**Framework of Delay Perception in 3DTI**

一、Defination：

Noticeability of delay：特定任务、特定延迟下，用户对延迟的感知程度。

Disruptiveness of delay：特定任务、特定延迟下，用户对延迟的厌烦程度。

[NOTE] Noticeability and disruptiveness of delay are two important factors that measured by most studies \cite{wu2009quality, schmitt2014influence, geerts2011we, schmitt2014asymmetric}. When a service is already within noticeable delay, we can appropriately increase the delay to have more room for smoothing \cite{xu2013exploiting}. On the other hand, when network resources are not sufficient, the acceptable delay become useful as a reference.

Quality：特定任务、特定延迟下，用户对传输质量的总体评价。

以上三个量是很多相关用户实验中要测量的量。

从上面两个量来看，我们当然很希望像telephone一样得到一个确切的值，作为延迟的参考，比如noticeable delay和tolerable delay。然而

[55] points out that delay perception is also different for the variety of users. So the concept below should take user percentage into account. 除了user以外，context等因素也会影响用户的评分。因此，如果我们需要给出一个确切的值作为延迟的参考的话，应该基于statistic的分析，要在数学统计上显著才行。

Noticeable delay：不能被大多数人感知到的延迟。我们借用[Exploiting Just-Noticeable Difference of Delays for Improving Quality of Experience in Video Conferencing]中对JND的定义来定义我们的noticeable delay.

JND describes a boundary beyond which the difference in audiovisual quality between the original and the new operating points will be statistically perceptible by humans.

Following Sat and Wah’s definition on JND [32], let p0 be the faction of subjects who correctly identify which of the two sessions has a longer MED. We define the 75% JND of delay as follows. [32] = [Statistical scheduling of offline  comparative subjective evaluations for real-time multimedia.] The definition is the same as the JND of delay when MED = 0ms.

[Statistical Scheduling of Offline Comparative Subjective Evaluations for Real-Time Multimedia]

Definition 1. Just Noticeable Difference (JND) of: When comparing a fixed and a variable on an operating curve, is the value for which 50% of the subjects perceive a difference in their quality.

Tolerable delay：大多数人能够接受的延迟，不影响体验。50% people feel untolerable.

二、Synchronization requirements levels：

我们认为，决定delay perception的主要因素是interaction中线索的强度。典型的线索有，同步性动作、conversation中的对话间距和overlap、对方的肢体动作和表情变化等等。

通过对过去工作的总结，我们根据线索划分task的延迟等级，分为forced synchronization、conversational cues、visual cues和lacking cues四个层级，他们分别需要大约50ms、150ms、400ms和若干秒的延迟要求。

如果考虑noticeable delay，那么一个task的延迟等级应该有其组成部分中要求最高的部分决定，比如允许聊天的下象棋这个task，既包含conversation，又包含visual cues，则它的noticeable delay应该是第二级conversation级的。

如果考虑tolerable delay，那么一个task的延迟等级应该由其主要组成部分决定，仍然以下象棋为例子，如果在下象棋的时候双方棋手的交流较少，虽然他们有语言交流，并能够从中敏锐得感受到延迟，但是task的主要组成部分是要根据对手的下法来做出反应，因此它的tolerable delay应该是第三级别visual cues级的，能在500ms左右的延迟都可以接受。

如果一个task同时满足多级的条件，以最强的一个条件为主。比如，允许聊天的下象棋这个应用即包含conversation，又包含视觉线索，则它属于第二个级，conversation级。

1、强制性同步动作15ms~50ms

定义：强制性同步动作指的是交互双方需要完成一些必须同一时间发生的动作，以保证交互的公平性、良好的效果等等。

分析：

在这种情况下，用户对远端数据延迟的感知能力应该几乎接近对本地延迟的感知能力。

我们知道，用户在3D游戏中，用户对自己肢体动作的延迟十分敏感，60FPS基本能满足大多数用户的需求，也就是15ms左右的延迟，因此我们推断，在此情况下，用户对延迟的要求的下限应该在15ms左右。而根据先前的工作，有论文指出同时count down的延迟要求在50ms，所以我们设定其上限在50ms左右。

As explained in [30], 100 ms is a well-known upper limit for users to feel that a system is reacting instantaneously. [30] = [Response times: the three important limits]

Also, in [14] we found citations to several studies defining a delivering delay threshold around 150–200 ms to keep an enjoyable shared experience in networked multi- player games. [14] = [The brave new world of multiplayer online games: synchronization issues with smart solution.]

