Characterizing voice and video traffic behavior over the Internet

Prasad Calyam, Chang-Gun Lee

Department of Electrical and Computer Engineering,
The Ohio State University, Columbus, OH 43210.
pcalyam@oar.net,cglee@ece.osu.edu

Abstract. In this paper, we present our research on characterizing voice and video traffic behavior in large-scale Internet Videoconferencing systems. We built a voice and video traffic quality measurement testbed to collect Videoconferencing traffic traces from several sites all over the world that were connected to our testbed via disparate network paths on the Internet. Our testbed also featured the H.323 Beacon, an H.323 session performance assessment tool we have developed, and various other open-source and commercial tools. Our findings obtained by analyzing the collected traffic traces demonstrate the impact of: 1) end-point technologies that use popular audio and video codecs and 2) network health status that is characterized by the variations of delay, jitter, lost and reordered packets in the network, on the end-user perception of audiovisual quality. The perceptual data used in our analysis includes both objective and subjective quality measures. These measures were collected from our testbed experiments for a few sample tasks involving various levels of human interaction in Internet Videoconferences.

1 Introduction

Technologies that support Voice and Video conferencing applications over IP networks (VVoIP) [1] [2] have already gained wide acceptance in todays end-user communities. Internet Service Providers and equipment vendors have developed large-scale conferencing systems in support of end-user communities [3] that have integrated VVoIP infrastructures into their data networks.

It is becoming increasingly necessary to develop VVoIP friendly network protocols and devices such as firewalls, Network Address Translators (NATs), packet shapers and call-admission control systems that augment existing VVoIP system functionalities, without however affecting end-user experience of audiovisual quality. Developing VVoIP friendly network protocols and system devices requires a sound understanding of the VVoIP traffic characteristics. Simulation studies fail to provide a realistic environment in which vital VVoIP traffic characteristics can be understood at a fundamental level. Hence, we have developed a novel measurement methodology for collecting real-world Videoconferencing traces. We have also developed techniques that could be used to effectively analyze the trace data in order to better understand the impact of factors such as end-point technologies and network health status on the end-user perception of audiovisual quality.

The end-user perception of audiovisual quality is measured using two popular methods: subjective quality assessments and objective quality assessments. Our characterization studies address both objective and subjective quality assessments. Subjective quality assessments involve playing a sample audiovisual clip to a number of human participants. Their judgment of the quality of the clip is collected and used as a quality metric. However, objective quality assessments do not rely on human judgment and involve automated procedures such as signal-to-noise ratio (SNR) measurements of original and reconstructed signals and other sophisticated algorithms to determine quality metrics. The reader is referred to [4] for additional details relating to the above two methods.

Many of the earlier attempts to characterize VVoIP traffic [5] [6] focused on voice traffic studies alone, and among these only few addressed issues of traffic characterization in the purview of end-user perception of audio quality. Studies such as [7] [8] discussed QoS issues relating to video traffic, in only a LAN environment, with a minimal setup of test cases and end-points. Their main focus was on studying frame rates issues, again without emphasis on end-user perception of the video quality for the various test scenarios. Other studies such as [9] chose a wider group of participants for obtaining perceptual quality evaluations of video. However, the results of their study were limited in scope since they covered only jitter and packet loss issues for streams originating from a single video codec (MPEG-1).

To the best of our knowledge, our study is the first to comprehensively characterize both audio and video traffic streams with the following considerations:

- The use of various network health scenarios and multiple end-point technologies (Ex., codecs, vendors, pc-based vs. appliance-based equipment) and diverse end-user demographics to obtain subjective and objective audiovisual quality assessments
- A study of the above audiovisual quality assessments for various Videoconferencing tasks that feature various levels of end-user interactions routinely seen in Internet Videoconferences

For the purpose of our study, we analyzed more than 300 VVoIP traffic traces collected by performing one-on-one testing in a LAN environment and on the Internet with over 26 sites located across multiple continents connected via disparate network paths. The network paths included research networks, commodity networks, and last-mile cable modem, DSL modem and satellite network connections. The collection of the traces involved a systematic emulation of various network health scenarios characterized by various values of network delay, jitter and loss. The traces were collected for a set of routine Videoconferencing Tasks that featured typical end-user interaction levels.

