Effect of Presenting Video as a Baseline During an American Sign Language Animation User Study

Pengfei Lu
The Graduate Center, CUNY
City University of New York
Doctoral Program in Computer Science
365 Fifth Ave, New York, NY 10016
+1-212-817-8190
pengfei.lu@qc.cuny.edu

Hernisa Kacorri
The Graduate Center, CUNY
City University of New York
Doctoral Program in Computer Science
365 Fifth Ave, New York, NY 10016
+1-212-817-8190
hkacorri@gc.cuny.edu

ABSTRACT
Animations of American Sign Language (ASL) have accessibility benefits for many signers with lower levels of written language literacy. Our lab has conducted several prior studies to evaluate synthesized ASL animations by asking native signers to watch different versions of animations and to answer comprehension and subjective questions about them. As an upper baseline, we used an animation of a virtual human carefully created by a human animator who is a native ASL signer. Considering whether to instead use videos of human signers as an upper baseline, we wanted to quantify how including a video upper baseline would affect how participants evaluate the ASL animations presented in a study. In this paper, we replicate a user study we conducted two years ago, with one difference: replacing our original animation upper baseline with a video of a human signer. We found that adding a human video upper baseline depressed the subjective Likert-scale scores that participants assign to the other stimuli (the synthesized animations) in the study when viewed side-by-side. This paper provides methodological guidance for how to design user studies evaluating sign language animations and facilitates comparison of studies that have used different upper baselines.

Categories and Subject Descriptors
H.5.2 [**Information Interfaces and Presentation**] User Interfaces – *evaluation/methodology*; K.4.2 [**Computers and Society**]: Social Issues – *assistive technologies for persons with disabilities*.

General Terms
Design, Experimentation, Human Factors, Measurement.

Keywords
Accessibility Technology for People who are Deaf, American Sign Language, Animation, Baseline, User Study.

1. INTRODUCTION
For various educational and language exposure reasons, a majority of high school graduates (typically at age 18-21) who are deaf in the U.S. have lower-than-average levels of literacy in written English; specifically, the average is a fourth-grade (age 10) English reading level or below [24]. So, many adults who are deaf have difficulty reading text that may appear on websites, captioning, or other media. More than half a million people in the U.S. use American Sign Language (ASL), a language with a distinct word order, linguistic structure, and vocabulary than English [19]; many adults have more sophisticated fluency in ASL than in English. Thus, presenting information as computer animations of ASL can make information and services accessible to deaf people with lower English literacy, as explained in [7].

While videos of human signing can be used in some applications and websites, there are limitations. If the information is frequently updated, it may be prohibitively expensive to continually re-film a human performing ASL for the new information. Computer synthesized animations allow for frequent updating, automatic production of messages (via natural language generation or machine translation techniques), wiki-style applications in which multiple authors script a message in ASL collaboratively, or scripting of messages by a single human author for presentation in an anonymous fashion (that does not reveal the face of the human author, as would happen in a video of them performing ASL). Assembling video clips of individual signs together to synthesize ASL messages does not allow for sufficient control, blending, and modulation to produce smooth transitions between signs, subtle motion variations in sign performances, or proper combinations of facial expressions with signs. Thus, animated virtual human characters are used by sign language synthesis research systems.

As part of our research on ASL animation, our laboratory has conducted studies to evaluate the understandability and naturalness of ASL animations. Typically, we ask native ASL signers to view our animations and then answer comprehension and subjective Likert-scale questions about the animations. We have developed several novel methodologies for conducting such studies, including: protocols to screen for native ASL signers [9], scripts of ASL stimuli that contain specific linguistic phenomena of interest [10, 11], comprehension questions presented in ASL with answer choices presented with images and photos corresponding to each choice (to enable participation by signers with limited English skills) [8], etc. In this paper, we investigate another important methodological issue; specifically, we examine the question of what type of upper baseline should be presented in an ASL animation evaluation study.

ASSETS'12, October 22–24, 2012, Boulder, Colorado, USA.

Our evaluation studies drive our research, as we explore alternative mathematical models for specifying the movements of an animated character to produce a clear and understandable animation. For instance, we have conducted studies that focus on the use of space around a signer to represent entities under discussion, the movement of verb signs, etc. [8, 9, 11]. Typically, in a study, we wish to compare whether one version of our animations (based on one mathematical model) out-performs another version of animations (based on a different model). We sometimes use comprehension questions in our evaluation studies; so, the scores that participants achieve depend not only on the animation stimuli shown but also on the difficulty of the comprehension questions. Thus, we often included an upper baseline (a third type of animation for comparison purposes) in our studies to make the results more meaningful. Usually, we have used a computer animation (of a visually identical animated character) whose movements were carefully specified by a native ASL signer with animation software experience. The rationale for this choice was that we believed the movements of the character specified by a human animator were an "ideal" that we sought to achieve with our animation models used for automatic synthesis.

Other researchers at conferences have sometimes asked why we have not included videos of humans performing ASL sentences as the upper baseline in our studies. We had been wary of including videos of humans because we were concerned that this might lead participants to focus on the superficial appearance of the human and the signing character, not on their movements, which were the focus of our research program. For instance, our experimental studies typically include additional time for the participants to provide unstructured feedback comments about the animations they have seen that day, and the comments about aspects of the character movement are most useful for our research. As a lab studying the linguistic movements of the body during ASL (and not the computer graphics issues related to how to display human figures), we were not focusing our research on the photorealism or graphical qualities of the character, but rather on its movements.

