A VoIP Measurement Infrastructure

Ian Marsh
SICS
Sweden
ianm@sics.se.

Fengyi Li
KTH IMIT
Sweden
d97-fli@nada.kth.se.

Abstract

Time, day, location and instantaneous network conditions largely dictate the quality of Voice over IP calls. In this paper we describe a VoIP measurement infrastructure to measure the delay, loss and jitter of simulated phone calls on the Internet. We measure the quality by transmitting a simulated voice call between chosen sites and carefully recording the subsequent packet arrivals at the receiver. We have gathered more than 25,000 sample VoIP sessions from ten global sites. This is our second and more detailed attempt at measuring VoIP quality. This second phase has also focused on the effects of packet size, network asymmetry and silence suppression on measuring jitter, delay and loss. We have made the sessions and tools available for future investigations. Generally the quality of VoIP is excellent within the US and Europe and has improved since our last measurements. Finally this paper concludes with what we have learnt from two efforts of measuring VoIP quality on Wide Area Networks.

1 Introduction

No standardised scheme exists for monitoring VoIP quality. Many efforts are underway to measure and monitor the quality, however many have been solo ones. The goal of this work has been to determine what parameters are important and how they should be measured. As a side product we have measured the quality of VoIP in 2002 and compared it with our previous attempt three years previously, described in [HHM99].

A secondary goal has been to gather a large repository of sample traces which can be used by other researchers for their own purposes. We have focused on the three important factors of quality, namely loss, delay and jitter, however by making the sessions available researchers can look at other factors such as the loss patterns e.g. Bolot [BCG95] or more recently [JS02]. This paper starts with a brief description of the current measurement infrastructure in the next section. Within it we show the locations of the hosts used now and those used three years ago. Also included are details of the sample file we used to probe the connections and a summary of data we gathered as a result of this probing. We give some examples of what our infrastructure can measure and these are presented in the form of general results for loss, delay and jitter in section 3. The related work section compares results obtained by other researchers in similar efforts to ours. In the scope of this work this is important as we are interested in how quality has changed over the last years. Finally, we round off the paper with future directions for measurements in Section 5 and some conclusions in Section 6.

2 The Measurement Infrastructure

2.1 Simulating Phone Calls

Our approach is to periodically send a pre-recorded phone call between two hosts over the Internet and measure how the packets arrive at the receiver. Since we are interested primarily in Voice over IP, we focus on loss, jitter and the one-way delay. These are three of the parameters used in the E-model standardised by the ITU-T[Int98]. We have also recorded the routes between the chosen hosts. The measurement procedure did not include any signalling, only the packet data transport. In the telephony world, this would correspond to only the speaking portion of the call, not the dialling of the number, the delay or ringing the remote phone. Details of the call characteristics are given in the top-half of Table 1.

2.2 Intermittent Probing

As mentioned this is the second attempt to measure VoIP quality on a large scale, so we have had chance to improve on the measurement infrastructure. The improvements can be summarised as:

- A full-mesh topology between available hosts
- More informed choice of hosts and more variation in distance, hops and time-zones
- Secure and automatic invocation of sessions
- Investigate asymmetry, packet size and silence suppression
- Clear separation of the infrastructure and VoIP tool

Our hosts are all computers at universities. Ideally we would like to have considered sites which are connected to a commercial network as well. The time of day measurements would probably show the opposite effect of an academic network behaviour, the busy times being in the evenings rather than the normal daytime load, however such hosts were not available to us.

The geographic locations of the hosts for the measurements are shown in Figure 1. Also shown are the hosts we used when we performed similar measurements in 1999, shown as squares. As can be seen only four sites remain for us to make fair quality comparisons, however it is doubtful whether all the links and routers are the same so an identical comparison is probably not feasible. Using ten fully interconnected hosts

