Towards the Understanding of Human Perceptual Quality in Tele-Immersive Shared Activity

Zixia Huang, Ahsan Arefin, Pooja Agarwal, Klara Nahrstedt, Wanmin Wu
Department of Computer Science
University of Illinois at Urbana-Champaign
{zhuang21, marefin2, pagarwal, klara, wwu23}@illinois.edu

ABSTRACT

ITU-T G.1070 [7] is widely cited for evaluating the subjective quality of the video conferencing, but its findings cannot be applied to the tele-immersion. The reasons are two fold. First, a tele-immersive system offers end users an unmatched realistic and immersive experience by allowing them to collaborate in the joint virtual space. Second, the human activities in the shared space are not limited to the conferencing conversation. In this paper, we conduct a user study with 19 participants to investigate the human perceptions of two tele-immersive shared activities, where media samples of different qualities are evaluated using the comparative category rating method [9] in case of each activity. We compare our subjective results to those presented in G.1070, and demonstrate heterogeneous human perceptual impacts in different activities.

Categories and Subject Descriptors

H.1.2 [**Information Systems**]: Models and Principles: Human factors; H.4.3 [**Information Systems Applications**]: Communications Applications: Computer conferencing, teleconferencing, and videoconferencing

General Terms

Experiment, Measurement

Keywords

3D Tele-immersion, Subjective Quality Assessment

1. INTRODUCTION

Researchers usually propose objective metrics to describe the quality of service (QoS) of media applications in various dimensions. However, these QoS metrics alone are unable to characterize the human perceptions, and it is difficult to formulate their combined effects in a closed form [1, 4]. Hence, subjective evaluations are needed to describe real user experience in media applications, and to guide the system adaptations.

Lots of subjective studies [2, 14, 15] have employed the *absolute category rating* (ACR) method proposed in ITU-T BT.500 [5], in which participants observe one single media sample and give an ACR score from 1 to 5 (a higher score is better). However, the problem of ACR is that a standard rating scale is missing due to the absence of a *reference sample* (i.e., a prescribed sample with the best possible quality). Thus, the participants in the studies usually give a score based on their own expertise. This leads to the non-uniform distributions of rating scores, which can invalidate the subjective results.

To address the ACR drawback, ITU-T P.910 [9] proposes an alternative assessment method, in which participants now observe two media samples of different qualities and give a comparative rating score using the *comparative category rating* (CCR) (details in Section 2). When one of the media sample in a comparison pair is the reference, CCR is reduced to a *degradation category rating* (DCR) method [9]. ITU-T G.1070 [7] and G.107 [6] have followed the subjective methods to assess real-time interactive video and audio conferencing.

Unfortunately, the subjective results obtained in G.1070 are unable to be applied to tele-immersive (TI) applications. The reasons are two fold. First, multiple distributed people, located in different geographical locations, collaborate in a joint virtual space in the tele-immersion. This immersive setting offers a realistic perception, which cannot be matched by traditional conferencing systems. Second, end users are not limited to conferencing conversation in the tele-immersive shared activities (TISA) [3]. They can also perform other tasks, e.g., remote education and online gaming.

Contributions. The limitations of G.1070 have motivated us to evaluate the human perceptions in TISA. In this paper, we conduct a user study, where 19 people are invited to the subjective comparison tests. We make three contributions. First, we propose a systematic methodology to demonstrate the CCR effectiveness in the subjective evaluations. Second, we investigate the human perceptual heterogeneity between TISA and video conferencing. Third, we show diverse impacts of TI activities on user satisfactions.

Our previous subjective studies on TISA either used the ACR method [15] to evaluate the interactive system quality, or employed the CCR method where we focused on an non-interactive environment [16].

MMSys'12 February 22–24, 2012, Chapel Hill, North Carolina, USA.

Table 1: Abbreviations and Definitions.

