Supporting Enterprise-Grade Audio Conferencing on the Internet

Krishna Ramachandran and Sunitha Beeram

Citrix Online
Santa Barbara, USA

Abstract. This paper evaluates if the Internet can support enterprise-grade audio conferencing. For our investigation, we collect real-world traffic traces from the audio conferencing solution in GoToMeeting, a popular online meeting application. We analyze these traces using *Conference MOS*, a new metric proposed in this paper to measure the quality of an audio conference. Our results indicate that a majority of users experience good conference quality. A small percentage of users that experience poor quality because of high delay and loss would be better served with PSTN. This leads us to believe that an enterprise-grade solution should adopt a combined VoIP and PSTN deployment strategy over a VoIP only solution.

1 Introduction

The market for audio conferencing and its integration with enterprise applications is growing fast. For example, in Europe, it is projected to reach USD 712 Million in 2010 [3], while in the United States it is expected to be up to USD 3 Billion by 2013 [5].

Audio conferencing has traditionally been available via the Public Switched Telephone Network (PSTN). Enterprise users use their PSTN phones to dial into meetings hosted on conferencing bridges. With the rapid evolution of VoIP technologies and the increasing penetration of broadband access in residential and business settings, audio conferencing over VoIP is fast becoming a viable alternative.

Models for VoIP conferencing include *dial-in conferencing* and *peer-to-peer conferencing* [18]. In dial-in conferencing, end-points call into a Voice Conferencing Bridge, which performs audio mixing and conference control. In the peer-to-peer model, the end-points are responsible for these tasks. An example of the latter approach is Skype [6], which by most measures epitomizes VoIP telephony on the Internet.

Judging by the popularity of Skype, one can hypothesize that peer-to-peer mixing is well suited for conferencing. However, supporting *enterprise-grade* quality presents a set of critical challenges, not seen in typical peer-to-peer applications. We define enterprise-grade conferencing as one that is scalable, available and has Mean Opinion Score (MOS) [9] quality rating higher than 3.5.

In terms of scalability, enterprises require that audio conferencing solutions support a large number of callers, sometimes up to a thousand users within a single conference [2]. The processing and packetization of audio are complex operations. For example, Skype, which operates by mixing audio at the end-points, can only support a

S.B. Moon et al. (Eds.): PAM 2009, LNCS 5448, pp. 143–152, 2009.

conference up to 25 users [6], because of this complexity. In order to support a large number of users, powerful servers are required to perform audio processing and encoding. Dial-in conferencing is amenable to such provisioning.

The second challenge has to do with availability. Enterprises expect high availability, *99.999%* being the holy-grail. Users in a *large* meeting can join and leave whenever they choose. In a peer-to-peer setting, handling user churn and failures is challenging, both in design and implementation. In contrast, dial-in conferencing is simpler to design and implement, operate, troubleshoot in case of problems, and, finally, provision for fault-tolerance and reliability.

In this paper, we are interested in investigating how well an enterprise-grade audio conferencing solution performs on the Internet. Toward this goal, we characterize the performance of our dial-in conferencing solution that provides audio conferencing for Citrix GoToMeeting [1], a popular online meeting application used by thousands of businesses around the world. GoToMeeting is a Sofware as a Service (SaaS) application that is used to collaborate online, give presentations and for online training. For our study, we collect five days of traffic traces from Aug 5-9, 2008 from a subset of voice conferencing bridges that are part of our VoIP infrastructure. Our data set contains over 9000 VoIP flows, which total approximately 230000 minutes of talk time.

For our analysis, we propose the Conference Mean Opinion Score (CMOS), a new metric that helps characterize audio quality in a conference. The MOS of a conference depends not only on the number of users in a conference and their network connections, but also on the time varying set of active speakers. Existing metrics fail to account for these considerations.

