Supporting Real Time VBR Using Dynamic reservation Based on Linear Prediction[*]

A. Adas
Department of Electrical Engineering
Georgia Institute of Technology
Atlanta, GA 30332

GIT–CC–95/26

August 8, 1995

Abstract

A dynamic bandwidth allocation strategy to support variable bit rate (VBR) video traffic is proposed. This strategy predicts the bandwidth requirements for future frames using either adaptive or non-adaptive least mean square (LMS) error linear predictors. The adaptive technique does not require any prior knowledge of the statistics, nor assumes stationarity. Several reservation schemes and pro-active congestion approach are also presented. Analysis using six one-half hour video traces indicate that prediction errors for the bandwidth required for the next frame are almost white noise.

By reserving bandwidth equal to the predicted value, only the prediction errors need to be buffered. Because the errors are almost white noise, small buffers size, high utilization, and small delay are achieved. Simulation results show that for the same expected cell loss, buffers size is reduced by more than a factor of 100 and network utilization is increased more that 250 % as compared to a fixed service rate.

College of Computing
Georgia Institute of Technology
Atlanta, Georgia 30332–0280

[*]This research was supported in part by the National Science Foundation under grant NCR-9396299.

Contents

1 Introduction

Variable bit rate (VBR) video traffic is expected to be one of the major applications that need to be supported by broadband packet-switched networks. Several studies on VBR video traffic indicate the existence of a slowly decaying auto-correlation structure [4, 5, 6, 16].
Providing efficient transport and Quality of Service (QoS) guarantees for VBR video is non-trivial in packet-switched networks. For instance, correlated traffic dramatically increases the queue length statistics at a multiplexor [1, 11, 12, 13]. Supporting VBR video traffic at a deterministic fixed service, not close to the peak, usually results in large buffers, large delay, and large delay jitter.

Because bandwidth in ATM networks can be allocated on demand, dynamic bandwidth allocation and re-negotiation during the connection lifetime has been considered in [9, 15].
In this paper we use the correlation structure of VBR video traffic to predict the bandwidth required for future frames. Linear prediction minimizing the mean square error is studied on six half hour video traces. Adaptive and non-adaptive techniques are considered for predicting the bandwidth required by future frames. Adaptive techniques do not require knowledge of the video statistics, and do not assume stationarity, so, they can be used for on-line real-time applications.

The order of the linear predictor is small (12 or less) and experiments show that it does not increase with the size of the VBR video traffic trace. The prediction errors resemble white noise, or at most short memory, but the marginals have a heavy tail.

Using fixed service rate, for highly correlated input traffic with a heavy tail, if not served at a rate close to the peak, causes large queues, large delays, and excessive cell loss [1, 13, 17].
In contrast, by reserving bandwidth at least equal to the predicted value, only the errors in the prediction needs to be buffered. Because the errors resemble white noise or at most short memory, small buffers, high utilization and small delays can be achieved. Results show that the buffer size is reduced by more than a factor of 100 as compared to a traditional deterministic fixed service rate reservation. For a given expected loss, the utilization increased from 0.3 for deterministic rate to 0.9 for a dynamic rate strategy, a 300 % increase in utilization.

These results suggest that fixed bandwidth allocation for VBR video will not achieve acceptable utilization. On the other hand, using prediction to reserve bandwidth dynamically, the problem is changed from supporting slowly decaying auto-correlated input traffic stream, to servicing and buffering the residuals (errors) of the prediction. The same prediction approach can be used to predict N frames ahead. The value of N depends on the video correlation structure, the errors, and the delay in negotiating the rate with the network.

The objectives of this paper are:

(a) To investigate the feasibility and performance of linear prediction algorithm to forecast VBR video traffic using both adaptive and non-adaptive techniques.

(b) To compare queuing and delay performance of fixed and dynamic bandwidth allocation.

Potential applications include: (1) Predicting and changing leaky bucket parameters. For MPEG streams, the predictions are found to better for individuals frames (I,P, and B), so multiple leaky buckets suggested in [17] in conjunction with forecasting may be used. (2) Dynamic reservation in a multi-access channels, e.g., in Metropolitan area networks.

The rest of the paper is organized as follows. Section 2 describes the non-adaptive mean square error linear predictors. Section 3 describes the least mean square (LMS) and the normalized least mean square (NLMS) adaptive linear predictors. Section 4 investigates the feasibility and the performance of forecasting different empirical video traces. It also discusses the effect of the size of the trace on the performance of the forecast. Section 5 addresses the performance of dynamic bandwidth allocation schemes based on linear prediction, both adaptive and non-adaptive, and compares them to the traditional deterministic fixed service rate. Section 6 discusses the preliminary idea of a pro-active congestion control scheme and Section 7 presents the conclusions.

2 Minimum Mean Square Error Linear Predictor

The $k - step$ linear predictor is concerned with the estimation (prediction) of $x(n+k)$ using a linear combination of the current and previous values of $x(n)$[7]. Thus, the p^{th} order linear predictor has the form:

$$\hat{x}(n+k) = \sum_{l=0}^{p-1} w(l)x(n-l) \tag{1}$$

where $w(l)$, for $l = 0, 1 \cdots, p-1$ are the linear prediction filter coefficients. This can be represented as shown in Figure 1.