Abbr	Definitions
TISA	Tele-immersive shared activity
CMOS	Subjective metric: comparative mean-opinion-score
CCR	Comparative category rating
DCR	Degradation category rating
PESQ	Perceptual evaluation of speech quality
HRD	Human response delay
CONV	Conferencing social conversation activity
COLL	Collaborative gaming activity
x_V	Objective metric: multi-view video frame rate
x_A	Objective metric: PESQ
x_D	Objective metric: interactivity factor
x_S	Objective metric: audio-visual synchronization skew
$\vec{x}$	4-dimensional objective quality point
$\vec{x}^*$	The optimal reference of TISA sample
EED_V	End-to-end delay for multi-view videos
EED_A	End-to-end delay for audio frame
C	CCR rating score set

2. TISA QUALITY METRICS

The goal of this paper is to use CCR to evaluate the diverse human perceptual quality in heterogeneous TISA. In order to capture different characteristics of TISA, we study two representative activities in our subjective evaluations: conversation-oriented (CONV) tasks and collaborative gaming (COLL) in this paper. The CONV activity describes the conferencing scenario with a social conversation, where participants at both systems are talking to each other with slow motion movement (Fig. 1(a)). In COLL (Fig. 1(b)), two distributed participants are playing the "rock-paper-scissor" game in the virtual environment.

We then conduct subjective user study and evaluating the TISA samples of different qualities in case of each activity. The following four steps are needed in realizing this goal: (1) identifying user-observable objective metrics to capture different TISA quality dimensions; (2) preparing TISA samples based on these objective quality metrics with different values; (3) specifying the subjective rating scales used in the user study; and (4) identifying subjective quality metrics to evaluate the collected user data. In this section, we will investigate both objective and subjective quality metrics for TISA evaluations. A summary of mathematical denotations used for the rest of this paper is presented in Table 1.

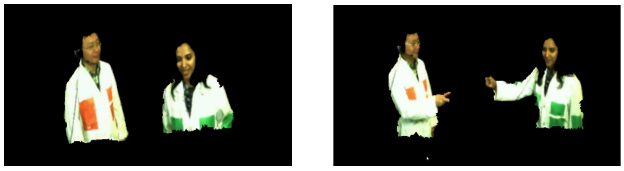
2.1 Objective Metrics

• Media Signal Quality

The media signal quality in TISA includes the audio quality x_A and the multi-view video quality x_V. Both metrics can be degraded by jitters and losses over the wireline and wireless networks.

For both wideband and narrowband audios, we use the *Perceptual Evaluation of Speech Quality* (PESQ) metric defined in ITU-T P.862 [8] to approximate x_A. PESQ allows the automatic computation of the quality of a (degraded) audio signal in the presence of the original reference. It returns a score ranging from 1 to 4.5. A larger PESQ means the (degraded) audio signal is more approximate to the reference, and hence a better audio intelligibility.

There are lots of factors deciding the multi-view video quality (rendered on the 2D screen): the multi-view video frame rate, the spatial resolution, the encoding quality and the number of views available in TISA. In this paper, we simplify the problem by only focusing on the multi-view video frame rate x_V. A larger x_V means a greater motion smoothness and hence a better video signal quality. We reduce the TISA sample space by assuming a fixed spatial resolution, encoding quality and view number in our study.



(a) CONV: conferencing conversation (b) COLL: rock-paper-scissor gaming

Figure 1: Two TISA evaluated in our study.

• Synchronization Quality

The audio and multiple multi-view video streams can experience different *end-to-end delays* (EED) between two distributed users. We assume that the set of time-correlated multi-view video frames is synchronized before it is sent to the display renderer for the purpose of accurate multi-view video rendering. Hence, we only investigate the impact of the resulting audio-visual synchronization skew x_S on human perceptions. We use EED_V to represent the duration between the time that a time-correlated multi-view video frame set is synchronously captured at the camera, and the time that it is displayed on the screen. EED_A is used to denote the duration between the microphone and speaker for an audio frame. Hence, x_S can be represented as:

$$x_S = \text{EED}_V - \text{EED}_A \tag{1}$$

Note that $x_S > 0$ means the audio is ahead of video, and that $x_S < 0$ means the audio is behind.

• Interactivity

In CONV, the perception of a user on the interactivity is impacted by the delayed response of the remote site. A user can become impatient when the response delay accumulates, and the remote person becomes more distant. Doubletalks [4] may be introduced at an extremely long delay, when the user begins to repeat his statement, assuming his previous words are dropped during the transmission. Hence, the interactivity attribute can be characterized by the *response delay* (x_D), which is incurred by the EED of local media streams (denoted as $\overline{\text{EED}}$) to the remote site, the duration required for the remote user to think of a response (i.e., human response delay (HRD) [4]), and EED of the remote streams traveling back to the local site. Fig. 2 shows the concept. Mathematically, x_D that a local user experiences can be represented as:

$$x_D = \overline{\text{EED}}^{U1\rightarrow U2} + \text{HRD}^{U2} + \overline{\text{EED}}^{U2\rightarrow U1} \tag{2}$$

where $U1$ and $U2$ represent the local and remote users, and HRD^{U2} is the $U2$'s HRD.