This paper explores what type of upper baseline should be used in user-based experimental studies of sign language animations – specifically, whether videos of a human signer would be a better upper baseline than a character animated by a human. Section 2 surveys evaluation methodologies currently used in the research community, and section 3 provides additional motivation for this research. Sections 4 and 5 describe our experimental study in which we replicate a study previously published in 2010 [10] and replaced the upper baseline in that study with a video of a human signer, and the results of the experiment study. Section 6 discusses our conclusions and future work.

2. RELATED WORK

Few researchers have explicitly discussed methodological aspects of sign language animation user-studies. We searched the literature for examples of studies where we can identify the use of particular baselines (the original authors may not have discussed their studies in these terms). Researchers studying animations of *non-signing* virtual humans are also discussed briefly in this survey below. We see that evaluation conducted can be organized into three categories with regard to the upper baseline used.

The first category is where no upper baseline exists. Although evaluation against a baseline usually results in more meaningful scores, many user-studies don't include any baselines. To test the feasibility of their approach, researchers ask users to evaluate the human animation under development without any baselines for comparison [22], conduct their experiment in multiple steps trying to improve at every step [3], or improve the parameters of their animation models through repeated presentation of an animation with slightly modified parameters values [3].

The second category is research in which videos of a human are used as an upper baseline for comparison to the animation being generated. For example, Kipp et al. compare avatars to human signers [13]. While not used as an upper baseline, an instructional video of a native signer appeared in the interface of the software presenting animations in studies conducted by Schnepp et al. [23, 21]; the video's use may have impacted participants' scores given to animations. Researchers studying non-signing virtual characters have also sometimes used videos of humans as a baseline for comparison, e.g. [1, 4].

The last category is research in which animation was used as an upper baseline in the evaluation; this seems to be the most popular approach in sign language animation (and non-signing virtual human) research. The similarity of appearance between the virtual characters in the "upper baseline" animation and the character in the animation under evaluation seems to vary across the studies. Further, this category can be divided into two subcategories by the way the upper baseline animation is created and manipulated.

The first covers upper baseline animations controlled by a human animator without any motion-capture data. As discussed above, up to now, this is the approach we have favored in our prior studies, in which we asked a human animator to carefully produce an animation of a virtual human to serve as our upper baseline [11, 15, 18]. Researchers studying the animation of non-signing virtual human characters (performing gestures along with speech) have employed a similar methodology, e.g. [2].

Finally, many researchers create their upper baseline animation by combining an animation tool, an animator, and data from a real human (collected via motion capture). Some researchers use a virtual human for the upper baseline that is visually identical to the character used in the animations being evaluated ([5, 6]), and some use a visually different character ([12] and part of the experiment in [5]). Kipp et al. [14] showed participants sign language animations produced by a variety of techniques. Non-sign-language animation researchers have also used virtual humans driven by motion-capture as upper baselines [20].

3. PRIOR RESULTS & HYPOTHESES

Given the diversity of study designs in the sign language animation research community, it is useful to understand the advantages of using human-video or animation upper baselines. In order to quantify the effect of showing videos of human signers in an ASL animation evaluation study, we needed to conduct an identical study in two ways: (1) once *without* using videos of human signers as an upper baseline and (2) once *with* such videos. Since we have up to now (with one exception, discussed below) conducted studies in which an animation produced by a skilled human animator was the upper baseline, we decided to replicate a previously conducted study, replacing the upper baseline with a video of human signer performing identical ASL stories as the animated character. We can examine how the comprehension scores, Likert-scale subjective evaluation responses, and feedback comments in the study may differ. Before discussing the details of this study replication in sections 4 and 5, we first wanted to

discuss a pair of prior studies that almost (but not quite) had this same structure. The results of this prior pair of studies formed our hypotheses for the work presented in sections 4 and 5.

An ongoing project at our laboratory is to collect a large sample of sign language sentences using motion-capture equipment worn by native signers. We wanted to evaluate whether we had correctly calibrated our motion-capture equipment to obtain clear movement data from the human signers. So, in [16], we evaluated virtual human animations based on our motion-capture recordings by comparing them to an upper baseline, which consisted of an animation of a virtual human character designed by a human animator. A few years later, in [17], we conducted a similar study to evaluate virtual human animations based on our motion-capture recordings, but we used a different upper baseline: in this new study, we showed a video of the human wearing the motion-capture equipment during the recording session. In both studies, native ASL signers who saw the animations/videos answered comprehension questions and Likert-scale subjective evaluation questions; by comparing how the scores in the two studies change, we gain insight into the effect of using a different upper baseline.

Changing the upper baseline did not produce a difference in the comprehension question scores for the other stimuli in the study (the motion-capture-based animations), which had similar scores in both studies [17]. More interesting was the effect on the Likert-scale subjective scores (1-to-10 for naturalness of movement, perception of understandability, and grammatical correctness of the animations). In the later study (with the human video upper baseline), the motion-capture-based animations received lower subjective scores than they had in the prior study (with the animation upper baseline). We speculated that seeing a video of real human as one of the stimuli being evaluated in a study led participants to assign lower subject ratings to the animations [17]. That is, none of the animations subjectively seemed as good when shown in comparison to videos of real people signing. But this was just speculation: to determine if including videos of humans as a baseline in a study would produce a depressive effect on subjective scores for other stimuli (and what the magnitude of this effect would be), we would need to conduct a carefully controlled pair of studies that were identical in all aspects, except for one of them including videos of real humans as an upper baseline (as we discuss in sections 4 and 5 of this paper). The pair of studies described above [16, 17] was not a sufficient test because the script of the stories in the two studies was not identical; further, the animations evaluated were motion-capture-based animations, not synthesized ASL animations (our lab's primary focus).

