Assessing the Quality of Voice Communications over Internet Backbones

Athina P. Markopoulou, *Member, IEEE,* Fouad A. Tobagi, *Fellow, IEEE,* Mansour J. Karam, *Member, IEEE*

***Abstract*— As the Internet evolves into a ubiquitous communication infrastructure and provides various services including telephony, it will be expected to meet the quality standards achieved in the public switched telephone network. Our objective in this paper is to assess to what extent today's Internet meets this expectation. Our assessment is based on delay and loss measurements taken over wide-area backbone networks and uses subjective voice quality measures capturing the various impairments incurred. First, we compile the results of various studies into a single model for assessing the VoIP quality. Then, we identify different types of typical Internet paths and we study their VoIP performance. For each type of path, we identify those characteristics that affect the VoIP perceived quality. Such characteristics include the network loss in a path and the delay variability that should be appropriately handled by the playout scheduling at the receiver. Our findings indicate that although voice services can be adequately provided by some ISPs, a significant number of Internet backbone paths lead to poor performance.**

I. Introduction

The Internet is evolving into a universal communication network and it is contemplated that it will carry all types of traffic, including voice, video and data. Among them, telephony is an application of great importance, particularly because of the significant revenue it can generate. In order for the Internet to constitute an attractive alternative to the traditional Public Switched Telephone Network (PSTN), it must provide high quality "Voice over IP" (VoIP) services. Our objective in this paper is to assess to what extent today's Internet meets these high quality expectations. In the process, we identify those aspects that may lead to poor voice quality.

Our approach in addressing this problem has three main characteristics. First, we use delay and loss measurements collected by means of probes sent between measurement facilities at five different US cities, connected to the backbone networks of seven different providers. These measurements correspond to a large number of paths (43 in total), (7) different ISPs and a long period of time (17 days); thus they are representative enough of Internet backbones. Second, we use subjective voice quality measures that take into account the various impairments. For this purpose, we compile into a single model the results of several studies conducted for specific voice impairments. Furthermore, we use a methodology for rating telephone calls that takes into account the variability of the transmission impairments with time. Third, we take into account the effect of different components of a VoIP system and, in particular, we consider the playout scheduling.

This study is limited to Internet backbones; nevertheless the results obtained are very useful. Backbone networks are an important part of the end-to-end path for both (i) long distance VoIP calls and (ii) calls that are serviced by a combination of a switched telephone network in the local area and Internet backbones for the long haul. Although backbone networks are known to be sufficiently provisioned to cause negligible degradation on data traffic, our study shows that a large number of the Internet paths exhibited poor VoIP performance, mainly due to high delay and high delay variability. Furthermore, if stringent communication requirements (such as interactivity levels suited for business conversations) are imposed, these paths become totally unacceptable for telephony use. Paths with low delay and low delay variability exhibit in general excellent performance. However, even those networks occasionally experience long periods of loss that can affect voice communications.

As far as the VoIP system is concerned, we consider both fixed and adaptive playout scheduling schemes. In both cases, we identify a tradeoff in quality between packet loss (due to late arrival) and increased delay in the playout buffer; this allows one to determine an appropriate choice of playout delay that takes into account this trade-off. With regards to adaptive playout schemes, we find that the practicality of adaptive schemes is hindered by the sensitivity of their performance to the proper tuning of their parameters and by the strong dependence of the optimum parameter values to the specific delay characteristics of each path.

The paper is organized as follows. In Section II, we describe the components of a VoIP system and the impairments they introduce. In Section III, we present the quality measures used for assessing the impairments in the network and our approach for rating a telephone call. In Section IV, we describe the probe measurements and their delay and loss characteristics. In Section V we apply our methodology to the traces, obtain and discuss numerical results pertaining to phonecalls quality. Section VI concludes the paper.

II. VoIP system

VoIP refers to voice communication over IP data networks. In this section, we identify and describe the various components of a VoIP system, shown in Fig. 1, and the impairments they introduce in voice communications.

Athina Markopoulou is with SprintLabs, Burlingame, CA. (Email: amarko@stanfordalumni.org.) This study was conducted when she was a Ph. D. student in Electrical Engineering at Stanford University. Fouad Tobagi is with the Department of Electrical Engineering, Stanford University, Stanford, CA. (Email: tobagi@stanford.edu.) Mansour Karam is with RouteScience Technologies Inc., San Mateo, CA. (Email: mans@routescience.com)

This work has been partially supported by RouteScience Technologies, Inc. It has also been made possible by a university equipment grant from IBM. RouteScience also provided the measurements used in this study.

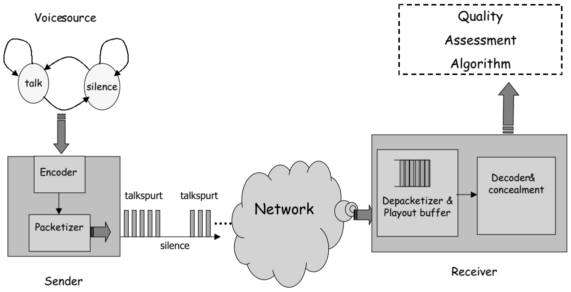


Fig. 1. VoIP System

A. Components of the VoIP system

Speech is an analog signal that varies slowly in time (with bandwidth not exceeding 4KHz). The speech signal alternates between *talkspurts and silence periods*, which are typically considered to be exponentially distributed; Sriram and Whitt in [2] used mean 352 ms for talkspurts and 650 ms for silences. For the purpose of transmission over networks, the speech analog signal is converted into a digital signal at the sender; the reverse process is performed at the receiver. In an interactive conversation, the participating parties switch turns in taking the sender and receiver roles.

There are many *encoding schemes* that have been developed and standardized by the ITU. The simplest is the sample-based G.711 which uses Pulse Code Modulation (PCM) and produces a digitized signal of 64 Kbps. CELP-based encoders provide rate reduction (i.e. 8 Kbps for G.729, 5.3 and 6.4 Kbps for G.723.1) at the expense of lower quality and additional complexity and encoding delay, [5].

Further reduction in the data rate can be achieved using *Voice Activity Detection* (*VAD*). The resulting talkspurts and silences have been shown to also follow roughly exponential distributions, with a mean that depends on the specifics of the VAD algorithm. For example, VAD tends to elongate talkspurts by a period, called the hangover time, in order to prevent speech clipping. First in [3], Brady used a long hangover and reported exponential talkspurts and silences with mean 1.2 and 1.8 sec respectively. A nice review together with a discussion on the on/off voice patterns resulting from modern voice coders can be found in [4]. In general, small hangover time results in small talkspurts and silences (200-400ms and 500-700ms on average respectively) while a large hangover results in larger durations (around 1-2 sec).

The encoded speech is then *packetized* into packets of equal size. Each such packet includes the headers at the various protocol layers (namely, the RTP (12B), UDP (8B) and IP (20B) header as well as Data Link Layer headers) and the payload comprising the encoded speech for a certain duration.

As the voice packets are sent over an *IP network*, they incur variable delay and possibly loss. In order to provide a smooth playout at the receiver despite the variability in delay, a *playout buffer* is used. Packets are held for a later playout time in order to ensure that there are enough packets buffered to be played out continuously. Any packet arriving after its scheduled playout time is discarded. There are two types of playout algorithms: fixed and adaptive.

A fixed playout scheme schedules the playout of packets so that the end-to-end delay p (including both network and buffering) is the same for all packets. It is important to select the value p so as to maximize the quality of voice communications. Indeed, a large buffering delay decreases packet loss due to late arrivals but hinders interactivity between the communicating parties. Conversely, smaller buffering delay improves interactivity but causes higher packet loss in the playout buffer and degrades the quality of speech. The value of fixed end-to-end delay should be chosen based on knowledge of the delay in the network. However, such an assessment may not always be possible or the statistics of delay may change with time. For these reasons, adaptive playout is considered.

Extensive work has been conducted on adaptive playout schemes, that monitor the network delay and its variations and adjust accordingly the playout time of voice packets. In [6], a number of algorithms were proposed, that consist of monitoring network delays, estimating the delay d_{av} and delay variation v using moving averages, adapting the playout time to $p = d_{av} + 4v$ at the beginning of each talkspurt but keeping it constant throughout a talkspurt. In addition, it was also proposed to detect delay spikes and adapt p faster during the spike periods. The scheme proposed in [7] improved over the previous one by using a delay percentile rather than a moving average, to estimate the network delay; the improvement achieved came at the expense of increased state and processing. In [8], the prediction of network delays was further improved by minimizing the normalized mean square prediction error. A second group of playout algorithms adapt the value of delay p on a packet-per-packet, instead of a talkspurt-per-talkspurt, basis and thus allows for capturing delay variations even within a single talkspurt. The scheme in [9] was the first to follow this approach, but it did not take into account the voice signal itself and the pitch of the speech signal was affected by the playout speed. The work in [10] used a time scale modification technique to preserve the pitch. It is interesting to note, that similar issues for maintaining smoothness for voice, have also been addressed in different contexts; the work in [11], dealt with workstation scheduling for real audio.

The content of the received voice packets is delivered to the *decode*r which reconstructs the speech signal. Decoders may implement *packet loss concealment (PLC)* methods that produce replacement for lost data packets, [12]. Simple PLC schemes simply insert silence, noise or a previously received packet. More sophisticated schemes attempt to find a suitable replacement based on the characteristics of the speech signal in the neighborhood of the lost packet(s). They may be interpolation-based (and try to match the waveform surrounding the lost portion) or regeneration-based (by being aware of the structure of the codec and exploiting the state of the decoder).

Although not evaluated in our study, it is worth mentioning that audio tools may include additional error resiliency mechanisms, [13]. These may include transmission of layered or redundant (FEC) audio, interleaving frames in packetization, retransmissions (if the end-to-end delay budget permits it), feedback to signal the sender to switch rate or encoder.

