Playout Scheduling and Loss-Concealments in VoIP for Optimizing Conversational Voice Communication Quality

Batu Sat
Dept. of Electrical and Computer Engineering
and the Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
batusat@uiuc.edu

Benjamin W. Wah
Dept. of Electrical and Computer Engineering
and the Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
wah@uiuc.edu

ABSTRACT

In this paper, we present new adaptive *playout scheduling* (POS) and *loss-concealment* (LC) schemes for delivering high and consistent *conversational voice communication quality* (CVCQ) perceived by users in real-time VoIP systems. We first characterize the delay and loss conditions of an IP network and a human conversation in a VoIP system. We then identify the attributes that affect the human perception of CVCQ, which include *listening-only speech quality* (LOSQ), *conversational interactivity* (CI), and *conversational efficiency* (CE). We investigate their trade-offs with respect to system-controllable *mouth-to-ear delays* (MEDs) and the amount of *redundant piggybacking*. Finally, we evaluate our adaptive POS and redundancy-based LC schemes by packet traces collected in the PlanetLab.

Categories and Subject Descriptors

H.4.3 [**Information Systems Applications**]: Communications Applications—*Computer conferencing, teleconferencing, and videoconferencing*

General Terms

Algorithms, Design, Human Factors, Performance

Keywords

Multimedia Communication, Voice over IP, Perceptual Conversational Quality, Just Noticeable Difference

1. INTRODUCTION

The use of VoIP for carrying real-time voice data over any IP network, public or private, has significant impacts on the multi-billion dollar telecommunication industry. Its promise on less expensive phone calls with comparable quality, as well as the proliferation of the broadband Internet, has increased its worldwide adoption.

MM'07, September 23–28, 2007, Augsburg, Bavaria, Germany.

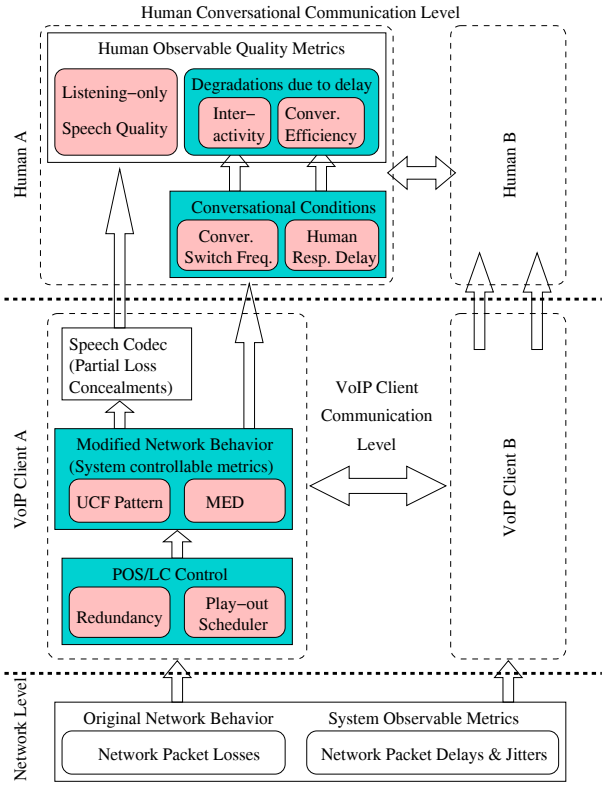

Figure 1: Architecture of a VoIP system.

An ideal VoIP session should allow participants to have an experience that closely resembles a face-to-face conversation, with signals transmitted with perfect quality and without delay. In practice, a VoIP session may experience degradations when late or lost voice frames cannot be recovered at the receiver at their scheduled playout times.

To overcome variations in the arrival times of IP packets, a receiver often employs a jitter-buffer and a playout scheduler. Here, the *mouth-to-ear delay* (MED) is the total delay a speech frame incurs at the sender, the network, and the receiver jitter buffer before it is played. Figure 1 depicts the interactions among the components of a VoIP system, the underlying IP network, and the human participants. It also shows the dependencies among the system-observable, system-controllable, and user-observable metrics. (Some of the terms and details are defined later in the paper.)

A voice conversation between two persons consists of a series of alternating speech segments with *talk-spurts* and *silence periods*. The quality of the conversation can, therefore, be evaluated by examining the quality of the one-way speech and that of the interactive conversation.

In a one-way transmission of speech, the *listening-only speech quality* (LOSQ) improves as MED increases. A user's perception of LOSQ mainly depends on the intelligibility of the speech heard, since the user lacks a reference to the original sequence. Intelligibility, on the other hand, depends on many factors other than signal degradations incurred during transmission. The topic of the conversation, the commonality of the words used, and the familiarity of the speakers can all affect intelligibility. To mitigate these subjective effects in evaluating LOSQ, formal mean-opinion-score (MOS) tests (ITU P.800) [9] are usually conducted by a panel of listeners who only listen to pre-recorded segments. However, due to the time-consuming and non-repeatable nature of MOS tests, listening-only perceptual speech quality is commonly evaluated by PESQ (ITU P.862) [10].

MOS and PESQ of one-way speech improve as MED increases because packets that experience long delays will eventually arrive before their scheduled playout times. Further, packets lost in the network can be recovered by *loss-concealment* schemes (LC) that send redundant copies of these packets in subsequent ones. With a sufficiently long MED and a large number of redundant copies, a perfect LOSQ can be achieved. The fraction of those frames that cannot be recovered is captured by the *unconcealed frame rate* (UCFR) [14]). Note that degradations in LOSQ as a function of MED also depend on the codec used: low bit-rate codecs tend to be less robust to packet losses, especially when consecutive frames are lost.

The quality of an interactive conversation, however, does not depend on LOSQ alone. The *G.114 guidelines* [4] prescribe a one-way MED of less than 150 ms to be desirable for a voice-communication system and more than 400 ms to be unacceptable. However, they do not specify a metric for measuring the effect of delays, nor do they give trade-offs that lead to conversations of high perceptual quality.

Two metrics we have found to be important in an interactive conversation are the *conversational interactivity* (CI) and the *conversational efficiency* (CE). Informally, CI is related to the silence period experienced by a person before hearing the other party's response, as well as the duration waited by the person after hearing the other party. (More formal definitions are shown in Section 3.) In contrast to a face-to-face conversation, as MED increases, the silence periods when switching between the two parties are no longer symmetric. When the asymmetry in the response times increases, humans tend to have a degraded perception of interactivity. Another effect of increased MEDs is the lower CE, since it takes longer time to accomplish a task with respect to the same conversation in a face-to-face setting.

The degradations due to delays may also depend on the conversational condition, such as the type of conversation being carried out and the conversational switching frequency (Section 3). For example, a social conversation may have less frequent switches between the parties, and the degradations due to long MEDs are perceived less severely. In contrast, in a business or mission-critical conversation in which the direction of the conversation switches more frequently, there is an increased need for face-to-face like interactivity.

Currently, there is no standard that relates MEDs to user-perceptible conversational-quality metrics. To facilitate the evaluation of conversations under different conditions, we propose in this paper a *conversational voice communication quality* (CVCQ) that relates LOSQ, CI, and CE. In general, LOSQ is a non-decreasing function of MED, but CI and CE are non-increasing functions of MED. Hence, there is an optimal MED at which there are proper trade-offs among LOSQ, CI, and CE. At this MED, the system achieves a high LOSQ by playing an adequate number of frames in time and maintains an adequate CI and CE in the conversation. As the optimal MED may change with network condition, it will need to be dynamically adapted by the *playout scheduling* (POS) algorithm in real time in order to achieve high and consistent CVCQ (Section 3).

