EVALUATION OF CONVERSATIONAL VOICE COMMUNICATION QUALITY OF THE SKYPE, GOOGLE-TALK, WINDOWS LIVE, AND YAHOO MESSENGER VOIP SYSTEMS

Batu Sat and Benjamin W. Wah

Department of Electrical and Computer Engineering
and the Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
{batusat, wah}@uiuc.edu

ABSTRACT

In this paper, we evaluate the *conversational voice communication quality* (CVCQ) of VoIP systems, both from the user and the system perspectives. We first identify the metrics for CVCQ, which include listening-only speech quality (LOSQ), conversational interactivity (CI), and conversational efficiency (CE). These depend on the mouth-to-ear delays (MEDs) between the two clients. Based on packet traces collected in the PlanetLab and on the dynamics of human interactive speech, we study four popular VoIP client systems: Skype (v2.5), Google-Talk (Beta), Windows Live Messenger (v8.0), and Yahoo Messenger with Voice (v8.0), under various network and conversational conditions.

1. INTRODUCTION

Software-based VoIP systems allow interactive voice communications between two or more parties by utilizing best-effort public IP networks. Due to the unreliable and time-varying characteristics of these networks, VoIP systems need to monitor and control their operations in order to provide good perceptual quality to users. Since there is no current standard for measuring CVCQ, we consider a set of user-observable attributes in our evaluations. These include listening-only speech quality (LOSQ), conversational interactivity (CI), and conversational efficiency (CE) [1]. We evaluate four commonly used VoIP software clients under realistic simulations of conversations between two parties. Our simulations are based on packet traces [2] collected in the PlanetLab that are typical of network conditions observed in the Internet.

2. CONVERSATIONAL VOICE QUALITY

We discuss in this section four metrics for measuring the user perceived CVCQ of a VoIP system.

IEEE WORKSHOP ON MULTIMEDIA SIGNAL PROCESSING, 2007.

2.1. Listening-Only Speech Quality

A user's perception of LOSQ mainly depends on the intelligibility of the speech heard because the user lacks a reference to the original speech signals. 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 effect intelligibility. To mitigate these subjective effects in the evaluation of LOSQ, formal mean-opinion-score (MOS) tests (ITU P.800) are usually conducted by a panel of listeners who only listen to pre-recorded speech segments.

Due to the expensive, time consuming and non-repeatable nature of MOS tests, LOSQ is commonly evaluated by PESQ (ITU P.862). Because it has been shown to have high correlations to subjective MOS tests for a variety of land-line, mobile and VoIP applications, it can be used to evaluate VoIP systems in a fast and repeatable way. Using the standard MOS terminology in ITU P.800.1, a conversion from PESQ to the standard listening-quality metric (MOS_{LQO}) can be done using equations in ITU P.862.1.

2.2. Mouth-to-Ear Delay/Conversational Interactivity

MED is an important element of conversational speech quality due to its effects on human perception in an interactive communication. It consists of delays incurred in a) speech encoding, b) packing speech frames into packets at the sender, c) the network, d) the play-out (jitter) buffer at the receiver, and e) decoding. Of these delays, encoding, decoding and packing delays are fixed and negligible. To smooth out non-deterministic network delays, jitter buffers are often employed at the receiver to control packet-level delays and to keep MED constant during a speech segment.

The *G.114 Guidelines* state that a one-way delay of less than 150 ms is desirable in a voice communication system and that more than 400 ms is unacceptable. These infer that

CVCQ is a monotonically non-increasing function of MED. Note that, even though the guidelines present a general understanding of the effect, they do not specify a metric to measure the effect that is comparable to LOSQ MOS_{LQO}.

On the other hand, the *E-model* (ITU G.107) considers the effect of one-way delay in the evaluation of conversational speech quality, but is only designed to assist service providers during the planning process. The transmission rating factor, R, of the E-model, is on a psycho-acoustical scale, where the effects of different degradations are additive and is defined as follows:

$$\begin{aligned} R &= R_o - I_s - I_d - I_e + A \\ I_d &= I_{dte} + I_{dle} + I_{dd} \\ MOS_{CQE} &= 1 + 0.035R + \tfrac{7R(R-60)(100-R)}{10^6} \end{aligned} \quad (1)$$

where R_0 is the basic SNR, and I_s (*resp.*, I_d, I_e, and A) is the simultaneous impairment (*resp.*, delay impairment, equipment impairment, and advantage) factor.