The major findings from our study are as follows:

- A majority of users experienced good quality in their conferences. The median CMOS was approximately 4.27. Only 11.75% of users had CMOS less than 3.5.
- A majority of VoIP connections to the conferencing bridges experienced good network conditions. The median MOS of these connections is 4.30. Only 11% had MOS less than 3.5.
- Unlike VoIP paths in the backbone that are well provisioned for high capacity [12,14], we find that end-to-end paths can experience high delay and loss. These factors resulted in 11% of VoIP connections to have MOS less than 3.5.

This paper offers the first measurements-based characterization of VoIP conferencing on the Internet based on real-world traffic measurements. Past VoIP studies [12,13,14,16] primarily investigated the performance of peer-to-peer VoIP flows. A second contribution of this paper is the Conference MOS, a metric for the estimation of audio quality in a conference.

2 Related Work

Several studies [12,13,14,16] have investigated the performance of VoIP flows in sections of the Internet. Bolot [13] in 1995 found that loss was a result of network congestion. Markopoulou et. al. [16] in 2001 probed 43 backbone paths between five cities in the US. They found that some of these paths experienced delay spikes and intermittent

outages lasting 0.5-2 seconds. In a more recent study, Birke et. al. [12] monitored the backbone of an Italian ISP and found loss, not delay, to be the principal contributor to bad quality. Majority of VoIP calls placed over the ISP backbone were of good quality. Boutremans et. al. [14] reported that loss is a result of failures in backbone links because of either inefficiencies in the routing protocol, faulty router software or human errors.

To the best of our knowledge, there exists no measurements-based characterization of a real-world audio conferencing solution. Birke et. al. [12] studied VoIP calls placed in the backbone network of an Italian ISP. However, they did not focus on audio conferencing and did not consider end-to-end VoIP flows that span multiple domains.

Proposals [8,17] exist to estimate the Mean Opinion Score (MOS) [9] of a conversation between two participants. To the best of our knowledge, there is no solution for the estimation of the MOS in a conference. Conference MOS depends on the number of speakers, the quality of their network connections and the active speaker set at any given time. The solution offered in this paper takes these factors into account.

3 Conference MOS Estimation

We base our estimation of the conference MOS on the *E-model* [8], a widely used analytic model standardized by the International Telecommunications Union (ITU). The E-model predicts MOS by taking into account various factors, such as equipment quality, encoder used, packet delay and loss. We use the E-model because it has two main advantages: (1) it is *non-intrusive*, i.e., it does not require comparison of the original voice signal against a degraded one; and (2) it is *parametric*, i.e., it predicts MOS from measured properties, such as echo, delay and loss. In contrast, *signal-based* models [17] function by processing the output voice signal. Parameteric models, in general, are easier to apply than signal-based methods. Further, where user privacy is a critical requirement, such as in enterprise solutions, the use of a signal-based method is typically not possible.

3.1 E-model Overview

The basic principle behind the E-model is that the perceived effects of impairments are additive. The E-model results in a quality rating R on a scale of 0 in the worst case to 100 in the best case for narrow-band codecs and from 0 to 129 for a mix of narrow-band and wide-band codecs[10]. Figure 1 shows the relation between MOS and R for wide-band codecs.

R is computed as: $R = 129 - I_d - I_e$, where 129 is used, because our data set has flows that are encoded using *iSAC* [4], a wide-band codec; I_d and I_e are impairment factors that represent the quality degradation due to delay and loss respectively. Notice that the two impairments are isolated into separate terms even though the delay and loss events may be correlated in the network. This is because the E-model computation assumes that the impairment contributions corresponding to the network events are separable.