While not a perfect test, this prior pair of studies helped us formulate hypotheses of how displaying videos of humans as an upper baseline would affect the results of a study evaluating synthesized ASL animations. We hypothesize the following:

- **H1:** A human video upper baseline will receive higher comprehension question accuracy scores than an animated-character upper baseline produced by a human animator.
- **H2:** The upper baseline used (human video or animated character) would not affect the comprehension questions accuracy scores for the other stimuli shown in the study.
- **H3:** A human video upper baseline will receive higher Likert-scale subjective scores than an animated-character upper baseline.
- **H4:** Using a human video upper baseline will depress the subjective Likert-scale scores that participants assign to the other stimuli (the synthesized animations) in the study.

4. REPLICATING A STUDY FROM 2010

Because the pair of studies discussed above [16, 17] had more differences between them than merely the addition of a human video as an upper baseline, they didn't allow us to isolate how this aspect of the study design affected the collected data. Thus, we selected a study originally presented at ASSETS 2010, and we decided to replicate this study, with the only change being the use of a human video as an upper baseline, instead of an animation.

In [10], we evaluated a model we designed for synthesizing the movements of "inflected" verb signs, whose movements depend on locations in the space around the signer where the verb's subject and object have been previously set up. We wanted to know how understandable ASL animations, in which the verbs were produced using our new model, would be, compared to: (1) a lower baseline consisting of "uninflected" versions of the verb signs (the unvarying/uncustomized dictionary form of each sign whose movement doesn't indicate subject/object) and (2) an upper baseline consisting of animations of inflected versions of each verb produced by a native ASL signer human animator.

We conducted an evaluation study with native ASL signers that consisted of two parts: In part 1, participants viewed animations of a virtual human character telling a short story in ASL. Each story included instances of the inflected verbs. Fig. 1 shows a story transcript; colors indicate locations around the signer where the verb's subject/object are located. After watching each story animation (of one of three types: inflected, uninflected, animator-produced) one time, participants answered multiple-choice comprehension questions. Questions focused on whether they understood and remembered the subject and object of each verb. For each story viewed, participants also responded to three 1-to-10 Likert-scale questions about how grammatically correct, easy to understand, or naturally moving the animation appeared.

In part 2 of the study, participants viewed three versions of an animation of a single ASL sentence side-by-side on one screen, as depicted in Fig. 2(a). The sentences shown side-by-side were identical, except for the version of the verb which appeared in each: version produced by our mathematical model, uninflected version of the verb (lower baseline), or version of the verb carefully created by a native ASL signer using animation software [25] (upper baseline). The participants could re-play each animation as many times as they wished. Participants were asked to focus on the verb and respond to a 1-to-10 Likert-scale question about its grammaticality, understandability, and naturalness in each of the three versions of the sentence. We used the methodology for similar ASL evaluation studies in [8, 9].

HELLO, MY NAME #CHARLIE.
I YOUR NEW BOSS.
MY ASSISTANT #JEFF THERE$_{RED}$
WILL GIVE$_{RED}$ →$_{YOU}$ YOU NEW INFORMATION.
WHEN YOUR OFFICE READY,
MY MANAGER #BOB THERE$_{BLUE}$ WILL TELL→$_{YOU}$ YOU.
YOUR NEW PHONE HE$_{RED}$ WILL GIVE$_{RED}$ →$_{YOU}$ YOU.
I SORRY.
HE$_{RED}$ LOSE YOUR KEY.
HE$_{BLUE}$ SCOLD→$_{RED}$ HIM$_{RED}$.
NOW, HE$_{BLUE}$ FORCE HIM$_{RED}$ PAY NEW KEY.

Fig. 1. Script for a story shown in the study.

To evaluate the four hypotheses listed in section 3, we needed to replicate our 2010 study [10], using the same set of passages, and questions. The only difference in our new 2012 study is that we replace the upper baseline animations from the 2010 study with videos of a human signer. Specifically, we recorded a human performing the 9 stories (with inflected versions of the verbs) for part 1 of the study and the 12 sentences (with inflected versions of the verbs) for part 2 of the study. For example, the top row in Fig. 2 shows what the participants saw in the part 2 (side-by-side comparison) in 2010 and the lower row (Fig. 2(b)) shows what they saw in 2012. All the other animations and their sequencing in this pair of studies were identical. All of the instructions and interactions for the study, in both 2010 and 2012, were conducted in ASL by a native signer, who is professional interpreter.

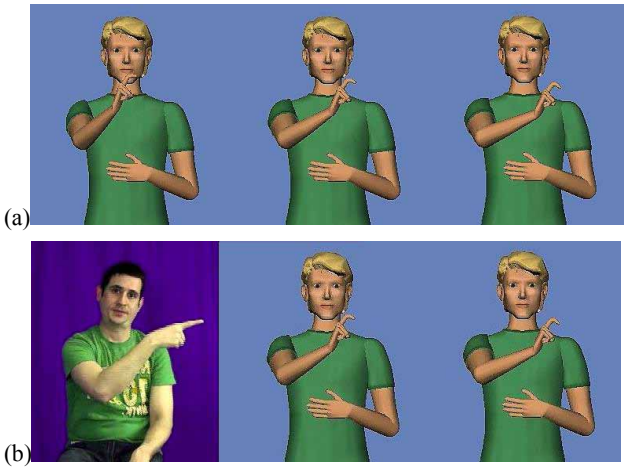

Fig. 2. Screenshots of the side-by-side comparison portion of the studies as shown to participants in (a) 2010 and (b) 2012.

