

# Analyzing Voice Quality in Popular VoIP Applications

Batu Sat and Benjamin W. Wah  
*University of Illinois at Urbana-Champaign*

This article presents a technique for comparing the speech quality of VoIP systems on networks suffering from congestion.

Voice over IP (VoIP), which provides real-time speech communication between two users in a way that closely resembles a face-to-face conversation, has had a significant impact on the multibillion-dollar telecommunication industry. The promise of less expensive phone calls with comparable quality and better features than public switched telephone networks (PSTNs) has accelerated VoIP's adoption, both in businesses and homes. Its better integration with various forms of collaborative communications, such as instant messaging, email, and voicemail, has made it suitable for future communication solutions, but there remain several issues that are unique to VoIP systems and require further analysis.

One issue is that a VoIP node can reside on any one of several types of hardware interfaces, such as a laptop, PDA, smartphone, or dedicated handset. Another issue is that when a conversation is conducted over the Internet, speech segments can experience delays, jitters, and losses. The quality of a conversation depends on two factors that are directly or indirectly perceived by users: the quality and the latency of the one-way speech segments received. In contrast to face-to-face conversations, the delays incurred in the reception of VoIP speech segments can lead to asymmetry in silence durations in between turns and cause inefficiency in communication. In these cases, each user will experience speech segments that are

separated by silence periods of alternating long and short durations. This asymmetry might lead to a perception that the other user is responding slowly to the conversation.

Due to path-dependent, nondeterministic, and nonstationary network behavior, the factors that affect conversational quality might vary over time and counteract each other. For example, the one-way quality and the delay incurred in the transmission of speech segments from the mouth of a speaker to the ear of a listener (*mouth-to-ear delay*, or MED) counteract each other in terms of their effects on conversational quality. On the one hand, speech segments will have a higher chance of being received and consequently provide better one-way quality if the receiver waits longer. On the other hand, this additional delay will result in a longer MED, which leads to lower perceptual quality.

The impact of delays on conversational quality also depends on the turn-switching frequency. For instance, an MED of 300 milliseconds (ms) can significantly degrade a conversation's symmetry and efficiency if participants take frequent turns, but will be virtually imperceptible if users take a long time (say 10 seconds) in each turn. Variations in one-way quality and latency might cause doubletalk and interruptions that further degrade conversational quality. Ideally, the trade-offs between delay and quality should be dynamic and respond to changing conditions, with the receiver either adapting its MED in order to achieve a consistent speech quality or keeping a consistent MED but allowing the speech quality to vary.

While the evaluation of speech communication systems has been an important field for both academia and industry for decades, the introduction of VoIP systems, has created a new set of issues that require new evaluation methods. To date, there has been only a small number of comprehensive evaluations of commonly used VoIP systems. In our previous work, we have attempted to take on this task with increasing levels of analytical sophistication.<sup>1,2</sup> The most straightforward analysis approach is through subjective tests. However, subjective tests cannot be used in large-scale experiments due to their large overhead, the high costs of listening experts, and their unrepeatable nature.

In addition, the evaluation of some commercial VoIP systems is hampered by their

proprietary nature. Most of these systems use codecs and algorithms not freely available for testing. As a result, it's impossible to obtain some of the critical parameters, such as the amount of packets unavailable at the decoder due to network losses or delays. We therefore must evaluate current systems by treating them as black boxes whose input and output waveforms are the only information available. Even so, both subjective and objective metrics are important in the evaluations because each alone is inadequate.

### **Network environment**

Our experiments show that public IP networks exhibit path-dependent, unreliable, and time-varying characteristics. Table 1 (next page) summarizes some of the results from our experiments conducted on PlanetLab (see <http://www.planet-lab.org>). The results are based on 71 unique connections among 22 nodes, six of which are in North America, eight in Europe, and eight in Asia. Because 59 of these connections are intercontinental, we interpret the observations accordingly. We have classified each connection by a triplet of delay, jitter, and loss. About one-third of these connections have low delay, low jitter, and low loss (L, L, L). Except for (L, H, H) and (H, H, H), there are at least four connections in each of the remaining six classes. For testing VoIP systems under different conditions, we have chosen a representative connection in each class. Our study leads to the following observations.

There are two events that cause the quality degradation of received speech frames. In some cases, packets carrying speech frames might be lost in the network, either in single packets or in multiple, consecutive packets. It's also possible for packets to be delayed beyond a point when they are too late for playback. In both cases, the receiver won't recover these packets without redundant transmissions. The Internet's loss behavior can change in a matter of seconds, and stationary models<sup>3</sup> aren't helpful for tracking these quickly changing conditions. Figure 1a depicts the temporal changes in packet losses for a connection with medium loss rate, where loss rates are calculated over a sliding window of one second. The data shows that the loss rates fluctuate between 3 and 51 percent and are unpredictable. The use of a 1-second averaging window is meaningful because several words can be

uttered within this interval, and the words can be unintelligible if several consecutive packets are lost.

Most intracontinental connections in North America and Europe have mean propagation delays of less than 75 ms, whereas most intercontinental connections and some intracontinental connections within Asia exhibit delays in excess of 150 ms. The delays experienced by IP packets can change quickly in a short interval (called jitters), increasing by hundreds of milliseconds from the delay of the previous packet in a packet period of 30 ms. These conditions are commonly referred to as *delay spikes*, which indicate sudden congestion in an intermediate router on the packets' path. When congestion is resolved, multiple consecutive packets can be received almost instantaneously. These consecutive packets experience decreasing delays until the delay value reaches the level before the spike. This behavior indicates that the congested router has emptied its buffers quickly after the congestion.

Figure 1b depicts the temporal changes in network delays for a Taiwan–US international connection with high jitters. It shows that several spikes can occur within 1 second, either in an individual or in a coupled fashion. To limit degradations caused by jitters, VoIP clients commonly employ play-out schedulers (POSS) that adjust the time waited before playing out the received speech frames. These schemes will incur additional play-out delays and extend the MED.

### **Conversational dynamics**

In a two-party conversation, each participant takes turns speaking his or her thoughts and listening to the other participant. There is a silence period during turn-taking (switching) when the current speaker ceases talking and the listener begins to speak. Both participants in a face-to-face conversation perceive this silence period. That is, users have a common reality in the perception of the sequence and the timing of events. This common reality is split into two realities that are experienced by the two participants when the conversation is carried out over a channel with delays. In this section, we analyze the effects of delays on the conversational dynamics.

Table 1. Internet traces collected in July and August 2007. Each set is from one source sent to seven destinations over a 10-minute duration with a 30-millisecond packet period.

Set	Characteristics (L, H, M)*			Hour (central standard time)	Source (location)	Destination (# in Asia, US, Europe, respectively)	Mean delay (ms)		JT60** (%)		Loss rate	
	Delay	Jitter	Loss rate				Min.	Max.	Min	Max	Min	Max
1	L	L	L	20:00	US	1, 2, 4	42.2	94.6	0.00	0.15	0.00	0.00
2	H	L	L	18:00	China	0, 3, 4	107.3	190.4	0.00	3.5	0.00	0.01
3	H	L	H	23:00	Hong Kong	0, 3, 4	101.2	204.3	0.00	1.64	14.7	22.7
4	H	H	L	22:00	Taiwan	1, 3, 3	198.0	280.4	68.3	72.2	0.14	0.22
5	M	L	L	20:00	Czech Rep.	2, 3, 2	56.0	158.4	0.45	0.97	0.00	3.39
6	M	H	L	17:00	US	2, 2, 3	74.9	170.9	5.2	6.2	0.00	4.33
7	M	L	H	1:00	Hong Kong	1, 3, 3	85.4	195.9	0.00	1.6	15.3	22.8
8	M	L	M	11:00	Canada	2, 2, 3	52.4	147.3	0.00	0.83	0.00	16.9
9	M	M	L	5:00	UK	2, 3, 2	26.5	139.9	0.00	8.10	0.00	3.2
10	H	M	M	1:00	China	0, 4, 3	103.7	198.9	1.2	6.6	1.9	8.6
11	M	M	M	8:00	Hungary	3, 2, 2	22.6	190.6	0.00	79.0	0.00	25.1

\* Delays are classified into low (< 100 ms), high ( $\geq 100$  ms), and mixed (a combination of both). Similarly, jitters are classified into low (< 5 % in JT60), high ( $\geq 5$  % in JT60), and mixed; and losses into low (< 5 %), high ( $\geq 5$  %), and mixed.