In actuality, R takes into account several other impairments, such as echo, background noise and imperfections in end-user equipment. These are not represented above

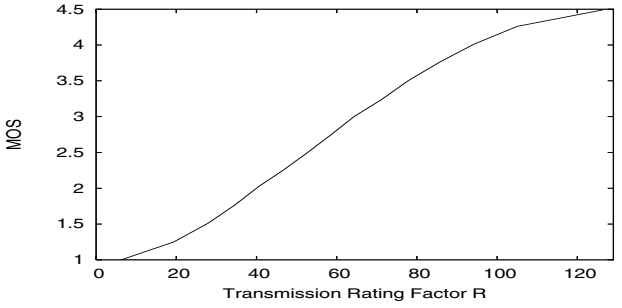

Fig. 1. Mapping between R and MOS for wide-band codecs

because we are primarily interested in assessing the suitability of the Internet for audio conferencing, whereas these factors do not depend on the network [15].

Impairment values for the iSAC codec are not available in ITU standards. Hence, we derived these values empirically as described in ITU specifications [7]. We conducted conversation-opinion tests [7] by varying either the network delay or loss. We used ten pairs of human subjects in our testing. For delay, the two test subjects took turns counting numbers as fast as possible; delay effects are most noticable during such a test. For loss, each subject read ten lines from a random paragraph. Each reading lasted about thirty seconds which is sufficient time for a subject to observe any degradations [17]. Figure 2 plots the average of the impairment values obtained from the ten experiments.

3.2 Conference MOS

Existing metrics fail to account for two important considerations when assessing the quality of a meeting hosted on a Voice Conferencing Bridge (VCB): (1) the voice quality perceived by a participant not only depends on the quality of its path to the VCB, but also on the number of *speakers* and the quality of the network paths between the speakers and the VCB; and (2) the set of speakers is likely to change on a frequent basis during the course of the meeting; therefore, quality measurements should periodically account for the up-to-date speaker set.

The following metric accounts for the above considerations:

$$cmos_p^t = \frac{\sum_{i=0}^{m} mos_{ip}}{m}$$

where $cmos_p^t$ is the *short-term* conference MOS for participant p for time period t, mos_{ip} is the MOS, computed using the E-model, on the path between speaker i and participant p traversing through the VCB v, and m is the number of speakers where $p \notin m$ if p is a speaker. For the computation of mos_{ip}, the path delay is given as $D = d_{iv} + d_{vp}$, where d_{iv} and d_{vp} are delays between i to v and v to p respectively. The path loss is given as $L = 1-(1-l_{iv})*(1-l_{vp})$ where l_{iv} and l_{vp} are the respective loss probabilities.

The parameter t controls the number of speakers that are included in the quality estimation. In the above equation, equal weight is given to each speaker in the time period t. If t is large, such weighting can skew the result. For example, if t is a minute,

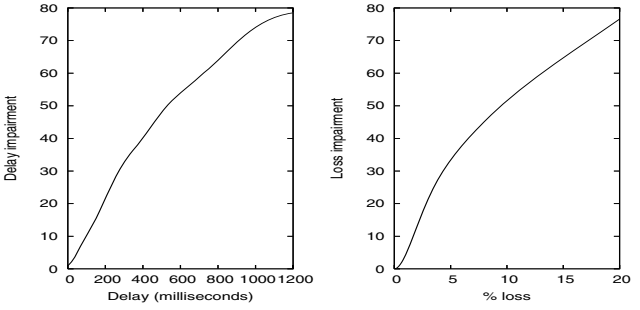

Fig. 2. Delay and loss impairment values

a speaker with a poor network connection who speaks for a very short time, say five seconds, in an otherwise good quality conversation, can cause the overall quality to be underestimated. We set t to be eight seconds in our implementation. This setting is long enough to reflect the time a person would take to form an opinion about the quality. Yet, it is short enough that averaging effects as noted above are minimized.

Long-term conference MOS is the mean of the short-term conference MOS values. Mean short-term MOS has been found to be a good predictor of long-term MOS [17].

4 Measurement Methodology

Our goal for the measurement study is to investigate if a dial-in conferencing solution is well suited for enterprise-grade audio conferencing on the Internet. In order to achieve this goal, we monitored and characterized the performance of the GoToMeeting conferencing solution using the Conference MOS metric. Our data set includes traces from end points around the world. Therefore, we believe that our results will likely be representative of other similar solutions.

