Joint Source/Channel Coding of Statistically Multiplexed Real-Time Services on Packet Networks

Mark W. Garrett, *Member, IEEE*, and Martin Vetterli, *Senior Member, IEEE*

Abstract—**We investigate the interaction of congestion control with the partitioning of source information into components of varying importance for variable bit-rate packet voice and packet video. High-priority transport for the more important signal components results in substantially increased objective service quality. Using a Markov chain voice source model with simple PCM speech encoding and a priority queue, simulation results show a signal-to-noise ratio improvement of 45 dB with two priorities over an unprioritized system. Performance is sensitive to the fraction of traffic placed in each priority, and the optimal partition depends on network loss conditions. When this partition is optimized dynamically, quality degrades gracefully over a wide range of load values. Results with DCT encoded speech and video samples show similar behavior. Variations are investigated such as further partition of low-priority information into multiple priorities. A simulation with delay added to represent other network nodes shows general insensitivity to delay of network feedback information. A comparison is made between dropping packets on buffer overflow and timeout based on service requirements.**

I. Introduction

REAL-time services such as voice and video have traditionally been coded for constant rate transport on circuit switched networks. This satisfies the requirement for low delay and low delay variance, but compromises quality and efficiency [1] since voice and video are fundamentally variable rate sources. Packet-switched networks work well for bursty data services, providing efficient operation through statistical multiplexing. This introduces variable delay and possible congestion loss, to which data services tend to be tolerant. The emerging solution for multiservice networks (i.e., B-ISDN, a.k.a. ATM) provides packet transport at a low level, allowing for statistical multiplexing to accommodate data services. For real-time services, circuit emulation is provided over the packet transport with peak allocation and fixed delay. Several scheduling and traffic control techniques have been proposed which permit this type of solution [2]-[5].

In this paper, we consider the use of statistical multiplexing for real-time services. Each source is allocated bandwidth less than its peak requirement, which saves cost by increasing network utilization, but results in nonzero probability of loss. While for data any lost packet must be recovered by retransmission, voice and video cannot wait for replacement of errored packets. They can tolerate some loss however, especially if losses can be restricted to less important signal components. This works because voice and video services have the property that *all bits are not equal*, unlike data where bits are indistinguishable. It, thus, becomes useful to devise coding schemes that separate signal components by significance for prioritized transport. Such *channel coding* may be done in tandem with *source coding* which provides information compression. Statistical multiplexing works for data because of delay tolerance, and for real-time services because of loss tolerance.

For packet video where a large bandwidth allocation (per user) is required, statistical channel sharing has been advocated [6] and may be necessary to make broadband video services economically viable. The provision of services on wide area networks is complicated by the significant cost of facilities. Unlike a local area data network, the design must include efficient use of bandwidth. Fiber optic transmission systems provide cheap bandwidth for present services and traffic requirements, but there is strong economic pressure to add new services, such as video, which will again make bandwidth a dear commodity. It then becomes attractive to expend some extra cost for more complex switching and control in return for a larger savings due to bandwidth efficiency.

The partition of information into two priorities introduces a new parameter, α, which is the fraction of traffic given high priority [7]. We use simple simulation models of an aggregate source and priority queue to show that a) the use of priority has a dramatic effect on the performance of real-time services, and b) that performance is quite sensitive to the value of the priority partition parameter, α. We find that for a given statistically multiplexed load, where there is a positive loss rate, an optimal choice of α exists and its value depends strongly on the loss conditions. We also investigate a system where α is dynamically controlled by network feedback. This is shown to yield superior performance.

A. Joint Source/Channel Coding

The idea that a source uses multiple network priorities for a single service may be understood from the point of view of signal processing. The separation principle of information theory [8] states that source and channel coding can be designed independently under certain idealized circumstances. In practice, however, it will often be advantageous to code for the source and channel jointly.

Manuscript received April 1991; revised August 1992; recommended for transfer from the IEEE TRANSACTIONS ON COMMUNICATIONS by the IEEE/ACM TRANSACTIONS ON NETWORKING Editor-in-Chief. M. Vetterli's work was supported by the National Science Foundation under Grants CDR-84-21402 and MIP-90-14189 and Bellcore. This paper was presented in part at the Third International Workshop on Packet Video, Morristown, NJ, March 1990.

M. W. Garrett is with Bell Communications Research, Morristown, NJ 07960. (email: mwg@faline.bellcore.com)

M. Vetterli is with the Department of Electrical Engineering and Center for Telecommunications Research, Columbia University, New York, NY 10027.

IEEE Log Number 9206170.

Multiresolution signal processing is a technique of successive approximations to a signal [9]. An ordered set of subsignals may be derived such that the first is a coarse approximation, and addition of subsequent components increases quality with decreasing marginal improvement. Techniques such as subband, pyramid, and transform coding [10]–[13] involve multiresolution decomposition. These are typically used for bandwidth reduction (source coding) because higher-order components or coefficients have lower variance and may be compressed by entropy coding. This lower variance also means these components contribute less to the signal in a mean squared error sense. Therefore, by using lower priority and concentrating losses on these subsignals, a channel coding procedure is created that is well matched to the source coding. (Also, the signal quality may be further protected via use of more powerful error correcting codes on the high-priority components.) In packet video, such joint source/channel coding has been used [14]–[17] and is sometimes called hierarchical, embedded, or layered coding.

Recursive coding schemes such as DPCM, although effective for compression, are less amenable to multiresolution decomposition. DPCM can be modified to be multiresolution-compatible with some loss of performance due to suboptimal prediction [18]. However, the quality may be sensitive to even a slight high-priority packet loss rate.

In general, joint source/channel coding includes any adaptation of the coding procedure to measured or anticipated network conditions. This reflects the fact that the coding method which yields highest quality will change depending on available resources. To demonstrate this property, we choose a simple one-parameter code, but a more complex dependence of the coding algorithm on the network is also possible.

B. Related Work

Much queueing theoretic work has been published recently [19]–[24] showing that packet voice is significantly more bursty and ill-behaved than Poisson sources. The queueing analysis is also more difficult, and efforts have been made to show where simpler models may be used effectively [25]–[27]. These analyses tend to measure performance in terms of expected delay or probability of loss by overflowing a finite FIFO buffer. A rich theory exists for these performance measures mainly because they are appropriate for evaluating data networks or arrival of voice calls to a circuit switch. For packet voice and video, however, we must be concerned with the tail distribution of delay rather than expectation, and loss should occur through timeout rather than buffer overflow. Furthermore, since subjective quality depends strongly on what information is lost, it is incumbent upon the performance analyst to use objective measures which differentiate among signal components rather than simple packet delay or loss statistics.