\*\* JT60 are jitters larger than 60 ms with respect to mean delay.

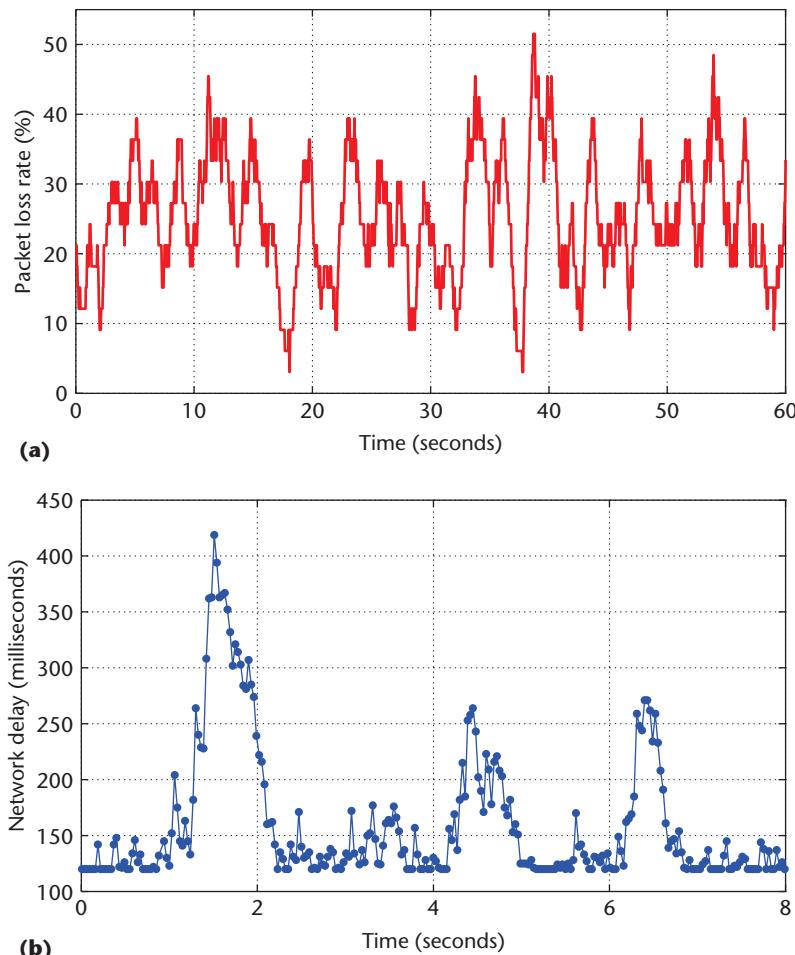


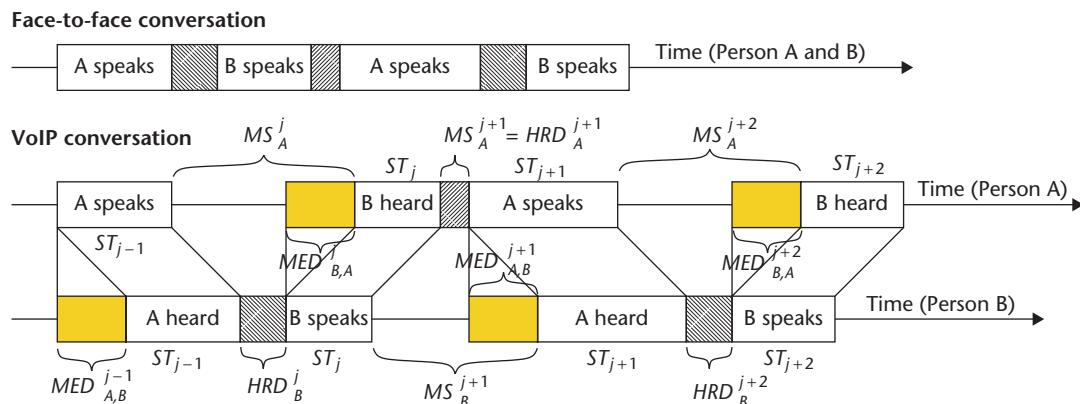
Figure 1. Traffic behavior (delay and loss rate) of two PlanetLab connections in Table 1: (a) trace set 8 from Canada to Portugal and (b) trace set 4 from Taiwan to US.

### Response delay and mutual silence

We first define the silence durations observed during turn-taking. Because there are two perspectives, A's and B's, we start from the perspective of the current speaker. We define *human response delay* from B's perspective ( $HRD_B$ ) as the period after B perceives that A has stopped talking and before B starts talking, during which B thinks about a response to A's speech. However, the same event is perceived as longer from A's perspective, which we define as  $MS_A^j$ , the *mutual silence* (MS) before the  $j$ th single-talk speech segment ( $ST_j$ ) is spoken and heard. Let  $MED_{A,B}^j$  be the MED between A's mouth and B's ear for transmitting  $ST_j$  from A to B, the relation among MS, HRD, and MEDs is as follows (see Figure 2):

$$\begin{aligned} MS_A^j &= MED_{A,B}^{j-1} + HRD_B^j + MED_{B,A}^j, \\ MS_A^{j+1} &= HRD_A^{j+1}, \\ MS_B^j &= HRD_B^j, \\ MS_B^{j+1} &= MED_{B,A}^j + HRD_A^{j+1} + MED_{A,B}^{j+1} \end{aligned}$$

During a VoIP session, a user doesn't have an absolute perception of MED because the user doesn't know when the other person starts talking. However, by perceiving the indirect effects of MED, such as MS, the user can deduce the existence of MED. In the rest of this section we present other metrics that capture the effects of delay on conversational dynamics and that users can perceive.



**Figure 2.**  
*Conversational dynamics in face-to-face and VoIP communications.*

### Conversational symmetry

Symmetry is related to the activities of the entities that affect each other. In the context of speech communication, turn-taking is the interaction between the participants. For this reason, we define conversational symmetry (CS) based on the user perceptible MS between turn-taking. Because the perception of temporal events is user dependent, we define *conversational symmetry of A* ( $CS_A$ ) as the ratio of the maximum and the minimum MSs experienced recently by A:

$$CS_A = \frac{\max_j MS_A^j}{\min_j MS_A^j}, \quad CS_B = \frac{\max_j MS_B^j}{\min_j MS_B^j}$$

In a face-to-face conversation, MS and HRD are perceived to be equal; thus,  $CS_A$  and  $CS_B$  are approximately 1. However, as the round-trip delay increases, the silence periods perceived during turn-taking are no longer symmetric. If the asymmetry in the perceived response times increases, humans tend to have a degraded perception of symmetry that will result in degradation of the conversation's quality. One possible effect is that, if A perceives that B is responding slowly, then A tends to respond slowly as well.