The difficulty of evaluating real-time conversations is that the relation among LOSQ, CI, and CE is very complex and cannot be expressed in closed form. Moreover, the evaluations may have to be carried out by subjective tests rather than by objective measures. One property to our advantage is that minor differences in each metric are not noticeable to humans. As a result, it is possible to discretize each metric into regions in such a way that quality differences are noticeable across two adjacent regions but not within a region.

Previous work. There have been several studies on adaptive POS schemes that aim to balance the number of late packets for playout and the jitter-buffer delays that packets wait before their scheduled playouts. Figure 2 depicts some proposed POS algorithms and their performance on an international connection. It shows the temporal changes in packet delays and the corresponding estimated playout delays. It also depicts the system- and user-observable quality metrics that contribute to CVCQ.

Open-loop schemes. These use heuristics for picking some system-controllable metrics (such as MED), based on network statistics available at the time. For example, Algorithms 1-3 [11] (Figure 2) calculate running estimates of the mean (d) and the variations (v) in network delays and choose a playout delay $p = d + 4v$ at the beginning of each talkspurt. Algorithm 4 [11] improves the estimations by tracking delay spikes and avoids long sequences of unconcealed lost frames. These algorithms are less robust because they are open-loop schemes and do not optimize a target. Further, they do not consider the effects of the codec used on LOSQ.

Closed-loop schemes with intermediate quality metrics. In our previous studies [13], we have proposed closed-loop schemes that adapt the amount of redundancy for LC using an intermediate metric. Although it is robust to dynamic network conditions, choosing such a metric proves to be difficult. The metric must be easy to compute at run time and be tied to a target objective. Algorithm 3 [12] (Figure 2) follows a similar approach by controlling an intermediate metric based on the late-loss rate collected in a window.

Closed-loop schemes with end-to-end quality metrics. The E-model (ITU G.107) [3] estimates the conversational quality based on objective network and system attributes, such as the loss rate, codec used, one-way delay, and echo level. It is used in a closed-loop framework [1] to jointly optimize POS and FEC-based LC. However, the study assumes a quasi-static Gilbert packet-loss process, a stationary observable delay distribution, and mutually independent delays and losses. These assumptions can be shown to be invalid for a variety of Internet connections (Section 2).

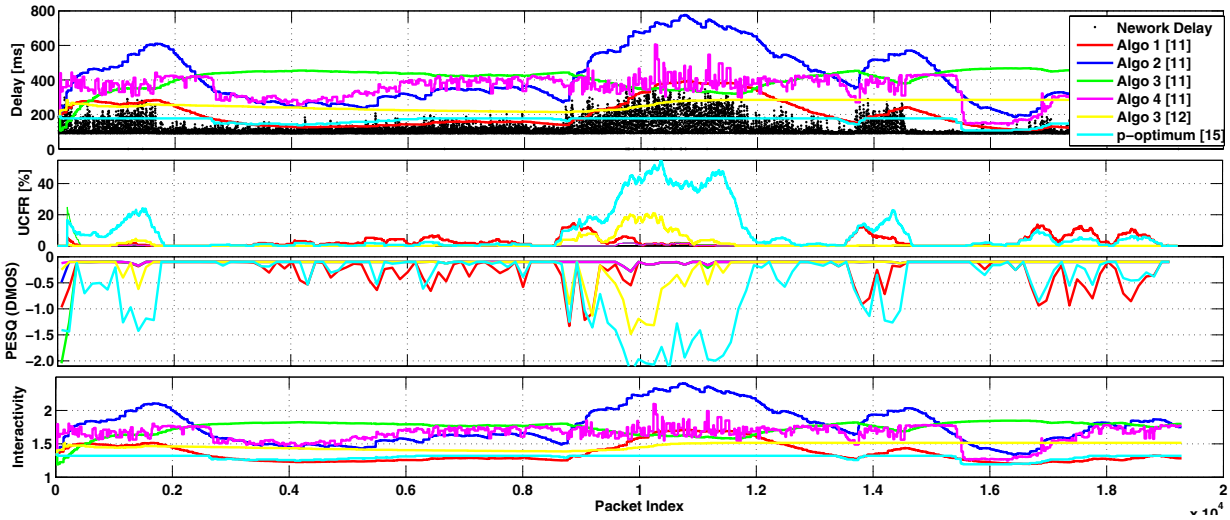

Figure 2: Delays over the US-Switzerland connection based on traces collected in the PlanetLab. Also shown are the previous playout schedules and their impact on UCFR, LOSQ (measured by PESQ), and CI.

Table 1: Traces collected in Dec. 2005: 5-min duration, 60-ms period, 100-byte payload, $\Delta \doteq |LR_{t+2sec} - LR_t|$.

Path	Source		Destination		One-Way Delay [ms]				Jitter wrt Mean (%)		Δ Loss Rate (in %)			
	Location	IP Address	Location	IP Address	Min	Max	Mean	Std.Dev.	$J > 60$ms	$J > 30$ms	Max	Mean	Std.Dev.	E[Δ]
1	Netherlands	130.37.198.243	China	219.243.200.53	170	223	172	2.4	0.0	0.1	18.2	5.3	3.9	4.2
2	Brasil	200.19.159.34	Wisconsin	198.133.244.146	121	379	129	17.1	2.0	5.7	12.1	1.7	2.6	2.1
3	Michigan	141.213.4.202	Spain	138.100.12.149	62	237	62	4.3	0.1	0.2	48.5	0.8	4.9	1.3
4	California	169.229.50.12	Germany	132.187.230.2	107	423	114	69.5	0.9	0.9	0.0	0.0	0.0	0.0
5	China	219.243.200.53	Wisconsin	198.133.244.146	120	148	123	3.2	0.0	0.0	63.6	33.5	10.9	15.9
6	Italy	130.136.254.21	Canada	142.103.2.1	99	257	102	40.5	0.3	0.3	15.2	1.2	2.1	1.9
7	Korea	143.248.139.168	Brasil	200.129.0.162	166	250	169	38.7	0.4	0.4	0.0	0.0	0.0	0.0
8	China	219.243.201.17	Ohio	129.22.150.90	110	237	112	3.4	0.1	0.1	30.3	8.0	5.8	5.8
9	Netherlands	130.161.40.154	Hong Kong	137.189.97.17	142	927	249	138.3	25.5	31.4	6.1	0.1	0.8	0.2
10	Wisonsin	198.133.244.146	China	219.243.200.53	121	134	122	1.8	0.0	0.0	42.4	17.4	7.8	9.8

Another study [15] proposes to use the E-model but separately models the effects of the loss rate and the codec on the listening-only portion of conversational quality. By training a regression model to estimate LOSQ (measured by PESQ) with respect to the loss rate, a POS algorithm *p-optimum* was proposed to adapt the playout delay on each talk-spurt in order to optimize LOSQ (Figure 2). For simplicity, the model was trained by a Bernoulli loss model and does not employ, nor is designed to work in conjunction with, a redundancy-based LC scheme. Because unconcealed lost frames can be bursty, such a model under-estimates the degradations due to lost frames. Moreover, these frames cannot be recovered by adjusting playout delays alone.