On the other hand, the interactivity attribute in COLL is mainly evaluated by the *collaborative* performance of the two participants involved in the task. Here, "collaborative" means that two participants are following each other to achieve a mutual goal. A person (called *initiator*) initiates a gesture, and the other person (called *follower*) must exactly follow at the same time (i.e., HRD ≈ 0). The two roles can be swapped during the activity. Because of the bi-directional EEDs of the media streams between the two parties, the *response delay* x_D that an initiator perceives can be described as the timing mismatch in the collaboration on

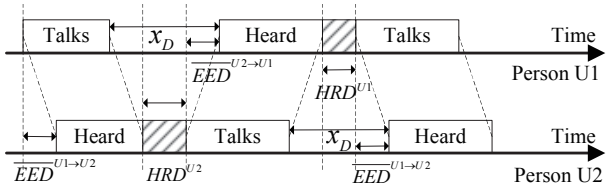


Figure 2: Interactivity in conversation-oriented activity.

his/her own rendering display (Fig. 3). In this case, x_D can be formulated as:

$$x_D = \overline{\text{EED}}^{U1\to U2} + \overline{\text{EED}}^{U2\to U1} \quad (3)$$

Because EED_A and EED_V may be different between two sites, we follow ITU-T G.1070 [7] and give both metrics an equal weight in computing $\overline{\text{EED}}$ in Eqn. 2 and 3, i.e.,

$$\overline{\text{EED}} = (\text{EED}_A + \text{EED}_V)/2 \quad (4)$$

• Combined Impacts

The overall human subjective perceptions of TISA are impacted by the combined impacts of the above user-observable quality attributes, which can be described by a 4-dimensional objective quality space with each objective quality *point* $\vec{x}$ in the space representing:

$$\vec{x} = \{x_V,\ x_A,\ x_D,\ x_S\} \quad (5)$$

In our user study, we create TISA samples with different configurations $\vec{x}$ (i.e., different values in one or multiple dimensions in $\vec{x}$). Throughout this paper, we use *frames per second* (fps) for the unit of the multi-view video frame rate x_V, *milliseconds* (ms) for the response delay x_D and the audio-visual sync skew x_S, and [1, 4.5] for the audio quality x_A.

2.2 Subjective Metrics

We focus on the subjective assessment tests in which two media samples with different configurations $\vec{x}$ are given to the participants consecutively, and each participant employs the CCR scale to compare the two samples. We use a comparison voting score set of C ={3, 2, 1, 0, -1, -2, -3}. This score set represents the scoring values to indicate that the quality of the first sample is {*much better*, *better*, *slightly better*, *same*, *slightly worse*, *worse*, *much worse*} than that of the second sample. We then process the votes using the following metrics.

We compute the average of people voting scores as the comparative mean-opinion-score (CMOS), as defined in ITU-T P.910 [9]. 95% confidence intervals are also calculated, assuming the t-distribution of the votes.

3. DESCRIPTIONS OF USER STUDY

Based on the discussion in Section 2 and 3, we present the configurations of our user study in assessing the subjective quality of two TI activities (i.e., CONV and COLL).

3.1 Methodology

Our user study investigates the end user experience at various TISA qualities and shows the CCR effectiveness. To find the mappings from the objective quality metrics (Section 2.1) to subjective space (Section 2.2), we create TISA samples with different configurations $\vec{x} = \{x_V,\ x_A,\ x_D,\ x_S\}$ (Eqn. 5). However, the value of $\vec{x}$ can be continuously changing in its 4-dimensional space, and thus, there can be

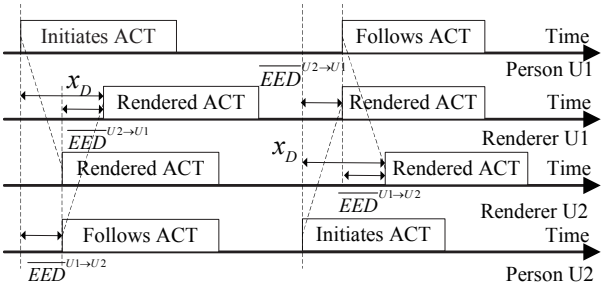


Figure 3: Interactivity in collaborative activity (ACT).