### Conversational efficiency

Another effect of communicating over a channel with delays is that it takes longer to

accomplish a task with respect to the same conversation in a face-to-face setting. Because VoIP providers frequently charge users according to a conversation's duration, a task will cost more when a channel suffers from delays. This effect is especially pronounced in international and mobile phone calls, in which both network delay and per-minute charges are higher. We define *conversational efficiency* (CE) as the ratio of the duration the participants actively speak or listen to the total call duration:

$$CE = \frac{\text{total speaking time} + \text{total listening time}}{\text{total time including silence}}$$

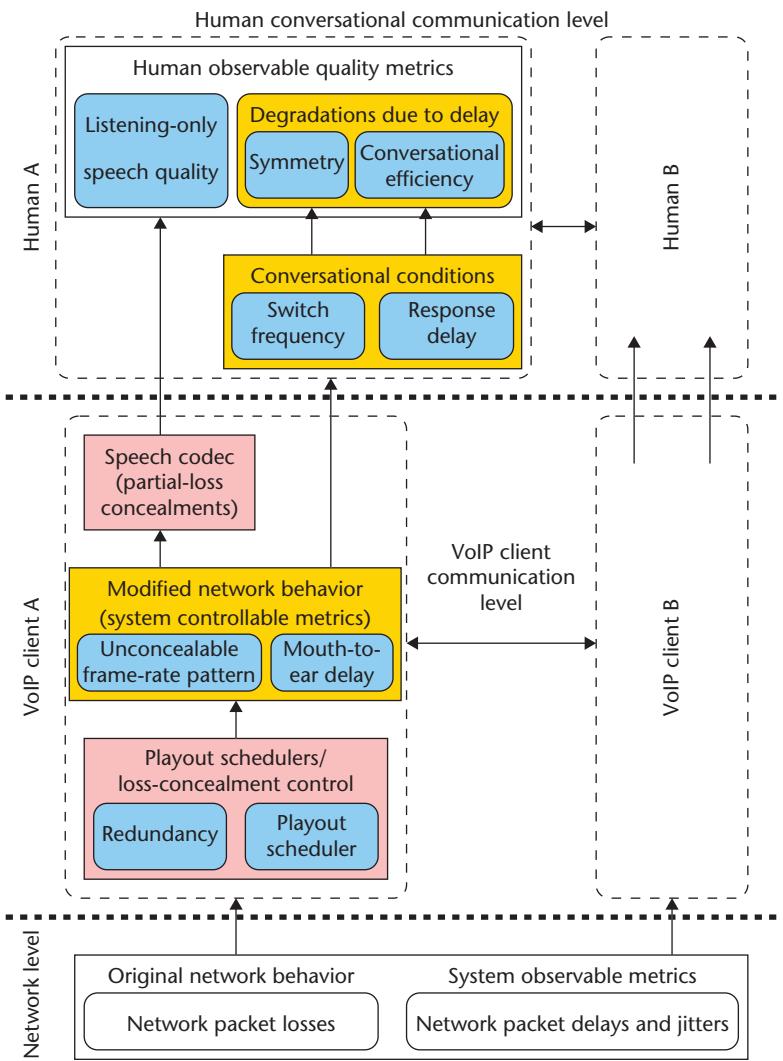
Table 2 shows the statistics for three face-to-face conversations of different average ST durations. Note that CS depends on the value of HRD; that is, if HRD is shorter, then users perceive more of a loss of symmetry due to MED. Likewise, when ST is longer, users perceive less of a loss of efficiency. Hence, MS, CS, and CE are user-perceived metrics that can be calculated objectively, whereas MED is a system-controlled metric that intimately affects those user-perceived metrics.

### Double-talk

In case of a large spike in network delays, if the system doesn't detect the spike and adapt its MED in time, a considerable number of consecutive frames can be lost for a duration that

**Table 2.** Statistics of three face-to-face conversations.

Conversation type	Average single-talk duration (ms)	Average HRD duration (ms)	No. of switches	Total time (sec.)
Fast	1,706	552	7	17.5
Medium	3,055	710	7	29.4
Slow	5,502	827	7	49.8



**Figure 3. Architecture showing interactions among the VoIP clients, the network, and the communicating humans.**

the listener can perceive. Depending on the spike's size and frequency, an utterance, a word, or even a sentence can be inaudible or unintelligible due to the unavailability of frames for playback. If this scenario occurs during a speech utterance, the listener can either assume that the speaker has stopped and start uttering his or her response, or ask the speaker to repeat the last words.

In either case, the speaker, unaware of the listener's difficulty, would most likely continue speaking and cause a collision of speech or unintentional interruptions (double-talk). Depending on the situation, the person observing the collision struggles to resolve the problem by waiting longer for the other to respond or repeating the previously spoken utterances. Further, the two parties might have a different perception of the point where the interruption happens, disrupting the rhythm of a natural

interactive conversation and causing both confusion and degradation in perceived quality. Two researchers identified these double-talk degradations in the 1970s, with both conducting subjective experiments and concluding that double-talk and confusion increases with increased channel delays.<sup>4,5</sup>

#### Adaptation of human behavior

In case of extreme difficulties in communication, such as extreme delays in getting a response or extremely low listening quality, users can either hang up and redial or change their talking style. The style change usually involves talking slowly, talking in longer batches, or deserting the wait-for-acknowledgment gestures. Users who are forced to take these behavioral-adaptation measures feel that their additional effort significantly decreases their satisfaction with the call. Further, this behavioral change might not be acceptable in some languages, cultures, and business-related or mission-critical communications. We are, however, not proposing objective measures to capture the effects of delay on double-talk and the adaptation of human behavior, as these effects are too subjective and heavily depend on the users and the conversations.

#### VoIP client architecture

This section describes the architecture of a general VoIP client and its interactions with the network and human users (see Figure 3). Its main components include the POS that controls the MED, and the loss-concealment scheme and the speech codec that affect the quality of the speech signals received.

#### Play-out scheduling

To buffer irregular packet arrivals (jitters) and to achieve smooth playback of speech frames, VoIP systems commonly employ a POS at the receiver. The POS maintains a consistent MED by controlling the time waited by each packet received so the utterance is played back in the same pace in which it was spoken. Depending on the network delay and jitter conditions, some packets might arrive later than their scheduled times; this information is unavailable for the decoder in generating the speech waveform. In response to changes in network conditions, it's possible to delay the play-out schedule and adjust the MED to