4.1 VoIP Infrastructure

Our VoIP infrastructure consists of GoToMeeting [1], a collaboration application installed on End Points (EPs), and Voice Conferencing Bridges (VCBs) geographically distributed at various data centers. The GoToMeeting application allows users to collaborate online, give presentations and provide online trainings. A GoToMeeting user can participate in an audio conference either via VoIP, with a mic and headset, or dial in using a PSTN phone. PSTN gateways facilitate the mixing of PSTN and VoIP traffic.

The VCBs receive packetized audio from the EPs over RTP, perform audio mixing and transmit the mix to the EPs. The EPs initiate conferences on VCBs using SIP. VCB selection is performed using a propreitary algorithm. Closeness, in terms of network latency, to the organizer of a conference is one parameter in the algorithm. Note that a VCB located in a country will have connections from users outside that country if conference participants are situated abroad.

4.2 Data Sets

We collected five days of data from Aug 5-9, 2008 from a subset of VCBs that are part of our VoIP infrastructure. These VCBs are located in data centers on the east and west coasts of the United States. Our data set is made up of packet traces and VCB log files. The traces contain RTP headers and RTCP packets. We did not record the RTP payload in order to ensure user privacy. RTP headers have contributing source (CSRC) information, which gives the list of speakers encoded in the RTP packet. The speaker list is utilized to compute the Conference MOS. The RTCP packets carry bidirectional loss and delay information. The log files contain conference information, such as a conference's creation time and its lifetime.

Our data set contains 9394 VoIP flows from end points around the world. These flows connected to 7436 conferences. Given that a conference consists of multiple users, one would expect the number of VoIP flows to be substantially higher than conferences. In the case of these conferences, some users chose to call in via PSTN.

VoIP flows are encoded using *iSAC* [4], a wide-band codec (16Khz sampling rate) from Global IP Solutions. These flows totalled 231000 minutes of talk time. Figure 3 shows the arrival of VoIP connections and the creation of conferences per hour over the five day period. A conference is created on a VCB when a VoIP/PSTN flow arrives for that conference. PSTN arrivals are not shown on this graph. The arrivals follow a typical diurnal pattern with peaks during the day. The night-time load is about one third that of the day. As expected, less activity is seen over the weekends. The peak number of VoIP arrivals per hour in our data set is between 150 to 200 whereas the peak number of conferences created per hour is between 100 and 140. At certain times, such as seen in the friday section of the graph, the number of VoIP connections is only slightly higher than conferences, because some users in these conferences joined using their PSTN phones.

Figure 4 shows the lifetime of conferences and VoIP connections as a CDF. The median and average lifetimes of VoIP flows are 10 minutes and 24 minutes respectively, whereas the median and average lifetimes of conferences are approximately 30 minutes and 49 minutes respectively. The lifetime of a VoIP flow is typically shorter because a user might join or leave a conference at any time.

Birke et. al. [12] measured the average VoIP flow duration within an Italian ISP to be of much shorter duration (approximately 2 minutes). We attribute the difference to the application use-cases. The Italian ISP is mainly used for one-to-one social/business calls [12], whereas conferences in GoToMeeting involve users who collaborate online on a specific topic.