In prior work, we had developed methodologies to ensure that responses given by participants are as ASL-accurate as possible. In [9], we discussed the importance of participants being native ASL signers and the study environment being ASL-focused with little English influence; we developed questions to screen for native ASL signers. For the 2010 study, ads were posted on New York City Deaf community websites asking potential participants if they had grown up using ASL at home or had attended an ASL-based school as a young child. Of the 18 participants recruited for the study, 12 participants learned ASL prior to age 5, and 4 participants attended residential schools using ASL since early childhood. The remaining 2 participants had been using ASL for over 15 years, learned ASL as adolescents, attended a university with classroom instruction in ASL, and used ASL daily to communicate with a significant other or family member. There were 12 men and 6 women of ages 20-56 (average age 30.5).

For our new study in 2012, we also recruited 18 native ASL signers as participants using similar techniques. Of the 18 participants, 16 participants learned ASL prior to age 5, and 10 participants attended residential schools using ASL since early childhood. The remaining 2 participants had been using ASL for over 13 years, learned ASL as adolescents, attended a university with classroom instruction in ASL, and use ASL on a daily basis to communicate with a significant other or family member. There were 12 men and 6 women of ages 22-49 (average age 32.8).

To produce the human video upper baseline, we recorded the videos from a native signer in our studio, where we asked the signer to sit on a stool in front of a blue curtain to match the background color in the animations we presented to the participants. The human signer also wore a green t-shirt on the day of the recording, which was similar to a virtual human character. The camcoder was placed facing the signer at his head height, which matches the perspective of the virtual human in the animations we presented to the participants in the experiment. We used one large monitor in front of the signer to display the story scripts (like the example in Fig. 1) during the recording. The signer had time to memorize and practice each of the scripts prior to the recording session. All of the instructions and interactions for the recording session were conducted in ASL by another native signer (a research assistant in our lab) sitting behind the camcorder – this was important to ensure that the signing being recorded was as fluent as possible. To produce videos which have the same time duration as the upper baseline animations that had been used in 2010, we asked the native signer being recorded to practice several times before the recording, and we used a stopwatch to measure how many seconds he took for each story during the practice and recording. Finally, after making several recordings of each story, we picked the one video recording of the story with the closest time duration to the upper baseline animation from 2010. We cropped and resized the video files to match the height/width of the 2010 upper baseline animations – and to approximate the same placement of a human in a the video frame as how the virtual human character had appeared in the animation in 2010. The framerate and resolution of the video was identical to the animation from 2010.

For this study, it was not only important that the sequence of signs performed in the story should match the animations, but also the locations in the surrounding signing space where the human or character points (to represent entities under discussion) needed to be identical. Fig. 3 illustrates how we set up small colored paper squares around the studio (with colors that matched the script in Fig. 1) to guide the human where to point or where to aim the motion path of inflecting verb signs during the recording session.

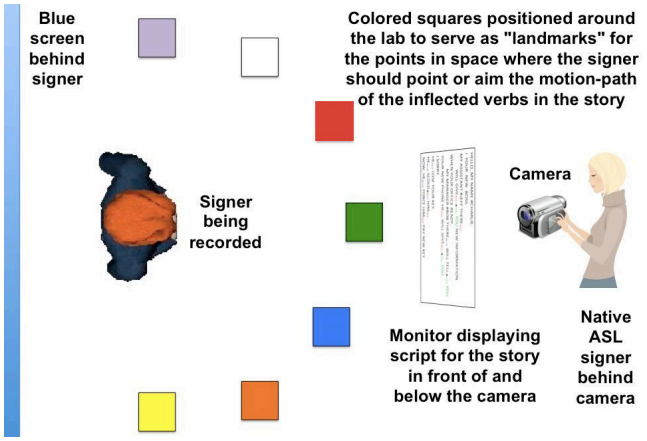

Fig. 3. Diagram of an overhead view of recording studio.

In order to serve as an effective upper baseline for comparison in a study, we would like to "control" as many of the variables of the ASL performance as possible – so that it is mostly the variable we care about which differs between the upper baseline and the

animation under primary evaluation scrutiny (which for us, was the synthesized animation using our verb inflection model). Producing a video recording of a human that "matched" the animations being shown as stimuli in the study was very difficult. At a minimum, we needed the signer to perform the same "script" of signs as the other (animation) versions of stimuli shown in the study, and we had to employ the colored squares described above to indicate where the signer should point or how the verbs motion paths should be aimed. Since ASL has no standard written form, we had to explain our notation scheme to the participant being recorded. Because the stories were a bit complicated (an average of 55 signs in length, included 3-5 main characters set up at various locations in the signing space, with 3-5 inflected verbs per story), the signer required a lot of practice in order to perform each story smoothly. Even when we asked the signer to try not to look at the scripts too much during the recording process, the signer still needed to glance at the script occasionally during the performance, which produces a somewhat infelicitous video with the signer's eyes glancing between the monitor displaying the story transcripts and the camcorder.

Further, the script notation does not capture all of the subtleties of performance that are part of ASL; it is merely a loose sketch of what must be signed. We also had to let the signer know how to control the speed of the signing, facial expressions, torso movement, head movement, etc. The signer had to practice before he was able to finish a story in a certain numbers of seconds. We also asked the signer not to add embellishments, e.g., additional emotional facial expressions, which hadn't appeared on our virtual human character's face. This coaching and scripting process is a delicate "balancing act" – on one hand, we want to record a natural, fluent version of the sentences from the human signer, but on the other hand, we want to control as many variables as possible so that they are held constant between our upper-baseline video and our animation being evaluated. Some participants in the study noticed problems in the human video, e.g., commenting "… person signs well but need little [more] facial expression." Other participant comments appear in section 5.