The delay impairment factor is further divided into I_{dte} and I_{dle} that, respectively, estimate the impairment due to the talker and listener echoes, and I_{dd} that estimates the degradation caused by too-long absolute delay even with perfect echo cancellation. By using (1) to convert R into MOS_{CQE}, MOS is found to decrease by one when MED is increased from 0 to 400 ms (Figure 1 in [1]).

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 *combined E-model and PESQ*, a conversational quality metric MOS_c was proposed [3]. Here, PESQ is converted into the scale of R and substituted into the E-model to represent the combined impairments due to I_e. A subsequent study used regression models to predict the contribution of PESQ. The model corrects the speech dependency part of the E-model but does not consider the effects of conversational conditions.

In an NTT study [4], 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. The study revealed that the degradation in MOS is more pronounced when a task requires shorter single-talk durations, but did not consider the effect of losses.

A utility function was proposed in [5] to represent the effects of MED, where after some MED threshold the conversation is perceived to be half-duplex and quality degrades. However, the goal of the study is to incorporate the effect of MED on the choice of FEC, rather than studying the effects of MED on conversational quality.

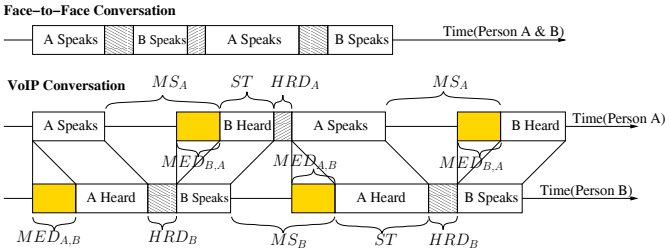

Figure 1: Delays in face-to-face and VoIP communications.

2.3. Human Perception of Mouth-to-Ear Delays

In a face-to-face voice conversation, users have a common reality in the perception of the sequence and the timing of events. However, as is illustrated in Figure 1 [1], voice over a communication channel with delays may lack a common perspective and may lead to multiple realities.

We define *human response delay* from B's perspective (HRD_B) as the time 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 *mutual silence* (MS_A). 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 perceiving the indirect effects of MED, such as MS, CI, and CE, the user can deduce the existence of MED. Note that the silence periods when switching between the two persons in VoIP are no longer symmetric.

The asymmetry between HRD and MS leads to a new CI, which is an important component of CVCQ. 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 propagation delay increases. If the asymmetry in the perceived response times increases, humans tend to have a degraded perception of CI that will result in the degradation of the quality of the conversation. (One possible effect is that, if A perceives that B is responding slowly, then A tends to respond slowly as well!)

Table 1: 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.

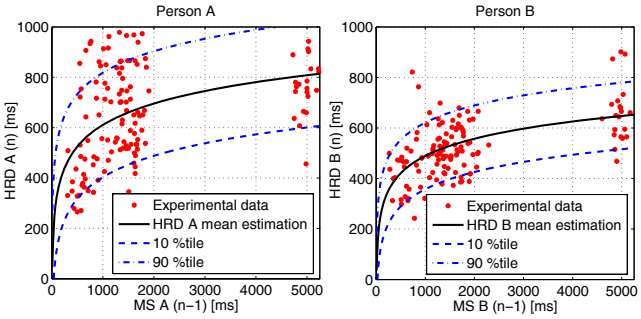

Figure 2: MS experienced and the next HRD.

Another effect of MED on a VoIP conversation is that it takes longer to accomplish a task when there is delay in the communication channel (Figure 1). 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 as the *conversational efficiency*:

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

Table 1 illustrates the statistics of two face-to-face conversations with different average ST durations. Assuming that $HRD_{VoIP} = HRD_{F2F}$ and does not change with MED (an assumption to be relaxed in the next section), CE can be estimated as a function of MED as follows:

$$CE \approx \frac{TotalF2FTime}{TotalF2FTime + (\#switches) * MED}. \quad (5)$$

Eq. (5) implies that changes in CE are almost undetectable for small MEDs and slow switching frequencies. However, CE decreases with increasing MEDs and short average single-talk durations (or large number of switches).

Due to space limitations, we do not present the effects due to double-talk, echo, and temporal changes in MED.

Based on the metrics identified in CVCQ, we evaluate in this paper four commonly used VoIP client software systems using trace-driven simulations.

