DESIGN OF A PACKET VOICE TRANSMISSION SYSTEM

by

H. Tanaka * C. K. Chan M. Dressler V. R. Dhadesugoor

C. V. Chakravarthy ** D. L. Schilling

Department of Electrical Engineering
The City College of New York
New York, N. Y. 10031

Abstract

This paper describes the design of a packet voice network and the results of the evaluation tests performed. The packet voice network was simulated on a PDP-11/34 computer for real time operation. Adaptive delta modulators were used as source encoders. The average packet transmission rate and the subjective quality of the processed speech are presented.

I. Introduction

As the development of computer networks proceeds, the need for voice transmission facilities over packet switched networks has been growing, especially for use in teleconferencing which is a natural communication tool between people. Up to this date a network voice protocol has been developed for the ARPA network and some measurements have been performed to determine the delay time distribution of packets. Similar research has been performed on several other networks.

It is well known that conversation becomes difficult if the round trip delay is greater than a few hundred milliseconds. In large packet switched networks, such as the ARPA network, the round trip delay can easily be greater than hundreds of ms, especially when the number of hops and the packet rate become large. Moreover, the delay time changes greatly from packet to packet. Researchers in ISI [1] showed that the average delay time as well as the variance becomes large if the packet rate exceeds 10 packets/sec on the ARPANET. In addition, the packet arrival sequence may be different from that transmitted. To cope with this situation, every packet is assigned a time stamp which designates the output time of the packet (network voice protocol). To resequence the packets, using the time stamps, requires the use of buffers at the receiving end. This increases the average delay time of the packet leading to the degradation of conversational quality. As for the packet error, (the probability that some erroneous packets are received) it is relatively small because of error control which is usually used between adjacent switching nodes.

In this study, we evaluated the conversational speech quality in a situation where the round trip delay can change greatly, and we propose the design of a packet voice transmission system. We have simulated a real time packet voice transmission system and performed certain evaluation tests to determine the quality of the processed speech. The parameters used in these tests are delay time distribution, packet loss rate and silence detection algorithm. We have used the Song Voice Adaptive Delta Modulator (SVADM) at the source encoder.

II. Packet Voice Transmission System

The system diagram of a generalized packet voice transmission system is shown in Fig. 1. The voice waveform signal is encoded into a binary sequence and fed into the packetizer. The packetizer examines the bit stream, detects the start and the end of speech, packs the bits and makes up a sequence of packets. At the same time, it assigns the time stamp to each packet whose value designates the starting time of the packet. Packets which are generated by the packetizer are passed to the packet switched network in which every packet is delayed randomly and discarded with some probability (which simulates packet loss probability), and finally delivered to the receiver. A sequence regenerator buffers the packets, checks the value of time stamps with the present time, and makes up the output bit stream.

a. Voice/Silence Detection Scheme in Packetizer

Although the speech waveform is transmitted in a digital format, the bit stream during silent periods is neglected. Consequently, the voice/silence detection scheme plays an important role in reducing the effective packet rate. The detection method used is shown in Fig. 2. The input bit stream is processed in groups of 16 bit words. Every incoming word is stored in a shift register whose word size is fixed. It is then compared with several fixed bit patterns which are the typical bit streams at silent periods, and the result (match or no match) is stored in another shift

---

* on leave from Department of Electrical Engineering, The University of Tokyo, Tokyo, Japan
** on leave from Department of E & ECE, I. I. T., Kharagpur 721302, India

This work is partially sponsored by ARPA under grant MDA - 903 - 78 - C - 0182

register of entry length L_{max}. After that, the total number of matches in this register is compared with some constant whose optimal value is dependent on the present input mode.

When in the silent mode, the number of matches in the shift register is compared with a constant V_o. If the number is less than V_o, the start of the active speech is detected and packetization begins. At the head of the first packet, a number of the previously stored words (pre-offset) is inserted to preserve the start of speech. In voice mode, the number of matches is compared with a constant S_o. If the number is greater, the end of speech (silence) is detected. At that time, some input words (post-offset) previously stored in the packet buffer are discarded to shorten the packet length.

b. Sequence Regeneration Scheme

The delay time of each packet through the network varies from packet to packet. Therefore, the order of received packets does not always match the order of those transmitted. Furthermore, the packet location on the time axis may fluctuate from the original. When the variance becomes large, we cannot neglect its effect on the quality of speech. The limit of the variance for which we do not need any form of sequence regeneration is fixed by subjective evaluation of conversational speech quality.