Table 2: Discretization of quality metrics in $\vec{x}$. HRD = 800 ms is used in computing x_D in CONV (Section 3.2). x_V is rounded to the nearest integer in the evaluation.

Metric	Unit	Discretiziation
x_V	fps	2.5, 5, 7.5, 10, 12.5, 15, 17.5, 20
x_A	[1, 4.5]	2.0, 4.0
x_S	ms	0, ± 75, ± 150, ± 225
x_D (CONV)	ms	1000, 1200, 1400, 1600, 1800, 2000, 2200, 2400, 2600
x_D (COLL)	ms	120, 180, 240, 300, 360, 420, 480, 540, 600

infinite number of options for $\vec{x}$. In this study, we discretize each metric within $\vec{x}$ (Table 2) according to the characteristics of real media traffic in the Internet.

We then ask the participants to compare TISA samples of the same activity in each test. We employ CCR rating scale as discussed in Section 2.2. We process the user subjective feedback accordingly. Due to the page limit, in this paper we will only present the impact of each quality dimension in $\vec{x}$ by keeping values in other dimensions fixed. The evaluation results concerning multi-dimensional quality tradeoffs (i.e., a media sample is better than another sample in some quality dimensions, but worse in other dimensions [4]) will be deferred to a future full-version paper.

3.2 Preparation of TISA Samples

We let two participants be situated at different sites and conduct activities through the TI system. The two sites are in the same local area network (LAN), so the outputs should be assumed to have no video and audio signal degradation with minimal latency and perfect synchronization. We record the distortion-free audio and video at both sites. For the video, because the TI system eventually displays the multi-view images on the 2D screen, we record the 2D video including both participants which is exactly shown on the screen (using the *xvideocap* software[1]) instead of the original multi-view images. For the audio, we mix the audio talkspurts of the two parties (using the *Virtual Audio Cable* software[2]), and *xvideocap* can also be utilized to record the mixed audio, with an automatic synchronization with the video.

We create TISA samples for both CONV and COLL applications. In CONV, we follow our previous VoIP study and use a HRD of 800 ms, and an average talkspurt duration of 2732 ms in our simulation [4]. In COLL, the average duration of talkspurts is 856 ms. In this study, the reference

[1] `http://xvidcap.sourceforge.net`
[2] `http://software.muzychenko.net/eng/vac.htm`

sample with the best-possible (called *optimal* in this paper) quality, assuming two sites are communicating in LAN, is $\vec{x}^* = \{20,\ 4.0,\ 800,\ 0\}$ for CONV, or $\vec{x}^* = \{20,\ 4.0,\ 0,\ 0\}$ for COLL.

We now assume that one TI site is local and the other is remote. We introduce the delay and sync skews for the remote streams, and impose degradations on its media signal quality (reduced x_V and x_A). The qualities of local audio and images remain untouched. The degraded TISA sample $\vec{x}$ describes the objective quality of the remote streams.

3.3 Setup of User Study

19 participants (average age: 26) are involved in our user study, and are trained to use CCR scales consistently before the subjective tests. They are required to sit 1.5 meter apart from a 61-inch NEC screen (resolution: 1280x720), and to rate TISA samples at different $\vec{x}$ values. The video is rendered at a resized resolution of 640x360 (original resolution: 420x240). The audio is played at a DELL AY410 2.1 speaker. To simulate a real TISA involvement, these observers are told to be assuming themselves sitting closely to the person in the local site so they can pay more attentions to the (degraded) quality of the remote person.

There are a total of 240 comparisons of TISA samples (with different configurations $\vec{x}$) within the whole test. Participants are able to pause at any time throughout the test. There are 10-second idle pauses between two consecutive comparisons, so that observers have sufficient time to consider their votes.

4. EVALUATION RESULTS

In this section, we present our CCR-based user study results. We will show the two TI activities (CONV and COLL) have heterogeneous impacts on human perceptions. As a comparison, we will also discuss the results from existing subjective studies (and particularly, G.1070).