Some work has included more innovative models which take advantage of the multiresolution characteristic of voice by dropping low-order bits in response to congestion [27]. In other studies, voice is coded and prioritized to yield smooth degradation with loss [28], [29] or the fixed delay constraint is explicitly included [30]. Only very preliminary analytic modeling of variable-rate video sources has been done to date [31]–[34], because of both the complexity of the output process and the large variety of coding methods.

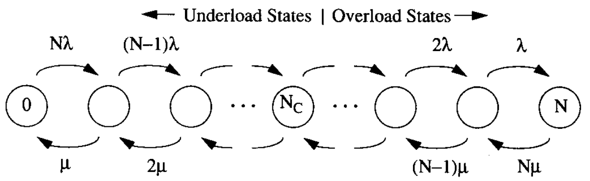


Fig. 1. Markov chain source model.

Our work derives mainly from that of Yin [7], [35], [36], which shows that priorities can be used advantageously for packet voice services. He shows several techniques for separating the voice signal into two priorities, and for controlling the loss of low-priority packets during periods of congestion. We expand on this first by using signal-to-noise ratio (SNR) as the performance measure [37], [38]; second, by examining the optimization of α in detail; and third by experimenting with real voice and video samples. SNR is meaningful for real-time services because it accounts for the variable significance of signal components in a way that is consistent with human perception. It also combines loss rates of different components into a single metric. We study this system by simulation, and so are free to use packet timeout as a loss mechanism. This is more natural to the services, and would be very difficult to pursue using Markov queueing analysis.

The outline of the paper is as follows: The next section describes the source and queue models, followed by a development of the SNR performance measure in Section III. The main simulation results are given in Section IV for fixed and dynamically variable values of α. In Section V, we describe variations on this theme including multiple priority levels and network feedback-path delay. Examples using real voice and video samples are given in Section VI.

II. System Description

A. Source Model and Coding

We construct a simple system which includes a variable-rate source model and a priority queueing discipline. The Markov chain shown in Fig. 1 has been widely used to model an aggregation of packet voice sources with silence suppression [39], and may be used to capture the basic behavior of packet video [31]. The Markov process, $n(t)$, is the number of sources active at time t out of N sources in the system. For each speaker, the alternating periods of activity and silence are exponentially distributed with average durations of $1/\mu$ and $1/\lambda$, respectively. This gives the state transition rates from any state n as $(N-n)\lambda$ toward state $n+1$ and $n\mu$ toward state $n-1$. The state N_C corresponds to the number of active sources which can be transported without loss on the given channel capacity, C. The states $n : n \leq N_C$ are thus underload states, while $n : n > N_C$ are overload states.

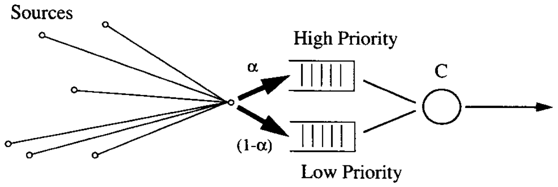


Fig. 2. Simple network model with priority queue.

Each source produces periodic samples of b bits. The samples are divided into two parts, which are packetized separately, forming two levels of packet priority. The most significant αb bits are placed into high-priority packets and the remaining $(1-\alpha)b$ bits into low-priority packets. (Note: for this kind of coding, α is restricted to the values k/b, $k \in \{0, 1 \ldots b\}$.) This prioritization uses the PCM sample as a natural multiresolution signal [40]. Since packets are assumed to be of fixed length, the time to generate high- and low-priority packets will depend on α. High- and low-order parts of a sample may be transmitted at slightly different times, especially for values of α near zero or one. Each packet is assumed to carry the value of α used so the samples can be properly decoded, and a field indicating the beginning, middle, or end of an active period. This allows a processor at a queueing point inside the network to track the number of active sources, $n(t)$, which are routed through that point. Alternatively, $n(t)$ can be estimated by measuring the number of packets received in one interpacket time interval.

B. Queueing Discipline and Loss Mechanism

The network is modeled in Fig. 2 as a single-hop two-priority queue. Note that the individual sources which make up the aggregate process, $n(t)$, are not necessarily co-located. Active sources contribute packets to the priority queue with independent phases. The server is work-conserving, and serves fixed-length packets periodically. Priority is nonpreemptive (i.e., a low-priority packet *in service* is not interrupted for an arriving high-priority packet). The channel capacity, C, is sufficient for no more than N_C sources, without loss. The queue has the property that, within each priority, packets retain their order but across priorities they do not. Reordering at the receiver is not a complicated problem. Arriving packets are stored in FIFO queues by priority, and at the time a sample is to be reconstructed, its components are either at the head of line in each queue, or are too late and will be discarded upon arrival. No sorting is necessary.

We use a single-hop network model to capture the basic characteristics of interest. A multiple hop network will have more complex behavior, and it becomes both harder to simulate and more difficult to interpret the results. If congestion is rare, the probability that packets from a given source encounter two congested nodes is much less than that of encountering one congested point. Short queueing delays at noncongested nodes and propagation time can be modeled as a constant added delay (see Section IV-B). Also, in a prioritized system, the second congested node encountered has a smaller effect than the first one. At the second congestion point, any undamaged connections in the cross-traffic will lose their low-priority packets before any high-priority traffic is discarded.

For a real-time service, it is important that the packet loss mechanism be modeled appropriately. The time between coding voice or video information at the source and reconstruction at the destination is fixed, and represents the useful lifetime for each packet (denoted t_L). Packets delayed longer than t_L will be discarded by the destination, so they might as well be discarded by the network using a lifetime enforcement mechanism. In most queueing analyses, loss measurements are derived from the probability of overflowing a finite buffer. The timeout loss mode can be assumed to dominate over the buffer overflow loss mode because buffers with delays exceeding t_L can be easily constructed. Even for a large system with $N_C = 200$, $b = 12$ (other parameters are given below in Table I), the buffer space needed to ensure timeout before buffer overflow (240 kBytes) would easily fit on one chip. This mechanism benefits network congestion control because it ensures that out-of-date information is not transported. Also, once an overload condition subsides, the congestion dissipates more quickly, allowing normal operation to resume.

III. SNR AS A PERFORMANCE MEASURE

The design of a system which provides any service *efficiently* requires a measure of quality that is relevant to the service. A simple packet loss probability may be adequate for measuring data transport performance because lost packets must be retransmitted. For voice and video, however, not all losses need to be recovered. Further, perceived quality can be dramatically improved if losses are imposed on the less significant components. Separate measures of loss rate by priority is of limited use since one does not know how to compare losses of different priorities. We, therefore, propose the signal-to-noise ratio (SNR), which combines the two loss rates, accounts for the variable significance of information, and better reflects perceived quality for these services. Since SNR is a long-run average measure, however, it does not convey loss correlations which can affect subjective quality.