B. Voice impairments in networks

The quality of voice communication is affected by a number of factors. First, voice encoding affects the quality of speech. Second, in the case of VoIP, the transmission of packet voice over a network is subjected to packet loss in network elements causing degradation in the quality of voice at the receiver. Further loss is incurred in the playout buffer at the receiver, caused by network delay jitter. Third, the interactivity between the communicating parties is affected by the delays incurred in the network. Indeed, a large delay may lead to “collisions”, whereby participants talk at the same time. To avoid such collisions, the participants can talk in turns, and thus take longer to complete the conversation. To achieve a good level of interactivity, the end-to-end delay (from mouth-to-ear) should be maintained below a certain maximum delay, typically on the order of 100 ms. Longer delays become noticeable, and the longer the end-to-end delay is, the lower is the degree of interactivity. The end-to-end delay encompasses: (i) the delay incurred in encoding (referred to as algorithmic delay), (ii) the delay incurred in packetization (function of the amount of speech data included in a packet), (iii) the delay incurred in the path from the sender to the receiver (propagation time, transmission time over network links, and queuing delays in network elements), (iv) the delay incurred in the playout buffer, and finally (v) the delay incurred in the decoder (usually negligible). Fourth, the presence of echo in various situations could be a major source of quality degradation in voice communication, [14]. One cause of echo is the reflection of signals at the four-to-two wire hybrids; this type of echo is present when a voice call involves a combination of a VoIP segment in the Internet and a circuit segment in the switched telephone network. Another cause of echo is in PC-based phones (typically equipped with a microphone and loud-speakers), whereby the microphone at the remote end picks up the voice played on the loud-speakers and echoes it back to the speaker. Voice echo is not perceptible if the end-to-end delay is very short (below 10 ms.) However it becomes annoying for longer end-to-end delays. The effect of echo can be mitigated by cancellation placed close to the cause of echo.

III. Assessment of Voice Communications in Packet Networks

In order to assess the quality of voice communication in the presence of impairments, it is crucial to study the individual as well as collective effects of the impairments, and produce quantitative measures that reflect the subjective rating that listeners would give. This subjective quality measure is also referred to as Mean Opinion Score (MOS) and is given on a scale of 1 to 5 , as defined in [15]. Fig. 2 shows the mapping of MOS to user satisfaction, as reported in [16], [17]. A MOS rating above 4.0 matches the level of quality available in the current Public Switched Telephone Network; a rating above 4.3 corresponds to the best quality whereby users are very satisfied; and a rating between 4.0 and 4.3 corresponds to a high quality level, whereby users are satisfied. A MOS rating between 3.6 and 4.0 corresponds to a medium quality level whereby some users are dissatisfied. A MOS rating in the range between 3.1 and 3.6 corresponds to a low level of quality whereby many users are dissatisfied. A MOS rating in the range between 2.6 and 3.1, the level of quality is poor whereby nearly all users are dissatisfied. And finally, a MOS below 2.6 is not recommended.

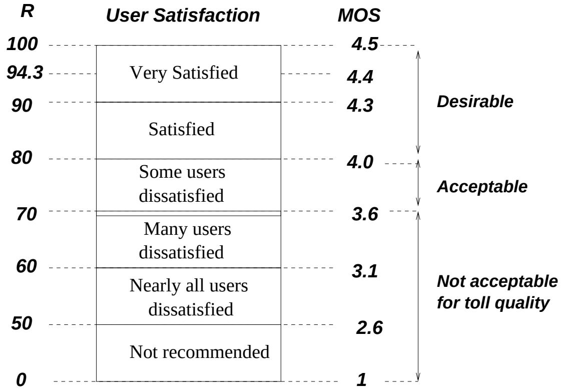


Fig. 2. Mean Opinion Score and its relation to user satisfaction and the Emodel rating R, according to G.107/Annex B and G.109.

Numerous studies have been conducted to assess the effect on voice quality of various impairments under various conditions. Some of them have also been compiled into reports and recommendations published by standards organizations. In this section, we give an overview of some of these studies, we summarize the results obtained therein in order to complete the evaluation space as well as to confirm their consistency, and we describe our approach for VoIP quality assessment.

Before proceeding, it worths commenting on the validity of MOS itself as a technique. MOS is valuable in that it addresses the human perceived experience, which the ultimate measure of interest. However, it should be used carefully: it is closely tied to the conditions and the specific goals of each experiment and thus difficult to generalize. Furthermore, current subjective testing methods are known to have a number of weaknesses, a review of which can be found in [18]. In our study, we use the results of classic subjective tests, as our starting point, in order to assess the loss and delay impairments in the Internet. We make every possible attempt to apply them under the same conditions they have been obtained. The steps we take will be similar, even if more accurate subjective tests become available in the future. As we discuss in the section III.E and III.F, our methodology successfully addresses many of the issues mentioned in [18], including the time varying nature of Internet impairments, the different aspects of voice quality and the recency effect. In addition, we supplement our assessment in terms of MOS with raw loss and delay measurements.

TABLE I

STANDARD ENCODERS AND THEIR CHARACTERISTICS

Standard	Codec	Rate	Frame	Lookahead	MOS
	type	(Kbps)	(ms)	(ms)	intr.
G.711	PCM	64		0	4.43
G.729	CS-ACELP	8	10	5	4.18
G.723.1	ACELP	5.3	30	7.5	3.83
G.723.1	MP-MLQ	6.3	30	7.5	4.00

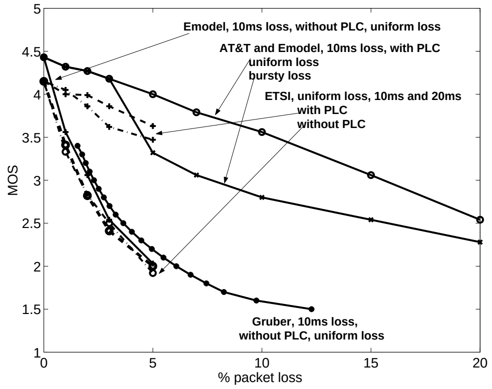


Fig. 3. G.711 quality under packet loss conditions as reported by various studies. The packet size is 10ms in all cases.

A. Degradation in speech quality

The degradation in speech quality due to the encoder is summarized in Table I. The quality after compression, without considering the effect of packet loss, is often referred to as intrinsic quality MOS_{intr}. As can be seen from the table, lower rate encoders result in lower MOS values.

We now address the effect of packet loss, which results in speech clipping, to voice subjective quality. Figures 3, 4 and 5 summarize the results of various studies for G.711, G.729 and G.723 respectively.

Among the earliest work in this area is that by Gruber and Strawczynski back in 1985 [19]. They addressed the effects of speech clipping and variable speech burst delays incurred in dynamically managed voice systems, using PCM encoding and speech activity detection. Of relevance to our study are the results pertaining to speech clipping, whereby speech clips of fixed durations are uniformly distributed across time, with loss rates ranging from 0 to 20%. The results for 10ms loss duration are plotted in Figure 3. Concealment was not considered, as such techniques did not exist for PCM at that time.

The benefit of error concealment has been studied for G.711 under both uniform and bursty loss conditions, considering packets containing 10 ms of speech, and packet loss rates up to 20%, [20]. Results pertaining to G.728 and G.729 with their standard packet loss concealment and loss rates only up to 5%, can also be found in [20]. Also, the study by Perkins et al., in [21], characterized the subjective performance of G.729 in wireless and wired networks under various conditions, including channel bit errors, environmental noise, and frame erasures (up to 3% loss rate); the results agree with those published in [20].

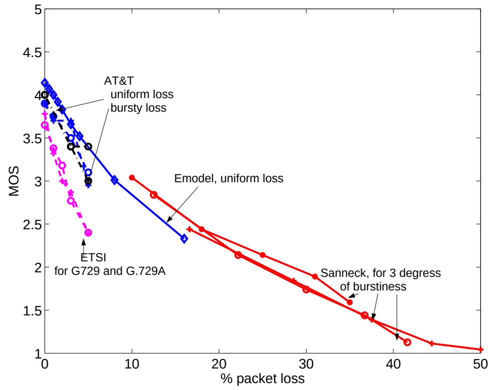


Fig. 4. G.729 quality under packet loss conditions, as reported by various studies. The packet size is $20ms$ and packet loss concealment was used in all cases.

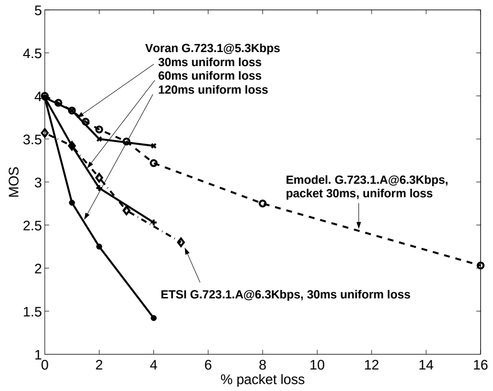


Fig. 5. G.723.1 quality under packet loss conditions, as reported by various studies. The packet size is 30ms and packet loss concealment was used in all cases.

Sanneck et. al. [22] also assessed the effect of loss on G.711 and G.729, using a Gilbert model to simulate bursty loss conditions, for loss rates up to 50%. The evaluation for G.711 was performed with and without loss concealment. The evaluation for G.729 was performed with the standard concealment as well as with newly proposed concealment schemes. The results pertaining to G.729 with the standard PLC are shown in Fig. 4. G.729 appears to be less sensitive than G.711 to the degree of burstiness, which is attributed to the robustness of the loss concealment method in G.729.

Voran studied the effect of loss on voice encoded with G.723.1 with VAD at 5.3 Kbps, [23]. Uniform loss at rates 0-4% were considered, with speech loss durations equal to a single frame (30 ms), two consecutive frames (60 ms) and 4 consecutive frames (120 ms). The results are shown in Fig. 5. The higher initial MOS compared to the other studies pertaining to G.723 is due to the different encoding schemes considered by the Emodel (MP-MLQ) and by [23](ACELP).