5. RESULTS AND COMPARISON

This section discusses and compares the results obtained in the original 2010 study and the new 2012 study – this includes the comprehension-question and Likert-scale scores collected in part 1 of the studies (after a participant viewed a story one time) and the Likert-scale scores collected in part 2 of the studies (in which participants assigned a score to each of the three sentences which they viewed side-by-side). In Fig. 4, 5, and 6, which display the results, the thin error bars in each graph display the standard error of the mean. Green colors indicate data collected in 2010, and purple, in 2012. Animator10 and Video12 were the upper baselines, Uninflect10 and Uninflect12 were the lower baselines, and Mode110 and Mode112 were the versions of the animations produced using our verb inflection model. Note that Uninflect10 and Uninflect12 were identical stimuli, the only difference was that the evaluation scores were collected in either the 2010 or 2012 study – likewise for Mode110 and Mode112.

To check for statistical significance, one-way ANOVAs were used for comprehension-question data, and Kruskal-Wallis tests for Likert-scale scores. The following comparisons were planned and conducted: (1) all three values from 2010, (2) all three values from 2012, (3) Video12 and Animator10, (4) Mode112 and Mode110, and (5) Uninflect12 and Uninflect10. Any of these planned comparisons that were statistically significant ($p<0.05$) have been marked with a star in Fig. 4, 5, and 6.

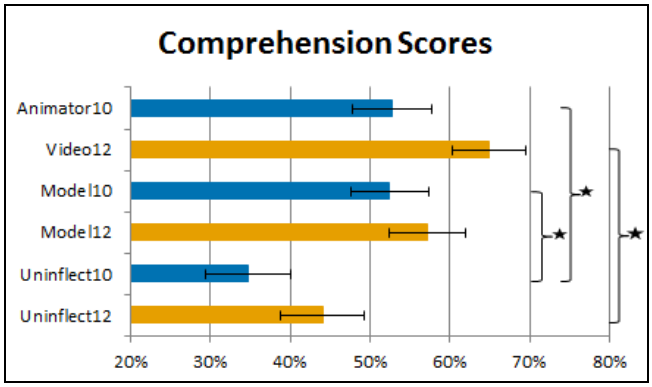

Fig. 4. Results of Comprehension Scores

Fig. 4 displays the comprehension-question accuracy scores from "part 1" of the 2010 and 2012 studies. There was no significant difference between Animator10 and Video12; so, hypothesis H1 was not supported. This was a surprising result: our videos of a human signer did not achieve higher comprehension scores than the animations of a virtual human with the verbs carefully animated by a human. This indicates that our upper baseline used in the 2010 study was a reasonable choice. Another surprising result (though not statistically significant) was that the 2012 scores for Model and Uninflected seemed a little higher. While no story was displayed more than one time during the study, we speculate that seeing a video of a human performing some of the ASL stories (with 3-5 characters set up in space and extensive use of inflected verbs) may have helped participants grasp the idea of the *overall genre* of the stories shown in the study, and perhaps this led to better comprehension scores for Model and Uninflect.

Since the differences between Mode110/Mode112 and between Uninflect10/Uninflect12 were not significant, then hypothesis H2 was partially supported – changing the upper baseline didn't significantly affect these scores. Of course, the support for H2 isn't clear. In 2010, there was a significant difference between Mode110 and Uninflect10, in 2012, this significant difference could no longer be observed. So, H2 is only partially supported.

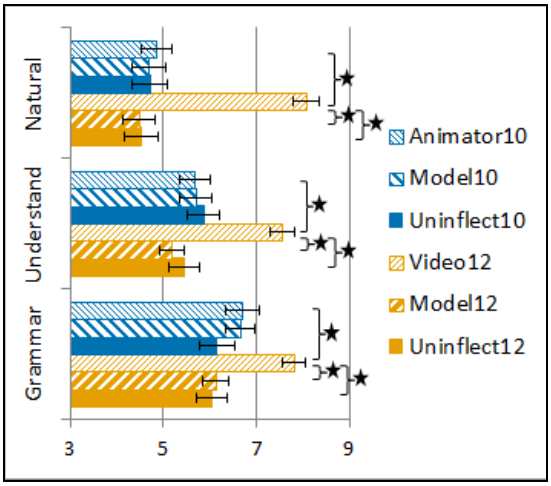

Fig. 5. Results of Grammaticality, Understandability, and Naturalness Likert-Scale Scores in the Two Studies

Fig. 5 displays the 1-to-10 Likert-scale subjective scores for grammaticality, understandability, and naturalness from "part 1" of the studies. Video12 had significantly higher grammaticality, understandability, and naturalness scores than Animator10 – thereby supporting hypothesis H3, that video of a human would get higher subjective Likert-scale scores than a virtual character animated by a human. Given that there was no significant difference in the comprehension scores between these, it was interesting that the subjective scores were significantly different. Participants subjectively preferred the videos, although there was no significant improvement in the comprehension scores.

In a similar vein, we note that Video12 had significantly higher Likert-scale scores than Mode112 (Fig. 5) but did not have significantly higher comprehension scores than Mode112 (Fig. 4). Videos of human signers seem to get higher subjective scores than do animations of virtual characters, but there isn't always a significant benefit in the comprehension scores. Given this observation, it is reasonable for future ASL evaluation studies to include both comprehension and Likert-scale subjective questions, since they seem to be measuring different aspects of animations.

The results in Fig. 5 did not support hypothesis H4; there was no significant depression in the Likert-scale scores for Model or Uninflect when we used the video upper baseline in 2012. When we examine the Likert-scale scores obtained during side-by-side comparisons in Fig. 6, we will see some contradictory results.