First, we compute the probabilities of receiving the high- and low-order portions of each voice sample. Denote p_{HL} as the probability of receiving both the high- and low-order parts of sample p_{OO} as the probability of total loss, and p_{HO} and p_{OL} as the probability of receiving the high- and low-order parts alone, respectively. Then, the high- and low-priority loss rates, p_{lh}, p_{ll}, which can be measured in the simulation, can be expressed as

$$\begin{aligned} p_{lh} &= p_{OO} + p_{OL} \\ p_{ll} &= p_{OO} + p_{HO}. \end{aligned} \tag{1}$$

To find expressions for p_{OO}, p_{HO}, p_{OL}, and p_{HL} in terms of p_{lh} and p_{ll}, we assume $p_{OL} = 0$ because this represents the case where the high-priority packet is lost while the corresponding low-priority packet is delivered. (The simulations verify the correctness of this assumption.) For multihop networks, this can be expected to break down somewhat but should remain reasonably valid if both priorities follow the same route. From (1) and setting the total probability to unity,

TABLE I
SIMULATION PARAMETERS

$1/\lambda$	600 ms
$1/\mu$	400 ms
b	12 bits/sample
R	8 kHz sample rate
L	48 bytes/packet
N_C	24 equivalent channels
t_L	100 ms (Simulation time = 3000 s)

we find

$$p_{OO} = p_{lh} \qquad p_{HO} = p_{ll} - p_{lh}$$
$$p_{OL} = 0 \qquad p_{HL} = 1 - p_{ll}\ . \tag{2}$$

The signal and noise energy functions are defined as

$$J_s = \sum_x p_x x^2 \tag{3}$$

$$J_n = E\left[\sum_x p_x (x - \hat{x})^2\right]. \tag{4}$$

The probability distribution of sample values, p_x, is taken to be uniform, which is approximately correct for companded voice samples (counting talkspurts only). The sample value, when the low-order bits are lost, $\hat{x}$, depends on α and the expectation (E) over the sum accounts for p_{lh} and p_{ll}. Details of this computation are given in the Appendix. The signal-to-noise ratio, in decibels, is

$$\text{SNR} = 10\log(J_s/J_n) \quad \text{dB}. \tag{5}$$

We thus effectively combine the two packet loss probabilities into a single measure of performance which is strongly related to the service (rather than the network) and therefore better reflects user perception of quality.

IV. SIMULATION RESULTS

The simulation was run with the parameters shown in Table I. A small system of 24 equivalent channels and a large sample size of 12 bits were chosen to bring out the structure of the results while keeping simulation time reasonably low. The source model parameters, λ and μ, are chosen to agree with a packet voice model and, therefore, depend on the energy threshold level used in the silence suppression mechanism. We found several sets of parameters in the literature with speaker activity factors ranging from 33% to 48% [19], [20], [25], [41], [42]. We use the median values with $s = 40\%$ voice activity. The simulation time is chosen large enough that the confidence intervals are generally negligible. For example, in Fig. 4 the 95% confidence intervals are only measurable at the knee of each curve, at about 1.5 dB. At low values of G, the results are very sensitive because almost no packets are lost. The points at $G = 1.75$ have, therefore, been simulated for 10 ks instead of 3 ks.

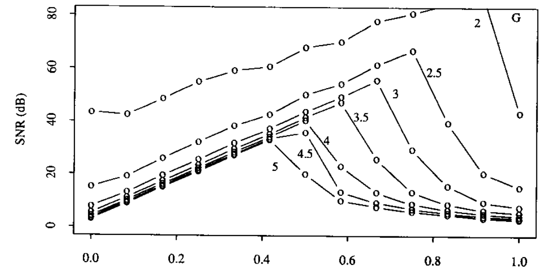


Fig. 3. Performance versus coding parameter α for various load values.

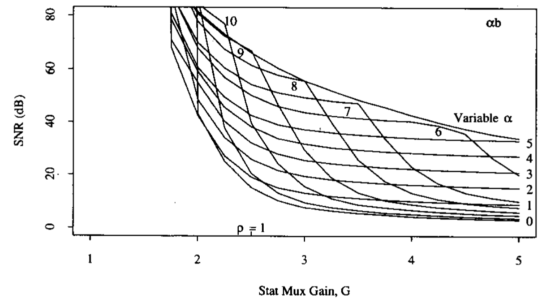


Fig. 4. Performance versus statistical multiplexing gain, G, for all values of αb and the variable α protocol.

A. Fixed α System

We first simulate the system with the fraction of high-priority traffic, α, fixed. The results in Fig. 3 show SNR plotted against α. Several curves are given for different values of G. At the extreme values of $\alpha = 0$ and $\alpha = 1$, all traffic is in one priority class, and the queueing discipline is exactly FIFO. When traffic is divided into two priorities, the performance is always better, with the optimum α depending on G (and, therefore, on p_l). We can understand the behavior shown in Fig. 3 as follows: At low values of α, only a small fraction of traffic has high priority, and it is effectively protected from loss. As α increases, more bits are placed in high-priority packets, and the SNR improves despite the increased low-priority loss rate. At some point, enough traffic has shifted to high-priority packets that high-priority loss becomes significant. Performance is very sensitive to p_{lh} and the SNR drops rapidly with further increase in α. As will be shown below, the best performance is reached when there is flow matching between high–priority traffic and available resources, i.e., $\alpha \approx N_C/n(t)$.

We redisplay the same results in Fig. 4 with SNR plotted against G and the number of high-priority bits, αb, as a parameter. As expected, the performance generally degrades with increasing load. At the point where high-priority packets start to be lost, each curve drops much more sharply. For each value of G, the curve with the highest SNR identifies the best value of α. The unprioritized cases ($\alpha b = 0$ or $\alpha b = 12$) appear as the lower envelope of the family of curves, showing that prioritization improves performance for any α. For loads up to $G = 3$, the difference in SNR, with and without priority, is about 45 dB.

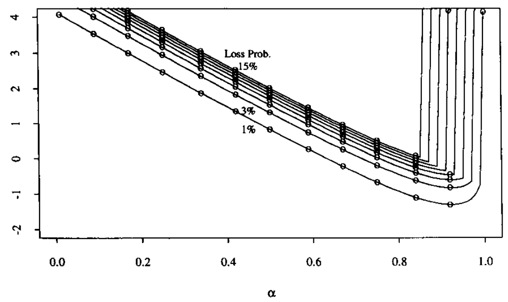


Fig. 5. Noise energy versus α for several values of loss probability p_l. Allowed values of α are indicated by circles.

