Network Latency and Operator Performance in Teleradiology Applications

Johannes N. Stahl, Wyatt Tellis, and H.K. Huang

Teleradiology applications often use an interactive conferencing mode with remote control mouse pointers. When a telephone is used for voice communication, latencies of the data network can create a temporal discrepancy between the position of the mouse pointer and the verbal communication. To assess the effects of this dissociation, we examined the performance of 5 test persons carrying out simple teleradiology tasks under varying simulated network conditions. When the network latency exceeded 400 milliseconds, the performance of the test persons dropped, and an increasing number of errors were made. This effect was the same for constant latencies, which can occur on the network path, and for variable delays caused by the Nagle algorithm, an internal buffering scheme used by the TCP/IP protocol. Because the Nagle algorithm used in typical TCP/IP implementations causes a latency of about 300 milliseconds even before a data packet is sent, any additional latency in the network of 100 milliseconds or more will result in a decreased operator performance in teleradiology applications. These conditions frequently occur on the public Internet or on overseas connections. For optimal performance, the authors recommend bypassing the Nagle algorithm in teleradiology applications.

KEY WORDS: teleradiology, computer supported cooperative work, internet, network latency, Nagle algorithm.

TELERADIOLOGY applications often include an interactive conferencing mode that enables 2 or more physicians to discuss a problematic or doubtful case by means of remote control mouse pointers. These pointers are used to point to the findings on the images that are being discussed over a voice communication channel such as a telephone. Although this method enhances the quality and reliability of the teleradiology procedure, it also is a complex task that requires accurate audiovisual coordination.

Increasingly, public interconnected networks, such as the Internet, are being used for teleradiology. One of the pitfalls in using the Internet and the underlying network protocol TCP/IP is the network latency. Network latency is the time a data packet takes to travel from the origin to the destination. On the Internet, a data packet passes through many intermediate nodes, each of which delays the packet through internal processing and routing mechanisms. Wood et al[1] reported network round-trip-times (which are approximately twice the latency) between 18 and 424 milliseconds, depending on the type of connection.

In addition to the latencies that occur on the network itself, the TCP/IP protocol uses an internal optimization scheme called the Nagle algorithm,[2] which can delay the delivery of small data packets by up to 350 milliseconds even before they are sent to the network. The purpose of this algorithm is to improve the efficiency of the communication over many intermediate nodes by collecting small data packets into a single larger packet. The downside of this method is that small data packets are not sent immediately when they are generated, but later when the large packet is complete. Because the data packets used to transmit the position of the remote cursor between 2 teleradiology systems are very small (only a few bytes), teleradiology applications can be affected by the Nagle algorithm. This results in jerky movements of the remote cursor, because the cursor position messages do not arrive continuously, but intermittently, as the large network packets are transmitted.

Combined, the network latency and the Nagle algorithm can lead to an audiovisual dissociation with the effect that there is a temporal discrepancy between the position of the remote mouse cursor and the verbal description received over the voice channel, usually a telephone. The purpose of this report is to determine the possible effects of this dissociation on the performance of the user in a teleradiology context.

MATERIALS AND METHODS

To measure the operator performance, we needed to design test tasks that resemble the operation of a teleradiology system

From the Laboratory for Radiological Informatics, University of California, San Francisco, CA.

This work was partially supported by the National Library of Medicine Contract No. N01-LM-6-3547 and by a scholarship from the Deutsche Forschungsgemeinschaft (Sta 517/1-1).

Address reprint requests to Johannes N. Stahl, MD, Radiologische Klinik, Abt. Radiodiagnostik, Universitätskliniken, 66421 Homburg/Saar, Germany.

0897-1889/00/1303-0006$10.00/0
doi:10.1053/jdim.2000.8058

and can be measured quantitatively. We also needed a teleradiology system that can systematically simulate different network latencies.

We modified a teleradiology system developed at our institution,[3] which uses the User Datagram Protocol (UDP) for the transmission of the remote cursor position. This protocol has a latency of less than 10 milliseconds on a local area network (LAN). We added a programmable delay loop to the program that can simulate various network conditions by delaying outgoing data packets by an adjustable time.