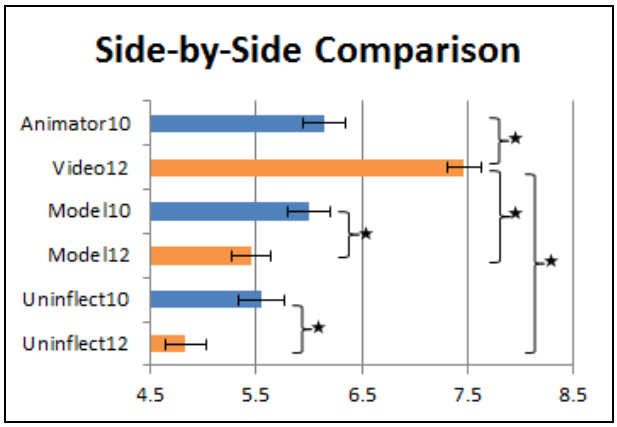

Fig. 6. Results of Side-by-side Comparison Scores

Fig. 6 displays the Likert-scale subjective scores collected from participants in 2010 and 2012 during "part 2" of the studies (the side-by-side comparison of identical sentences, with different versions of the verb in each, which could be replayed many times). Video12 is significantly higher than Animator10, further supporting hypothesis H3 (human videos would get higher Likert-scale subjective scores than animation upper baselines).

In Fig. 6, hypothesis H4 *was* supported: Using a human video upper baseline depressed the subjective Likert-scale scores that participants gave to the animations. Mode112 was significantly lower than Mode110, and Uninflect12 was significantly lower than Uninflect10. The magnitude of this depression is 10%-20%. This is not a surprising result; when looking at videos of humans in direct comparison to animations of a virtual human character, it is reasonable that participants would feel that the animations are less natural/grammatical. What is surprising is that we had not observed any significant depression in Fig. 5 when looking at the Likert-scale data from part 1, in which participants assigned a Likert-scale subjective score to a story that they had just watched.

One explanation for this result may be that the depressive effect may depend on whether participants are assigning Likert-scale subjective scores to videos in a side-by-side direct comparison (as in part 2, Fig. 6) or sequentially throughout a study (as in part 1, Fig. 5). Perhaps in the side-by-side setting, the video looks very distinct from the other two stimuli, which are both animations. Another possible explanation for this result may be that during part 1 of the study, when watching a story one time and then answering the comprehension questions, the participants may have been very focused on the task of trying to understand and remember as much information as possible from the stories. Thus, they may have been less focused subjectively on the superficial appearance of the animations/videos.

Based on the feedback comments in 2012, participants indicated they felt comfortable with the experiments and animations/videos presented, writing: "It was overall good," "It is interesting to watch it," "That's a learning experience for me to get used to ASL animation," and "Would love to see the software!" The most frequent comments from participants were on the topic of the signing/appearance of the virtual human when comparing to the real human in the videos; some participants wrote: "The robot needs to look/sign more like a human," "Some pointing names were a little confusing," "When robot fingerspell a person name need little clearer," "The position of the eyes isn't exactly helpful," and "Some signs I could not understand like 'you,' it is easily overlooked. I didn't realize it was referring to "you" the whole experiment." Another theme in the feedback comments was that the facial expression of the virtual human character needs improvement, writing: "Lack of facial expression makes it quite difficult for me to understand the avatar", "I would like to see more facial expression but like eyebrows, mouth movement," etc.

As discussed in section 4, when creating baselines for comparison to animations in a study, a balance must be achieved between matching the content of the stimuli across versions and allowing for natural signing. Some of the comments of participants in the study indicated that in a few cases, we were not successful at this. Specifically, when producing the script for the human to perform in the video recordings, we included every sign that was performed by the virtual human character in the upper baseline animations from 2010. When a signer sets up points in space to represent entities under discussion, the signer may refer to these items later in the conversation by pointing to them. Because the movement path of an inflected ASL verb indicates the location around the signer where the subject and object of the verb are established, it is common (but not required) for signers to omit pointing to the subject/object before/after the verb (because the location in space that represents those entities is already indicated by the motion-path of the verb). The human animator who produced our upper baseline animations in 2010 still included some extra "pointing" to these locations, and so we included them in the script given to the human signer in 2012. In the feedback comments in 2012, some participants said: "Most verbs shouldn't end with the pointing of the finger (or direction) as the action already indicated that much," "too many endings were a pointing, it threw off my attention a lot," etc. What is interesting is that no participants criticized this in 2010, thus, when they saw a human signer performing this extra pointing movement, it felt more unnatural and warranted a comment at the end of the study.

6. CONCLUSIONS AND FUTURE WORK

This paper gives methodological guidance on the use of upper baselines in user-based evaluations of sign language animations. By replicating a past study and replacing the animated-character upper baseline in that study with a video of a human signer, we quantified how the evaluation scores collected were affected by this modification. This research has two key contributions: (1) This paper provides guidance for *future* sign language animation researchers who are designing a user-based evaluation study. They can make a more informed choice of which type of upper baseline to use for comparison in their study. (2) This paper provides guidance for readers of *previously published* studies who are trying to compare the results of studies that had used different upper baselines (animated-character or human-video). Given the results of our paper, it is easier for readers to understand how the scores might have been affected by the different study design.

