Understanding the Human Perceptions 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

Both comparative category rating (CCR) and degradation category rating (DCR) methods [20] have been heavily employed in the subjective evaluations of media systems. The resulting metrics, comparative mean-opinion-score (CMOS) and degradation mean-opinion-score (DMOS), can be used to describe the system subjective quality. However, the subjective metrics may work unsuccessfully when the variance of participant votes is large. The diversity in human interests can appear due to the tradeoffs of multiple quality dimensions, which concurrently dominate the overall quality of the media system. In this paper, we conduct a user study with 19 participants to evaluate the subjective quality of two tele-immersive shared activities (TISA), where media samples of different qualities are evaluated in case of each activity. Our study aims at (1) showing the effectiveness and limitation of CMOS and DMOS using real subjective data, and (2) demonstrating the heterogeneous impacts of TISAs on human perceptions.

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 aspects. However, these QoS metrics alone are unable to characterize the human perceptions, and it can be difficult to formulate their combined effects in a closed form. Hence, subjective evaluations are needed to evaluate real quality of experience (QoE) in media applications and guide the system adaptations.

Lots of subjective studies [6, 10, 41, 43, 44, 46] have employed the *absolute category rating* (ACR) method proposed in ITU-T BT.500 [15], in which participants observe one single media sample and give an ACR score from 1 to 5 (a higher score is better). The average of user voting scores is computed as the *mean-opinion-score* (MOS). 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 [20] proposes an alternative assessment method, in which participants now observe two media samples and give a comparative rating score. This can be either the *degradation category rating* (DCR) in which a degraded sample is compared against a reference sample, or the *comparative category rating* (CCR) in which any two media samples with different qualities are compared together. Participants give voting scores in the comparison process (details in Section 3), and the resulting average score is either the *degradation mean-opinion-score* (DMOS) or the *comparative mean-opinion-score* (CMOS). In this sense, DCR can be looked at as a subset of CCR, and CMOS can be used to approximate DMOS (Section 3.2). Several studies [13, 16, 17, 34] have utilized DCR and CCR in their subjective evaluations.

While CCR (and DCR) can generally perform far better than ACR in terms of rating scaling uniformity, we argue that its resulting subjective metric CMOS is unable to capture the variance of user votes. The problem is that the quality of a media system can be concurrently dominated by multiple quality dimensions (i.e., video frame rate, one-way delay, etc.). Hence, the tradeoffs among these dimensions in a comparison test can trigger the diversity of human preferences (Section 3.2), which has been demonstrated in our past VoIP studies [13, 34]. Note that the problem can only happen in CCR when multi-dimensional quality tradeoffs exist, so neither media sample in the comparison is the reference. No other study has investigated the CCR issue in the interactive video systems though.

Contributions. The problem of CMOS in capturing the user interest diversity has motivated us to evaluate the hu-

Technical report, University of Illinois at Urbana-Champaign
Submitted in Dec 12, 2011.

Table 1: Abbreviations and Definitions.

Abbr	Definitions
TISA	Tele-immersive shared activity
MOS	Subjective metric: mean-opinion-score
CMOS	Subjective metric: comparative mean-opinion-score
DMOS	Subjective metric: degraded mean-opinion-score
VAR	Subjective metric: variance of participant votes
CCR	Comparative category rating
DCR	Degraded 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 macroframe 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 video macroframe
EED_A	End-to-end delay for audio frame
C	CCR rating score set
N_{total}	Total number of votes
$N_{>,=,<0}$	Number of votes which give a score $>$, $=$ or <0
N_{th}	Threshold of vote number as inconclusive

man perceptions in the tele-immersive shared activity (TISA) [14] (described in Section 2). In this paper, we conduct a user study and invite 19 people to participate in the subjective comparison tests. Our main contributions are four fold. First, we propose a systematic methodology to demonstrate the effectiveness and limitation of CMOS and DMOS metrics. Second, we show that the CMOS is not sufficient to describe the subjective comparison results where tradeoffs of multiple quality dimensions are involved. Third, we present that human perceptions can be affected heterogeneously in different TI activities. Fourth, we conclude that there is a demand for a new metric to interpret the subjective comparison results to address the CMOS limitation. Proposing such a subjective metric, however, is beyond the scope of this paper.

Our previous subjective studies on TISA either used ACR [43, 44, 46] to evaluate the interactive system quality, or employed CCR where we only focused on the impact of a single quality dimension (e.g., 3D video depth in [45]) in an non-interactive environment.

Outline. We give a brief description of TISA in Section 2. We investigate its objective and subjective quality metrics and present a survey of existing subjective quality assessment studies in Section 3. We describe our user study configurations in Section 4. We analyze our subjective findings and discuss their implications on the system design in Section 5. Section 6 concludes the paper. A summary of mathematical denotations used for the rest of this paper is presented in Table 1.

2. TISA BACKGROUND

Interactive tele-immersive (TI) applications can offer a joint holographic environment where distributed users at different geographical locations are able to conduct shared activities with an unmatched realistic experience. Unlike the commercial teleconferencing or telepresence systems [1, 2, 4], where users may find themselves talking to the screen, TI applications can enhance the traditional remote communication style by allowing full-body interactions in an immersive collaboration with multi-sensory feedback. Apart from the conferencing capability, useful applications have also been found in medical consultation, cyber-archeology and collaborative gaming (Fig. 1) [5, 11, 43].

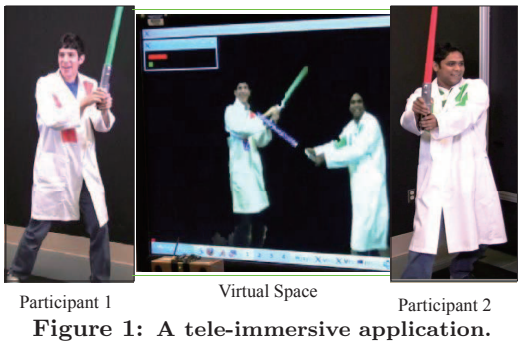


Figure 1: A tele-immersive application.

