TRAFFIC SMOOTHING EFFECTS OF BIT DROPPING IN A PACKET VOICE MULTIPLEXER

K. Sriram and D. M. Lucantoni

AT&T Bell Laboratories, Holmdel, N.J. 07733

ABSTRACT

Recent studies have shown that the superposition of packet sequences generated by packetized voice sources with speech detection exhibit high burstiness due to inherent correlations between successive interarrival times in the superposition stream. In a packet voice multiplexer without bit dropping, these correlations tend to cause significantly larger queueing delays and packet losses than would be predicted for a Poisson input stream. In this paper, we examine the performance of a packet voice multiplexer queue in which the less significant bits on voice packets are dropped during states of congestion in the multiplexer. Using the results of simulation and analytical modeling, it is illustrated that bit dropping on voice packets significantly smooths the superposition packet voice traffic by speeding up the packet service rate during critical periods of congestion in the queue. This phenomenon renders it possible to approximate the superposition process by a Poisson process for analyzing a packet voice multiplexer with bit dropping. The multiplexer is modeled using an $M/D/1/K$ model in which D denotes the deterministic but state dependent nature of service. By comparison with a simulation, this model is shown to produce quite accurate performance predictions.

1. *INTRODUCTION*

Significant effort is currently being devoted to the development of packet oriented technologies for integrated multiplexing and switching of voice and data[1] [2] [3] [4] [5] [6]. Packetization of voice makes it possible to carry voice along with data on a single integrated packet-switched network. The advantages of integrated packet voice/data networks are many, e.g. the efficient sharing of transmission and switching facilities, the capacity advantage due to statistical multiplexing, and the potential evolution toward a fully integrated network which would provide image and video services as well. Some recent studies[7] [8] [9] have also focused on hybrid multiplexing schemes in which voice is serviced using synchronous time division multiplexing while data is serviced asynchronously through packet queues. In this memorandum, we concentrate on integrated networks in which voice packets are also queued and therefore served asynchronously. However, generally voice packets would be given priority over data so as to ensure a low delay for voice. Voice packets, however, do get jittered in time due to the variable queueing delays at the access interface as well as inside the network at the switching nodes enroute. A suitable delay build-out strategy is employed at the receiver to eliminate the delay jitter before playing out the voice to the listener[10].

In order to traffic engineer an integrated voice/data network, it is first necessary to understand the characteristics of the packet traffic generated by voice and data sources. In particular, it is of interest to know a way to characterize the superposition of packet arrival streams generated by a number of voice and/or data sources. Several recent studies have dealt with this issue in some detail[11] [12] [13] [14] [15] [16] [17]. The work reported here is a continuation of the authors' previous studies[11] [12] and, in particular, this memorandum investigates the smoothing effects on the superposition traffic in a packet voice multiplexer with bit dropping for congestion control. The bit dropping scheme, based on an embedded voice coding technique[18] [19] [20] [21] [22] [23] takes advantage of the fact that voice is tolerant to bit dropping to a certain extent. Based on subjective quality consideration, it is well recognized that it is better to drop bits on voice packets rather than drop whole packets[24] [25] (for a given ratio of carried over offered information).

Recent studies have shown that the superposition of packet sequences generated by packetized voice sources with speech detection exhibit high burstiness (relative to a Poisson process) due to inherent correlations between successive interarrival times in the superposition stream. In the papers by Sriram and Whitt[11] and Heffes and Lucantoni[12], the burstiness due to correlations has been investigated by using the indices of dispersion for counts and intervals. In a packet voice multiplexer without bit dropping, these correlations tend to cause significantly larger queueing delays and packet losses than would be predicted by a Poisson model. In this paper, we examine the performance of a packet voice multiplexer in which the less significant bits in voice packets are dropped during states of congestion in the multiplexer. It is illustrated here that bit dropping on voice packets significantly smooths the superposition packet voice process by speeding up the packet service rate during critical periods of congestion in the queue. In fact, this paper seeks to establish a very significant hypothesis, namely, *in a packet voice multiplexer with bit dropping, the superposition packet arrivals can be viewed as a Poisson process from a packet delay and queue length perspective.* The hypothesis is conjectured first by

CH2534-6/88/0000-0759 $1.00 © 1988 IEEE

looking at the structure of correlations in the superposition arrival process characterized by the index of dispersion for intervals (see Section 3). Then it is verified by comparing the results of exact simulations of the multiplexer with those of analytical modeling. The importance of the hypothesis is due to fact that the virtual smoothing of the packet arrival process as a result of bit dropping on the voice packets has significant implications in terms of effecting delay and buffer overflow reductions in packet voice/data multiplexers. This result (i.e. traffic smoothing due to bit dropping) also makes it easier to develop tractable analytical models for the multiplexer for the general mixed voice/data scenarios, and helps avoid undue conservatism in traffic engineering.

We model the packet voice multiplexer with bit dropping as a queueing system with a state dependent server whose service speeds up during states of congestion. The specific model used is an $M/D/1/K$ model in which D denotes the deterministic but state dependent nature of packet service times. The model enables us to compute various quantities of interest for performance characterization, e.g., the carried mean bit rate for a voice call, the queue length distribution, the packet loss due to queue overflows, the mean and variance of voice packet delay, etc. Our numerical examples predict capacities in terms of the number of speech sources that can be supported for different speech activity rates, line transmission rates and performance requirements. The numerical results also provide guidelines for traffic engineering and for selecting design parameters such as buffer size, bit dropping thresholds, etc.

Some simulation studies on packet voice multiplexers with congestion control by bit dropping were recently published in the literature[26] [27] [28]. In these papers, the thresholds for dropping bits had been chosen to be high (relative to queue fill values) and as a result the reduction in packet queueing delay is not significant. In the present memorandum, it is shown that by a prudent choice of bit dropping thresholds, the queueing delay in the multiplexer can be reduced effectively during congestion at the expense of moderate bit dropping with only a slight reduction in voice quality. Further, this memorandum also successfully demonstrates the validity of a Poisson model with state dependent service times for the packet voice multiplexer with bit dropping. Also some very useful insight is provided into the relationship between the bit dropping thresholds and the effect of correlations in the superposition packet voice process. The validity of a Poisson model is supported by observing that bit dropping tends to maintain a moderate buffer occupancy in the queue so that not too many packet arrivals can interact in the queue, and hence the effect of correlations on queueing delays is insignificant.

In Section 2, a packet voice multiplexer with bit dropping for congestion control is described. In Section 3, drawing upon our previous results[11] [12], we provide a foresight into the efficacy of a Poisson model for a packet voice multiplexer with bit dropping. In Section 4, we describe the $M/D/1/K$ and present its analysis. Numerical examples and discussion are presented in Section 5, and the conclusions are stated in Section 6.