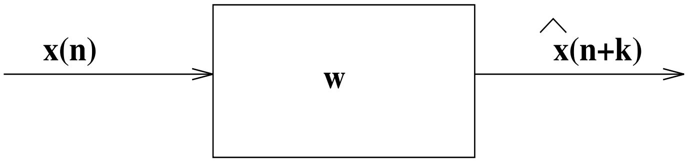


Figure 1: Linear predictor

Let :

$$\begin{aligned} \mathbf{w} &= [w(0), w(1), \cdots, w(p-1)]^T, \\ \mathbf{x}(n) &= [x(n), x(n+1), \cdots, x(n-p+1)]^T, \\ e(n) &= x(n+k) - \hat{x}(n+k). \end{aligned} \tag{2}$$

From (1) and (2)

$$e(n) = x(n+k) - \mathbf{w}^T\mathbf{x}(n). \tag{3}$$

The optimal linear predictor in the mean square sense is one which minimizes the mean square error ξ, where:

$$\xi = E\{e^2(n)\}.$$

Since ξ is a quadratic function, it has a unique minimum. Therefore, the vector $\mathbf{w}$ that minimizes ξ is found by taking the gradient, setting it equal to zero and then solve for $\mathbf{w}$:

$$\begin{aligned} \nabla\xi &= \nabla E\{e^2(n)\} \\ &= -2E\{e(n)\mathbf{x}(n)\} = 0. \end{aligned}$$

Substituting the value for $e(n)$ from (3):

$$\nabla\xi = -2E\left\{\left(x(n+k) - \mathbf{w}^T\mathbf{x}(n)\right)\mathbf{x}(n)\right\} = 0.$$

Taking expectation and writing it in matrix form, we have:

$$\mathbf{R_x}\mathbf{w}^T = \mathbf{P}(k) \tag{4}$$

where,

$$\mathbf{R_x} = \begin{bmatrix} r_x(0) & r_x(1) & \ldots & r_x(p-1) \\ r_x(1) & r_x(0) & \ldots & r_x(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ r_x(p-1) & r_x(p-2) & \ldots & r_x(0) \end{bmatrix}, \mathbf{P}(k) = \begin{bmatrix} r_x(k) \\ r_x(k+1) \\ \vdots \\ r_x(k+p-1) \end{bmatrix},$$

and $r_x(k)$ is defined as $E\{x(n+k)x(n)\}$.

The equations in (4) are the Wiener-Hopf equations for linear prediction. For a 1-step linear predictor ($k = 1$), the set of linear equations defined in (4) are equivalent to the set of linear equations used to fit a p^{th} order autoregressive (AR) process with the exception of a minus sign [7]. The solution of the linear equations in (4) requires the knowledge of the auto-correlation of $\mathbf{x}(n)$ and it also assumes wide sense stationarity, i.e., the mean, variance, auto-covariance of $\mathbf{x}(n)$ do not change with time. It also requires inverting $\mathbf{R_x}$ whose size depends on the order of the linear predictor, p.

3 Adaptive Least Mean Square Error Linear Predictor

We consider in this section the method of least mean square error linear predictor (LMS) [8]. The LMS is an adaptive approach. It does not require any prior knowledge of the auto-correlation of the sequence. Therefore, it can be used as an on-line algorithm for forecasting bandwidth. The operation of an adaptive linear predictor is shown in Figure 2. The prediction coefficients $\mathbf{w}(n)$ are time varying. The errors, $\{\mathbf{e}(n)\}$, are fed back and used to adapt the filter coefficients in order to decrease the mean square error. $e(n)$, $\mathbf{x}(n)$, and $\mathbf{w}$ are defined as in (2).

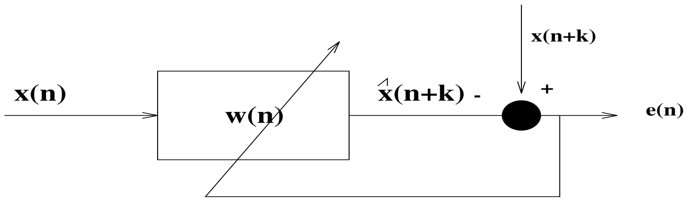


Figure 2: Adaptive Linear Predictor

Following is a summary of the algorithm:

- Start with an initial estimate of the filter (prediction) coefficients $\mathbf{w}(0)$.
- For each new data point compute $\nabla\xi$, where

$$\nabla\xi = -2E\{e(n)\mathbf{x}(n)\}.$$

In practice, the statistics are not known and may change with time, Therefore, the expectation is replaced with an estimate. The simplest estimate is the one point sample average e(n)$\mathbf{x}$(n).

- Update $\mathbf{w}(n)$ by taking a step of size 0.5μ in the negative gradient direction (this will point to the bottom of the error surface) see [8]. The update equation for LMS filter coefficients:

$$\begin{aligned}\mathbf{w}(n+1) &= \mathbf{w}(n) - 0.5\mu\nabla\xi && (5)\\ &= \mathbf{w}(n) + \mu e(n)\mathbf{x}(n). && (6)\end{aligned}$$

If $\mathbf{x}(n)$ is stationary, $\mathbf{w}(n)$ converges in the mean to the optimal solution $\mathbf{R_x w} = \mathbf{P}$ [7, 8]. The only thing left is what value of μ one should use. As has been shown in [8], LMS will converge in the mean if $0 < 1/\mu < 2/\lambda_{max}$, where λ_{max} is the maximum eigenvalue of $\mathbf{R_x}$.
A normalized LMS **(NLMS)** is a modification to the LMS algorithm where the update equation is:

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \frac{\mu e(n)\mathbf{x}(n)}{||\mathbf{x}(n)||^2},$$

where $||\mathbf{x}(n)||^2 = \mathbf{x}(n)^T\mathbf{x}(n)$. The advantage of using NLMS over LMS is that it is less sensitive to the step size μ. If $0 < \mu < 2$, then NLMS will converge in the mean [8]. Using large μ results in a faster convergence and quicker response to signal changes. However, after convergence, the prediction coefficients will have large fluctuations. On the other hand using small μ results in slower convergence and less fluctuation after convergence, i.e., there is a trade off. Since at time n the value of $x(n+k)$ is not available to compute $e(n)$, $e(n-k)$ is used instead. For example, the 1-step linear predictor update equation becomes:

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \frac{\mu e(n-1)\mathbf{x}(n-1)}{||\mathbf{x}(n-1)||^2}$$

Replacing $e(n)$ by $e(n-1)$, the gradient estimate is $e(n-1)\mathbf{x}(n-1)$ instead of $e(n-1)\mathbf{x}(n)$. Our experiments showed that using $e(n-1)\mathbf{x}(n)$ instead, does not always converge to the optimal solution.