Problem statement. To overcome the limitations in previous studies, we develop POS and LC schemes as closed-loop control schemes that optimize the trade-offs among the observable LOSQ, CI, and CE metrics in order to deliver high and consistent CVCQ.

2. NETWORK ENVIRONMENT

Public IP networks exhibit dynamic path-dependent characteristics in real-time transmissions. Table 1 summarizes some of our results on experiments conducted in the PlanetLab. In this section, we present our observations on the system-observable network conditions, define POS/LC-control schemes, and present the trade-offs among the system-controllable quality metrics. The relation among these components are also shown in Figure 1.

In a VoIP session, the clients send and receive voice packets (or frames) in a bidirectional fashion. where each client acts as a sender and as a receiver depending on the direction of the flow. Let T be the period that network packets are transmitted by the sender, and s_i (*resp.*, r_i) be the sending time (*resp.*, arrival time) of the i^{th} packet. Hence, $n_i = r_i - s_i$ is the network delay experienced by the packet. The receiver stores the packet in a jitter-buffer until it is played at t_i, where $b_i = t_i - r_i$ is the buffer delay the packet spends in the buffer. Here, p_i, the playout delay for the i^{th} packet, is the total elapsed time from sending to playout.

To smooth the irregular arrivals of packets and to ensure that the playout delay of each frame is constant, the receiver schedules the playout times for all the frames T ms apart, with no gaps or overlaps in the played-out stream. However, since n_i is random for each frame and unknown by the receiver in advance, there is no guarantee that all frames arrive at the receiver before their playout times.

A frame is called *unconcealed* if it is unavailable to be played at the receiver at its scheduled playout time (or equivalently, if its network delay is longer than its scheduled playout delay). Speech frames may be lost for two reasons. First, the UDP packets carrying them may be lost, either in isolation or in bursts. In this case, we consider $n_i = \infty$ for the i^{th} packet. Second, due to delay spikes, a packet may be delayed beyond a point when it is too late to be played.

Observations. a) Packet-loss conditions may change in a matter of seconds, and stationary models [5] are not capa-

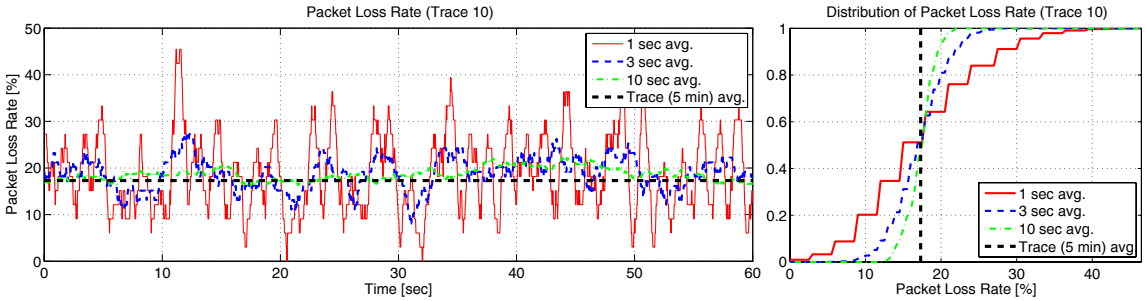

Figure 3: Temporal changes and distributions of loss rates (averaged over sliding windows of 1, 3 and 10 seconds, respectively) for Trace 10 (medium loss rate).

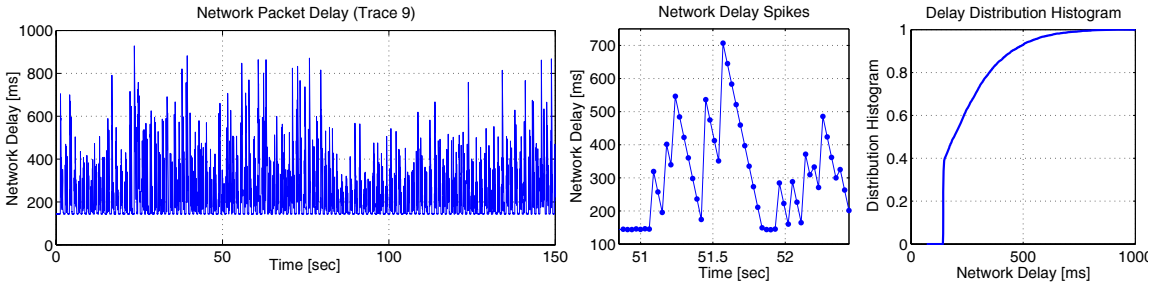

Figure 4: Temporal changes, spikes and distribution of packet delays for Trace 9 with high jitters.

ble of tracking fast changing conditions and for controlling transmission parameters in real-time. Figure 3 depicts the temporal changes in the packet-loss rate for a connection with a medium loss rate. The plot, as well as the loss-rate distribution, show that the loss rates spread from 0% to 40% for 1-second moving averages. These high variations have strong effects on the consistency of quality.

b) Network delays can change even faster than loss rates, increasing hundreds of milliseconds from that of the previous packet in a matter of a packet period (e.g. 30 ms). These conditions are refereed to as delay spikes, which is caused by a sudden decrease in the buffers in one of the routers on the path of the packets. After the spike, multiple consecutive packets may be received almost instantaneously when the congested router empties its buffers quickly. This causes those consecutive packets after the spike to experience less and less delay until the delay value reaches the level before the spike. Figure 4 depicts the temporal changes in delay behavior for an international connection with high jitters. We observe that within a second, several spikes can occur, either in an individual or in a coupled fashion.

The behavior of delay spikes and their effects on those inaudible segments of speech in real-time VoIP transmissions can be evaluated by the distribution of the heights and the frequency of the spikes. Here, the height of a spike is related to the duration of an inaudible segment when there are inadequate jitter buffers.

c) Although intra-US connections usually do not suffer from the above events, most inter-continental connections and some intra-Asian and intra-European connections suffer from one event or the other, and some suffer from both.

Trade-offs on system-controllable metrics. Our experiments show that a significant number of international connections exhibit packet-loss rates of more than 5%, which cause perceptible LOSQ degradations in the decoded speech. These losses cannot be concealed by adjusting the playout schedule alone or by the codec, but require redundant copies of each frame to be sent in subsequent packets. Further, the POS at the receiver needs to be informed of the redundancy degree in order to delay the playout schedule of all the frames, since each subsequent packet that contains a redundant copy is sent T ms later than the previous packet.

We define UC_i, the *unconcealment indicator* of frame i, to be zero when the loss of frame i can be concealed because either the original frame or a redundant copy is received before its scheduled playout time p_i; that is,

$$UC_i(p_i, R_i) \quad = \quad \begin{cases} 0 & \text{if } (n_i + (R_i - 1)T) \leq p_i \\ 1 & \text{otherwise,} \end{cases} \qquad (1)$$

where R_i is the redundancy degree of frame i (or the number of copies transmitted in the original and subsequent packets). In practice, R_i is an integer between 1 and 4.

We further define the *unconcealable frame rate* ($UCFR_i^W$) as the fraction of those unconcealed frames among a window of W frames ending with frame i:

$$UCFR_i^W(\bar{p}, \bar{R}) \quad = \quad \frac{1}{W} \sum_{j=i-W+1}^{i} UC_j(p_j, R_j),$$

where $\bar{p}$ and $\bar{R}$ are the W most recent system-controllable metrics in a vector form.