Several among the above-mentioned studies and several other studies have been compiled into documents published by the ITU, [16], [24], [17], [25] and ETSI, [26]. Work initiated at ETSI, resulted in the development of a group of standards by ITU-T in 1996, known as the “Emodel”, [16], [24], [25]. Recommendation G.113 [25] collected results from studies that applied packet loss to G.711, G.723 and G.729. These are shown in Fig. 3, 4 and 5 respectively using the label “Emodel”.

The results for G.711 with packet loss concealment, for both uniform and bursty loss, are taken from [20]. In addition, a curve for G.711 without packet loss concealment is provided, which agrees with the results for $10ms$ packet obtained by [19], see Fig. 3. Work along the same lines is still ongoing in ETSI and a recent technical report is [26], dated in 2000; we plot the results contributed by Nortel Networks for G.711, for G.729 and G.729A, and for G.723.1, packet sizes of $10ms$, $20ms$ and $30ms$ and using PLC, in Fig. 3, 4 and 5 respectively, under the label "ETSI".

Discussion. One can make the following observations, looking at the above results. The encoding scheme affects the intrinsic MOS quality (before any loss) and therefore the maximum allowed packet loss in the network to sustain acceptable quality. However, the slope of MOS degradation seems to be the same for comparable experiments. For packet loss concealment and $10ms$ loss duration, MOS drops by roughly $1-1.5$ unit every 10% of packet loss; in experiments without packet loss concealment, MOS drops much faster, by roughly 1 unit every 1% of packet loss. Larger loss durations result in increased degradation. Finally, bursty loss seems to affect the resilience of G.711 but not that of G.729.

B. Loss of interactivity

In 1991, a study by NTT assessed the loss of interactivity due to large end-to-end delay, in echo free telephone circuits, [28]. Various amounts of delay was introduced and Mean Opinion Scores, conversational efficiency and detectability thresholds were obtained, using groups of subjects varying with various degrees of expertise. Six conversational modes ("tasks") were considered, each having a different switching speed between the communicating parties and thus a different sensitivity to delay. The most stringent task is Task 1, where people take turns reading random numbers as quickly as possible. On the other extreme, Task 6 is the most relaxed type, free conversation.

Recommendation G.114, published in 2000, also focused on the loss of interactivity due to delay, assuming echo free environments, [27]. Traditionally, a one way delay up to $400ms$ was considered acceptable for planning purposes; recommendation G.114 emphasized that this is not the case for highly interactive conversations and declares $150ms$ acceptable for most applications in echo free environments. G.114/Annex A estimates for the delay incurred in various components of circuit and IP networks are provided. In Annex B, results from the above-mentioned [28] and other similar studies are collected.

The Emodel[1] standards also provide a formula for calculating the loss of interactivity as function of the one way delay, in the absence of echo, [16]. The degradation in MOS as delay increases, as reported by all three sources, is shown in Fig. 6.

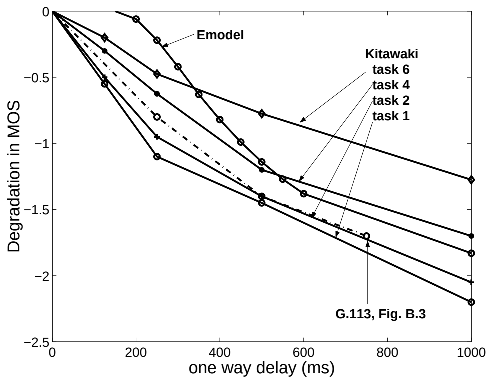


Fig. 6. Loss of interactivity due to one way delay in echo free environments, as reported by various studies.

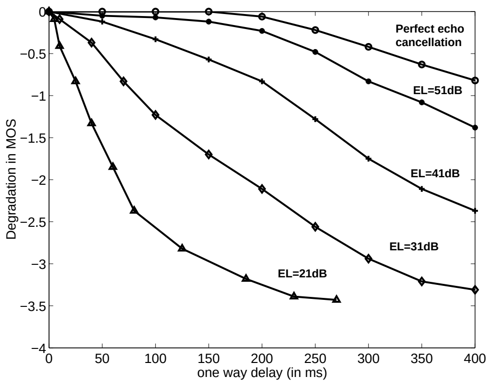


Fig. 7. Degradation in MOS due to echo, according to the Emodel (G.107).

C. Echo impairment

As discussed in Section II-B, echo can cause major quality degradation, if it is not adequately canceled. Its effect is amplified by large delays. The Emodel provides formulas that allow to calculate the impairment due to talker ($I_{dte}(m2e, EL_2)$)and listener ($I_{dle}(m2e, EL_1)$) echo respectively, given some transmission parameters. $m2e$ stands for the one way or "mouth to ear" delay; EL_1, EL_2 are the echo losses in dB at the points of reflection and their value depends on the echo cancellation used. $EL = \infty$ (infinite echo loss) corresponds to perfect echo cancellation. $EL = 51\,dB$ corresponds to a simple yet efficient echo controller. Fig. 7 shows the degradation in MOS, due to the combined talker and listener echo on the path.

D. Using the Emodel to combine all impairments

As mentioned above, the Emodel started as a study by ETSI and got standardized by ITU-T in [16][24][25]. Comprehensive studies of the Emodel can be found in [29] and [30]. It is a computational model that uses transmission parameters to predict the subjective quality of voice quality. It gives an overall rating R for the quality of a call, on a scale from 0 to 100, whose translation to quality and MOS is shown in Fig. 2. The Emodel combines different impairments based on the principle that the perceived effect of impairments is additive, when converted to the appropriate psycho-acoustic scale (R).

$$R = (R_o - I_s) - I_d - I_e + A \quad \textbf{\textit{(1)}}$$

The details of equation (1) are as follows. R_0 is the basic signal-to-noise ratio based on send, receive loudness, electrical and background noise. I_s captures impairments that happen simultaneously with the voice signal, such as sidetone and PCM

[1] The Emodel [16] (and later studies based on it such as that of Cole et.al., [29]) are lenient in the sense that they predict no degradation for delay below $150ms$. A possible explanation is that the Emodel curve does not take into account the aspect of the different conversational modes (or tasks) and the expertise of the subjects that participated in the subjective experiments.

quantizing distortion. Both R_0 and I_s terms are intrinsic to the transmitted voice signal itself and do not depend on the transmission over the network. Thus, they are irrelevant for the purpose of comparing VoIP to PSTN calls. I_d and I_e capture the degradation in quality due to delay related impairments (loss of interactivity and echo) and distortion of the speech signal (due to encoding and packet loss), respectively. A is the advantage factor; it accounts for lenient users, who accept some degradation in quality in return for the ease of access, e.g. when using cellular or satellite phone. For the purpose of comparison to PSTN calls, this factor is set to 0.

The Emodel is important in our study for two reasons. First, it quantifies the MOS degradation due to delay (I_d) and loss (I_e) impairments. In addition, the Emodel models the effect of noise and other speech related impairments, thus allowing us to take them into account without going into detail. Second and most important, the Emodel combines all the impairments into a single rating, using additivity in the appropriate scale R.

In summary, we obtain a MOS rating as follows. First, we assess the degradation in speech quality (at the encoder and due to packet loss in the network and in the playout buffer) using the curves for G.711, G.729 and G.723 in Figures 3, 4 and 5, respectively. We are particularly interested in the bursty loss which is the case in the Internet traces. In the Emodel terminology, this first step means that we calculate the I_e factor. Second, we assess the loss of interactivity using the NTT study and the strict (task 1) and lenient (task 6 or free conversation) tasks in Fig. 6. We assess the degradation due to echo in the path, using the Emodel curves in Fig. 7. In the Emodel terminology, this second step means that we calculate the I_d factor as $I_d = I_{d,echo}(m2e, EL) + I_{d,interactivity}(m2e)$. Third, we calculate the overall rating R from (1) and we translate it to MOS. In the rest of the paper, we present results in terms of MOS but the underlying calculations are in the R scale.Throughout the paper, we use interchangeably I_e and "degradation in MOS due to speech distortion", I_d and "degradation in MOS due to delay".

E. Applying the above data to Internet Traces

To appropriately use the above data to assess the performance of Internet traces, we have to make sure that we apply them for the same conditions under which the subjective results have been obtained. There are some important conditions underlying those experiments. First, the durations of speech samples used are carefully chosen. (ITU-T recommendation P.800 states that speech samples in the order of 2-3 seconds must be used to assess subjective speech quality, [15]. Conversational tests in the order of 1 minute have been used in [28] and in ITU-T G.114 to assess interactivity.) Second, the loss pattern was considered uniform in all but one experiment (which considered bursty loss up to 100 ms). However, there is no guarantee that these assumptions hold for Internet packet loss and after applying playout buffering. Third, the impairment remains the same throughout the experiment. One cannot apply the I_d and I_e curves to evaluate phonecalls lasting several minutes, during which impairments vary considerably.

A natural approach to address the first and third considerations, is to divide the call duration into *fixed time intervals* and assess the quality of each interval independently. Appropriate interval durations could be those used in the experiments or the talkspurt durations. The second consideration has to do with the burstiness in loss. As subjective results for long and arbitrarily bursty loss durations (which is the case in the Internet) are not available, we consider that any performance evaluation in terms of MOS should be supplemented with statistics about the loss durations themselves. In particular, loss durations above 100ms are difficult to conceal at the receiver, lead to loss of entire phonemes and they should be reported as glitches.

An attempt to address together the burstiness and the non-stationarity of Internet impairments is the one proposed in [31].They defined high and low loss periods or variable durations, called "bursts" and "gaps" respectively.[2] The use of variable intervals appropriately addresses the burstiness in the following ways. First, the loss during gaps is enforced to be uniform by the definition of a gap. During burst periods, we use the I_e curves for bursty loss. Second, by dynamically partitioning each trace into its own gaps and bursts, we emphasize the periods of high loss, as opposed to calculating the loss rates over arbitrarily long intervals and smoothing them out. Due to real-time processing constraints in a commercial system, [31] made some computational simplifications: tracking an average gap and burst instead of the actual values. In our offline analysis, we use the idea of bursts and gaps but we avoid these simplifications.