Specifically, we examined four hypotheses in this study:

- **H1: Not supported.** Videos didn't get higher comprehension scores than our animated-character upper baseline. This indicates that an animated-character with proper movements can be an effective upper baseline for comprehension studies.
- **H2: Partially supported.** Changing to a video upper baseline didn't significantly affect the comprehension scores for the other stimuli, but a statistically significant difference between the "Model" animations and the "Uninflect" lower baseline animations in 2010 was no longer observed in 2012.
- **H3: Supported.** Human videos received higher Likert-scale subjective scores than an animated-character upper baseline.
- **H4: Should be split into two hypotheses** – for sequential collection of Likert-scale subjective evaluation scores during a task also involving answering comprehension questions (H4a) or during simultaneous side-by-side comparison (H4b)**.**
  - **H4a: Not supported.** Using a human video upper baseline did not depress the subjective Likert scale scores that participants assign to the other stimuli during part 1 of our studies. Perhaps this was due to signers being asked to also answer comprehension questions during part 1 (and thus were less attuned to the subjective animation quality) or perhaps the depressive effect on Likert-scale subjective scores doesn't occur during sequential stimuli presentation.
  - **H4b. Supported.** A significant depression was measured in the Likert-scale scores of the "Model" and "Uninflect" animations during the side-by-side comparisons.

In short, the results presented in this paper indicate that either form of upper baseline is potentially valid for use in a user-study to evaluate animations of sign language – there are merely some effects on the scores collected that must be taken into account when comparing the results across studies. Thus, researchers should consider the goals of their research and the specific aspects of sign language animation that they are focusing on when selecting an appropriate upper baseline. Researchers studying computer graphics issues relating to the visual appearance of a virtual human for sign language animations may wish to include videos of humans as an upper baseline – since this would serve as an "ideal" of photorealism. Further, researchers who wish to convey to a lay audience the overall understandability of their sign language animations (i.e., the current state of the art) may prefer using videos of humans as an upper baseline because it makes it easier to communicate to a lay audience the current quality level of their animations. Alternatively, researchers who are studying linguistic issues for sign language animations (e.g., sequencing of signs, the speed/timing of signs, the movement paths of the hands during certain signs, the timing of facial movements that relate to the signs, etc.) may find an animated-character baseline more useful to their research. For researchers who are not studying computer graphics issues related to the character's appearance, more useful data may be obtained from comparing their animations to an upper baseline of an animated character – thereby isolating the movement/timing from the appearance. For researchers who are not adjusting appearance aspects of a character, such a baseline may serve as their "ideal" of the correct movements and timing for a virtual character. The downside is that it is more difficult to convey to non-specialists the current quality level of their animations, because they are not being directly compared to a human video. Of course, researchers with enough time/resources to conduct a study with participants to compare a larger number of groups of animations and videos may wish to include both forms of upper baseline.

Researchers should also consider that while using a video upper baseline may yield evaluations that are easier for a lay audience to understand, it could lead to misconceptions. The human might do things that ASL animation technology will not be able to do in the next decade or two, e.g., automatically planning subtle emotional aspects of facial movement, automatically constructing complex 3D classifier predicate expressions, etc. Researchers using video upper baselines would need to explain these limitations to manage the expectations of a lay audience being presented their results.

Further, producing a human video that is a good upper baseline is harder than researchers may expect (see section 4). Depending on the specific linguistic phenomena that you are trying to keep constant across all of the versions of animations/videos shown in a study, the human performer's task is difficult to impossible. In our study, much work was required to assure that the human had identical sign sequencing, identical subject/object and pointing locations, and approximately similar overall time duration as the animations to which it was being compared. If a researcher needed to produce an upper baseline that also held constant some aspect of signing that is very detailed (e.g. precise millisecond timing of speed/pauses, exact height of the eyebrows, etc.), then asking a human to exactly perform this could be nearly impossible.

In future work, we want to explore the reason why no depressive effect was measured for H4a: Was it because the participants had also been asked to perform a comprehension task or because they were provided Likert-scale subjective ratings sequentially (not side-by-side)? A follow-up study could disambiguate this. When a depression of Likert-scale scores does occur due to a video upper baseline, we are also interested in determining if this could lead to a "compression" of the scores for the "middle" stimuli (animation being evaluated) and the lower baseline (as these two values are compressed downward toward the lower end of the Likert scale). If so, this would be undesirable because it could be more difficult to distinguish differences between the lower baseline and the animation being evaluated. Unfortunately, we could not address this issue in the current paper because the Likert-scale scores for Uninflect10 and Mode110 were already indistinguishably close (there was little room for them to get any closer in our 2012 study.) In future work, we would need to use as a starting point a study with lower baseline, our model, and upper baseline animations with a significant difference between all three cases in the Likert-scale scores. By replicating such a study, we could determine if the use of a video upper baseline led to this "compression" effect.

7. ACKNOWLEDGMENTS

This material is based upon work supported in part by the US. National Science Foundation under award number 0746556 and award number 1065009, by the PSC-CUNY Research Award Program, by Siemens A&D UGS PLM Software through a Go PLM Academic Grant, and by Visage Technologies AB through a free academic license. We would like to thank our advisor Matt Huenerfauth for his support and contributions to this paper. Jonathan Lamberton assisted with the recruitment of participants and the conduct of experimental sessions described in this paper.

8. REFERENCES

[1] Ahlberg, J., Pandzic, I.S., You, L. 2002. Evaluating face models animated by MPEG-4. In *I.S. Pandzic, R. Forchheimer (eds.), MPEG-4 facial animation: the standard, implementations and applications*, Wiley & Sons, 291–296.

[2] Bergmann, K. 2012. The production of co-speech iconic gestures: empirical study and computational simulation with virtual agents. Dissertation, Bielefeld University, Germany.