The remainder of the paper is organized as follows: Section 2 describes the methodology for collecting the traffic traces. Section 3 discusses our analysis of the collected traffic traces. Section 4 concludes the paper.

2 Measurement Methodology

A novel measurement methodology was developed as part of our study. The methodology addressed challenges involved in coping with the network dynamics for efficient and accurate collection of traffic traces from world-wide sites for a specific set of experimental conditions. The following subsections describe in detail the various challenges dealt by our methodology.

2.1 World-wide Testbed Setup

Our world-wide testbed (shown in Fig. 1) was used to collect the VVoIP traffic traces from numerous sites (shown in Fig. 2) that participated in one-on-one testing to rate audiovisual quality for various Videoconferencing tasks. Prior to our testing with each site, we gathered various types of information such as the participant profile, experience of the participant with VVoIP, site-network connectivity, and end-point technology routinely used for audio and video conferences . The participants included various demographics of end-users: Videoconferencing Co-ordinators, Network Engineers, Instructors, Graduate Students and IT Managers. Internet connectivity at the sites included LAN, research, commercial and last-mile DSL, Cable Modem and Satellite network links. Participating sites used various commercial and open-source VVoIP end-point technologies. The various audio codecs observed to be used at the endpoints included GSM, G.711 and G.722 and the various video codecs included H.261, H.262 and H.263.

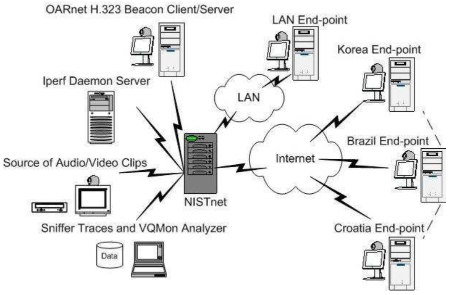

Fig. 1. Overall World-wide Test Setup

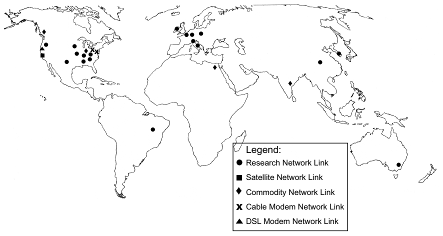

Fig. 2. World-wide sites participating in the testing

Many tools were used to obtain the VVoIP traffic traces and other performance data necessary for analysis. Table 1 shows the various tools used and their purpose in the testing. Measurement data was collected for both LAN and Internet tests. The basic difference between the LAN and Internet tests in terms of the NISTnet settings for generating different network health scenarios was that- in the LAN environment, the delay, jitter and loss values configured on the NISTnet were the overall values in the end-to-end path. In the Internet tests, the network paths had inherent values of delay, jitter and loss. Hence to configure any necessary end-to-end delay, jitter and loss values for the Internet tests, a network path characteristics pre-determination step was required for each site. To accurately obtain the inherent network path characteristics, we used a heuristic approach using the data obtained from the OARnet H.323 Beacon and NLANR Iperf. The end-to-end delay, jitter and loss values configured on the NISTnet were obtained by deducting inherent path characteristics values from the LAN settings.

No	**Tool Name**	**Tool Description**	**Purpose of Tool in Testing**
1.	OARnet H.323 Beacon	H.323 session performance measurement tool	For network path characterization and subjective quality ratings
2.	Ethereal	Network traffic sniffer and trace analysis tool	To capture traffic traces for analysis
3.	Telchemy VQMon	Network traffic trace analyzer for VoIP Systems	For objective quality ratings
4.	NISTnet	Network-emulation tool	To introduce delay, jitter and loss in the end-to-end traffic paths
5.	NLANR Iperf	TCP/UDP throughput measurement tool	For network path characteristization

Table 1. Tools Used in the Testing

Amongst the tools used in the testing, the OARnet H.323 Beacon [11] is an application-specific network measurement tool that we have developed. It provides H.323-protocol specific evidence and other information necessary to troubleshoot H.323 application performance in the network and at the host (end-to-end).