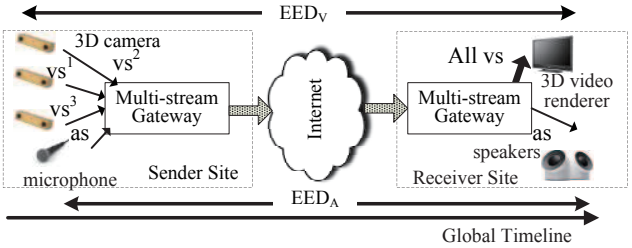


Figure 2: A tele-immersive dissemination topology. vs: video stream, as: audio stream.

The characteristics of a TI application (Fig. 2) include the capturing of an audio stream and multiple video streams of the local participant, the real-time dissemination of the media data, and the rendering of the remote audio and multi-view videos. The application also displays the local video on its own screen to emulate a joint virtual space. To provide seamless TISA, an ideal TI system should offer an experience similar to a face-to-face room interaction, where the users expect the in-sync audio and video information with perfect intelligibility and minimal latency. But in reality, the imperfections of wireline and wireless networks may only offer a downgraded user experience [14]: an out-of-sync audio-visual rendering, an imperfect audio intelligibility and video motion smoothness, and a degraded interactivity. In Section 3, we will identify both objective and subjective metrics to capture the TISA quality in multiple dimensions.

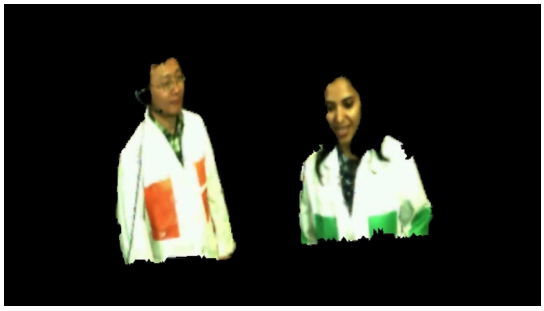
Depending on the shared interactive activities using the TI application, the media system can have heterogeneous demands on various quality dimensions. Thus, it is interesting to understand how the perceptual quality is impacted by the heterogeneity of TISA. We study two representative TISAs: conversation-oriented (CONV) tasks and collaborative (COLL) gaming activities.

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. 3(a)). Generally speaking, TI users in CONV pay attention to the audio intelligibility of the conversation more than the image quality.

In COLL (Fig. 3(b)), two distributed participants are playing the "rock-paper-scissor" game in the virtual environment. In this application, the visual timing mismatch is more important to human perceptions.

3. TISA QUALITY METRICS

The goal of this paper is to evaluate the effectiveness and limitation of CMOS and DMOS under the combined impacts of multiple quality dimensions on human perceptions



(a) CONV: conferencing scenario with social talk

(b) COLL: rock-paper-scissor collaborative gaming

Figure 3: Two TI applications evaluated in our user study.

in TISA. This can be achieved by conducting subjective user study and evaluating the TISA samples of different qualities. 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 paper, we will investigate both objective and subjective quality metrics for TISA evaluations. We will also present a survey of existing subjective quality assessment studies for media systems at the end of the section.

3.1 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 [19] 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 video *macroframe*[1] 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 video macroframe 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.

.

• Synchronization Quality

The audio and multiple multi-view video streams can experience different *end-to-end delays* (EED) between two distributed users. An EED includes the accumulated latencies incurred at the Internet and end systems.

We assume that the video macroframe 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 video macroframe is synchronously captured at the camera, and the time that it is displayed on the screen (Fig. 2). EED_A is used to denote the duration between the microphone and speaker for an audio frame (Fig. 2). Hence, x_S can be represented as:

$$x_S = \text{EED}_V - \text{EED}_A \quad (1)$$

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

[1] A video *macroframe* represents a set of multi-view frames belonging to different video streams, capturing the same physical object at the same time from different camera directions.

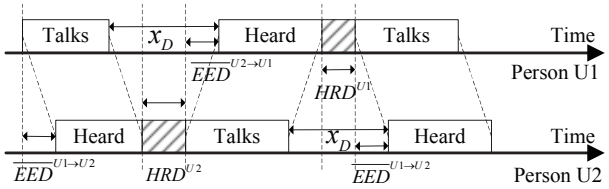


Figure 4: Interactivity in conversation-oriented activity.

• Interactivity

In the conversation-oriented activity, 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 [13] 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) [13]), and EED of the remote streams traveling back to the local site. Fig. 4 shows the concept. Mathematically, x_D that a local user experiences can be represented as:

$$x_D = \overline{\text{EED}}^{U1\to U2} + \text{HRD}^{U2} + \overline{\text{EED}}^{U2\to U1} \quad (2)$$

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 a collaborative activity 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 follow. The two roles can be swapped dur-

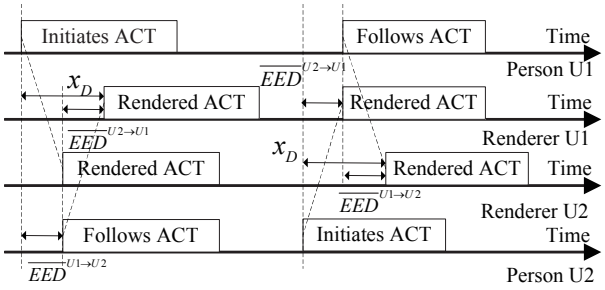


Figure 5: Interactivity in collaborative activity (ACT).

ing 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 his/her own rendering display (Fig. 5). 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 [17] 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 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 macroframe rate x_V, *milliseconds* (ms) for the respond delay x_D and the audio-visual sync skew x_S, and [1, 4.5] for the audio quality x_A. All the four metrics are user-observable. An example would be $\vec{x} = \{10,\ 3.0,\ 1100,\ 50\}$, representing a TISA setting with a video macroframe rate of 10 fps, an audio quality of 3.0, a response delay of 1100 ms, and an audio-visual sync skew of 50 ms.

3.2 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 voting scores using the following three metrics.

