The performance of a packet-switched wide-area network for integrated data and voice service is investigated under the assumption of a traffic mix resulting in network resources mainly loaded with data traffic. The analysis focuses on two issues impacting the transport delay, namely probability of allocating network resources to one traffic type in contention cases and reduction of voice redundancy depending on network congestion level. The algorithm used for reducing voice redundancy bases on embedded ADPCM coding and a particular bit assembling within voice packets.

1. INTRODUCTION

As a first step toward services integration in packet-switched networks, voice and data integration is envisaged, (1), (2), (3), (4), (5). In this paper the problem of handling both voice and data is addressed by referring to CCITT X.25 protocol for data application and by adopting a voice-oriented, X-series compatible and simplified protocol for voice, described in (6) and (7), thus exploiting the advantages deriving from protocol commonality.

Among other issues needed to be considered in order to adapt the X.25 protocol for voice handling, flow control plays a significant role. Due to the high interaction between partners in a voice session and the stringent requirements on transfer time, excessive delay in information delivery is to be faced by speeding up voice transport. This can be achieved, i.a. by making use of priorities in allocating network resources, by reducing voice load at the expense of fidelity, or by a combination of the two.

In (8) various schemes for adapting the voice sources encoding rate to the network congestion level, as reported by the sinks, are examined in a voice only environment.

In this paper, which expands under several aspects a preliminary analysis reported in (9), both random allocation of network resources according to different probabilities and load adaptation are considered. While allocation of resources is ruled by servicing dedicated voice and data queues, load adaptation is pursued by keeping the encoding rate at the source fixed and dropping off voice bits as packets travel through the network, depending on actual congestion level. Such an action aims at meeting delay constraints. Prerequisite for its adoption is that it does not corrupt too much the affected speech segments, i.e. does not make objectionable the speech quality at the relevant application level. Bit dropping is operated through an algorithm activated locally to the switching nodes in the long distance network. The reduction is accomplished by adopting embedded ADPCM coding, (10), conveniently assembling single bits of voice samples inside each packet and intervening on less significant of them.

In the analysis emphasis is given to the bursty nature of voice packet arrival process. Although various investigations address characterization of arrival statistics for packet voice, they mainly deal with voice streams flowing through a single network facility, (11), (12), (13). Due to the complexity of the analysis that considers routing through tandem links and 99.4th percentile end-to-end delays, a simulation approach has been adopted.

The paper is organized as follows: Section 2 presents the voice handling protocol and the load reduction algorithm; Section 3 describes the network model used in the analysis; Section 4 presents results referred to various cases; finally, Section 5 reports on the major findings.

2. VOICE HANDLING ISSUES

2.1 X.25-compatible Protocol Features

As for the voice handling protocol, reference will be made to the mentioned simplified, X.25-compatible version, (6), (7). Accordingly, error control is missing at both layer 2 and 3, and flow control is replaced by dropping off less significant bits comprised in the information field as a means of affecting the packet transport delay. In particular, layer 3 functions are limited to packet addressing and routing according to the virtual circuit technique, and possibly to bit dropping. X.25 layer 2 procedures, besides those performing zero stuffing and destuffing and error detection, are inhibited. Current X.25 flag patterns are retained to ensure bit transparency and frame synchronization. In addition, new bit configurations in the address field are used to discriminate between voice and data frames. Information frames do not need to support send and receive sequencing, and error detection is the only action to be taken at layer 2.

2.2 Signal Encoding and Packetizing

Packetizing delay corresponding to 16 ms speech and 32 kb/s ADPCM coding, (10), have been assumed. Using of 4 bits for expressing the difference between adjacent samples, a voice packet consists of 128 samples for a total of 512 bits.

