_3, are limited in that they do not sufficiently cover the large space of adaptation strategies, where time-varying quality, network conditions and buffer capacity are all tied together. Therefore, without large and more diverse subjective databases, introducing more sophisticated continuous-time inputs or deploying more complex neural networks will yield relatively small performance gains. We have recently launched a large subjective experiment to collect an adequate amount of such data, which will allow us to leverage even more sophisticated learning techniques as in [79] and potentially incorporate other inputs, such as quality switching. In the future, we envision building predictive models that exploit realistic network information extracted from the client side, i.e., developing databases and prediction models based on realistic network traces and bandwidth availability patterns. Ultimately, we seek to deploy methods that can perceptually optimize bitrate allocation and/or network and bandwidth usage, and that can be readily deployed in large streaming architectures.

TABLE XIII

MEDIAN PERFORMANCE FOR VARIOUS CONTINUOUS-TIME FEATURE SETS ON D_2 WHEN USING THE NARX LEARNER. NOTE THAT USING FEATURES $R_1 + R_2$ DEFINES THE RN-QoE PREDICTOR WHILE $R_1 + R_2 + M$ GIVES THE RMN-QoE PREDICTOR

Model	NARX			
Features/Metric	RMSE	OR	DTW	SROCC
R_1	4.65	9.03	4.00	0.94
R_2	8.38	31.15	7.29	0.82
M	6.74	23.12	6.39	0.82
R_1+R_2	4.41	8.14	4.02	0.95
R_1+M	4.86	12.12	4.26	0.92
R_2+M	6.41	21.53	6.17	0.84
R_1+R_2+M	4.49	6.84	4.08	0.93

APPENDIX I
IMPLEMENTATION DETAILS

The design of continuous-time QoE predictors often involves deciding upon a number of architecture-specific settings, including an imputation strategy, the activation function and the training algorithm. Next, we discuss these aspects and conclude with a note on computational complexity.

A. Inputs of Different Length

An important consideration when implementing the proposed model is accounting for different input durations. For example, while video quality predictions are computed on all frames of normal playback [16], the R_1 input (in the presence of rebuffering events) will have longer durations. While it is possible to train and evaluate the GN and GR QoE Prediction models without imputing missing VQA response values during rebuffering events, we found it useful to develop an imputation scheme that defines same-sized inputs for each test video. In previous studies, playback interruption has been found to be at least as annoying as very low bitrate distortions [21]; hence we selected imputed VQA values corresponding to very low video quality. Imputing with zeros is not a good idea;

TABLE XIV

MEDIAN PERFORMANCE FOR VARIOUS TIME-SERIES ENSEMBLE METHODS APPLIED ON THE CLASS OF RM-PREDICTORS ON DATABASE D_2 - DIRECT COMPARISON WITH HUMAN PERFORMANCE ("REF" ROW)

Model Type	RMN				RMR				RMH			
Model/Metric	RMSE	OR	DTW	SROCC	RMSE	OR	DTW	SROCC	RMSE	OR	DTW	SROCC
best	4.90	2.14	4.75	0.91	6.46	7.10	6.38	0.87	6.03	7.76	9.76	0.75
avg	4.34	0.00	3.85	0.95	5.74	2.13	4.55	0.93	4.61	1.33	5.85	0.87
med	4.46	0.00	3.71	0.94	5.56	1.08	3.86	0.95	4.39	1.35	6.23	0.86
mod	4.33	0.00	3.79	0.94	5.48	1.05	3.94	0.94	4.41	1.33	6.37	0.85
DTW-single	4.55	0.00	4.02	0.94	5.62	1.33	4.00	0.94	4.52	1.13	7.61	0.84
DTW-prob	4.40	0.00	3.78	0.95	5.62	1.18	3.96	0.95	4.57	1.16	5.72	0.87
ref	3.91	0.00	4.60	0.93	3.91	0.00	4.60	0.93	3.91	0.00	4.60	0.93