例子：

（1）猜拳

猜拳是时序影响公平性的游戏的一个典型的例子。[Inter-destination multimedia synchronization: schemes, use cases and standardization

]曾将公平性作为延迟的一个重要的评判标准。在一个50ms端到端延迟的系统中，如果两名用户在玩猜拳的游戏，则至少有一方会感觉对方慢出了至少50ms，从而可能抱怨游戏不公平。

（2）合奏

合奏指的是两端的用户使用两个不同的乐器或者一个“shared”的乐器同时演奏，后者举例来说是钢琴的四手联弹。想象两个身处异地的钢琴家，需要为一场演出磨合一首四手联弹的曲子，只要他们各自拥有一台类似的钢琴，就可以利用3DTI技术将这两台钢琴fuses成一台虚拟的钢琴。在这种情况下，系统的延迟要求就会非常的高，因为合奏对同步的要求是很高的，受过专业训练的音乐家能够判断出x毫秒的误差。

[The Effects of Latency on Ensemble Performance]

A coping strategy was discovered that allowed the performers to maintain a solid tempo up to 50-70 msec of delay. The strategy can be quickly summarized as a leader - follower relationship. 如果有指挥家，则存在50~70ms延迟的策略，也就是说，如果系统提供同步提示，那么50~70毫秒的延迟是可以忍受的。When delay is between 10-20 msec each way, it may be providing a stabilizing effect on the tempo. 10-20 msec of delay may be better for ensemble performance than 0 msec of delay。10~20ms的one-way端到端延迟是最优的，甚至不不存在延迟更优。

（3）数倒计时

数倒计时[Evaluation on Perceptual Audiovisual Delay using Average Talkspurts and Delay]的论文。如果两名远端的用户正在一起倒计时数数，每个用户都会期望听到对方的声音和自己的声音同时发出。如果系统的one way端到端延迟在50ms，那么总会有一个用户听到自己的声音和对方的声音差距在50毫秒以上，从而可能发现延迟的存在。

2、conversation notice:100ms standard 150ms tolerable 300ms

Turn talking model:

[A simplest systematics for the organization of turn-talking for conversation 1974]

In face-to- face situations we have learned to unconsciously use the timing of the small pauses in speech to manage a conversation and infer reactions from our interlocutors.

[Turn-taking in Human Communication – Origins and Implications for Language Processing 2016]

Concept: overlap

A turn is 2s in average. Switching of speakers is rapid, with a typical gap or offset of 200 ms.

Across languages, the modal response time (gaps between turns) is around 200 ms, the average duration of a single syllable.

Moreover, the language production system is notoriously slow – preparation before output begins takes 600 ms for a single word if primed [7,8], approximately 1000 ms if not [9], and around 1500 ms for a short clause [10]

Much of this latency is caused by the slow encoding of phonological forms and articulatory gestures (for a range of factors influencing latency of response see [11]).

Turn talking系统有时候会依赖于prediction，从而导致有可能出现overlap。不同语言预测的可能性不一样，我们说的是中文，以主谓宾为主，主宾谓为辅，这可能导致对话间隔更难预测，或者是更长。

However, the turn-taking system relies on prediction. Languages also differ in the size of the units or increments that must be planned in advance of beginning to speak – these are large if the language is verb-initial, but small if the language is subject-initial [40].