4 Forecasting Empirical MPEG-I VBR Traces

We use six long empirical video traces (about half an hour each). obtained from the public ftp site *ftp-info3.informatik.uni-wuerzburg.de*. The sequences represent a wide range of video applications. These video sequences are encoded according to the MPEG-I standard. MPEG-I has three frame types: I, P, and B. I frames use intra-frame coding, P frames use intra-frame coding and motion compensation based on previous frames, and B frames use intra-frame and motion compensation based on previous and next frames. The encoder uses a periodic frame pattern which is called a Group of Pictures (GOP). The GOP pattern used for the video sequences was IBBPBBPBBPBB. For more details about the encoder parameters see [18]. Table 1 shows the video traces, their mean, and maximum frame sizes. The traces were used to test the forecast methods.

Since I, P, and B frame types have different statistical characteristics, we separate them and forecast each frame type separately. Each video sequence is split virtually into I, P, and B subsequences. Table 2 shows the $\sum e^2(n)/\sum x^2(n)$ and the order (p) of a 1-step linear predictor using the set of Wiener-Hopf equations defined in (4). The smaller the $\sum e^2(n)/\sum x^2(n)$, the better the forecast. The 1-step linear predictor is used to predict the required bandwidth of the next frame.

Sequence	Mean (bits)	Max (bits)
Soccer	25,110	190,836
Jurassic Park	13,078	119,010
Gold Finger	24,308	244,592
Talk Show	14,573	106,383
Star Wars	15,599	185,628
Terminator	10,904	79,600

Table 1: Encoded sequences used

The Akaike information criterion (AIC) [7, 19] was used to choose the best order not greater than 12. The AIC criterion associates a cost function with the order of the filter. The maximum order, $p = 12$, was chosen after experimenting with different orders and checking if the auto-correlation function of $e(n)$ was close to that of a white noise. The order of NLMS was kept the same as that of the Wiener-Hopf equations (4) for a fair comparison of the two methods.

It appears from Table 2 that the I-frames can be predicted more accurately than P or B frames. For example, the $\sum e^2(n)/\sum x^2(n)$ for the I frames subsequence of the Talk Show trace was 0.004, which is a signal to noise ratio (SNR) of 250, compared to 0.115 and 0.013 for the P and B subsequences, respectively.

Figures 3(a)-(c) show the forecasted and actual subsequences for the I, P, and B frames, of the Terminator trace respectively. The Terminator video sequence has the largest $\sum e^2(n)/\sum x^2(n)$. Nonetheless, closer inspection reveals that the forecasted values appear close to the actual values except at sharp transitions which are most likely scene changes. Figures 4(a)-(f) show the auto-correlations for the I, P, and B subsequences and the residuals of the forecast for the same trace. Although, the input subsequences are highly correlated, the residuals after forecasting resemble white noise. This was true for all of the six video sequences tested.

The primary difficulty in supporting VBR video traffic at a fixed rate results from the fact that VBR traffic is correlated and has heavy tail. Highly correlated input process with a heavy tail, if served at a fixed rate not close to the peak, causes large queues, large delays, and excessive cell loss [1, 13, 17]. Hence, by reserving bandwidth at least equal to the prediction, only the errors of the prediction need to be buffered. Since the errors resemble white noise or at most short memory, smaller buffers, less delays, and higher utilization are expected when compared to traditional fixed rate reservations.

Adaptive NLMS Algorithm

The results of the NLMS prediction algorithm are also presented in Table 2. Figure 5 shows the forecasted and the actual I subsequence and the auto-correlation of the residuals for the Terminator trace. The results appear to be close to those obtained by using the Wiener-Hopf equations (4) in which prior knowledge of the auto-correlation is needed. However, some correlations do remain in the residuals (Figure 5) and this most likely results in slightly larger $\sum e^2(n)/\sum x^2(n)$ (Table 2). Therefore, the NLMS adaptive algorithm can be used on-line to predict the rate of VBR traffic without the need to know the auto-correlation of the video stream in advance.

Long-range dependent processes can be represented by an autoregressive (AR) process with infinite order [14]. Therefore, to approximate a long range dependent process with an AR

Sequence	Subsequence	$\sum e^2(n)/\sum x^2(n)$ Wiener-Hopf	$\sum e^2(n)/\sum x^2(n)$ NLMS	Order
Jurassic Park	I	0.011	0.013	7
	P	0.110	0.120	12
	B	0.046	0.052	12
Star Wars	I	0.013	0.016	8
	P	0.220	0.260	12
	B	0.101	0.087	12
Gold Finger	I	0.015	0.020	11
	P	0.047	0.053	12
	B	0.018	0.022	12
Terminator	I	0.023	0.028	12
	P	0.124	0.140	12
	B	0.097	0.109	12
Talk Show	I	0.004	0.006	11
	P	0.115	0.138	12
	B	0.013	0.016	12
Soccer	I	0.027	0.033	2
	P	0.036	0.042	3
	B	0.042	0.053	12

Table 2: The $\sum e^2(n)/\sum x^2(n)$ for the video sequences and the order of the linear predictor used

process, as the size of the realization of the process increases, the order of the AR process should be increased [2]. The effect of increasing the size of the VBR trace on the mean square error is studied by: (1) the first half of the trace is forecasted and the order of the filter is obtained; (2) the same filter order is used on th entire trace. Table 3 shows the results for the Terminator sequence using the two methods of forecast, Wiener-Hopf (Equation(4)) and NLMS. We observe that the difference due to trace size is negligible. Moreover, the auto-correlation in both cases was found to be close to white noise. Therefore, it appears that the size of the trace does not affect the performance of that of a linear predictor, at least for these traces.

