Recurrent and Dynamic Networks that Predict Streaming Video Quality of Experience

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

***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 human observer is the end-user; 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 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 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 learn 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.

***Index Terms*—subjective quality, objective quality, quality of experience, video quality assessment, 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]–[6] that divide streaming video content into chunks (represented at various quality levels); whereby the quality level (or representation) to be played next 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.

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 ARMA models [7], [8] 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 excited 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.

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 networks. Building on preliminary work in [9], [10], we advance progress towards this goal by devising efficient QoE predictions engines employing dynamic networks including recurrent neural networks, NARX [9], [10] and Hammerstein Wiener models [11], [12]. 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

C. G. Bampis and A. C. Bovik are with the Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, USA (e-mail: bampis@utexas.edu; bovik@ece.utexas.edu). Zhi Li and Ioannis Katsavounidis are with Netflix Inc. (e-mail: {zli,ikatsavounidis}@netflix.com). This work is supported by Netflix Inc.

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, 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 the various 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 with discussions regarding 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]–[18]. 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) [19]–[25], reduced-reference (RR) [26] or no-reference (NR) [27]–[32], depending on whether all or part of a pristine reference video is used in the assessment process. Besides video quality degradations, retrospective QoE is also affected by playback interruptions; hence retrospective predictive models have been proposed that compute global rebuffering-related features, such as the number or durations of rebuffering events [33], [34]. Hybrid approaches that model video quality degradations and rebuffering events have very recently been studied, resulting in models like SQI [35] and the learning-based Video ATLAS [36].

Continuous-time QoE prediction has received much less attention and is a more challenging problem. In [11], a Hammerstein-Wiener dynamic model was used to make continuous-time QoE predictions on videos afflicted only by dynamic rate changes. In [10], 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 [37], 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 [12]. Only rebuffering-aware inputs were considered, using a simple model selection strategy. 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 [9], 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 experimental validation carried out only on a single dynamic model on a single subjective database. Since predictive models designed or learned and tested a specific dataset run the risk of inadvertant "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 prediction models. Optimal model parameters can significantly vary across different test videos; hence carefully designed 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 network
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 the *optimal, continuous-time* human performance

Our goal here is to surmount 1-7, 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 [15] may cause annoying visual artifacts [13], [14] thereby affecting subjective QoE.
2) Playback interruption: frequent or long rebuffering events adversely affect subjective QoE [33]. Compared to degradations on visual quality, rebuffering events have remarkably different effects on subjective QoE [15].

3) Memory (or hysteresis) effects: Recency [15], [38] 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 [39].

Broadly, subjective QoE "is a non-linear aggregate of video quality, rebuffering information and memory" [9]. Recently, the learning-driven Video ATLAS model [36] proposed to combine these different sources of information to predict QoE in general streaming environments where rebuffering events and video quality changes are commingled. However, 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:

(a) At least three types of "QoE-aware inputs" must be fused: VQA model responses, rebuffering measurements and memory effects.
(b) 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 [40] or PSNR. QoE-rich information can reduce the number of necessary inputs and boost the general capabilities of the QoE predictor.
(c) Dynamic networks with memory are able to capture recency (or memory) which is an inherent property of QoE.
(d) These dynamic networks 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.
(e) 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. In the following sections, we motivate and explain the unique features of this new method.

IV. The GN-QoE Predictor

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) ST-RRED is used as the VQA model. Previous studies [9], [15], [36], [41], have shown that ST-RRED is an excellent indicator of video quality. As was done in [10], it is straightforward to augment the GN-QoE Predictor by introducing additional QoE-aware inputs, if they verifiably contribute QoE prediction power. 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 side video-information.
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 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 [9], [42], [43]. 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 $(\mathbf{u}_t)$:

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

where $f(\cdot)$ is a non-linear function of previous inputs $\{y_{t-1}, y_{t-2}, ..., y_{t-d_y}\}$, and previous (and current) external variables $\{\mathbf{u}_t, \mathbf{u}_{t-1}, \mathbf{u}_{t-2}, ..., \mathbf{u}_{t-d_u}\}$, where d_y is the number of lags in the input and d_u is the number of lags in the external variables.

In a NARX model, there are two types of inputs: past outputs that are fed back as future inputs to the dynamic network, 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 [44] 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 RHS 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 [10]. 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 network to be trained is feed-forward; hence static backpropagation can be used [45].

It has been shown [10] 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

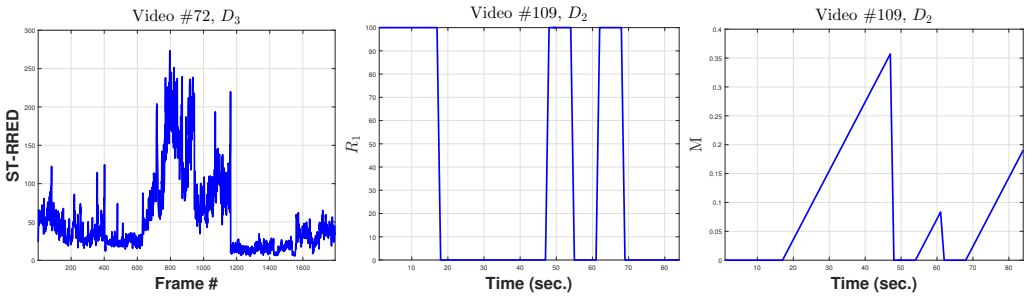


Fig. 1: Examples of the proposed continuous time QoE variables measured on videos from all of the considered databases. Left to right: ST-RRED computed on video #72 of the LIVE-NFLX Video QoE Database (denoted by D_3), and R_1 and M on the LIVE Mobile Stall Video Database-II (denoted by D_2).

inputs and external variables in a non-linear fashion. Given that the problem is of medium size, we chose the Levenberg-Marquardt [46], [47] algorithm to train the model [48]. To reduce the chances of overfitting in the OL training step, we used an early stopping approach [49]: the first 80% of the samples were used to train the OL NARX, while the remaining 20% were used to validate it.