TABLE XV

MEDIAN PERFORMANCE FOR VARIOUS TIME-SERIES ENSEMBLE METHODS APPLIED ON THE CLASS OF G-PREDICTORS ON D_3 USING ST-RRED - DIRECT COMPARISON WITH HUMAN SCORES ("REF" ROW)

Model Type	GN				GR				GH			
Model/Metric	RMSE	OR	DTW	SROCC	RMSE	OR	DTW	SROCC	RMSE	OR	DTW	SROCC
best	0.29	5.48	28.39	0.78	0.38	9.65	29.52	0.69	0.24	2.37	25.56	0.76
avg	0.26	0.00	23.08	0.86	0.39	3.30	21.63	0.79	0.19	0.00	10.19	0.87
med	0.25	0.00	22.52	0.86	0.30	2.21	19.35	0.80	0.15	0.00	9.16	0.88
mod	0.25	0.00	22.03	0.85	0.30	2.17	19.94	0.80	0.14	0.00	9.28	0.89
DTW-single	0.26	0.00	19.99	0.86	0.31	3.10	21.09	0.80	0.16	0.00	13.82	0.85
DTW-prob	0.25	0.00	21.48	0.86	0.30	2.32	19.17	0.81	0.16	0.00	9.46	0.89
ref	0.20	0.00	10.71	0.90	0.20	0.00	10.71	0.90	0.20	0.00	10.71	0.90

some video quality models never approach such low values while others (such as ST-RRED) correspond lower values to better video quality. For simplicity, we picked the min (or the max) value of the video quality prediction corresponding to the worst quality level encountered over the entire video as the nominal VQA input value during playback interruptions. To recognize causality, we could also pick the min (or max) VQA values up until the rebuffering event occurs; we found that this did not greatly affect the final results. This imputing step is required only on the LIVE-NFLX dataset.

B. Activation Function

We experimented with various activation functions: logistic sigmoid (logsig), hyperbolic tangent sigmoid (tansig) and linear (purelin) and we also tried various combinations of them in the hidden and output layers. We carried out ten experiments and computed the median OR on D_1 and D_2. For D_1, we used $d_u = 10$, $d_y = 10$, a single hidden layer with 8 neurons and ST-RRED as the VQA model. For D_2, we used $d_u = 6$, $d_y = 6$, a single hidden layer with 8 neurons and the features R_1, R_2 and M. As shown in Table XI, using tansig for the hidden layer and purelin for the output layer proved to be good choices (in terms of OR) for this task on both databases. Other evaluation metrics produced similar results.

C. Training Algorithm

We compared the default Levenberg-Marquardt algorithm against other training algorithms [62]. Table XII shows that using the Levenberg-Marquardt (trainlm) performed very close to the best performing method on D_1 (trainbfg) and was significantly better on D_2. This suggests that the use of a general training algorithm such as Levenberg-Marquardt is sufficient for QoE prediction.

D. Computational Complexity

The proposed continuous-time QoE predictors require calculating perceptual VQA models, training and testing the neural network. Therefore, besides calculating the VQA feature, these neural networks can be trained offline and take up only a small computational overhead. To demonstrate this, we fixed the NARX architecture to $d_u = 10$ and $d_y = 10$ lags, $H = 8$ hidden nodes and a single hidden layer, then calculated the compute time for SSIM, for training and for testing the GN-QoE predictor on all 112 videos in D_3 (see Table XIV). All of the timing experiments were carried out on a 16.04 Ubuntu LTS Intel i7-4790@3.60 GHz system. Both the NARX and SSIM implementations used unoptimized Matlab code.

We found that calculating SSIM and training the neural network take up considerably more time (291 sec. and 5 sec. respectively) than testing it (0.04 sec.). Notably, calculating SSIM takes much more time than training, since we deployed relatively simple neural networks. For adaptive streaming applications, where the reference video and its compressed versions are readily available, the VQA measurements and the neural network training can be carried out in an offline fashion. Trained model values and associated VQA values can be sent to the client as part of the metadata and then the client side can perform such QoE predictions in real-time. Compared to simply calculating the VQA values, the only (online) computational overhead of the proposed predictors is the testing step, which is relatively fast. If the client side has low computational power, these operations could also be carried out by proxy "QoE-aware" servers.