Since according to embedded ADPCM coding, each subsequent bit used for representing the difference between two samples adds resolution to the representation, this technique lends itself for voice load reduction and graceful degradation of speech quality. In order to ease possible dropping of less significant bits, a packet is structured into a certain number of fields, each containing bits of equal significance related to different samples. Fig. 1 shows the structure of a voice packet as assumed in the analysis. The upper part of the figure represents the sequence of 128 voice samples before assembling. The lower part shows how the single bits are moved into the various fields, 16, into which a packet is assumed to be partitioned. In particular, the bits in the first position in each sample are ordered in sequence in the first four fields (each field consisting of 32 bits), the bits in the second position in each sample are ordered in the subsequent four fields, and so on, so that the significance of fields decreases from one packet extreme to the other.

Dropping a certain number of fields, starting from the extreme corresponding to less significant of them, amounts to reducing the redundancy of the affected samples in the same way as if an encoding rate lower than the original one were used.

2.3 Signal Reduction Algorithm

In each node, before transmitting a packet, the state of the relevant output queue is inspected. The queue state is expressed by the number of bits associated with the information fields of the voice packets waiting for transmission, i.e.
the bits belonging to the packet in the process of being transmitted are not counted. If this number exceeds a stated threshold, the packet to be transmitted is not allowed to contain more than a specified number of fields. Should the actual number of fields be greater than allowed, then so many fields are dropped (and lost) as necessary to meet the limitation. Packets may be reduced up to a specified extent, i.e. once the maximum reduction has been reached, subsequent reduction is inhibited. A generalization of the algorithm consists in setting several thresholds for describing the state of the queue, with associated number of fields allowed for transmission. In this analysis, threshold values are chosen as integer multiples of a basic threshold, the parameter of the algorithm. A packet may be reduced up to half the original size and allowed fields linearly decrease by one unit passing from one threshold to the upper next threshold.

3. THE MODEL

Elements of the queueing network model, represented in Fig. 2, are switching nodes and transmission channels. Nodes are provided with memories of three types: buffer storage for incoming frames waiting to be processed, memories for storing data packets temporarily inhibited to transmission due to window flow control, memories for storing packets waiting for transmission facilities to become available. Full-duplex transmission mode is assumed and simultaneous transmission activity on both ways is modelled through distinct and independent servers, one for each direction. Voice and data frame contending for network resources, both switching and transmission, join separate queues. Inside each queue the servicing algorithm is FCFS, whereas for the inter-queue servicing algorithm random allocation of resources according to specified probabilities is considered. A data frame enters the network through the input data queue in the origin node and, after being locally processed, it either joins the data queue associated with the transmission channel, or it is held in the node, according to whether the window flow control allows for transmission or not. Upon reaching the downstream node, a data frame joins the input data queue and, after processing, an acknowledgement for the upstream node is issued. Piggybacked acknowledgements on information frames are considered, as well as use of a same frame for collectively acknowledging multiple data frames. A voice frame enters the network through the voice input queue and after obtaining service it is routed to the voice output queue associated with the transmission channel leading to the subsequent node. Depending on the state of the voice workload waiting for service, a reduction in the information field of the packet in the process of being transmitted may be operated.

3.1 Data Flow Control

Admission into the network of single data packets is controlled by monitoring the packet population, both voice and data, within the long distance network at the instant of packet arrival. Data packets are denied access when the population exceeds a specified threshold. Backpressure is exercised of halted data packets on data sources and the process of data packet arrival is interrupted; the process is resumed when data packets are allowed again into the network. In case of backpressure, check for admission population level is performed at randomly and exponentially distributed intervals, long on the average half the data packet interarrival time. The rationale for this admission control is that voice applications should not experience delay which hampers natural speech interaction, and that data service may be—to a certain extent—sacrificed to meeting voice service requirements. Link-by-link window flow control within the long distance network is performed at layer 2. Use of extended numbering is assumed and the considered window size is 128. Dedicated acknowledgement packets of 80 bits are employed, when data frames to be used for piggybacking are not timely available.

3.2 Environment

Ample buffer space and an error-free environment are assumed, thus excluding recovery actions due to out-of-sequence or information contents corruption, respectively.

3.3 Switching Nodes and Transmission Channels

Processing of a frame in a node, independently of the type, takes a constant time, $125$ ms, i.e. nodes can handle up to 8000 packets per second. Full-duplex transmission facilities in the long distance network operate at 2.048 Mb/s. Transmission time is proportional to (fixed) frame length.