As is shown in Figure 1, UCFR depends on MED, the degree of redundant piggybacking in LC [13], and the network conditions. Recall that MED is the total delay a speech frame incurs at the sender, the network, and the receiver jitter buffer before it is played. It includes the playout delay and the delays incurred during encoding, packetization, and decoding. As the latter delays are negligible and deterministic, we use MED and playout delay interchangeably.

Figure 5 depicts the trade-offs among the playout delays, redundancy degrees, and UCFR under several network conditions for a family of POS schemes with fixed playout de-

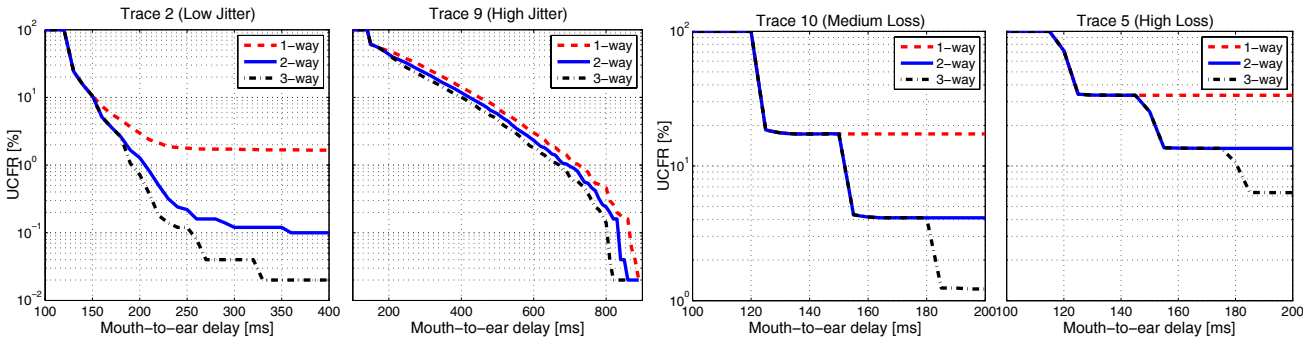

Figure 5: Trade-offs among MED, UCFR, and LC (degrees of piggybacking) for traces of different properties.

lays. It shows that the MED and the redundancy needed for achieving a given UCFR is connection dependent and cannot be configured initially at system design time.

Adaptation of system-controllable metrics. As delay and loss conditions can change quickly in a connection, especially under bursty conditions, the MED and the redundancy needed for achieving a given UCFR is time-varying and may be difficult to determine a priori. Figure 6 depicts the short-term (3 sec) and long-term (10 min) trade-offs between MED and UCFR for the various POS algorithms presented in Figure 2. It shows that the 3-sec average UCFR is widely varying and inconsistent, which can can lead to low-quality speech segments that are perceptible. As variations in quality is considered an important form of degradation, it is essential to monitor the status in real time and to adapt MEDs frequently in order to achieve a favorable trade-off between long-term average UCFR and short-term quality.

However, there are some limitations on the adaptation of the playout schedule. Changing the schedule of frames require the modification of the output waveform of a speech segment. Current codecs are not designed to change their MEDs within a talk-spurt in a way that is transparent to listeners. External alterations of the waveform within a talk-spurt [7] are possible but are computationally expensive and may result in degradations in perceptual quality. For these reasons, we consider adapting MED only at the beginning of each talk-spurt, which when performed in moderation, is not perceptible to listeners and does not limit our ability to adapt to changing network conditions. In the rest of this paper, we refer to the playout scheduling decision as one made on a talk-spurt basis and denote the playout delay (or MED) for talk-spurt k as p^k.

3. CONVERSATIONAL QUALITY

A user's perception of the quality of a VoIP system depends on the metrics that can be directly or indirectly perceived. Figure 1 shows the two components of human perceptible quality: LOSQ on the quality of one-way speech and the degradations due to delay. In this section, we study the dynamics that affects conversational quality. For simplicity, we assume that echo cancellations have been done.

Listening-only Speech Quality. As is shown in Figure 1, LOSQ depends on the speech codec used and UCFR, which in turn depends on MED, the degree of redundant piggybacking in LC [13], and the network conditions. Low-bit-rate codecs that remove frame dependencies for coding efficiency are less robust to unconcealable losses, especially

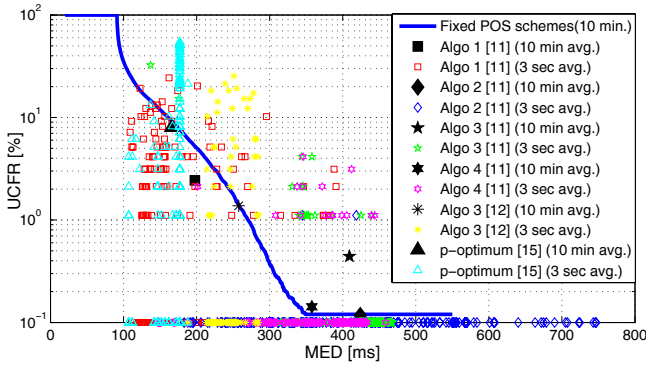

Figure 6: Trade-offs between MED and UCFR for the US-Switzerland connection in Figure 2 (averaged over 10-min. and 3-sec. windows).

when consecutive frames are lost. Most codecs employ some form of loss concealment in their decoders. However, these schemes are not adequate to guarantee LOSQ at an acceptable level for all conditions. Our previous work (Figure 2 in [14]) has shown that a PESQ of 3.00 or more is achieved only when UCFR is smaller than 3% for ITU G.729 and 5% for iLBC [2]. Hence, packet-level loss concealments are needed to provide consistent and high LOSQ for users. We use the iLBC codec in our system due to its relatively better robustness against losses.

Previous work on conversational quality. *The E-model* (ITU G.107) [3] has been designed to assist service providers during the planning process of a communication system. It uses R, a transmission rating factor on a psycho-acoustic scale that can be converted into an equivalent MOS measure [3], to model the effect of one-way delays on conversational speech quality. Figure 7 depicts the effect of MED on MOS in the E-model for a perfect listening-only speech. Because the metric calculated in the E-model is speech-independent and is based on tabulated values on the effects of the codec used and packet losses in the average sense, it alone is not adequate for capturing CVCQ in a real-time conversation.

Combined E-model and PESQ. A conversational quality metric MOS_c was proposed to combine the E-model and PESQ [15]. Here, PESQ is converted to the scale of R and substituted into the E-model to represent the impairment due to I_e (for codec and packet losses). However, since PESQ requires both the original and the degraded speech waveforms, the metric can only be measured off-line or in

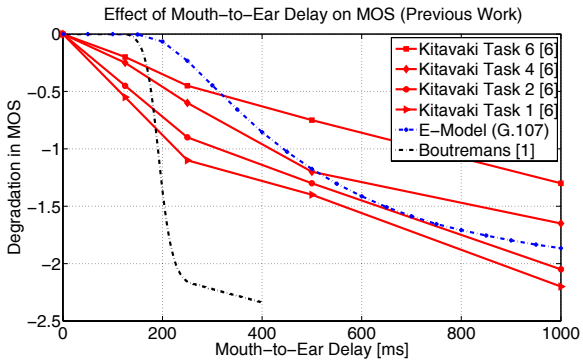

Figure 7: Effects of MED on MOS [6], E-model [3], and a conversion by Boutremans and Boudec [1].