The GN-QoE predictor uses ST-RRED as its VQA feature which, compared to SSIM, is a significantly better-performing VQA model [54], but its computational overhead may limit its potential in some practical applications. However, efficient approximations to ST-RRED that are implemented in the spatial domain are available [80].

APPENDIX II
ADDITIONAL EXPERIMENTAL ANALYSIS

In this section, we study in greater detail continuous-time performance bounds, the effects of using different rebuffering-related inputs for D_2 and provide more detailed results in Tables XVI, XVII and XIX.

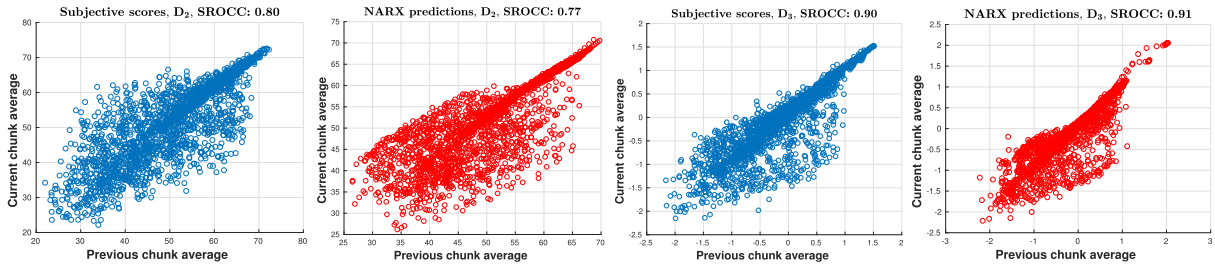

Fig. 12. Relationship between current and previous subjective and objective scores on D_2 and D_3. The objective predictions are able to capture the effects of recency.

TABLE XVI

MEDIAN PERFORMANCE FOR THE CLASS OF V-QOE PREDICTORS ON D_1

Model Type	VN				VR				VH			
Model/Metric	RMSE	OR	DTW	SROCC	RMSE	OR	DTW	SROCC	RMSE	OR	DTW	SROCC
NIQE [35]	8.61	34.79	29.72	0.54	9.71	42.84	49.42	0.33	8.95	42.78	55.86	0.27
PSNR	6.76	25.07	24.37	0.72	8.10	36.16	35.55	0.56	7.19	29.51	37.49	0.67
VMAF [32]	4.95	12.38	17.80	0.89	6.42	24.05	27.86	0.73	6.44	23.03	27.42	0.81
MS-SSIM [76]	4.07	5.73	15.89	0.91	5.79	17.64	23.67	0.73	7.50	31.82	44.86	0.59
SSIM [77]	4.02	5.45	14.22	0.90	6.07	17.43	24.13	0.74	7.32	30.69	41.78	0.67
ST-RRED [33]	4.25	5.90	15.21	0.90	6.98	20.81	27.22	0.71	5.40	15.31	27.09	0.87

TABLE XVII

MEDIAN PERFORMANCE FOR VARIOUS ENSEMBLE METHODS APPLIED ON THE CLASS OF V-PREDICTORS ON D_1 USING ST-RRED

Model Type	VN				VR				VH			
Model/Metric	RMSE	OR	DTW	SROCC	RMSE	OR	DTW	SROCC	RMSE	OR	DTW	SROCC
best	4.25	5.90	15.21	0.90	6.98	20.81	27.22	0.71	5.40	15.31	27.09	0.87
avg	3.64	5.24	14.11	0.91	4.99	15.59	17.64	0.85	4.86	14.69	16.72	0.90
med	3.69	5.24	14.01	0.91	4.23	9.15	16.31	0.90	4.85	13.99	16.46	0.90
mod	3.76	4.55	14.26	0.91	4.17	8.47	16.22	0.90	4.82	13.99	20.92	0.90
DTW-single	3.92	5.59	14.01	0.90	4.24	8.81	17.06	0.90	5.02	15.04	18.52	0.89
DTW-prob	3.67	5.25	14.11	0.91	4.20	10.51	16.35	0.89	4.84	14.69	16.72	0.90