4.1 Media Signal Quality

• Audio Signal Quality

The audio PESQ (i.e., x_A), as its name suggests, is computed on a psycho-acoustic scale which is already able to describe the real human subjective perceptions on audio signals. That is to say, when we fix x_V, x_D and x_S as optimal, we are able to approximate the impairment of x_A as:

$$\text{CMOS}(x_A) = 4.5 - x_A \tag{6}$$

Here, 4.5 is the maximal-possible value of x_A. Our CCR findings are aligned with the PESQ results, thus showing the CCR effectiveness (data not plotted due to space limit).

• Video Signal Quality

Previous work. G.1070 estimates the video signal quality based on the coding distortion and packet losses robustness. The standard focuses on the video image artifacts by assuming the availability of some loss concealment mechanisms within the 2D video codec. These metrics, however, are inapplicable to the current multi-view video codec used in our TI testbed. On the other hand, [11] utilizes an exponential model to identify the impact of 2D video frame rate on the video signal degradations. Because the TI 3D multi-view videos will eventually be rendered on a 2D display, this mathematical model lays a theoretical foundation for our study.

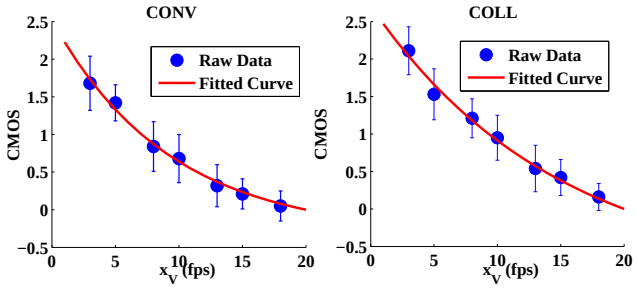


Figure 4: CMOS and 95% confidence intervals comparing optimal reference $\vec{x}^*$ (first sample in the comparison) and x_V-degraded samples (but optimal x_A, x_D, x_S).

Table 3: Fitting results for Eqn. 7.

TISA	Q	c	MSE
CONV	2.52	2.16	0.01
COLL	2.71	1.35	0.01

Our results. Fig. 4 shows the CMOS results comparing $\vec{x}^*$ to the samples with different degraded x_V while keeping other quality dimensions optimal. We modify the exponential model in [11] to find the fitting curve describing the mapping from x_V to the corresponding CMOS:

$$\text{CMOS}(x_V) = Q - Q \times \frac{1 - e^{-c \times x_V / x_V^{\max}}}{1 - e^{-c}} \tag{7}$$

In this equation, c is the slope of the curve, which describes the impact of x_V changes on the CMOS. A smaller c will introduce a larger degradation to CMOS at the same x_V. Q represents the maximum-possible impairment of x_V. $x_V^{\max}$ is set to be 20 fps, the maximum multi-view video frame rate in our study. We want to find the best fitting parameters Q and c of the exponential curve. We utilize the nonlinear fitting tool in Matlab (*nlinfit* function) to compute Q and c. The fitting results as well as the corresponding mean squared error (MSE) are shown in Table 3. Because c is smaller in COLL, an equal x_V decrease can cause more perceptual degradations in COLL than CONV. The reason is due to more frequent body movement in the COLL activity.

In both cases, the reasonable confidence interval lengths (within $\pm 0.2 \sim \pm 0.32$) show the CCR effectiveness.

4.2 Synchronization Impairment

Previous work. We focus on the audio-visual lip synchronization. There have been many studies working on the subjective perceptions of synchronization impairment, but none of them can be directly used in our TISA scenario.

For video conferencing, G.1070 uses a linear form to describe the human perceptual impairment of the lip skew (i.e., x_S) on a dedicated videophone terminal with a maximum screen size of 4.2 inch. Their proposed coefficients characterizing synchronization impairment are, however, independent of the media signal quality.

For on-demand videos, Steinmetz and Nahrstedt [12] recommend an in-sync region of a maximum 80-ms skew for a video, and they show that an out-of-sync skew of more than 160 ms is unacceptable. But their study assumes perfect media signal quality during the synchronization evaluation, and it does not take into account the impact of the video content heterogeneity.