[3] Davidson, M. J., Alkoby, K., Sedgwick, E., Berthiaume, A., Carter, R., Christopher, J., Craft, B., Furst, J., Hinkle, D., Konie, B., Lancaster, G., Luecking, S., Morris, A., McDonald, J., Tomuro, N., Toro, J. and Wolfe, R. 2000. Usability Testing of Computer Animation of Fingerspelling for American Sign Language. *Presented at the 2000 DePaul CTI Research Conference*, Chicago, IL, November 4, 2000.

[4] Garau, M., Slater, M., Bee, S., and Sasse, M. A. 2001. The impact of eye gaze on communication using humanoid avatars. In *SIGCHI'01*, Seattle, USA. ACM, NY, USA.

[5] Gibet, S., Courty, N., Duarte, K., and Le Naour, T. 2011. The SignCom system for data-driven animation of interactive virtual signers: Methodology and evaluation. *ACM Trans. Interact. Intell. Syst.* 1, 1, Article 6 (October 2011), 23 pgs.

[6] Huenerfauth, M. 2006. Generating American Sign Language Classifier Predicates For English-To-ASL Machine Translation. Doctoral Dissertation, Computer and Information Science, University of Pennsylvania.

[7] Huenerfauth, M., Hanson, V. 2009. Sign language in the interface: access for deaf signers. In C. Stephanidis (ed.), *Universal Access Handbook*. NJ: Erlbaum. 38.1-38.18.

[8] Huenerfauth, M., Lu, P. 2012. Effect of spatial reference and verb inflection on the usability of American sign language animation. In *Univ Access Inf Soc*. Berlin: Springer.

[9] Huenerfauth, M., Zhao, L., Gu, E., Allbeck, J. 2008. Evaluation of American sign language generation by native ASL signers. *ACM Trans Access Comput* 1(1):1-27.

[10] Huenerfauth, M., Lu, P. 2010. Modeling and Synthesizing Spatially Inflected Verbs for American Sign Language Animations. In *Proceedings of The 12th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2010),* Orlando, Florida, USA. New York: ACM Press.

[11] Huenerfauth, M., Lu, P., and Rosenberg, A. 2011. Evaluating Importance of Facial Expression in American Sign Language and Pidgin Signed English Animations. In *Proceedings of The 13th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2011)*, Dundee, Scotland. New York: ACM Press.

[12] Kennaway, J. R., Glauert, J. R. W. and Zwitserlood, I. 2007. Providing signed content on the Internet by synthesized animation. *ACM Trans. Comput.-Hum. Interact.* 14, 3, Article 15 (September 2007).

[13] Kipp, M., Heloir, A., Nguyen, Q. 2011. Sign language avatars: animation and comprehensibility. In *H. Vilhjálmsson, S. Kopp, S. Marsella, K. Thórisson (eds.), Intelligent Virtual Agents* (Vol. 6895). Springer, 113-126.

[14] Kipp, M., Nguyen, Q., Heloir, A., and Matthes, S. 2011. Assessing the deaf user perspective on sign language avatars. In *Proceedings of ASSETS'11*, Dundee, Scotland. ACM, New York, NY, USA, 107-114.

[15] Lu, P., Huenerfauth, M. 2011. Synthesizing American Sign Language Spatially Inflected Verbs from Motion-Capture Data. *Second International Workshop on Sign Language Translation and Avatar Technology (SLTAT)*, in conjunction with *ASSETS 2011*, Dundee, Scotland.

[16] Lu, P., Huenerfauth, M. 2010. Collecting a Motion-Capture Corpus of American Sign Language for Data-Driven Generation Research. *Proceedings of the First Workshop on Speech and Language Processing for Assistive Technologies (SLPAT), The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2010)*, Los Angeles, CA, USA.

[17] Lu, P., Huenerfauth, M. 2012. Collecting and Evaluating the CUNY ASL Corpus for Research on American Sign Language Animation. Manuscript submitted for publication.

[18] Lu, P., Huenerfauth, M. (2012, in press). Learning a Parameterized Lexicon of American Sign Language Inflecting Verbs from Motion-Capture Data. *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies (SLPAT), Human Language Technologies: The 12th Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2012)*, Montreal, Canada.

[19] Mitchell, R., Young, T., Bachleda, B., & Karchmer, M. 2006. How many people use ASL in the United States? Why estimates need updating. *Sign Lang Studies*, 6(3):306-335.

[20] Pražák, M., McDonnell, R. and O'Sullivan, C. 2010. Perceptual evaluation of human animation timewarping. In *ACM SIGGRAPH ASIA 2010 Sketches (SA 2010)*. ACM, New York, NY, USA, Article 30, 2 pages.

[21] Schnepp, J. and Shiver, B. 2011. Improving Deaf Accessibility in Remote Usability Testing. In *Proc. of ASSETS'11*, Dundee, Scotland. ACM, New York, 255-256.

[22] Schnepp, J., Wolfe, R. and McDonald, J. 2010. Synthetic Corpora: A Synergy of Linguistics and Computer Animation. *Fourth Workshop on the Representation and Processing of Sign Languages, LREC 2010.* Valetta, Malta.

[23] Schnepp, J., Wolfe, R., Shiver, B., McDonald, J. and Toro, J. 2011. SignQUOTE: A Remote Testing Facility for Eliciting Signed Qualitative Feedback. *2nd Int'l Workshop on Sign Language Translation & Avatar Technology,* Dundee, UK.

[24] Traxler, C. 2000. The Stanford achievement test, 9th edition: national norming and performance standards for deaf & hard-of-hearing students. *J Deaf Stud & Deaf Educ* 5(4):337-348.

[25] VCom3D. 2012. Homepage. http://www.vcom3d.com/