F. Assessing Phonecalls.

Apart from assessing short intervals, we would also like to simulate the rating that a user would give after talking on the phone for several minutes. Such a duration consists of multiple short intervals.

Independent MOS rating of each short interval t has been shown to correlate well with the continuous instantaneous rating of the call, [33]. Evaluating each interval leads to transitions between plateaus of quality, as represented by the dashed line in Fig. 8. However, transitions between periods of high and low loss are perceived with some delay by the listener. For example in Fig. 8, a human would perceive the changes in quality according to the smooth solid instead of the step-like dashed line. This effect is known as *recency effect*. The instantaneously perceived I_e is considered by [31] to converge toward the $I_e(loss)$ for a gap or burst, following an exponential curve with time constants $T_{bad} = 5\,sec$ for the high loss and $T_{good} = 15\,sec$ for the low loss periods. The constants are based on the study in [32].

The last step is to compute the overall rating at the end of a call, based on the instantaneously perceived quality during the

[2] If the number of consecutive received packets between two successive losses is less than a minimum value g_{min}, then the sequence of the two lost packets and the intervening received packets is regarded as part of a burst; otherwise, part of a gap. The choice of g_{min} becomes then important. At one extreme a small g_{min} would give small burst durations with high packet loss rate; on the other hand, a large g_{min} would group neighboring losses into one burst with smaller loss rate but averaged over a larger period of time. As the loss in a gap is $(100/gmin)\%$, we choose $g_{min} \geq 1sec$ that leads to $\leq 1\%$ loss in gaps and to meaningful durations in the order of a few seconds.

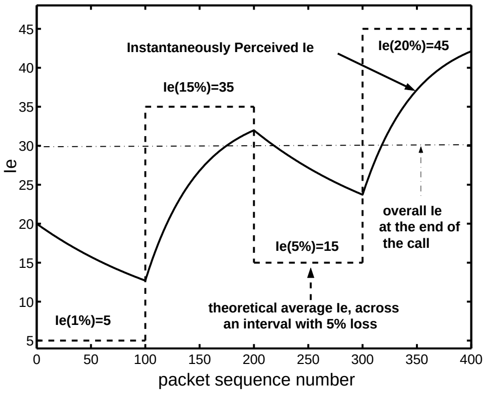


Fig. 8. Transitions between periods of high and low loss. Theoretical vs. instantaneously perceived I_e (i.e. MOS degradation due to loss).

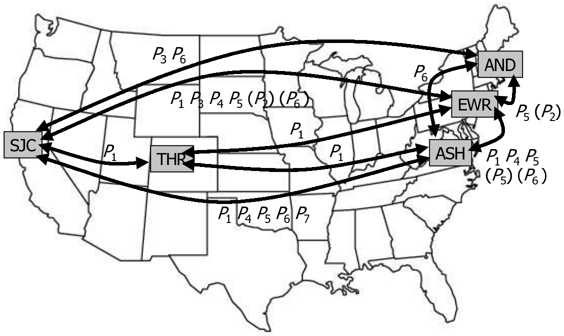


Fig. 9. Probes measurements

call. It has been shown in [33] that, at a first approximation, the overall rating is the time average of the instantaneously perceived MOS. In [31], the final rating is further adjusted to include the effect of the last significant burst and had good correlation with subjective results, [34], [35]. Notice however, that an individual might forget some bad moments in the middle of the call, that a network provider might be interested in monitoring and eliminating. To highlight these bad moments, we also report the worst quality experienced during a call.

In summary, we use variable bursts and gaps, the recency effect, and the overall rating at the end of a phonecall.

IV. Internet Measurements

In this section we describe the measurement experiment and the main delay and loss characteristics observed over the backbone networks of 7 Internet providers in continental U.S. An extensive characterization and modeling of these measurements can be found in our follow-up work in [36], [37].

A. *Related Work*

There has been extensive work on measurements and characterization of delay and loss in the Internet. This research topic continuously evolves along with the evolution of the network and the applications. Of interest to this study are delay and loss measurements over the public Internet, and backbone networks in particular, with respect to speech and multimedia transmission.

In the early 90s, Bolot et. al. sent audio traffic and measured the delay and loss incurred. In [38] and by the same authors, delay variability was found to have the form of spikes and was modeled it as the result of multiplexing the audio flow with an Internet interfering flow. In [39], the audio loss process was found to consist mostly of isolated packets. This is not necessarily the case neither in today's Internet nor in our measurements, [36], [37]. In [40], multicast measurements were used to study the temporal and spatial correlation of packet loss in the MBONE, exploiting the multicast tree topology. A recent study, [41], conducted a large scale experiment where they streamed MPEG-4 low rate video to clients located in more than 600 cities and provided statistics for the quality of the streaming sessions. Finally, another relevant recent study is [42], which focused on link failures on Sprint's backbone network and their effect on VoIP quality. The same group followed up on this work with [43]; they measured and characterized link failures which result in routing reconfigurations, possible service disruption and packet loss similar to the ones we observe.

B. *Probe Measurements*

Our study is based on delay and loss measurements provided by RouteScience Inc. Probes were sent by and collected at measurement facilities in 5 major US cities: San Jose in California (SJC), Ashburn in Virginia (ASH), Newark in New Jersey (EWR), Thornton in Colorado (THR) and Andover in Massachusetts (AND). 43 paths in total were used, obtained from seven different providers, which we refer to as $P_1, P_2, ..., P_7$ for anonymity purposes. The measurement setup is shown in Fig. 9. E.g. the bidirectional arrow drawn between SJC and AND means that measurements were collected from SJC to AND and from AND to SJC using providers P_3 and P_6. All paths are backbone paths, connected to the measurement facilities through either T3 or T1 links. Paths for all providers are two ways, except for those shown in parenthesis.

The probes were 50 Bytes each and were sent every $10\,ms$ [3] from Tuesday 2001/06/26 19:22:00 until Friday 2001/06/29 00:50:00 UTC.[4] GPS was used to synchronize senders and receivers and the network delays were inferred by subtracting the sender's from the receiver's timestamp. The load generated by the probes was insignificant and did not affect the delay and loss characteristics of the networks.

[3] By taking into account the providers' access bandwidths we are able to compute the transmission time and infer delays for any voice packet size from the probe delays. The 10ms sending interval is small enough to simulate the highest rate a VoIP encoder/packetizer might send packets at. By appropriately omitting probes we can simulate lower packet rates or silence periods. For example, by omitting 100 consecutive probes, we simulate a silence period of 1 sec. Also, by omitting every other probe packet, we can simulate voice packets sent every 20ms.

[4] We have also studied a similar data set, also collected by RouteScience, for 14 days (from 04:53:08 on 12/1/2000 until 23:59:59 UTC on 12/14/2000) using the same providers and three of the measurement facilities, namely SJC, EWR and ASH. The advantage of the earlier over the current measurement set is that it covers a longer time period. Its main drawback is that probes were sent at 100ms intervals, which are larger than those used by VoIP encoders. All the results we present are based on the current, fine granular data set. In the context of this paper, the earlier set of measurements was only useful to validate that our current findings, are true over a longer time period.

C. Observations on the Traces

1) Network Loss: Let us first discuss the loss events found in the measurements. Only one out of the 43 paths had consistently no loss during 2.5 days. All the other paths incurred loss with characteristics that vary among different providers and sometimes also between paths of the same provider.

For four paths, belonging to the same provider (P_3), single packets were lost regularly, at 0.2% rate and for the entire measurement period. For the remaining 38 paths, loss was concentrated over relatively short periods of time, at rates ranging from 10 to 100%. However, averaged over the entire measurement period, loss appears to be low: no more than 0.26% of all packets are lost on any path.

In addition to the loss rate, of interest to voice applications is the loss duration. We define as loss duration the period of time during which all probes are lost; this would result to a voice segment of the same duration being lost. Loss durations varied from 10ms (1 voice packet lost) up to 167 sec! Six out of seven providers experienced particularly long loss durations, in the order of tens of seconds, which we call outages. Outages happened at least once per day; for the six paths of provider P_4, they were a recurrent phenomenon. For two providers, the outages were correlated with changes in the fixed part of the delay; in example is shown in Fig. 12(a). The change in delay was in the order of 1-2 ms and it would not be noticed if it were not accompanied by loss. Other outages happened simultaneously on many paths of the same provider. Finally, some outages were repeated at the exact same time both days.

Based on the long outage durations and on the over-provisioning of IP backbones, we attribute the outages to link failures rather than to congestion. Link failures happen due to various reasons, such as linecard or router crashes, fiber cuts, maintenance operations. Typically, routing protocols need at least 5/15 sec to converge to a new configuration when a link goes down /up respectively. During this reconfiguration period, forwarding may be disrupted and voice packets may be lost. The reader is referred to [43] for a study of link failures and for the timers used during routing protocols convergence. The changes in the fixed part of the delay that were observed to accompany an outage are good indications of routing changes. Furthermore, the repetition of outages at the exact same time of each day indicates daily maintenance operations. Simultaneous outages on many paths of the same provider, indicate a failure of a shared link.

2) Delay Characteristics: As far as delay is concerned, there are two characteristics of interest: the fixed and the variable part of the delay. The *fixed part* of the delay consists of propagation and transmission delay and it is low (i.e. below the noticeable 100 -150 ms) on the backbone networks under study. Indeed, transmission delay is negligible on high speed backbone routers. Propagation delay is below 10 ms for communication on the same coast and in the range of 32-45 ms for coast-to-coast. Surprisingly enough, there are paths for which the fixed delay was as high as 78 ms, which is twice as large as the coast-to-coast minimum delay. This suggests that routing may not follow the shortest path. Unfortunately, we have no routing data available to verify this claim. Additional contributions to the total end-to-end delay can come from slow access links, from packetization at the sender and from playout delay at the receiver. In addition, *delay variability* leads to further packet drops in the playout buffer due to late arrivals.