3.4 Traffic Issues

Bidirectional, balanced exchange of information is assumed between partners involved in a voice call or data session. Traffic mix resulting in 76% data and 30% voice workload on transmission channels, unless otherwise specified, has been taken as reference. Voice signal statistics in (14) are used to model voice traffic. Due to the different dynamics related to the arrival of calls and packets, the number of calls simultaneously in progress has been kept fixed. Voice calls are individually modelled as alternating "activity" and "silence" pha
ses. Activity phases (talkspurts) correspond to the emission of a burst of frames, whose number depends on talkspurt duration, exponentially distributed, and voice encoding rate. The total length of a voice frame is 592 bits.

As modelling of data traffic is concerned, data frames are individually spaced and data sessions are collectively considered. The length of information and acknowledgement data frames is 1096 and 80 bits, respectively. Voice frames belonging to the same talkspurt are spaced by the packetizing time, whereas an exponential distribution is assumed for the time between submission to the network of a data frame, or last voice packet in a burst, and start packetizing of the next frame.

3.5. Performance Metrics and Targets

The reference end-to-end connection is depicted in Fig. 3.

While the long distance network is modelled as described above, the access network is modelled by considering that each access side typically comprises a low speed line, 64 Kb/s, connecting the user with a packet handler, and a high speed line, 2.048 Mb/s, from the packet handler to the long distance network. The influence of the access network is accounted for by assuming that high and low speed lines behave as M/M/1 queues. Typical utilization for low and high speed lines ranges between .3 to .5 and .7 to .8 respectively. By combining extreme utilization values, two options for the access network are referred to in what follows, namely "low utilization" i.e. .3 and .7 and "high utilization", i.e. .5 and .8. The design values related to these options are to be intended in case no backpressure on data is exercised and as related to .8 and .9 design utilization of the channels in the long distance network respectively. Different utilization of the channels reflects into different utilization of the access network segments.

Delay in the long distance network is defined as the time elapsed from joining the input queue in the origin node until leaving the destination node. In the case of data traffic, this delay also includes the time spent at the network entry point waiting for admission into the origin node queue.

The end-to-end delay is a variable component (time spent for accessing transmission and processing facilities) and constant components (packetizing, transmission and switching time). The constant components are assumed to contribute 44 ms to the end-to-end delay, resulting from the following breakdown: 32 ms for encoding and decoding, 10 ms due to propagation, 2 ms for processing in the packet network interfaces (1 ms each). As for the variable components, (theoretical) delay distributions in the access network are combined by convolution with the (experimental) distributions for the long distance network.

A value of 250 ms for the 99.9th percentile voice end-to-end delay has been taken as service target, (15).