Our results. In Fig. 5 (1) and (3), we carry on experiments to evaluate the lip skew impairment at the optimal

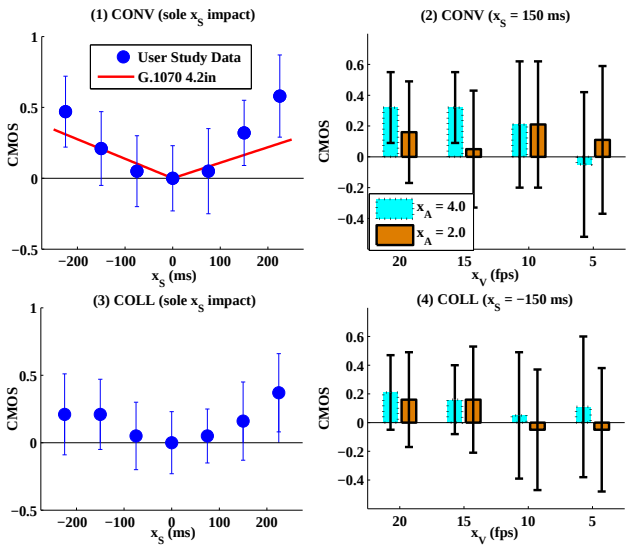


Figure 5: CMOS results and 95% confidence intervals. (1) and (3) show the comparisons between the optimal reference $\vec{x}^*$ (first sample in the comparison) and a sample with a degraded x_S (but optimal x_V, x_A, x_D). (2) and (4) show the two samples with different x_S, but same x_V, x_A, and $x_D = 800$ ms for CONV and 0 ms for COLL. The first sample in the comparison is $x_S = 0$ and the second is $x_S = \pm 150$ ms.

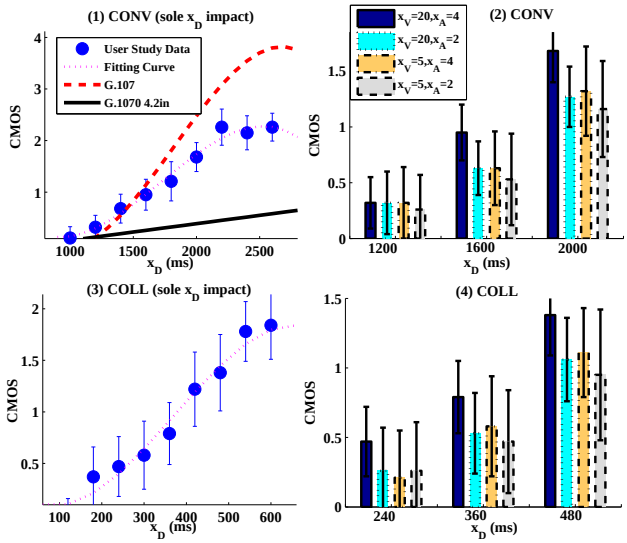


Figure 6: CMOS results and 95% confidence intervals. (1) and (3) show the comparisons between the optimal reference $\vec{x}^*$ (first sample in the comparison) and a sample with degraded x_D (with optimal x_V, x_A and x_S). G.107 and G.1070 delay curves are also drawn in (1). (2) and (4) show the results for two samples with different x_D, but same x_V, x_A, and $x_S = 0$. The first sample in the comparison is $x_D = 800$ ms for CONV and 0 ms for COLL, and the second is with the degraded x_D as indicated in the figures.

x_V, x_A and x_D, when compared to $\vec{x}^*$. We show the CMOS results and 95% confidence intervals at different x_S. In Fig. 5 (2) and (4), we evaluate the impact of x_V and x_A on the synchronization quality. Due to the space limit, we only show the results at selected x_S values ($x_S = 150$ ms for CONV and -150 ms for COLL) with different x_V and x_A options, compared to the samples of $x_S = 0$ with the same media signal quality. We have three observations.

First, our limited study reflects that the heterogeneous TI activities can affect the synchronization perfection. Generally, the degradation of a lip skew in the COLL environment is smaller than that in CONV with the same skew, because (1) the talkspurt durations in COLL are much shorter, and (2) people are focusing on the visual collaborative activity more than talkspurts in COLL. The lengths of confidence intervals are comparable in the two applications.

Second, our study exhibits that people are more tolerant of video ahead of audio ($x_S < 0$) than audio ahead of video ($x_S > 0$). The reason is that the talkspurt durations in TISA are generally much shorter than those in on-demand videos, so a lip skew at the end of an utterance is more noticeable. Fig. 5 shows that a late video portion at the time that an utterance has been fully played has a greater perceptual impact than a late audio portion. Our findings are aligned with the results in [12].