In our study, we employed both the subjective and objective quality assessment methods to determine end-user perception of audiovisual quality for various network health scenarios. To obtain subjective quality assessment scores from the test participants, we extended the Quality Assessment Slider methodology presented in [4] and developed our own slider that was integrated into our H.323 Beacon Client GUI. Participants ranked the audiovisual quality for the various Videoconferencing tasks on a scale of 1 to 5, which is the Mean Opinion Score (MOS) ranking technique.

To obtain objective quality assessment scores, we utilized the Telchemy VQMon tool [12] that implements the E-Model and uses traffic traces obtained for the various Videoconferencing tasks as an input for the analysis. The E-Model is a well established computational model that uses transmission parameters to predict the subjective quality. Though, the E-Model fundamentally addresses objective quality assessment of voice traffic, our collected data shows reasonable correlation of the subjective quality assessment scores for audiovisual quality provided by the participants and the objective quality assessment scores provided by VQMon. Section 3.1 of this paper discusses the correlation results. The reader is referred to [13] for additional details relating to E-Model components.

2.2 Quantizing Network Health Scenarios

Our definition of network health refers to the effect of the combined interaction of delay, jitter and loss. The three network parameters co-exist for every path in the Internet at any given point of time. Regulating any one of the parameters affects the other parameters and ultimately the Quality of Service (QoS) perceived by the end-user in terms of H.323 application performance. [14] illustrates a real-world example where resolving a loss problem in an Intercampus DS3 led to a decrease in the observed loss but unexpectedly led to an increase in the overall jitter levels in the network.

Table 2. Nine Network health scenarios for NISTnet Settings

	S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**
Delay	G	A	P	G	G	P	G	G	A
Jitter	G	A	P	G	P	G	G	A	G
Loss	G	A	P	P	G	G	A	G	G

Legend:
G: Good, A: Acceptable, P: Poor, S1-S9: Scenario (1-9)

In our design of experiments, we employed a full factorial design for the 3 parameters i.e. we emulated 3^3=27 scenarios that covered every possible network health status; each status comprising of different levels of delay, jitter and loss. The levels chosen for the scenarios were derived from our earlier work [10] where the performance bounds of the network parameters were mapped to Good, Acceptable and Poor grades of quality as experienced by an end-user. The Good grade corresponds to delay values of (0-150) ms, jitter values of (0-20) ms and loss values of (0-0.5) %. The Acceptable grade corresponds to delay values of (150-300) ms, jitter values of (20-50) ms and loss values of (0.5-1.5) %. The Poor grade corresponds to delay values > 300ms, jitter values > 50ms, loss values > 1.5%.

We performed extensive LAN tests that covered all of the 27 scenarios and selected 9 scenarios shown in Table 2 for Internet measurements whose results in essence reflected the results of the 27 scenarios. To understand this selection process, we can consider an example that uses a three-letter notation which will be used in the remainder of this paper to indicate network health status. A (GGG) notation implies values of (delay, jitter, loss) in the good performance bounds and in that order. Letters A and P are used to indicate values in the acceptable and poor performance bounds respectively. Hence, a GGG condition refers to the best provisioned network path and a PPP condition refers to the worst provisioned network path, that can be seen in real-networks today.

It can be intuitively determined that the PGG or GPG scenario and the PPP scenario results which are amongst the 9 selected scenarios can be used to deduce the results that could be expected from the PPG, GPP, PGP scenarios. Another such example is - the results of GAA, AGA, AAG could be deduced from the results of the GGA or GAG scenarios and the AAA scenario results. Our process of selecting 9 scenarios out of all the possible 27 scenarios significantly reduces the time involved in each subjective assessment session with the end-users and also makes it more practical for obtaining data from a large number of test sites.

2.3 Videoconferencing Tasks

For each of the 9 scenarios described in Section 2.2, a Videoconferencing task was assigned. A Videoconferencing task can basically be any activity that takes place in a routine Videoconference. A casual conversation or an intense discussion or even a class lecture can qualify as a Videoconferencing task. We associated every Videoconferencing task with an activity level based on the interaction intensity. The activity levels were: low, moderate, or high. A low activity level refers to scenarios where the participant is passively viewing the remote stream with minimal interaction. A moderate activity level refers to scenarios where the participants discuss short questions and answers. A high activity level refers to scenarios where the participants are involved in a heated debate or where the participants are taking lecture notes that require undivided attention and audiovisual stream clarity. There is significant literature recommending strategies for tasks that could be part of Subjective and Objective assessments of audiovisual quality [4]. All of them recommend that in addition to passive viewing for assessment of audiovisual quality, the participants must be presented with realistic scenarios. Key guidelines proposed in the above literature were followed in the tasks creation, tasks ordering, participants training to score the audiovisual quality and overall environment setup for the assessment.

