Assessing and Improving User QoE of Video Telephony

Yong Liu
Electrical & Computer Engineering Department
Polytechnic School of Engineering
New York University

Abstract

Video telephony is an important application. It is extremely challenging to deliver good user Quality of Experience (QoE) over the best-effort Internet. In this paper, we summarize our previous work on measurement study of commercial video telephony solutions and outline our current and future work on assessing and improving user QoE of video telephony.

1 Background and Motivation

The Internet has fundamentally changed the way people communicate, from emails, text-messages, blogs, tweets, to Voice-over-IP calls, etc. We are now experiencing the next big change: *Video Telephony*. Although video telephony was originally conceived in 1920s, largely due to its stringent bandwidth and delay requirements, it has little success, especially in the end-consumer market, until very recently. The proliferation of video-capable consumer electronic devices and the penetration of increasingly faster residential network accesses paved the way for the wide adoption of video telephony. Video telephony is more than two-party video chat. It also supports Multi-party Video Conferencing (MPVC) between multiple geographically distributed participants by transmitting audio and video signals among them in realtime. Compared with voice, video is much more bandwidth-demanding. While Skype encodes high-quality voice at data rate of $40kbps$, a Skype video call with good quality can easily use up bandwidth of $1Mbps$ [4]. Compared with video streaming, video telephony has much *tighter delay constraints*. While seconds of buffering delay is often tolerable in video streaming, in video telephony, user Quality-of-Experience (QoE) degrades significantly if the one-way end-to-end video delay goes over 350 milli-seconds [1]. To deliver good video telephony QoE over the *best-effort* Internet, video telephony solutions have to cope with user device and network access heterogeneities, dynamic bandwidth variations, and random network impairments, such as packet losses and delays. All these have to be done through video generation and distribution *in realtime*, which makes the design space extremely tight.

2 Measurement of Commercial Solutions

We recently conducted a comprehensive measurement study on three most popular video conferencing systems on the Internet: Facetime/iChat, Google+ Hangout, and Skype. *Our study was focused on their key design choices and their delivered user QoE. The ultimate goal is to understand how different design choices impact video telephony QoE perceived by end consumers.* All the systems use proprietary protocols and encrypt data and signaling. There is very limited public information about their architecture, video encoding and distribution algorithms. Through a series of carefully designed black-box measurements, we unveiled important information about their key design choices [4, 2, 3].

Distribution Architecture: Google+ is purely server-based. Each user is assigned to a proxy server. A user uploads her audio and video to her proxy and downloads voice and video of other users from her proxy. There is no direct communication between users. iChat employs simple P2P design. Each user always uploads her audio and video to the conference initiator, which then redistributes it to all other users. Essentially, the conference initiator acts as a server, but with only limited bandwidth. iChat cannot sustain good video quality when the number of users is more than three. Facetime only offers two-party video call, it uses direct P2P transmission between two users whenever possible. Skype employs a hybrid solution for MPVC.[1] Audio is distributed using P2P: each user uploads her audio to the conference initiator, which mixes all audio streams into one stream, then distributes it to all other users. Video is distributed using servers: each user uploads her video to a server, which then directly relays the video to other users.
Source Video Coding: Google+ employs layered video coding to address receiver heterogeneity. A source generates multiple video layers using temporal and spatial scalability. Different receivers get different numbers of video layers, matching their download capacities. In iChat, each source only generates one video version at a rate determined by the download capacity of the weakest receiver. In Skype, a source may generate multiple video versions and send different versions to different subsets of receivers.
Data Transport and Loss Recovery: For audio and video transmission, all the systems mostly use UDP at the transport layer.[2] At the application layer, Google+ and iChat/FaceTime employ RTP, while Skype uses its own protocol. They all adapt transmitted video quality and data rate to the available network bandwidth, packet loss and delay. Skype employs an overly aggressive FEC scheme for loss recovery. Due to the inability to differentiate congestion losses from random losses, Skype might increase its FEC ratio upon congestion, leading to a vicious-congestion-cycle [3]. The redundant traffic ratio in Google+ is small. iChat uses retransmission, but Facetime uses FEC for loss recovery.
User Quality-of-Experience: All systems can deliver video rate in a wide range from $30kpbs$ to $950kpbs$ in wired line and mobile wireless networks. User perceived end-to-end video delay is significantly longer than the end-to-end network delay. This suggests that video encoding, FEC coding, data buffering, and video decoding also incur significant delays. While those systems are robust against up to 10% random losses [2], they are highly vulnerable to bursty losses and long packet delays on wireless links with weak receptions [3]. Finally, the audio and video misalignment is more noticeable in Skype than in Google+ and Facetime/iChat, which is mainly due to its separated audio and video distributions.

3 On-going and Future Work

Leveraging on our measurement study, we are working on the following items regarding assessing and improving user QoE of video telephony.

- **Root cause analysis of video telephony quality degradation.** Video telephony is vulnerable to various network impairments, such as bandwidth dips, random and bursty packet losses, and congestion delays, etc. We are investigating how each of them contribute to quality degradation of video telephony, including low resolution and low frame-rate video, long audio and video delays, and video and audio freezes, etc. After identifying the root causes, we can design better video coding and transmission strategies as well as better network control and management to deliver better video telephony QoE.

- **Subjective study of video telephony quality.** Users have different preferences and tolerances in video telephony quality and degradations. We will conduct subjective study of real video telephony users to understand how they value different aspects of video telephony quality. The ultimate goal is

[1] Skype two-party video call uses direct transmission between the two users whenever possible.
[2] We did observe infrequent use of TCP, which, we believe, is mainly for getting around of firewalls blocking UDP traffic.

to establish Mean Opinion Score (MOS) type of video telephony quality models that relate user QoE of video telephony with various video encoding parameters and network QoS metrics.

- **Real-time bandwidth prediction and rate adaptation for video calls over cellular networks.** It is known that cellular links present highly varying network bandwidth and packet delays. If the sending rate of the video call exceeds the available bandwidth, the video frames may be excessively delayed, destroying the interactivity of video call. We are working on a cross-layer design of proactive congestion control, video encoding and rate adaptation to maximize the video transmission rate while keeping the one-way frame delays sufficiently low. Our system actively measures the available bandwidth in real-time by employing the video frames as packet trains. Using an online linear adaptive filter, it then makes a history-based prediction of the future capacity, and determines a rate budget for the video rate adaptation. Finally, it decides in real-time whether to send or discard an encoded frame, according to the budget, thereby preventing self-congestion and minimizing the packet delays.

References

[1] JANSEN, J., CÉSAR, P., BULTERMAN, D. C. A., STEVENS, T., KEGEL, I., AND ISSING, J. Enabling composition-based video-conferencing for the home. *IEEE Transactions on Multimedia 13*, 5 (2011), 869–881.

[2] XU, Y., YU, C., LI, J., AND LIU, Y. Video telephony for end-consumers: Measurement study of google+, ichat, and skype. In *ACM Internet Measurement Conference*. Best Paper Award.

[3] YU, C., XU, Y., LIU, B., AND LIU, Y. "Can you SEE me now?", A Measurement Study of Mobile Video Calls. In *Proceedings of IEEE INFOCOM* (2014).

[4] ZHANG, X., XU, Y., HU, H., LIU, Y., GUO, Z., AND WANG, Y. Profiling Skype Video Calls: Rate Control and Video Quality. In *Proceedings of IEEE INFOCOM* (March 2012). `http://eeweb.poly.edu/faculty/yongliu/docs/skype.pdf`.