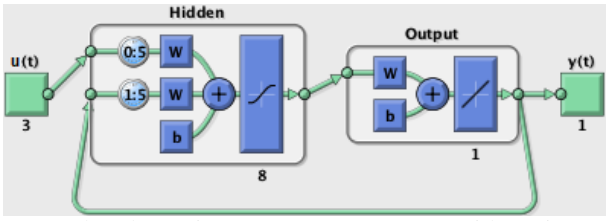


Fig. 2: The dynamic CL NARX system with 3 inputs, 8 neurons in the hidden layer and 5 feedback delays. Note that the recurrency of the NARX occurs in the output layer [45].

The GN-QoE Predictor follows a learning-driven approach which requires careful validation and design. However, 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 [50], [51] and neural network ensembles [52], [53]. 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 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 [7], [8], 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.

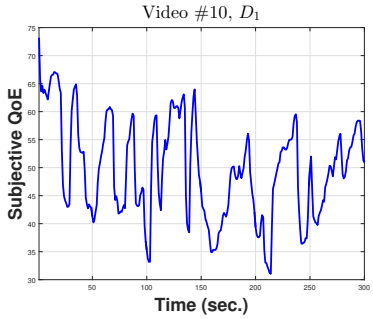


Fig. 3: Exemplar subjective QoE scores on video #10 from the LIVE HTTP Streaming Video Database (denoted by D_1).

2) Using 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 in complex ways, 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 the GN-QoE using κ different combinations of model orders d_u and d_y, λ different 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.

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 [54] 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 [15], we hypothesize that $\nu_i = \sum_j \mathbf{D}_{ij}$, 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

$$i_o = \arg\min_i \nu_i, \tag{2}$$

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 validation step). The second approach is to assign a probabilistic weight to each of the C candidate predictors, i.e.

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

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 [55], while median and mode ensembles are more robust against outliers [56].

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 [41]. Indeed, it allows the use of any VQA model (FR, RR or NR), depending on the available reference information. As in [9], [36], 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 network approaches. For example, Layer-Recurrent Neural Networks (denoted here as RNNs) [57] or the Hammerstein-Wiener (HW) dynamic model [11], [12] can be used instead of NARX, yielding GR-QoE and GH-QoE models. 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 I) may be considered as a special case of [11]. We summarize the proposed family of G-predictors and other predictors that use subset of these inputs, and their characteristics in Table I. 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. Learning GR-QoE

Recurrent Neural Networks (RNNs) [57] have recently gained popularity due to their successful applications to various tasks such as handwriting recognition [58] and speech recognition [59]. 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. An example of the network is shown in Fig. 4.

TABLE I: 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.

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

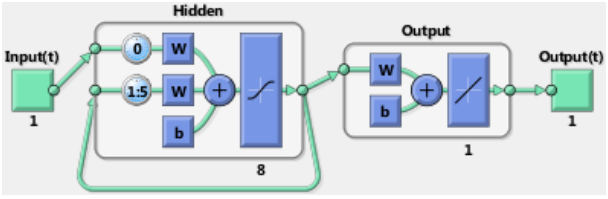


Fig. 4: The dynamic RNN approach with 1 input, 8 neurons in the hidden layer and 5 layer delays. Note that the recurrency of the RNN occurs in the hidden layer rather than in the output layer [45].

Given that the amount of available subjective data is insufficient to train a deep network, we decided to train relatively simple RNN networks, i.e., 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. Learning GH-QoE

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 [11] or rebuffering events [12]. 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.

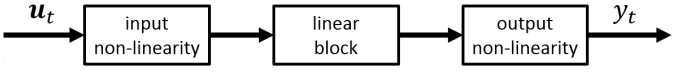


Fig. 5: The HW dynamic approach.

The family of G- QoE predictors (see Table I) 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. Experimental Setup

A. Subjective Video QoE Databases

In [11], 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 [60] using a different database (denoted by D_2), where diverse rebuffering patterns were inserted into 24 different video contents of various spatial resolutions. Unlike [11], 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 [4]–[6]. Towards bridging this gap, we built the new LIVE-NFLX Video QoE Database [15] (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. Examples of test videos from all three databases are shown in Fig. 6.

B. Validation Framework

Next we discuss our validation scheme. Notably, the proposed recurrent models are highly non-linear; hence the traditional time series model estimation techniques used in ARMA models [7] 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 II). However, 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 or validating, we have found that different initial weights produce different results for GN and GR QoE Predictors. As a result, the performance of both the GN and GR QoE Predictors should be estimated across initializations. By comparison, previous continuous-time QoE prediction models [9], [11], [12] have used a single model order.

Here we propose a novel validation framework that is suitable for *streaming video* QoE prediction. This idea builds on a simpler approach that was introduced in [10]. Let $i = 1 \dots N$ index the video in a given subjective 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 the same video content. Depending on which subjective database is used, we applied the following steps. From the LIVE-NFLX dataset, we removed all videos having either the same content or the same distortion pattern [9]. From the other two databases, we removed all videos having the same content. This process yielded a set of training QoE time series for each test video.

Next, we further divided the training set further into a second training set and a validation set for testing. 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, 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 validation step. To account for different weight initializations, we repeated the training process T times; then averaged the performances across initializations.

TABLE II: 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.

Network	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

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 [20] and in hybrid models of retrospective QoE [35], [36], 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 [10]–[12].

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.

