HUMAN-CENTRIC MULTI-LAYER SYNCHRONIZATION SCHEME WITH INTER-SENDER SYNCHRONIZATION SKEW CONTROL

Zixia Huang, Klara Nahrstedt

Ke Liang

Department of Computer Science
University of Illinois at Urbana-Champaign
{zhuang21, klara}@illinois.edu

Advanced Digital Sciences Center
University of Illinois at Urbana-Champaign
ke.liang@adsc.com.sg

ABSTRACT

Next-generation teleimmersive environments allow full-fledged collaboration of people in different geographical locations. The unique user experience poses new inter-sender synchronization requirement, which has not been investigated in past studies. The real-time nature of teleimmersive applications also introduce a tradeoff between synchronization quality and interactive quality during online synchronization adaptation. In this paper, we conduct subjective user study and demonstrate that inter-sender synchronization is important to human perception in teleimmersive activities. We also show that people exhibit contradicting opinions at the presence of synchronization and interaction tradeoff. Based on real subjective feedback, we propose a novel human-centric multi-layer synchronization scheme with inter-sender synchronization skew control for multi-site teleimmersion.

1. INTRODUCTION

A multi-site teleimmersive system is an application with multiple sensors of different media modalities (e.g., video, audio, haptics), capturing each site room environment from different angles at the same time. For example, a doctor and a patient in remote teleimmersive rooms conduct tele-physiotherapy session, and a medical student watches the *session* by receiving sensory data from the doctor and patient in synchronized fashion within certain latency. This imposes new hierarchical multimedia synchronization requirements, which involve four **layers** of synchronization interactions. (1) **Intra-stream synchronization** prescribes the synchronous presentation of media frames within each sensory stream at the receivers, according to their original captured timeline at the multimedia sensors. (2) **Intra-media synchronization** represents the synchronization of sensory streams from multiple media devices of the same media modality, e.g., multi-camera synchronization. (3) **Intra-bundle synchronization** describes the synchronization of multiple media modalities within a multi-modal multi-stream *bundle*, e.g., audio-visual lip synchronization. (4) **Intra-session synchronization** includes both *inter-receiver* and *inter-sender* synchronization

within a teleimmersive session.

There were numerous subjective research on intra-stream, intra-media, intra-bundle and inter-receiver synchronization evaluation [1, 2], but no work has ever investigated inter-sender synchronization, a new feature unique to teleimmersive activities. For example, if a remote student is watching the doctor diagnose the back of the patient in tele-physiotherapy, and the multimedia streams from remote doctor and patient experience heterogeneous Internet latency, their out-of-sync diagnostics activities will mislead the remote student and cause a poor learning experience. Inter-sender synchronization is very important to teleimmersive environments, and users demand a realistic involvement when observing in-pace virtual collaboration of remote parties. We will focus on inter-sender synchronization and investigate its human perceptual requirement in our study.

Multimedia synchronization adaptation, with inter-sender synchronization skew control, can usually be achieved by inserting a playout buffer before media presentation to compensate for Internet delay variation at each receiver site [1, 2], but the adaptation of playout buffer can create a tradeoff between synchronization quality and interactive quality. For example, a large buffer usually improves intra-stream synchronization quality, by allowing more media frames to arrive before scheduled presentation, and aligning them according to their original captured timestamps. But at the same time, interactive quality is degraded as the result of a prolonged buffer. Existing studies, e.g., ITU-T G.1070 [3], are unfortunately not able to capture this tradeoff.

In this paper, we investigate new subjective requirements on teleimmersive synchronization. We only consider inter-sender synchronization at the intra-session layer, and assume the inter-receiver synchronization problem has been studied in [1, 4, 5]. We conduct a user study and demonstrate that there is a perceptual need for inter-sender synchronization. We also show the tradeoff between synchronization and interactive quality, and study how the tradeoff will affect synchronization adaptation. Based on real perceptual feedback, we propose a new human-centric multi-layer synchronization scheme with inter-sender skew control for multi-site teleim-

mersive activities.