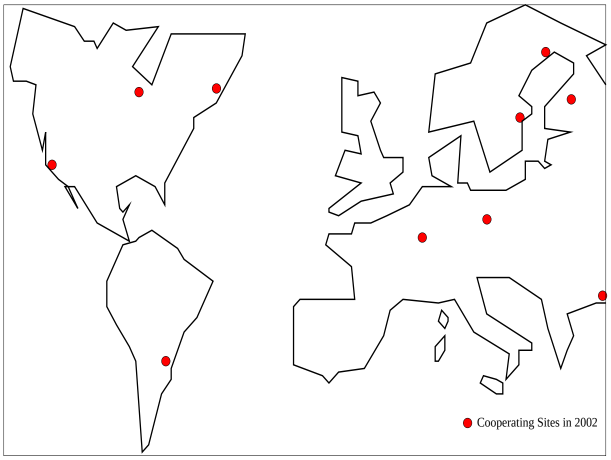


Figure 1: Locations of Test Sites

and automatic invocation enabled us to collect more than 25,000 sample sessions over a period of 15 weeks, the data we obtained is summarised in the bottom half of Table 1. As far as we know, this is the largest repository of VoIP sample sessions made available.

Used Data	**Details**
Call duration	$\sim$ 150 secs
Coding	8 bit PCM Ulaw
File Size	584480 bytes
No silence sup.	3653 packets
Silence sup.	2043 packets (57%)
Obtained Data	
Measurement duration	15 weeks
Total trace files	25000
Normal tests	14000
Bidirectional tests	5000
Packet size tests	4000
Silence Suppression tests	2000

Table 1: Used And Obtained Data

2.3 Measurement Core

To send, receive and measure loss, delay and jitter for VoIP one needs an RTP enabled tool. The tool sends the pre-recorded conversation from one host to its peer(s[1]) and

[1] We have not considered multicast as part of these experiments.

measures the quality parameters per session. The results (in the form of a trace file) are sent back to the central site for processing and to be stored in the central repository. This process is handled entirely by Sicsophone [HHM02], a low-delay VoIP tool modified for these measurement purposes. RTCP is also included in the implementation for delay measurements, however we chose to process the traces from the sessions packet by packet. This gives us finer granularity when investigating loss patterns for example.

2.4 Log File Format, Timestamping And Storage

One advantage of using Sicsophone is its ability to save detailed log files. For each transmission, Sicsophone creates a log file at the receiver, recording each packet event. The silence periods are also recorded, as an 'empty' event. The trace file format used in this work is essentially a dump of the RTP header. An example of the log file format used in this work is shown in Figure 2.

```
I 1017112866 661181
T 1017112866 862086
E 1017112866 883304 172 32869 0 0 160
E 1017112866 902002 172 32869 1 0 320
E 1017112866 917461 172 32869 2 0 480
E 1017112866 936997 172 32869 3 0 640
-------------------------------------
1 2          3      4   5     6 7 8
```

Figure 2: Log File Format

Field 1 is the event which follows 'I' indicates a session start, 'T' an empty event, and 'E' data has arrived, 2 and 3 are Unix Epoch timestamps. Field 4 is the packet size (160 byte payload + 12 bytes for the RTP header), field 5 is the RTP header, 6 the sequence number and 7 and 8 timestamps. Once the session is completed the file is compressed and securely copied to the central site for processing.

2.5 Timestamping Sessions

We save all the sessions in subdirectories named by the sender plus receiver in the form of host1_host2. We therefore currently have (9 x 9) - 9, 72 subdirectories. The file name is the timestamp at the central location, Sweden, when the sessions were started. Swedish time is GMT+1 hour for winter time and GMT+2 for summer time. This makes it easy to compare files taken at the same time, for example, when investigating asymmetry, silence suppression or even the effect of packet size . We simply look for the files with the same name but in different directories, one named by host1_host2 and another named by host2_host1.

For plots when the local time is important (i.e. pattern over the day) then we can use the time in the files themselves (fields 2 and 3 are the Unix epoch) to get the local time at a particular location, e.g. as in Figure 4. In experiments such as these a clear design of what should be measured clearly influences how the data should be stored for the most efficient retrieval. This is even more important where real-time processing has to be done.