Method	Subsequence	$\sum e^2(n)/\sum x^2(n)$ Half	$\sum e^2(n)/\sum x^2(n)$ All
Wiener-Hopf	I	0.0243	0.0245
	P	0.1240	0.1250
	B	0.0960	0.0970
NLMS	I	0.030	0.028
	P	0.140	0.143
	B	0.113	0.109

Table 3: The $\sum e^2(n)/\sum x^2(n)$ for the Terminator video trace using half and all the trace

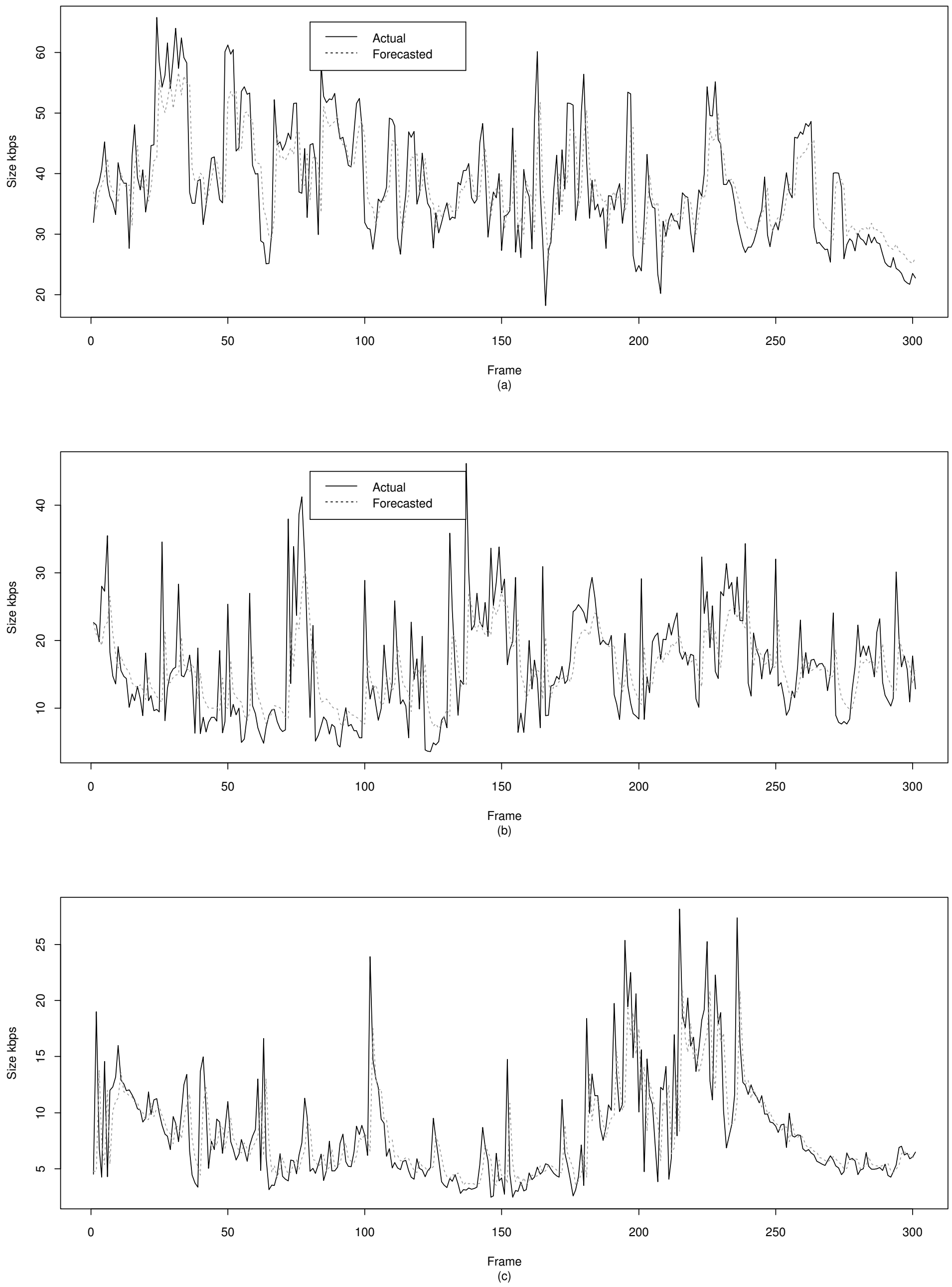


Figure 3: Actual and forecasted subsequence for the Terminator trace. (a) I subsequence; (b) P subsequence; (c) B subsequence