Third, Fig. 5 (2) and (4) show that both x_V and x_A do impact the synchronization quality. We find that the lengths of confidence intervals are much larger ($> \pm 0.4$) as x_V and x_A degrade (e.g., x_V =5 fps or $x_A = 2.0$). When the multi-view video frame rate lowers, the motion jerkiness becomes the dominant factor degrading the human perceptions, and thus, a lip skew can be difficult to tell. On the other hand, when x_A is small, the poor audio intelligibility also creates a hard time for users to differentiate a lip skew, and an incomplete utterance can cause misperception on the synchronization quality.

4.3 Interactivity

Previous work. Both G.1070 and G.107 study the interactivity (delay impairment) in video and audio conferencing. G.107 uses a complex sixth-order model to describe the VoIP delay impairment (which is independent of audio quality). G.1070, on the other hand, employs a linear function to present the delay impacts in the video conference. The standard shows that the delay degradation is much smaller than VoIP applications.

Our results. We conduct tests for evaluating the x_D impairment. These include two sets of comparisons. In the first set, we study the sole x_D impact at the optimal x_V, x_A and x_S. We show the corresponding CMOS and confidence intervals by referencing $\vec{x}^*$ in Fig. 6 (1) and (3). The G.107 and G.1070 findings are also plotted in CONV as a comparison. In the second set, we study the effects of the media signal quality on the x_D perception. Due to space limit, Fig. 6 (2) and (4) only show the results for selected x_D values. There are several observations.

First, we follow [13] and use a third-order polynomial model to describe the CMOS degradations due to x_D. The results are shown in Fig. 6 (1) and (3).

$$\text{CMOS}(x_D) = a_0 + a_1 \cdot x_D + a2 \cdot x_D^2 + a_3 \cdot x_D^3 \qquad (8)$$

Table 4 presents the fitting results both activities as well as the corresponding MSE. Generally for CONV, $x_D < 1200$ ms is desired (CMOS < 0.5) and $x_D > 2000$ is bad (CMOS > 1.5). For COLL, $x_D < 200$ ms is desired (CMOS < 0.5) and $x_D > 400$ is bad (CMOS > 1.5). Hence, the COLL application requires a higher demand for interactivity than CONV. This is because people in COLL attach more importance to the visual timing mismatch in the collaboration. The derived curves prove the CCR effectiveness in describing human perceptions.

Table 4: Fitting results for Eqn. 8.

TISA	a_3	a_2	a_1	a_0	MSE
CONV	1.033^{-9}	5.342^{-6}	-0.007	3.036	0.010
COLL	-1.945^{-8}	2.163^{-5}	-0.003	0.231	0.009

Table 5: Comparisons for CONV and COLL characteristics. Note that H/L mean comparatively more/less important between the two application.

	x_V	x_A	x_D	x_S
CONV	L	H	L	H
COLL	H	L	H	L

Second, we find that our CONV findings are in between the G.107 and G.1070 delay curves. The reason is that a user in a VoIP application (G.107) usually lacks a perception of the activities of the remote party. So the local person is prone to assuming the remote talkspurts have been dropped by the Internet at a delayed response, and may repeat his/her utterances which can cause doubletalks. On the other hand, a person in either a video conferencing (G.1070) or a TI session is able to see what the remote user is doing, and hence he/she is more tolerant of the delay. But in TISA, because both people are located in an immersive environment, a higher demand for interactivity is expected, compared to the video conferencing. In addition, the delay results that G.1070 obtains are somewhat too conservative.

Third, we demonstrate that the media signal quality does affect the interactivity perception, as in Fig. 6 (2) and (4). The figures show that, a delayed response has less impacts on human perceptual degradations (smaller CMOS in the figures) in an environment with reduced video motion smoothness and audio signal intelligibility.

5. IMPLICATIONS TO TI SYSTEM DESIGN

The above discussions imply two important aspects in designing an interactive TI system.

TISA heterogeneity. A good media system should not only be able to adapt to Internet dynamics, but also be built upon the heterogeneous characteristics of TI applications to meet the real user demands. From Section 4, we qualitatively conclude the perceptual importance for the two TI activities in Table 5. Compared to COLL, CONV generally requires a higher demand for the audio signal intelligibility and the constrained lip skew, but a lower expectation on the video motion smoothness and interactivity. These characteristics should be addressed in the system design.