When the variance is greater than the limit, the use of a sequence regeneration scheme is unavoidable. The scheme which we propose is as follows:

Let us assume the delay time distribution is as in Fig. 3. Packets with delay time less than T_s are stored in buffers: those with delay time greater than T_s are discarded. T_s is the absolute constant delay time of the packets between the source encoder and the destination decoder. At time T_s stored packets are outputs to the decoder.

The real shape of the delay distribution curve is shown to be similar to Fig. 3, [2], with most of the delay time concentrated near the minimum. Although the probability of occurence of large T_s is rather small, the distribution spreads to the very large delay time region. If P_e is to be very small, T_s can become sufficiently large so that the round trip delay becomes intolerable. P_e, which is the probability that the delay time is greater than T_s, gives the effective packet loss probability due to long delay time.

c. Encoding of Voice

To encode the speech waveform an adaptive delta modulation scheme has been used. The encoding algorithm used, is the Song voice adaptive delta modulator (SVADM) [3] given by:

$$1)\ X(k+1) = X(k) + S(k+1) \quad (1)$$

$$2)\ S(k+1) = |S(k)| \cdot e(k) + S_o \cdot e(k-1) \quad (2)$$

$$3)\ e(k) = \mathrm{Sgn}\,[M(k) - X(k)] \quad (3)$$

At the k^{th} interval, X(k) is the estimate of the incoming analog signal,

S(k) is the step size,

e(k) is the digital output of the encoder,

M(k) is the input signal

S_o is the minimum step size (constant)

Sgn function is defined by

$$\mathrm{sgn}(y) = \begin{cases} +1 \text{ for } y \geq 0 \\ -1 \text{ for } y < 0 \end{cases} \quad (4)$$

This encoding scheme was shown to be very efficient and to perform better than the continuously variable slope delta modulator (CVSD) [4].

III. The Simulator

The block diagram of the packet voice transmission simulator is shown in Fig. 4. The functions of packetizer, packet network and sequence regenerator are all performed by the PDP 11/34 computer. This simulator has been used for real time system evaluation.

a. Hardware Configuration

A PDP 11/34 minicomputer was used along with a DR-11 digital input/output interface to connect external devices to the computer. The specification of the control device used as interface (using 28 TTL Logic I.C.'s), between the DR-11 and a pair of encoder/decoder is as follows:

- 16 bit parallel input/output to/from computer for each channel.
- 16 bit parallel to/from serial conversion.

Bit streams from both encoders are stored bit by bit in shift registers (16 bit words), parallel transferred to the input buffer of the DR-11 and read into the computer memory. As the same clock is supplied to both encoders, input data for each channel is made up at the same time and read into memory sequentially. Data is read out of the computer after every read-in operation. From the output buffer of the DR-11 two words are placed into shift registers, one word for each channel, and continuous bit streams are generated for the decoders of both channels.

b. Software Configuration

The operation of the simulator program is shown in Fig. 5. The input/output processes are shown in Figs. 6(a) and 6(b) respectively. The program consists of 800 machine language instructions. The data area comprises 4K bytes (256 blocks) of packet buffer control blocks, and 16K bytes of packet buffer area for each channel, making up 36K bytes in total. After the read/write operation, the processing is performed sequentially for each channel. The processing sequence for each packet is as follows:

1. Voice detection (if in silence mode)
2. Allocation of packet buffer
3. Random Delay time generation
4. Insertion of packet buffer into the proper location of output-packet chain
5. Word collection

6. Silence detection (if in voice mode)
7. Comparison of the assigned output-time with present time and decision to output
8. Outputting of either words from packet buffer or silence patterns.

To perform these tasks. We use 3 packet buffer chains. A new packet buffer is acquired from the idle buffer chain, and an incoming word is stored in the buffer. The packets in the output chain are stamped with the output time and arranged in increasing order for transmission. If a new packet is created and the output time is assigned, the packet should be inserted into the proper location in the output packet chain by searching the chain. Process no4 (above) requires considerable processing time. For example, the number of packet buffers which exist in the computer can be greater than 40 in some cases. The margin which is permitted in each cycle for word processing is limited. 'Cycle' is the time unit from an input of a channel to the next input of the same channel. All time values are normalized to this unit. Processes no. 3 and no. 4, which are done at the time of new packet creation are time-divided into several sequential tasks, each of which is executed within a single word processing cycle. If N cycles of search operation are required to find the location, N+3 cycles in total are needed to complete the processing.

c. Output Time Generation for Each Packet

The arrival time of each packet can be calculated as follows:

$$T_{arv} = T_{create} + T_{min} + T_{random} \qquad (5)$$

Where T_{create} is the time when the packet is created, T_{min} is the minimum delay time of the packet switched network, and T_{random} is a random delay time. For the distribution function of T_{random}, 2 kinds of functions were assumed.