2.6 Scheduling and Invocation

To perform the tests over a regular time periods, we need to initiate the sessions at given times. This is done by *crontab*, a Unix facility for scheduling jobs to run them at particular times. Generally in our tests, a session between a certain pair of hosts is scheduled to run once per hour. We chose this frequency as it did not place high loads on remote machines which we did not control. An alternative would have been to implement the time scheduling and invocation within Sicsophone itself. However, this would not allow us to use other tools instead of Sicsophone. By using Unix scripts, we can use other tools for the sending and receiving the audio streams. Our primary motivation to implement the infrastructure in this manner was to be able to use a tool which we could not modify or had features we could implement ourselves without too much effort. For example Realaudio streams could be used for the tests by simply substituting Sicsophone with Realaudio tools.

Sampling more often than once an hour would give us finer granularity measurements, however it would not be truly representative of a typical usage pattern. One improvement in our measurement infrastructure would be to include call patterns similar to those on a telephone network [Sch02]. This essentially would capture the calling pattern of users, however this is not probably the same for IP telephony. Because of pricing differences people tend to speak longer due to the lower tariffs. One weakness of our use of the Unix scheduling is that the calls are invoked at the same time for each host, so for example Belgium always sends at HH:13 to Argentina, some randomisation would be an improvement.

3 Results Obtained

3.1 Overall

We show some example of the results that can be gathered by using such an infrastructure. More detailed results on VoIP quality can be found in [Li02]. As mentioned this is the second time we have conducted such tests, so it is natural to begin with a comparison of the measured qualities in 1999 and 2002. We should stress that in 1999 we only had hundreds of measurements, whilst in 2002 we gathered many thousands. The results in Table 2 are for the four hosts which were available both times.

Quality	1999	2002	% diff. (+/-)
Jitter	45.1ms	22.6ms (± 13.7)	-50.0%
Loss	1.2%	0.5%	-58.3%
Delay	115 ms	84.95ms (± 44.85)	-26.1%

Table 2: 1999 and 2002 VoIP Quality Differences

3.2 Loss

Table 3 shows the loss for connections between all the sites. The top number in the cell is the average loss percentage and the lower value is the standard deviation. From this table we can see that most of the connections have very low loss rates (less than 1%) except the connections where the Turkish or Argentinian sites are involved. We can guess this is due to the high bandwidth networks which are rarely congested. From the table in the case of the Turkish host note we cannot say if the loss is due to congested international links between them and the other sites or access problems at the site, for example many students on a LAN sharing a low-bandwidth connection to the Internet. However from looking at the patterns over a number of days, particularly the weekend (not shown), we can see the same loss patterns, thus assuming the students are not working all weekend we can attribute the problem to congested International links rather than access to bandwidth at the LAN level.

receiver / sender	Mass.	Mich.	Cali.	Belg.	Finl.	Swed.	Germ.	Turk.	Arg.
Mass.	/	0.066 (0.60)	0.121 (0.99)	0.124 (0.83)	0.096 (0.76)	0.037 (0.20)	0.0 (0.0)	4.88 (4.65)	8.971 (7.21)
Mich.	0.021 (0.15)	/	0.186 (1.10)	0.020 (0.11)	0.099 (1.09)	0.126 (2.19)	0.186 (0.90)	2.98 (1.92)	6.54 (7.06)
Cali.	0.057 (0.26)	0.122 (1.93)	/	0.196 (0.75)	0.604 (1.38)	0.191 (0.26)	2.822 (3.01)	4.367 (2.42)	8.932 (8.20)
Belg.	0.0 (0.0)	0.0 (0.0)	1.170 (0.97)	/	0.004 (0.01)	0.0 (0.0)	0.171 (0.66)	3.768 (2.67)	NA
Finl.	0.001 (0.10)	0.017 (0.25)	0.660 (1.39)	0.072 (0.25)	/	0.001 (0.01)	0.0 (0.0)	3.177 (1.73)	7.494 (6.49)
Swed.	0.0 (0.0)	0.025 (0.37)	0.069 (0.10)	0.081 (0.25)	0.001 0.01	/	0.0 (0.0)	2.984 (0.95)	NA
Germ.	0.0 (0.0)	0.002 (0.01)	2.515 (1.87)	0.004 (0.01)	0.0 (0.0)	0.0 0.0	/	3.732 2.48	NA
Turk.	8.129 (2.77)	7.960 (2.91)	7.676 (6.83	7.093 (3.98)	7.786 (2.70)	8.389 (3.14)	7.965 (3.10)	/	NA
Arg.	0.506 (1.38)	0.521 (1.46)	0.618 (1.76)	0.480 (1.38)	0.529 (1.39)	0.028 (0.09)	0.056 (0.10)	5.822 (2.97)	/