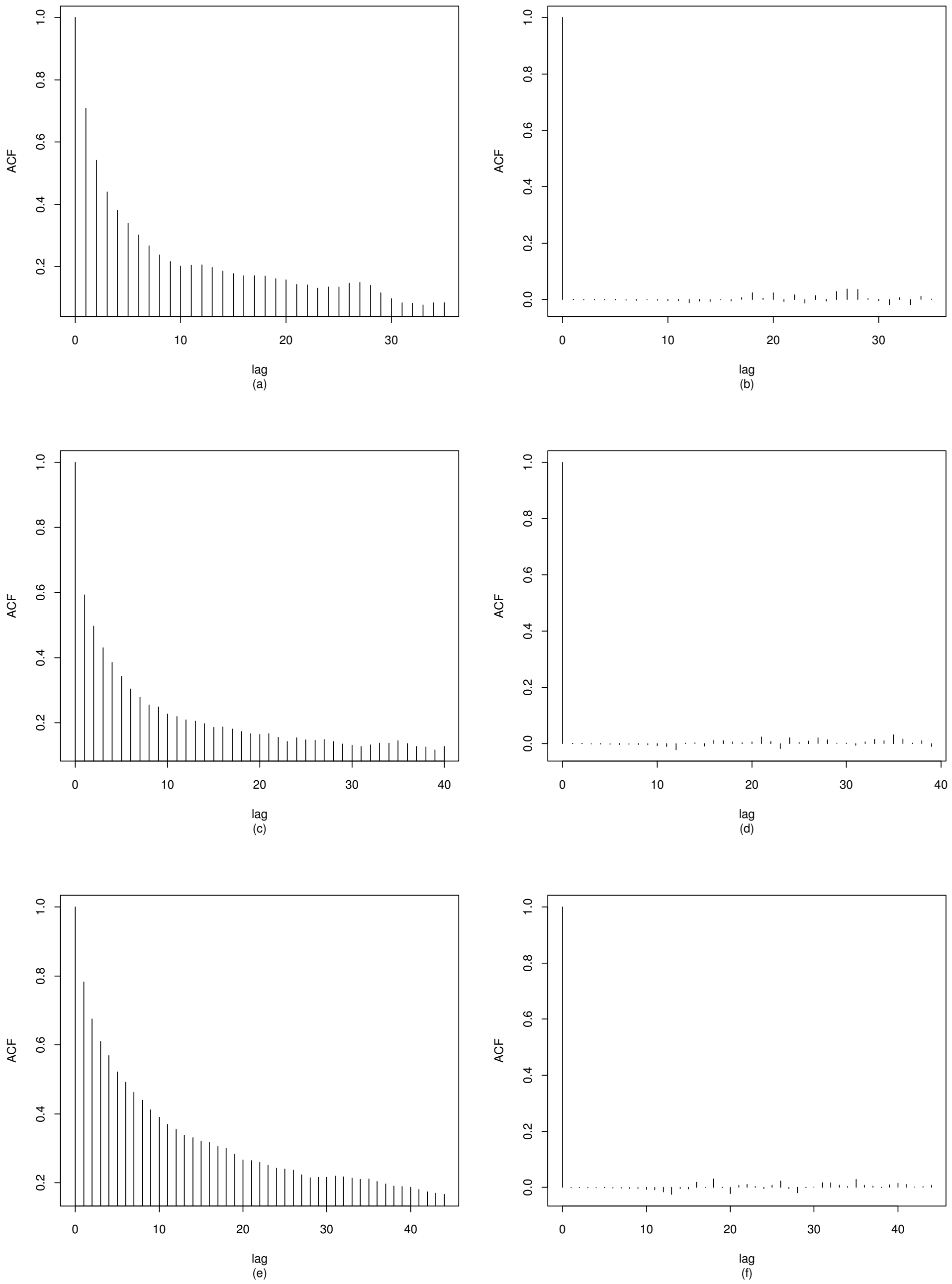
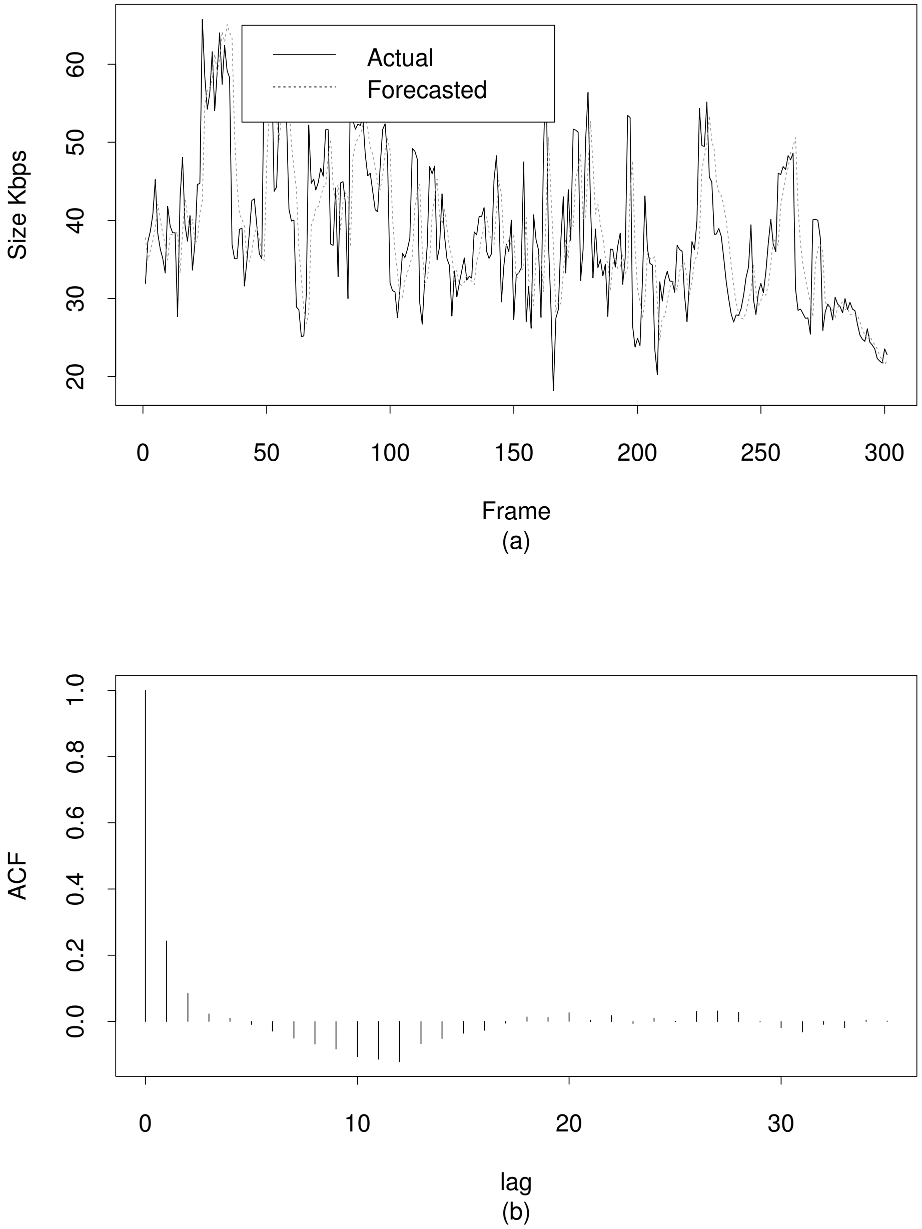