2. DESCRIPTION OF A PACKET VOICE MULTIPLEXER WITH BIT DROPPING FOR CONGESTION CONTROL

A schematic of a packet voice multiplexer is shown in Fig. 1. Each voice source is packetized, and the superposition of all these sources constitutes the packet arrival process to the queue in the multiplexer. In general, the multiplexer would also service various data sources not shown in the Fig. 1. It is assumed here that the traffic present in the multiplexer is all bit droppable voice. Assume that the queue is finite with a buffer capacity of K packets, and that the service discipline is first-in first-out (FIFO). Each voice source is sampled at 8 kHz rate and encoded using an embedded ADPCM scheme at 32 kbps rate. One hundred and twenty eight samples are collected over a 16 ms interval, and they are organized into a packet as shown in Fig. 2. All the least significant bits from the 128 samples are contained in block #1 of the packet, the next significant bits are contained in block #2, and the two most significant bits are contained in blocks #3 and #4. We assume that a 10 byte header is used to incorporate a range of information about the packet such as its destination, the time-stamp information for build-out at the receiver, and other protocol related information . One bit in the header may be designated for indicating whether or not a packet is bit droppable. The multiplexer would in general have the capability to distinguish a voice-band data (VBD) call from a voice call, and can set this bit to prevent bit dropping on a VBD call. In this memorandum, we assume for simplicity that the traffic offered to the multiplexer is all voice. It is also assumed that voice silence detection is employed and packets are generated only during talkspurt periods as shown in Fig. 3. No packets are generated during silence periods. However, restoration of silence periods is automatically accomplished at the receiver as part of the play-out strategy that uses time-stamp information in the packet headers. The parameter values used in Fig. 3 are obtained from the work of May and Zebo[29].

The bit dropping algorithm used in the multiplexer is described in Table 1. Let L denote the current queue fill in packets. The voice packet is structured as shown in Fig. 2. The bit dropping on the packets is done at the output of the queue, i.e. at the server (see Fig. 1) just prior to transmission. When L is smaller than the first threshold, no bit dropping is done. When L exceeds the first threshold Q_1, but is still smaller than the second threshold Q_2, the block #1 containing the least significant bits (see Fig. 2) is dropped. When L exceeds Q_2, the blocks #1 and #2 are both dropped thus reducing the information in the packet down to the two most significant bits. The transmitted voice packet sizes and the corresponding service times (on a T1 link) are shown in Table 1 for each congestion state. A packet is lost only if the queue is full when it arrives. One may consider an algorithm which drops the whole packet at the server if a third threshold, say Q_3, is exceeded. We do not include this case in this study but the model can be easily extended to incorporate the same.

3. A FORESIGHT INTO THE EFFICACY OF A POISSON MODEL

It was shown in our earlier work[11] [12] that a Poisson approximation for the superposition voice packet arrival process underestimates the delays considerably in a packet voice multiplexer without bit dropping. The work had provided some very useful insight into the nature of the superposition process via the index of dispersion characterization. In this section, we draw upon this insight to present a foresight into the efficacy of a Poisson arrival approximation for a multiplexer which does bit dropping for congestion control.

Fig. 3 illustrates the packet arrival process due to a single voice source. Sriram and Whitt showed that this process is very bursty in the sense that its inter-arrival time has a large squared coefficient of variation c^2 (i.e., the ratio of variance to square of mean value). In fact, the value of c^2 is 18.1 for this process which is very large considering that a Poisson process has $c^2 = 1$.

The superposition of packet sequences generated by n voice sources is well characterized by the indices of dispersion for counts and intervals. To help the discussion in this section, let us restate here the definitions of these quantities. For this purpose, let $\{X_k,\ k \geq 1\}$ represent the sequence of interarrival times in the superposition of n voice sources. Let $S_k = X_1 + X_2 + ... + X_n$ denote the sum of k consecutive interarrival times. The index of dispersion for intervals (IDI), which we also call the k-interval squared coefficient of variation sequence, is the sequence $\{c_k^2,\ k \geq 1\}$ defined by

$$c_k^2 = \frac{k\ Var\,(S_k)}{[E\,(S_k)]^2} = \frac{k\ Var\,(X_1) + 2\sum_{j=1}^{k-1}(k-j)\ cov\,(X_1,\ X_{1+j})}{k\,[E\,(X_1)\,]^2} \tag{1}$$

where $cov\,(X_i,\ X_j) = E\,(X_iX_j) - E\,(X_i)\,E\,(X_j)$. Let $N(t)$ represent the counting process associated with the arrival process in consideration, i.e. $N(t)$ denotes the number of arrivals in time (0,t). Then the index of dispersion for counts (IDC), I(t), is a function defined as follows

$$I(t) = \frac{var\ [N(t)]}{E\ [N(t)]}\ ,\quad t > 0 \tag{2}$$

Several interesting properties of the IDI sequence $\{c_k^2,\ k \geq 1\}$ and the IDC function $I(t)$ were explored with specific reference to the superposition packet voice arrival process. In Figs. 4 and 5, the IDI sequence and the IDC function are shown, respectively, for the superposition packet voices process. Sriram and Whitt have discussed the notion of *relevant time scale* extensively. They observed that over a short time interval, the superposition process does look like a Poisson process which is characterized by $I(t) = 1$ for $t \geq 0$, and $c_k^2 = 1$ for $k \geq 1$. The c_k^2 curves indicate that positive correlations exist between the intervals, and the cumulative effect of these correlations causes c_k^2 to be substantially larger than 1 for large k. Thus, while on the one hand the single interval in the superposition process tends to an exponentially distributed variable as the number of component voice sources n increases (see Appendix in Sriram and Whitt[11]), on the other hand the positive correlations over many consecutive intervals cause the process to substantially deviate from a Poisson distribution. The notion of *relevant time scale* now comes into play strongly. When the superposition voice packet arrival process is offered as input to a multiplexer without bit dropping congestion control, a significant number of packets can queue up in the multiplexer at high utilization, and correlations between the consecutive intervals play a significant role in influencing the queueing behavior. This is the reason why the actual delays are significantly higher than those predicted by the Poisson model at high loads (see Figure 6). However, if a bit dropping congestion control is employed in the voice packet multiplexer, then the queue is prevented from building up to large values. Consequently, not too many packet intervals can interact in the queue, and hence the values of the IDI and IDC are relevant only over a small range where they tend to imitate a Poisson like behavior (see Figs. 4 and 5). The bit dropping mechanism speeds up the service at the critical moments and prevents the build up of a large backlog of packets in the queue. As long as only a few packets interact in the queue at any time, the exponentially distributed single interval primarily controls the queue behavior and the effect of the correlations would be insignificant. These are the intuitive reasons for the potential efficacy of a Poisson model for the superposition packet voice process in a multiplexer with bit dropping.

4. THE $M/\tilde{D}/1/K$ MODEL AND IT'S ANALYSIS

We now present an analysis of the packet voice multiplexer with bit dropping (see Fig. 1) assuming a Poisson approximation for the superposition arrival process and a deterministic but state-dependent service time (see Fig. 2 and Table 1 in Section 2). Our approach is based on a Markov chain characterization of the queue length at departure epochs. The steady-state queue length probabilities are derived and they are used to compute the performance measures of interest, such as mean and variance of queue length (buffer occupancy) and packet delay, probability of packet loss due to buffer overflow, mean bits per sample for voice, etc.

4.1 Computations of Queue Length Probabilities

Let L_j denote the number of packets in the system at departure epochs (just after a departure). Let A_j denote the number of packet arrivals during the jth service time. Then the following recursion holds for L_j:

$$L_{j+1} = \min\ \{K-1,\ ([L_j - 1]^+ + A_{j+1})\} \tag{3}$$

where K is the finite buffer capacity in packets and where $[x]^+ = x$ when $x > 0$ and $[x]^+ = 0$ when $x \leq 0$.