Fig. 6: Exemplar videos from the three QoE databases used in the experiments on: (a). LIVE HTTP Streaming Video Database, (b). LIVE Mobile Stall Video Database-II, (c). LIVE-NFLX Video QoE Database.

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 and a predicted QoE waveform w. 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 subject QoE contains strong time dependencies and inherent non-stationarities.

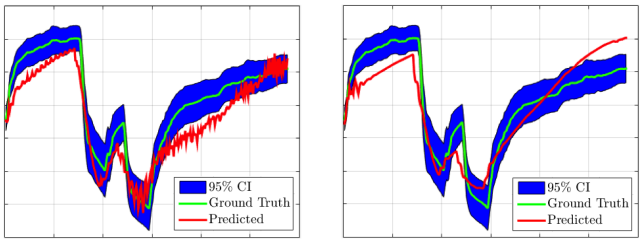
There are other evaluation metrics that are more suitable for time series comparisons, e.g., the root mean squared error (RMSE), which captures the overall signal fidelity, and the outage rate (OR) [11], which measures the frequency of occurrence of the predicted values falling outside twice the confidence interval of g. Since we are interested in capturing QoE trends, the dynamic time warping (DTW) distance could also be employed [9], [15], [54]. 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.
3) The 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. 7, 8 and 9. Figure 7 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 8 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. 9). Clearly, any single evaluation metric is likely to insufficiently descriptive of performance; hence we report all three of these metrics, along with the SROCC to draw a clearer picture of relative performance.

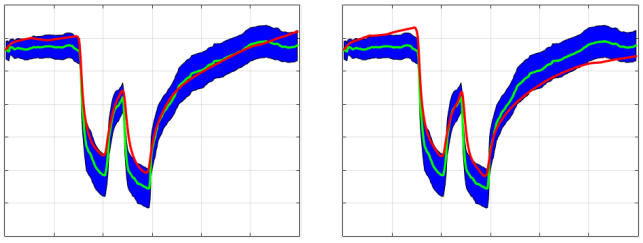
D. Continuous-time Performance Bounds

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 bound on prediction performance of the ith test video. Finally, to obtain a single bound performance over a given database, we calculated the median value over all test videos.



(a) OR = 5.9028 (b) OR = 13.1944

Fig. 7: Vertical axis: QoE; horizontal axis: time. OR is an intuitive metric; but it does not adequately describe the prediction error behavior between the confidence bounds.



(a) DTW = 2.9571 (b) DTW = 19.5554

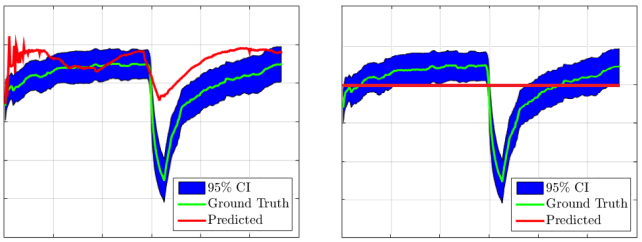
Fig. 8: Vertical axis: QoE; horizontal axis: time. DTW better reflects the temporal trends of the prediction error although it is harder to interpret.

E. 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

TABLE III: Median performance metrics for the class of V- QoE predictors on database D_1. The best result per evaluation metric for each dynamic network is in boldface.

Network Type	VN				VR				VH			
Model/Metric	RMSE	OR	DTW	SROCC	RMSE	OR	DTW	SROCC	RMSE	OR	DTW	SROCC
NIQE [28]	8.6113	34.7917	29.7234	0.5439	9.7064	42.8378	49.4228	0.3273	8.9473	42.7778	55.8639	0.2710
PSNR	6.7613	25.0694	24.3740	0.7172	8.1003	36.1616	35.5527	0.5561	7.1937	29.5105	37.4949	0.6669
VMAF [25]	4.9518	12.3776	17.7980	0.8909	6.4212	24.0541	27.8561	0.7300	6.4390	23.0345	27.4229	0.8076
MS-SSIM [61]	4.0672	5.7343	15.8909	**0.9057**	**5.7919**	17.6431	**23.6744**	0.7325	7.4958	31.8182	44.8634	0.5886
SSIM [62]	**4.0224**	**5.4545**	**14.2184**	0.8951	6.0686	**17.4324**	24.1329	**0.7406**	7.3203	30.6897	41.7781	0.6722
ST-RRED [26]	4.2451	5.9028	15.2087	0.9044	6.9833	20.8136	27.2188	0.7059	**5.3953**	**15.3103**	**27.0865**	**0.8658**



(a) RMSE = 0.3627 (b) RMSE = 0.3294

Fig. 9: Vertical axis: QoE; horizontal axis: time. RMSE effectively captures the overall signal fidelity; but it does not effectively account for the local temporal structure of the prediction error.

frames of normal playback [9], 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 [15]; hence we selected imputed VQA values corresponding to very low video quality. Imputing with zeros is not a good idea; 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.

VIII. Experimental Results

In this section, we thoroughly evaluate and compare between the different approaches. 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.

A. Qualitative Experiments

We begin by visually evaluating the different models on a few videos from all three QoE databases. Figure 10 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 models yielded similar outputs. In such cases, it may be that forecasting ensembles yield little benefit.

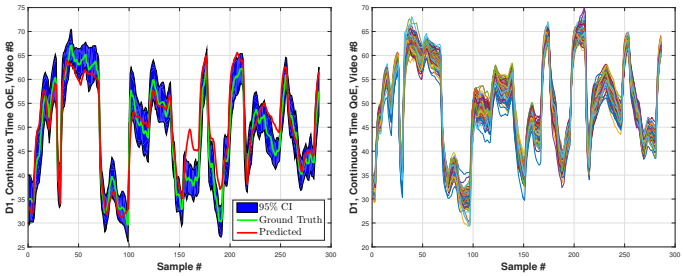


Fig. 10: The VN-QoE Predictor on video #8 of database D_1. Left: prediction using the best cross-validated model; right: predictions from all the models.