[Co-Constructing Non-Mutual Realities: Delay-Generated Trouble in Distributed Interaction]

We are particularly interested in how delay affects communicants’ experience of the conversation. Consider a hypothetical conversation between two remote collaborators. One person asks her collaborator a question, which he answers as soon as he hears; his response then travels back to her. She thus hears the response as coming after a gap determined by double the length of the delay inherent in the technology, a gap she can interpret in a number of ways. He, however, thinks he has answered promptly, but may now perceive a gap before receiving her acknowledgment. What is said and heard by users on each side of the communications link is thus different, but in such a way that neither side is aware of the discrepancy. To put it another way, communicants are not co-present to the communication in the same way. This has, as we shall see, far-reaching consequences.

Belows:

[18] = [Effects of transmission delay and access delay on the efficiency of verbal communication] (1960s)

[22] = [Characterizing, predicting, and measuring video-mediated communication: A conversational approach.] (1990s)

[8] = [Video teleconferences versus face-to-face meetings] (1982)

[24] = [Why do users like video?]

As demonstrated by Krauss & Bricker [18], small audio delays of 300ms can have detrimental effects on communication processes, and delays as large as 900ms can severely impact a pair’s ability to communicate. O’Connaill and Whittaker [22] found that audio delays between 410ms and 720ms led to reduced use of back-channels, fewer interruptions, and less overlapping speech. Cohen [8] described how a simultaneous 705ms delay of audio and video resulted in longer conversational turns and decreased overlap between utterances, and Tang & Isaacs [24] found that a one-way delay of 570ms severely disrupted turn- taking behaviors. In summary, the work on audio delay and its impact on collaborative performance tends to find that delays below 300ms pose little problem. Delays between 450ms and 700ms can severely impact communication and coordination processes. And delays greater than 700ms drastically impact communication, coordination, and overall task performance.

用户可以通过对方反应的时间推断出延迟的大小，从过去的论文看来，just noticeable延迟大约在100ms左右，150ms是行业的标准，而延迟在300ms左右会导致用户的厌烦。

在3DTI中，我们能够参考以上相关工作，但是不能照搬照抄。因为也有论文指出，更多更丰富的信道可能会使得用户对延迟的感知更加敏感，而另一方面，又会使得用户对延迟的容忍程度提高。因此，在我们这个框架中，将conversation作为一个重要的层级，但是它在3DTI中的具体表现则要通过进一步的实验来验证。[Video increases the perception of naturalness during remote interactions with latency] delay disrupts conversations more in audio-only communication than in audio- video communication. [Why do users like video] can also support this conclusion.]

[ASYMMETRIC DELAY IN VIDEO-MEDIATED GROUP DISCUSSIONS]

Our data shows that already one person with a high delay, affects the experience of the whole group as strong as the person with delay.

This paper support that 100ms is noticeable.

[Exploiting Just-Noticeable Difference of Delays for Improving Quality of Experience in Video Conferencing]

Definition: JND describes a boundary beyond which the difference in audiovisual quality between the original and the new operating points will be statistically perceptible by humans

We notice in the current Internet that increasing MED as well as reducing packet rate can help reduce the delay-aware loss rate in congested connections.

Starting from the operating point of an existing system, we increase its MED to within JND in order to have more room for smoothing network delay spikes as well as recovering lost packets, without incurring noticeable degradation in interactivity

Guidelines like ITU G.114 [One-way transmission time. Recommendation G.114, 1996.] focus on interactivity and suggest that MEDs less than 150 ms is desirable and more than 400 ms is unacceptable.

the delay requirement depends on the speed and turn frequency of a conversation.

A general survey on measuring QoE in audio [Quality of experience of VoIP service: A survey of assessment approaches and open issues] and video [Standardization activities in the ITU for a QoE assessment of IPTV] can be found in the literature.

[Evaluation on perceptual audiovisual delay using average talkspurts and delay]

support 100ms as noticeable delay.