Let the packet service times be denoted by D_4, D_3 and D_2 corresponding to congestion states defined by $0 \le L_j \le Q_1$, $Q_1 < L_j \le Q_2$ and $Q_2 < L_j \le K-1$, respectively (see Section 2, Table 1 and Fig. 2). Here Q_1 and Q_2 are specified queue thresholds for bit dropping. The bit dropping algorithm operates at the output of the queue, i.e., the packet service rate is determined according to the state of the queue at the time a packet actually enters into service (or transmission on the link). Let the sequences $\{a_i,\ i \ge 0\}$,$\{b_i\ i{\ge}0\}$ and $\{c_i,\ i{\ge}0\}$ denote the probabilities of the number of arrivals during deterministic service times D_4, D_3 and D_2 respectively. The following equations hold for these sequences:

$$a_i = \frac{e^{-\lambda D_4}\,(\lambda D_4)^i}{i!}\ ,\ i{\ge}0$$

$$b_i = \frac{e^{-\lambda D_3}\,(\lambda D_3)^i}{i!}\ ,\ i{\ge}0$$

$$c_i = \frac{e^{-\lambda D_2}\,(\lambda D_2)^i}{i!}\ ,\ i{\ge}0 \qquad (4)$$

where $\lambda = N r$ denotes the average total packet arrival rate for N voice sources, while each source has an average rate r. Now the state transition probability matrix, P , for the Markov chain denoted by $\{L_j\ ;\ j \ge 1\}$ is given by:

$$\begin{array}{c|cccccccccccccccc|}
 & 0 & 1 & 2 & \cdots & Q_1{-}1 & Q_1 & Q_1{+}1 & \cdots & Q_2{-}1 & Q_2 & Q_2{+}1 & \cdots & K{-}2 & K{-}1 \\
\hline
0 & a_0 & a_1 & a_2 & & \cdot & \cdot & \cdot & & \cdot & \cdot & \cdot & & & 1{-}\sum \\
1 & a_0 & a_1 & a_2 & & \cdot & \cdot & \cdot & & \cdot & \cdot & \cdot & & & 1{-}\sum \\
2 & 0 & a_0 & a_1 & & \cdot & \cdot & \cdot & & \cdot & \cdot & \cdot & & & 1{-}\sum \\
\cdot & & & & & & & & & & & & & & \\
\cdot & & & & & & & & & & & & & & \\
\cdot & & & & & & & & & & & & & & \\
Q_1{-}1 & & & & a_0 & a_1 & a_2 & a_3 & & \cdot & \cdot & \cdot & & & \cdot \\
Q_1 & & & & & a_0 & a_1 & a_2 & & \cdot & \cdot & \cdot & & & \cdot \\
Q_1{+}1 & & & & & & b_0 & b_1 & & \cdot & \cdot & \cdot & & & \cdot \\
\cdot & & & & & & & & & & & & & & \\
\cdot & & & & & & & & & & & & & & \\
\cdot & & & & & & & & & & & & & & \\
Q_2{-}1 & & & & & & & & b_0 & b_1 & b_2 & b_3 & & & \cdot \\
Q_2 & & & & & & & & & b_0 & b_1 & b_2 & & & \cdot \\
Q_2{+}1 & & & & & & & & & & c_0 & c_1 & & & \cdot \\
\cdot & & & & & & & & & & & & & & \\
\cdot & & & & & & & & & & & & & & \\
\cdot & & & & & & & & & & & & & & \\
K{-}2 & & & & & & & & & & & & c_0 & c_1 & 1{-}\sum \\
K{-}1 & & & & & & & & & & & & & c_0 & 1{-}c_0
\end{array} \qquad (5)$$

where $\sum$ denotes the sum of the rest of the elements of the row.

The steady state queue length probability vector $\boldsymbol{\pi}$, whose elements are defined as $\pi_i = Prob.\ \{L_j = i\}$, $0 \le i \le K-1$, is obtained by solving:

$$\boldsymbol{\pi} P = \boldsymbol{\pi} \qquad (6)$$

4.2 Probability of Packet Loss

For calculating the packet loss probability, we follow an approach similar to the one given in Gross and Harris[30] for an M/G/1/K model. However, for the system under study the state dependent nature of service times must be taken into consideration. Let $\{p_n,\ 0{\le}n{\le}K-1\}$, $\{q'_n,\ 0{\le}n{\le}K\}$, and $\{q_n,\ 0{\le}n{\le}K-1\}$ denote system state probabilities at an arbitrary time, an arrival point (any), and an arrival point given that the packet joins queue, respectively. By matching the effective rates of arrival and departure, Gross and Harris have shown that

$$\lambda\ (1-q'_K) = (1-p_0)\mu \qquad (7)$$

where μ is the mean packet service rate and q'_K is the packet loss probability. This equation also holds for the $M/D/1/K$ system under study provided that μ is computed properly with bit dropping taken into consideration. Gross and Harris have derived the following equation for p_0 which also holds for our $M/D/1/K$ model:

$$p_0 = \frac{\pi_0}{\pi_0 + \rho} \qquad (8)$$

where $\rho = \lambda/\mu$. Combining (7) and (8), we get

$$q'_K = \frac{\pi_0 + \rho - 1}{\pi_0 + \rho} \qquad (9)$$

Now we turn to the computation of μ for our model. To do this, we focus on the departure epochs and look at the service time for the packet to be served next. If the system is empty at a departure epoch, then the system waits until the next packet arrives and serves it in time D_4. If the system is in any of the states 1 through $K-1$ at a departure epoch, the next packet to be served goes immediately into service and its service time will be D_4, D_3 or D_2 in accordance with the bit dropping algorithm of Table 1. Therefore, the mean time to service a packet is given by

$$\mu^{-1} = \pi_0 D_4 + D_4 \sum_{i=1}^{Q_1-1} \pi_i + D_3 \sum_{i=Q_1}^{Q_2-1} \pi_i + D_2 \sum_{i=Q_2}^{K-1} \pi_i \quad (10)$$

This completes our derivation of the packet loss probability.

4.3 Voice Quality (Mean Bits Per Sample)

Here we first derive the fractions, f_4, f_3 and f_2, of packets that receive service at 4 bit, 3 bit and 2 bit per sample rates, i.e. with service times D_4, D_3 and D_2, respectively. Recognizing that $\{\pi_i,\ 0 \le i \le K-1\}$ are departure state probabilities and following the same observations as we did in Section 4.2, it is seen that

$$f_4 = \sum_{i=0}^{Q_1-1} \pi_i$$

$$f_3 = \sum_{i=Q_1}^{Q_2-1} \pi_i$$

$$f_2 = \sum_{i=Q_2}^{K-1} \pi_i \quad (11)$$

Because the voice sources are independent and identical, the above fractions which refer to all packets transmitted on the link, are also applicable to each individual voice source. Hence, the mean bits per sample for the voice packets *transmitted on the link* on for each call is given by

$$\bar{b} = 4f_4 + 3f_3 + 2f_2 \quad (12)$$

To compute the mean bits per sample for each call over all its packets *received at the multiplexer* , we need to include a term to account for lost packets. This quantity is given by

$$\bar{b}' = (1-q'_K)\bar{b} \quad (13)$$

where q'_K is the packet loss probability that we computed in Section 4.2.

In general, the values of q'_K would be negligibly small up to fairly high values of loads so that $\bar{b}' \approx \bar{b}$. Therefore, in the remainder of this memorandum we use the term mean bits per sample of $\bar{b}$ as defined in (12), unless stated otherwise.