Figure 4: Auto-correlation of the actual subsequence and the residuals of the Terminator trace. (a)& (b) I subsequence; (c) & (d) P subsequence; (e) & (f) B subsequence



(a)

(b)

Figure 5: NLMS method. (a) Actual and forecasted I subsequence for the Terminator; (b) The auto-correlation of the residuals for the same subsequence.

5 Dynamic Bandwidth Allocation Using the Forecast

The primary difficulty in supporting VBR video traffic at a fixed rate results from the fact that VBR traffic is correlated and has heavy tail. Highly correlated input process with a heavy tail, if served at a fixed rate not close to the peak, causes large queues, large delays, and excessive cell loss [1, 13, 17]. Hence, by reserving bandwidth at least equal to the prediction, only the errors of the prediction need to be buffered. If the predicted value is less than the actual value, the difference is buffered. If the predicted value is more, extra bits can be transmitted from the buffer. Since the errors resemble noise or at most short memory, then smaller buffers, less delays, and higher utilization are expected when compared to traditional fixed rate reservations.

If all the predicted bandwidth is not available from the network, the video encoders need to decrease their bit rate to not over load the network as in [10].

Simulation Results

A single server with a FIFO queue was simulated. The traces of the actual video sequences were used as the input process. Two scenarios were studied through the simulation: (1) fixed service rate and (2) dynamic service rate based on prediction. To be able to compare the two scenarios, the bandwidth is assumed to be available from the network. Hence, any bandwidth needed will be granted. This is done only to compare the two scenarios for the same video transmission quality. The buffer size used is a multiple of the maximum frame size in the video trace and the $E[loss]$ is defined as:

$$E[loss] \quad = \quad \frac{L}{N} \times 100,$$

where L is the number of dropped cells and N is the total number of cell arrived. Table 4 shows the $E[loss]$ for a fixed service rate for different utilization levels (0.9 and 0.8) and different buffer sizes. The relation between buffer size and the $E[loss]$ is almost linear. and increasing the buffer size was not efficient in reducing the $E[loss]$.
Table 5 shows E[loss] for dynamic bandwidth allocation. To achieve the utilization (U), the value used for reservation is the predicted value $\hat{x}(n+1)$ divided by the required U. Another way to achieve a specific utilization is by adding a constant C to each predicted value, i.e., reserve $\hat{x}(n+1) + C$. This approach did not achieve the same performance as the former approach, therefore, it was not used.

Figure 6 (a) shows the queue length process for the fixed capacity approach for the Jurassic Park trace. The queue length is large and stays large for a long period of time. This is due to both high correlation and heavy tail distribution of the input stream. For example the queue builds up and reaches a value greater than 15000 kbits and it stays large for a considerable period of time. Therefore, if the buffer was less than 15000 kbits, excessive losses will occur. Even if one can afford this buffer size, the problem of delay remains. For example, cells that arrive after their play back point are not useful. These cells consume network resources and may cause other cells not to meet their delay bounds. Figure 6(b) shows the queue length for the same video trace (Jurassic Park) but this time with dynamic allocation. As Figure 6(b) shows, the maximum queue length is decreased considerably from larger than 15000 kbits to approximately 140 kbits. This is a reduction by a factor of more than a 100. Moreover, the queue length process does not build up, and does not stay high for a long time because the residuals are uncorrelated. Therefore, if a buffer less than 140 kbits is used, the losses will not be large because the area under the queue length process curve is very small.

Let us consider the maximum utilization that can be achieved for a specific $E[loss]$. Using simulations, we found as an example, for the Soccer video trace and for $E[loss] \leq 0.025$, the maximum utilization using the fixed approach was 0.3, while for the dynamic approach it was 0.9. This significant increase in utilization was observed for all the six video traces tested.
It appears from the results that fixed bandwidth allocation for VBR video will not achieve acceptable utilization. On the other hand, using prediction to reserve bandwidth dynamically, the problem is changed from supporting highly correlated input traffic stream, to servicing and buffering the residuals (errors) of the prediction which resemble white noise or short memory. This approach achieved a significant increase in the network utilization and decreased the buffer needed.
The same approach can be used for predicting N frames ahead. The value of N depends on the structure of the video, the errors, and the delay in negotiating the rate with network. The above prediction strategy can be used either at the network side or at the user side. An example of how to implement and use the prediction is given in Section 6.

Enhanced DBW Allocation using Forecast Errors

An enhancement to the dynamic bandwidth allocation based on prediction may be achieved by using the knowledge of previous errors to enhance the reservation, for example:

- Use LMS or NLMS to predict $\hat{x}(n+1)$.
- If $e(n) > 0$, then reserve $\hat{x}(n+1) + e(n)$.

If $e(n)$ is positive, the reserved value is less than the actual frame size and $e(n)$ bits are known to be not transmitted. These bits are most likely buffered. Hence, by reserving the error of the previous prediction and the next predicted frame size, the $e(n)$ bits buffered are guaranteed transmission over the next interval. Thus, the maximum delay will be one frame and the maximum buffer size needed will be at most the maximum positive prediction error.