Table 3: Packet Loss (Percentage and Deviation)

3.3 Delay

ICMP times are derived from `ping` measurements and using the RTCP sender and receiver reports implemented in Sicsophone. We sent ICMP packets of 160 bytes at 20ms intervals as well as using the RTP stream of Sicsophone to measure the round-trip delay, we used two methods for comparison. To obtain the one-way delay, we divide the round-trip time by two. The question arises if asymmetry can give vastly different delays for the two directions. In Table 4 the diagonal separates the direction of any two hosts. The tests were run over a number of weeks. The top number is the one way delay derived by RTCP reports, while the lower one is the one-way ICMP echo request and reply time (standard deviations in parenthesis).

receiver / sender	Mass.	Mich.	Cali.	Belg.	Finl.	Swed.	Germ.	Turk.	Arg.	NZ
Mass.	/	38.02 (17.14) 37.54 (16.88)	54.15 (15.81) 55.80 (18.90)	67.11 (15.47) 67.15 (14.70)	97.09 (2.60) 98.00 (4.35)	99.46 (8.51) 98.90 (5.67)	58.39 (5.00) 58.29 (5.08)	388.2 (43.17) 415.5 (61.69)	NA 99.64 (4.91)	149.0 (61.99) 159.8 (59.38)
Mich.	36.44 (15.36) 34.55 (12.73)	/	40.38 (4.50) 40.73 (4.30)	63.54 (4.24) 63.51 (4.44)	88.17 (7.98) 86.83 (2.40)	86.73 (4.71) 85.61 (2.13)	63.61 (8.20) 60.95 (3.71)	358.9 (44.94) 390.9 (38.95)	NA 112.1 (10.55)	124.3 (52.14) 111.9 (8.60)
Cali.	54.45 (16.67) 55.21 (18.57)	40.58 (5.07) 41.35 (4.81)	/	81.04 (2.24) 83.84 (0.49)	105.9 (2.98) 108.3 (1.84)	107.5 (2.39) 107.7 (2.47)	81.53 (1.80) 82.07 (0.65)	386.9 (60.46) 419.2 (35.43)	NA 123.9 (12.44)	81.90 (9.74) 81.24 (2.87)
Belg.	65.24 (10.13) 62.17 (10.13)	63.37 (3.34) 60.68 (3.22)	84.01 (1.33) 81.53 (1.53)	/	31.32 (0.60) NA	33.41 (0.17) 30.23 (0.17)	16.58 (10.36) NA	341.1 (24.61) 337.0 (15.14)	NA 136.5 (7.12)	152.1 (4.30) NA
Finl.	97.84 (4.23) 97.71 (2.81)	86.75 (1.86) 86.55 (2.01)	109.9 (4.70) 109.6 (2.84)	NA 30.72 (0.32)	/	13.62 (1.00) 13.44 (0.65)	26.72 (7.27) 24.07 (5.63)	321.2 (39.32) 309.4 (19.67)	NA 161.5 (12.23)	168.2 (8.68) 170.0 (6.20)
Swed.	99.28 (8.76) 97.77 (5.74)	84.91 (1.94) 84.72 (0.26)	105.6 (2.11) 106.5 (0.76)	33.28 (0.39) 32.75 (0.29)	13.49 (0.53) 13.00 (0.00)	/	29.77 (12.72) 23.50 (0.00)	322.2 (30.34) 333.6 (23.08)	NA 165.6 (17.90)	169.9 (13.08) 176.7 (16.23)
Germ.	63.47 (9.57) 62.39 (9.57)	60.35 (0.54) 60.29 (1.03)	84.40 (9.97) 82.40 (1.42)	NA 11.07 (0.24)	27.75 (7.33) 27.50 (7.23)	29.24 (7.60) 25.87 (4.75)	/	300.7 (39.66) 318.5 (66.39)	NA 149.8 (15.68)	158.7 (15.87) 159.3 11.74
Turk.	379.1 (47.09) 380.6 (26.30)	387.9 (35.46) 379.8 (28.80)	410.9 (43.89) 425.1 (32.50)	330.2 (28.62) 334.1 (19.89)	318.9 (42.44 332.1 (25.10)	311.1 (8.28) 322.1 (9.38)	378.2 (49.26) 369.2 (37.13)	/	NA 490.8 (25.80)	486.5 (43.40) 473.8 38.21
Arg .	NA 117.0 (30.77)	NA 146.7 (44.18)	NA 152.0 (47.75)	NA NA	NA 164.1 (27.20)	NA 160.9 (47.65)	NA 180.5 (50.36)	NA NA	/	NA NA
NZ .	166.6 (117.7) 156.3 (113.0)	110.4 (15.31) 111.4 (6.69)	80.01 (6.03) 80.95 (3.54)	NA NA	162.8 (7.11) 167.1 (3.35)	173.9 (5.71) 172.1 (4.29)	153.4 (6.67) 153.3 (1.73)	463.1 (69.13) 482.3 (30.62)	NA 206.4 (10.22)	/