3 Analysis Of Traffic Traces

To obtain the necessary data from the traces, we used the RTP analysis module of the Ethereal network protocol analyzer. We filtered-out the audio and video streams from the other LAN traffic and decoded the audio and video packets information.

3.1 Subjective versus Objective Quality Rankings

Figs. 3 and 4 show the box-plots for the various subjective and objective quality rankings determined from the experiments, respectively. The box-plots indicate the minimum, maximum, 1st quartile, 3rd quartile, median and mean values of the rankings obtained for the data aggregated from all the sites for the 9 network health scenarios. We can analyze the results in relation to the horizontal lines marked at MOS values of 3 and 4 which are the boundaries for acceptable and

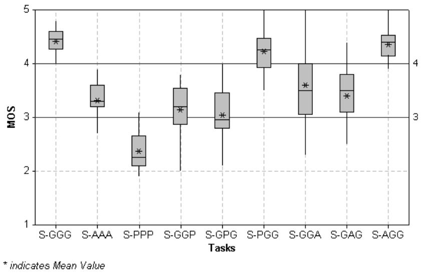

Fig. 3. Subjective Quality Rankings for the 9 Network Health Scenarios

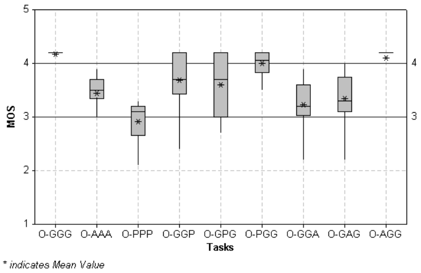

Fig. 4. Objective Quality Rankings for the 9 Network Health Scenarios

poor grades of perceived audiovisual quality respectively. It can be noticed that both in the case of subjective and objective assessments, the MOS was least affected by the values of delay being in the acceptable or poor performance bounds. A more detailed discussion relating to this observation can be found in Section 3.3. From Figs. 3 and 4, the degradation of the perception of audiovisual quality can be seen to be more sensitive in the case of values of jitter in the acceptable or poor performance bounds compared to that of loss values. This observation illustrates that jitter is the most dominating factor amongst delay, jitter and loss that affects end-user perception of audiovisual quality.

The Pearson-Correlation values for delay, jitter and loss for the subjective and objective quality rankings when compared on an individual basis were 82.7%, 73.7% and 71.2%, respectively. The correlations can be observed to be in the above-average to strong range; a 100% indicates maximum attainable strength of correlation. Our correlation results are impressive considering the fact that they correspond to the aggregate of the data points obtained from experiments involving various demographics of users, various types of codecs, and various network link speeds spanning multiple continents as described in Section 2.1. In the case of our LAN results,we observed a strong correlation between the subjective and objective rankings. This observation can be attributed to the lack of diversity in the demographics of the subjects and the limited end-point technologies available in our LAN testbed.

3.2 Packet-size Distributions versus Activity Levels

Understanding traffic characteristics in terms of packet-size distributions is important since it has implications on the end-to-end performance achieved by the traffic streams. A better understanding of how the network handles various packet size distributions could help in determining important trade-offs at the application level. We explored the packet size distributions for Videoconferencing tasks involving high, moderate and low activity levels. Fig. 5 shows the video packet-sizes distribution data aggregated for all the sites traces for the high-activity level in the GGG network health setting. The video packet-size distributions for moderate and low activity levels were nearly similar to the high-activity level characteristics.