• Average Score: CMOS/DMOS

We compute the average of people voting scores for CMOS according to ITU-T P.910 [20]. Note that when one of the TISA samples in the user study comparison is the reference sample (denoted as $\vec{x}^*$, will be discussed in Section 4.2), we can use CMOS to approximate DMOS [15]. DMOS is defined as the perceptual quality impairment of a TISA sample from the reference $\vec{x}^*$. To simplify the descriptions for the rest of the paper, we use DMOS instead of CMOS, whenever $\vec{x}^*$ is one of the samples in the user study comparison. We use CMOS for all other cases, where $\vec{x}^*$ is not involved.

In this study, DMOS is always ≥ 0. A quality point with a smaller DMOS means a closer quality to $\vec{x}^*$, and hence, a better quality[2].

Previous subjective studies [13, 16, 17, 34, 35] present subjective findings using CMOS and DMOS for their efficiency and simplicity. But the major limitation is that both metrics compute the average scores, and thus, are unable to describe the potential diversity of user votes. A group of participants may output contradicting opinions[3] in a subjective comparison, in which a media system is evaluated under the quality tradeoffs (e.g., a comparison between a TISA sample with a better interactivity but a worse media signal quality, and another sample with a more satisfactory media signal quality but a poorer interactivity). Because participants pay different attention to heterogeneous quality dimensions, they can have different preferences when measuring the two TISA samples. Similar findings have been demonstrated in our previous VoIP studies [13, 34].

• Distribution of User Votes

To address this issue, we compute the distribution of user votes ($N_{>0},\ N_{=0},\ N_{<0}$), representing the percentage of voting score that is greater than, equal to, and less than 0. The following equation is satisfied:

$$N_{>0}\ +\ N_{=0}\ +\ N_{<0} = N_{\text{total}} \quad (6)$$

Here, N_{total} is the total number of votes. For example, we have 5 votes for score $2 \in C$, 4 votes for score $1 \in C$, 5 votes for score $0 \in C$, and 5 votes for score $-1 \in C$ out of 19 participants, then $(N_{>0},\ N_{=0},\ N_{<0}) = (9, 5, 5)$ and $N_{\text{total}} = 19$.

• Inconclusiveness

Oftentimes, we are unable to derive from ($N_{>0},\ N_{=0},\ N_{<0}$) and tell with confidence that the majority (say, more than 50%) of votes will agree that one media sample is better/worse than the other, or the qualities of two samples are about the same. For example, if $(N_{>0},\ N_{=0},\ N_{<0}) = (6,\ 7,\ 6)$, we are having an *inconclusive* situation. We call this subjective comparison *inconclusive* outcome. Here, we rely on the hypothesis test (proposed in our past study [13]) to identify whether the distribution ($N_{>0},\ N_{=0},\ N_{<0}$) is inconclusive.

The basic idea of [13] is that we compute the voting probabilities:

$$(p_{>0},\ p_{=0},\ p_{<0}) = (N_{>0}/N_{\text{total}},\ N_{=0}/N_{\text{total}},\ N_{<0}/N_{\text{total}}) \quad (7)$$

We then model ($p_{>0},\ p_{=0},\ p_{<0}$) using a multinomial distribution with 3 possible outcomes, assuming the independence of participants. We selectively combine two options within the 3 outcomes, and have an equivalent binomial distribution that represents the *for* or *against* probabilities of the opinion. We conduct hypothesis testing on whether $(p_i, \sum_{j\neq i} p_j)$ (p_i/p_j can be either $p_{>0}$, $p_{=0}$, or $p_{<0}$) is drawn randomly from a binomial distribution: binomial$(N_{\text{total}}, p \geq 0.5)$.

[2]We follow ITU-T G.1070 to define DMOS. Some other studies [3] may have a reciprocal definition by prescribing that a larger DMOS mean a better perceptual quality.

[3]Contradicting opinions mean that participants do not agree on which one is better within the comparison of two samples. Some participants give a positive comparison score, while others can output a negative score.

Table 2: A survey of subjective quality assessment. The types of studied media system includes conferencing application (Conf), video-on-demand (VOD) and TISA. VSQ: video signal quality, which includes the video frame rate (FR), the spatial resolution (RES) and the encoding quality (ENC). ASQ: audio signal quality. Three subjective methods are classified: ACR, CCR and DCR. Y: representing the corresponding quality dimension is studied.

	Type	Studied Media		VSQ			ASQ	Interactivity	Sync	Method	Comments
		Video	Audio	FR	RES	ENC					
[17]	Conf	2D	Y	Y	Y	Y	Y	One-way delay	Y	DCR	Independent
[16]	Conf		Y				Y	One-way delay		DCR	Independent
[13]	Conf		Y				Y	Response delay		CCR	Dependent
[12]	Conf		Y				Y	Response delay		CCR	Dependent
[34]	Conf		Y				Y	Response delay		CCR	Dependent
[30]	VOD	2D		Y						ACR	
[37]	VOD	2D	Y						Y	ACR	
[9]	VOD	2D	Y	Y					Y	ACR	Dependent
[7]	VOD	2D		Y	Y	Y				CCR	Dependent
[36]	VOD	2D				Y				DCR	
[29]	VOD	2D				Y				ACR/DCR	
[47]	VOD	2D		Y						DCR	
[48]	VOD	2D		Y						DCR	
[27]	VOD	2D		Y		Y				ACR	Independent
[24]	Conf	2D		Y	Y					ACR	Dependent
[50]	VOD	2D		Y	Y	Y				ACR	Dependent
[26]	VOD	2D		Y						ACR	
[32]	VOD	2D		Y						ACR	
[40]	Conf/VOD	2D	Y			Y	Y			DCR	Dependent
[41]	VOD	2D	Y	Y					Y	ACR	Dependent
[6]	VOD	2D	Y	Y			Y			ACR	Dependent
[10]	Conf	2D	Y	Y			Y			ACR	Dependent
[21]	Conf	2D	Y	Y			Y	One-way delay		ACR	Dependent
[25]	Conf	2D	Y	Y			Y			ACR	Dependent
[31]	VOD	2D							Y	ACR	
[42]	VOD	2D	Y	Y		Y	Y			ACR	Independent
[33]	Conf		Y					One-way delay		ACR	
[8]	Conf		Y					One-way delay		ACR	
[23]	Conf		Y					One-way delay		ACR	
[18]	Conf		Y					One-way delay		ACR	
[28]	Conf	2D	Y						Y	ACR	
[46]	TISA	3D								ACR	
[44]	TISA	3D						One-way delay		ACR	Independent
[43]	TISA	3D	Y				Y			ACR	Independent
[45]	TISA	3D	Y				Y			CCR	Independent