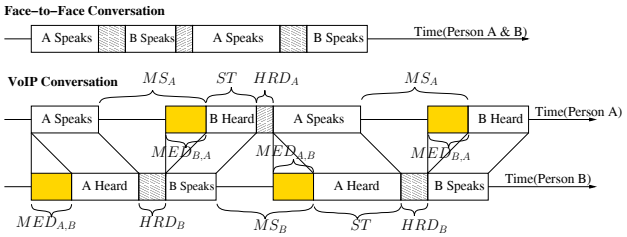

Figure 8: Conversational dynamics.

Table 2: Statistics of two face-to-face conversations.

Conversation Type	Avg. single-talk duration	Avg. HRD duration	# of switches	Total Time
Social	3,737 ms.	729 ms.	7	35 sec.
Business	1,670 ms.	552 ms.	15	35 sec.

an intrusive way. A subsequent study was proposed to use regression models to predict PESQ on-line, which was discussed in Section 1. The solution addresses the issue on the tabulated effects of the codec and losses in the E-model, but does not consider the effects of conversational conditions.

Call Clarity Index (CCI). The ITU P.561 and P.562 [8] standards specify the objective parameters to be collected via INMD (in-service non-intrusive measurement device) for analyzing and interpreting INMD voice service measurements. The model, developed by British Telecom in 1998, relates those parameters collected by INMD to customer-opinion prediction. It was designed for PSTN networks whose delays are short and constant throughout a conversation. However, it is not suitable for long-delay packet-switched Class D networks that may include possibly non-linear and time variant signal processing devices, such as echo control and speech compression. Thus, there is currently no customer opinion model for VoIP transmissions that considers all aspects required by P.561.

Other previous work. In an NTT study [6], conversational experiments were conducted in the form of tasks by two parties using a voice system with adjustable delays. The tasks studied range from reading random numbers, to verifying city names, and to free conversation with varying average single-talk duration. Subjective-quality results reveal that the degradation in MOS is more pronounced when a task requires shorter single-talk durations (Figure 7). However, the study does not consider the effect of losses.

A utility function [1] was proposed to represent the effects of MED, in which a conversation is perceived to be half-duplex and quality degrades suddenly after some MED threshold (Figure 7). The goal of the study was to incorporate the effect of MED on the choice of FEC, rather than studying the effects of MED on conversational quality.

Conversational dynamics. The quality of a conversation in a VoIP system depends on the naturalness and the rhythm of the conversation. The dynamics is different between a face-to-face conversation and one over a network with delays. In a face-to-face conversation, users have a common reality in the perception of the sequence and the timing of events. However, as is illustrated in Figure 8, a conversation over a delayed channel may lack a common perspective and may lead to multiple realities.

We define HRD_B (*human-response delay* from B's perspective) as the duration after B perceives that A has stopped talking and before B starts talking, during which B thinks about how to respond to A's speech. However, the same delay is perceived to be longer from A's perspective, which we call MS_A (*mutual silence* from A's perspective). The relation between MS_A and HRD_B, where $MED_{A,B}$ is the MED between A's mouth and B's ear, is as follows:

$$\begin{aligned} MS_A &= MED_{A,B} + HRD_B + MED_{B,A} \\ \text{and } MS_B &= MED_{B,A} + HRD_A + MED_{A,B}. \end{aligned} \quad (2)$$

During a VoIP session, a user does not have an absolute perception of MED because the user does not know when the other person will start talking. However, by interactively perceiving the indirect effects of MED, such as MS and CE, the user can deduce the existence of MED. In short, MS, CI, and CE are user-perceptible quality metrics that are intimately affected by MED, a system controllable metric. Next, we formally define CI and CE.

Conversational interactivity (CI). Based on user observable metrics, we define the *interactivity factor* (CI_i^j) of single-talk speech segment j (ST_j) from person i's perspective to be the ratio of MS_i observed by i before ST_j is heard and HRD_i waited by i after ST_j is heard:

$$CI_A^j = \frac{MS_A^{j-1}}{HRD_A^j}, \qquad CI_B^j = \frac{MS_B^{j-1}}{HRD_B^j}. \quad (3)$$

In a face-to-face conversation, CI would be approximately 1. However, CI increases as the round-trip delay increases. If the asymmetry in the perceived response times increases, humans tend to have a degraded perception of interactivity that will result in the degradation of the conversational quality. One possible effect is that, if A perceives that B is responding slowly, then A tends to respond slowly as well.

Conversational efficiency (CE). Another effect of MED on a VoIP conversation is that it takes longer to accomplish a task when there are communication delays (Figure 8). We define the ratio of the time a conversation takes in a face-to-face setting to the time to carry out the same conversation in a VoIP setting to be the relative CE:

$$CE = \frac{\sum_j \sum_{A,B}(ST + HRD_{F2F})}{\sum_j \sum_{A,B}(ST + HRD_{VoIP} + MED)}. \quad (4)$$

Since a conversation over a network is charged according to its duration, the same conversation might cost more for a network with longer MEDs. This effect is especially pronounced in international and mobile phone calls, where both the network delay and the per-minute price are higher.

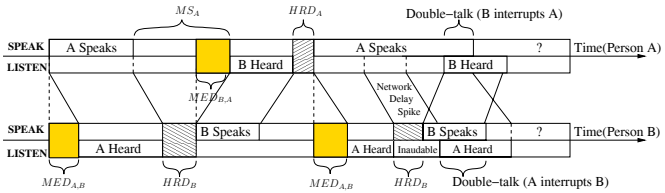

Figure 9: Occurrence of a double-talk due to a lack of adequate reaction to network-delay spikes.

Table 2 shows the statistics of two face-to-face conversations with different average ST durations. The following estimates CE as a function of MED for the VoIP sessions that correspond to the conversations in Table 2, assuming that $HRD_{VoIP} = HRD_{F2F}$, independent of MED.

$$CE \approx \frac{TotalTime}{TotalTime + (Num.of switches) * MED}. \quad (5)$$

Note that changes in CE are almost undetectable for small MEDs and slow switching frequencies, but that CE decreases with increasing MEDs and even more rapidly when the average single-talk duration is short.

Double Talk. When large spikes in network delays are not detected by a VoIP system, the MED will not be adjusted, and a considerable amount of consecutive frames may be lost for a duration that is perceived by the receiving client. Depending on the duration and the frequency of the spikes, an utterance, a word, or even a sentence may be inaudible or unintelligible at the receiver. If this scenario occurs during a speech utterance, the listener either assumes that the speaker has stopped talking and starts uttering his/her own response, or asks the speaker to repeat the last words or sentence. In either case, the initial speaker, unaware of the lost perception of the listener, would most likely continue speaking and cause a collision of the speeches (double-talk). A person observing this collision struggles to resolve the situation either by waiting longer for the other person to respond or by repeating the previously spoken utterances. Further, as depicted in Figure 9, the order of the utterances may be perceived differently by different parties, disrupting the rhythm of a natural interactive conversation and causing confusion and degradation in the perceived quality.