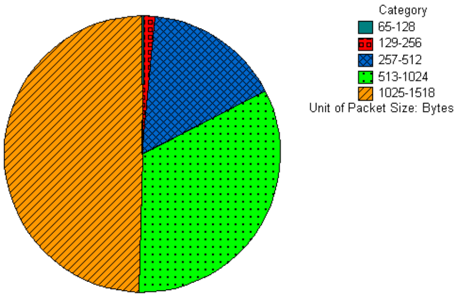

Fig. 5. Video Packet-sizes Distribution for High-activity level and GGG Network Health Scenario

This similarity was also observed in all of the other network health scenarios. Our observations are in accord with the fact that packet sizes generated at the sources are mainly affected by the details and movements encoded in the transmitted images. Since routine Videoconferencing tasks do not generally contain video with high temporal aspects (e.g. Soccer or Hockey sports telecast videos), similarity in the aggregation of packet size distributions for the various network health scenarios and high, moderate and low activity levels in our experiments can be expected. The above packet-size distribution characteristics can be safely used to represent routine Videoconferencing traffic while developing network simulations to model VVoIP behavior as seen in the Internet.

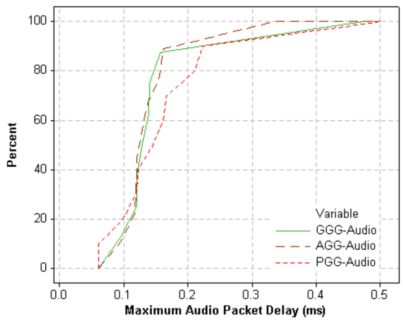

Fig. 6. Cumulative Distribution Function of Maximum Inter-packet Audio Delay

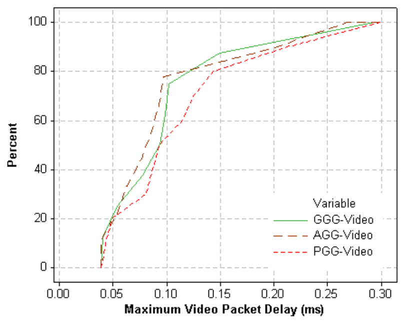

Fig. 7. Cumulative Distribution Function of Maximum Inter-packet Video Delay

3.3 Audio and Video Inter-packet Delays

Figs. 6 and 7 illustrate the impact of good, acceptable and poor grade network delay values in the network on the maximum inter-packet delays aggregated from all the sites for the audio and video streams at the application level, respectively. It should be noted that the maximum inter-packet delay is different from the end-to-end delay values. The inter-packet delay is a measure of the time difference between 2 succesive audio or video packets as observed at the network layer. End-to-end delay indicates the amount of time a packet takes to travel from the sender's application to the receivers's application. The components that contribute to the end-to-end delay include: (a) compression and transmission delay at the sender, (b) propation, processing and queuing delay in the network and (c) buffering and decompression delay at the receiver.

It can be observed from Figs. 6 and 7 that the cdf curves are almost close to one-another, i.e. the maximum packet delays of audio and video packets at the application level are not affected by the delay magnitudes (GGG, AGG, PGG) in the end-to-end network path. This observation concurs with the conclusion presented in Section 3.1 that large values of end-to-end delay do not affect the end-user perception of audiovisual quality. Irrespective of the delay conditions in the network, the packets get relatively offset as they traverse the network. Common use of interactive Videoconferencing seen in satellite IP network environments, where one-way delay values reach upto a second, can also be referred to affirm our above conclusion. However, it should be noted that large end-to-end delay values hamper the interactivity of human conversations in a Videoconference though not the end-user perception of audiovisual quality.

3.4 Inter-packet Jitter versus Network Health and Packet Sizes

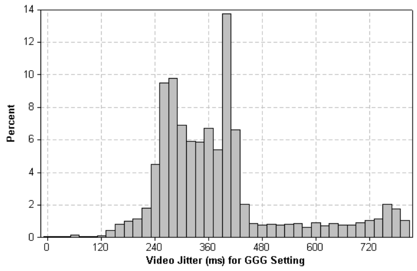

Fig. 8. Probability Mass Function of Inter-packet Video Jitter for the GGG Network Health Scenario

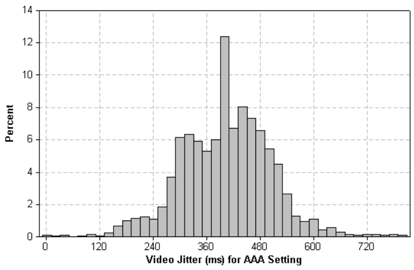

Fig. 9. Probability Mass Function of Inter-packet Video Jitter for the AAA Network Health Scenario

Figs. 8 - 10 show the probability mass function or the occurrence frequency of the inter-packet jitter data aggregated from all the sites for the GGG, AAA and

PPP network health scenarios. It should be noted that the inter-packet jitter does not correspond to end-to-end jitter. It is a value observed at the network layer without any smoothing by jitter buffers at the application level. The application layer jitter buffers compensate the large inter-packet jitter values and reduce the end-to-end jitter values using sophisticated adaptive buffering schemes.