Derivation details can be found in [13]. Here, we directly reach the conclusion: a comparison is inconclusive if no number N_i in ($N_{>0}$, $N_{=0}$, $N_{<0}$) (i.e., N_i can be either $N_{>0}$, $N_{=0}$ or $N_{<0}$) satisfies

$$\sum_{g=0}^{N_i} \left(\begin{array}{c} N_{\text{total}} \\ g \end{array} \right) \cdot 0.5^g \cdot 0.5^{N_{\text{total}}-g} \geq \alpha \tag{8}$$

where α is the significance level. We assume N_{th} is the minimal number of N_i satisfying the above equation. For 90% (resp. 80%, 70%) significance, every number in ($N_{>0}$, $N_{=0}$, $N_{<0}$) should be less than $N_{\text{th}} = 12$ (resp. 11, 10) out of $N_{\text{total}} = 19$ at an inconclusive comparison.

3.3 Related Subjective Studies

The purpose of the subjective quality assessment is to find a mapping from the objective metrics to the subjective opinions. We have conducted a survey for existing studies on the subjective evaluations of media systems. The work can be broadly divided into two categories. In the first category, the papers (e.g., [29, 33, 37]) investigate the mapping of a single quality dimension to user experience, by assuming all other quality dimensions as optimal values. In the second category, the studies identify the cross/combined effects of multiple quality dimensions on the overall human perceptions. Representative work are [12, 13, 16, 17, 34, 44].

Table 2 presents a list of existing subjective studies, including the type of applications, studied media, quality dimensions, and their subjective rating methods. The type of media applications can be TISA, traditional video/audio conferencing or on-demand videos. The studied media can either be video (2D or 3D), or audio. The identified quality dimensions are the video signal quality (including the frame rate, the spatial resolution, and the encoding quality), the audio signal quality, the interactivity (either one-way delay [16, 17, 44] or VoIP conversational response delay [12, 13, 34] is studied), and the synchronization quality. The subjective rating methods can be either ACR, CCR or DCR. The comments column specifies whether the cross impacts (dependency) of the multiple quality dimensions are identified in the study.

Among these work, perhaps ITU-T G.107 [16] and G.1070

Table 3: Discretization of quality metrics in $\vec{x}$. HRD = 800 ms is used in computing x_D in CONV (Section 4.2). 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	950, 1150, 1350, 1550, 1750, 1950, 2150, 2350, 2550
x_D (COLL)	ms	120, 180, 240, 300, 360, 420, 480, 540, 600

[17] are those that are closest to our study because both standards investigate the cross/combined impact of multiple quality dimensions similar to what we have identified in Section 3.1. But they are only for VoIP and 2D video conferencing. None of these work is able to describe the combined effect of multi-view video macroframe rate, audio quality, response delay and audio-visual synchronization skew in a TI setting.

In Section 4, we will discuss the findings from the selected surveyed work in Table 2 (ITU-T G.107 and G.1070 particularly), and compare them to our subjective results.

4. 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 applications.

4.1 Methodology

Our user study investigates both TI applications (i.e., CONV and COLL) discussed in Section 2 in order to evaluate the effectiveness and limitations of subjective metric CMOS and DMOS in both activities. To find the mappings from the objective quality metrics (Section 3.1) to subjective space (Section 3.2), 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 infinite number of options for $\vec{x}$. In this study, we discretize each metric within $\vec{x}$ (Table 3) according to the characteristics of real media traffic in the Internet. To further reduce the TISA sample space, only the bolded numbers in the table are investigated when evaluating the cross impacts of multiple quality dimensions.

In our user study, we ask the participants to compare TISA samples of the same application in each test. We employ CCR rating scale as discussed in Section 3.2. We divide our tests into two categories, and process the user subjective feedback accordingly.

Category I: we only consider the impact of a single quality dimension in $\vec{x}$ by keeping values in other dimensions fixed. The diversity of user votes is expected to be small, and mutually contradicting votes are unlikely. So we focus on presenting the CMOS or DMOS, and show the effectiveness of both scores.

Category II: we compare TISA samples with quality tradeoffs, and show the diversity of user opinions. In this case, we will discuss both CMOS and the distribution of the user votes. We will show the limitation of CMOS by identifying the inconclusiveness of the subjective comparisons in the study.

4.2 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[4]) 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[5]), 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 [13], and use a HRD of 800 ms, and an average talkspurt duration of 2732 ms in our simulation [13]. In COLL, the average duration of talkspurts is 856 ms. In this study, the reference sample with the best-possible 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.

4.3 Setup of User Study

19 participants (average age: 26) are involved in our user study. They are required to sit 1.5 meter apart from a 61-inch NEC screen (system 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 more than 100 TISA samples (with different configurations $\vec{x}$) and their comparisons 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.

5. EVALUATION RESULTS

In this section, we present the findings of both test categories (Section 4.1) from our user study. We focus on the effectiveness (conclusiveness) of CMOS or DMOS in Category I, while addressing the limitation (inconclusiveness) of the CMOS metric in Category II. We will show the two TI application CONV and COLL have heterogeneous impacts on human perceptions. Based on the subjective findings, we will then discuss their implications to the system design, and conclude there is a need for a new subjective metric to describe the inconclusive comparisons.

[4] `http://xvidcap.sourceforge.net`

[5] `http://software.muzychenko.net/eng/vac.htm`

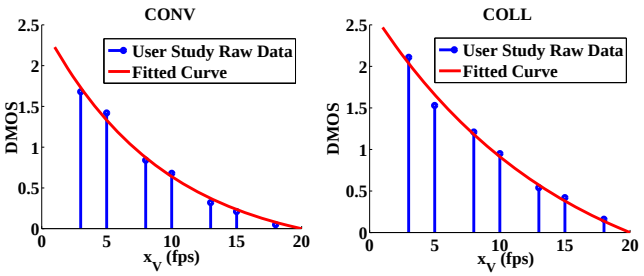


Figure 6: DMOS results for comparing x_V-degraded samples (with different x_V but optimal x_A, x_D and x_S) to the optimal reference $\vec{x}^*$.