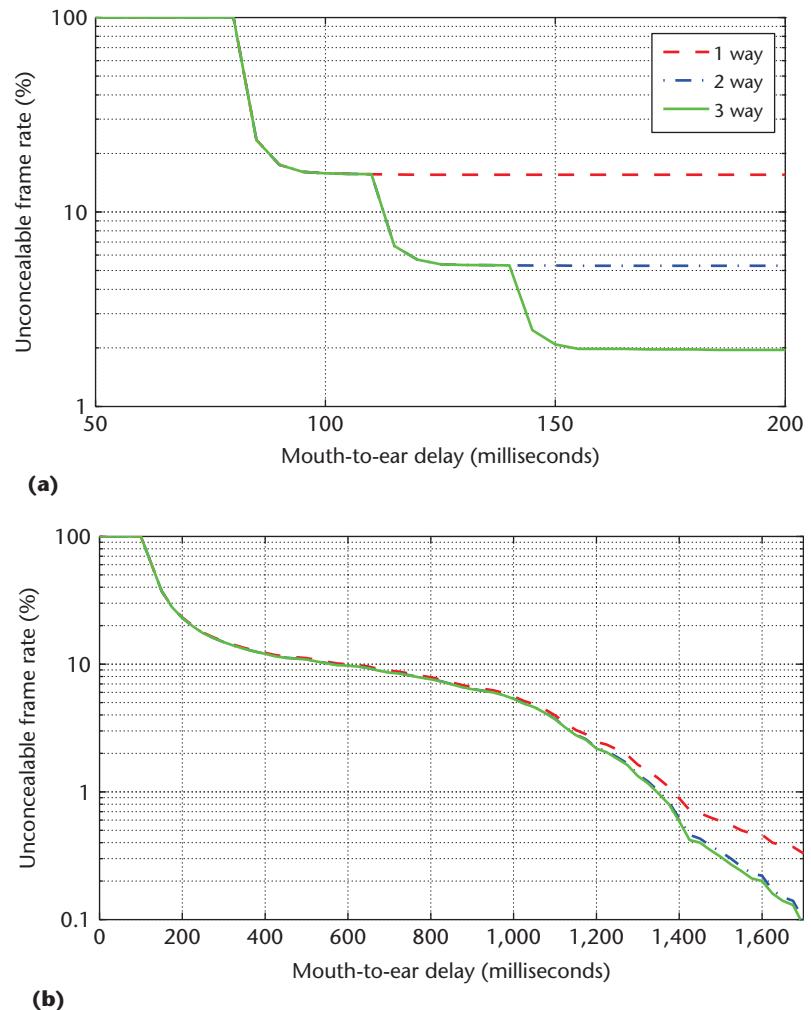
allow more packets to have a chance to arrive in time.

Several studies have addressed the design of adaptive play-out scheduling algorithms, most of which use previously observed network conditions to decide on the MED at the beginning of a speech segment. Some algorithm designs adjust play-out scheduling within a speech segment by using time-scale modification. Play-out scheduling algorithms can be as simple as an open-loop heuristic or as complex as optimizing an end-to-end objective metric that estimates the conversational quality (for example, E-Model as defined by ITU Recommendation G.107). It's difficult to deduce the play-out scheduling algorithm used in the four VoIP clients studied in this article because we can't decode the content of the packets received in each client.

### Loss concealment

VoIP clients commonly employ a redundancy-based loss-concealment scheme for limiting unconcealable losses. These usually require the coordination of the sender and the receiver clients to control losses at the receiver in a closed-loop fashion. They might require sending auxiliary information in every speech packet. The redundancy can be as simple as duplicating previously transmitted packets or as complex as forward-error correction, multi-description, or layered coding. One simple scheme is to piggyback a redundant copy of one or more of the past packets transmitted when sending the current packet. This technique is feasible in VoIP because most speech frames are small, and the maximum transmission unit on the Internet is large enough to allow transmission of multiple frames in the same packet.

Piggyback-based loss concealments will incur additional delays because the POS must wait for those redundant copies to arrive before declaring that the original packet is lost. Figure 4 depicts the trade-offs among MED, redundancy degrees, and the *unconcealable frame rate* under two network conditions and a fixed play-out delay. It shows that the MED and the redundancy needed for achieving a given *unconcealable frame rate* is connection-dependent. In particular, piggyback-based loss concealment is not always effective in high-jitter connections (see Figure 4b).



Some VoIP systems we tested exhibit increases in their packet payload in response to network losses, but we cannot deduce the specific loss-concealment scheme used or evaluate their effectiveness because these systems use proprietary codecs and variable bit rates, which change their payload according to speech input. Moreover, as mentioned previously, the packet contents are unavailable because they are encrypted.

### Speech codecs

Speech codecs reduce the bit rate needed when transmitting the speech waveform. They range from simple-waveform codecs, which mimic the waveform shape, to complex codecs that model human speech production. They generally aim to maintain high speech quality while reducing the bit rate by five to 20 times with respect to the original pulse-code-modulation representation.

**Figure 4. Trade-offs between mouth-to-ear delay and unconcealable frame rate for two PlanetLab connections in Table 1, where the additional play-out delay is equal to the number of periods corresponding to the redundancy degree: (a) trace set 8 from Canada to Portugal and (b) trace set 4 from Taiwan to US.**

Speech codecs were initially designed for wireless and transoceanic speech transmissions with scarce network resources. Their role in VoIP systems, however, is different because network use on the current Internet is less of a concern. The most important feature in current VoIP systems is the ability to withstand lost or late packets. It is difficult to test the codecs used in existing VoIP systems because they are either proprietary or unknown.

### Other approaches

MED is an important element that affects conversational speech quality. It consists of the delays incurred in speech encoding, packing speech frames into packets at the sender, network transmission, play-out buffers at the receiver, and decoding. Among these, delays due to encoding, decoding, and packing are fixed and don't significantly contribute to MED. The component controlled by the VoIP system is the jitter-buffer delay and the delay in waiting for redundant information to arrive.

As mentioned previously, subjective tests led to the conclusions that MED affects user perception of conversational quality, and that longer one-way delays increase the dissatisfaction rate of users. However, because only a few constant delays were used in these tests, the conclusions cannot be directly applied to evaluating VoIP systems that might have long and varying MEDs.

ITU Recommendation G.114 states that a *one-way delay* of less than 150 ms is desirable in a speech communication system and that a delay of more than 400 ms is unacceptable.<sup>6</sup> However, G.114 doesn't specify a metric for measuring the effect that can be combined with listening-only speech quality. Hence, it can't be used directly for evaluating VoIP systems.

### Objective metrics

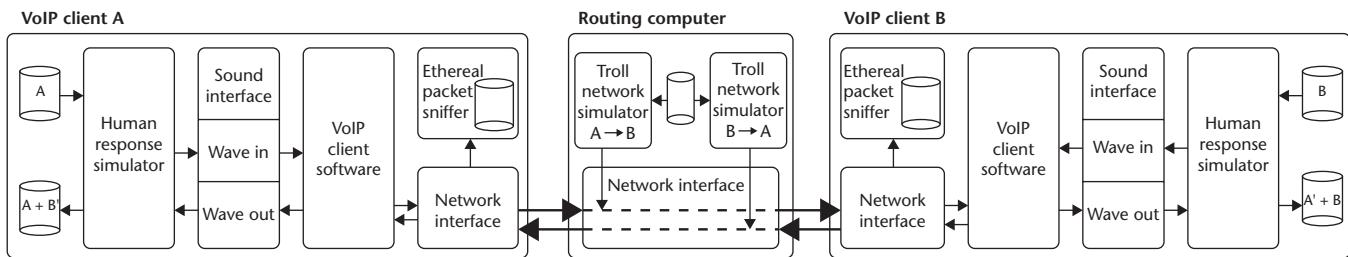
The ITU has several recommendations on objective methods for evaluating conversational quality with an absolute category rating. The perceptual evaluation of speech quality (ITU P.862) is an objective measure to determine a speech segment's quality by comparing the original and the degraded waveforms. Because it only assesses the one-way quality and not delay effect, it must be used in conjunction with other metrics when evaluating conversational quality.

The *E-Model* (ITU G.107) estimates the conversational quality of a speech communication system by considering the effects of the speech encoder, packet loss, one-way delay, and echo. The model was designed for network-planning purposes and assumes the independence of one-way speech quality and delay degradations. This assumption leads to a conclusion that the same MED affects the conversational quality in the same way for conversations, from slow to fast turn-switching, as well as for conversations from low to high one-way quality.