TABLE IV: Wilcoxon significance test [63] (using significance level $\alpha = 0.05$) on various VQA models applied on database D_1 when the OR metric was used to assess the VN-QoE Predictor. 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. 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	-	-	0	0
MS-SSIM	1	1	-	-	-	-
SSIM	1	1	1	-	-	-
ST-RRED	1	1	1	-	-	-

By contrast, Fig. 11 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. However, in this example, the RMH-QoE Predictor performed well.

The example in Fig. 12 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 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.

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). Table III summarizes the performance of all three dynamic approaches when using leading VQA models: PSNR, NIQE

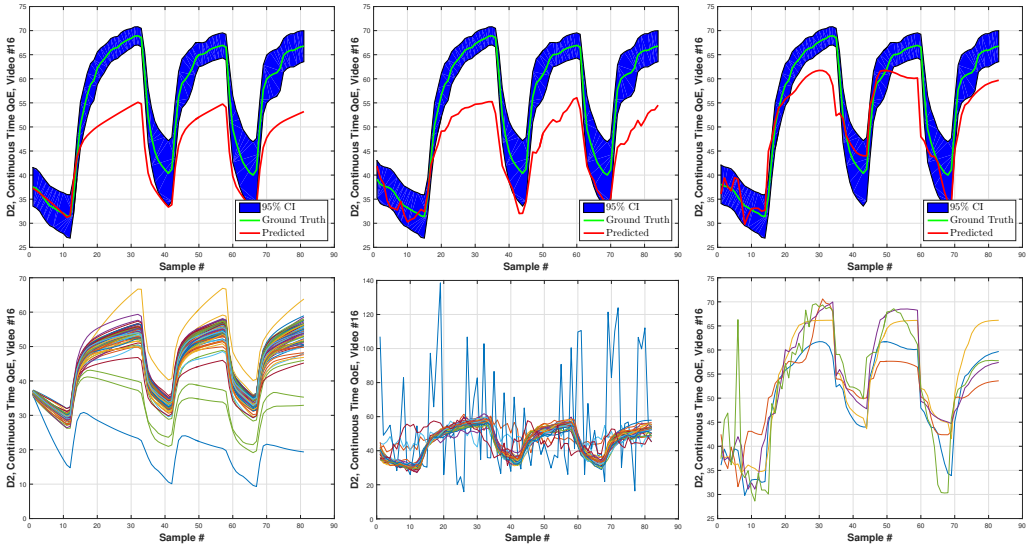


Fig. 11: Columns 1 to 3: The RMN-, RMR- and RMH-QoE Predictors applied to video #16 of database D_2. First row: prediction using the best cross-validated model; second row: predictions from all models.

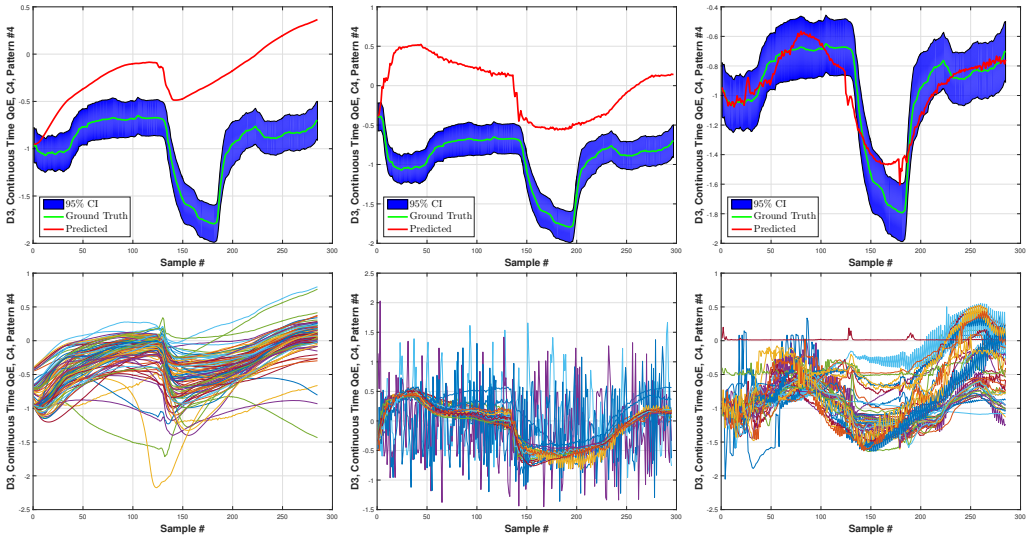


Fig. 12: Columns 1 to 3: The GN-, GR- and GH-QoE Predictors applied on pattern #4 of database D_3. First row: prediction using the best cross-validated model; second row: predictions from all the models.

[28], VMAF (version 3.1) [25], MS-SSIM [61], SSIM [62] and ST-RRED [26]. Unsurprisingly, NIQE performed the worst across all dynamic approaches; but it is a no-reference frame-based video quality metric. PSNR delivered the second worst performance, but it does not capture any perceptual quality 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. We also carried out statistical validation tests, as shown in Table IV. We found that for the VN model, the performance differences between MS-SSIM, SSIM and ST-RRED were not statistically significant; but all three of them performed better than VMAF 3.1, PSNR and NIQE. These results show that perceptual VQA models, when combined with dynamic networks that learn to conduct continuous-time QoE prediction, do not perform equally well; hence deploying high performance VQA models can greatly contribute to improved QoE prediction. Among the three compared dynamic approaches, the VN-QoE Predictor consistently outperformed the VR and VH models. It has been previously demonstrated [42] that the NARX architecture is less sensitive than RNN models when learning long-term dependencies. Notably, VH performed poorly when fed by all of the VQA models other than ST-RRED.