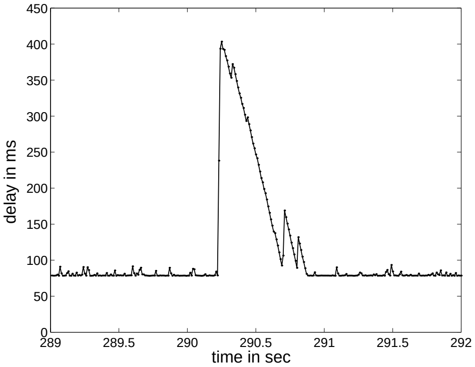


Fig. 10. Example of delay spike, frequently appearing on provider P_1 (THR-P_1 -ASH, Wed 06/27/01 UTC).

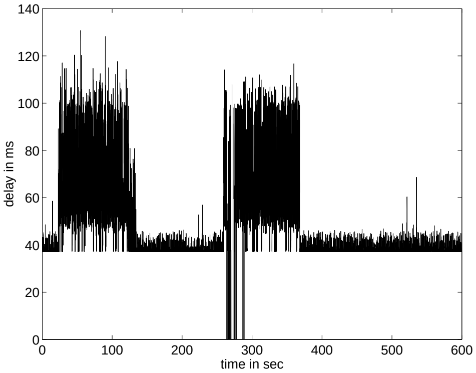
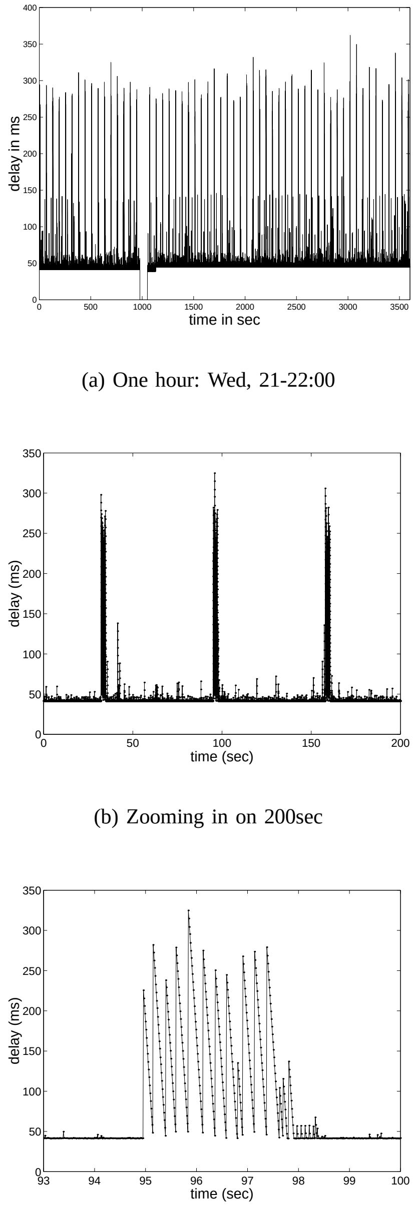


Fig. 11. 10 minutes on the path from EWR to SJC, using provider P_2, on Thu 06/28/01. The delay distribution alternates between 2 states. The second transition is accompanied by a 30 sec period of loss (157 clips, the longest is 1.5sec long).

Delay variability had always the shape of spikes. An example is shown in Fig. 10: there is a sudden increase in delay, followed by a roughly 45 degrees decrease. Different paths had different height/ frequency/clustering of spikes, but spikes are the dominant delay pattern on all the backbone networks we studied.

Delay variability may be caused by queuing (in which case delay pattern should be random) or due to other router-specific operations (in which case the delay pattern is more regular). On most paths, we observed very limited delay variability due to queuing, as is expected on well provisioned backbone networks. However, in some cases (mainly on providers P_1 and P_2), there were increased delay percentiles during business hours compared to night, indicating increased traffic load.

We also observed many regular patterns. A first example, that we call *higher plateaus*, is shown in Fig. 11. Such events happen on paths of provider P_1 and P_2, last for several minutes and are sometimes accompanied by long loss. The second example is the *periodic clustered delay spikes* shown in Fig. 12. The periodicity of this patterns as well as the magnitude of their spikes makes it difficult to explain through queuing and



(a) One hour: Wed, 21-22:00

(b) Zooming in on 200sec

(c) Zooming in on 7 sec

Fig. 12. Periodic Delay Pattern on EWR-P_4-SJC on Wed 06/27/01. Clusters of delay spikes (spikes are 300-350sec high and clusters last 3sec each), are repeated every 60-70 sec.

interleaving with interfering traffic. Furthermore, this periodic pattern is repeated every 60-70 sec on all six paths of the same provider for the entire measurement period. We attribute this perfectly periodic pattern to a router operation (such as debug options turned on, implementation-specific internal tasks) or to network control traffic (such as periodic routing table updates). During those periods, a router may stop forwarding traffic to serve higher priority tasks, resulting to the observed spikes.

3) Characteristics per Provider: The paths have a consistent behavior across the days observed. Furthermore, paths of the same provider have in general the same delay and loss pattern, whether they are short or long distance. This is intuitively expected as backbone network elements are shared by many paths of the provider. In Table II, we present the 43 paths grouped per provider.

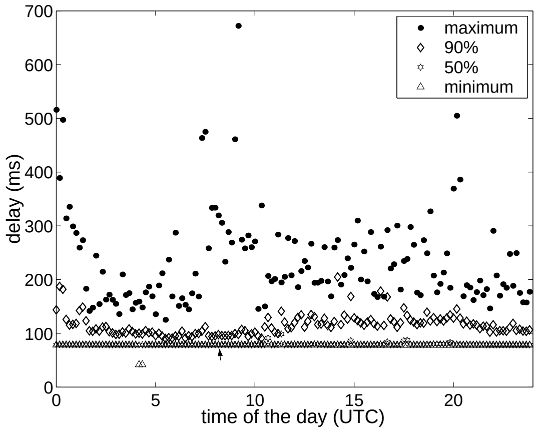


Fig. 13. Example path of provider P_1 (THR-P_1-ASH) on Wednesday 06/27/01. The delay percentiles are computed for 10 minutes intervals.

The ten long distance paths of provider P_1 exhibit high delay values, both for the fixed and the variable part. The delay variability comes in the form of single spikes, Fig. 10, or in the form of clusters of spikes, that last for five to ten minutes and appear during peak hours. Delay statistics for a typical path of provider P_1 are computed in 10 minutes intervals for an entire day and shown in Fig. 13. These paths can lead to acceptable performance if an appropriate playout scheme absorbs the delay variations. The two remaining P_1 paths are inherently poor, due to particularly high delays (as high as 800 ms) and many outage events (almost one every hour).

Delay on the two paths of provider P_2 alternates between two states. During the off-peak hours, delay in the long distance path is in the [37, 50]ms range; during the busy hours it jumps to “higher plateaus” in the range of [37,120] ms that last several minutes, as in Fig. 11, or several hours. Loss, at high rates and long durations, happens at the transitions between these states; an example is shown in the same figure.

The paths of provider P_3 have good performance. One of them had no loss at all. In the remaining 5 paths, single packets (10ms speech) were lost regularly every 5 seconds on average, (or at. 0.2% rate). These can be concealed without any perceived effect on the voice application. Delay and delay variability are also low.

Delay on all six paths of provider P_4 follows the periodic pattern of Fig. 12. Clusters of spikes 250-300ms high and lasting 3 sec each, are repeated every 60-70sec. The perfect periodicity on all paths and for the entire measurement period as well as the height of the spikes, hint toward network control traffic or a periodic operation specific to the routers used by this provider. These paths also exhibit outages in the order of 10s of seconds, 3-4 times per day, accompanied by changes in the fixed delay; an example is shown in Fig. 12(a).

The six paths pertaining to provider P_5 have also low delay variability in the order of 2-10ms. Occasional spikes can become higher during business hours. Two of the long distance paths experience no regular loss across the entire day. The other

TABLE II
CONSISTENT CHARACTERISTICS PER PROVIDER

Provider	number of paths	Distance	delay variability	typical loss duration (ms)	long loss (outages)		VoIP quality
					duration (sec)	times per day	
P_1	8	long	high	10-500	1-15	1-2	can be fair
	4	long	high	0	25-40	10	poor
P_2	1	short	two different states	0-20	1.5	3	good, except for state transitions
	1	long		200-400	5-15	2-3	
P_3	2	short	low	10	20	1	good
	4	long	low		6-20	2	
P_4	2	short	periodic spike clusters	10-100	15-30	3-5	poor
	4	long					
P_5	2	short	low	~200 or 0	1.1-2.5 (on all paths)	1 (on all paths)	good
	4	long	low				
P_6	4	short	very low	20	1.5-12	1-2	excellent
	5	long	very low	0-100			
P_7	2	long	very low	0	116, 167	1	excellent

TABLE III
SUMMARY OF SIMULATIONS SETTINGS

Component	Options Considered in Simulations
Talkspurt	exponential, mean=1.5sec, min=240ms
Silence	exponential, mean=1.5sec
Call duration	exponential with mean: 3.5 min (business), 10 min (residential)
Compression	G.711, G.729
Tasks	Task 1 (strict), Emodel (average)
Playout	Fixed throughout a call (various values) Spike-Det (as in [6]) “Improved“ (as in [36])
Traces used	THR-P_1-ASH, 15 min (Wed 14:00-14:15) THR-P_1-ASH, 1hour (Wed 14:00-15:00) THR-P_1-ASH, 1 day (Wed 0:00-23:59) EWR-P_6-SJC, 1 day (Wed 0:00-23:59) SJC-P_4-ASH, 1 hour (Wed 20:00-21:00)

four paths incur loss durations approximately 200ms (18-24 packets) at negligible rates. All six paths incur 1.1-2.5sec loss, at the same time (2:40 on Thu 06/28/01).