5 Traffic Analysis

Our goal is to study the MOS of conferences hosted on our VoIP infrastructure. We calculate a participant's conference MOS and *connection MOS* using the technique outlined in Section 3.2 and the E-model respectively. The Connection MOS is simply the MOS on the path between the participant and the Voice Conferencing Bridge (VCB). We calculate connection MOS because a participant's connection to the VCB is a key contributor to its conference MOS. Further, connection MOS is a good baseline for

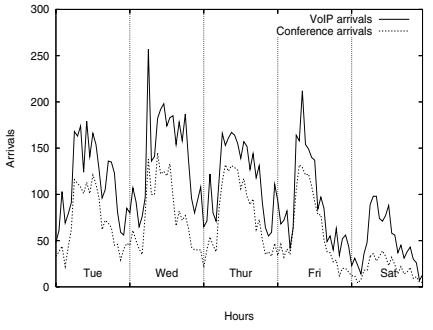

Fig. 3. Arrivals of VoIP connections and conferences

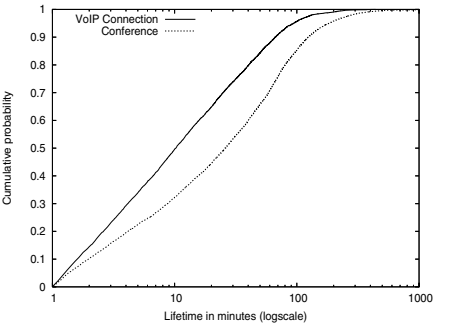

Fig. 4. Lifetime of VoIP connections and conferences

comparison purposes. Conference MOS and connection MOS higher than 3.5 is classified as good quality, whereas anything lower is undesirable.

We wrote a tool that computes these MOS values using: (1) conference details captured in log files; (2) loss and delay measurements in RTCP packets; and (3) speaker information from RTP packet headers.

5.1 Results

Figure 5 plots the average connection MOS value and average conference MOS value as a CDF. The average is computed over the lifetime of a participant's flow by taking the mean of the respective instantaneous MOS values, which are calculated every eight seconds (motivated in Section 3.2). Average MOS has been found to be a good predictor of long-term MOS [17]. This paper does not investigate variations in instantaneous MOS values. We plan to expore such variations as future work.

We make two observations from this figure. First, the median conference MOS is approximately 4.20, while the median connection MOS is about 4.30. Only 11.75% of participants have conference MOS less than 3.5. 11% of participants have connection MOS less than 3.5. *This result implies that a majority of participants, in general, experienced good quality for conferences hosted on the GoToMeeting VoIP infrastructure.*

Second, conference MOS appears lower than connection MOS. This is because the conference MOS depends not only on the participant's connection to the VCB, but also on the network connections of the speakers in the conference, one or many of which can adversely impact the participant's conference MOS.

In order to further explore the relationship between conference MOS and connection MOS, we plot in Figure 6 the conference and connection MOS values for hundred randomly selected participants. Points on the line $x = y$ indicate all connections with similar conference and connection MOS. Points above this line are cases with higher connection MOS than conference MOS. There are no points below this line because a participant's conference MOS can only be as good as its connection MOS. As can be seen in the graph, some participants have connection MOS higher than 3.5, but conference MOS less than 2.5. For such cases, the remaining participants in their respective

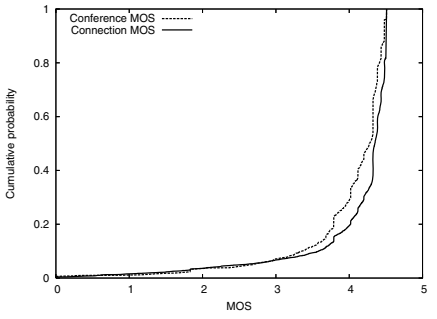

Fig. 5. Connection and conference MOS as a CDF

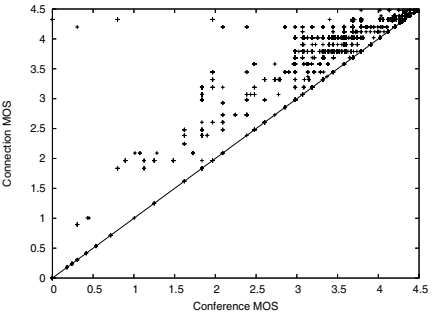

Fig. 6. Participant's connection and conference MOS as a scatter plot

conferences have likely very poor connections to the VCB and, as a consequence, impact the conference MOS in a severe manner.