TABLE XVIII

MEDIAN PERFORMANCE FOR VARIOUS ENSEMBLE METHODS APPLIED ON THE CLASS OF RM-PREDICTORS ON D_2

Model Type	RMN				RMR				RMH			
Model/Metric	RMSE	OR	DTW	SROCC	RMSE	OR	DTW	SROCC	RMSE	OR	DTW	SROCC
best	4.49	6.84	4.08	0.93	6.33	21.08	5.74	0.89	5.66	16.22	9.04	0.75
avg	4.01	0.00	2.99	0.97	5.59	11.48	3.83	0.95	4.20	3.71	5.43	0.88
med	3.88	0.00	2.93	0.97	5.38	6.62	3.19	0.96	3.79	4.29	5.73	0.87
mod	3.93	0.00	3.03	0.96	5.34	7.60	3.23	0.96	3.88	4.03	5.65	0.86
DTW-single	4.15	0.00	3.03	0.97	5.39	7.25	3.36	0.95	3.99	3.88	6.84	0.86
DTW-prob	3.91	0.00	2.96	0.97	5.33	7.25	3.31	0.96	4.05	3.38	5.10	0.88

TABLE XIX

MEDIAN PERFORMANCE FOR VARIOUS ENSEMBLE METHODS APPLIED ON THE CLASS OF G-PREDICTORS ON D_3 USING ST-RRED

Model Type	GN				GR				GH			
Model/Metric	RMSE	OR	DTW	SROCC	RMSE	OR	DTW	SROCC	RMSE	OR	DTW	SROCC
best	0.28	16.31	26.53	0.81	0.37	22.55	28.58	0.72	0.22	6.19	25.45	0.77
avg	0.24	8.31	19.82	0.88	0.29	14.87	20.11	0.81	0.15	0.33	8.08	0.90
med	0.24	6.66	21.65	0.89	0.29	13.90	18.47	0.82	0.11	0.00	7.43	0.91
mod	0.24	3.92	20.60	0.88	0.28	13.90	19.23	0.81	0.10	0.00	6.98	0.91
DTW-single	0.25	6.02	19.75	0.89	0.30	14.77	21.28	0.82	0.13	0.00	12.25	0.87
DTW-prob	0.24	6.54	20.00	0.89	0.29	14.31	18.90	0.82	0.12	0.00	7.45	0.91

A. Details on Continuous-Time Performance Bounds

Following the steps described in Section VII-D, we compared the best performing combination (RMN-QoE Predictor) against an upper bound, i.e., human performance, using $S = 10$ shuffles. Table XIII shows that the RMN-QoE Predictor outperformed both the RMR- and RMH-QoE Predictors, and its performance in terms of RMSE came close to the reference human performance. We found this difference to be statistically significant; hence there is some room for improvement. However, the performance in terms of OR was very good when any of the ensemble methods was considered. Surprisingly, the DTW and SROCC performances were not always inferior to human scores, and sometimes these differences were statistically significant.

Comparing the objective prediction scores between Tables XVIII and XIII, we discovered that, when using only a subset of the subjective scores as ground truth, the performance of the objective prediction models was reduced. This may be explained by the fact that subjects do not always agree with each other; hence using all of the subjective scores reduces both the objective and subjective uncertainty.

As in D_2, we also report the results compared against human performance in Table XIV for D_3. We drew similar observations as in Table XIII: the objective predictions tend to get worse while human performance usually upper bounds model performance. It is intriguing that combining the different GH-QoE forecasts delivered RMSE scores better than human performance - a difference which we found to be statistically significant. When objective prediction models are trained on subjective data, human performance should generally be superior to or at least statistically equivalent to objective predictions. However, this upper bound may be violated when we consider post-processed forecasting ensembles: human performance is the upper bound only on time-series predictions generated by an *individual* model. Our observation may be explained by the design of these two QoE databases. Database D_2 includes only rebuffering events, while D_3 involves a mixture of rebuffering and compression; a task that is even more challenging for human subjects. Therefore, the difficulty of the tasks may increase subjective uncertainty per test video; an uncertainty for which simple averaging of the continuous scores across subjects may not always be the best method of aggregating them. This reinforces our growing belief that simply averaging continuous QoE responses disregards the inherent non-linearities in these responses [21].