We designed a synthetic test image that consists of white numbers 1 to 50 randomly distributed on a black background (Fig 1). We developed 2 test tasks that were performed by test persons using a pair of teleradiology systems in separate rooms and talking to each other over a telephone.

A. The first test person randomly points to a number on the image and waits for the second person to correctly communicate the number through the telephone. After the number has been identified by the second person, the first person points to another randomly selected number. This procedure is repeated for 40 numbers out of the 50 numbers on the test image. The time required to correctly identify all 40 numbers is measured with a chronometer.

B. The first test person moves the cursor across the numbers. In approximately 5-second intervals, without halting the movement, the first person instructs the second person over the telephone to identify the number under the cursor at that instant. Both test persons quietly write down the number and continue until 10 numbers are recorded. The number sequences noted by the 2 test persons are compared, and the number of errors is counted.

Both tasks simulate 2 radiologists identifying and describing structures on an image by combined visual and audible communication. The first task (task A) was designed to avoid mistakes by having the test persons cross check the numbers over the telephone and to measure the effects of the latency on the time required to perform the task. The second task (task B) measures

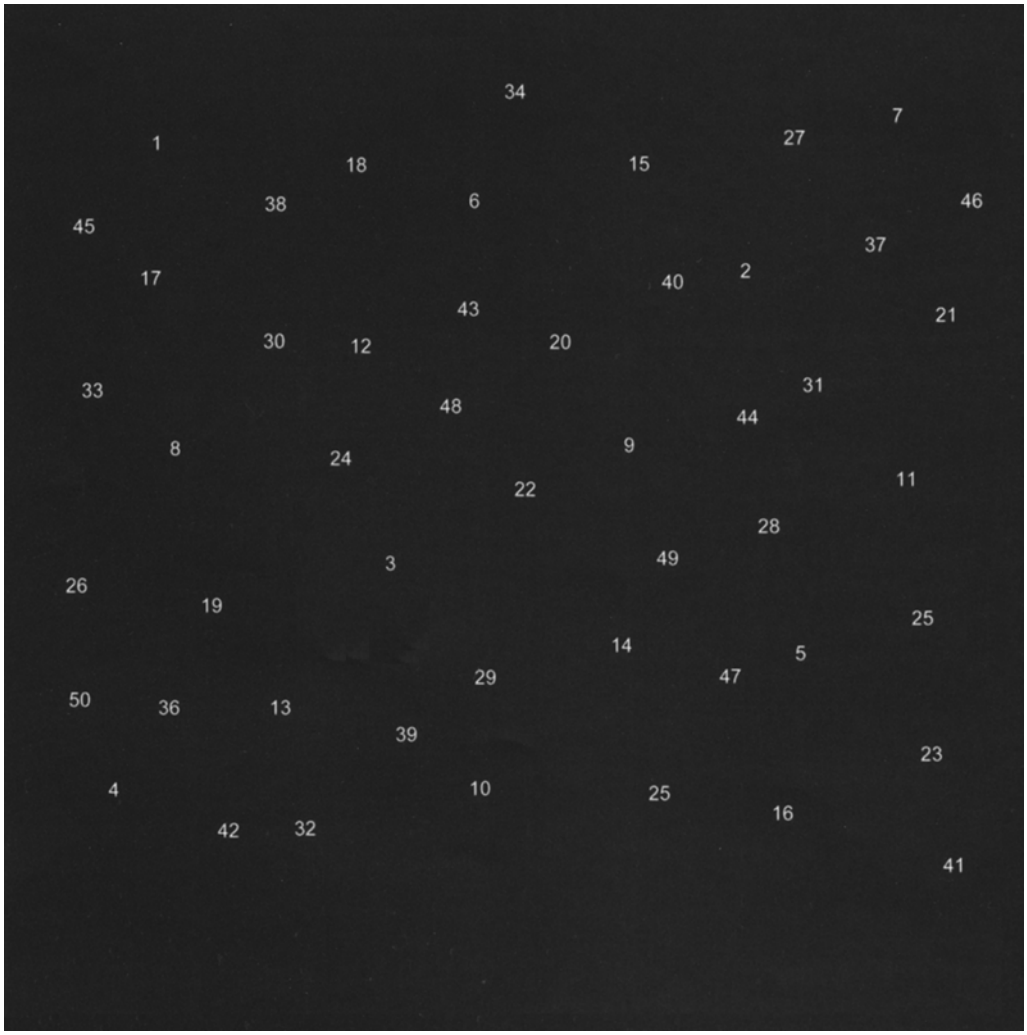

Fig 1. Test image with randomly distributed numbers.

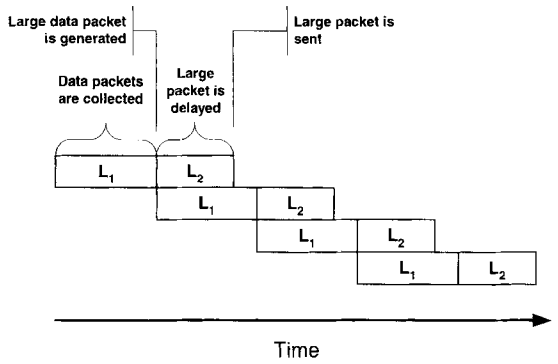

Fig 2. Timing diagram of the simulated composite latency. During the first L_1 period data packets are collected. When the timer L_1 elapses, a large packet is generated and delayed by the time L_2. The timer L_1 is reset and newly generated data packets are collected. When the timer L_2 elapses, the large data packet is sent. The delay loops L_1 and L_2 overlap each other.

the number of communication errors introduced by the latencies when no cross checking or audible feedback is performed.