Discussion

One may consider updating only the filter coefficients whenever the predicted value is less than the actual value. This approach is a more conservative one aiming at reducing the marginal tail of the errors. Simulation results showed that the $E[loss]$ was higher than the NLMS approach. This is due to the fact that the errors generated were correlated. Therefore, it is important for any prediction scheme used to have white noise residuals. Also, it's worth noting that the filter coefficients used for prediction emphasize the last value. This is especially true for the I frames. Thus, one may model the I frame subsequence as a non stationary autoregressive integrated moving average (ARIMA) with a difference d equal to one [3].

6 Pro-Active Congestion Control

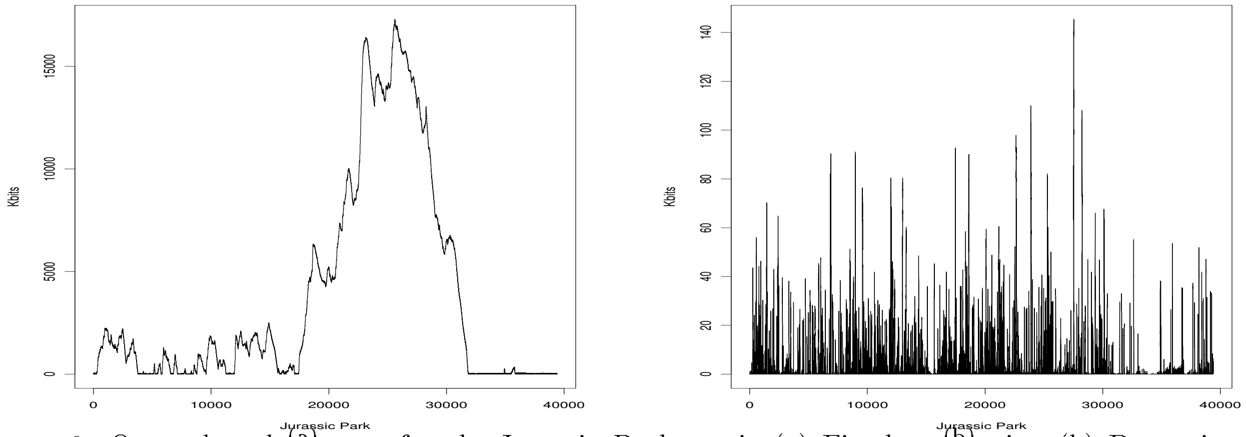
This section discusses how a prediction algorithm can be used to achieve high utilization and prevent congestion in the network. One possible approach is to use a framing strategy. Time is split into frames, each frame corresponds to a time interval T. Each video source predicts the bandwidth required for the next k frames. The sources divide the predicted value by a factor of α, where $0 < \alpha \leq 1$. The larger is α, the higher utilization that can be achieved, but more likely larger delays will occur. Therefore, each source can determine for it self the value

U	Buffer size	Jurassic Park	Star Wars	Gold Fingers	Terminator	Talk	Soccer
0.9	1	9.38	14.82	9.04	6.81	7.34	9.62
	2	8.20	13.36	8.13	5.31	6.54	8.23
	3	7.43	12.31	7.57	4.44	5.94	7.16
	4	6.81	11.49	7.12	3.83	5.51	6.23
	5	6.34	10.86	6.73	3.80	5.13	5.56
0.8	1	5.93	11.2	5.32	3.85	4.34	5.99
	2	4.93	9.91	4.54	2.58	3.79	4.75
	3	4.27	9.00	4.10	1.96	3.34	3.84
	4	3.78	8.32	3.73	1.51	3.05	3.17
	5	3.41	7.84	3.40	1.18	2.81	2.65

Table 4: E[loss] for different utilization when bandwidth reserved is fixed

U	Buffer size	Jurassic Park	Star Wars	Gold Fingers	Terminator	Talk	Soccer
0.9	1	0.005	0.0360	0.0068	0.0450	0.000	0.0222
	2	0.000	0.0000	0.0000	0.0000	0.000	0.0000
	3	0.000	0.0000	0.0000	0.0000	0.000	0.0000
0.8	1	0.000	0.0002	0.0003	0.0037	0.000	0.0003
	2	0.000	0.0000	0.0000	0.0000	0.000	0.0000
	3	0.000	0.0000	0.0000	0.0000	0.000	0.0000

Table 5: E[loss] for different utilization when bandwidth reservation is dynamic



(a) (b)

Figure 6: Queue length process for the Jurassic Park movie (a) Fixed reservation (b) Dynamic reservation

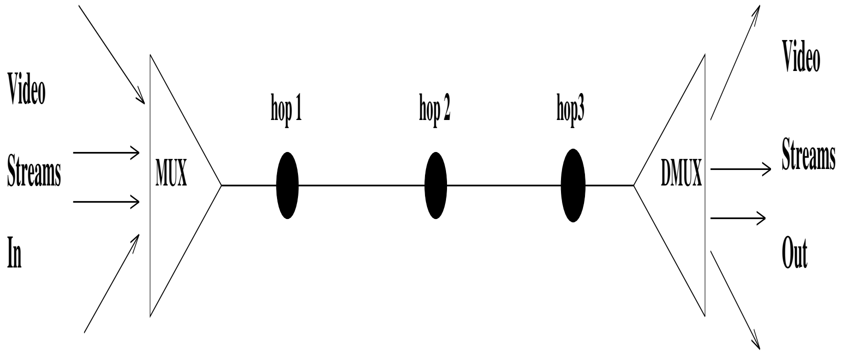


Figure 7: Implementation

of α, based on the accuracy of the forecast and the tradeoff between the service price and QoS. The value of α can also be adaptive, i.e., could change with time, as the accuracy or the QoS requirements of the forecast changes.