4. RESULTS

The analysis relates to the reference connection of Fig. 3, i.e. the long distance network consists of three nodes connected through two transmission channels. Data traffic is only exchanged between directly connected nodes, whereas voice traffic is exchanged between any node pair and hence it is also routed through one transit node. Symmetry for the offered load, both voice and data, submitted from any node is assumed and destination nodes are addressed equitably. The packet population limiting the admission into the
long distance network is equal to 1000. Fig. 4 and Fig. 5 show how servicing voice traffic with probability .65, i.e. twice more frequently than data traffic, respectively impacts the average and the 99.9th percentile delays. Delays in the figure are shown against the channel utilization as it results by increasing both voice and data offered load while keeping constant their proportion in the traffic mix. Data load increases by increasing the packet arrival rate, whereas voice load increases by increasing the number of simultaneous voice sessions. On the horizontal axes mapping between channel utilization and number of voice sessions is also shown. Since the length of voice packets is practically half the length of data packets, this could be viewed as a "fair" policy for allocating network resources. As results demonstrate, the two traffic types actually obtain quite different service. In Fig. 4 the voice average delay in the long distance network increases very slowly by increasing channel utilization, whereas the data delay increases steeply. In the same figure the data delay corresponding to the absence of voice load is also depicted. The difference between the data delays in the two cases increases very rapidly and, for 95% utilization, the ratio between them is almost one order of magnitude in favour of the case in which only data are handled. This phenomenon is due to a great extent to the bursty nature of voice traffic. As a demonstration, by replacing the voice packet arrival process with a Poisson process, the difference reduces considerably, as it is evidenced in the same figure. This phenomenon is due to a great extent to the bursty nature of voice traffic. As a demonstration, by replacing the voice packet arrival process with a Poisson process the difference reduces considerably, as it is evidenced in the same figure. This observation is in line with the finding in (13), according to which at high resource utilization a Poisson process does not satisfactorily model the compound arrival process of multiple voice sources. No reduction effect is detectable for voice traffic whose delay remains practically unaffected, as a demonstration that voice service protection is satisfactorily robust. Fig. 5 evidences the combined effect of delays in the long distance and in the access network on the 99.9th percentile end-to-end delay. The voice end-to-end delay is only very marginally affected by the delay experienced in the long distance network under both low and high utilization options for the access network, and only slightly increases with the load. Anyway, for the high utilization option, voice service requirements are no more met above 90% channel utilization. On the contrary, the 99.9th percentile data end-to-end delay increases steeply above 90% utilization, due to the dramatic increase of the related delay in the long distance network. In the same figure it is also shown both the 99.9th percentile long distance network and the 99.9th percentile end-to-end delays, for the case in which only data are handled. As a general observation, the delay contributed by the access network conditions the 99.9th percentile end-to-end delay up to channel utilization in the long distance network between 85% and 90%. Beyond that limit the behaviour of the long distance network clearly appears dominant. As the trend exhibited by the voice delay shows, there is practically no room left for improvement of voice handling and hence redundancy reduction would not pay for voice. In conclusion, voice experiences very low delay and service is well within requirements, except for a slight deviation in the case of high channel utilization in the long distance network combined with the high utilization option for the access network. However this result is obtained at the expense of data traffic which is severely penalized. This situation requires that data traffic be favoured in accessing network resources, for example by adopting less unbalanced servicing probabilities for data and voice. Figure 6 to 9 refer to a case in which the two traffic types are serviced with the same probability, i.e. .5. As for the average delay in the long distance network, Fig. 6, when no voice reduction is operated, results indicate that the difference between voice and data delay reduces to a great extent, with voice delay experiencing a drastic increase relative to the former case, although still below the data delay. Data are also handled somewhat worse for utilization up to 90% and benefit of some improvement above that value. The explanation is that parithetically handling voice and data is detrimental to voice because traffic peaks cannot be handled timely enough with the incurred delay increase.
On the other hand, the share of voice traffic is apparently too small to let the data traffic profit of equitable access to network resources, at least for moderate utilization.

In this utilization range, data traffic even suffers from voice not leaving the long distance network more quickly. At higher utilization, the load intensity of the two traffic types overrides any other influence and equitable access operates in the intended direction for data. A similar improvement for data traffic is also exhibited by the 99.9th percentile end-to-end delay, Fig. 7. In the figure the present case is contrasted with the former one for comparison's sake.

Poor service for voice may be faced in this case by voice redundancy reduction. In Fig. 6 the effect of reducing redundancy as a function of the reduction threshold is shown. As it is apparent, the objective of impacting the voice delay is reached and - according to expectations - the reduction is more pronounced as the utilization increases and the reduction threshold decreases. For the lower bit reduction threshold (2000), the average delay sensibly narrows the delay shown in the former case.

An indication of the voice redundancy reduction paid for obtaining this result is given in Fig. 8, where the percentage reduction of voice packet depending on reduction threshold is plotted against the channel utilization. In the figure distinction is also made between one-hop and two-hop routed packets. In the worst case, i.e. 2000 bit reduction threshold and two-hop routing, voice packets reduction does not exceed 8%, except for limiting utilization values.

Fig. 8 Percentage voice redundancy reduction for .5 voice servicing probability.