3. A MODEL OF INTERACTIVE CONVERSATION

Typical human response delays range from 300 ms to 800 ms in most cases, depending on the language, social status, relation of the parties, and the task achieved through the conversation. In a related ITU P.59 standard, artificial conversational speech is modeled to have mutual silence periods, between the end of A's speech and the beginning of B's speech, to have a geometric distribution with 508 ms mean. The standard, however, does not consider the asymmetric perception of MS in the case of MED, nor the process of adaptation of human behavior under different conditions.

In our experiments, we have investigated the relation between the user-observed MS and the user controlled HRD. By controlling delays in a VoIP setting, we measured the steady-state MS and HRD of multiple conversations for a group of humans, where the delay is kept constant within a conversation and changed from one experiment to another. Most of the results lie in the typically observed MS range, and some are measured with exceedingly long 2-sec MED. Figure 2 depicts the experimental data, which is fitted to the following models for both users. Here, HRD increases monotonically with MS by a logarithmic mean, in addition to a random component with a normal distribution.

$$HRD_A^j = -191 + 98 * log(MS_A^{j-1}) + \mathcal{N}(0, 103) \quad (6)$$
$$HRD_B^j = -238 + 123 * log(MS_B^{j-1}) + \mathcal{N}(0, 162).$$

The dynamics of HRD changing with MS can be modeled by a Markov process using (2) and (6), starting with an initial HRD based on a user's expectation. HRD can be proved to converge to a finite value as the conversation proceeds, for any given MED and initial HRD. This means that the process is stable, and CE converges to a non-zero value.

4. SYSTEM EVALUATION

Figure 3 depicts our test-bed that consists of two client computers running the VoIP software and a routing computer for simulating the network using PlanetLab traces.

Each VoIP client processes a speech waveform and sends UDP packets to the other client. We use two human-response-simulator (HRS) software we have developed in the two end-clients where the VoIP software is running. These simulators talk, listen, and respond appropriately, taking turns by using pre-recorded single-talk speech segments from two conversational recordings (Table 1). The recordings used in the simulated responses consist of one with a faster conversational switching frequency and one with a slower frequency that represent, respectively, a business and a social conversation. The HRD used in the simulations is based on (6). Due to space limitations, we only present the results on the business conversation.

Using the acoustical information collected from the two clients of the VoIP connection, Table 2 presents the LOSQ in terms of PESQ, CE, and segment-based CI. We have conducted experiments on four packet traces [2] and an ideal connection with no loss and no delay. We have also shown the statistics of the MEDs for each connection.

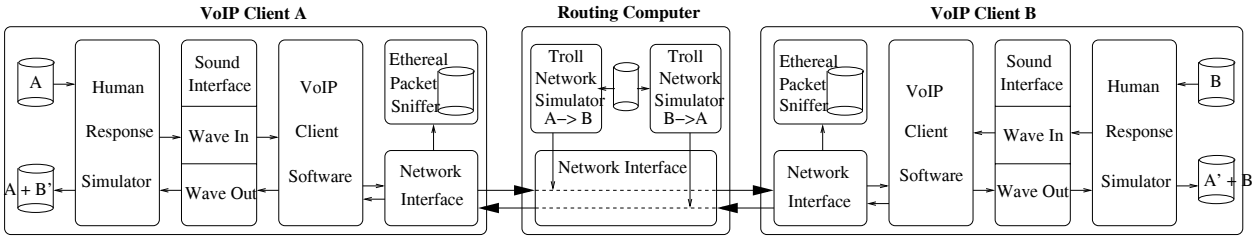

Figure 3: Our test-bed to emulate a two-way interactive voice communication using traces collected in the PlanetLab.