We now study the efficacy of ensemble forecasting approaches. Table V shows that NARX again performed better than the other networks across all ensemble methods. However, 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. 10), given that more robust estimators such as the non-parametric mode produced results similar to the mean ensemble which can be sensitive to outliers. Unlike VN and VH, using better ensemble estimators improved the performance of VR predictions. This may be explained by the larger uncertainty involved in the VR

TABLE V: Median performance metrics for various time series ensemble methods applied on the class of V- predictors on database D_1 using ST-RRED. The best result overall is in boldface. The naming convention of the ensemble methods is as follows: "best": pick best (from 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).

Network Type	VN				VR				VH			
Model/Metric	RMSE	OR	DTW	SROCC	RMSE	OR	DTW	SROCC	RMSE	OR	DTW	SROCC
best	4.2451	5.9028	15.2087	0.9044	6.9833	20.8136	27.2188	0.7059	5.3953	15.3103	27.0865	0.8658
avg	**3.6415**	5.2448	14.1071	**0.9113**	4.9875	15.5932	17.6393	0.8482	4.8605	14.6853	16.7186	0.8978
med	3.6851	5.2448	**14.0136**	0.9110	4.2293	9.1525	16.3132	0.8980	4.8455	13.9860	16.4650	0.8957
mod	3.7551	**4.5455**	14.2622	0.9094	4.1696	8.4746	16.2175	0.8985	4.8216	13.9860	20.9198	0.8958
DTW-single	3.9151	5.5944	**14.0136**	0.8997	4.2386	8.8136	17.0555	0.8953	5.0198	15.0350	18.5177	0.8898
DTW-prob	3.6727	5.2448	14.1049	0.9111	4.1989	10.5085	16.3487	0.8938	4.8414	14.6853	16.7181	0.8960

predictions. Notably, determining the single best predictor using DTW in eq. (2) performed better than the predictions based on the "best" model parameters during validation. This verifies our earlier observation: the optimal model may vary over different data splits. The probabilistic weighting scheme in eq. (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 VI). Overall, the NARX learner again produced the best results. The RMN-QoE Predictor outperformed both the RMR and RMH-QoE Predictors. Notably, using ensemble methods greatly improved OR (by more than 10% for both the RMR and RMH models), DTW and SROCC across all dynamic network approaches. Using an ensemble method other than the mean led to a drop of OR by almost 5% in the case of the RMR-QoE Predictor. The DTW-based probabilistic weighting scheme further improved the RMH-QoE Predictor's prediction performance. 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.

Using the previous results, we were able to compare among different objective metrics. Following the steps described in Section VII-D, we compared the best performing combination (RMN-QoE Predictor) against the upper bound, i.e., human performance, using $S = 10$ shuffles. Table VII 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. We review this observation in Section VIII-D.

Comparing the objective prediction scores between Tables VI and VII, 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.

It has been shown [10] that combinations of VQA inputs (e.g. ST-RRED combined with SSIM) can deliver improved results. Here we investigate this claim by studying the effects of using different combinations of rebuffering-related inputs. We selected NARX as the network architecture and performed QoE predictions using a number of inputs ranging from one to three, as shown in Table VIII. We also used the parameters from Table II. 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.

TABLE VIII: Median performance metrics 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. The best result per evaluation metric is in bold.

Network	NARX			
Features/Metric	RMSE	OR	DTW	SROCC
R_1	4.6529	9.0333	**4.0030**	**0.9374**
R_2	8.3789	31.1519	7.2924	0.8216
M	6.7418	23.1209	6.3926	0.8208
R_1+R_2	**4.4114**	8.1367	4.0149	0.9469
R_1+M	4.8595	12.1199	4.2619	0.9177
R_2+M	6.4060	21.5301	6.1705	0.8405
R_1+R_2+M	4.4928	**6.8439**	4.0751	0.9268

Note that 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 QoE database D_3.

D. Quantitative Experiments - D_3

We investigated the performance of ensemble methods on the class of G- predictors applied on the more complex problem of QoE prediction when both rate drops and rebuffering occur (see Table IX) by using database D_3. We found that overall, the GH-QoE Predictor performed 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. Overall, all forecasting ensembles greatly improved the performance of all

TABLE VI: Median performance metrics for various time series ensemble methods applied on the class of RM- predictors on database D_2. The best result per evaluation metric is in boldface.

Network Type	RMN				RMR				RMH			
Model/Metric	RMSE	OR	DTW	SROCC	RMSE	OR	DTW	SROCC	RMSE	OR	DTW	SROCC
best	4.4928	6.8439	4.0751	0.9268	6.3343	21.0842	5.7353	0.8897	5.6628	16.2211	9.0351	0.7526
avg	4.0087	**0.0000**	2.9912	**0.9689**	5.5937	11.4835	3.8338	0.9494	4.2038	3.7050	5.4336	0.8808
med	**3.8806**	**0.0000**	**2.9338**	0.9655	5.3789	6.6156	3.1905	0.9632	3.7896	4.2892	5.7285	0.8708
mod	3.9249	**0.0000**	3.0279	0.9618	5.3419	7.5962	3.2318	0.9600	3.8766	4.0270	5.6462	0.8599
DTW-single	4.1542	**0.0000**	3.0277	0.9661	5.3938	7.2500	3.3605	0.9517	3.9885	3.8839	6.8397	0.8579
DTW-prob	3.9066	**0.0000**	2.9589	0.9681	5.3280	7.2500	3.3126	0.9644	4.0471	3.3761	5.1034	0.8839

TABLE VII: Median performance metrics for various time series ensemble methods applied on the class of RM- predictors on database D_2 - direct comparison with human performance. The best result is in boldface.