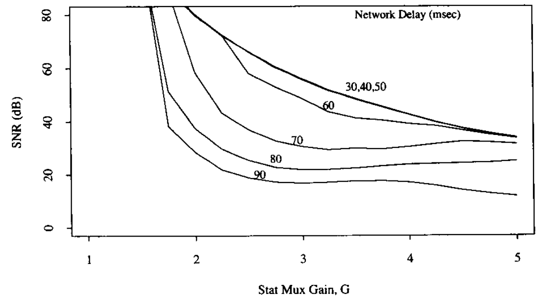


Fig. 6. Performance of variable α system with network delay.

Simulations are carried out for G ranging from 1 to 5. This corresponds to an average offered load of ρ = 0.4 to 2. We choose to look at the high load and overload regions for several reasons. First, the most "valid" region to study is where $G > 1$ and $\rho < 1$, i.e., sources are statistically multiplexed but the system is not overloaded on average. The system has interesting behavior here, and it is easier to understand in a larger context. Second, although we examine the random source process given N (or G), there is also the random process of call arrivals. That is, N varies with time, and the system may be in overload for some period with still acceptable performance. In this system with a timeout loss mechanism, no unstable behavior is observed at $\rho = 1$ (G = 2.5). Third, even though quality degrades for persistent overload, this may be considered a feature where the quality of service depends on system load. By analogy to the telephone system, this would mean that rather than getting "all trunks busy" on Mother's Day, a customer would get a line with a bit more noise. This identifies a tradeoff to be exploited between concentrating total loss of service on a few customers, and distributing a small degradation over all customers.

B. System with Dynamic Control

Here, we consider a system where α is continuously optimized according to the current source state $n(t)$. A heuristic argument for determining the best coding partition, α_{opt}, is that as much traffic as possible should be given high priority, subject to the condition that none of it is lost. This results in a flow-matching procedure at each time t, such that the fraction of traffic given high priority α does not exceed the fraction of traffic which can be supported $N_C/n(t)$. With the constraint that bits not be divided (i.e., αb is an integer), this yields

$$\alpha_{\text{opt}} = \frac{1}{b}\left\lfloor \frac{bN_C}{n(t)} \right\rfloor. \tag{6}$$

This heuristic is confirmed by minimizing the noise energy with respect to α, as shown in Fig. 5. Details of the calculation appear in the Appendix. To a first approximation, the optimization of α results in all capacity being used by high-priority traffic or $p_{lh} \approx 0$ and $p_{ll} \approx 1$.

The system was simulated with sources using α_{opt} determined by feedback from the network. The resulting curve, labeled "variable α" in Fig. 4, lies slightly above the envelope of the fixed α curves. The variable α curve does not coincide exactly with the fixed α curves because, for a given value of G, the instantaneous load is varying with time. The dynamically adaptive system tracks the source, and the high-priority traffic can always be very close to the available capacity.

The variable α curve does not have the characteristic "knee" of the fixed α cases because high-priority loss is avoided. Performance degrades gracefully with load. α_{opt} is calculated from $n(t)$ and fed back to all sources without delay (except packet construction time). This constitutes an optimal source/channel coding which performs better than any of the fixed α systems.

We next simulate a system where a constant delay is added to each packet between the source and the queueing point, to represent additional network queueing and propagation delay. These results are shown in Fig. 6. The SNR is quite insensitive to network delay values less than 60 ms, which is a substantial fraction of the packet lifetime. When feedback delay is very long, the feedback information becomes irrelevant. For instance, the 70 ms delay curve behaves similarly to the fixed $\alpha b = 4$ case. With such long delay, the advantage of adaptation is defeated.

V. Variations

A. Multiple Priorities

In the variable α case, each source matches the high-priority traffic rate to an estimate of its share of the network capacity. The low-priority traffic only gets service when random fluctuations cause sources to underestimate available capacity, or capacity is left because $\alpha_{\text{opt}}b$ is rounded down to an integer value. The low-priority traffic that is served carries bits of different significance, and may itself be partitioned into two prioritized parts. This makes three priority levels. Rather than send the lowest component at a third priority, the source may discard this information immediately. We refer to this mechanism as "source dropping." In Fig. 7, results are given for three cases with source dropping. The best performance results when only one bit per sample is carried in low-priority packets. High-priority packets are constructed adaptively using $\alpha_{\text{opt}}b$ bits per sample as before. Another curve shows that if the low priority is limited to 4 bits per sample, the SNR is 2 or 3 dB lower. The variable α curve, where low priority is not limited, is shown for comparison. If no low-priority

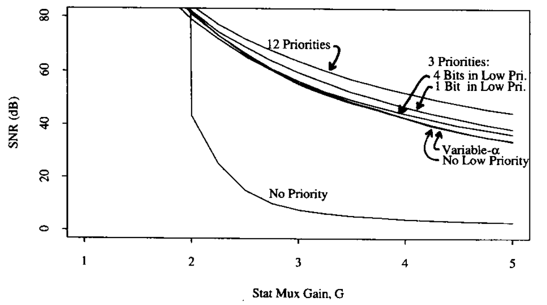


Fig. 7. Variations with 1, 2, 3, and 12 priorities.

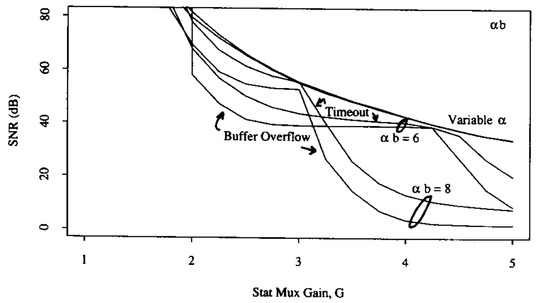


Fig. 8. Comparison between timeout and buffer overflow loss mechanisms. (αb = 6, 8, and variable α protocol.)

packets are sent at all, the performance, interestingly, is about equal to the variable α case. This case corresponds to a single priority with the source rate being adjusted to the network availability [28], [44], [45]. The simple two-priority, variable α case performs so badly (by this comparison) because too much of the small capacity available to the low priority is wasted on very insignificant information.

The topmost curve shows behavior with the 12 bits sent in 12 separate priorities. Here, the system complexity is shifted from having to keep track of the control information $n(t)$ or $\alpha_{\rm opt}$ to having to support 12 levels of priority for a single service. (Presumably, this service with its priorities would be embedded in a larger structure of resource allocation and scheduling [4].) In Jain's nomenclature, this shift is from "source-based" to "switch-based" congestion control [46]. The 12-priority curve appears 4 to 6 dB better than the best source-dropping case, which in turn is about 4 dB better than two priorities with variable α. The unprioritized case is shown for comparison. Overall, Fig. 7 shows that the gains of simple prioritization are significant, with smaller additional gains as the system is made more sophisticated.