2. SYNCHRONIZATION FORMULATION

2.1. Continuous Multimedia Data Model

In a multi-site teleimmersive session, each site sends a bundle of time-correlated multi-modal media data. We denote the bundle of site x as u^x . The bundle u^x incorporates sensors with different media modalities. By letting m_i^x be the i -th media modality of site x , the media bundle u^x is represented as: $u^x = \{m_1^x, m_2^x, \dots\}$. For example, we use $i = 1$ or ‘V’ to represent video modality, $i = 2$ or ‘A’ for audio modality, $i = 3$ or ‘H’ for haptic modality, and etc. Multiple media sensors of the same modality (e.g., multi-camera array) can capture an object at the same time, but from different angles. Each sensor produces a sensory stream $s_{i,j}^x$ (j is the stream index), i.e., $m_i^x = \{s_{i,1}^x, s_{i,2}^x, \dots\}$. Each sensory stream $s_{i,j}^x$ constitutes a sequence of media frames from the same sensor.

2.2. Definition of Synchronization Quality

A **synchronization skew** in continuous multimedia is defined as the delay difference of two time-correlated *media objects* (e.g., sensory stream or media modality), traveling from the media sources to the current location. One of the objects is usually the **synchronization reference**. Because of the multi-layer synchronization hierarchy, a media object can be represented in multiple forms, meaning that the synchronization references must change accordingly at different layers. Here, we adopt a more reasonable approach by defining multiple skews for different layers respectively.

Intra-stream synchronization. In an interactive application like teleimmersion, the intra-stream synchronization reference is usually prescribed as the first frame at the start of the media presentation. All other media frames should follow this reference and be presented with the same *end-to-end latency* from their captured time at the media sensors to the presentation time at the receivers (i.e., zero intra-stream skew), in order to minimize media distortion. A playout buffer is often employed to smooth Internet jitter. However, there can be multiple media frames *unavailable* at their scheduled presentation due to extremely large jitter and limited buffer size. To represent the quality of intra-stream synchronization, we define the term *unavailable rate*, denoted by $UR(s_{i,j}^x)$, as the percentage of media frames of media stream $s_{i,j}^x$ that do not arrive at a receiver site in time and thus, cannot be presented.

Intra-media synchronization. We denote $D(s_{i,j}^x)$ as the experienced latency of $s_{i,j}^x$ (after playout buffer). The intra-media synchronization skew $\Delta D(s_{i,j}^x)$ w.r.t. the *reference stream* $s_{i,*}^x$ (*: reference stream index) is defined as:

$$\forall x, i, j: \Delta D(s_{i,j}^x) = D(s_{i,j}^x) - D(s_{i,*}^x) \quad (1)$$

Intra-bundle synchronization. Because sensory streams within a media modality can experience heterogeneous latencies, we prescribe that the latency of a media modality is defined as the latency of the intra-media reference stream within

this modality, in order to best match human perceptual interests, i.e., $D(m_i^x) = D(s_{i,*}^x)$. Thus, the intra-bundle synchronization skew of m_i^x w.r.t. the *reference media modality* m_*^x (*: reference media modality index) is defined as:

$$\forall x, i: \Delta D(m_i^x) = D(m_i^x) - D(m_*^x) \quad (2)$$

Intra-session synchronization skew. Similar to the intra-bundle layer, we prescribe that the latency of a bundle is defined as the latency of the intra-bundle synchronization reference within the bundle, i.e., $D(u^x) = D(m_*^x)$. Given the *reference site* * (*: reference site index), the inter-sender synchronization skew is:

$$\forall x: \Delta D(u^x) = D(u^x) - D(u^*) \quad (3)$$

3. HUMAN PERCEPTUAL EVALUATION

A perceivable inter-sender synchronization error causes a poor experience at third-party observers when they watch remote parties conduct an out-of-sync activity in the teleimmersive space. A larger inter-sender skew will even fabricate a wrong collaborative activity that can mislead the observers. Synchronization control is important in real-time teleimmersive applications, but its tension with interactive quality also affects human perception. In this section, we conduct subjective assessment on inter-sender synchronization and the tradeoff between synchronization and interactive quality, two aspects that have never been studied in previous multimedia synchronization research. Due to space limit, we focus on audio and video media modalities in our user study, but such simplification does not shadow the complexity and significance of the synchronization problems that we investigate.

3.1. Related Work

Past subjective studies showed that all four synchronization layers reflect different aspects of human perceptual requirements. An error in intra-stream synchronization can cause temporal media distortion (e.g., image jerkiness or audio pitch [6]), while an intra-media synchronization skew will introduce a false experience of spatial correlations during media presentation (e.g., a visual mismatch between multi-view images when there is an intra-media skew beyond 80 ms [7]). On the other hand, an intra-bundle synchronization skew can contribute to a perceivable time inconsistency across different media modalities (e.g., a maximum of 80 ms audio-visual skew will not be perceived [2]), and an error in inter-receiver synchronization beyond 150 ms in networked chorus can cause annoying experience [8].