Provider P_6 exhibits low delay variability (typically within 2 ms) and negligible loss. The only problem on these paths are 1.5-12 sec periods of 50% loss. There are also frequent changes in the fixed delay which are not accompanied by loss.

The two long distance paths of provider P_7 have practically no delay variability and exhibit excellent performance except for a single outage that happened at the same time on both directions (119 sec in one and 167 sec in the other direction) and preceded a change in the fixed part of the delay. A few single packets dropped (5 in 2.5 days) and delay spikes, as infrequent as every 10 minutes, are negligible .

V. NUMERICAL RESULTS

In this section we apply the assessment methodology of Section III to the traces of Section IV. We first go through the analysis of an example path. Then, we present results for example paths of different providers.

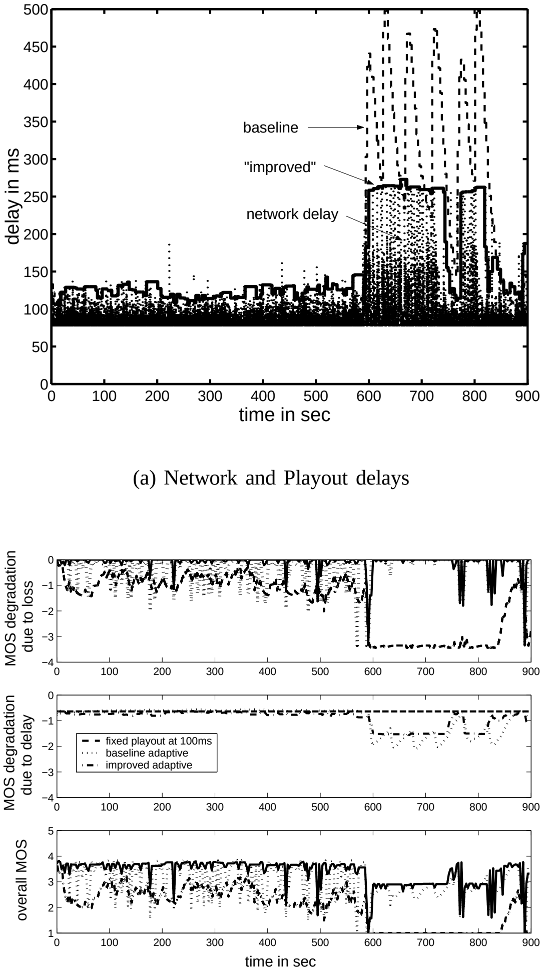
Our choices for the VoIP system are summarized in Table III. For both the talkspurts and the silences distributions, we use exponential with mean 1.5 sec; the minimum talkspurt duration is 240 ms, as suggested in [44]. As far as the playout buffer scheme is concerned, we considered both fixed and adaptive. The objective of this paper is not to design a new playback scheme or to exhaustively evaluate all existing ones, but to use realistic schemes to evaluate VoIP performance. We chose to implement the adaptive schemes proposed in [6] because they are well known, computationally light (thus suitable for simple implementations), yet able to follow the network delay variations. In particular, we used the “spike-detection” algorithm with its default parameters as the baseline adaptive scheme.

A. Example Path

Let us first consider the example trace of provider P_1 and a call taking place from 14:00 until 14:15 on 06/27/01. The selected trace exhibits large delay variations and a period of sustained loss. Fig. 14(a) shows the network delays and the playout times using fixed and adaptive playout. Fig. 14(b) shows the corresponding perceived quality. These results were achieved using G.711, which has a high intrinsic quality, an adequate echo cancellation ($EL = 51dB$) and requiring medium interactivity.

Let us first consider a fixed playout at 100ms. The quality is acceptable during the first 10 minutes, but not during the last 5 minutes, as shown in Fig. 14. Clearly, the larger the playout delay, the smaller the loss due to late arrivals but the larger the delay impairment. However, the overall MOS is a combination of both impairments and there exists a trade-off between loss and delay, shown in Fig. 15, leading to an optimal value of the playout delay that maximizes the overall $MOS(loss, delay)$. A similar loss-delay tradeoff holds under any VoIP configuration. However, the optimal delay value and the maximum achievable MOS may differ. For example, G.729, which starts at a lower intrinsic quality, can achieve maximum $MOS = 3$ and thus cannot be carried at acceptable quality levels during the considered 15 minutes. Similarly, a strict interactivity requirement (“Task 1”) or an acute echo (e.g. $EL = 41\,\mathrm{dB}$), would lead to $maximum\,MOS \cong 3$, which is unacceptable. An appropriate fixed value for the entire 15 minutes is around $200\,ms$. However, a more appropriate choice is 130 ms for the first 10 minutes and 250 ms for the last five minutes.

Adaptive schemes adjust more frequently, i.e. every talkspurt.



(a) Network and Playout delays

(b) Time varying impairments and instantaneous quality

Fig. 14. An example of 15min call (Wed 06/27/01, 14:00-14:15, UTC) and its perceived quality for fixed and adaptive playout. Path THR-P_1-ASH.

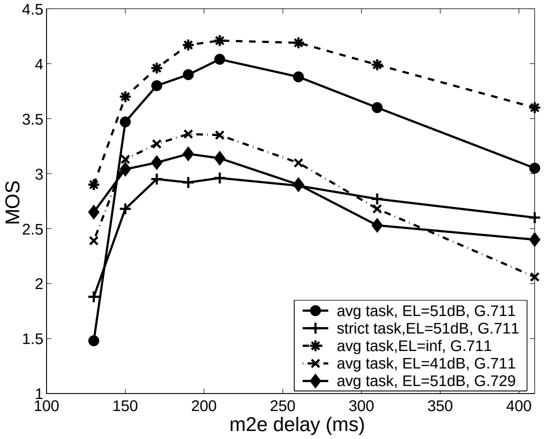


Fig. 15. Combined MOS (including both loss and delay impairments) at the end of the 15 min example call. Various configurations are considered: task (average or strict), $EL = \{\infty,\ 51dB,\ 41dB\}$, compression scheme (G.729 or G.711). Fixed playout is applied during the entire call.

TABLE IV

TUNING THE PARAMETERS AND THE EFFECT ON THE PERFORMANCE OF THE BASELINE ADAPTIVE PLAYOUT OVER THE 10 FIRST MINUTES OF EXAMPLE TRACE P_1.

$weight$	threshold	% loss	avg	avg
α	ENTER		delay	MOS
0.99802	100ms	3.29%	103 ms	2.7
same	50ms	2.35%	108 ms	2.88
same	30ms	0.8667%	132 ms	3.18
same	20ms	0.17%	166 ms	3.26
0.875	20ms	7.55%	98 ms	1.82
0.90	100ms	6.82%	98 ms	1.96
0.95	same	5.1%	102 ms	1.48
0.98	same	2.6%	113 ms	2.66
0.99	20ms	1.5%	125 ms	2.92
0.99	30ms	2.62%	109 ms	2.67

The baseline adaptive scheme operated near the optimal region achieving $max\,MOS = 3.6$ for an average delay of $122\,ms$. Fig. 14 shows the playout delays and the resulting perceived quality. In Fig. 14(b), we show the loss impairment (due to loss in the network and in the playout buffer), the delay impairment and the overall MOS. The baseline algorithm fails at the following points. First, it tries to follow the network delays too close during the first 10 minutes, thus leading to significant loss rates and many clips of small durations. Second, it results in long loss durations in the transition around 600sec. Third, the $p = d + 4v$ formula leads to significant over-estimation of the delay, and thus delay impairment, during the last 5 minutes.

Noticing these problems, we tried to tune the baseline playout for this trace and we found that its performance is very sensitive to the tuning. Table IV shows that the performance of the algorithm is poor not only for the default parameters used in Fig. 14, but also for a wide range of the parameters.

This motivated us to design our own playout schemes that would be appropriate for these backbone delay variations, [36]. One of the schemes we considered was a percentile-based algorithm, similar to [7]. As a further improvement, we dynamically adjusted the percentile to achieve the maximum MOS as a function of both delay and loss, see Fig. 15. The performance of the "improved" scheme is shown in Fig. 14 together with the fixed and the baseline scheme. Indeed, the "improved" scheme achieves a tight upper bound of network delay in Fig. 14(a) and it notices fast the sudden change in delay pattern. Thus, it results in low loss and delay impairments and high overall MOS, see Fig. 14(b). In the context of this study, we evaluate the traces using existing algorithms.

Having discussed one call in detail, let us now consider many calls initiated at random times, uniformly spread over an entire hour, e.g. from 14:00 to 15:00. We consider exponentially distributed call durations as in [4]. 150 short ($3.5\,min$ mean) and 50 long ($10\,min$ mean) durations simulate business and residential long distance calls, respectively. To rate each call, we use both the minimum MOS during the call (that a network operator might want to eliminate)and the more lenient rating at the end (that a human would give), as discussed in Section III-F. Fig. 16 shows the cumulative distribution (CDF) of ratings for the 200 calls, using both measures. If fixed playout is

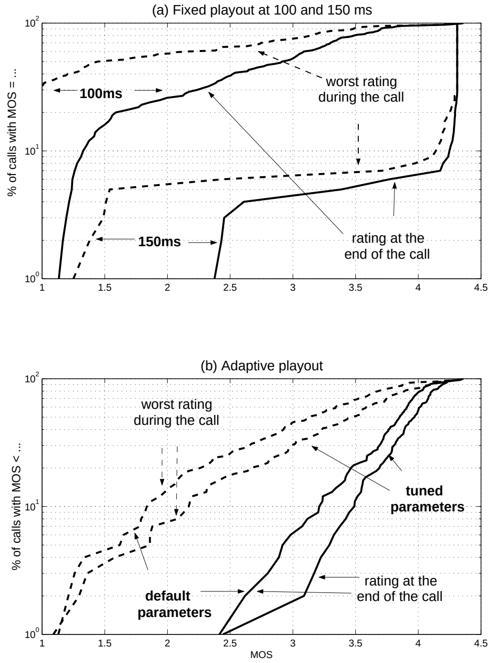


Fig. 16. CDF of call ratings in one-hour period (Wednesday 06/27/01, 14:00-15:00) on a loaded path (THR-P1-ASH).