Adaptation of Human Behavior. In case of extreme difficulties in comprehension, such as extreme delays in getting a response or extremely low listening quality, users either hang up and re-dial, or change their speech style in order to ease the effort needed. This style change usually involves talking slowly, talking in longer batches, or waiting for acknowledgment gestures. Users who are forced to take these behavioral changes generally have significantly lower satisfaction of the conversation. Further, this behavioral change might not be acceptable in some languages, cultures, and business-related or mission-critical communication tasks.

Trade-offs on metrics of conversational quality. The CVCQ of a conversation (or a portion of it) can be represented by the following triplet:

$$CVCQ = \{LOSQ(MED, R), CE(MED), CI(MED)\}$$

that corresponds to a point in a 3-dimensional space, where each axis represents a human perceptible quality attribute.

Figure 10 depicts the trade-off between CE and CI as a function of MED for the conversations in Table 2. We see

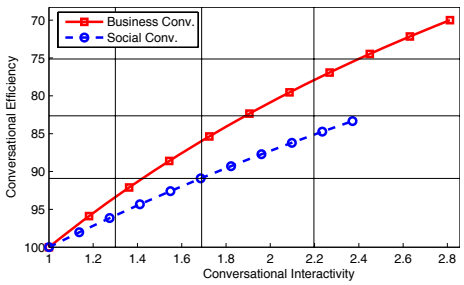

Figure 10: Effect on CI and CE when MED changes for the two conversations in Table 2.

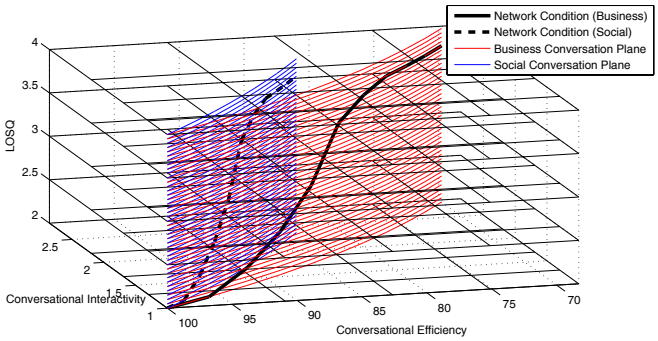

Figure 11: The planes representing the conditions of the conversational type. The curve on each plane represents the conditions imposed by the network.

that the degradations in CI and CE are less pronounced for the social conversation with a lower switching frequency.

Figure 11 illustrates the trade-offs among CI, CE, LOSQ, and the system-controllable MED in CVCQ. The red and blue planes parallel to the LOSQ axis represent the conditions imposed by the conversational type (business versus social in Table 2). For given network and conversational conditions, the black curve on each plane represents the trade-offs among CI, CE, and LOSQ, when parameterized by the system-controllable MED and subject to the constraints imposed by the network and the conversation.

The trade-offs in Figure 11 are very complex and cannot be represented in a closed form because they involve subjective judgments on the quality of a conversation. Further, it is not possible to give a total order of all the alternatives because the trade-offs among the metrics lie on a Pareto optimal boundary. In other words, it may not be possible to compare two conversations, one with high LOSQ but low CI and CE and another with high CI and CE but low LOSQ. To this end, we propose to use *just noticeable difference* (JND) as a vehicle to represent the discrete nature of the trade-offs and their partial orders. We only compare a conversation belonging to a JND block with that of its immediate neighbors where one of the CVCQ attributes has changed.

JND is a concept that is commonly used to explain the sensitivity of human perception to sensory inputs, such as pitch of sound and intensity of light. In general, JND is refereed to a difference in the physical sensory input that results in the detection of the change 50% of the time. For a variety of sensory inputs, including the perception of temporal durations, JND has been shown to obey a ratio with respect to the original input level. Here, we use JND as a

Table 3: User responses in comparison MOS tests.

User Response	CMOS Score
A is strongly preferred over B	−2
A is preferred over B	−1
A and B are preferred equally	0
B is preferred over A	1
B is strongly preferred over A	2

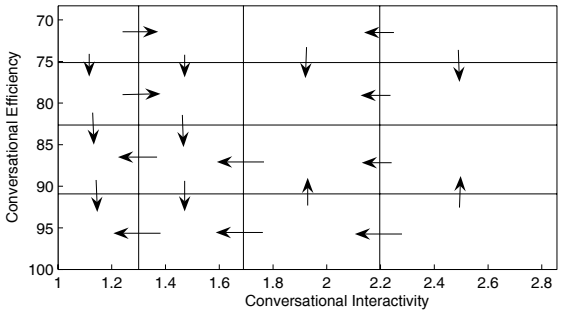

Figure 12: Relative user opinion on the CI-CE plane. The size of an arrow indicates the degree of preference of one alternative over another.

framework to discretize the CVCQ space into a finite set in order to help quantify the user opinion, without limiting the accuracy of the representation.

Let k_{CI} be the ratio of increased CI that is considered indistinguishable from CI_{JND} in a given JND block. Similarly, let k_{CE} be the ratio of decreased CE that is considered indistinguishable from CE_{JND} in this block. That is,

$$k_{CI} = \frac{CI}{CI_{JND}}, \quad k_{CE} = \frac{CE_{JND}}{CE} \text{ within a JND block.}$$

In this paper, we use $k_{CI} = 1.3$ and $k_{CE} = 1.1$. For instance, starting with $CE_{JND} = 1.0$, those CE's greater than $\frac{1.0}{k_{CE}} = 0.91$ are in the same JND block. Likewise, starting with $CI_{JND} = 1.0$, those CI's greater than $1.0k_{CE} = 1.3$ are also in the same block. With respect to LOSQ, since LOSQ based on PESQ is already calculated on a psycho-acoustic scale, we use a linear scale of 0.5 to discretize LOSQ. Using this approach, Figure 11 represents CVCQ as one of 64 blocks.

$$CVCQ_{JND}(JND_{LOSQ}, JND_{CI}, JND_{CE}), JND_x \in \{1, 2, 3, 4\}.$$

We conduct comparative listening tests off-line in order to generate decisions for guiding POS at run time. However, the CVCQ curve depends on the conversational and network conditions and change over time. In order to provide a complete partial ordering of all JND blocks, we collect comparative user opinion in our listening tests for all six neighbors with respect to each of the 64 JND blocks. To eliminate possible effects of the speakers and the conversational topic on the listeners' opinion of quality, we use the same speech segments extracted from a real conversation. We then modify the frame loss pattern, HRD, and MED in order to meet the LOSQ, CI, and CE criteria of each specific JND block.

Table 3 defines the *comparison MOS* between two conversations (B compared to A) in our listening tests (similar to ITU P.800 Annex E, Comparison Category Rating method):

$$CMOS(A \rightarrow B) \in \{-2, -1, 0, 1, 2\}. \quad (6)$$

Figure 12 depicts a subset of the results of the comparative listening tests. Here, preferences are represented as arrows, with each of an appropriate length towards the direction of the preferred block and drawn normal to the boundary.

In general, users prefer conversations with higher LOSQ, better CI, and higher CE. Further, when a particular metric is poor with respect to other metrics, a user places more importance to improving the metric that is suffering more than those metrics that are relatively good.