3.2. Description of Subjective Study

We let two participants be situated at different sites and conduct activities in the teleimmersive space. The two sites are in the same local area network, so the outputs should be assumed minimal latency. We record distortion-free videos at both sites. We divide our tests into two categories, and process the user subjective feedback accordingly.

Case 1: Inter-sender synchronization. Two people are playing rock-paper-scissor game in the teleimmersive activity, when a third-party person needs to decide who wins. We insert a controlled fixed latency over one of the sender sites, and the other remains undistorted. We perform perceptual evaluation at the presence of inter-sender synchronization errors.

Case 2: Tradeoff between synchronization and interactive quality. Two people are chatting in the activity. Each site sends one video stream, and controlled variable latency is only introduced over one of the sites. A third-party person, who sits very close to the undistorted speaker, is observing their conversation. We emulate a real multi-layer synchronization buffer adaptation scenario with inter-sender skew control, where we improve the video intra-stream synchronization quality at the sacrifice of interactive quality. In other words, we reduce the unavailable rate of video packets at their scheduled presentation by increasing the buffering latency and the overall end-to-end latency. We would like to study the subjective feedback of this tradeoff.

In both cases, we ask 19 participants (average age: 26) to compare video samples in each test. All video samples are rendered at a resolution of 640x360.

3.3. Subjective Metrics

We conduct subjective assessment tests where two media samples are given to the participants consecutively, and the comparative category rating (CCR) scale [9] is employed to compare the two samples. We use a comparison voting score set of $C = \{3, 2, 1, 0, -1, -2, -3\}$, indicating 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 three metrics.

• Average Score: CMOS

We compute the average of people voting scores for comparative mean opinion score (CMOS) according to ITU-T P.800 [9]. CMOS was often used in previous studies for its efficiency and simplicity. But the major limitation is that the metric computes the average score, and thus, is unable to describe the potential diversity of user votes. A group of participants may output contradicting opinions¹ in a subjective comparison, in which a media system is evaluated under the quality tradeoff (e.g., a comparison between a teleimmersive sample with a better interactive quality but a worse synchronization quality, and another sample with a more satisfactory synchronization quality but a poorer interactive quality). Because participants pay different attentions to heterogeneous quality dimensions, they can have different preferences when measuring the two samples.

¹Contradicting 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.

• 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:

$$N_{>0} + N_{=0} + N_{<0} = N \quad (4)$$

N 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 $N = 19$.

• 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 to identify whether the distribution $(N_{>0}, N_{=0}, N_{<0})$ is inconclusive. The basic idea is that we compute the voting probabilities:

$$(p_{>0}, p_{=0}, p_{<0}) = (N_{>0}/N, N_{=0}/N, N_{<0}/N) \quad (5)$$

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: $\text{binomial}(N, p \geq 0.5)$.

Derivation details can be found in [10]. 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

$$\sum_{g=0}^{N_i} \binom{N}{g} \cdot 0.5^g \cdot 0.5^{N-g} \geq \alpha \quad (6)$$

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 $N_{th} = 12$ (resp. 11, 10) out of $N = 19$ at an inconclusive comparison.

3.4. Subjective Evaluation Results

• Case 1

Fig. 1 presents the comparison studies for media samples with different inter-sender synchronization skews. The first media sample is always with zero inter-sender skew, while the second sample is degraded by a skew as specified in each caption. We show that CMOS monotonically increases as the inter-sender synchronization worsens.

Fig. 1. Comparison results ($N_{>0}$, $N_{=0}$, $N_{<0}$) for case 1. Each caption presents the comparison between two media samples with different inter-sender skews. x-axis shows three categories of comparison votes (voting score > 0 , $= 0$ and < 0), and y-axis describes the number of votes in each category: ($N_{>0}$, $N_{=0}$, $N_{<0}$). The resulting CMOS is also averaged and indicated in each comparison figure.

To evaluate the conclusiveness, we use $\alpha = 70\%$ significance in Eqn. 6. Because $N = 19$, this returns $N_{th} = 10$. Hence, all comparisons reach conclusive results.