Table 4: ICMP and RTCP Delays in milliseconds ($\pm$ std. dev.)

3.4 Jitter

Jitter is the statistical variance of the packet interarrival time. In the RFC 1889 jitter is defined to be the mean deviation (smoothed absolute value) of the packet spacing change between the sender and the receiver. If packet i is sent from the sender with timestamp s_i and arrives at receiver at time r_i, then for two packets i and j, the difference of the packet spacing d may be expressed as:

$$d = (r_j - s_j) - (r_i - s_i) = (r_j - r_i) - (s_j - s_i).$$

Sicsophone sends packets of the same size at each interval which means that $s_j - s_i$ is constant. The difference of the packet spacing d is used for calculating the interarrival jitter. According to RFC, the interarrival jitter should be calculated continuously as each packet i is received. Using d for one particular packet the interarrival jitter j_{i-1} for the previous packet $i - 1$ can be calculated according to the formula:

$$j_i = j_{i-1} + (|d(i-1,i)| - j_{i-1})/16.$$

This algorithm is the optimal first-order estimator and the parameter $1/16$ gives good noise reduction (according to the RFC). Figure 3 plots the average jitter values for connections grouped into regions. We have chosen US, EU, US to EU and Argentina as geographic groups. We chose this grouping to highlight the difference in jitter *to* Argentina. The measurements were taken simultaneously. Each 'region' contains approximately 1200 data points, i.e. for every connection we have about 400 jitter measurements.

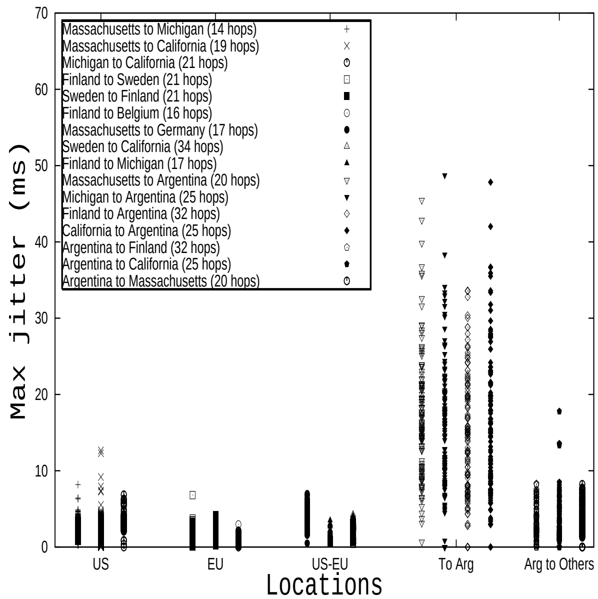


Figure 3: Average Jitter

3.5 Network Asymmetry

Our results show, although asymmetry exists it is not so serious in terms of VoIP quality as we had expected or reported in 1999. Even in the last case the difference is 60ms, not too disastrous to South America. In our 5000 bidirectional tests, the loss, delay and jitter differences were usually small. For example, the loss values were up to 0.5-1.25% in opposite directions. Figure 4 is an example of route asymmetry where loss is concerned; a matrix of the hop counts in both directions is given in [Li02].