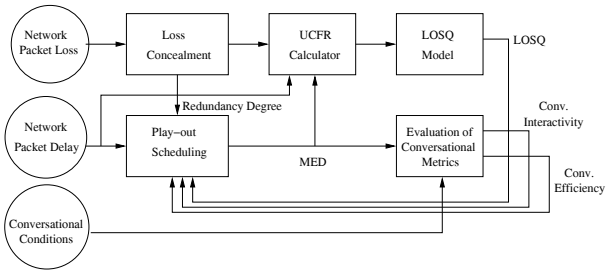

Figure 13: Proposed system control by POS/LC.

4. VOIP SYSTEM CONTROL VIA POS/LC

The goal of our system control is to mitigate the degradations caused by the imperfections of the underlying network and to provide high and consistent conversational quality to users. Based on the trade-offs between those system-controllable metrics as a function of the network condition (Section 2) and the trade-offs between those user-observable metrics as a function of the system-controllable metrics (Section 3), we discuss in this section our POS/LC-control schemes that provide end-to-end trade-offs among the network condition and the user-perceptible quality metrics.

In our POS/LC-control schemes, we invoke our LC and POS in a closed-loop fashion (Figure 13) in order to dynamically pick an appropriate redundancy degree for each packet and a playout schedule for each talk-spurt.

Our LC scheme uses the network-loss information to select a redundancy level at the receiver and relays that decision to the sender. In our previous work [13], we have observed that packet losses can be effectively concealed to an acceptable and stable level by piggybacking redundant copies of the original information in subsequent packets. Since each coded frame is very small, redundant piggybacking will not increase the payload of each packet to exceed the MTU. Here, we use 2% as our target unconcealed packet-loss rate after redundant piggybacking:

$$R_{i+FBD} = \min\{R \mid UCFR_i^W(\bar{p}, \bar{R}) \leq 2\%\}. \quad (7)$$

where FBD is the feedback delay in number of packets when relaying the redundancy degree to the sender, and $W = 100$ (≈ 3 sec) is the length of the recent window of frames used to calculate UCFR. To evaluate only the effects of network losses (not late losses), we use ∞ for all the members of $\bar{p}$. We also assume a constant redundancy degree throughout the W packets in our calculation.

Our POS scheme uses the network-delay information, redundancy decision, and the predictions of CVCQ metrics for an upcoming talk-spurt in order to select a suitable playout schedule for that talk-spurt. In deciding the optimal playout delay for an upcoming talk-spurt, POS needs to estimate the plane in the 3-dimensional CVCQ space (Figure 11) that corresponds to the predicted conversational condition and

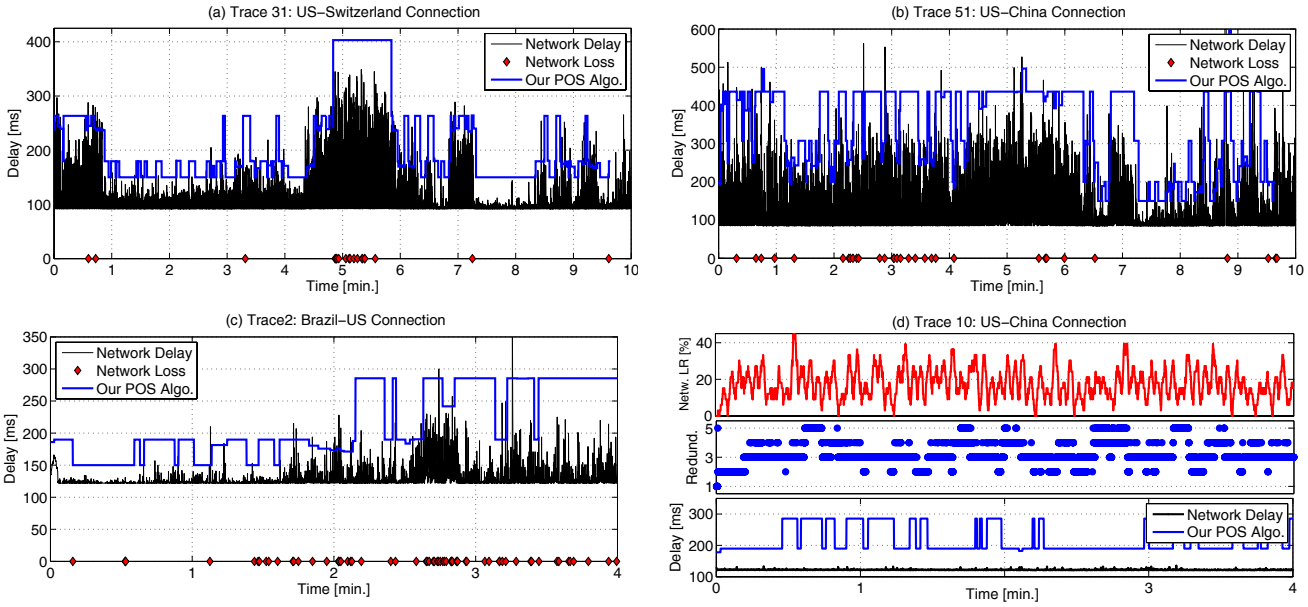

Figure 14: Network delays and POS/LC-control decisions made by our JND-based POS and LC schemes for 4 international connections: (a) and (b) suffer from high and time-varying delay spikes; (c) suffers from low packet losses and moderate time-varying delays; and (d) suffers from medium packet losses.

the curve on that plane that corresponds to the predicted LOSQ under various network conditions. These estimations are heuristic and are not intended to be optimal but rather in facilitating the use of the JND-based decision making. Once the CVCQ curve has been estimated, POS finds a JND block that is Pareto-optimal among its neighbors, which leads to an MED within the optimal block. The tasks carried out in our POS scheme consists of three components:

a) *Predicting CI and CE for the next talk-spurt.* The conversational condition is usually stable and only changes slowly during a conversation. Hence, CI and CE, as a function of candidate MEDs, can be estimated accurately by monitoring the silence and the voiced durations at the receiver. The predicted CI and CE allow us to establish a plane in the 3-dimensional CVCQ space (Figure 11).

b) *Predicting LOSQ for the various network conditions in the next talk-spurt.* Given the plane in the CVCQ space for the next talk-spurt, we need to estimate the LOSQ curve on that plane with respect to various network conditions.

The E-model is not suitable for this task because its estimate of LOSQ (represented in I_e) is based on some tabulated effects of the speech codec used and the loss rate. The level of accuracy in I_e is very crude and is inadequate.

Another possibility is to evaluate LOSQ using PESQ that can more accurately capture the effects of losses. Because the evaluation of PESQ requires both the original as well as the degraded speech waveforms, it can be computed either at the sender by relaying network-loss statistics periodically there, or at the receiver by conducting an intrusive transmission of a reference signal to the receiver. Both alternatives incur delays and transmission overheads and are impractical under dynamic network conditions.

In our approach, we approximate PESQ as a function of UCFR by conducting off-line experiments. Using patterns of loss frames derived from Internet traces, we learn to classify different network conditions and the corresponding PESQ. Because real traces exhibit bursty loss patterns, our approach leads to significantly more accurate estimates of PESQ as compared to those in the previous work [15] that uses IID loss patterns. By combining the delay distribution statistics collected at run-time with the LOSQ (PESQ) model, we can accurately estimate the relation between LOSQ and MED in the overall CVCQ trade-off curve. We do not show the details here due to space limitation.

c) *Adapting MED to the predicted network and conversational conditions.* We define $\mathcal{S}$ to be the set of JND blocks that the CVCQ trade-off curve passes through. $\mathcal{S}$ is an ordered list of triplets, each corresponding to the coordinates of a JND block. To ensure neighborliness, only one member of the triplet can be different from the previous triplet in the list. The list always starts at (4,1,1) that corresponds to the block where MED is the lowest and LOSQ is the worst. It ends at a block whose $JND_{LOSQ} = 1$ (best LOSQ).