5.1 Category I: 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:

$$\text{DMOS}(x_A) = 4.5 - x_A \tag{9}$$

Here 4.5 is the maximal value of x_A. This equation demonstrates the conclusiveness of DMOS under the impact of single quality dimension x_A. Note that, with the same audio frame loss rate, x_A can still vary when different audio codecs are employed.

• Video Signal Quality

Previous work. ITU-T 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, [30] 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 render on a 2D display, this mathematical model lays a theoretical foundation for our study. DMOS has been proven effective under the impact of single quality dimension x_V in both G.1070 and [30].

Our results. Fig. 6 shows the DMOS results comparing $\vec{x}^*$ to the samples with different degraded x_V while keeping other quality dimensions optimal. We use the exponential model in [30] to find the fitting curve describing the mapping from x_V to the corresponding DMOS:

$$\text{DMOS}(x_V) = Q - Q \times \frac{1 - e^{-c \times x_V / x_V^{\max}}}{1 - e^{-c}} \tag{10}$$

In this equation, c is the slope of the curve, which describes the impact of x_V changes on the DMOS. A smaller c will introduce a larger degradation to DMOS 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 video macroframe 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 4. 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. Again, our study shows the DMOS conclusiveness by computing the distribution of user votes (details not presented) using Eqn. 8 in Section 3.2.

Table 4: Fitting results for Eqn. 10.

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

5.2 Category I: Synchronization Impairment

Previous work. As discussed in Section 3, 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 on-demand videos, Steinmetz and Nahrstedt [38] 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. [9] evaluates the synchronization in the mobile terminal with a maximum of QCIF image size. The paper shows that the synchronization threshold is affected by the video frame rate.

For video conferencing, ITU-T 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.

Our results. In Fig. 7 (1) and (4), we carry on experiments to evaluate the lip skew impairment at the optimal x_V, x_A and x_D, when compared to $\vec{x}^*$. We show the DMOS results at different x_S. In Fig. 7 (2-3) and (5-6), we evaluate the impact of x_V and x_A on the synchronization quality. We show the CMOS results for $x_S = \pm 150$ ms with different x_V and x_A values, compared to the samples of $x_S = 0$ with the same media signal quality. We have four observations.

First, the variance of user voting scores is small (distributions of user votes are not presented), and we show both DMOS and CMOS conclusiveness in evaluating the impact of single quality dimension x_s in our study.

Second, our limited study reflects that the heterogeneous TI applications 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.

Third, contrary to the findings of on-demand videos in [9], 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. 7 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 Steinmetz and Nahrstedt's results in [38].

Fourth, Fig. 7 (2-3) and (5-6) show that both x_V and x_A

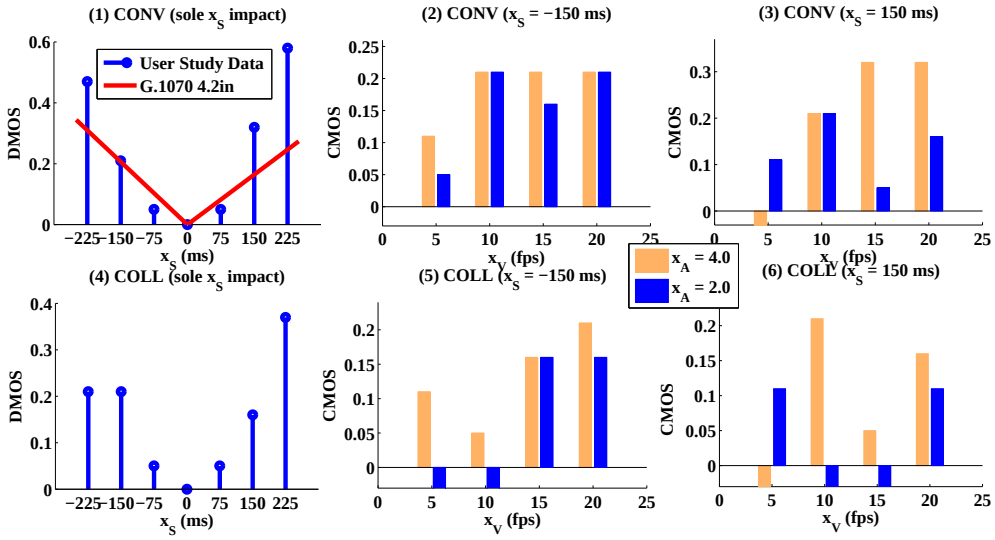


Figure 7: (1) and (4) show the DMOS of a sample with a degraded x_S (but optimal x_V, x_A and x_D) compared to the optimal reference $\vec{x}^*$. (2-3) and (5-6) show the comparison results CMOS of two samples with different x_S, but same x_V, x_A, and $x_D = 800$ ms for CONV and 0 ms for COLL. The first sample in the comparison is $x_S = 0$ and the second is $x_S = \pm 150$ ms.

do impact the synchronization quality. We find that x_S is less noticeable at a smaller $x_V = 5$ fps than a better x_V. This is because the motion jerkiness becomes the dominant factor degrading the human perceptions, and thus, a lip skew can be difficult to tell. We also observe that when the audio signal is degraded (i.e., $x_A = 2.0$), the CMOS results in the figures do not follow the same distribution as the cases with perfect audio quality $x_A = 4.0$, but instead, exhibit some randomness. This shows that the poor audio intelligibility also creates a hard time for users to differentiate a lip skew, and that an incomplete utterance can cause misperception on the synchronization quality.

5.3 Category I: Interactivity

Previous work. Previous studies on the interactivity (delay impairment) can be divided into two categories based on their applications.

For packet-switched telephone network, Kiatawaki and Itoh [23] study the pure delay effect on speech quality, and their results show that one-way delays are detectable and can influence listeners' subjective assessment. Richards [33] and Brady [8] conclude from their subjective evaluations that longer delays can decrease the user satisfaction rate.