The E-Model oversimplifies the situation because there will be less emphasis on the perception of delay when the one-way quality is low, but there will be more emphasis on the asymmetry and inefficiency of the conversation when the speech quality is high. Likewise, the effect of MED is more pronounced in a conversation with a high turn-taking frequency. Other shortcomings of the E-Model include its crude estimation of the codec and loss effects and a lack of characterization of the variations in speech quality and delay. In short, the E-Model is inadequate for capturing the trade-offs among the factors that affect conversational quality.

The *Call Clarity Index* (ITU P.561 and P.562) estimates the customer's opinion of a speech-communication system in a way similar to the E-Model. Although it provides models for PSTN systems, it doesn't have a user-opinion model for packet-switched networks suffering from long delays and that rely on nonlinear and time-variant signal-processing devices, such as echo control and speech compression. As a result, it's unsuitable for evaluating the conversational quality of VoIP systems.

Due to the proprietary nature of commercial VoIP systems, their evaluation using the E-Model or the Call Clarity Index is not even possible. Some of the numbers required in these metrics are unavailable because either the codecs are proprietary or the number of late or lost packets at the decoder is unavailable. Hence, the evaluation of current systems must be done by treating them as black boxes whose input and output waveforms represent the only information available. Objective metrics based on this information, however, are limited.



### Subjective metrics

A user's perception of speech quality mainly depends on the intelligibility of the speech heard, largely because the user lacks a reference point for the original speech signals. Intelligibility, on the other hand, depends on many factors beyond signal degradations incurred during transmission. The topic of the conversation, the commonality of the words used, and the familiarity of the speakers all can have an impact on intelligibility. To mitigate these subjective effects in the evaluation of speech quality, it's possible to rely on a panel of listeners who don't participate in the conversation but listen to prerecorded speech segments in mean-opinion-score (MOS) tests (from ITU P.800).<sup>7</sup>

In a 1991 NTT study, researchers conducted conversational experiments by having two parties perform tasks using an adjustable-delay speech system.<sup>8</sup> The tasks helped study several scenarios, including reading random numbers, verifying city names, and free conversation with varying average single-talk duration. Subjective quality results revealed that the degradation in MOS is more pronounced when a task requires shorter single-talk durations. However, the study doesn't consider the impact of losses and variations in delay. It's therefore not directly applicable to the evaluation of VoIP systems. Other researchers proposed a utility function to represent the effects of MED in cases in which a conversation is perceived to be half-duplex with much lower quality after some MED threshold.<sup>9</sup> The study focused on selecting a proper forward-error correction based on the MED, rather than on the effects of MED on conversational quality.

ITU P.800 describes a method for evaluating the  $MOS_{CQs}$  (conversational quality subjective) score of a conversation. The method requires asking two subjects to converse over a communication system to complete a specific task, such as arranging a meeting or describing a picture. The subjects are asked to rate the quality

using an *absolute category rating*. The opinions of several users are then averaged to obtain the conversational quality of the system being tested.

There are several shortcomings to this approach when applied to the evaluation of VoIP systems. First, the subjects must both complete a task and evaluate the conversation quality, but the cognitive attention required to evaluate the quality can hinder the completion of the task, and vice versa. Second, the task's type and complexity affect the quality perception of the conversation. Tasks requiring faster turn-taking can be more adversely affected by transmission delays than other tasks. Third, there is no reference in subjective evaluations, and the absolute category rating highly depends on the listeners' expertise.

Moreover, the results of current subjective tests aren't useful as a measure for relative comparisons. If system A's absolute rating is better than B's rating, it doesn't lead to the conclusion that if the subjects were asked to compare the two systems, they would have found A to be of better quality. Likewise, the difference in absolute ratings doesn't translate into the relative difference in quality of the two systems. In addition, the results are difficult to repeat, even for the same subjects and the same task.

### Evaluating conversational quality

Figure 5 depicts our testbed for emulating two-way interactive speech communication, where there are two computers running VoIP software and a router for emulating the network condition. The testbed aims to address the repeatability of standard subjective tests described in ITU P.800. To address the repeatability of network conditions, we modified the router's Linux kernel to intercept User Datagram Protocol packets carrying encoded speech packets between two VoIP clients. In conjunction with modifying the kernel, we use a troll program to drop or delay intercepted

**Figure 5. Our testbed for emulating two-way interactive speech communication.**

packets in each direction according to the traces collected on the PlanetLab.

To address the repeatability of conversational conditions, we developed human-response-simulator (HRS) software that runs on the two client computers. HRS is capable of simulating any conversation with prerecorded speech segments, each using the segments that belong to one side of the conversation. We configured one of the HRSs to start the conversation; then both HRSs take turns speaking their respective segments. Each HRS listens to the speech played by the VoIP client and responds after waiting appropriately. The HRS interfaces with the VoIP client software via a Virtual Audio Cable (see <http://nrcde.ru/music/software/eng/vac.html>) that allows digital waveform transfer to and from the VoIP client without quality loss. Each HRS records the waveforms spoken and heard by the respective client to which it's interfaced.

We think of this setup as a pair of speech-response systems that converse with each other by following a script in such a way that the conversation can be repeated almost exactly for the same VoIP system under the same condition. The testbed ensures that when two VoIP systems are tested under the same conditions, the variations in quality are only due to system differences. Hence, it enables us to compare the relative quality of the two systems.

After collecting the recordings, we asked human subjects to evaluate two conversations recorded under the same condition using the Comparative Category Rating scale in ITU P.800. In addition, we extracted objective data, such as perceptual evaluation of speech quality (PESQ), MED, CS, and CE, from each recording. Finally, we presented the recordings randomly to the human subjects, who did not know the system or the network conditions used in each recording.

We applied our evaluation method to four VoIP systems: Skype, Google Talk, Windows Live Messenger, and Yahoo Messenger. We conducted the tests using six simulated traces from Table 1 with distinct network conditions, and a new trace on an ideal network with no loss and no delay. Along with the three conversations in Table 2, we tested 21 distinct combinations for each system.

For our subjective tests, we compared each pair of systems (six comparisons for four

systems) under each combination of network and conversational conditions. Doing so resulted in 126 comparisons. With six human subjects, we carried out a total of 756 subjective tests. We observed that the opinions from different users relating to the same pair of systems and test conditions fell within three Comparative Category Rating values of each other in 83 percent of the cases. These numbers indicate we can have confidence in the test results.

#### Objective metrics

We observed that the performance of the various VoIP systems is comparable under ideal network conditions. However, performance started to deviate as we introduced more delay, jitter, and loss. These differences indicate that different trade-offs are made by each system to overcome network impairments. We further observed that Windows Live Messenger is superior in terms of listening-only speech quality (measured by PESQ), that Skype has consistently larger MEDs, and that Google Talk generally has shorter MEDs and better CS and CE, as indicated in Table 3.

The PESQ quality of the systems under the ideal network condition is a strong indication of the intrinsic performance of their respective speech codecs without loss concealment. When the network suffers from delays and jitters, the PESQ observed reflects the combined performance of the play-out scheduling scheme for concealing late packets and the robustness of the speech codec to unconcealed late packets. We observed that the systems don't have widely different MEDs for the three conversations of different turn-taking frequencies and the same network conditions. This result indicates that the play-out scheduling schemes of these systems don't optimize their MEDs in response to different turn-taking conditions.

#### Subjective metrics

We carried out extensive comparative subjective evaluations of the six system pairs under the same network and conversational conditions. (The complete results are not shown here.) Figure 6a (on page 56) illustrates the distribution of the opinions for each of the system pairs. Figures 6b and 6c further present the opinions under different network conditions. To get a reasonable number of samples in each distribution, we combined the results

**Table 3.** Objective evaluations of four VoIP systems. The best quality results for each of the four systems are shown in bold.