These poor connections are a part of the 11% of participants with connection MOS less than 3.5 in Figure 5. Some of these connections could be from international participants. Figure 7 plots the CDF of the average connection MOS for all participants, but also segmented into participants within the US and ones outside the US. We observe that international participants have relatively poor quality connections compared to local participants. The median connection MOS for local participants is 4.39, while with international participants, it is slightly lower at 4.11. 16.74% of international participants had connection MOS less than 3.5 while only 9.19% of local participants had this low value. As such, the lower quality is to be expected given the transcontinental delay and loss on paths to the VCBs in the US. However, almost 10% of US participants also had connection MOS less than 3.5.

In order to understand why these connections had low MOS, in Figure 8, we plot the average delay and loss impairments as a scatter plot for all participants with connection MOS less than 3.5. The graph is divided into four quadrants. We observe that loss and delay contribute to MOS in varying proportions. For example, points in the second quadrant (Q2) indicate that high delay is the principal contributor to low MOS; points in quadrant four (Q4) show that high loss is the principal contributor, whereas points in the third quadrant (Q3) indicate that high loss and high delay contribute in approximately equal proportions; points in the first quadrant (Q1) indicate that although delay and loss were low individually, a combination of the two leads to bad audio quality. Our finding differs from that of other studies [12,14] which conclude that loss, not delay, is the main reason for poor quality. These studies were limited to backbone networks which are typically provisioned for high capacity; end-to-end paths are likely limited because of the "last-mile".

5.2 Discussion

Our results indicate that the Internet is suitable for enterprise-grade audio conferencing. While our results are specific to the VoIP infrastructure we monitored, we believe that our conclusion is applicable to other similar deployments.

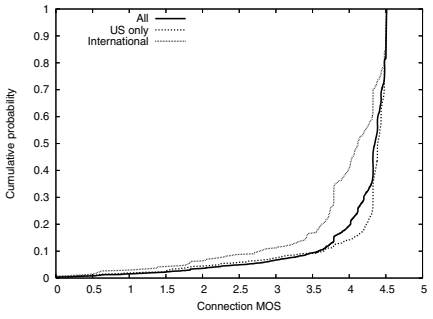

Fig. 7. Connection MOS for US and international participants

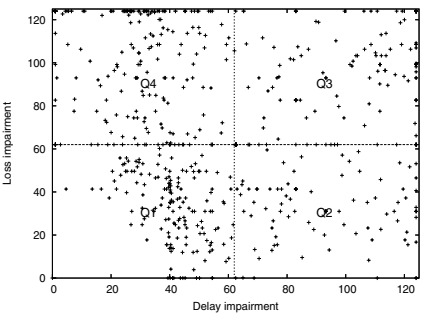

Fig. 8. Average impairment factors for participants with connection MOS less than 3.5

Participants with poor network connections will be the "weakest-links" in achieving good conference quality. Strategies to reduce their influence include the placement of VCBs in the proximity of such participants and to utilize sophisticated routing techniques [19]. For example, one strategy could be to deploy VCBs in an overlay [11] that covers hotspots around the world and to interconnect these VCBs with high bandwidth connections in order to mix participants that span hotspots. While such a strategy could alleviate some of the poor quality, it will likely fail to eliminate all such cases.

Therefore, we believe that in order to support enterprise-grade conferencing, a solution that allows users with poor network connections to dial in using their PSTN phones is a better alternative. The Internet backbone is provisioned for high capacity. It is conceivable that the "last-mile" is similarly provisioned in the near future. Until such time, enterprise-grade audio conferencing solutions should utilize a deployment strategy that combines VoIP and PSTN than a pure VoIP solution.