At the beginning of a talk-spurt, POS considers the possible adaptations of MED. Given the CVCQ trade-off curve represented as a set of JND blocks in $\mathcal{S}$, POS starts from the block that corresponds to the MED of the last talk-spurt. It has three possible decisions while staying in $\mathcal{S}$: i) do not change MED; ii) increase MED to cross into a neighboring block; ii) decrease MED to cross into the neighboring block in the other direction. This decision is made in accordance to the CMOS measure in (6). By traversing multiple blocks in $\mathcal{S}$, one at a time, the system will eventually settle in a Pareto-optimal block among the user-perceptible metrics, where there is no incentive to cross into a neighboring block.

There are a range of MEDs at which the system operates in the optimal JND block, as all these operating points are indistinguishable by users. To improve the robustness of POS against uncertainty in changing network conditions and to prevent the operating point from drifting to a suboptimal block, we heuristically calculate the MED used by a linear combination of the shortest and the longest MEDs allowed in that block. Our heuristic emphasizes longer MEDs when the network condition may change rapidly.

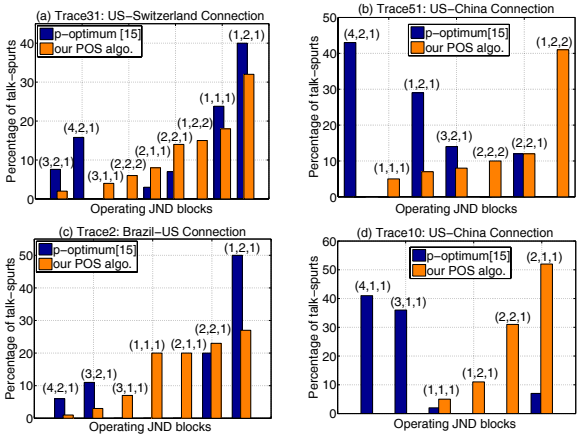

Figure 15: Percents of talk-spurts that our proposed schemes and the p-optimum algorithm operate in each user-perceptible JND block for 4 connections.

5. EXPERIMENTAL RESULTS

In this section, we present the performance of our POS/LC-control schemes for a simulated VoIP session using traces from four international connections collected in the PlanetLab. In our tests, we simulated a sender-receiver pair carrying a business conversation (Table 2), while assuming a symmetric sender-receiver pair in the other direction with the same control schemes. The sender used the iLBC codec with a 30-ms framing option and sent one encoded voice frame in each UDP packet. After receiving a moving window of packets, the receiver piggybacked the LC decision to the sender in the next reverse-path speech packet. The sender then adapted its redundancy degree in the next forward-path packet sent. For each talk-spurt, the receiver adjusted the MED at the beginning as needed, by adding or skipping silence segments in the received waveform. It limit the length of the skipped silence segments to 30% of the mutual-silence duration preceding the talk-spurt adjusted.

Figure 14 depicts the decisions made by our POS/LC-control schemes, along with the delay and loss conditions for the four connections. We observe that our POS algorithm schedules MED values between 150 ms and 450 ms and tracks the changing network conditions closely, while making discrete adjustments when needed in order to keep the conversational quality in a user preferred state. We show the packet-loss rates and redundancy decisions in detail for Connection (d), which exhibits medium-loss rates. We omit this information for the other connections, since their loss rates are low and consequently their redundancy degree is rarely greater than 1.

Figure 15 compares the performance of our POS/LC-control schemes and the p-optimum algorithm [15]. It depicts the percent of time that each algorithm operates in each of the user-perceptible JND blocks for the four connections in Figure 14. We represent the JND blocks as a discretized CVCQ triplet (LOSQ,CI,CE), where (1,1,1) (*resp.* (4,4,4)) corresponds to the best (*resp.* worst) conversational condition. For Connections (a) and (c), we observe that p-optimum operates more in the most preferred JND block with respect to our algorithm. However, p-optimum operates considerable amounts of time in the least preferred JND blocks (25% in (a) and 17% in (c)), whereas our scheme rarely does. The inconsistency of p-optimum is further depicted for Connections (b) and (d), where the majority of the talk-spurts are in the least preferred JND block. The poor performance of p-optimum is due to its optimistic estimates of the codec robustness to consecutive losses (for Connection b) and a lack of an accompanying redundancy-based loss-concealment scheme (for Connection d).

In summary, our JND-based POS/LC-control schemes perform well by delivering consistent and desirable CVCQ trade-offs for the network conditions tested. When compared to the p-optimum algorithm, it has more consistent behavior across connections of different losses and delays and does not operate in blocks with very low LOSQ.

6. REFERENCES

[1] C. Boutremans and J.-Y. L. Boudec. Adaptive joint playout buffer and FEC adjustment for Internet telephony. In *Proc. IEEE INFOCOM*, volume 1, pages 652–662, 2003.

[2] S. Anderson, et. al. Internet low bit rate codec (iLBC). http://www.ietf.org/rfc/rfc3951.txt, Dec. 2004.

[3] ITU. G.107. The E-model, a computational model for use in transmission planning.

[4] ITU G.114. One-way transmission time.

[5] S. Fosse-Parisis, J.-C. Bolot and D. Towsley. Adaptive FEC-based error control for Internet telephony. In *Proc. IEEE INFOCOM*, vol:3, pp:1453-1460, 1999.

[6] N. Kiatawaki and K. Itoh. Pure delay effect on speech quality in telecommunications. *IEEE Journal on Selected Areas of Comm.*, 9(4):586–593, May 1991.

[7] Y. J. Liang, N. Faber, and B. Girod. Adaptive playout scheduling and loss concealment for voice communication over IP networks. *IEEE Trans. on Multimedia*, 5(4):532–543, Dec. 2003.

[8] ITU P.562. Analysis and interpretation of INMD voice-service measurements.

[9] ITU P.800. Methods for subjective determination of transmission quality.

[10] ITU P.862. Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs.

[11] R. Ramjee, J. Kurose, D. Towsley, and H. Schulzrinne. Adaptive playout mechanisms for packetized audio applications in wide-area networks. In *Proc. 13th IEEE Annual Joint Conf. on Networking for Global Communication*, volume 2, pages 680–688, 1994.

[12] D. T. S. B. Moon, J. Kurose. Packet audio playout delay adjustment: performance bounds and algorithms. *Multimedia Systems*, 6(1):17–28, Jan. 1998.

[13] B. Sat and B. W. Wah. Speech- and network-adaptive layered G.729 coder for loss concealments of real-time voice over IP. In *IEEE Int'l Workshop on Multimedia Signal Processing*, Oct. 2005.

[14] B. Sat and B. W. Wah. Analysis and evaluation of the Skype and Google-Talk VoIP systems. In *Proc. IEEE Int'l Conf. on Multimedia and Expo*, July 2006.

[15] L. Sun and E. Ifeachor. New models for perceived voice quality prediction and their applications in playout buffer optimization for VoIP networks. In *Proc. IEEE Communication*, volume 3, pages 1478–1483, 2004.