B. Timeout versus Buffer Overflow Loss Mode

The timeout loss mechanism is more natural for a real-time service and should offer some performance advantage over simple buffer overflow. In Fig. 8, results are given for a pure buffer loss mode where the buffer size corresponds to the time constraint of 100 ms (For our parameters, $Q_{\rm max} = R \cdot b \cdot N_C \cdot t_L/L = 600$ packets.) Note that the variable α cases match closely, indicating that buffer availability is largely irrelevant. High-priority traffic is served without much queueing delay, and low-priority traffic is almost all dropped by either mechanism. With fixed α, the two schemes are closest at the critical point where loss moves from low to high priority, again because queueing is not significant here. At other points, the timeout mechanism provides significantly better performance than buffer overflow.

TABLE II
STATISTICS OF EIGHT DCT COEFFICIENTS FOR VOICE SAMPLES

k	relative variance	entropy	α
1	4.3233	4.80	0.223
2	2.5577	4.23	0.436
3	0.7431	3.16	0.588
4	0.2198	2.36	0.702
5	0.0950	2.07	0.802
6	0.0466	1.72	0.885
7	0.0134	1.35	0.950
8	0.0011	1.04	1.000

VI. DCT-CODED VOICE AND VIDEO

In this section, we measure the effectiveness of prioritization using actual samples of voice and video with more realistic coding. We replace the PCM-embedded code with the discrete cosine transform (DCT). Here, each block of 8 samples for voice, or 8 × 8 picture elements for video, are transformed into a set of coefficients which are ordered according to frequency (in one or two dimensions). These signals have correlation properties that result in higher variance for the lower-frequency coefficients. This means that entropy coding techniques are effective in reducing the amount of information necessary to transmit the higher-order coefficients. Also, these contribute less to the signal in a mean squared error sense. Thus, the DCT is appropriate for joint source/channel coding.

We construct our experiment for voice as follows: Forty speech samples comprising four different sentences spoken by both male and female speakers are coded using an eight-sample DCT. The variances of each coefficient are shown in Table II to decrease monotonically, indicating that more energy is contained in the lower frequencies. The eight coefficients are then each quantized to 8 bits and assigned either low or high priority according to a threshold. After measuring the distribution of values, a Huffman code table is generated for each coefficient separately. This assigns a variable-length codeword to each of the 256 values, which minimizes the entropy, or average number of bits necessary to transmit the coefficient. For each value of k in Table II, the entropy represents a proportion of the traffic which is attributed to that coefficient, so the corresponding α is the entropy of the bits from 1 to k normalized to the total entropy.

Given α and p_l, we use the approximation of (13) (see the Appendix) to find p_{lh} and p_{ll}. High- and low-priority coefficients are discarded randomly and replaced with their mean values according to these loss rates, and the SNR is calculated from (3)–(5), where the sum is over the actual sample distribution rather than over a uniform distribution. For each case, four random patterns of loss were averaged together.

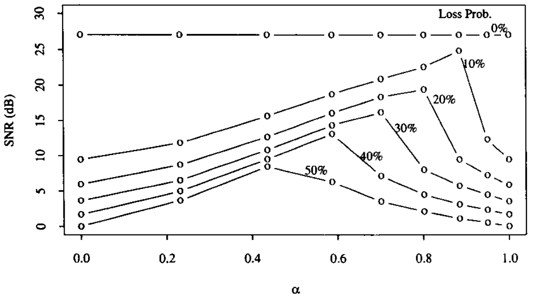


Fig. 9. DCT voice performance versus α for several values of loss probability (8 point DCT coding).

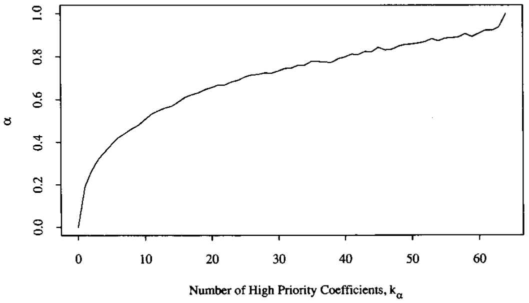


Fig. 11. α as a function of number of high-priority coefficients for video experiment.

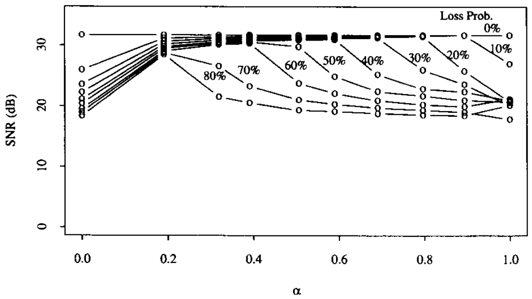


Fig. 10. DCT video performance versus α for several values of loss probability (8 × 8 DCT coding).

Fig. 9 shows the results, which bear a strong similarity to those of Fig. 3. The performance is given for various values of overall loss probability p_l (which can be related to G).

For packet video, we construct a similar but more complex experiment on ten random frames taken from a movie. Each monochrome frame consists of 480 × 512 pixels with 8 bits of gray-scale resolution. Each frame is divided into blocks of 8 × 8 pixels, and each block is encoded using the DCT. The coefficients are ordered according to their distance from the origin and quantized. With the large number of zero-valued coefficients, the video bandwidth is reduced by run-length encoding. The first k_α coefficients are given high priority, and the rest low priority. Within each priority, Huffman codewords are assigned to nonzero coefficient values and to runs of consecutive zero-valued coefficients. The high-priority stream then typically consists of some nonzero coefficients and a codeword indicating the number of remaining zeros. The low-priority stream usually has only zero-run codewords. When there is activity in high frequencies, then nonzero DCT coefficients will appear in the low priority. Each stream is segmented into packets of L bytes.

The code is first run without loss to determine the value of α which corresponds to each partition of coefficients k_α. Then, losses are imposed at the rates given by (13). Each low-priority packet is discarded with probability p_{ll}. In order to maintain correlation of losses, when low-priority loss occurs, the high-priority packet being processed concurrently is dropped with probability p_{lh}/p_{ll}. (This is the conditional probability of high-priority loss given low-priority loss.) The noise energy J_n is averaged over the ten images, and the SNR is calculated. By using the approximation of (13), we ignore the stochastic behavior of the video source, which may affect performance significantly. These results are intended to show the effect of prioritization rather than absolute performance.

Results for the video experiment are shown in Fig. 10. The SNR does not vary as much as in previous cases, but the peak performance clearly occurs at a value of α which decreases with increasing load. The compression of this curve is due partly to the extreme concentration of information in the lowest coefficients as shown in Fig. 11. Of the 64 coefficients, the first accounts for 19% of all traffic, the second 7%, etc.