There is an intuitive way of computing the mean bits per sample values in overload without even using a queueing model. This can be done by observing that at high loads the transmission capacity would be saturated by the carried bits of information and the header in the transmitted packets. Hence, assuming that no packets are dropped due to buffer overflow, it is easy to see that the following hyperbolic equation holds for the mean bits per sample, $\bar{b}$, as a function the number, N, of active voice sources:

$$(S\bar{b} + H)\,N r = C \quad (14)$$

where S is the number of voice samples per packet, H is the number of bits in the header, r denotes the packet arrival rate per voice source, and C represents the transmission capacity. We observed that this approximation matched very well with the results of the $M/D/1/K$ model as well as the simulations for mean bits per sample values in overload.

4.4 Mean and Standard Deviation of Queue Length and Waiting Time

Because the arrivals form a Poisson process in our model, the state probabilities at departure epochs, $\{\pi_n,\ 0 \le n \le K-1\}$, are the same as those at arrival epochs, $\{q_n,\ 0 \le n \le K-1\}$, given that the arrival joins the queue. Hence, the mean, EQ, and standard deviation, σ_q, of queue length for transmitted packets is given by

$$EQ = \sum_{i=1}^{K-1} i\pi_i \quad (15)$$

$$\sigma_q = \sum_{i=1}^{K-1} i^2\pi_i - (EQ)^2 \quad (16)$$

The quantity EQ also represents the mean queue length at departure epochs. Hence, regardless of the state dependent nature of the service, Little's formula holds and so the mean waiting time EW is given by

$$EW = \frac{EQ}{\lambda(1-q'_K)} \quad (17)$$

The waiting time distribution or its variance is not easy to compute. This is because the waiting time of a packet is a function of the service times of the packets ahead of it, and in this system, these service times are a function of the number of packet arrivals that occur during the waiting time of the packet in consideration. (Recall that in a normal M/G/1 system, i.e. one without state dependent service, the service times of the packets in front are always independent.) So we need to use an approximation to compute the standard deviation of delay.

We will now describe an approximation for the standard deviation that we found to give fairly satisfactory results. Note that the mean packet service time is given by μ^{-1} as expressed in (10). Approximate the mean squared value of waiting time by assuming that the service time for each packet served is a fixed quantity μ^{-1}. Hence,

$$E(W^2) \approx \sum_{i=1}^{K-1} (i\mu^{-1})^2 \, \pi_i \tag{18}$$

$$\begin{aligned} var(W) &= E(W^2) - (EW)^2 \\ &\approx \mu^{-2}\left[\sum_{i=1}^{K-1} (i)^2 \, \pi_i - \left[\sum_{i=1}^{K-1} i\pi_i\right]^2\right] \\ &= \mu^{-2}\sigma_q^2 \end{aligned} \tag{19}$$

$$\sigma_w \approx \frac{\sigma_q}{\mu} \tag{20}$$

where σ_w denotes the standard deviation of waiting time. We tried other approximations including some bounds, but found (19) to be most suitable by means of comparing analytical and simulation results.

5. *NUMERICAL EXAMPLES AND DISCUSSION*

In this section, we demonstrate the validity of our $M/D/1/K$ model by comparing numerical results obtained by the analytic model with those obtained by an exact simulation model. In the simulation model, the packet generation process for each source is modeled exactly as shown in Fig. 3. The queueing and bit dropping mechanisms are also simulated exactly as described in Section 2. The simulation is written in FORTRAN and is run on a Cray-1. We let the simulation run for about 15 minutes of real-time operation of the multiplexer. Typically, one to two million voice packets are serviced in the multiplexer during a 15 minute period. The performance measures of interest are sampled at intervals of a minute starting at the 10 minute epoch. The resulting six samples are averaged. We repeated the 15 minute runs several times with different seed values in each run for all the random number generators, and found that the results were very close to each other. The confidence intervals are extremely narrow and hence not shown in our plots. The confidence intervals are extremely narrow is for two reasons: (i) we simulate a very large number (one to two million) packet arrivals in each simulation run, and (ii) the bit dropping mechanism significantly diminishes the burstiness in the packet voice arrival process.

We will now state the parameter values used in the simulation and analytical models. Two different cases of speech activity rates are considered in the numerical examples: 22 packets per second (pps) and 26.25 pps, corresponding to 35% and 42% activity factors, respectively. The mean speech talkspurt and silence lengths are assumed to be 352 ms and 650 ms for the first case, and 420 ms and 580 ms for the second case, respectively. The first case of speech activity parameter values above applies when nothing is stated explicitly about the same. When N voice calls are active, the λ for the $M/D/1/K$ model is given by $\lambda = \alpha N$ packets per second, where α denotes the average packet arrival rate per source. The packet sizes and their service times during various congestion states are as shown in Table 1, Section 2. The values of the bit dropping thresholds are assumed to be (13,26,52) packets or, equivalently, (5,10,20) milliseconds unless stated otherwise. The third value in this set is not a threshold for dropping bits, rather it is the finite buffer capacity for storing voice packets in the queue. Threshold values specified in milliseconds can be converted to packet counts by using the appropriate link transmission rate and the voice packet size. The link transmission rate is assumed to be 1.536 Mbps except when stated otherwise.

The mean bits per sample vs load curves obtained by the model and by simulation are compared in Fig. 7. Associated with the mean bits per sample values are the fractions of packets with 4,3 and 2 bits per sample, i.e., f_4, f_3 and f_2. The curves for these fractions obtained by the model and by simulations are compared in Fig. 8. The corresponding mean queue length and mean packet waiting time curves are shown in Figs. 9 and 10. We note that the model compares closely with exact simulations in all cases (i.e. Figs. 7-10). The Poisson model yields slightly higher delays at low and medium values of the load. This is because the single interval in the superposition process dominates the long term correlation effects at these loads. Note that just a few packet intervals interact in the queue at low or medium loads (see Fig. 9). For the positive correlations between successive packet intervals to produce an effect, a large number of packets must interact in the queue (see Fig. 5). However, the bit dropping mechanism essentially prevents this from happening even at high loads. The service speeds up during congestion, thereby maintaining a low queue occupancy. Thus bit dropping eliminates or significantly diminishes the effect of the burstiness due to correlations.

For a comparison between multiplexers with and without bit dropping, we focus on the mean delay vs. load curves of Figs. 6 and 10. At a link utilization value of about 95 percent (130 voice sources in Fig. 6), the mean delay for the case without bit dropping is about 30 milliseconds which is more than an order of magnitude higher than the corresponding mean delay for the case with bit dropping (120 voice sources in Fig. 10). (The header effect was included in computing Fig. 10 but not in Fig. 6. Hence, the 95 percent utilization in Fig. 10 occurs at 120 voice sources.) In Fig. 6, also note that for a multiplexer without bit dropping, the Poisson approximation under-estimates the delay by an order magnitude at a 95 percent utilization. For the case with bit dropping, the Poisson approximation for mean delay is very good at moderate loads and above (see Fig. 10). The slight over-estimation of the delay in Fig. 10 by the Poisson model in the range of light to moderate loads is not of much concern from a design point of view. This is because the voice build-out delay value in the receiving access interface would be normally selected on the basis of network delay performance in overload.