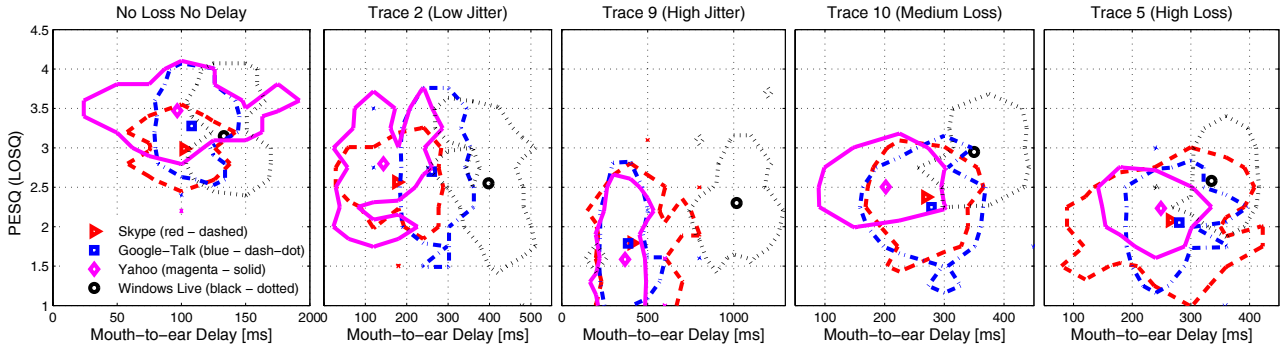

Figure 4: PESQ-MED: Mean values and contours of regions containing 90% of the samples for each system and each trace.

Table 2: CVCQ evaluations of four VoIP systems for three Internet and one ideal connections.

Trace	VoIP System	PESQ			MED [ms]			CI	CE
		10%	mean	90%	10%	mean	90%	mean	
No loss, no delay	Skype	2.675	2.988	3.287	85	102	114	1.29	90.2
	GTalk	2.753	3.278	3.700	94	108	121	1.31	89.4
	Yahoo	3.146	3.472	3.694	56	97	129	1.29	88.2
	WinLive	2.585	3.149	3.798	127	133	146	1.38	88.1
2 (medium jitters)	Skype	2.039	2.566	2.989	82	176	254	1.51	82.1
	GTalk	2.081	2.699	3.332	238	261	288	1.69	82.6
	Yahoo	2.002	2.799	3.478	65	144	217	1.43	82.3
	WinLive	1.757	2.551	3.389	314	397	463	2.08	79.2
5 (high losses)	Skype	1.576	2.081	2.555	154	265	342	1.51	81.4
	GTalk	1.384	2.052	2.581	223	280	356	1.54	80.7
	Yahoo	1.861	2.233	2.534	189	249	286	1.44	82.3
	WinLive	2.201	2.580	2.980	301	335	360	1.77	75.4
9 (high jitters)	Skype	1.103	1.793	2.394	275	410	560	1.65	76.2
	GTalk	1.163	1.789	2.450	286	386	489	2.00	77.9
	Yahoo	0.905	1.585	2.353	251	369	465	2.06	76.4
	WinLive	1.795	2.300	2.759	897	1017	1160	4.06	66.6
10 (medium losses)	Skype	1.799	2.376	2.838	235	269	317	1.68	82.2
	GTalk	1.691	2.247	2.762	244	278	320	1.57	82.1
	Yahoo	2.158	2.505	2.908	145	202	254	1.50	82.9
	WinLive	2.494	2.946	3.404	318	350	372	1.96	80.8

Figure 4 depicts the experimental results under various levels of network delay and loss adaptation. Yahoo tries to minimize MEDs whenever possible, without sacrificing much on LOSQ. On the other hand, Windows Live uses higher MEDs (by larger jitter buffers) in general in order to achieve the highest and more consistent LOSQ. Finally, Google-Talk and Skype use conservative and more consistent MEDs but their LOSQ can fluctuate when network conditions change over time.

In short, Windows Live is more robust to packet losses by using higher MEDs, which suggests that it uses a redundancy-based loss-concealment scheme. However, Windows Live's choice of high MEDs of over 1 sec in high-jitter conditions leads to unacceptable CI and CE (Table 2).

5. REFERENCES

[1] B. Sat and B. W. Wah, "Playout scheduling and loss-concealments in VoIP for optimizing conversational voice communication quality," in *Proc. ACM Multimedia*, Augsburg, Germany, Sept. 2007.

[2] 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.

[3] 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*, 2004, vol. 3, pp. 1478–1483.

[4] N. Kiatawaki and K. Itoh, "Pure delay effect on speech quality in telecommunications," *IEEE Journal on Selected Areas of Communication*, vol. 9, no. 4, pp. 586–593, May 1991.

[5] C. Boutremans and J.-Y. Le Boudec, "Adaptive joint playout buffer and FEC adjustment for Internet telephony," in *Proc. IEEE INFOCOM*, 2003, vol. 1, pp. 652–662.