B. Rebuffering-Related Inputs

It has been shown [17] that combinations of VQA inputs (e.g. ST-RRED combined with SSIM) can deliver improved results. Here we investigate the effects of using different combinations of rebuffering-related inputs. We selected NARX as the dynamic model, and performed QoE predictions using a number of inputs ranging from one to three, as shown in Table XV. We also used the parameters from Table III. Notably, we found that only using the R_1 input contributed significantly greater prediction power than R_2 and M; it is capable of effectively capturing rebuffering effects and is suitable for being used alone in the GN-, GR- and GH-prediction models. Combining all three inputs improved the OR by only 2%. This suggests that R_1 is an efficient descriptor of the effects of rebuffering events on QoE.

C. Additional Tables

In this section we include the earlier described Tables XVI, XVII, XVIII and XIX.

D. Modeling Recency

To conclude this Appendix, we now show that the NARX-driven GN-QoE predictor is indeed able to capture recency effects in subjective QoE. To do so, we collected the GN-QoE predictions from D_2 and D_3, then performed a moving average operation, i.e., we averaged the predictions (and the subjective ground truths) at evenly-spaced moments separated by 10 and 5 seconds on D_2 and D_3 respectively, using corresponding sliding window sizes of 5 and 2.5 seconds respectively. Figure 12 shows that both the subjective and objective scores are very strongly correlated with preceding time averages, indicating that the objective GN-QoE predictions are indeed able to capture the effects of recency in subjective QoE.

ACKNOWLEDGEMENT

The authors thank Anush K. Moorthy for fruitful discussions on models of human performance on continuous-time QoE. They would also like to thank A. Aaron and the entire Video Algorithms team at Netflix for their support of this work.

REFERENCES

[1] *Cisco Visual Networking Index, 2016–2021 White Paper*. Accessed: Dec. 5, 2017. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

[2] O. Oyman and S. Singh, "Quality of experience for HTTP adaptive streaming services," *IEEE Commun. Mag.*, vol. 50, no. 4, pp. 20–27, Apr. 2012.

[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," Informational Internet-Draft 2582, Sep. 2011.

[5] 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.

[6] A. Beben, P. Wiśniewski, J. M. Batalla, and P. Krawiec, "ABMA+: Lightweight and efficient algorithm for HTTP adaptive streaming," in *Proc. Int. Conf. Multimedia Syst.*, 2016, Art. no. 2.

[7] 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, 2015.

[8] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hoßfeld, and P. Tran-Gia, "A survey on quality of experience of HTTP adaptive streaming," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 469–492, 1st Quart., 2015.

[9] S. Tavakoli, K. Brunnström, J. Gutiérrez, and N. García, "Quality of Experience of adaptive video streaming: Investigation in service parameters and subjective quality assessment methodology," *Signal Process., Image Commun.*, vol. 39, pp. 432–443, Nov. 2015.

[10] M.-N. Garcia *et al.*, "Quality of experience and HTTP adaptive streaming: A review of subjective studies," in *Proc. Int. Workshop Quality Multimedia Exper. (QoMEX)*, Sep. 2014, pp. 141–146.

[11] S. Tavakoli, S. Egger, M. Seufert, R. Schatz, K. Brunnström, and N. García, "Perceptual quality of HTTP adaptive streaming strategies: Cross-experimental analysis of multi-laboratory and crowdsourced subjective studies," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 8, pp. 2141–2153, Aug. 2016.

[12] R. R. Pastrana-Vidal, J. C. Gicquel, C. Colomes, and H. Cherifi, "Sporadic frame dropping impact on quality perception," *Proc. SPIE*, vol. 5292, pp. 182–193, Jun. 2004.

[13] 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.

[14] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*. Hoboken, NJ, USA: Wiley, 2015.