VII. Conclusions

This study has demonstrated the effectiveness of prioritized transport of real-time services in packet-oriented networks. The use of joint source/channel coding to partition the signal into two priorities increases performance significantly, especially if the fraction of traffic placed in high-priority α can be adjusted to match traffic conditions. Further complexity in the priority scheme yields some additional gain.

We have assumed a timeout-driven loss mechanism, which is currently difficult to build for high-speed networks, because it requires sequential checking or sorting of packet lifetime fields. As VLSI technology improves, it will become feasible to use parallel hardware to actively discard packets which have timed out in queue. (For instance, the technology of content addressable memories is promising [47].) The main value of priorities is substantiated for a currently realizable buffer overflow loss mode, with somewhat inferior behavior.

We have attempted to include concepts and approaches from both signal processing and network protocol points of view. This work may be relevant to the current standardization efforts of both ATM network design and variable-rate packet video. Although initial ATM service offerings will probably use simple peak allocation, the possibility of significantly increasing capacity by modifying the switch queueing disciplines should not be ruled out. For voice traffic, the required bandwidth is small, so the use of compression or variable-rate coding may not be interesting. For video, however, the bandwidth requirement is significant. Early studies indicate a peak-to-average ratio (which is the potential cost reduction factor) in the range 1.5 to 4.5 [48]–[51]. This figure may increase as variable-rate video coding is refined.

Although our simulation has been parameterized for voice, it can be taken as a very simple model for video. A realistic model for video will be much more complex and difficult to develop than that for voice. We look forward to future work which applies the concepts developed here of joint source/channel coding (network/source joint optimization, layering, prioritization, etc.) to more sophisticated and realistic video experiments. The SNR performance measure developed here is an improvement over using packet loss rates for evaluating real-time services. However, it is an objective measure, and should be checked against subjective measures such as mean opinion score (MOS).

Data services have much higher burstiness than either voice or video [52], and a similar design approach might prove fruitful. Some data services are more sensitive to delay than others, and the traffic stream may be divided into substreams to be prioritized. Although loss is not acceptable (without retransmission), the parts identified as low priority would be those more tolerant of considerable delays, allowing the total stream to be compressed into a smaller bandwidth than would be required otherwise for good performance. This idea merits further investigation as LAN interconnection is expected to be a significant traffic component for B-ISDN [53].

VIII. Appendix

A. Noise Energy Calculations

Consider the sequence of sample values representable by b bits: $-(2^{b-1}-1)\ldots 0 \ldots 2^{b-1}$. We calculate the signal energy as

$$J_s = \frac{1}{2^b} \sum_{x=-(2^{b-1}-1)}^{2^{b-1}} x^2 = \frac{2^{2b}}{12} + \frac{1}{6}. \tag{7}$$

If the last $(1-\alpha)b$ bits are lost and replaced by their mean value of $2^{(1-\alpha)b-1}$, then the sequence of recovered sample values consists of $2^{\alpha b}$ groups of repeated values (i.e., the fine resolution has been lost). The differences between the samples and their recovered values form a linear sequence within each group, and the noise energy conditioned on low-priority loss is calculated as

$$\begin{aligned} J_{n*} &= \frac{1}{2^b} \sum_{x=-(2^{b-1}-1)}^{2^{b-1}} (x-\hat{x})^2 \\ &= \frac{1}{2^b} 2^{\alpha b} \sum_{y=1}^{2^{b(1-\alpha)}} \left(y - 2^{b(1-\alpha)-1}\right)^2 \\ &= \frac{2^{2b(1-\alpha)}}{12} + \frac{1}{6}. \end{aligned} \tag{8}$$

This corresponds to the energy of a signal where samples are quantized to $b(1-\alpha)$ bits. Thus, for $\alpha = 1$, $J_{n*} \approx 0$, and for $\alpha = 0$, $J_{n*} = J_s$. (Note: If we had chosen the sample values symmetrically about zero, we would have had $J_{n*} = 0$ with $\alpha = 1$.) Using (2), the expected noise energy of (4) is then given by taking a weighted average over the two possible loss modes: total loss or low priority loss.

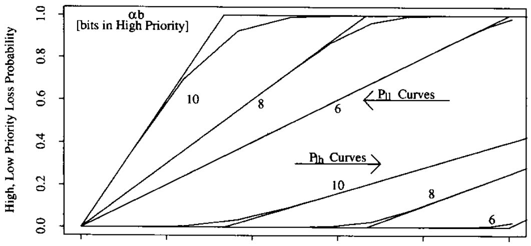


Fig. 12. Priority loss rates as a function of total-loss rate simulation results and idealized curves ($b = 12$).

$$J_n = p_{lh} J_s + (p_{ll} - p_{lh}) J_{n*}. \tag{9}$$

It can be shown that this generalizes to multiple priorities as

$$J_n = \sum_r J_{n*}(r) \left(p_{lr} - \sum_{q=1}^{r-1} p_{lq} \right) \tag{10}$$

with

$$J_{n*}(r) = \frac{2^{2(b-m(r))}}{12} + \frac{1}{6} \tag{11}$$

where $m(r)$ is the number of sample bits in the rth priority, and p_{lr} is the corresponding loss probability.

B. Calculation of Optimal α

Considering the priority queue of Fig. 2, we construct an idealized relation between the loss probabilities for high- and low-priority packets p_{lh}, p_{ll} and the overall probability of loss p_l. Since the stochastic variations of the source are significantly smoothed by the buffer, we expect that for low-loss rates, all loss is imposed on low-priority packets. (This is less true for very small systems, $N < 24$.) High-priority packets are then only lost after the low-priority traffic is completely discarded. This deterministic approximation gives the relations:

$$p_{ll} = \begin{cases} \min\left(\frac{p_l}{1-\alpha}, 1\right), & \alpha \neq 1 \\ 0, & \alpha = 1 \end{cases} \tag{12}$$

$$p_{lh} = \begin{cases} \max\left(0, \frac{p_l - (1-\alpha)}{\alpha}\right), & \alpha \neq 0 \\ 0, & \alpha = 0. \end{cases} \tag{13}$$

These are plotted in Fig. 12 for several values of αb along with simulation results. The system follows (13), diverging only at the "interesting" region where the loss rate matches the low-priority traffic volume ($p_l \approx 1 - \alpha$). We next use (13) as an approximation to the system behavior to estimate the optimal value of alpha. It is also used in Section VI to relate performance to p_l and α for DCT-coded voice and video samples.