6 Conclusion

In this paper, we investigate if the Internet can support enterprise-grade audio conferencing. We use the proposed Conference MOS metric to study the performance of the GoToMeeting audio conferencing solution. Our results indicate that a majority of users utilizing this solution experienced good conference quality. A minority would have been better served if they had dialed in using PSTN phones. This leads us to conclude that a deployment strategy that supports VoIP and PSTN would serve users better than a VoIP only solution.

Conference MOS is useful for analyzing the quality of an audio conference. We believe it can help network operators monitor and troubleshoot their deployments. Furthermore, it can be used to proactively notify a user of his poor quality connection so that he can take corrective measures, such as calling in with a PSTN phone.

The results presented in this paper shed much needed light on the performance of audio conferencing in the Internet. As future work, we plan to expand our monitoring effort to include a larger set of conferencing bridges and VoIP flows. We also plan to study the transient behavior of VoIP flows and its impact on quality.

Acknowledgments

Albert Alexandrov, Bruce Brown, Robert Chalmers, Nitin Gupta, Alan Knight, Ashwin Sampath and Chris Scheiman provided valuable technical support for the paper. The Operations and IT teams helped with traffic collection and storage.

References

1. Citrix GoToMeeting, http://www.gotomeeting.com
2. Citrix GoToWebinar, http://www.gotowebinar.com/webinar/pre/reliability.tmpl
3. Frost Conferencing Research, http://www.frost.com/prod/servlet/svcg.pag/ITCF
4. Global IP Solutions iSAC, http://www.gipscorp.com/files/english/datasheets/iSAC.pdf
5. North American Audio Conferencing Services Markets, http://www.mindbranch.com/North-American-Audio-R1-6595/
6. Skype Conference Calling, http://www.skype.com/business/features/conferencecall/
7. Methods for Subjective Determination of Transmission Quality. In: ITU-T P.800 (1996)
8. E-model, a Computation Model for use in Transmission Planning. In: ITU-T G.107 (2002)
9. Mean Opinion Score (MOS) Terminology. In: ITU-T P.800.1 (2003)
10. Provisional Impairment Factor Framework for Wideband Speech Transmission. In: ITU-T G.107 Amendment 1 (2006)
11. Amir, Y., Danilov, C., Goose, S., Hedgvist, D., Terzis, A.: 1-800-OVERLAYS: Using Overlay Networks to Improve VoIP Quality. In: ACM NOSSDAV, Stevenson, WA (June 2005)
12. Birke, R., Mellia, M., Petracca, M.: Understanding VoIP from Backbone Measurements. In: IEEE Infocom, Anchorage, AL (May 2007)
13. Bolot, J., Crepin, H., Vega-Garcia, A.: Analysis of Audio Packet Loss in the Internet. In: ACM NOSSDAV, Durham, NH (April 1995)
14. Boutremans, C., Iannaccone, G., Diot, C.: Impact of Link Failures on VoIP Performance. In: ACM NOSSDAV, Miami, FL (May 2002)
15. Cole, R., Rosenbluth, J.: Voice over IP Performance Monitoring. In: ACM SIGCOMM Computer Communication Review, vol. 31, pp. 9–24 (April 2001)
16. Markopoulou, A., Tobagi, F., Karam, M.: Assessment of VoIP Quality over Internet Backbones. In: IEEE Infocom, New York, NY (June 2002)
17. Rix, A., Beerends, J., Kim, D., Kroon, P., Ghitza, O.: Objective Assessment of Speech and Audio Quality — Technology and Applications. In: IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, pp. 1890–1901 (November 2006)
18. Rosenberg, J., Schulzrinne, H.: Models for Multi Party Conferencing in SIP. Internet Engineering Task Force (IETF), draft-ietf-sipping-conferencing-models.txt (January 2003)
19. Tao, S., Xu, K., Estepa, A., Fei, T., Gao, L., Guerin, R., Kurose, J., Towsley, D., Zhang, Z.: Improving VoIP Quality Through Path Switching. In: IEEE Infocom, Miami, FL (March 2005)