• Case 2

Fig. 2 shows several representative comparison results, where we study the tradeoff between video unavailable rate UR and the corresponding one-way end-to-end latency D_e (including buffering latency). 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. 2(6)). The quality tradeoff contributes to this diversity. Generally, if the perceptual degradation in one quality dimension of a teleimmersive 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. 2(6), 9 out of 19 participants prefer a better interactive quality, 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 which has a better intra-stream synchronization quality. As two quality points are moving apart on the tradeoff curve, and one dimension is gradually improving as another dimension is worsening comparably, the likelihood of outputting contradicting opinions is increasing. For example, Fig. 2(6) shows a greater voting diversity than Fig. 2(3) and (5).

Second, the interpretation of the average score CMOS may lack statistical significance at a large variance of user votes. For example, a CMOS = 0 in Fig. 2(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. 2(6) are completely different).

Fig. 2. Comparison results ($N_{>0}$, $N_{=0}$, $N_{<0}$) for case 2. Each caption presents the comparison between two media samples with different unavailable rate and experience latency pairs (UR, D_e). x-axis shows three categories of comparison votes (voting score > 0 , $= 0$ and < 0), and y-axis describes the number of votes in each category: ($N_{>0}$, $N_{=0}$, $N_{<0}$). The resulting CMOS is also indicated in each figure.

Third, when we use $\alpha = 70\%$ significance and $N_{th} = 10$ (same as case 1) to evaluate the inconclusiveness, we find that, except Fig. 2(1,4), all other comparisons are inconclusive.

• Remarks

Both conclusive and inconclusive subjective results are important considerations in designing a mature synchronization control scheme for best user experience. The conclusive findings, e.g., the inter-sender synchronization evaluation in case 1 and all past work described in Section 3.1, should directly be used to guide the synchronization adaptation and reduce the skews of all synchronization layers below a non-perceivable threshold, a number that can only be obtained from subjective tests in heterogeneous teleimmersive activities. The inconclusive opinions, e.g., the tradeoff evaluation between synchronization and interactive quality in case 2, cannot be described by existing standardized metrics like CMOS, so it is very difficult to represent the combined impact of the two quality dimensions in a closed form. A heuristic approach can be used to balance the tradeoff of this multi-dimensional quality during online control.

4. MULTI-LAYER SYNCHRONIZATION SCHEME WITH INTER-SENDER SKEW CONTROL

Based on our subjective findings, we develop a new buffer control scheme for multi-layer synchronization with inter-sender skew control, capable of supporting scalable number of participating sites, media modalities and sensory streams.

Previous studies only investigated inter-receiver skew control, and used a broad concept of *inter-stream synchronization* to cover both intra-media and intra-bundle synchronization [1].

4.1. Problem Formulation

We denote D_n as the latency of media data incurred at the sender and during Internet distribution, and D_e as the overall end-to-end delay. Hence, D_e includes D_n and the buffering latency. The synchronization control scheme must be able to decide D_e for all media data, that lead to bounded synchronization quality at each receiver site.

Recall from Section 3 that the upper bound of synchronization errors can be obtained by performing conclusive subjective tests for different teleimmersive activities. But the adaptation of D_e will create a tradeoff between synchronization and interactive quality. Because a closed form cannot be reached by subjective contradicting opinions, we have to take a heuristic approach by deciding the minimal D_e for each multimedia sensory stream within the session, such that we are able to smooth Internet jitter and bound multi-layer synchronization errors at the best interactive quality

At a given receiver, the problem can be formulated as:

$$\text{Goal:} \quad \forall x, i, j : \min D_e(s_{i,j}^x) \quad (7)$$

$$\text{Constraints:} \quad \text{UR}(s_{i,j}^x) \leq \text{UR}_{\text{TH}}(i) \quad (8)$$

$$|\Delta D_e(s_{i,j}^x)| \leq \delta_1(i) \quad (9)$$

$$|\Delta D_e(m_i^x)| \leq \delta_2(i, *) \quad (10)$$

$$|\Delta D_e(u^x)| \leq \delta_3 \quad (11)$$

In the above equations, Eqn. 7 specifies all media frames within a sensory stream $s_{i,j}^x$ should maintain a same D_e to guarantee intra-stream synchronization. We count within each stream the number of media frames that should have arrived at the receiver site. Eqn. 8 indicates the maximal unavailable rate of these media frames at their scheduled presentation, where $\text{UR}_{\text{TH}}(i)$ is the upper bound of the unavailability of the media modality i . Eqn. 9-11 specify the multi-layer skews, and $\delta_1(i)$, $\delta_2(i, *)$ and δ_3 represent the upper bound of intra-media, intra-bundle and intra-session (inter-sender) synchronization errors respectively, which can be derived from subjective tests.