Combining (9), (12), and (13), we calculate α_{opt} by minimizing the noise energy

$$J_n = \begin{cases} \frac{p_l}{1-\alpha} J_{n*} & \alpha \leq 1 - p_l \\ \frac{p_l - (1-\alpha)}{\alpha} J_s + \frac{1-p_l}{\alpha} J_{n*} & \alpha \geq 1 - p_l. \end{cases} \tag{14}$$

For any loss probability, J_n decreases with α over the range $\alpha \leq 1 - b^{-1}$ and increases for $\alpha \geq 1 - p_l$. Fig. 5 shows this function, with the allowed values of α indicated by circles. In the remaining range, $1 - b^{-1} < \alpha < 1 - p_l$, which only exists for small values of p_l, the slope of J_n changes from negative to positive (e.g., see the 1% curve in Fig. 5). Since no allowed values of α can exist in this region, the minimum J_n occurs at the allowed value of α given by (6).

References

[1] M. Nomura, T. Fujii, and N. Ohta, "Basic characteristics of variable rate video coding in ATM environment," *IEEE J. Select. Areas Commun.*, vol. 7, no. 5, pp. 752–760, June 1989.

[2] L. Zhang, "Virtual clock: A new traffic control algorithm for packet switching networks," in *Proc. ACM SIGComm Symp.*, Philadelphia, PA, Sept. 1990, pp. 19–29.

[3] S. J. Golestani, "Congestion-free transmission of real-time traffic in packet networks," in *Proc. IEEE INFOCOM*, San Francisco, CA, June, 1990, pp. 527–536.

[4] A. A. Lazar, A. Temple, and R. Gidron, "An architecture for integrated networks that guarantees quality of service," *Int. J. Dig. and Analog Commun. Syst.*, vol. 3, pp. 229–238, 1990.

[5] S. H. Lee, "An integrated transport technique for circuit and packet switched traffic," in *Proc. IEEE INFOCOM*, Mar. 1988, pp. 110–118.

[6] L. Turner, T. Aoyama, D. Pearson, D. Anastassiou, and T. Minami, Eds., *IEEE J. Select. Areas Commun.*, vol. 7, no. 5, June 1989.

[7] N. Yin, "Voice congestion control in integrated packet-switched voice/data sytems," Ph.D. Dissertation, Dept. Elect. Eng., Columbia Univ., New York, NY, 1988.

[8] R. E. Blahut, *Digital Transmission of Information.* Reading MA: Addison-Wesley, 1990.

[9] S. Mallat, "A theory of multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674–693, July 1989.

[10] M. Vetterli, "Multi-dimensional sub-band coding: Some theory and algorithms, signal processing," *IEEE Signal Proces.*, vol. 6, pp. 97–112, Feb. 1984.

[11] J. W. Woods and S. D. O'Neil, "Sub-band coding of images," *IEEE Trans. Signal Proces.*, vol. 34, no. 5, pp. 1278–1288, May 1986.

[12] P. J. Burton and E. H. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. 31, no. 4, pp. 532–540, Apr. 1983.

[13] N. S. Jayant and P. Noll, *Digital Coding of Waveforms.* Englewood Cliffs, NJ: Prentice-Hall, 1984.

[14] G. Karlsson and M. Vetterli, "Packet video and its integration into the network architecture," *IEEE J. Select. Areas Commun.*, vol. 7, no. 5, pp. 739–751, June 1989.

[15] M. Ghanbari, "An adaptive video codec for ATM networks," in *Proc. Third Int. Workshop on Packet Video*, Morristown, NJ, Mar. 1990.

[16] M. Nomura, T. Fujii, and N. Ohta, "Layered packet-loss protection for variable rate video coding using DCT, " in *Proc. Second Int. Workshop on Packet Video*, Torino, Italy, Sept. 1988.

[17] K. Shinamura, Y. Hayashi, and F. Kishino, "Variable bitrate coding capable of compensating for packet loss," in *Proc. SPIE Conf. Visual Commun. and Image Proces.*, Nov. 1988, pp. 991–998.

[18] D. J. Goodman, "Embedded DPCM for variable bit rate transmission," *IEEE Trans. Commun.*, vol. 28, no. 7, pp. 1040–1046, July 1980.

[19] C. J. Weinstein and E. M. Hofstetter, "The tradeoff between delay and TASI advantage in a packetized speech multiplexer," *IEEE Trans. Commun.*, vol. 27, no. 11, pp. 1716–1720, Nov. 1979.

[20] T. E. Stern, "A queueing analysis of packet voice," in *Proc. IEEE GLOBECOM*, San Diego, CA, Dec. 1983, pp. 2.5.1–2.5.6.

[21] Y. C. Jenq, "Approximations for packetized voice traffic in statistical multiplexer," in *Proc. IEEE INFOCOM*, Apr. 1984.

[22] J. N. Daigle and J. D. Langford, "Models for analysis of packet voice communication systems," *IEEE J. Select. Areas Commun.*, vol. 4, no. 6, pp. 847–855, Sept. 1986.

[23] R. C. F. Tucker, "Accurate method for analysis of a packet-speech multiplexer with limited delay," *IEEE Trans. Commun.*, vol. 36, no. 4, pp. 479– 483, Apr. 1988.

[24] S. Ganguly and T. E. Stern, "Performance evaluation of a packet voice system," *IEEE Trans. Commun.*, vol. 37, no. 12, pp. 394–397, Dec. 1989.

[25] H. Heffes and D. M. Lucantoni, "A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance," *IEEE J. Select. Areas Commun.*, vol. 4, no. 6, pp. 856–868, Sept. 1986.

[26] K. Sriram and W. Whitt, "Characterizing superposition arrival processes in packet multiplexers for voice and data," *IEEE J. Select. Areas Commun.*, vol. 4, no. 6, pp. 833–846, Sept. 1986.

[27] K. Sriram and D. M. Lucantoni, "Traffic smoothing effects of bit dropping in a packet multiplexer," *IEEE Trans. Commun.*, vol. 37, no. 7, pp. 703–712, July 1989.

[28] T. Bially, B. Gold, and S. Seneff, "A technique for adaptive voice flow control in integrated packet networks," *IEEE Trans. Commun.*, vol. 28, no. 3, pp. 325–333, Mar. 1980.

[29] D. W. Petr, L. A. DaSilva, and V. S. Frost, "Priority discarding of speech in integrated packet networks," *IEEE J. Select. Areas Commun.*, vol. 7, no. 5, pp. 644–656, June 1989.

[30] C. Yuan and J. A. Silvester, "Queueing analysis of delay constrained voice traffic in a packet switching system," *IEEE J. Select. Areas Commun.*, vol. 7, no. 5, pp. 729–738, June 1989.

[31] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson, and J. Robbins, "Performance models of statistical multiplexing in packet video communications," *IEEE Trans. Commun.*, vol. 36, no. 7, pp. 834–844, July 1988.