Trace class (delay, jitter, loss)	VoIP system	Conversation 1 w/fast turn-taking				Conversation 2 w/medium turn-taking				Conversation 3 w/slow turn taking			
		PESQ	MED	CS	CE	PESQ	MED	CS	CE	PESQ	MED	CS	CE
No, no, no	Skype	3.192	286	2.04	67	3.244	338	1.95	74	3.418	290	1.70	83
	GTalk	3.557	<b>130</b>	<b>1.47</b>	<b>71</b>	3.506	147	1.42	<b>78</b>	3.536	160	1.39	<b>85</b>
	Yahoo	3.553	140	1.51	<b>71</b>	3.676	<b>139</b>	<b>1.39</b>	<b>78</b>	3.785	151	1.37	<b>85</b>
	WinLive	<b>3.562</b>	171	1.62	70	<b>3.856</b>	154	1.43	<b>78</b>	<b>3.928</b>	<b>133</b>	<b>1.32</b>	<b>85</b>
L, L, L	Skype	3.328	319	2.15	66	3.119	541	2.52	71	3.254	392	1.95	82
	GTalk	3.371	<b>203</b>	<b>1.74</b>	<b>69</b>	<b>3.525</b>	368	2.04	74	3.092	<b>201</b>	<b>1.49</b>	<b>84</b>
	Yahoo	3.534	205	<b>1.74</b>	<b>69</b>	3.492	<b>203</b>	<b>1.57</b>	<b>77</b>	3.354	298	1.72	83
	WinLive	<b>3.675</b>	222	1.81	<b>69</b>	3.492	218	1.61	<b>77</b>	<b>3.746</b>	393	1.95	82
L, L, H	Skype	2.339	442	2.60	63	2.461	416	2.17	73	2.565	424	2.02	81
	GTalk	2.484	230	1.83	<b>69</b>	2.501	<b>265</b>	<b>1.75</b>	<b>76</b>	2.305	275	1.67	83
	Yahoo	2.502	<b>217</b>	<b>1.79</b>	<b>69</b>	2.755	276	1.78	<b>76</b>	2.485	<b>239</b>	<b>1.58</b>	<b>84</b>
	WinLive	<b>3.306</b>	336	2.22	66	<b>3.309</b>	340	1.96	74	<b>3.257</b>	321	1.78	83
L, H, L	Skype	2.693	408	2.48	64	2.882	487	2.37	72	<b>3.083</b>	420	2.02	82
	GTalk	3.145	<b>216</b>	<b>1.78</b>	<b>69</b>	3.145	<b>227</b>	<b>1.64</b>	<b>77</b>	2.854	<b>261</b>	<b>1.63</b>	<b>83</b>
	Yahoo	3.085	274	1.99	67	3.097	240	1.68	76	2.987	274	1.66	83
	WinLive	<b>3.454</b>	404	2.47	64	<b>3.512</b>	432	2.22	73	2.953	420	2.02	82
H, L, L	Skype	3.096	550	2.99	61	3.325	462	2.30	72	3.444	420	2.02	82
	GTalk	3.466	<b>281</b>	<b>2.02</b>	<b>67</b>	3.517	<b>279</b>	<b>1.79</b>	<b>76</b>	3.435	<b>287</b>	<b>1.69</b>	<b>83</b>
	Yahoo	3.531	283	2.03	<b>67</b>	3.464	305	1.86	75	<b>3.687</b>	301	1.73	<b>83</b>
	WinLive	<b>3.792</b>	313	2.13	66	<b>3.803</b>	315	1.89	75	3.647	309	1.75	<b>83</b>
H, L, H	Skype	2.619	535	2.94	61	2.564	504	2.42	72	2.564	503	2.22	81
	GTalk	2.639	<b>273</b>	<b>1.99</b>	<b>67</b>	2.666	<b>283</b>	<b>1.80</b>	<b>75</b>	2.469	<b>300</b>	<b>1.73</b>	<b>83</b>
	Yahoo	2.749	281	2.02	<b>67</b>	2.472	365	2.03	74	2.617	314	1.76	<b>83</b>
	WinLive	<b>3.060</b>	440	2.60	63	<b>3.251</b>	421	2.19	73	<b>3.286</b>	363	1.88	82
H, H, L	Skype	2.985	612	3.22	59	2.983	574	2.62	70	2.652	648	2.57	79
	GTalk	3.296	<b>399</b>	<b>2.45</b>	<b>64</b>	<b>3.151</b>	<b>410</b>	<b>2.15</b>	<b>73</b>	2.729	<b>397</b>	<b>1.96</b>	<b>82</b>
	Yahoo	3.022	544	2.97	61	3.068	487	2.37	72	2.841	573	2.39	80
	WinLive	<b>3.327</b>	595	3.15	60	2.937	589	2.66	70	<b>2.930</b>	748	2.81	78

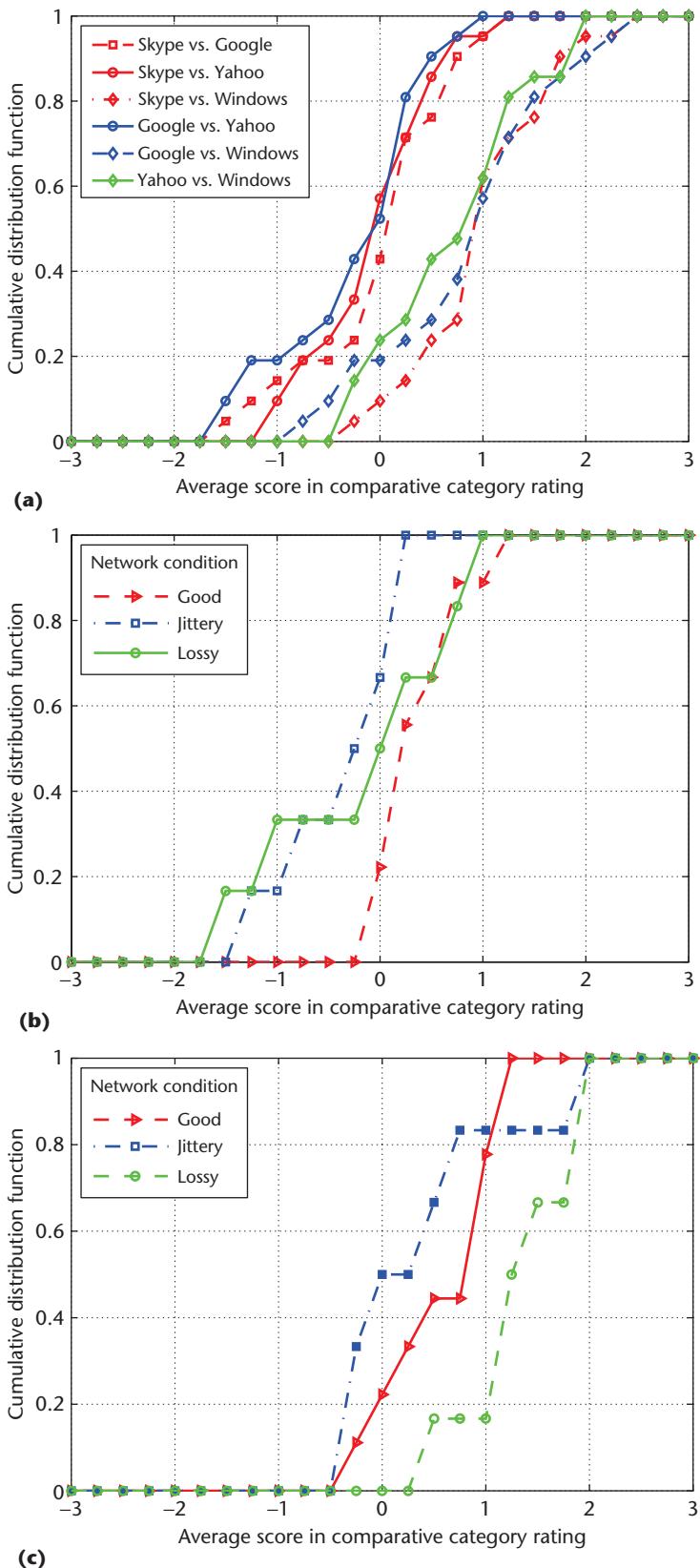
for some of the network conditions that are relatively similar in terms of their effects on performance: (N, N, N), (L, L, L), (H, L, L) for good conditions; (L, H, L), (H, H, L) for jittery conditions; and (L, L, H), (H, L, H) for lossy conditions.