Network Type	RMN				RMR				RMH			
Model/Metric	RMSE	OR	DTW	SROCC	RMSE	OR	DTW	SROCC	RMSE	OR	DTW	SROCC
best	4.9035	2.1376	4.7538	0.9102	6.4582	7.1004	6.3780	0.8666	6.0270	7.7637	9.7567	0.7462
avg	4.3345	**0.0000**	3.8451	0.9456	5.7435	2.1315	4.5460	0.9327	4.6125	1.3333	5.8474	0.8671
med	4.4583	**0.0000**	**3.7086**	0.9433	5.5579	1.0784	3.8582	**0.9461**	4.3935	1.3514	6.2293	0.8580
mod	4.3294	**0.0000**	3.7929	0.9410	5.4830	1.0531	3.9363	0.9439	4.4064	1.3333	6.3676	0.8501
DTW-single	4.5501	**0.0000**	4.0212	0.9422	5.6212	1.3260	3.9958	0.9378	4.5213	1.1303	7.6111	0.8402
DTW-prob	4.4013	**0.0000**	3.7769	0.9459	5.6201	1.1766	3.9615	**0.9461**	4.5650	1.1629	5.7202	0.8716
ref	**3.9060**	**0.0000**	4.6007	0.9338	**3.9060**	**0.0000**	4.6007	0.9338	**3.9060**	**0.0000**	4.6007	0.9338

dynamic networks.

As before, we also report the results compared against human performance in Table X. We drew similar observations as in Table VII: the objective predictions tend to get worse while human performance usually upper bounds model performance. However, 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 [15].

E. Quantitative Experiments - Activation Function, Training Algorithm

We also tested 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 the 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. Similar results were produced by other evaluation metrics.

TABLE XI: Comparison between different activation functions when training the NARX component using OR. Rows correspond to the activation function used in the hidden layer, while columns to the activation function in the output layer on D_1 (VN) and on D_2 (RMN).

Database	D_1 (VN)			D_2 (RMN)		
Activation	tansig	logsig	purelin	tansig	logsig	purelin
tansig	10.3793	20.2759	5.8966	10.5881	31.3844	7.6818
logsig	8.9655	22.5517	5.1034	10.3386	33.2626	7.9185
purelin	9.2759	31.2759	11.1034	26.0428	50.5526	5.481

We also compared the Levenberg-Marquardt algorithm against other training algorithms [48]. 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.

F. Discussion

Combining perceptual video quality models with rebuffering event measurements and memory can deliver promising continuous-time QoE prediction results. We found that a simple rebuffering-related input can capture the effects of rebuffering events and that advanced video quality metrics may improve on the predictive performance. Among the various design aspects of the developed predictors, the choice of the dynamic network contributed the most in the final results. Overall, we found that predictions using recurrent neural networks were unstable; hence ensemble methods turned out to be most efficient for those cases. Both the Hammerstein-Wiener and the NARX approaches are simple, but their performances varied: on D_1 and D_2 the NARX-based predictors were better than HW, while for D_3 the HW component improved upon NARX. Therefore, designing a successful continuous-time QoE predictor relies heavily on choosing a suitable dynamic model component. Notably, using ensemble prediction methods can help alleviate these dependencies by producing reliable and more robust forecasts. However, these improvements may not be significant if the individual forecasts are similar to each other.

TABLE IX: Median performance metrics for various time series ensemble methods applied on the class of G- predictors on database D_3 using ST-RRED. The best result is in boldface.

Network Type	GN				GR				GH			
Model/Metric	RMSE	OR	DTW	SROCC	RMSE	OR	DTW	SROCC	RMSE	OR	DTW	SROCC
best	0.2761	16.3109	26.5323	0.8127	0.3705	22.5546	28.5764	0.7170	0.2182	6.1912	25.4509	0.7685
avg	0.2433	8.3065	19.8192	0.8833	0.2903	14.8650	20.1141	0.8098	0.1535	0.3300	8.0770	0.9002
med	0.2386	6.6636	21.6525	0.8900	0.2873	13.8968	18.4652	0.8160	0.1084	0.0000	7.4343	0.9109
mod	0.2416	3.9183	20.5978	0.8765	0.2832	13.9032	19.2255	0.8143	**0.0958**	**0.0000**	**6.9761**	**0.9138**
DTW-single	0.2489	6.0227	19.7491	0.8915	0.2972	14.7661	21.2789	0.8199	0.1262	0.0000	12.2498	0.8716
DTW-prob	0.2380	6.5349	19.9997	0.8894	0.2871	14.3094	18.8939	0.8192	0.1203	0.0000	7.4475	0.9062

TABLE X: Median performance metrics for various time series ensemble methods applied on the class of G- predictors on database D_3 using ST-RRED - direct comparison with human scores. The best result is in boldface.

Network Type	GN				GR				GH			
Model/Metric	RMSE	OR	DTW	SROCC	RMSE	OR	DTW	SROCC	RMSE	OR	DTW	SROCC
best	0.2920	5.4823	28.3930	0.7822	0.3767	9.6446	29.5208	0.6944	0.2401	2.3664	25.5554	0.7628
avg	0.2581	**0.0000**	23.0780	0.8558	0.3083	3.3029	21.6328	0.7941	0.1872	**0.0000**	10.1901	0.8719
med	0.2451	**0.0000**	22.5154	0.8622	0.2997	2.2128	19.3513	0.8022	0.1492	**0.0000**	**9.1562**	0.8841
mod	0.2450	**0.0000**	22.0302	0.8493	0.2998	2.1736	19.9365	0.7981	**0.1419**	**0.0000**	9.2764	0.8928
DTW-single	0.2585	**0.0000**	19.9898	0.8595	0.3080	3.0972	21.0850	0.7956	0.1616	**0.0000**	13.8216	0.8467
DTW-prob	0.2468	**0.0000**	21.4827	0.8594	0.3012	2.3189	19.1690	0.8056	0.1585	**0.0000**	9.4583	0.8851
ref	0.1960	**0.0000**	10.7132	**0.9029**	0.1960	**0.0000**	10.7132	**0.9029**	0.1960	**0.0000**	10.7132	**0.9029**

TABLE XII: Comparison between different training algorithms using the NARX component on databases D_1 (VN) and D_2 (RMN). The number of maximum iterations was set to 1000. The best result is in boldface.