For Internet conference, ITU-T G.114 [18] prescribes that a one-way delay of less than 150 ms is desirable and a delay of more than 400 ms is unacceptable in a two-party VoIP. ITU-T G.107 uses a complex sixth-order model to describe the VoIP delay impairment. [13] and [34] study the impact of audio signal quality on the interactivity, and their results show that the combined effect on human perceptions cannot be described as a linear form (a result contrary to G.107). On the other hand, ITU-T G.1070 employs a linear function to present the delay impacts in the 2D 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 DMOS by referencing $\vec{x}^*$ in Fig. 8 (1) and (3). The G.107 and G.1070 findings are

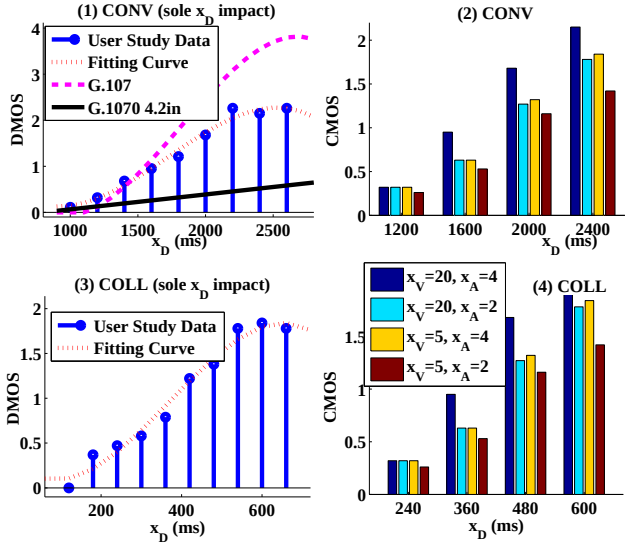


Figure 8: (1) and (3) show the DMOS under the impact of x_D of a sample (with optimal x_V, x_A and x_S) compared to the optimal reference $\vec{x}^*$. G.107 and G.1070 delay curves are also drawn in (1). (2) and (4) show the comparison results CMOS of two samples with different x_D, but same x_V, x_A, and $x_S = 0$. The first sample in the comparison is with optimal x_D and the second is with degraded x_D.

also plotted in CONV as a comparison. Because both studies only consider the impairment of one-way delay, we assume that the bi-directional EED is symmetric in computing x_D, meaning that $\overline{\text{EED}}^{U1\to U2} = \overline{\text{EED}}^{U2\to U1}$. In the second set, we study the effects of the media signal quality on the x_D perception. Fig. 8 (2) and (4) show the resulting CMOS. There are several observations from the figures.

First, we follow [39] and use a third-order polynomial model to describe the DMOS degradations due to x_D. The results are shown in Fig. 8 (1) and (3).

$$\text{DMOS}(x_D) = a_0 + a_1 \cdot x_D + a2 \cdot x_D^2 + a_3 \cdot x_D^3 \qquad (11)$$

Table 5 presents the fitting results both activities as well as

Table 5: Fitting results for Eqn. 11.

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

the corresponding MSE. Generally for CONV, $x_D < 1200$ ms is desired (DMOS < 0.5) and $x_D > 2000$ is bad (DMOS > 1.5). For COLL, $x_D < 200$ ms is desired (DMOS < 0.5) and $x_D > 400$ is bad (DMOS > 1.5). Hence, the COLL application requires a higher demand for interactivity than CONV. This is because people in the COLL attach more importance to the visual timing mismatch in the collaboration.

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 2D video conference (G.1070) or TISA 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. 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. 8 (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.

In all the figures, our results show the effectiveness of both CMOS and DMOS in representing the perceptual degradations under the single-dimensional impact of x_D.

5.4 Category II

In this study, we have done substantial subjective comparisons over the TISA samples with multi-dimensional quality tradeoffs. We focus on the tradeoff between x_V and x_D, which is most commonly seen in the real TI system over the Internet. The reason is that the data rate for multi-view videos is very high. By increasing x_V at a fixed bandwidth availability, the resulting higher visual data demand can generally introduce additional transmission (timing) overhead over the Internet and end systems, which in turn degrades (increases) x_D. Fig. 9 and 10 show some of the selected representative results, where both $x_A = 4.0$ and $x_S = 0$ are fixed. We compute the distribution of user votes ($N_{>0}$, $N_{=0}$, $N_{<0}$), as well as the corresponding CMOS. Several observations are to be noted.

First, we find a huge diversity of user votes in some of the comparisons (e.g. Fig. 9(6) in CONV and Fig. 10(4) in COLL). The multi-dimensional quality tradeoffs contribute to this diversity. Generally, if the perceptual degradation in one quality dimension of a TISA sample is not overshadowed by the enhancement of another dimension, users can output contradicting voting scores, because they can attach heterogeneous importance to different quality attributes based on their individual interests. For example, in Fig. 9(6), 9 out of 19 participants prefer a better interactivity, so they think the first sample is better. Another 9 participants like a smoother body motion in the video, so they argue for the second sample. As two quality points are moving apart on the tradeoff curve, and one one dimension is gradually improving as another dimension is worsening comparably, the likelihood of outputting contradicting opinions is increasing (e.g., Fig. 9(6) shows a greater voting diversity than Fig. 9(3) and (5)).

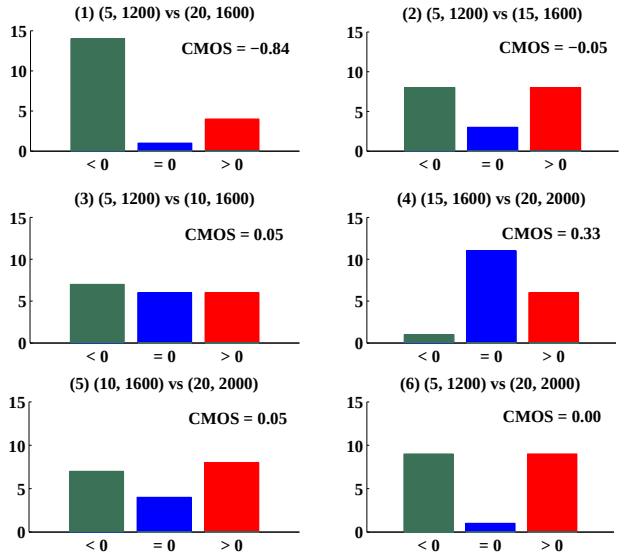


Figure 9: Comparison results ($N_{>0}$, $N_{=0}$, $N_{<0}$) for samples in CONV. All samples with $x_A = 4.0$ and $x_S = 0$. The caption format is (x_V^1, x_D^1) vs (x_V^2, x_D^2), where the four numbers are the x_V and x_D of the first and second samples.