From Fig. 7 we observe that the multiplexer can support 120 and 132 voice calls respectively at mean bits per sample values of 3.8 and 3.5, respectively. To get a feel for the system capacity, assume that a mean bits per sample value of 3.7 and a packet loss fraction of 10^{-3} are set as objectives to meet voice quality requirements. Then from Fig. 7 we note that, for a speech activity factor of 0.35, the multiplexer can support 120 active voice calls on a T1 while providing good voice quality. The corresponding multiplexer capacity is about 100 voice calls when the speech activity factor is 0.42 (see Tables 2,4). The packet loss fraction value is generally negligible up to fairly high loads (see Table 4). Hence, the mean bits per sample value is generally the limiting factor in determining capacities. Clearly, the bit dropping mechanism provides an overload control mechanism wherein voice quality is gracefully degraded in overload. In a multiplexer without bit dropping, when the system goes into overload the voice quality is severely degraded due to high packet loss. It is well known that from voice quality considerations, that it is better to drop bits rather than whole packets[24] (for the same carried to received information ratio). Hence bit dropping in a voice packet multiplexer is a very useful feature to have because it enhances capacity and provides a good overload protection.

The variance of the queue length and the waiting time are shown in Figs. 11 and 12. The comparisons between the modeling and the simulation results look reasonably acceptable. The variance initially increases with load but then decreases due to the fact that bit dropping smooths the variations in the queue fill. The variances of the queue length as well as the waiting time have to decrease to zero as the load approaches infinity due to the finite buffer effect. Some wiggles in the two curves (Figs. 11 and 12) at intermediate load values are possibly due to some complex interplay between the bit dropping thresholds and the queue fill values. As the typical queue length value approaches the first threshold, traffic smoothing is affected due to bit dropping and the variance shows a dip. But then as the load increases, the variance can go up again slightly due to saturation of the bit dropping effects at the first threshold. Then again the variance could experience another dip when the typical queue length approaches the second threshold.

Tables 2,3,4 and Figs. 13,14 show the sensitivity of the mean bits per sample value to bit dropping thresholds. The mean delay for voice packets increases in proportion to the bit dropping thresholds due to the randomness of the arrival process. However, the mean bits per sample value exhibits a certain degree of insensitivity to the threshold values (see Figs. 13,14). An obvious question to ask is how far can we reduce the thresholds to achieve the lowest possible delay while maintaining acceptable voice quality. Table 2 and Figs. 13,14 seem to provide an answer to this question. For example, for a 384 Kbps link using (20,40,60) ms threshold values do not provide any significant improvement over (5,10,15) ms values for the quality of voice in mean bits per sample, but the delay performance for the latter case would be significantly superior. Tables 2,3,4 show that a sharp decrease in threshold values (say, from Case 4 to Case 1) is undesirable because of the severe degradation of mean bits per sample value and increased packet loss probability at a nominal load of 96 active voice trunks. Also an increase in the threshold values (e.g. going from Case 4 to Case 5) may result in a slight improvement in mean bits per sample or packet loss values but at the same time the mean packet delay goes up significantly at nominal load values (see Tables 2,3,4). Thus these tables suggest that it is possible to make a prudent choice of values for the bit dropping thresholds so that the delay performance is optimized while the bit dropping and packet loss performance objectives are met as per requirements.

6. CONCLUSIONS

In this memorandum, we studied the performance of a packet voice multiplexer in which the less significant bits in voice packets are dropped during states of congestion in the multiplexer. Using the results of simulation and analytical modeling, we illustrated that bit dropping on voice packets significantly smooths the superposition packet voice process by speeding up the packet service rate during critical periods of congestion in the queue. This phenomenon renders it possible to approximate the superposition process by a Poisson process for analyzing a packet voice multiplexer with bit dropping. The multiplexer was modeled as a queueing system with a state dependent server whose service speeds up during states of congestion. The specific model used was an $M/D/1/K$ model in which D denotes the deterministic but state dependent nature of service. The model enabled us to compute various quantities of interest for performance characterization, e.g., the carried mean bit rate for a voice call, the queue length distribution, the packet loss due to queue overflows, the mean and variance of voice packet delay, etc. By comparison with simulations, this model was shown to produce quite accurate performance predictions. The results showed that significant capacity and performance advantages are gained in the multiplexer as a result of the bit dropping scheme.

7. ACKNOWLEDGEMENTS

The authors would like to thank W. Whitt for some very insightful discussions on burstiness of superposition processes, A. E. Eckberg and J. M. Holtzman for many helpful comments and A. Anastasio for assistance in developing the computer simulation experiments.