Database	D_1				D_2			
Metric	RMSE	OR	DTW	SROCC	RMSE	OR	DTW	SROCC
trainlm	3.9960	5.7241	15.3884	**0.9070**	**4.4253**	**7.4647**	**4.1464**	**0.9291**
trainbfg	**3.8958**	**4.8621**	**14.7250**	0.8995	6.3309	17.5301	6.6519	0.8134
trainrp	4.2611	7.7931	17.5826	0.8860	9.2138	29.2530	9.9632	0.7111
trainscg	4.1977	6.0345	16.5334	0.8850	6.5928	21.0884	7.2083	0.7899
traincgb	4.0063	5.3448	15.9283	0.8882	6.1350	18.3629	6.3425	0.8225
traincgf	4.2704	6.8621	16.5400	0.8837	6.1077	18.5646	6.5928	0.8212
traincgp	4.0702	5.8276	15.5916	0.8935	6.4566	21.0863	6.5722	0.8027
trainoss	4.4877	6.9655	18.1414	0.8722	7.1855	24.2692	7.2972	0.8025
traingdx	6.3337	17.7241	22.2605	0.7952	11.8730	38.4865	10.0407	0.6558

IX. Future Work

We deployed dynamic network models having simple structures, that process a small number of inputs rich in QoE-related information and which can be easily deployed in both FR and NR applications. We hope that this work will be useful to video QoE researchers as they address the challenging aspects of continuous-time video QoE monitoring. It would be interesting to investigate possible ways of introducing more inputs to the dynamic architectures, as in [10]. Methods of continuous-time QoE summarization are also of interest, i.e., learning-driven ways of aggregating subjective continuous-time QoE scores into a single subjective QoE score. Since, in this way, continuous time QoE predictions would be used as proxies for subjective QoE, we could then apply similar pooling techniques to QoE prediction algorithms.

In our preliminary experiments, we found that when our proposed QoE prediction engines are 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 the need for building more general and publicly available datasets. 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.

X. Acknowledgement

The authors thank Anush K. Moorthy for fruitful discussions on models of human performance on continuous-time QoE.

References

[1] http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html.

[2] "ISO/IEC FCD 23001-6 part 6: Dynamics adaptive streaming over HTTP (DASH)," MPEG Requirements Group, 2011.

[3] R. Pantos and W. May, "HTTP live streaming," 2016.

[4] Z. Li, X. Zhu, J. Gahm, R. Pan, H. Hu, A. C. Begen, and D. Oran, "Probe and adapt: Rate adaptation for http video streaming at scale," *Selected Areas in Communications*, vol. 32, no. 4, pp. 719–733, 2014.

[5] A. Beben, P. Wiśniewski, J. M. Batalla, and P. Krawiec, "ABMA+: lightweight and efficient algorithm for HTTP adaptive streaming," *International Conference on Multimedia Systems*, 2016.

[6] 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," *Computer Communication Review*, vol. 44, no. 4, pp. 187–198, 2015.

[7] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[8] T. C. Mills, *Time Series Techniques for Economists*. Cambridge University Press, 1991.

[9] C. G. Bampis, Z. Li, and A. C. Bovik, "Continuous Prediction of Streaming Video QoE using Dynamic Networks," *Signal Processing Letters,* accepted.

[10] C. G. Bampis and A. C. Bovik, "An augmented autoregressive approach to HTTP video stream quality prediction," European Signal Processing Conference, submitted.

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

[12] D. Ghadiyaram, J. Pan, and A. C. Bovik, "A time-varying subjective quality model for mobile streaming videos with stalling events," in *SPIE Optical Engineering+ Applications*, 2015, pp. 959 911–959 918.

[13] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. on Image Process.*, vol. 19, no. 6, pp. 1427–1441, June 2010.

[14] 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 Journal on Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 652–671, 2012.

[15] C. G. Bampis, Z. Li, A. K. Moorthy, I. Katsavounidis, A. Aaron, and A. C. Bovik, "Temporal effects on subjective video quality of experience," *Transactions on Image Processing,* submitted.

[16] J. Søgaarda, S. Tavakolib, K. Brunnströmcd, and N. Garcíab, "Subjective analysis and objective characterization of adaptive bitrate videos," in *IS&T International Symposium on Electronic Imaging*, 2016.

[17] N. Staelens, J. De Meulenaere, M. Claeys, G. Van Wallendael, W. Van den Broeck, J. De Cock, R. Van de Walle, P. Demeester, and F. De Turck, "Subjective quality assessment of longer duration video sequences delivered over HTTP adaptive streaming to tablet devices," *IEEE Transactions on Broadcasting*, vol. 60, no. 4, pp. 707–714, 2014.

[18] S. Winkler, "Analysis of public image and video databases for quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 616–625, 2012.

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

[20] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. on Image Process.*, vol. 19, no. 2, pp. 335–350, 2010.

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

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

[23] P. V. Vu, C. T. Vu, and D. M. Chandler, "A spatiotemporal most-apparent-distortion model for video quality assessment," in *IEEE International Conference on Image Processing*, 2011, pp. 2505–2508.