It can be observed from Figs. 8 - 10 that the mass of the inter-packet jitter in the graphs shifts in the direction of the larger jitter values as the network health deteriorates. For the GGG setting, the inter-packet jitter values are all clustered around 400ms, while for the AAA setting, though the highest frequency of inter-packet jitter values are around the 400ms range, there is a relatively higher spread of values greater than 400ms. For the PPP setting, a bimodal distribution can be observed with two significant peaks of inter-packet jitter values at about 400ms and 680ms. Depending on the video jitter buffer settings at the end-points, the packets with large inter-packet jitter values get dropped at the application level. It is obvious that a significant amount of drop can be expected for the PPP network health scenario compared to the GGG and AAA scenarios, which also explains the very low MOS scores shown in Figs. 3 and 4 for the PPP case.

Fig. 11 shows the impact of increasing packet sizes on inter-packet video jitter values for the GGG, AAA and PPP network health scenarios for all the sites. Although not entirely apparent in the case of AAA network health scenario, for the GGG and PPP network health scenarios, we can note that the inter-packet video jitter tends to decrease when higher packet-sizes are used. Hence, the optimization of the performance of VVoIP traffic streams may be achieved by using relatively larger packet-sizes. If we refer back to Fig. 5, we can notice that general Videoconferencing traffic utilize a higher percentage of large packet-sizes.

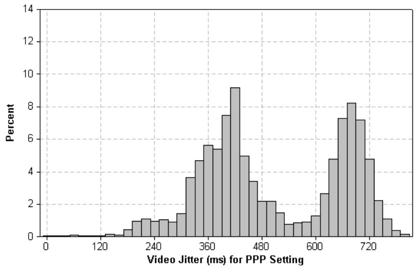

Fig. 10. Probability Mass Function of Inter-packet Video Jitter for the PPP Network Health Scenario

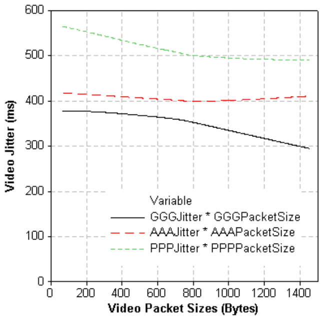

Fig. 11. Inter-packet Video Jitter versus Packet Sizes for the GGG, AAA and PPP Network Health Scenarios

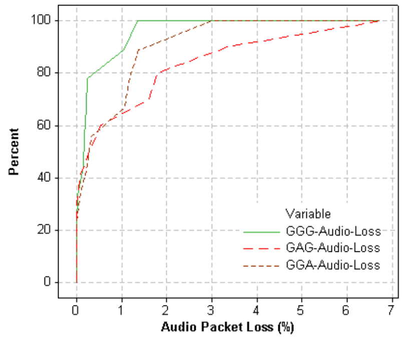

Fig. 12. Cumulative Distribution Function of Audio Packet Loss for Acceptable Jitter and Loss Levels

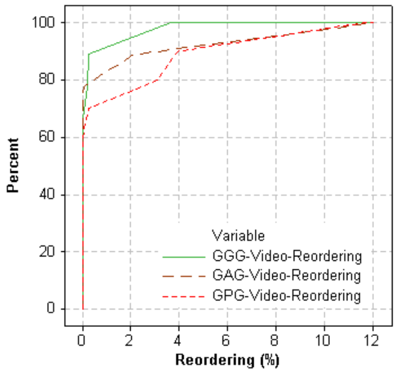

Fig. 13. Cumulative Distribution Function of Video Packet Reordering for Acceptable and Poor Jitter Levels

3.5 Application-level Packet Loss and Reordering versus Network Health

Fig. 12 shows the impact of jitter and loss (GAG and GGA) on audio packet loss data aggregated from all the sites at the application level. It is evident that network jitter has a higher impact on the audio packet loss compared to the impact of network delay or loss. A similar characteristic was observed in the case of video packet loss also at the application level. Although end-point jitter buffers can be made to dynamically adapt to any significant presence of network jitter for minimizing the jitter buffer discards for packets that do not meet the play-out deadline, they do not significantly ameliorate the stream performance because of the general high variability of the network characteristics in the timescale of a typical Videoconferencing session.