REFERENCES

1. J. S. Turner and L. F. Wyatt, "A packet network architecture for integrated services," in *Proc. IEEE GLOBECOM 83*, San Diego, CA, Nov. 1983, pp. 2.1.1-6.
2. W. Hoberecht, "A layered network protocol for packet voice and data integration," *IEEE J. Selected Areas in Commun.*, vol. SAC-1, no. 6, pp. 1006-1013, Dec. 1983.
3. J. J. Kulzer and W. A. Montgomery, "Statistical switching architectures for future services," *Proc. ISS '84*, Florence, Italy, May 1984, pp. 43A.1.1-6.
4. J. S. Turner, "Design of an integrated services packet network," presented at the Ninth Data Commun. Symp., Whistler Mt. B. C., Canada, Sept. 1985.
5. R. W. Muise, T. J. Schonfeld and G. H. Zimmerman, "Experiments in wideband packet technology," *Proc. Zurich Seminar on Digital Communication*, March 1986, pp. D4.1-D4.5.
6. C. M. Corbalis, "A design example of a T1-based fast packet voice switch," *Proc. of the IEEE International Conference on Commun.*, Seattle, June 1987, pp. 36-3.1-36.3.5.
7. S. R. Amstutz, "Burst switching—A method for distributed and integrated voice and data switching," *IEEE Commun. Mag.*, pp. 36-42, Nov. 1983.
8. K. Sriram, P. K. Varshney, and J. G. Shanthikumar, "Discrete-time analysis of integrated voice-data multiplexers with and without speech activity detectors," *IEEE Journal on Selected Areas in Commun.*, vol. SAC-1, no. 6, pp. 1124-1132, Dec. 1983.
9. P. O.'Reilly, "A fluid-flow approach to performance analysis of integrated voice-data systems with DSI," in *Proc. Int. Conf. Modeling and Tools for Performance Anal.*, Sophia-Antipolis, France, June 1985, pp. 121-136.
10. P. M. Gopal, J. W. Wong, and J. C. Majithia, "Analysis of Playout Strategies for Voice transmission using packet switching techniques," *Performance-Evaluation* 4 (1984), (North-Holland), pp. 11-18.
11. K. Sriram and W. Whitt, "Characterizing superposition arrival processes in packet multiplexers for voice and data," *IEEE Journal on Selected Areas in Commun.*, vol. SAC-4, no. 6, September 1986, pp. 833-846.
12. H. Heffes and D. M. Lucantoni, "A Markov-modulated characterization of packetized voice and data traffic and related statistical multiplexer performance," *IEEE Journal on Selected Areas in Commun.*, vol. SAC-4, no. 6, September 1986, pp. 856-868.
13. J. N. Daigle and J. D. Langford, "Models for analysis of packet voice communication system," *IEEE Journal on Selected Areas in Commun.*, vol. SAC-4, no. 6, September 1986, pp. 847-855.
14. Y. C. Jenq, "Approximations for packetized voice traffic in statistical multiplexer," *Proc. IEEE INFOCOM* Apr. 1984, pp. 256-259.
15. T. E. Stern, "A queueing analysis of packet voice," in *Proc. IEEE Global Telecomm. Conf.*., San Diego, CA, Dec. 1983, pp. 2.5.1-2.5.6.
16. M. L. Luhanga, "Analytical model of a packet voice concentrator," Columbia University, New York, NY, Tech. Rep. No. 1984-02.
17. S. Ganguly, "Analysis and performance evaluation of multiplexed packet voice and data," Ph.D. dissertation, Dep. Elec. Eng., Columbia University, New York, NY, June 1985.
18. T. Bially, B. Gold and S. Seneff, " A technique for adaptive voice flow control in integrated packet networks," *IEEE Trans. Commun.*, Vol. COM-28, March 1980, pp. 325-333.
19. T. Bially, A. J. McLaughlin and C. J. Weinstein, "Voice communication in integrated digital voice and data networks," *IEEE Trans. Commun.*, Vol. COM-28, September 1980, pp. 1478-1490.
20. R. V. Cox and R. E. Crochiere, "Multiple user variable rate coding for TASI and packet transmission systems," *IEEE Trans. Commun.*, Vol. COM-28, March 1980, pp. 334-337.
21. D. J. Goodman, "Embedded DPCM for variable bit rate transmission," *IEEE Tran. on Commun.*, COM-28, No. 7, July 1980.
22. Y. Yatsuzuka, "High-Gain Digital Speech Interpolation with ADPCM Encoding," *IEEE Tran. on Commun.*, COM-30, No. 7, April 1982, pp. 750-761.
23. J. M. Holtzman, "The interaction between queueing and voice quality in variable bit rate packet voice systems," *Proc. ITC 11*, Kyoto, Japan, Sept. 4-11, 1985, Paper 2.2A-4.
24. C. Dvorak and T. J. Schonfeld, "Wideband Packet Technology Performance," *Proc. of AT&T Symposium on Speech Processing Applications and Systems*, Session 4, October 1985.
25. D. O. Bowker and C. A. Dvorak, "Speech Transmission Quality of Wideband Packet Technology," *Proc. of IEEE GLOBECOM*, Tokyo, Japan, November 1987.
26. M. Listani and F. Villani, "Voice communication handling in X.25 packet switching networks," *Proc. IEEE GLOBECOM*, San Diego, Nov. 1983, Paper 2.4.
27. D. Grillo, F. Villani, M. Calabrese and R. Pietrojusti, "Impact of low and high bit rate coding in an integrated packet network," *Proc. International Switching Symposium*, Florence, Italy, May 1984, Paper 42B1.
28. D. Grillo, "Interactive voice information handling in wide area packet switched networks," *Proc. ITC 11*, Kyoto, Japan, Sept. 1985, Paper 5.4A-4.
29. C. E. May and T. J. Zebo, unpublished paper, AT&T Bell Laboratories, December 1981.
30. D. Gross and C. M. Harris, *Fundamentals of Queueing Theory* , 2nd ed., John Wiley & Sons, 1985, pp.279-285.

TABLE 1. DESCRIPTION OF THE BIT DROPPING ALGORITHM (L DENOTES THE CURRENT QUEUE FILL IN PACKETS).

CONGESTION STATE	CONGESTION CONTROL ACTION	TRANSMITTED VOICE PACKET	
		SIZE (BYTES)	SERVICE TIME (MS @ T1 Rate)
$0 \leq L \leq Q_1$	DROP NOTHING	74	.385
$Q_1 < L \leq Q_2$	DROP BLOCK #1	58	.302
$Q_2 < L \leq K-1$	DROP BLOCKS #1 & 2	42	.219

TABLE 2. SENSITIVITY TO BIT DROPPING THRESHOLDS: MEAN BITS PER SAMPLE

- SPEECH ACTIVITY RATE ≈ 42% (26.25 PPS PER SOURCE)
- THRESHOLDS:

Case 1: (3,5,8) packets ≈ (1,2,3) ms
Case 2: (5,10,15) packets ≈ (2,4,6) ms
Case 3: (8,16,24) packets ≈ (3,6,9) ms
Case 4: (13,26,52) packets ≈ (5,10,20) ms
Case 5: (20,40,90) packets ≈ (8,16,34) ms

# VOICE SOURCES	MEAN BITS PER SAMPLE				
	Case 1	Case 2	Case 3	Case 4	Case 5
60	3.94	4.00	4.00	4.00	4.00
72	3.87	3.96	3.99	3.99	4.00
84	3.74	3.89	3.96	3.99	4.00
96	3.57	3.73	3.83	3.90	3.95
108	3.34	3.47	3.54	3.59	3.61
120	3.10	3.15	3.18	3.19	3.19
132	2.87	2.86	2.85	2.85	2.85
144	2.67	2.60	2.57	2.56	2.56
156	2.51	2.41	2.36	2.31	2.31
168	2.42	2.27	2.20	2.12	2.10
180	2.29	2.17	2.11	2.02	1.94

TABLE 3. SENSITIVITY TO BIT DROPPING THRESHOLDS: MEAN PACKET DELAY (PARAMETERS SAME AS IN TABLE 2)

# VOICE SOURCES	MEAN VOICE PACKET DELAY (milliseconds)				
	Case 1	Case 2	Case 3	Case 4	Case 5
60	.65	.66	.68	.68	.68
72	.77	.85	.87	.90	.90
84	.91	1.12	1.29	1.41	1.45
96	1.06	1.49	1.97	2.69	3.55
108	1.21	1.90	2.80	4.32	6.60
120	1.31	2.27	3.49	5.50	8.17
132	1.38	2.52	3.98	6.44	10.01
144	1.41	2.67	4.23	6.86	10.53
156	1.41	2.70	4.33	7.09	10.52
168	1.39	2.70	4.34	7.81	11.65
180	1.36	2.65	4.30	9.02	16.26

TABLE 4. SENSITIVITY TO BIT DROPPING THRESHOLDS: PACKET LOSS (PARAMETERS SAME AS IN TABLE 2)

# VOICE SOURCES	VOICE PACKET LOSS FRACTION				
	Case 1	Case 2	Case 3	Case 4	Case 5
60	7.7×10^{-6}	0	0	0	0
72	6.5×10^{-5}	0	0	0	0
84	3.7×10^{-4}	1.2×10^{-6}	0	0	0
96	1.5×10^{-3}	2.3×10^{-5}	1.12×10^{-7}	0	0
108	4.8×10^{-3}	2.4×10^{-4}	5.4×10^{-6}	0	0
120	.012	1.4×10^{-3}	1.0×10^{-4}	0	0
132	.024	5.5×10^{-3}	9.0×10^{-4}	0	0
144	.043	.015	4.5×10^{-3}	5.16×10^{-6}	0
156	.066	.033	.015	2.5×10^{-4}	1.26×10^{-6}
168	.095	.058	.036	5.4×10^{-3}	7.95×10^{-4}
180	.13	.091	.069	3.7×10^{-2}	3.18×10^{-2}