4.2. Algorithm

At each receiver site, we decide D_e of all received sensory streams using the following three steps.

Step 1. By recording arrival time of all media frames, we first decide the lower bound of $D_e(s_{i,j}^x)$ for each sensory stream $s_{i,j}^x$ to satisfy Eqn. 8, and denote it as $\hat{D}_e(s_{i,j}^x)$. We sort all received sensory streams from all sender sites by their $\hat{D}_e$ value from the longest to the shortest, and denote the sorted list of $\hat{D}_e$ as $\{\hat{D}_e(s_{i,j}^x)\}$. We also find the mapped sensory stream list $\{\tilde{s}_q\}$ such that:

$$\forall q < q' : \hat{D}_e(\tilde{s}_q) > \hat{D}_e(\tilde{s}_{q'}) \quad (12)$$

Here, $\{\tilde{s}_q\}$ is a set that covers all multi-modal sensory streams from all remote sender sites, and q is the stream index.

Step 2. Based on $\{\hat{D}_e\}$, we update the actual D_e value ($D_e \geq \hat{D}_e$) of all sensory streams in the sorted list for multi-layer synchronization. For each stream $\tilde{s}_r$, we evaluate the multi-layer skews introduced by $\tilde{s}_r$ and $\tilde{s}_{r'}$ ($\forall r' = \{1, \dots, r-1\}$), where r and r' are both stream indices. If the synchronization satisfies Eqn. 9-11 constraints, we prescribe $D_e(\tilde{s}_r) = \hat{D}_e(\tilde{s}_r)$. Otherwise, we increase the buffer of $\tilde{s}_r$ and set $D_e(\tilde{s}_r)$ to be the smallest number which does not violate the multi-layer synchronization requirements against $\tilde{s}_{r'}$ ($\forall r' = \{1, \dots, r-1\}$). We work on all received sensory streams, and update the $\{D_e\}$ list accordingly.

Step 3. Due to multi-layer synchronization interactions, step 2 may be repeated multiple times before Eqn. 9-11 are achieved.

4.3. Evaluation Results

We rely on a real multi-site teleimmersive testbed to evaluate our multi-layer synchronization control scheme. While only one audio stream and three 3D video streams (representing three views) are configured at each site in our user study and evaluation testbed, our human-centric synchronization scheme can be generalized to other media modalities. To make our results repeatable, we prerecord at the senders the size of original audio and multiview video frames, so that we are able to transmit the same media data in every experiments. A network emulator is implemented between each sender and receiver sites, replaying variable delay distributions for media packets according to real Internet traces collected in Planet-Lab. To simplify the description, we only present three-site scenario where two sites (site 1 in USA and site 2 in China) are playing rock-paper-scissor game, and the other (site 3 in USA) is a passive observer.

We evaluate our synchronization control algorithm (Alg 1) at receiver site 3. As a comparison, we also study traditional single-reference inter-stream synchronization studies (Alg 2) [1, 11], where intra-media and intra-bundle synchronization skews are combinedly computed based on a single reference stream in Eqn. 9-10. In our Alg 1, we choose $s_{V,3}^1/s_{V,3}^2$ and $s_{A,1}^1/s_{A,1}^2$ as the video/audio reference stream, audio m_A^1/m_A^2 as the reference modality, and n^2 as the reference site. In previous Alg 2, we simply choose $s_{A,1}^1/s_{A,1}^2$ as the inter-stream synchronization reference for all video and audio streams. Based on our subjective findings, we use $\text{UR}_{\text{TH}}(\text{'V'}) = 4\%$, $\text{UR}_{\text{TH}}(\text{'A'}) = 1.5\%$, $\delta_1(\text{'V'}) = \delta_1(\text{'A'}) = 0$, $\delta_2(\text{'V'}, *) = \delta_2(\text{'V'}, \text{'A'}) = 80$ ms, and $\delta_3 = 60$ ms in Eqn. 8-11. We demand zero intra-media skew (i.e., $\forall i, \delta_1(i) = 0$) for minimal spatial distortion across all media modalities.

We update the control status every 8 seconds². Fig. 3 and 4 present the intra-media and intra-bundle synchronization

²8-second status update is for evaluation only. In real applications, the control status is updated during silence period to mitigate media distortion.