Based on the predicted values and α, each source will negotiate with the network the bandwidth required for future frames and send a reservation request. If the reservation is granted, the source encoder will not reduce the output rate. If the reservation is not granted, the video source must adapt and reduce the output bit rate.

For the network in Figure 7, each intermediate network node (hop) between source and destination will add all the requested bandwidth from all the video sources share the output path from that hop. If the total bandwidth requested is less than the link capacity, all requests will be granted. Otherwise, depending on how much capacity is available, the video sources (all or subset) will be asked to reduce their output bit rate. Other non delay sensitive traffic can be transmitted on the availability of the bandwidth, or a fixed bandwidth can be reserved for it.

7 Conclusions

The feasibility and performance of adaptive and non-adaptive linear predictors based on minimizing the mean square error are studied using long empirical video traces. The forecasting results for the 1-step linear predictor show that:

- Forecasted values appear to be close to the actual values. For example, the $\sum e^2(n)/\sum x^2(n)$ of the Talk Show I subsequence was less than 0.004, a SNR more than 250. However, sharp transitions in the video causes the error distribution to have a heavy tail.
- The prediction errors in all cases resemble white noise, or at most short memory.
- Although VBR video traffic is structured and correlated, the order of the filter does not increase with increasing the size of the sequence.

Dynamic bandwidth allocation schemes for VBR video based on adaptive and non-adaptive linear prediction were presented. The adaptive technique does not require any prior knowledge of the video and does not assume stationarity, so it can be used for on-line, real-time applications.

The results suggest that fixed bandwidth allocation for VBR video will not achieve acceptable utilization. On the other hand, using prediction and reservation of bandwidth dynamically changes the problem from supporting slowly decaying auto-correlated input traffic stream to servicing and buffering the residuals (errors) of the prediction. Errors, resembling white noise, allow for small buffer size, higher utilization, and small delay. Simulation results show that buffer size is reduced by more than a factor of 100 as compared to a traditional deterministic fixed service rate reservation. For a given expected loss, the utilization increased from 0.3 for deterministic rate to 0.9 for a dynamic rate strategy, a 300 % increase in utilization.

The same prediction approach should extend to forecasting N frames ahead. The value of N depends on the video correlation structure, the errors residual, and the delay in negotiating the rate with the network.

Acknowledgment

Special thanks to Amarnath Mukherjee for his ideas, supervision, and fruitful discussions through out this research.

References

[1] Adas, A. and A. Mukherjee, "On Resource Management and QoS Guarantees For Long Range Dependent Traffic," *Proc. IEEE INFOCOM,* pp 779-787, Boston, April 1995.

[2] Beran, J., "Statistics for long-memory processes," *Chapman & Hall,* New York, 1994.

[3] Box, G.E.P, G.M. Jenkins, "Time series analysis — forecasting and control," Prentice Hall, 1976.

[4] Beran, J., R. Sherman, M.S. Taqqu and W. Willinger, "Variable-bit-rate video traffic and long-range dependence," accepted for publication in *IEEE Trans. Networking,* 1993.

[5] Garrett, M., "Contributions toward real-time services on packet-switched networks," Ph.D. Thesis, Columbia University, 1993.

[6] Garrett, M., and W. Willinger, "Analysis, modeling and generation of self-similar VBR video traffic," *Proc. ACM Sigcomm,* pp 269-280, London, 1994.

[7] Hayes, M. H., "Statistical Signal Processing and Modeling," John Wiley I& Sons, 1996.

[8] Haykin, S., "Adaptive filter theory," Prentice Hall, 1991.

[9] Zhang, H. and E. Knightly "RED-VBR: A new approach to support delay-sensitive VBR video in packet-switched networks," it Proc. 5th International Workshop on Network and Operating System Support For Digital Audio and Video, Durrham, New Hampshire, 1995.

[10] Hemant, K., P. Mishra and A. Reibman, " An adaptive congestion control scheme for real-time packet video transport," *Proc. ACM sigcom,* pp 20-31, San Francisco, September 1993.

[11] S. Q. Li and C. L. Hwang., "Queue Response to Input Correlation Functions: Discrete Spectral Analysis," *IEEE/ACM Trans. on Networking,* Vol. 1, No. 5, Oct. 1993, pp. 522-533.

[12] S. Q. Li and C. L. Hwang., " Queue Response to Input Correlation Functions: Continuous Spectral Analysis,' ' *IEEE/ACM Trans. on Networking,* Vol. 2, No. 6, Dec. 1994, pp. 678-692.

[13] Livny, M., B. Melamed and A.K. Tsiolis, "The impact of auto-correlation on queueing systems," *Management Science,* pp 322-339, March 1993.

[14] Miller, J. "Forecasting Long Range Dependence," Ph.D. Thesis, Southern Methodist University, 1994.

[15] Pancha P., and M. El Zarki, "Bandwidth requirements of variable bit rate MPEG sources in ATM networks" *Proc. IEEE INFOCOM 1993,* pp 902-909, San Francisco, March 1993.

[16] Pancha P., and M. El Zarki, "Variable bit rate video transmission" *IEEE Commun. Mag.*, Vol.32, No.5, pp. 54-66, May 1994.

[17] Pancha P., and M. El Zarki, " Leaky bucket access control for VBR MPEG video," *Proc. IEEE INFOCOM 1995,* pp 796-803, Boston, April 1995.

[18] Rose, O. "Statistical properties of MPEG video traffic and their impact on traffic modeling in ATM systems," University of Wuerzburg. Institute of Computer Science Research Report Series. Report No. 101. February 1995.

[19] The S+ package, Version 3.0, Statistical Sciences, Inc., September 1991.