3.5.1 Silence Suppression

Using silence suppression means no packets are sent when the sender is silent; without silence suppression packets are sent even during silence. We have tested our links using

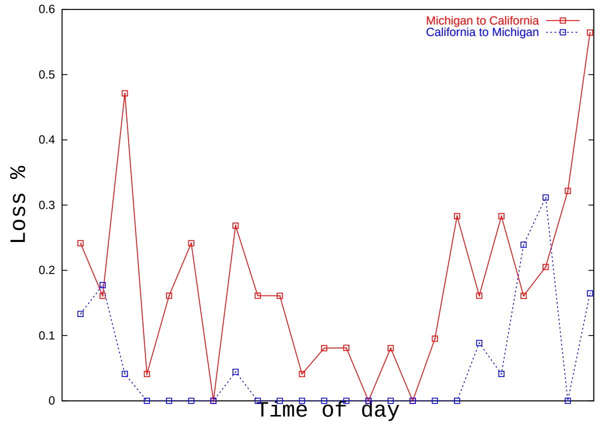


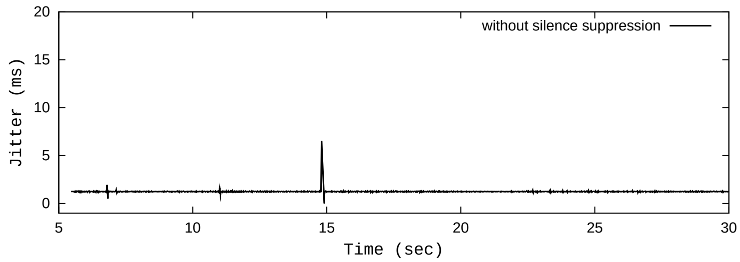
Figure 4: Route Asymmetry

both silence suppression and not. We found that using silence suppression did not have any significant influence on the network load, i.e. we could not cause any extra loss by not using it. Figure 5 plots the jitter values for a connection from Massachusetts to Michigan. With silence suppression the average jitter is 1.250 ms, while without the average is 1.281 ms. The difference is small. In total we performed 2000 silence suppression tests.

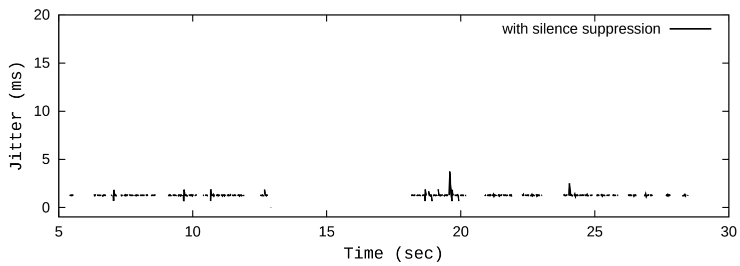
In high load situations (or low capacity) using silence suppression may influence the network load. Without silence suppression more data is sent so that we can get more sampling of the network, but the network is slightly more loaded. For good connections where the loss is very low one can choose to send without silence suppression. For bad connections such as to Argentina and Turkey, it is advisable to use silence suppression to reduce the load on the network.

4 Related Work

In this section we consider work done by other researchers and compare our results with theirs. In 1997, Maxemchuk and Lo measured the quality of intra state, inter country, and international Internet links [ML97]. They defined the quality of a connection as 'the fraction of time that the signal is received without distortion for intervals of time that are long enough to convey active speech segments'. The conclusions were that the Internet was capable of carrying voice data although there are some difficulties. They stated longer connections gave worse quality and the quality was much worse during busy hours. The same conclusions were drawn by Dong Lin in 1999 [Lin99]. She measured the end-to-end delay, packet loss and jitter of both US and international links, and concluded that packet size did not influence packet loss. Our conclusion agrees with hers even though our measurements were taken three years after hers. She also observed that the length of consecutive losses was small. In her tests, loss lengths



(a) Without silence suppression



(b) With silence suppression

Figure 5: With and Without Silence Suppression

of one and two packets were predominant, while in ours, it is one to three. Compared with her results, we observed lower delays, however since we did not have access to exactly the same links as she did we cannot say this categorically. Even our connections which are "longer" give lower delay but identical test should be conducted to verify this. In her tests three years ago, the round trip delay was under 300 ms for only the hosts inside US, while in our tests most of the links, except the Argentinian and Turkish host, have round-trip delay under 300 ms.