[15] T. C. Mills, *Time Series Techniques for Economists*. Cambridge, U.K.: Cambridge Univ. Press, 1991.

[16] C. G. Bampis, Z. Li, and A. C. Bovik, "Continuous prediction of streaming video QoE using dynamic networks," *IEEE Signal Process. Lett.*, vol. 24, no. 7, pp. 1083–1087, Jul. 2017.

[17] C. G. Bampis and A. C. Bovik, "An augmented autoregressive approach to HTTP video stream quality prediction." [Online]. Available: https://arxiv.org/abs/1707.02709

[18] C. Chen, L. K. Choi, G. de Veciana, C. Caramanis, R. W. Heath, Jr., and A. C. Bovik, "Modeling the time–varying subjective quality of HTTP video streams with rate adaptations," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 2206–2221, May 2014.

[19] 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, p. 95991-1–95991-8, Sep. 2015.

[20] 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.

[21] 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.

[22] J. Søgaard, S. Tavakolib, K. Brunnström, and N. García, "Open access subjective analysis and objective characterization of adaptive bitrate videos," in *Proc. IS&T Int. Symp. Electron. Imag.*, 2016, pp. 1–9.

[23] N. Staelens *et al.*, "Subjective quality assessment of longer duration video sequences delivered over HTTP adaptive streaming to tablet devices," *IEEE Trans. Broadcast.*, vol. 60, no. 4, pp. 707–714, Dec. 2014.

[24] S. Winkler, "Analysis of public image and video databases for quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 616–625, Oct. 2012.

[25] R. R. Pastrana-Vidal, J. C. Gicquel, J. L. Blin, and H. Cherifi, "Predicting subjective video quality from separated spatial and temporal assessment," *Proc. SPIE*, vol. 6057, pp. 276–286, Feb. 2006.

[26] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Process., Image Commun.*, vol. 19, no. 2, pp. 121–132, 2004.

[27] 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.

[28] M. H. Pinson, L. K. Choi, and A. C. Bovik, "Temporal video quality model accounting for variable frame delay distortions," *IEEE Trans. Broadcast.*, vol. 60, no. 4, pp. 637–649, Dec. 2014.

[29] K. Manasa and S. S. Channappayya, "An optical flow-based full reference video quality assessment algorithm," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2480–2492, Jun. 2016.

[30] P. V. Vu, C. T. Vu, and D. M. Chandler, "A spatiotemporal most-apparent-distortion model for video quality assessment," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 2505–2508.

[31] J. Y. Lin, T.-J. Liu, E. C.-H. Wu, and C.-C. J. Kuo, "A fusion-based video quality assessment (FVQA) index," in *Proc. Signal Inf. Process. Assoc. Annu. Summit Conf.*, Dec. 2014, pp. 1–5.

[32] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara. *Toward a Practical Perceptual Video Quality Metric*. Accessed: Dec. 5, 2017. [Online]. Available: http://techblog.netflix.com/2016/06/toward-practical-perceptual-video.html

[33] 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.

[34] Y. Kawayoke and Y. Horita, "NR objective continuous video quality assessment model based on frame quality measure," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2008, pp. 385–388.

[35] 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.

[36] 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.

[37] F. Yang, S. Wan, Y. Chang, and H. R. Wu, "A novel objective no-reference metric for digital video quality assessment," *IEEE Signal Process. Lett.*, vol. 12, no. 10, pp. 685–688, Oct. 2005.