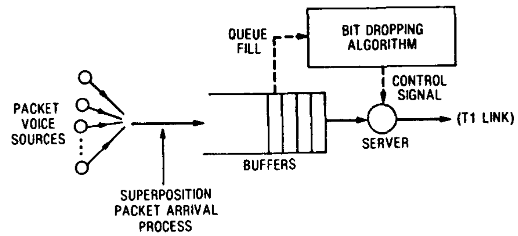

FIGURE 1. PACKET VOICE MULTIPLEXER WITH BIT DROPPING

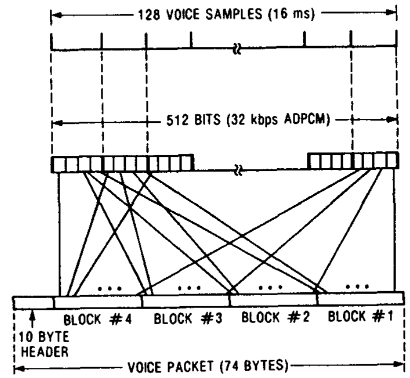

FIGURE 2. ORGANIZATION OF AN EMBEDDED-ADPCM-ENCODED VOICE PACKET

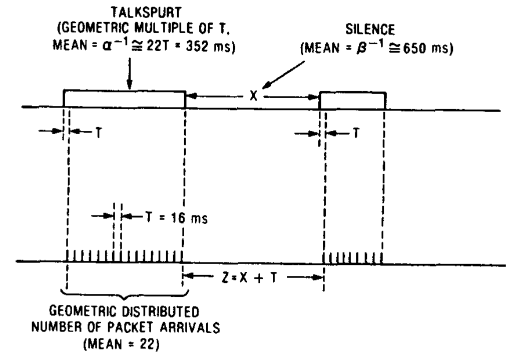

FIGURE 3. PACKET ARRIVAL PROCESS DUE TO ONE VOICE SOURCE

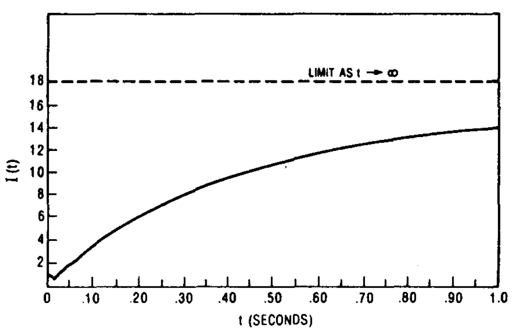

FIGURE 4. INDEX OF DISPERSION FOR NUMBER OF ARRIVALS IN (0,t) FOR VOICE PACKET ARRIVAL PROCESS

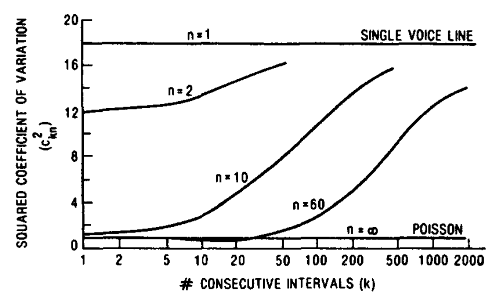

FIGURE 5. k-INTERVAL SQUARED COEFFICIENT OF VARIATION CURVES

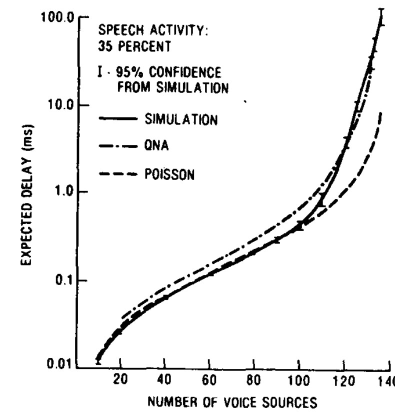

FIGURE 6. EXPECTED DELAY FOR PACKET VOICE MULTIPLEXER

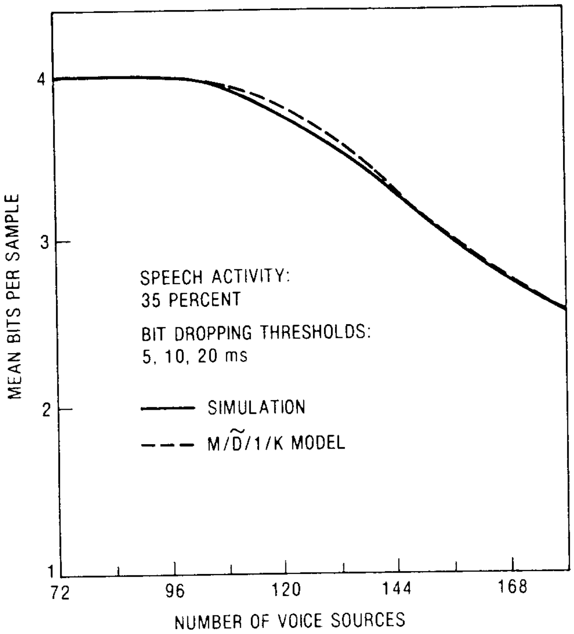

FIGURE 7. MEAN BITS PER SAMPLE AS A FUNCTION OF LOAD

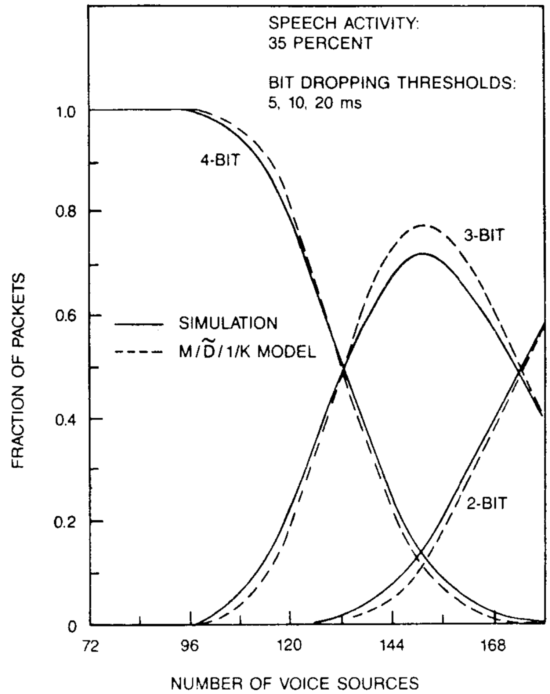