5 Possible Future Directions for VoIP Measurements

Since the quality is reasonably good for VoIP (and improving) we have highlighted a few potential areas for future investigations. We have not looked at hosts on commercial networks due to accessibility reasons. So the measurements should be considered from academic end-points. However many VoIP users will use their connections from home and the usage patterns may well be significant different from those we have measured. As stated in our measurements we usually send once an hour. More sampling will show the quality of the network more exactly, and in some cases it is important to see if "outages" occur causing either user to terminate the session. We should also try different call lengths. Currently our sample call lasts approximately 150 seconds (for 160 bytes/packet). An improvement would be to use short, medium and long duration calls, which would help verify some of the results reported in the related work

(Maxemchuk for example). We might see different effects for longer calls. We could also include signalling of the calls, since a failure at this stage would not result in the call being set-up. This is far more serious than a few percent packet loss for the user. A lot of work has been done to measure the quality of VoIP, but little attention has been paid to the reliability issues. More investigations should be done on the connectivity of the hosts. We have discovered at least one host resides on a network where traffic shaping is performed. This is useful information when looking for explanations of observed effects. Finally the log files can be used for further research. Since we have collected almost 25,000 log files, they can be used for looking at artifacts other than loss, delay and jitter.

6 Conclusions

We have constructed a VoIP measurement infrastructure for investigating quality of telephone calls on the Internet. We use simple Unix commands to start, terminate and gather data from scheduled sessions. We use a simple VoIP tool, Sicsophone to send a pre-recorded conversation between the participating sites. The infrastructure has been made robust by enscapsulating the session in Unix shell script. So far we have gathered more than 25,000 sessions, more or less automatically. We have used ten hosts located around the globe to conduct our tests. We have recently added one in New Zealand, but most importantly we have connected them in a full-mesh configuration and started VoIP sessions between them securely and automatically. There has been a focus on quality characteristics of loss, delay and jitter. In summary we can state VoIP is feasible today in 2002, but not on a global scale. From the sites which we had access to in 1999 and now, we can also say that the quality for VoIP is improving. Much of our effort has been devoted to handling exception cases, suh as machines being unavailable, configuration changes, and SSH versions, which are where the problems for VoIP lie. Generally the quality of the network is good, but the hosts and routers and more specifically changes to them cause many problems. If a VoIP system is to be successfully deployed it needs to be more *reliable* from our experience.

References

[BCG95] J. Bolot, H. Crepin, and A. Garcia. Analysis of audio packet loss in the internet. Lecture Notes in Computer Science, pages 163–174, Durham, New Hampshire, April 1995. Springer.

[HHM99] Olof Hagsand, Kjell Hansson, and Ian Marsh. Measuring internet telephone quality: Where are we today ? In *Proceedings of the IEEE Conference on Global Communications (GLOBECOM)*, Rio, Brazil, November 1999. IEEE.

[HHM02] Olof Hagsand, Kjell Hansson, and Ian Marsh. End system adaption to packet audio jitter. In *Submitted to ACM Multimedia 2002*, 2002.

[Int98] International Telecommunication Union. The E-model, a computational model for use in transmission planning. Recommendation G.107, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, December 1998.

[JS02] Wenyu Jiang and Henning Schulzrinne. Comparison and optimization of packet loss repair methods on voip perceived quality under bursty loss. Miami Beach, Florida, May 2002.

[Li02] Fengyi Li. Wide Area Measurements of Voice over IP Quality, June 2002.

[Lin99] Dong Lin. Real-time voice transmissions over the Internet. Master's thesis, Univ. of Illinois at Urbana-Champaign, 1999.

[ML97] N. F. Maxemchuk and S. Lo. Measurement and interpretation of voice traffic on the internet. Montreal, Canada, June 1997.

[Sch02] Henning Schulzrinne. When Will the Telephone Network Disappear?, June 2002.