1. Flat density function
2. Approximate function of the measure result for the ARPA Network [2].

Random number generation was realized by

$$X = C \cdot X \qquad (6)$$

where C = 37, and X is a 16 bit integer.

IV. System Evaluation

The system has been evaluated by conducting the following tests:

a. Variation of Parameters in Silence/Speech Detection

Some of the important parameters such as the average number of transmitted packets, and speech quality have been obtained by varying the parameters used in the silence/speech detection. Results appear in Fig. 7 (a), (b), (c). In addition, packet size distribution measurements show that more than 95% of the packets are of full size. Speech quality was categorized in the following way:

Excellent - not different from or better than (due to silence rejection) the original speech.
Very Good - slightly different from original with no chopping of voice.
Good - slight degradation of speech due to chopping.
Fair - continuous chopping of voice although speech is still intelligible.
Poor - uninteligible.

b. Subjective Evaluation of a Two-Way Conversation With Constant Network Delay

With the parameters for silence/speech detection set at the optimal and packet size of 128 bytes, the ease with which a two-way conversation can be carried out has been evaluated. This test is conducted with a fixed time delay introduced in the system. The subjects are asked to rate the system into various categories as indicated in Table 1, as follows:

Very Easy - not different from local telephone.
Easy - conversation manageable with time needed for adjustment.
Difficult - difficulty in conversing due to large round trip delay.

c. Network Performance as a Function of Packet Loss And Random Delay

The quality of speech, introducing probabalistic packet loss and random delay time (random arrival) with flat distribution from T_{min} to T_{max} has also been obtained. Results are available in Fig. 8.

V. System Design Methology

As a result of the delay time distribution and packet loss probability measurements a packet voice transmission can be designed. From these values the optimal system parameters for the speech/silence detection scheme can be obtained.

a. Speech/Silence Detection Scheme

The number of words reserved for future speech/silence decisions should correspond to from 10 to 30 ms of speech. If we use 16K bits/sec. of delta modulation, L_{max} must be greater than 30 words (30 ms). Therefore, 32 is selected as a good number for L_{max}. The optimal value of the pre-offset and the post-offset are 8 and 16 words respectively. Those for the threshold parameters V_o and S_o, to change the processing mode, are 3 and 10 words.

b. Time Stamp Handling

If the absolute delay time is greater than 200 ms., we usually have difficulty with conversation. If the variance of the delay time exceeds 24 ms, we should be forced to use sequence regeneration scheme such as time stamping, when sequence regeneration is used it is suggested that the resulting constant delay time T_s between encoder and decoder should be adjusted so that the probability of packet loss due to a large

delay time becomes less than 10^{-2}. After T_s is fixed, the number of buffers needed for sequence regeneration can be calculated as follows:

$$N_b = F \cdot T_s / P \qquad (7)$$

where P is the average length of the packets in bits.

VI. Conclusions

In the above discussion, we assumed that the network characteristics are fixed and can't be changed. As the development of packet transmission systems progresses, it is expected that packet networks will have packet voice capability. At that time, packet networks will be designed with the provision that 99% of the packets will have a coast to coast delay time less than 300 ms. With the progress of packet switching speeds, the average delay time induced by one packet switch can be less than 1 ms. Digital transmission bit rate of 10 Mbits, to connect packet switching facilities, may be reasonable in the future as well.

With coast to coast transmission delay of about 20 ms in case of terrestrial link, and 250ms in case of satellite, a packet network for voice, as well as data, transmission, will be easily achievable.

References

1) Cohen D., "Specifications for the Network Voice Protocol", NSC Note 68 (RFC741, N1C42444) Jan. 1976.
2) S. L. Cansey, E. R. Masler, E. R. Cole, "Some Initial Measurements of ARPANET Packet Voice Transmission" Conf Rec. NTC'78, Birmingham, Alabama, pp 12.2.1-12.2-5, Dec. 1978.
3) C. L. Song, J. Garodnick, D. L. Schilling, "A Variable Step Size Robust Deltamodulator", IEEE Trans. on Comm. Technol. Vol. Com-19, pp 1033-1046, Dec. 1971.
4) V. R. Dhadesugoor, C. Ziegler, D. L. Schilling, "Deltamodulation in Packet Voice Networks" to appear in IEEE Trans. Comm., Nov. 1979.