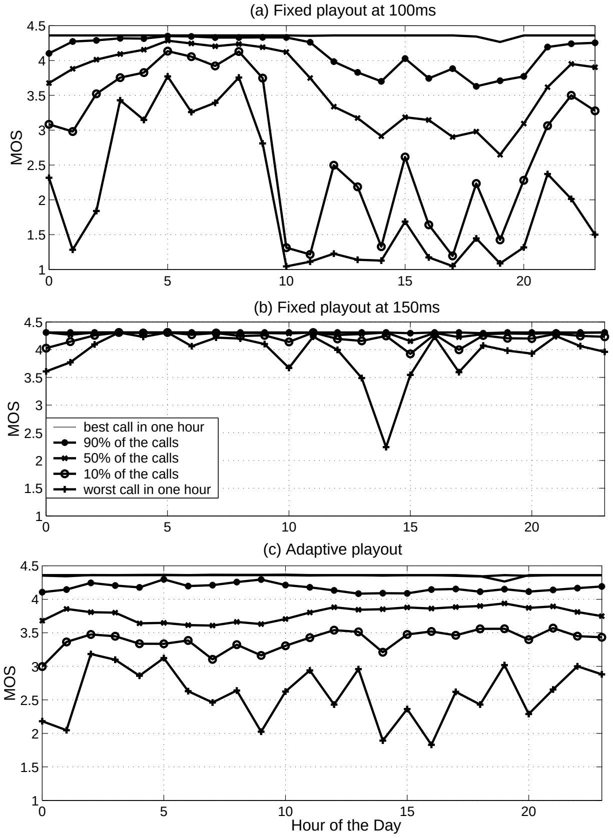


Fig. 17. Call quality statistics for every hour of an entire day (Wed 06/27/01) on a loaded path (THR-P_1-ASH). Playout used: (a) Fixed at 100ms (b) Fixed at 150ms (c) Adaptive with default parameters.

used, Fig. 16(a), then the choice of the fixed value becomes critical: $150\,ms$ is acceptable (only 6% of the calls have final rating below 3.6 and only 8% of them experience a period of $MOS < 3.6$) while $100\,ms$ is totally unacceptable (90% of the calls have rating at the end below 3.6). For the adaptive playout, Fig. 16(b), we observe the following: (i) the CDF is more linear than for the fixed scheme (ii) this performance is acceptable but still not excellent (10% of the calls have overall $MOS < 3.6$; 50% of them experience a period of $MOS < 3.5$ at least once) and (iii) tuning of the parameters does not lead to significant improvement.

While in Fig. 16 we plot the entire CDF, in Fig. 17 we consider only some percentiles (i.e. worst rating, 10%, 50%, 90%, 100%) of call ratings for each hour-bin of the entire day. E.g. the points in Fig. 17(a) for $Hour = 14$ are consistent with Fig. 16(a): out of the 200 calls between 14:00 and 15:00, the worst rating was 1.1, 10% of the calls had $MOS \leq 1.4\%$, 50% of the calls had $MOS \leq 3$, 90% of the calls had $MOS \leq 3.75$ and some calls had perfect rating.

Fig. 17(a) shows that a fixed playout at $100\,ms$ is unacceptable when the delays on the path are high, i.e. during the business hours. In practice, the choice of the fixed playout value should not be the same for the entire day, but should be infrequently adjusted. In Fig. 17(c), the adaptive playout had the same performance for the entire day including the business hours, because it was able to monitor the changes in the network delays. The reason why 10% of the calls in any hour still had $MOS < 3.5$, is the sensitivity of the scheme to the tuning of its parameters. The bad rating at 14:00 is due to network loss.

B. Other types of paths

We applied the same procedure to different types of paths. Paths of very low delay and low delay variability, mainly belonging to providers P_6 and P_7, achieve an excellent MOS at all times except for the rare cases when outages occur. Given that the fixed part of the delay on these paths is below 50 ms, a conservatively high fixed playout delay of 100-150 ms is sufficient to yield excellent performance, except for a few very high delay spikes. 90% of the calls on the example path of provider P_6 have $4 \leq MOS < 4.4$. Only two calls in the entire day have a low rating, because they overlapped with outages. The performance degrades when the adaptive playout tries to closely follow the network delay; this is unnecessary for these paths, where delay does not vary significantly.

In contrast, paths of provider P_4 exhibit periodic clusters of spikes that are as high as 250-300ms, see Fig. 12. The delay pattern is the same across the entire day and we examine a typical hour. If the baseline adaptive playout is used, then 20% of the calls have overall $MOS < 3.5$ and 80% of the calls experience $MOS < 3.5$ for some period. If a more interactivity is required, then the entire CDF degrades by approximately 0.8 unit of MOS. If an appropriately high fixed delay is chosen, performance improves: only 10% of the calls have $MOS < 3.5$. Because the spikes are 250-300ms high, they cannot be accommodated without loss in interactivity.

C. Discussion

1) On the Performance of Backbone Networks : Our study shows a large range of behaviors across backbone networks. (However, behavior was mostly consistent across paths of the same provider and similar for short and long distance paths.) There are some backbone networks that exhibit good characteristics and are already able to support voice communication at high quality levels. Other backbone networks exhibit undesirable characteristics, such as large delay spikes, periodic delay patterns, outages correlated with changes in the fixed part of the delay, loss simultaneously on many paths. These characteristics lead to poor VoIP performance. Using G.711 encoder with high intrinsic quality, good echo cancellation and low interactivity requirements, these paths are barely able to provide acceptable VoIP service ($MOS > 3.6$). Performance is even worse when interactivity requirements are strict (MOS decreases by roughly 0.5-1 units) or when echo is inadequately canceled. Support of G.729, which has lower intrinsic quality, is possible only on the good paths.

Action for improving the VoIP performance can be taken inside the network or/and at the end-systems. Most of the problems we identified in the backbone networks seem more related to reliability (e.g. link failures and routing reconfiguration), network protocols (routing protocol exchanges or other control traffic) and router operation (e.g. debug options, router "vacations"), rather than to traffic load. Therefore, in these high bandwidth environments, more effort should be put on understanding the network operation and improving reliability rather than on QoS mechanisms. To mitigate network induced impairments, the end-systems can also use some mechanisms, including PLC (to mitigate the effect of packet loss), playout scheduling (to absorb delay variability) and path diversity techniques. The effectiveness of such techniques is limited by the magnitude of the impairments introduced by the network.

2) On the Playout Buffer: In this paper, our intention was to consider some realistic playout schemes as part of the end-to-end VoIP system. We first consider fixed playout for a range of fixed playout delays, as a benchmark for comparison. for most of the paths, an appropriately high fixed value led to high overall perceived quality.

We then considered the adaptive schemes proposed in [6], in order to support VoIP in high delay paths and consider strict interactivity requirements. However, we observed that the baseline adaptive scheme did not perform well over the backbone networks under study. Furthermore, it was sensitive to the tuning of its parameters. The reason is that the delay pattern consists of spikes and there is no slow varying component to track in these backbone networks. We next tried a percentile-based approach, similar to the one proposed in [7], but also taking into account the loss-delay trade-off of Fig. 15, in order to choose the percentile that optimizes the overall perceived quality $MOS(loss, delay)$. In this paper, we only present sample results from this approach. In [36], we continue the work on playout scheduling in two directions. First, we demonstrate the need for making the algorithm learn the network delay pattern and adjust its parameters appropriately. Second, we design different modes of operation, depending on the user preference between loss and delay.

VI. Conclusion

In this paper, we assess the ability of Internet backbones to support voice communication. We compare and combine results from various subjective testing studies and we develop a methodology for assessing the perceived quality of a telephone call. A key asset in our study is the use of network measurements collected over backbones of major ISPs.

Although backbone networks are, in general, sufficiently provisioned and thus expected not to cause problems for data traffic, we find that this is not necessarily the case for voice traffic. Some backbone networks exhibit fairly good characteristics, leading to a confirmation that packet voice is a sound approach. Other backbones exhibit problems that seem mostly related to reliability and network operation. As long as problems exist but remain below a certain magnitude, some measures can be taken at the end-systems to mitigate their effect.

Acknowledgments

The authors would like to thank the anonymous reviewers of INFOCOM 2002 and ToN for their invaluable feedback. Special acknowledgment goes to Amit Vyas for helping with the processing of the traces for valuable discussions throughout the project. We are also grateful to RouteScience for providing the traces and for partially supporting this study.

References

[1] A. Markopoulou, F. Tobagi, M. Karam, "Assessment of VoIP quality over Internet Backbones", *in Proc. of IEEE INFOCOM 2002,* NY, June 2002.

[2] K. Sriram, W. Whitt, "Characterizing superposition arrival processes in packet multiplexers for voice and data", *IEEE Journal on Selected Areas on Communications, SAC-4(6): 833-846,* September 1986.

[3] P. Brady, "A technique for investigating on/off patterns of speech", *Bell Labs Tech.Journal, 44(1):1-22,* January 1965.

[4] W. Jiang, H. Schulzrinne, "Analysis of On-Off Patterns in VoIP and their effect on Voice Traffic Aggregation", *in Proc. ICCCN 2000.*

[5] R. Cox, "Three new speech coders from the ITU cover a range of applications", *IEEE Communications Magazine*, September 1997.

[6] R. Ramjee, J. Kurose, D. Towsley, H. Schulzrinne, "Adaptive playout mechanisms for packetized audio applications in wide-area networks", *in Proc. of IEEE Infocom 1994,* Toronto, Canada, June 1994.