Figure 6b shows that Skype performed similarly to Google Talk under lossy conditions; however, Skype performed more effectively under jittery conditions, whereas Google Talk was preferred under good network conditions. Figure 6c shows that Windows Live Messenger was preferred over Yahoo Messenger under most conditions, and that it was strongly preferred under lossy conditions. These results suggest that different systems employ different trade-offs in addressing losses and jitters.

### Comparative results

A classifier can help predict comparative subjective evaluation results using objective measures that can be easily collected by our testbed. A classifier has an advantage over time-consuming subjective tests, which are expensive to conduct and don't scale well with the number of systems compared. But we can't fully predict subjective results by objective means because the subjective results aren't totally consistent themselves. In this study, we employed a support vector machine (SVM) due to its speed and accuracy.<sup>10</sup> We use 22 inputs (features) that can be objectively obtained from the conversation recordings collected by our testbed.

For each of the two systems compared, we input their CS, CE, average PESQ, and MED of



**Figure 6. Distribution of subjective comparison scores of four VoIP systems:** (a) under all network conditions; and (b) Skype versus Google Talk and (c) Yahoo versus Windows Live Messenger under good, jittery, and lossy network conditions.

the conversation, and their variances across the speech segments and turns. To characterize a system's relative performance, we also input their ratio and difference of the average PESQs and MEDs. To characterize the conversational condition, we input the average single-talk duration and average HRD, as well as the turns-switching frequency. To characterize the network condition, we use the average network delay, the percentage of packets that exhibit jitters of more than 60 ms, and the average loss rate.

Once we obtained the distribution of the subjects' opinions, we applied hypothesis testing to determine the dominant opinion with statistical significance. Given  $K$  subjects have evaluated each comparison pair, we obtained the distribution denoted by the triplet  $(p_{-1}, p_0, p_1)$  that corresponds, respectively to the three opinions: A is better than B, A is about the same as B, and A is worse than B. In our hypothesis tests to determine if one opinion is dominant with statistical significance, we compared the number of responses choosing opinion  $i \in \{-1, 0, 1\}$  against a binomial distribution with nondominant opinion.

Specifically, we defined the null hypothesis ( $H_0$ ) as  $p_i$  drawn from binomial  $(K, p \leq 0.5)$ . If the null hypothesis can be rejected with 90 percent statistical significance, then opinion  $i$  is *dominant*. By construction, no two opinions can be dominant at the same time. However, it's possible that no opinion is dominant with 90 percent statistical significance. In that case, the comparison between A and B is *inconclusive*. We used the dominance information as the target value for the classifier.

Once we obtained the input features and the target labels, we used the radial-basis function as the kernel function to project the 22 dimensions to higher dimensions, where we search for a set of hyperplanes to separate the classes. We used a dynamic search tool in LIBSVM<sup>10</sup> to find the optimal kernel parameters. To ensure that we could generalize the results, we turned to cross-validation techniques commonly used in statistics. In  $n$ -fold cross validations, the sample set is randomly divided into  $n$  partitions, roughly of equal size. The  $n - 1$  partitions are then used to train the classifier, and the remaining partition is used to test its performance. The classifier is trained and tested  $n$  times, each time using a different partition as the testing set. The average classification rate is

**Table 4. Comparative subjective evaluations of pairs of VoIP systems and the prediction results of our support vector machine.**  
**In comparing systems A to B, we show the dominant opinion with 90 percent statistical significance: we indicate if A is better than, about the same as, worse than, or inconclusive with respect to B. Red symbols represent results that differ from the subjective results.**

System		Turn frequency	Subjective results						Prediction results: training data						Prediction results: unseen data								
pairs	(A vs. B)		NNN	LLL	LLH	LHL	HLL	HLH	NNN	LLL	LLH	LHL	HLL	HLH	HHL	NNN	LLL	LLH	LHL	HLL	HLH	HHL	
Skype vs.	Fast	?	?	<	?	?	>	>	?	?	<	?	?	>	>	?	?	>	?	?	>		
GTalk	Medium	?	<	>	>	?	?	<	?	<	>	>	?	?	<	?	<	>	>	?	>		
	Slow	≈	?	?	?	<	<	≈	≈	?	?	?	<	<	≈	≈	?	?	?	<	?	?	
Skype vs.	Fast	?	?	<	?	?	?	<	?	?	<	?	?	<	?	?	>	?	?	>	<		
Yahoo	Medium	?	?	<	?	?	≈	>	?	?	<	?	?	≈	>	?	?	<	?	?	≈	<	
	Slow	?	≈	>	≈	<	<	>	?	≈	>	≈	<	<	>	<	?	>	?	<	<	?	
Skype vs.	Fast	<	<	<	≈	?	<	<	<	?	<	?	<	<	?	<	<	<	<	<	<	<	
WinLive	Medium	<	<	<	<	<	<	<	<	<	<	<	<	<	<	<	<	<	<	<	<	<	
	Slow	<	<	<	<	<	<	<	?	<	<	<	<	<	?	<	<	<	<	<	<	?	
GTalk vs.	Fast	<	≈	<	>	?	<	≈	<	≈	<	>	?	<	≈	<	≈	<	>	?	<	≈	
Yahoo	Medium	≈	≈	?	<	≈	>	>	≈	≈	?	<	≈	>	>	<	?	?	?	?	≈	?	≈
	Slow	≈	≈	?	≈	≈	?	>	≈	≈	?	≈	?	<	?	≈	≈	?	≈	≈	?	?	
GTalk vs.	Fast	<	≈	?	<	?	<	<	<	≈	?	<	?	<	?	<	≈	<	?	<	?	<	
WinLive	Medium	<	<	<	<	<	<	<	?	<	<	<	<	<	?	<	<	<	<	<	<	?	
	Slow	<	<	<	?	<	<	?	<	<	<	?	<	<	?	<	<	<	?	<	<	?	
Yahoo vs.	Fast	?	<	<	<	?	<	<	<	?	<	?	<	<	?	<	?	<	?	?	?	<	
WinLive	Medium	<	<	<	<	<	<	<	?	<	<	<	<	<	?	<	<	<	<	<	<	?	
	Slow	<	<	<	?	<	<	<	?	<	<	<	<	<	?	<	<	<	<	<	<	?	

referred to as the *cross-validation score*. Because the classifier doesn't learn from the samples in the testing set, a high cross-validation score is interpreted as the ability of the classifier to generalize to samples with conditions not in the training set.