[24] J. Y. Lin, T.-J. Liu, E. C.-H. Wu, and C.-C. J. Kuo, "A fusion-based video quality assessment (FVQA) index," *Signal and Information Processing Association Annual Summit and Conference*, 2014.

[25] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric." http://techblog.netflix.com/2016/06/toward-practical-perceptual-video.html.

[26] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *Trans. on Circ. and Syst. for Video Tech.*, vol. 23, no. 4, pp. 684–694, 2013.

[27] Y. Kawayoke and Y. Horita, "NR objective continuous video quality assessment model based on frame quality measure," in *IEEE International Conference on Image Processing*, 2008, pp. 385–388.

[28] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.

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

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

[31] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *Trans. Img. Process.*, vol. 25, no. 1, pp. 289–300, 2016.

[32] A. C. Bovik, "Automatic prediction of perceptual image and video quality," *Proc. IEEE*, vol. 101, no. 9, pp. 2008–2024, Sept. 2013.

[33] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz, "Quantification of YouTube QoE via crowdsourcing," in *IEEE International Symposium on Multimedia*, 2011, pp. 494–499.

[34] D. Z. Rodriguez, J. Abrahao, D. C. Begazo, R. L. Rosa, and G. Bressan, "Quality metric to assess video streaming service over tcp considering temporal location of pauses," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 3, pp. 985–992, 2012.

[35] Z. Duanmu, Z. Kai, K. Ma, A. Rehman, and Z. Wang, "A quality-of-experience index for streaming video," *Selected Topics in Signal Processing*, 2016.

[36] C. G. Bampis and A. C. Bovik, "Learning to predict streaming video QoE: Distortions, rebuffering and memory," *Transactions on Image Processing,* submitted.

[37] H. Yeganeh, R. Kordasiewicz, M. Gallant, D. Ghadiyaram, and A. Bovik, "Delivery quality score model for internet video," in *IEEE International Conference on Image Processing*, Oct 2014, pp. 2007–2011.

[38] D. S. Hands and S. Avons, "Recency and duration neglect in subjective assessment of television picture quality," *Applied cognitive psychology*, vol. 15, no. 6, pp. 639–657, 2001.

[39] A. J. Greene, C. Prepscius, and W. B. Levy, "Primacy versus recency in a quantitative model: activity is the critical distinction," *Learning & Memory*, vol. 7, no. 1, pp. 48–57, 2000.

[40] K. D. Singh, Y. Hadjadj-Aoul, and G. Rubino, "Quality of experience estimation for adaptive HTTP/TCP video streaming using H. 264/AVC," *Consumer Communications and Networking Conference*, 2012.

[41] "On the robust performance of the ST-RRED video quality predictor," http://live.ece.utexas.edu/research/Quality/ST-RRED/.

[42] T. Lin, B. G. Horne, P. Tino, and C. L. Giles, "Learning long-term dependencies in NARX recurrent neural networks," *IEEE Trans. on Neur. Netw.*, vol. 7, no. 6, pp. 1329–1338, 1996.

[43] H. T. Siegelmann, B. G. Horne, and C. L. Giles, "Computational capabilities of recurrent NARX neural networks," *IEEE Trans. on Sys., Man, and Cyber.*, vol. 27, no. 2, pp. 208–215, 1997.

[44] K. S. Narendra and K. Parthasarathy, "Identification and control of dynamical systems using neural networks," *IEEE Transactions on neural networks*, vol. 1, no. 1, pp. 4–27, 1990.

[45] https://www.mathworks.com/help/nnet/ug/design-time-series-narx-feedback-neural-networks.html.

[46] K. Levenberg, "A method for the solution of certain non-linear problems in least squares," *Quarterly of applied mathematics*, 1944.

[47] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *Industrial and Applied Mathematics*, 1963.

[48] https://www.mathworks.com/help/nnet/ug/choose-a-multilayer-neural-network-training-function.html.

[49] https://www.mathworks.com/help/nnet/ug/improve-neural-network-generalization-and-avoid-overfitting.html.

[50] M. Leutbecher and T. N. Palmer, "Ensemble forecasting," *Journal of Computational Physics*, vol. 227, no. 7, pp. 3515–3539, 2008.

[51] G. P. Zhang, "Time series forecasting using a hybrid arima and neural network model," *Neurocomputing*, vol. 50, pp. 159–175, 2003.

[52] Z.-H. Zhou, J. Wu, and W. Tang, "Ensembling neural networks: many could be better than all," *Artificial intelligence*, 2002.

[53] A. Krogh, J. Vedelsby *et al.*, "Neural network ensembles, cross validation, and active learning," *Advances in neural information processing systems*, vol. 7, pp. 231–238, 1995.

[54] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series." in *KDD workshop*, 1994.

[55] J. H. Stock and M. W. Watson, "Combination forecasts of output growth in a seven-country data set," *Journal of Forecasting*, vol. 23, no. 6, pp. 405–430, 2004.

[56] N. Kourentzes, D. K. Barrow, and S. F. Crone, "Neural network ensemble operators for time series forecasting," *Expert Systems with Applications*, vol. 41, no. 9, pp. 4235–4244, 2014.

[57] J. L. Elman, "Finding structure in time," *Cognitive Science*, vol. 14, no. 2, pp. 179–211, 1990.

[58] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 5, pp. 855–868, 2009.

[59] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling." in *INTERSPEECH*, 2014, pp. 338–342.

[60] 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," *Global Conference on Signal and Information Processing*, 2014.

[61] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," *Asilomar Conference on Signals, Systems and Computers*, 2003.

[62] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. on Image Proc.*, vol. 13, no. 4, pp. 600–612, 2004.

[63] S. Siegel, *Non-parametric statistics for the behavioral sciences.* McGraw-Hill, 1956.