[38] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 289–300, Jan. 2016.
[39] A. C. Bovik, "Automatic prediction of perceptual image and video quality," *Proc. IEEE*, vol. 101, no. 9, pp. 2008–2024, Sep. 2013.
[40] M. Leszczuk, M. Hanusiak, M. C. Q. Farias, E. Wyckens, and G. Heston, "Recent developments in visual quality monitoring by key performance indicators," *Multimedia Tools Appl.*, vol. 75, no. 17, pp. 10745–10767, Sep. 2016.
[41] S. Van Kester, T. Xiao, R. E. Kooij, O. Ahmed, and K. Brunnström, "Estimating the impact of single and multiple freezes on video quality," *Proc. SPIE*, vol. 7865, Feb. 2011.
[42] T. Hoßfeld, S. Egger, R. Schatz, M. Fiedler, K. Masuch, and C. Lorentzen, "Initial delay vs. interruptions: Between the devil and the deep blue sea," in *Proc. Int. Workshop Quality Multimedia Exper.*, Jul. 2012, pp. 1–6.
[43] 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*, Dec. 2011, pp. 494–499.
[44] D. Z. Rodríguez, J. Abrahão, D. C. Begazo, R. L. Rosa, and G. Bressan, "Quality metric to assess video streaming service over TCP considering temporal location of pauses," *IEEE Trans. Consum. Electron.*, vol. 58, no. 3, pp. 985–992, Aug. 2012.
[45] Z. Duanmu, K. Zeng, K. Ma, A. Rehman, and Z. Wang, "A quality-of-experience index for streaming video," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 154–166, Feb. 2017.
[46] C. G. Bampis and A. C. Bovik, "Learning to predict streaming video QoE: Distortions, rebuffering and memory," *Signal Process., Image Commun.*, to be published. [Online]. Available: https://arxiv.org/abs/1703.00633
[47] K. D. Singh, Y. Hadjadj-Aoul, and G. Rubino, "Quality of experience estimation for adaptive HTTP/TCP video streaming using H.264/AVC," in *Proc. Consum. Commun. Netw. Conf.*, Jan. 2012, pp. 127–131.
[48] *Video Quality Experts Group (VQEG)*. Accessed: Dec. 5, 2017. [Online]. Available: https://www.its.bldrdoc.gov/vqeg/projects/audiovisual-hd.aspx
[49] *Parametric Bitstream-Based Quality Assessment of Progressive Download and Adaptive Audiovisual Streaming Services Over Reliable Transport*, document P.1203, Dec. 2017. [Online]. Available: https://www.itu.int/rec/T-REC-P.1203-201611-I
[50] A. Raake, M.-N. Garcia, W. Robitza, P. List, S. Göring, and B. Feiten, "A bitstream-based, scalable video-quality model for HTTP adaptive streaming: ITU-T P.1203.1," in *Proc. Int. Conf. Quality Multimedia Exper. (QoMEX)*, May/Jun. 2017, pp. 1–6.
[51] 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.*, Oct. 2014, pp. 2007–2011.
[52] D. S. Hands and S. E. Avons, "Recency and duration neglect in subjective assessment of television picture quality," *Appl. Cognit. Psychol.*, vol. 15, no. 6, pp. 639–657, 2001.
[53] A. J. Greene, C. Prepscius, and W. B. Levy, "Primacy versus recency in a quantitative model: Activity is the critical distinction," *Learn. Memory*, vol. 7, no. 1, pp. 48–57, 2000.
[54] A. C. Bovik, R. Soundararajan, and C. G. Bampis. *On the Robust Performance of the ST-RRED Video Quality Predictor*. Accessed: Dec. 5, 2017. [Online]. Available: http://live.ece.utexas.edu/research/Quality/ST-RRED/
[55] A. Rehman and Z. Wang, "Perceptual experience of time-varying video quality," in *Proc. Int. Workshop Quality Multimedia Exper.*, Jul. 2013, pp. 218–223.
[56] T. Lin, B. G. Horne, P. Tiňo, and C. L. Giles, "Learning long-term dependencies in NARX recurrent neural networks," *IEEE Trans. Neural Netw.*, vol. 7, no. 6, pp. 1329–1338, Nov. 1996.
[57] H. T. Siegelmann, B. G. Horne, and C. L. Giles, "Computational capabilities of recurrent NARX neural networks," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 27, no. 2, pp. 208–215, Apr. 1997.
[58] K. S. Narendra and K. Parthasarathy, "Identification and control of dynamical systems using neural networks," *IEEE Trans. Neural Netw.*, vol. 1, no. 1, pp. 4–27, Mar. 1990.