Stated differently, this means that in the worst case about 40 bits per voice packet are lost on the average, which maps into somewhat more than one field containing the less significant bits. It is worthy of noting that, by increasing utilization, two-hop routed packets proportionally lose less redundancy than one-hop routed packets. This suggests that reduction operates almost independently on the first and the second hop for two-hop routed packets at moderate utilization, whereas reduction in the first hop conditions reduction in the second hop at higher utilization. As a consequence, it may be induced that the effectiveness of the algorithm progressively reduces with the number of hops.

Fig. 9 shows the effect of voice reduction on the 99.9th percentile delay both in the long distance network and end-to-end. The delay in the long distance network exhibits a trend very similar to that of the average delay. Quite significant appears to be the effect of on the end-to-end delay which drops to values meeting the specified service target up to 95% utilization under the high utilization option for the access network and 2000 bit reduction threshold. Considering the relatively small voice redundancy reduction paid for the improvement obtained for both data and voice traffic, the reduction algorithm seems to perform efficiently.

Since the positive effects on delays are obtained by combining service probabilities and reduction of redundancy, one could infer that further gain in the end-to-end delay could be achieved by favouring data access to network resources beyond the parithetical situation.

Fig. 10 shows the effect of confining to .3 the probability of service for voice traffic, a proportion which equals the voice share in the traffic mix. As results demonstrate, this setting has disastrous consequences on both voice and data traffic. Two striking effects are evident. Firstly, voice traffic delays increase to such an extent that they overtake data delays, with the latter also increasing considerably with respect to the preceding case. Secondly, due to the vertical increase in the voice delays, the channel utilization is limited to 90%. The results shown relate to 20000 bit reduction threshold. The situation does not improve significantly for lower thresholds.

From this case it can be deduced that bursty voice traffic and unbalance in the voice and data packet sizes prevent data traffic from being strictly favoured in accessing network resources.

Besides delays, another issue is essential in evaluating the effectiveness of combined service priority and voice reduction policy, namely data throughput. Fig. 11 shows the data throughput against the channel utilization for the last two considered cases.

5.4A4-5
Throughput for .3 voice servicing probability linearly increases with utilization, except at almost limiting utilization values, with small impact from backpressure and negligible impact from voice reduction. Quite different is the behaviour for .3 voice servicing probability. In this case, throughput follows linearly with the utilization up to 90% and then drops steeply, due to backpressure. Consequently, strictly favouring data traffic reveals detrimental also with reference to data throughput. In the same figure it is also shown that deviation from ideal throughput characteristics due to bandwidth consumed by acknowledgement traffic is of little entity, independently of the case considered.

5. CONCLUSIONS

A manifolds of factors impact voice and data service in the considered model. The conclusions drawn from the analysis focused on probability of allocating network resources to one traffic in contention cases and reduction of voice redundancy, can be summarized as follows:

- carrying a mixture of voice and data in a packet switched network employing an X.25-derived protocol for voice, requires that delays for the two traffic types be traded off. Voice delay requirements are met if resources are allocated to one traffic according to a probability that is proportional to the packet length for the other traffic. Anyway, data need to be protected against bandwidth capture of voice, due to the bursty nature of related load submission process;

- protection of data traffic can be achieved by increasing the related servicing probability. Positive effects on data traffic obtained by equitably allocating network resources are balanced by service worsening for voice. To face this situation, reduction of voice redundancy through the analyzed algorithm appears effective in meeting specified voice service targets at the expense of a modest bit loss;

- pronounced unbalance in the service probability favouring data traffic are detrimental to both traffic types and reduce bandwidth utilization;

- the effectiveness of the algorithm is due to high responsiveness in levelling traffic peaks, made possible through continuous monitoring of the node output queues and an appropriate setting of the voice redundancy reduction threshold. This suggests that the algorithm is best employed in those situations where the bursty nature of voice traffic is retained;

- the 99.th percentile end-to-end delay appears to be heavily conditioned by the performance of the access network up to sustained channel utilization.

A certain exposure of the algorithm is associated with possible interventions on contiguous packets belonging to the same conversation segment. Related effects on the quality of the signal deserve a separate analysis which is beyond the scope of this paper.

6. REFERENCES