Table 6: 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, the interpretation of the average score CMOS may lack the statistical significance at a large variance of user votes. For example, a CMOS = 0 in Fig. 9(6) cannot tell with confidence whether a sample within a comparison is of the same quality with the other sample (actually the qualities of the two samples in Fig. 9(6) are completely different).

To evaluate the inconclusiveness, we use $\alpha = 70\%$ significance in Eqn. 8. Because $N_{\text{total}} = 19$, this returns $N_{\text{th}} = 10$. Hence, except Fig. 9(1,4) and Fig. 10(1,5,6), all other comparisons are inconclusive.

5.5 Implications to TI System Design

TISA heterogeneity. A good system design 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 the above discussions, we qualitatively conclude the perceptual importance for the two TI applications in Table 6. 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.

CMOS/DMOS drawbacks. Previous studies on VoIP

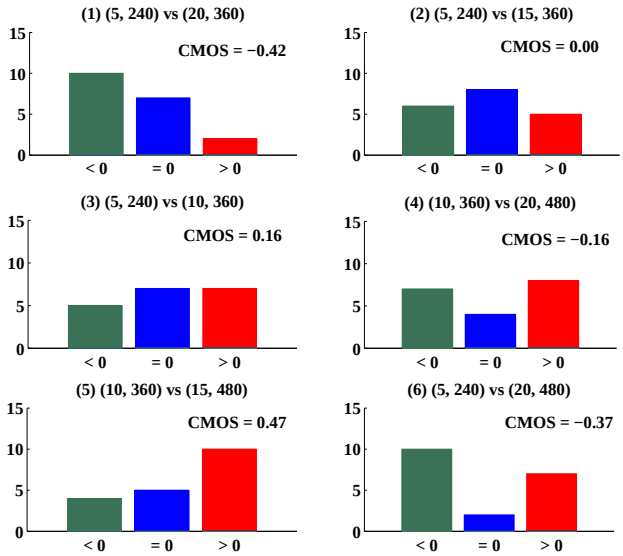


Figure 10: Comparison results ($N_{>0}$, $N_{=0}$, $N_{<0}$) for samples in COLL. All samples with $x_A = 4.0$ and $x_S = 0$. The caption format is (x_V^1, x_D^1) vs (x_V^2, x_D^2), where the four numbers are the x_V and x_D of the first and second samples.

[39] or video conferencing [22] usually propose adaptation algorithms that are based on the DMOS ordering in G.107 or G.1070. Here, we use our user study results to argue that the quality closed forms derived in both standards are only suitable for the subjective quality assessment of media samples, and the resulting DMOS ordering are not good for system adaptations.

First, multiple quality points, which are distant in the multidimensional Euclidean space, can lead to same or similar DMOS when they are compared to the optimal reference $\vec{x}^*$. For example in CONV, $\vec{x}^1 = \{12, 4.0, 0, 0\}$ in Fig. 6(1), $\vec{x}^2 = \{20, 4.0, 0, -225\}$ in Fig. 7(1), and $\vec{x}^3 = \{20, 4.0, 1300, 0\}$ in Fig. 8(1) all lead to DMOS of around 0.5. If we achieve adaptation based on the DMOS ordering, the system may jump among these operating points which can cause *flicker effects* (i.e., the perceptible change of media qualities). These flickers should be minimized, which would otherwise downgrade human perceptions [49].

Second, as we have discussed in Category II tests, the diversity of user votes under the tradeoffs of multiple quality dimensions, make it difficult to interpret the obtained CMOS scores with statistical significance. Because the comparison results can be inconclusive, a total ordering of multiple quality points may not be accessed, and only a partial order can be decided. Similar conclusions have also been reached in our previous VoIP studies [12, 13, 35].

A need for new subjective metric! Given the inconclusiveness of CMOS at the diversity of user voting scores, we conclude that there is a need to propose a new subjective metric to interpret the subjective results under multi-dimensional quality tradeoffs. Developing methodologies to propose such a metric is beyond the scope of this paper.

6. CONCLUSION

In this paper, we propose a methodology to evaluate the effectiveness and limitation of CMOS and DMOS metrics in two TI activities. We show that while both metrics are effective in presenting diverse human perceptual degradations under the impact of single-dimensional quality metric in heterogeneous TISAs, CMOS can lack statistical significance in expressing the inconclusive comparison results under multi-dimensional quality tradeoffs. Hence, we conclude there is a need to propose a new subjective metric to address the inconclusive issue.

7. REFERENCES

[1] Cisco Telepresence. Cisco Corporation. http://www.cisco.com.

[2] Skype. http://www.skype.com.

[3] Video Clarity Corporation, Understanding the JND scale white paper. http://www.videoclarity.com.

[4] Vidyo Telepresence. Vidyo corporation. http://www.vidyo.com.

[5] P. Bajcsy, K. McHenry, H.-J. Na, R. Malik, and et al. Immersive environments for rehabilitation activities. In *Proc. of ACM Int'l Conference on Multimedia*, pages 829–832, 2009.

[6] J. Beerends, C. De, and E. Frank. The influence of video quality on perceived audio quality and vice versa. *Journal of Audio Engineering Society*, 47(5):355–362, May 1999.

[7] A. D. Bimbo, S.-F. Chang, and A. Smeulders. Subjective evaluation of scalable video coding for content distribution. In *Proceedings of the 18th International Conference on Multimedia*, 2010.

[8] P. T. Brady. Effects of transmission delay on conversational behaviour on echo-free telephone circuits. *Bell System Technical Journal*, 50(1):115–134, Jan. 1971.