Ordering of subjective scores. Previous studies on VoIP or video conferencing [10, 13] usually propose adaptation algorithms based on the the (extended) quality models used in G.107 and G.1070. Here, we argue that the quality closed forms derived in both standards are only suitable for subjective quality assessment of media samples, and the resulting score orderings are not good for system adaptations.

As our study shows, multiple quality points, which are distant in the multidimensional Euclidean space, can lead to same or similar CMOS when they are compared to the optimal reference $\vec{x}^*$. For example in CONV, $\vec{x}^1 = \{12, 4.0, 0, 0\}$ in Fig. 4(1), $\vec{x}^2 = \{20, 4.0, 0, -225\}$ in Fig. 5(1), and $\vec{x}^3 = \{20, 4.0, 1300, 0\}$ in Fig. 6(1) all lead to CMOS of around 0.5. If we achieve adaptation based on the score ordering, we may switch between two operating configurations $\vec{x}$, which are close in the subjective score space, but are actually distant in terms of each quality dimension [3]. This can cause *flicker effects* (i.e., the perceptible change of media quality dimensions). The quality flickers should be minimized, which would otherwise downgrade human perceptions [3].

6. CONCLUSION

In this paper, we propose a systematic methodology to investigate the subjective quality in TISAs. We show the CCR effectiveness in evaluating the diverse human perceptual degradations in heterogeneous activities, under the impact of each quality dimension. However, existing subjective metrics are unable to capture the multi-dimensional tradeoffs (as proved in VoIP [4]). So the next step is to study this limitation in a TI setting, and to propose new subjective evaluation framework to address the issue.

Acknowledgement. We appreciate constructive comments from our shepherd, Prof. Sheng-Wei (Kuan-Ta) Chen, and the three anonymous reviewers. This research study is supported by NSF CNS 0834480, 0964081, 1012194, IIP 1110178, by UPCRC grant from Intel and Microsoft, and by Grainger Grant.

7. REFERENCES

[1] K.-T. Chen, C. C. Tu, and W.-C. Xiao. Oneclick: A framework for measuring network quality of experience. In *Proc. of IEEE INFOCOM*, 2009.

[2] O. Daly-Jones, A. Monk, and L. Watts. Some advantages of video conferencing over high-quality audio conferencing: fluency and awareness of attentional focus. *Int'l Journal of Human-Computer Studies archive*, 49(1), July 1998.

[3] Z. Huang and K. Nahrstedt. Perception-based media packet scheduling for high-quality tele-immersion. In *Proc. IEEE Int'l Conference on Computer Communications*, Mar. 2012.

[4] Z. Huang, B. Sat, and B. W. Wah. Automated learning of play-out scheduling algorithms for improving the perceptual conversational quality in multi-party VoIP. In *Proc. IEEE ICME*, 2008.

[5] ITU-BT.500. Methodology for the subjective assessment of the quality of television pictures, 2002.

[6] ITU-G.107. The E-model, a computational model for use in transmission planning, 2008.

[7] ITU-G.1070. Opinion model for video-telephony applications, 2007.

[8] 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, 2001.

[9] ITU-P.910. Subjective video quality assessment methods for multimedia applications, 2008.

[10] A. Meddahi, H. Afifi, and G. Vanwormhoudt. "MOSQoS": Subjective VoIP Quality for Feedback Control and Dynamic QoS. *Proc. of IEEE Int'l Conference on Communications*, 2006.

[11] Y.-F. Ou, T. Liu, Z. Zhao, Z. Ma, and Y. Wang. Modeling the impact of frame rate on perceptual quality of video. In *Proc. of IEEE ICIP*, pages 689–692, 2008.

[12] R. Steinmetz and K. Nahrstedt. Multimedia computing, communications and applications, Prentice Hall, 1995.

[13] L. Sun and E. Ifeachor. Voice quality prediction models and their application in VoIP networks. *IEEE Communications*, 3:1478–1483, 2004.

[14] A. Vatakis and C. Spence. Evaluating the influence of frame rate on the temporal aspects of audiovisual speech perception. *Neuroscience Letter*, 11(405), Sept. 2006.

[15] W. Wu, A. Arefin, Z. Huang, P. Agarwal, and et al. I'm the Jedi! - A case study of user experience in 3D tele-immersive gaming. In *Proc. IEEE ISM*, 2010.

[16] W. Wu and et al. Color-plus-depth level-of-detail evaluation metric for 3d teleimmersive video. In *ACM MM*, 2011.