Fig. 3. D_n and resulting D_e adaptation at n^3 , for sensory streams from n^2 . In (a),(d), the two D_e curves of Alg 1 and 2 overlap.

Fig. 4. (a) Snapshot of $\Delta D_n(s_{V,1}^2)$; (b) Percentile of synchronization skew: $\Delta D_n(s_{V,1}^2)$, $\Delta D_n(s_{V,2}^2)$ and $\Delta D_n(m_V^2)$; (c-d) Resulting synchronization skew due to receiver adaptation: $\Delta D_e(s_{V,1}^2)$, $\Delta D_e(s_{V,2}^2)$ and $\Delta D_e(m_V^2)$. In (c), $\Delta D_e(s_{V,1}^2) = \Delta D_e(s_{V,2}^2) = 0$.

Fig. 5. Synchronization skews based on Alg 1: $\Delta D_e(s_{V,1}^2) = \Delta D_e(s_{V,2}^2) = 0$, $\Delta D_e(m_V^2) = 80$ and $\Delta D_e(m_A^2) = 60$ ms.

results over a 40-second span. Fig. 3 shows D_n variation and the resulting D_e adaptation at the receiver site 3 for the four sensory streams from the sender site 2. Between two consecutive updates, the intra-stream synchronization is guaranteed by maintaining equal D_e of all media frames in each sensory stream. Measured at the receiver site 3, Fig. 4(a)(b) present the intra-media and intra-bundle skews ΔD_n introduced over the Internet, and Fig. 4(c)(d) demonstrate the multi-layer skews of ΔD_e after synchronization control. The traditional single-reference algorithm (Alg 2) is incompetent to bound the skews across video streams, while our proposed algorithm (Alg 1) can perform successful and consistent synchronization based on hierarchical references.

Because site 1 and 3 are both located in USA, the network between the two sites is characterized by low latency (≈ 40 ms) and little jitter (≈ 5 ms). Hence, we insert the buffering latency at the receiver site 3 over the media data from the sender site 1, such that the inter-sender skew is always equal to $\delta_3 = 60$ ms. Fig. 5 shows the resulting intra-media, intra-bundle and intra-session (inter-sender) synchronization skews based on Alg 1. The results demonstrate the effectiveness of our multi-layer synchronization control scheme.

To conclude, our subjective test proves the importance of

inter-sender synchronization to human perception in teleimmersive applications, but at the same time, shows that the current techniques are yielding inconclusive results during the presence of synchronization and interaction tradeoff. Our proposed human-centric multi-layer synchronization scheme takes into account the unique teleimmersive inter-sender synchronization requirement, and yields better results to address the tradeoff between synchronization and interactive quality. This work is supported by NSF CNS 10-12194.

5. REFERENCES

- [1] F. Boronat, J. Lloret, and M. Garcia, "Multimedia group and inter-stream synchronization techniques: a comparative study," *Information Systems*, 2009.
- [2] Z. Huang, K. Nahrstedt, and R. Steinmetz, "Evolution of temporal multimedia synchronization principles: A historical viewpoint," *ACM Transactions on Multimedia Computing Communications and Applications*, Oct. 2013.
- [3] ITU-G.1070, "Opinion model for video-telephony applications," 2007.
- [4] R. Brandenburg, H. Stokking, and et al, "RFC7272: Inter-Destination Media Synchronization (IDMS) Using the RTP Control Protocol (RTCP)," June 2014.
- [5] I. Vaishnavi, P. Csar, D. Bulterman, and et al, "From IPTV to synchronous shared experiences challenges in design: Distributed media synchronization," *Signal Processing: Image Communication*, vol. 26, no. 7, pp. 370–377, 2011.
- [6] B. Blesser, "Digitization of audio: a comprehensive examination of theory, implementation, and current practice," *Journal of the Audio Engineering Society*, 1978.
- [7] L. Goldmann, J. Lee, and T. Ebrahimi, "Temporal synchronization in stereoscopic video: Influence on quality of experience and automatic asynchrony detection," in *IEEE Int'l Conference on Image Processing*, 2010.
- [8] Y. Miyashita, Y. Ishibashi, and et al, "QoE assessment of group synchronization in networked chorus with voice and video," in *IEEE TENCON*, 2011.
- [9] ITU-P.800, "Methods for subjective determination of transmission quality," 1996.
- [10] Z. Huang, "The design of a multi-party VoIP conferencing system," M.S. thesis, University of Illinois at Urbana-Champaign, Urbana, IL, 2009.
- [11] 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.