[59] Mathworks. *Design Time Series NARX Feedback Neural Networks*. Accessed: Dec. 5, 2017. [Online]. Available: https://www.mathworks.com/help/nnet/ug/design-time-series-narx-feedback-neural-networks.html
[60] K. Levenberg, "A method for the solution of certain non-linear problems in least squares," *Quart. Appl. Math.*, vol. 2, no. 2, pp. 164–168, Jul. 1944.
[61] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *J. Ind. Appl. Math.*, vol. 11, no. 2, pp. 431–441, 1963.
[62] Mathworks. *Choose a Multilayer Neural Network Training Function*. Accessed: Dec. 5, 2017. [Online]. Available: https://www.mathworks.com/help/nnet/ug/choose-a-multilayer-neural-network-training-function.html
[63] Mathworks. *Improve Neural Network Generalization and Avoid Overfitting*. Accessed: Dec. 5, 2017. [Online]. Available: https://www.mathworks.com/help/nnet/ug/improve-neural-network-generalization-and-avoid-overfitting.html
[64] M. Leutbecher and T. N. Palmer, "Ensemble forecasting," *J. Comput. Phys.*, vol. 227, no. 7, pp. 3515–3539, Mar. 2008.
[65] G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," *Neurocomputing*, vol. 50, pp. 159–175, Jan. 2003.
[66] Z.-H. Zhou, J. Wu, and W. Tang, "Ensembling neural networks: Many could be better than all," *Artif. Intell.*, vol. 137, nos. 1–2, pp. 239–263, May 2002.
[67] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation and active learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 7. 1995, pp. 231–238.
[68] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Proc. KDD Workshop*, 1994, pp. 359–370.
[69] J. H. Stock and M. W. Watson, "Combination forecasts of output growth in a seven-country data set," *J. Forecasting*, vol. 23, no. 6, pp. 405–430, Sep. 2004.
[70] N. Kourentzes, D. K. Barrow, and S. F. Crone, "Neural network ensemble operators for time series forecasting," *Expert Syst. Appl.*, vol. 41, no. 9, pp. 4235–4244, Jul. 2014.
[71] J. L. Elman, "Finding structure in time," *Cognit. Sci.*, vol. 14, no. 2, pp. 179–211, Mar. 1990.
[72] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 855–868, May 2009.
[73] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proc. INTERSPEECH*, 2014, pp. 338–342.
[74] 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. Global Conf. Signal Inf. Process.*, Dec. 2014, pp. 989–993.
[75] S. Siegel, *Nonparametric Statistics for the Behavioral Sciences*. New York, NY, USA: McGraw-Hill, 1956.
[76] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, Nov. 2003, pp. 1398–1402.
[77] 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.
[78] C. G. Bampis, Z. Li, A. Moorthy, I. Katsavounidis, A. Aaron, and A. C. Bovik. (2017). *Netflix Video Quality of Experience Database*. [Online]. Available: http://live.ece.utexas.edu/research/LIVE_NFLXStudy/nflx_index.html
[79] H. Mao, R. Netravali, and M. Alizadeh, "Neural adaptive video streaming with Pensieve," in *Proc. SIGCOMM*, 2017, pp. 197–210.
[80] C. G. Bampis, P. Gupta, R. Soundararajan, and A. C. Bovik, "SpEED-QA: Spatial efficient entropic differencing for image and video quality," *IEEE Signal Process. Lett.*, vol. 24, no. 9, pp. 1333–1337, Sep. 2017.

Christos G. Bampis is currently pursuing the Ph.D. degree with the Laboratory for Image and Video Engineering with The University of Texas at Austin. His research interests include image and video quality assessment, and quality of experience in adaptive video streaming.

Zhi Li is currently with the Video Algorithms Group, Netflix Inc., improving video streaming experience for consumers and using knowledge of human video quality perception to design and optimize encoding/streaming systems.

Ioannis Katsavounidis is currently a Senior Research Scientist with the Video Algorithms Group, Netflix Inc., involved in video quality and video codec optimization problems.

Alan C. Bovik (F'96) is currently a Professor with The University of Texas at Austin. He has received many awards, including the 2015 Primetime Emmy Award from the Television Academy.