3、visual cues:

visual cues指的是对方用户的肢体动作、面部表情和交互行为。

As summary in [The Impact of Delayed Visual Feedback on Collaborative Performance].

The theory?: **Situation Awareness Theory** holds that visual information helps pairs assess the current state of the task and plan future actions [10, 11]

**Grounding Theory** suggests that visual information can serve as an unambiguous source of evidence that allows conversational partners to generate efficient speech and more easily assess a level of understanding [4, 6, 7].

[10] = [Toward a theory of situation awareness in dynamic systems]

[11] = [*Situation awareness analysis and measurement*]

[4] = [*Using language*. Cambridge University Press, New York, NY, US.]

[6] = [Definite reference and mutual knowledge]

[7] = [Referring as a collaborative process]

[21] = [Turning away from talking heads: The use of video-as-data in neurosurgery]

[3] = [How conversation is shaped by visual and spoken evidence]

[19] = [Visual information as a conversational resource in collaborative physical tasks]

(1) Situation Awareness Theory:

In order for collaboration to be successful, group members need to maintain an ongoing awareness of one another’s activities, the status of relevant task objects, and the overall state of the collaborative tasks. [10, 11]

For example, Nardi and colleagues [21] describe how a scrub nurse on a surgical team uses visual information provided on an overhead monitor to assess the task state and anticipate the instruments a surgeon will need. If the nurse notices the surgeon has nicked an artery during a surgery, she can immediately begin preparation of the cauterization and suture materials and have them ready to present before the surgeon asks for them. However, if the visual information were delayed for some reason, such tight-coordination would not be possible and precious seconds could be lost.

In a similar fashion, but at an even finer temporal level, Gutwin and colleagues [17] describe how task coordination is supported by the availability of visual information during a tightly-coupled collaborative task in which pairs need to quickly move computational objects within a shared 2D workspace.

(2) Grounding Theory:

It states that successful communication relies on a foundation of mutual knowledge or common ground [4, 6, 7]. Speakers form utterances based on an expectation of what a listener is likely to know and then monitor whether the utterance was understood. In return, listeners have a responsibility to demonstrate their level of understanding. Shared visual information serves to support both the initial generation of utterances as well as to provide evidence of comprehension [3, 19].

[Understanding How Network Performance Affects User Experience of Remote Guidance]

[The Impact of Delayed Visual Feedback on Collaborative Performance]

Previous results have demonstrated that visual information helps collaborative pairs to understand the current state of their task, ground their conversations, and communicate efficiently

They notice where one another’s attention is focused, and they use cues available in the local visual context to help them coordinate both the language they use as well as the actions they engage in.

In Study 1 we found that the visual delay had no impact on task performance when it was less than 939ms. However, for the range between 939 and 1798ms the delay impacted both the Primary and Plaid puzzles equally.

[The Use of Visual Information in Shared Visual Spaces: Informing the Development of Virtual Co-Presence] We show that having the shared visual space helps collaborators understand the current state of their task and enables them to communicate and ground their conversations efficiently.

[K inect-taped Communication: Using Motion Sensing to Study Gesture Use and Similarity in Face-to-face and Computer-mediated Brainstorming] We usually use other cues to get feedback from our interlocutors, e.g. gestures, body language and facial expression [TO CHECK]

（1）基于模仿的telecollaboration：搭积木。

（2）能看见对方的回合制游戏：下象棋。

4、lacking cues:

lacking cues指的是没有以上线索的情况。

比如

（1）只能专注在棋盘上的下棋游戏

三、一些关于延迟的建议：

Noticeability and disruptiveness of delay are two important factors that measured by most studies [63, 55, 16, 54]. When a service is already within noticeable delay, we can appropriately increase the delay to have more room for smoothing [64]. On the other hand, when network resources are not sufficient, the acceptable delay become useful as a reference.

1、在within noticeable 延迟的情况下，可以适当增加buffer来处理

2、在远远超出tolerable delay的情况下，如果可以，我们认为应该牺牲传输质量来满足延迟的要求。