With our SVM model, we could successfully predict 97.6 percent of the samples in our training set and 64.3 percent when using 10-fold cross validations. To further validate our results, we used new conversations and packet traces and applied our classifier to predict the subjective results.

Table 4 shows the dominant comparative opinions for the subjective experiments, SVM predictions of the training set, and SVM predictions of the unseen data. In our tests, we observed that all systems operated well under good network conditions, as the difference in performance among the systems is too close to be perceived. However, as we introduced network imperfections, there were clear user preferences in terms of conversational quality. Windows Live Messenger was strongly preferred over other systems under lossy conditions. In contrast, Skype was slightly preferred over Google Talk, and Windows Live Messenger

was slightly preferred over Yahoo Messenger under jittery conditions.

Furthermore, the distributions of predicted comparative opinions between system pairs matched closely to those obtained through subjective tests, even for unseen data. The results indicate that our SVM classifier can be used to comparatively evaluate the conversational quality of VoIP systems under a variety of network and conversational conditions.

## Future work

Following the research presented in this article, we developed a methodology to conduct off-line subjective tests to statistically identify the optimal MED value under a given set of network and conversational conditions. We are currently developing an overall methodology to conduct subjective tests to identify optimal MED values under any conditions observable at run time to help design POS control for VoIP systems. The classifier developed in this article can be applied in evaluating the mentioned or any newly designed VoIP system against other VoIP systems, including ones compared in this article, without the need for further subjective tests.

MM

## Acknowledgement

Research in this article was supported by National Science Foundation Grant CNS 08-41336.

## References

1. B. Sat and B.W. Wah, "Analysis and Evaluation of the Skype and Google Talk VoIP Systems," *Proc. IEEE Int'l Conf. Multimedia and Expo*, IEEE CS Press, 2006.
2. B. Sat and B.W. Wah, "Evaluation of Conversational Voice Quality of the Skype, Google Talk, Windows Live, and Yahoo Messenger VoIP Systems," *IEEE Int'l Workshop Multimedia Signal Processing*, IEEE Press, 2007.
3. J.-C. Bolot, S. Fosse-Parisis, and D. Towsley, "Adaptive FEC-Based Error Control for Internet Telephony," *Proc. IEEE Infocom*, vol. 3, IEEE CS Press, 1999, pp. 1453-1460.
4. P.T. Brady, "Effects of Transmission Delay on Conversational Behavior on Echo-Free Telephone Circuits," *Bell System Technical J.*, vol. 50, no. 1, 1971, pp. 115-134.
5. D.L. Richards, *Telecommunication by Speech*, Butterworths, 1973.
6. *ITU-T G-Series Recommendations, Transmission Systems and Media, Digital Systems and Networks*, Int'l Telecommunication Union; <http://www.itu.int/rec/T-REC-G/en>.
7. *ITU-T P-Series Recommendations, Telephone Transmission Quality, Telephone Installations, Local Line Networks*, Int'l Telecommunications Union; <http://www.itu.int/rec/T-REC-P/en>.
8. N. Kiatawaki and K. Itoh, "Pure Delay Effect on Speech Quality in Telecommunications," *IEEE J. Selected Areas of Communication*, vol. 9, no. 4, 1991, pp. 586-593.
9. C. Boutremans and J.-Y. Le Boudec, "Adaptive Joint Play-Out Buffer and FEC Adjustment for Internet Telephony," *Proc. IEEE Infocom*, vol. 1, IEEE CS Press, 2003, pp. 652-662.
10. C.-C. Chang and C.-J. Lin, "LIBSVM: A Library for Support Vector Machines," 2009; <http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf>.

**Batu Sat** is a PhD candidate at the Electrical and Computer Engineering Department, University of Illinois at Urbana-Champaign. His research interests include the development of evaluation and design methods for real-time multimedia communication systems. Sat has an MS in electrical engineering from UIUC. He's a student member of the IEEE and ACM. Contact him at [batusat@illinois.edu](mailto:batusat@illinois.edu).

**Benjamin W. Wah** is the Franklin W. Woeltge Endowed Professor of Electrical and Computer Engineering and Professor of the Coordinated Science Laboratory of the University of Illinois at Urbana-Champaign. His research interests include nonlinear search and optimization, multimedia signal processing, and computer networks. Wah has a PhD in computer science from the University of California, Berkeley. He is a Fellow of the American Association for the Advancement of Science, ACM, and IEEE. Contact him at [wah@illinois.edu](mailto:wah@illinois.edu).

**IEEE Annals**  
of the History of Computing

From the analytical engine to the supercomputer, from Pascal to von Neumann, from punched cards to CD-ROMs—the *IEEE Annals of the History of Computing* covers the breadth of computer history. The quarterly publication is an active center for the collection and dissemination of information on historical projects and organizations, oral history activities, and international conferences.

[www.computer.org/annals](http://www.computer.org/annals)

## Call for Papers

### Editorial Board

**Sethuraman Panchanathan**

*Editor in Chief*

**Susanne Boll**

**Daniel Ellis**

**Farshad Fotouhi**

**Forouzan Golshani**

**William Grosky**

**Yu Hen Hu**

**Jane Hunter**

**Frank Nack**

**Yong Rui**

**Dorée Duncan Seligmann**

**John R. Smith**

**Qibin Sun**

**Utz Westermann**

**Heather Yu**

**Wenjun Zeng**

*IEEE MultiMedia* serves a broad readership, including researchers, technology developers, practitioners, end users, and designers of multimedia systems and applications.

Topics covered by this publication include design, development, and applications of multimedia systems, as well as their deployment and use. Original contributions on novel applications of multiple media, tutorials, and case studies are of particular interest. Timely topics collected into special theme issues are encouraged.

*IEEE MultiMedia* magazine seeks original articles discussing research as well as advanced practices in hardware and software, spanning the range from theory to working systems. Example topics include

- multimedia aspects of
  - sensory information processing,
  - mobility and transport,
  - services computing,
  - content-based retrieval and media mining,
  - multimodal interfaces and human-computer interaction,
  - real-time computing,
  - ontologies and semantics, and
  - standards;
- multimedia authoring and creation;
- pervasive media services;
- content protection and security;
- content adaptation;
- media personalization;
- interactive and experiential computing;
- immersive and virtual environments; and
- multimodal biometrics

as related to such domains as biomedicine, arts and entertainment, education, the environment, disability services, commerce, and enterprise-wide systems.

Articles should be approximately eight magazine pages with roughly five figures or images, where a page is approximately 750 words and an average-sized image counts for 150 words. Please limit the number of references to the 10 to 12 most relevant. Also consider providing background materials in sidebars for nonexpert readers. Visit *IEEE MultiMedia*'s author guidelines and author resources pages (links found at <http://www.computer.org/multimedia>) for more information on our requirements.

Articles must be submitted through the IEEE Computer Society's online manuscript service, Manuscript Central, at <https://mc.manuscriptcentral.com/cs-ieee>. From there, you can log into the Author Center and upload your submission. Once your manuscript is uploaded, you can view it online to check the status at any time. If you have any questions regarding Manuscript Central, contact the magazine assistant at [mm-ma@computer.org](mailto:mm-ma@computer.org).

*IEEE MultiMedia* also welcomes proposals for special issues on timely topics related to the magazine's scope. For further information and to discuss possible projects, please contact

Sethuraman Panchanathan  
Editor in Chief, *IEEE MultiMedia*  
Arizona State University  
Tempe, AZ 85287-5406  
[panch@asu.edu](mailto:panch@asu.edu)