FIGURE 8. FRACTIONS OF PACKETS WITH 4,3,2 BITS PER SAMPLE

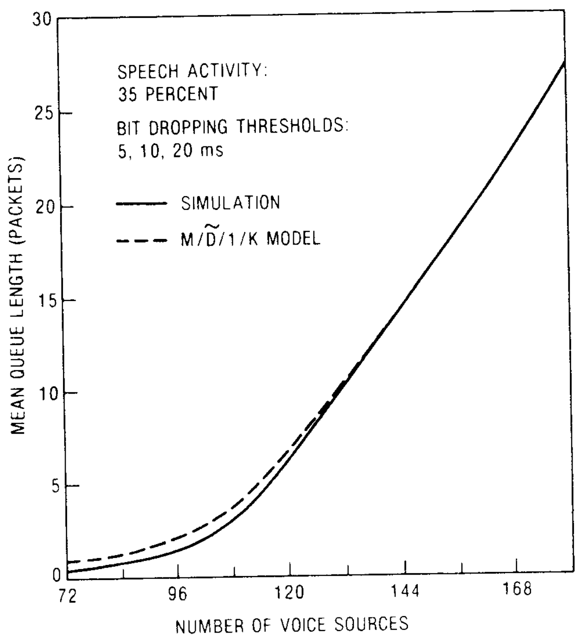

FIGURE 9. MEAN QUEUE LENGTH AS A FUNCTION OF LOAD

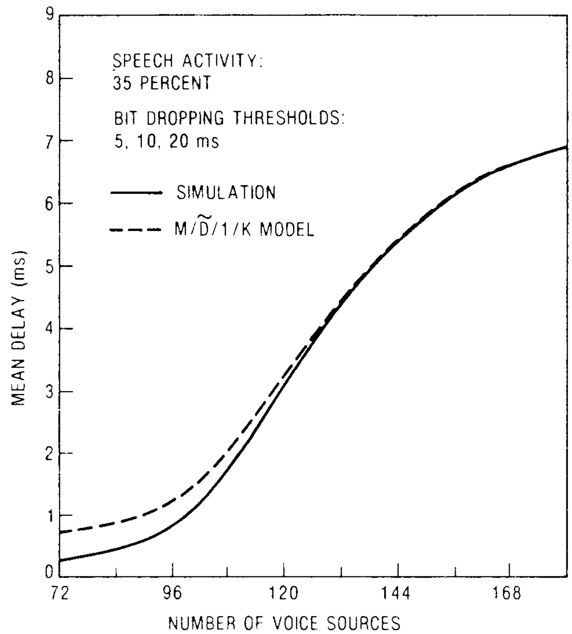

FIGURE 10. MEAN VOICE PACKET DELAY AS A FUNCTION OF LOAD

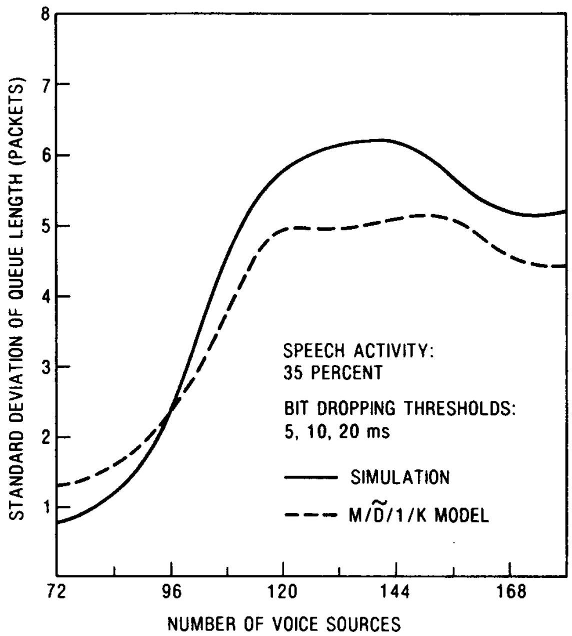

FIGURE 11. STANDARD DEVIATION OF QUEUE LENGTH AS A FUNCTION OF LOAD

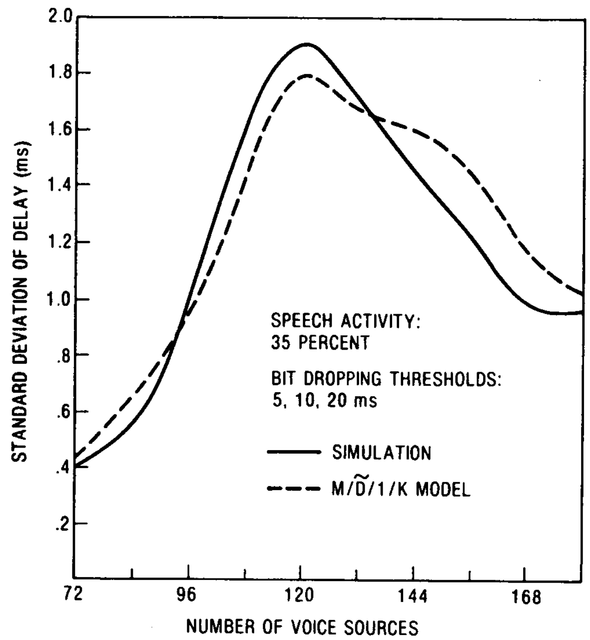

FIGURE 12. STANDARD DEVIATION OF PACKET DELAY AS A FUNCTION OF LOAD

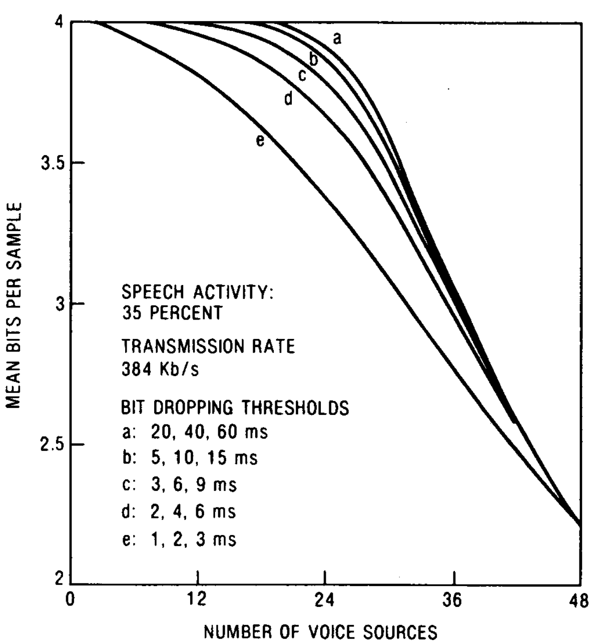

FIGURE 13. SENSITIVITY OF MEAN BITS PER SAMPLE TO BIT DROPPING THRESHOLDS AT 384 Kbps TRANSMISSION RATE

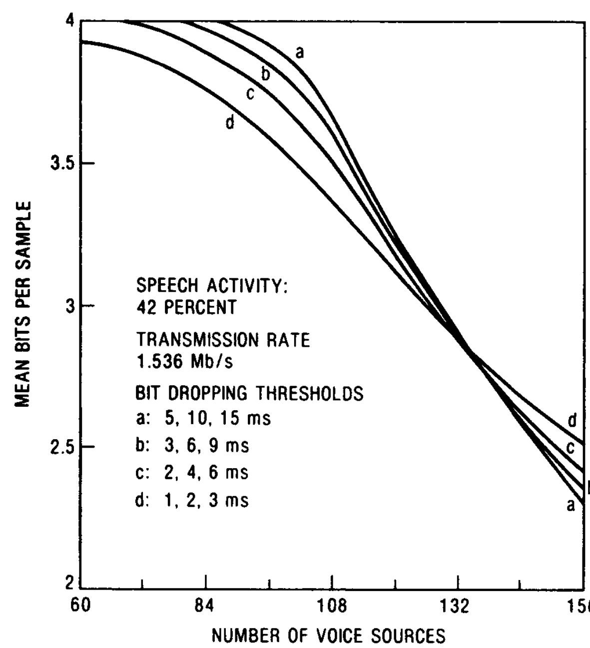

FIGURE 14. SENSITIVITY OF MEAN BITS PER SAMPLE TO BIT DROPPING THRESHOLDS AT 1536 Kbps TRANSMISSION RATE