Packet reordering also has been known to cause performance bottleneck issues such as lack of lip-synchronization in VVoIP applications. Lack of lip-synchronization refers to the case where the audio and video streams are not multiplexed in a synchronized fashion at the playback time resulting in a perceived mismatch in the audio and video content. Jitter being the most crucial factor amongst delay, jitter and loss in the context of VVoIP application performance, we investigate the impact of acceptable and poor network jitter levels on the packet reordering seen at the application level. Fig.13 shows the cdf of the packet reordering with the increase in the network jitter. It can be observed that even for the GGG scenario reordering occurs. The reordering becomes more pronounced for the GAG and GPG scenarios. The existence of high values of reordering could make the VVoIP traffic more non-deterministic at the end-points. This non-deterministic nature impacts the buffering and smooth-playback of the audio and video streams which ultimately affects the end-user perception of audiovisual quality as shown in Fig. 3 and 4 for the GAG and GPG cases.

4 Conclusion and Future Work

In this paper, we presented our carefully designed and executed experiments for obtaining VVoIP traffic traces from a testbed involving 26 sites world-wide. The results of our experiments demonstrated the impact of the various end-point technologies and network parameters on the end-user experience of audio-visual quality while using VVoIP applications. We also presented a comparative analysis of the objective and subjective quality assessments for various network health diagnostics. Further, we analyzed the end-to-end performance variations of VVoIP applications by characterizing packet size distributions, packet delay, jitter, loss and reordering, for both audio and video streams.

We are currently using our above experimental methodology and site contacts to study the impacts of high-speed real-time multimedia data streams at high-speeds (> 768Kbps). We are also planning to extend our observations to develop a mathematical model that suggests the impact of various network and end-point technologies on the end-user perception of audiovisual quality involving VVoIP applications.

References

1. ITU-T Recommendation H.323, “Infrastructure of audiovisual services- Systems and terminal equipment for audiovisual services”, 1999.
2. M. Handley, H. Schulzrinne, E. Schooler, and J. Rosenberg, “SIP: Session Initiation Protocol,” IETF RFC 2543, 1999.
3. Megaconferences: Worlds largest Internet Conferences - http://www.megaconference.org
4. J. Mullin, L. Smallwood, et.al, “New techniques for assessing audio and video quality in real-time interactive communications”, IHM-HCI Tutorial 2001.
5. I. Marsh, F. Li, “Wide-Area measurements of Voice over IP quality”, QoFIS 2003.
6. A. Markopoulou, F. Tobagi, M. Karam, “Assessment of VoIP quality over Internet backbones”, IEEE INFOCOM 2002.
7. V. Chandrashekar, “A framework for Quality of Service analysis of IP-based video networks”, Masters Thesis, North Carolina State University, 2003.
8. R. Finger, A. Davis, “Measuring video quality in Videoconferencing systems”, Wainhouse Research Whitepaper, 1998.
9. M. Claypool, J. Tanner, “The effects of jitter on the perceptual quality of video”, ACM Multimedia, 1999.
10. P. Calyam, M. Sridharan, W. Mandrawa, P. Schopis, “Performance measurement and analysis of H.323 traffic”, PAM 2004.
11. P. Calyam, W. Mandrawa, M. Sridharan, A. Khan, P. Schopis, “H.323 Beacon: An H.323 application related end-to-end performance troubleshooting tool”, ACM SIGCOMM NetTs 2004.
12. A. Clark, “Modeling the effects of burst packet loss and recency on subjective voice quality”, http://www.telchemy.com
13. ITU-T Recommendation G.107, “The E-Model: A computational model for use in transmission planning”, 1998.
14. A. Indraji, D. Pearson, “The Network”, Internet2 Commons Site Co-ordinator training, 2003.