[7] S. Moon, J. Kurose, D. Towsley, "Packet audio playout delay adjustment: performance bounds and algorithms", *ACM/Springer Multimedia Systems, vol. 6, pp.17-28,* January 1998.

[8] P. DeLeon, C. Sreenan, "An adaptive predictor for media playout buffering", in *Proc. of IEEE ICASSP* 1999, March 1999.

[9] C. J. Sreenan, J.-C. Chen, P. Agrawal, B. Naderdran, "Delay reduction techniques for playout buffering", *IEEE Transactions on Multimedia, Vol.2, no.2 pp. 88-100,* June 2000.

[10] Y. Liang, N. Farber, B. Girod, "Adaptive Playout Scheduling and Loss Concealment for Voice Communications over the networks", *IEEE Transactions on Multimedia,* April 2001.

[11] I. Kouvelas, V. Hardman, "Overcoming workstation scheduling problems in a real-time audio tool", *in Proc. of USENIX 1997,* Anaheim CA, USA, January 1997.

[12] C. Perkins, O. Hodson, V. Hardman, "A survey of packet loss recovery techniques for streaming audio", *IEEE Network*, Sept./Oct. 1998.

[13] UCL, Department of Computer Science, Robust Audio Tool (RAT), *http://ww-mice.cs.ucl.ac.uk/multimedia/software/rat/*

[14] T. Kostas, M. Borella, I. Sidhu, G. Schuster, J. Grabiec, "Real-time voice over packet-switched networks", *IEEE Network*, January/February 1998.

[15] *ITU-T Recommendation P.800*, "Methods for subjective determination of transmission quality", August 1996.

[16] *ITU-T Recommendation G.107,* "The Emodel, a computational model for use in transmission planning", December 1998.

[17] ITU-T *Recommendation G.109*, "Definition of categories of speech transmission quality", Sept. 1999.

[18] A. Watson, M. Sasse, "Measuring perceived quality of speech and video in multimedia conferencing applications", *in Proc. of ACM Multimedia 1998*, Bristol, UK, Sept 12-16, 1998.

[19] J. Gruber, L. Strawczynski, "Subjective effects of variable delay and speech clipping in dynamically managed voice systems", *IEEE Trans. on Communications*, *vol. 33, No.8*, Aug.1985.

[20] R. Cox, M.Perkins, "Results of a subjective listening test for G.711 with frame erasure concealment", *AT&T contr. to T1A1.7/99-016*, May 1999.

[21] M. Perkins, K. Evans, D. Pascal, L. Thorpe, "Characterizing the subjective performance of the ITU-T 8 Kbps Speech Coding Algorithm - ITU-T G.729", *IEEE Comm. Magazine*, Sept. 1997.

[22] H. Sanneck, L. Le, A. Wolisz, "Intra-flow loss recovery and control for VoIP", *in Proc of ACM Multimedia 2001*, Ontario, Canada 2001.

[23] S. Voran, "Speech quality of G.723.1 coding with added temporal discontinuity impairments", *Proc. of ICASSP* May 2001.

[24] *ITU-T Recommendation G.108*, "Application of the Emodel: a planning guide", September 1998.

[25] *ITU-T Recommendation G.113*, "Transmission impairments due to speech processing", February 2001.

[26] *ETSI, TIPHON project*, TR 101 329-6 "Actual measurements of network and terminal characteristics and performance parameters in TIPHON networks and their influence on voice quality", July 2000.

[27] *ITU-T Recommendation G.114*, "One way transmission time", May 2000.

[28] N. Kitawaki, K. Itoh, "Pure delay effects on speech quality in telecommunications", *IEEE Journal on Selected Areas in Communications*, *vol . 9, no.4*, May 1991.

[29] R.G.Cole, J. Rosenbluth, "Voice over IP performance monitoring", *Computer Communications Review, V.4, no.3*, April 2001.

[30] V. Vleeschauwer, J. Janssen, G. Petit, F. Poppe, Alcatel Technical Report "Quality bounds for packetized voice transport", *Alcatel Tech. Report*, 1st Quarter of 2000.

[31] A. Clark, "Modeling the effects of burst packet loss and recency on subjective voice quality", *Proc. of IP Telephony Workshop*, March 2001.

[32] France Telecom R&D, "Continuous assessment of time-varying subjective vocal quality and its relationship with overall subjective quality", *ITU ST 12*, *Contr. COM 12-94-E*, July 1999.

[33] France Telecom R&D, "Study of the relationship between instantaneous and overall subjective speech quality for time-varying quality speech sequences: influence of the recency effect", *ITU Study Group 12*, *contribution D.139*, May 2000.

[34] A. Clark, R. Liu, "Comparison of TS 101 329-5 Annex E with PAMS and PSQM", Temp.Doc. 061 for *TIPHON#23*, July 2001.

[35] A. Clark, R. Liu, "Comparison of TS101 329-5 Annex E with Emodel", Temp.Doc. 062 for *TIPHON#23*, July 2001.

[36] A. Markopoulou, "Assessing the Quality of Multimedia Communications over Internet Backbones", *Ph.D. Dissertation, Stanford University*, October 2002.

[37] F. Tobagi, A. Markopoulou, M. Karam, "Is the Internet ready for VoIP?", *in Proc. IWDC 2002*, Capri, Italy, September 2002.

[38] J.-C. Bolot, "Characterizing end-to-end packet Delay and Loss in the Internet", *in Proc. of ACM SIGCOMM 1993*, pp.289-298, San Francisco, CA, USA, September 1993.

[39] J.-C. Bolot, H. Crepin, A. V. Garcia, "Analysis of Audio Packet Loss in the Internet", *in Proc. of NOSSDAV 1995*, Durham NH, USA, 1995.

[40] M. Yajnik, S. Moon, J. Kurose, D. Towsley, "Measurement and Modeling of the Temporal Dependence in Packet Loss", *in Proc. of IEEE INFOCOM 1999*.

[41] D. Loguinov, H. Radha, "Performance of Low Bitrate Internet Video Streaming", *in Proc. IEEE INFOCOM 2002*, New York, USA, June 2002.

[42] C. Boutremans, G. Iannaccone, C. Diot, "Impact of link failuires on VoIP performance", *in Proc. ACM NOSSDAV 2002*, Miami Beach, FL, USA, March 2002.

[43] G. Iannaccone, C. Chuah, R. Mortier, S. Bhattacharyya, C. Diot, "Analysis of link failures in an IP backbone", *in Proc. of ACM Internet Measurement Workshop*, Marseille, France, November 2002.

[44] W. Jiang, H. Schulzrinne, "QoS measurement of real time multimedia services in the Internet", *Columbia Univ. Report CUCS-015-99*, 1999.

PLACE
PHOTO
HERE

Athina Markopoulou (M '98) received the B.S. degree in Electrical and Computer Engineering from the National Technical University of Athens in 1996. She received the M.S. and Ph.D. degrees in Electrical Engineering from Stanford University in 1998 and 2002 respectively. She is currently a postdoctoral visitor at Sprint ATL. Her research interests include network performance measurement and modeling, and multimedia traffic over packet networks.

PLACE
PHOTO
HERE

Fouad Tobagi (M'77-SM'83-F'85) received the Engineering Degree from Ecole Centrale des Arts et Manufactures, Paris, France, in 1970 and the M.S. and Ph.D. degrees in Computer Science from the University of California, Los Angeles, in 1971 and 1974, respectively. From 1974 to 1978, he was a Research Staff Project Manager with the ARPA project at the Computer Science Department, University of California, Los Angeles, and engaged in research in Packet Radio Networks, including protocol design, and analysis and measurements of packet radio networks. In June 1978, he joined the faculty of the School of Engineering at Stanford University, Stanford, California, where he is Professor of Electrical Engineering and Computer Science. In 1991, he co-founded Starlight Networks, Inc., a venture concerned with multimedia networking, and served as Chief Technical Officer until August 1998. His research interests have comprised packet radio and satellite networks, local area networks, fast packet switching, multimedia networking and networked video services, and multimedia applications. His current interests include voice and video communication over the Internet, wireless and mobile networks, network design and provisioning, and network resource management. Dr. Tobagi is a Fellow of the IEEE for his contributions in computer communications and local area networks. He is the winner of the 1981 Leonard G. Abraham Prize Paper Award in the field of Communications Systems for his papser "Multiaccess Protocols in Packet Communications Networks" and co-winner of the IEEE Communications Society 1984 Magazine Prize Paper Award for the paper "Packet Radio and Satellite Networks". He has served as Associate Editor for Computer Communications in the IEEE Transactions on Communications for the period 1984-1986, Editor for Packet Radio and Satellite Networks in the Journal of Telecommunications Networks for the period 1981-1985, Co-Editor of the special issue on Local Area Networks of the IEEE Journal on Selelected Areas in Communications (November 1983), Co-Editor of Advances in Local Area Networks, a book in the series Frontiers in Communications (New York: IEEE Press), Co-Editor of the special issue on Packet Radio Networks of the Proceedings of the IEEE (January 1987), and Co-Editor of the special issue on Large Scale ATM Switching Systems for B-ISDN of the IEEE Journal on Selected Areas in Communications (October 1991). He has also served as Co-Editor of Advances in Local Area Networks, a book in the series Frontiers in Communications (New York: IEEE Press). He is currently serving as editor for a number of journals in High Speed Networks, wireless networks, multimedia and optical communications. He is a member of the Association for Computing Machinery and has served as an ACM national Lecturer for the period 1982-1983. He is co-recipient of the 1998 Kuwait Prize in the field of Applied Sciences.

PLACE
PHOTO
HERE

Mansour Karam (M '98) received the B. Engineering degree in Computer and Communications Engineering from the American University of Beirut, Beirut, Lebanon, in 1995 and the M.S. and Ph.D.degrees in Electrical Engineering from Stanford University in 1997 and 2001, respectively. He is currently a Research Engineer at RouteScience Technologies. His research interests include the support of multimedia applications in wired and wireless networks and routing control over the Internet.