Both tasks were performed by 5 test persons (nonradiologists, 22 to 30 years old) and repeated for 2 sequences of simulated latencies. The first sequence used a simple, constant latency simulating the latency of the physical network. We used the latencies 0, 250, 500, 750, and 1,000 milliseconds.

In the second sequence, we used a composite latency consisting of a variable portion L_1 simulating the Nagle algorithm, and a constant portion L_2 simulating the network latency. The variable portion L_1 will collect data packets into 1 large packet for the programmed time. Once the timer L_1 has elapsed, the large data packet is delayed by the time L_2 and then sent over the network. The 2 delay loops operate overlapped, so that while the timer L_2 delays the large packet, the timer L_1 is restarted, and newly generated small packets are collected. Please refer to Fig 2 for timing details. For the second sequence, we used combined latencies of 200, 250, 300, 400 and 500 milliseconds. We omitted the long latencies of 750 and 1,000 milliseconds, because the first sequence already showed that these latencies affect the performance intolerably. Instead, we used different combinations of L_1 and L_2 to simulate the combined latencies of 400 and 500 milliseconds.

RESULTS

Our first finding was that the results are very reproducible, although interindividual differences exist. Table 1 shows the results for task A with different constant latencies. When the latency was set to 0 the test persons needed, on average, 64 seconds to complete task A. Our expectation was that this time would increase by 10 seconds for each additional 250 milliseconds of latency (250 ms/number × 40 numbers/task = 10 s/task). This prediction is shown as the dotted line in Fig 3. The figure shows that with 250 milliseconds latency the test persons performed the task actually faster than predicted, whereas latencies of 400 milliseconds and above result in a worse-than-predicted performance. This difference is statistically significant (the times required to complete task A, normalized to the predicted value, differ for the latencies 250 milliseconds and 500 milliseconds with $P = .019$ in paired t test).

Table 1. Average Times Required to Complete Task A With Simulated Constant Latency

Latency (ms)	0	250	500	750	1000
Linear prediction (s)	63.8	73.8	83.8	93.8	103.8
Time required (s)	63.8	71.2	87	101	116.2
Standard deviation (s)	3.35	4.87	7.35	10.20	7.89

NOTE. n = 5.