[9] I. Curcio and M. Lundan. Human perception of lip synchronization in mobile environment. In *Proc. of IEEE Int'l Symposium on a World of Wireless, Mobile and Multimedia Networks*, 2007.

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

[11] M. Forte and G. Kurillo. Cyberarchaeology - experimenting with teleimmersive archaeology. In *Proc. of Int'l Conference on Virtual Systems and Multimedia*, 2010.

[12] Z. Huang. The design of a multi-party VoIP conferencing system. M.S. thesis, University of Illinois at Urbana-Champaign, Urbana, IL, 2009.

[13] 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 Int'l Conference on Multimedia and Expo*, pages 493–496, July 2008.

[14] Z. Huang, W. Wu, K. Nahrstedt, A. Arefin, and R. Rivas. Tsync: A new synchronization framework for multi-site 3d tele-immersion. In *Proc. ACM Workshop on Network and Operating Systems Support for Digital Audio and Video*, June 2010.

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

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

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

[18] ITU-G.114. One-way transmission time, 2003.

[19] 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.

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

[21] M. Jackson, A. H. Anderson, R. Mcewan, and J. Mullin. Impact of video frame rate on communicative behaviour in two and four party groups. In *Proc. of ACM Conference on Computer Supported Cooperative Work*, pages 11–20, 2000.

[22] A. Khan, L. Sun, E. Jammeh, and E. Ifeachor. Quality of experience-driven adaptation scheme for video applications over wireless networks. *IET Journal of Communications*, 4(11):1337–1347, 2010.

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

[24] J. Kies, R. Williges, and M. Rosson. Evaluating desktop video conferencing for distance learning. *Elsevier Computers and Education*, 28(2).

[25] H. Knoche and H. D. Meer. Compensating for low frame rates. In *In CHI extended abstracts on Human factors in computing systems*, pages 1553–1556, 2005.

[26] Z. Lu, W. Lin, C. Boon, S. Kato, and et al. Measuring the negative impact of frame dropping on perceptual visual quality. *Human Vision and Electronic Imaging X, SPIE*, 5666:554–562, 2005.

[27] M. Masry and S. S. Hemami. An analysis of subjective quality in low bit rate video. In *Proc. of IEEE Int'l Conference on Image Processing*, 2001.

[28] L. Mued, B. Lines, S. Furnell, and P. Reynolds. The effects of lip synchronization in IP conferencing. In *Proc. IEE Int'l Conference on Visual Information Engineering*, pages 210–213, 2002.

[29] T. Oelbaum, H. Schwarz, M. Wien, and T. Wiegand. Subjective performance evaluation of the SVC extension of H.264/AVC. In *Proc. of IEEE Int'l Conference on Image Processing*, 2008.

[30] 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 Int'l Confernce on Image Processing*, pages 689–692, 2008.

[31] S. Par and A. Kohlrausch. Sensitivity to auditory-visual asynchrony and to jitter in auditory-visual timing. *Human Vision and Electronic Imaging V*, 3959:234–242, June 2000.

[32] R. Pastrana-Vidal, C. Jean, C. Colomes, and H. Cherifi. Sporadic frame dropping impact on quality perception. *Human Vision and Electronic Imaging IX, SPIE*, 5292:182–193, 2004.

[33] D. L. Richards. *Telecommunication by Speech.* London, UK: Butterworths, 1973.

[34] B. Sat and B. W. Wah. Playout scheduling and loss-concealments in VoIP for optimizing conversational voice communication quality. In *Proceedings of ACM Int'l Conference on Multimedia*, pages 137–146, Sept. 2007.

[35] B. Sat and B. W. Wah. Statistical scheduling of offline comparative subjective evaluations for real-time multimedia. *IEEE Transaction on Multimedia*, 11(6):1114–1130, Oct. 2009.

[36] K. Seshadrinathan, R. Soundararajan, A. Bovik, and L. K. Cormack. Study of subjective and objective quality assessment of video. *IEEE Transaction on Multimedia*, 19(6):1427–1441, June 2010.

[37] R. Steinmetz. Human perception of jitter and media synchronation. *IEEE Journal on Selected Areas in Communications*, 14(1):61–72, 1996.

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

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

[40] A. Vahedian, M. R. Frater, and J. F. Arnold. Impact of audio on subjective assessment of video quality in videoconferencing applications. *IEEE Transaction on Circuits System and Video Technology*, 11(9):1059–1062, 2001.

[41] 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.

[42] S. Winkler and C. Faller. Perceived audiovisual quality of low-bitrate multimedia content. *IEEE Transaction on Multimedia*, 8(5):973–980, 2006.

[43] 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 Int'l Symposium on Multimedia*, Dec. 2010.

[44] W. Wu, A. Arefin, R. Rivas, K. Nahrstedt, R. Sheppard, and Z. Yang. Quality of experience in distributed interactive multimedia environments: Toward a theoretical framework. In *Proceedings of ACM International Conference on Multimedia*, 2009.

[45] W. Wu and et al. Color-plus-depth level-of-detail evaluation metric for 3d teleimmersive video. In *Proc. of ACM Int'l Confernce on Multimedia*, 2011.

[46] W. Wu, Z. Yang, and K. Nahrstedt. A study of visual context representation and control for remote sport learning tasks. In *Proc. of World Conference on Educational Multimedia, Hypermedia and Telecommunications*, 2008.

[47] S.-F. C. Y. Wang and A. C. Lou. Subjective preference of spatio-temporal rate in video adaptation using multi-dimensional scalable coding. In *Proc. of IEEE Int'l Conference on Multimedia and Expo*, 2004.

[48] M.-M. Yadavalli, G. and S. Hemami. Frame rate preferences in low bit rate video. In *Proc. of IEEE Int'l Conference on Image Processing*, 2003.

[49] A. Zaccaria and M. Bitterman. The effect of fluorescent flicker on visual efficiency. *Journal of Applied Psychology*, 36(6):413–416, Dec. 1952.

[50] G. Zhai, J. Cai, W. Lin, X. Yang, W. Zhang, and M. Etoh. Cross-dimensional perceptual quality assessment for low bit-rate videos. *IEEE Transaction on Multimedia*, 10(7):1316–1324, Nov. 2008.