[32] P. Pancha and M. El Zarki, "Modeling video sources for resource allocation in ATM based B-ISDN," in *Proc. Third Int. Workshop on Packet Video*, Morristown, NJ, Mar. 1990.

[33] S.-S. Huang, "Source modeling for packet video," in *Proc. IEEE INFOCOM*, Mar. 1988, pp. 1262–1267.

[34] P. Sen, B. Maglaris, N.-E. Rikli, and D. Anastassiou, "Models for packet switching of variable-bit-rate video sources," *IEEE J. Select. Areas Commun.*, vol. 7, no. 5, pp. 865–869, June 1989.

[35] N. Yin, T. E. Stern, and S. -Q. Li, "Performance analysis of a priority-oriented packet voice system," in *Proc. IEEE INFOCOM*, June 1987.

[36] N. Yin, S. -Q. Li, and T. E. Stern, "Congestion control for packet voice by selective packet discarding," *IEEE Trans. Commun.*, vol. 38, no. 5, pp. 674–683, May 1990.

[37] M. W. Garrett and M. Vetterli, "A joint source/channel model for real time services on ATM networks," in *Proc. Third Int. Workshop on Packet Video*, Morristown, NJ, March 1990.

[38] M. Vetterli and M. W. Garrett, "Joint source/channel coding for real time packet services," in *Proc. Australian Video Commun. Workshop*, Melbourne, Australia, July 1990.

[39] S. -Q. Li, "A new performance measurement for packet voice transmission in burst and packet switching," *IEEE Trans. Commun.*, vol. 35, no. 10, pp. 1083–1094, Oct. 1987.

[40] N. S. Jayant and S. W. Christensen, "Effects of packet losses in waveform coded speech and improvements due to an odd–even sample-interpolation procedure," *IEEE Trans. Commun.*, vol. 29, no. 2, pp. 101–109, Feb. 1981.

[41] P. T. Brady, "A statistical analysis of on–off patterns in 16 conversations," *Bell Syst. Tech. J.*, vol. 47, pp. 73–91, Jan. 1968.

[42] R. L. Easton, P. T. Hutchinson, R. W. Kolor, R. C. Mondello, and R. W. Muise, "TASI-E communications system," *IEEE Trans. Commun.*, vol. 30, no. 4, pp. 803–807, Apr. 1982.

[43] J. G. Gruber, "Delay related issues in integrated voice and data networks," *IEEE Trans. Commun.*, vol. 29, no. 6, pp. 786–800, June 1981.

[44] J. M. Holtzman, "The interaction between queueing and voice quality in variable bit rate packet voice systems," in *Proc. ITC 11*, Sept. 1985

[45] M. Listani and F. Villani, "Voice communication handling in X.25 packet switching networks," in *Proc. IEEE GLOBECOM*, Dec. 1983

[46] R. Jain, "Myths about congestion management in high-speed networks," Tech. Rep. DEC-TR-726, Digital Equip. Corp., Littleton, MA, Oct. 1990.

[47] A. J. McAuley and C. J. Cotton, "A self-testing reconfigurable CAM," *IEEE J. Solid-State Circ.*, vol. 26, no. 3, Mar. 1991.

[48] W. Verbiest, L. Pinoo, and B. Voeten, "Statistical multiplexing of variable rate video sources in asynchronous transfer mode networks," in *Proc. IEEE GLOBECOM*, Dec. 1988, pp. 7.2.1–7.2.6.

[49] F. Kishino, K. Manabe, Y. Hayashi, and H. Yasuda, "Variable bit-rate coding of video signals for ATM networks," *IEEE J. Select. Areas Commun.*, vol. 7, no. 5, pp. 801–806, June 1989.

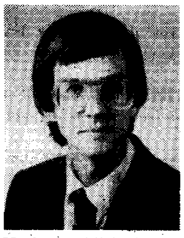
[50] M. Ghanbari and D. E. Pearson, "Statistical behavior of VBR-coded television pictures," in *Proc. Second Int. Workshop on Packet Video*,

Torino, Italy, Sept. 1988.

[51] H. S. Chin, J. W. Goodge, J. W. R. Griffiths, and D. J. Parish, "Statistics of video signals for viewphone type pictures," in *Proc. Second Int. Workshop on Packet Video*, Torino, Italy, Sept. 1988.

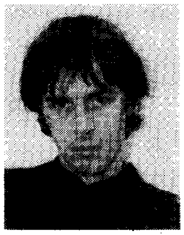
[52] W. E. Leland and D. V. Wilson, "High time-resolution measurement and analysis of LAN traffic: Implications for LAN interconnection," in *Proc. IEEE INFOCOM*, Apr. 1991.

[53] C.F. Hemrick, R.W. Klessig, and J.M. McRoberts, "Switched multimegabit data service and early availability via MAN technology," *IEEE Commun. Mag.*, vol. 26, no. 4, pp. 20–28, Apr. 1988.



Mark W. Garrett (S'82–M'84) received the B.S., M.S., and M.Phil. degrees in electrical engineering from Columbia University, New York, NY, in 1982, 1984, and 1991, respectively. He is presently pursuing his Ph. D. degree at the same university.

In 1984, he joined the Communication Science Research Division at Bell Communications Research, where his work concentrated on local and metropolitan area network design. He constructed the first media access control (MAC) chip to run at over 300 Mb/s (using the Fasnet LAN protocol). He has published a dozen research articles on packet communications. His current interests are in packet video and traffic control for multiservice packet networks.



Martin Vetterli was born in Switzerland in 1957. He received the Dipl. Elec.-Ing. degree from the Eidgenossische Technische Hochschule Zurich, Switzerland, in 1981, the Master of Science degree from Stanford University, Stanford, CA, in 1982, and the Doctorat es Science degree from the Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland, in 1986.

In 1982, he was a Research Assistant at Stanford University. From 1983 to 1986, he was a Researcher at the Ecole Polytechnique. He has worked for Siemens and AT&T Bell Laboratories. In 1986, he joined Columbia University, New York, NY, where he is currently Associate Professor of Electrical Engineering, member of the Center for Telecommunications Research, and Co-Director of the Image and Advanced Television Laboratory. He is a member of the MDSP Committee of the IEEE Signal Processing Society and of the Editorial Boards of *Signal Processing*, *Image Communication*, and *Annals of Telecommunications*. He received the Best Paper Award of EURASIP in 1984 for his paper on multidimensional subband coding, the Research Prize of the Brown Bovery Corporation (Switzerland) in 1986 for his thesis, and the IEEE Signal Processing Society's 1991 Senior Award (DSP Technical Area) for a 1989 Transactions paper with D. LeGall on filter banks for subband coding. His research interests include wavelets, multirate signal processing, computational complexity, signal processing for telecommunications, and digital video processing.