With the composite delay, the results were almost identical (Table 2 and Fig 4). An interesting observation was that there is no significant difference between the different combinations of L_1 and L_2 and an equivalent constant latency of $(L_1 + L_2)$. This indicates that although the variable latency L_1 delays the packets on average only by $L_1 \times 0.5$, the effect on the performance in task A is the same as that of a constant delay of L_1.

The tests with task B showed a sharp increase in the number of errors for latencies above 400 milliseconds (Fig 5). In these tests, there was no significant difference in the number of errors between the sum of L_1 and L_2 and an equivalent constant delay, except for the 750-millisecond delay (Table 3).

DISCUSSION

Our results show that when performing complex audiovisual communication tasks dissociation between the visual and the audible input caused by network latencies results in possible communica-

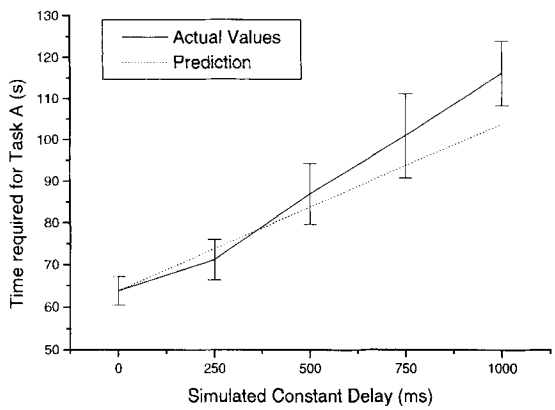

Fig 3. Results for task A with simulated constant delay (solid line). The predicted values are shown as a dotted line.

Table 2. Average Times Required to Complete Task A With Simulated Composite Latency

Total latency (ms)	250	400	500	200	300	400	500
Variable latency (L_1, ms)	249	250	249	100	200	300	400
Constant latency (L_2, ms)	1	150	251	100	100	100	100
Linear prediction (s)	73.8	79.8	83.8	71.8	75.8	79.8	83.8
Time required (s)	72.4	79.2	88.6	68.8	74.4	82.4	89.6
Standard deviation (s)	5.32	6.18	5.68	4.38	4.62	4.34	6.66

NOTE. N = 5.

tion errors and slower performance of the human operators. Latencies below 300 milliseconds are barely noticeable by the test persons, which is reflected in the better-than-predicted performance for the 250-millisecond latency. A possible explanation for this phenomenon is that 300 milliseconds are close to the reaction time for voice commands.

When the latency exceeds 400 milliseconds, the performance becomes increasingly worse than what could be predicted by simply adding the cumulated latency to the original performance. This coincides with a subjective loss of confidence expressed by the test persons.

This also is true for variable latencies that are caused by the Nagle algorithm. Our results suggest that regarding the operator performance, the effects of the Nagle algorithm are identical to a constant latency of the same magnitude as the maximum latency of the Nagle algorithm (usually 300 milliseconds). We proved this finding by performing an additional test series for task A with the TCP protocol instead of the UDP protocol, without additional simulated latency. The average time required for task A under these conditions was 77.2 ± 5.36 seconds, which is slightly longer than with a 300-millisecond constant delay (compare Table 2).

This is important insofar that when TCP is used as the communication protocol for teleradiology applications, the Nagle algorithm will impose a minimum effective latency of 300 milliseconds, which adds to any unavoidable latency caused by the physical network. Under these conditions, the remaining headroom to keep the total latency under 400 milliseconds becomes very narrow.

It could be debated whether the test tasks that we used are an adequate simulation of real-world teleradiology and whether the effects of the network latency on the performance will be the same under real-world conditions. Our results show that the test persons could compensate for very long latencies by increasing their concentration and by slowing the verbal communication. Although these mechanisms can maintain the communication, they also may lead to increased fatigue and a loss of confidence.

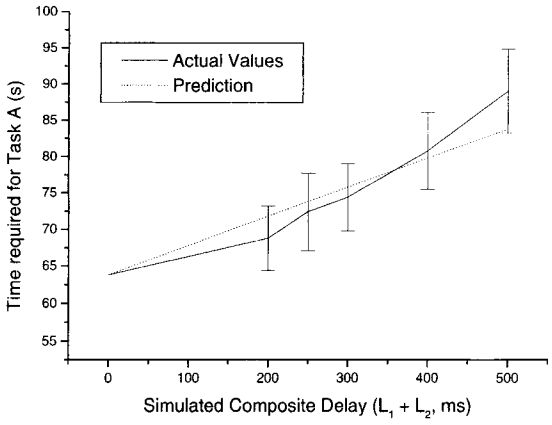

Fig 4. Results for task A with simulated composite latency (solid line). The predicted values are shown as a dotted line. Please note the different scale of the x-axis compared with Fig 3. The values for the different combinations of L_1 and L_2 are combined in this graph, because they do not differ significantly.

CONCLUSION

Interconnected public networks such as the Internet can be a useful tool for teleradiology applica-

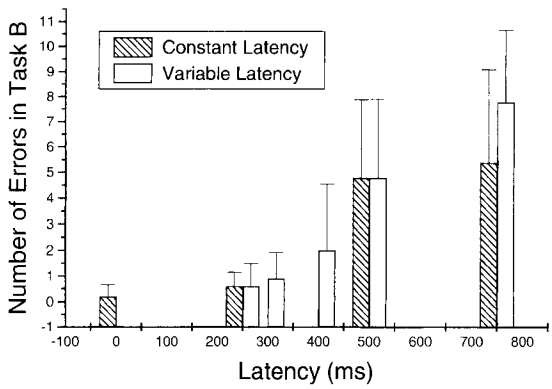

Fig 5. Number of errors (incorrect numbers out of 10 total numbers) encountered during task B, depending on the network latency.

Table 3. Average Number of Errors (Incorrect Numbers out of 10) Encountered During Task B for Constant Latencies (left, $L_1 = 0$) and Composite Latencies (right, $L_1 > 0$)

Total latency (ms)	0	250	500	750	250	300	400	500	750
Variable latency (L_1, ms)	N/A	N/A	N/A	N/A	200	250	250	250	250
Constant latency (L_2, ms)	0	250	500	750	50	50	150	250	500
Average no. of errors	0.2	0.6	4.8	5.4	0.6	0.9	2	4.8	7.8
Standard deviation	0.45	0.89	3.11	3.71	0.55	1.02	2.55	3.09	2.86

NOTE. N = 5.

tions in many regards. However, the designers of such applications should be aware of the potential pitfalls. We could show that network latencies above 400 milliseconds could significantly degrade the performance of teleradiology users in simulated tasks. Because the Nagle algorithm used by TCP/IP implementations already imposes a latency of about 300 milliseconds, the recommended 400 milliseconds can be exceeded under certain network conditions.

Although future implementations of the Internet such as the Next-Generation Internet (NGI) certainly will provide faster performance and minimized latency, it is nevertheless our conclusion that for optimum performance, the Nagle algorithm should be bypassed in teleradiology applications, if possible. This can be done by either disabling the algorithm, which some socket implementations allow, or by using the UDP protocol instead of TCP for the remote cursor. The latter approach, although more efficient, requires additional strategies to compensate for possible packet loss or packet reordering that can occur with UDP communication.

REFERENCES

1. Wood FB, Cid VH, Siegel ER: Evaluating internet end-to-end performance: Overview of test methodology and results. J Am Med Inform Assoc 5:528-545, 1998

2. Nagle J: Congestion control in IP/TCP internetworks. Internet Engineering Task Force (IETF) RFC 896, *http://www.ietf.org/rfc/rfc0896.txt,* 1984

3. Zhang J, Stahl JN, Song KS, et al: Real-time teleconsultation with high resolution and large volume medical imaging, in Horii SC, Blaine GJ (eds): Medical Imaging. San Diego, CA, SPIE, 1998(3339), pp 185-190