An Automated Technique for Real-Time Production of Lifelike Animations of American Sign Language

John McDonald[1], Rosalee Wolfe[1], Jerry Schnepp[2], Julie Hochgesang[3],
Diana Gorman Jamrozik[4], Marie Stumbo[1], Larwan Berke[3],
Melissa Bialek[1], Farah Thomas[1]

[1]DePaul University, Chicago, IL, {jmcdonald, wolfe}@cs.depaul.edu,
{mstumbo, mbialek1, fthomas}@mail.depaul.edu
[2]Bowling Green State University, Bowling Green, OH, schnepp@bgsu.edu
[3]Gallaudet University, Washington, D.C., {julie.hochgesang, larwan.berke}@gallaudet.edu
[4]Columbia College Chicago, Chicago, IL, dgorman@colum.edu

ABSTRACT

Generating sentences from a library of signs implemented through a sparse set of key frames derived from the segmental structure of a phonetic model of ASL has the advantage of flexibility and efficiency, but lacks the lifelike detail of motion capture. These difficulties are compounded when faced with real-time generation and display. This paper describes a technique for automatically adding realism without the expense of manually animating the requisite detail. The new technique layers transparently over and modifies the primary motions dictated by the segmental model, and does so with very little computational cost, enabling real-time production and display. The paper also discusses avatar optimizations that can lower the rendering overhead in real-time displays.

1. INTRODUCTION

One approach to sign language synthesis creates animations using a library of signs [**1**] [**2**] [**3**] [**4**] [**5**]. The signs are procedurally combined into phrases and sentences using tags akin to annotation tags used in linguistic analysis of sign language. In a sign language, the phonological parameters of lexical items can change depending on their usage, and new motions can be layered, depending on structural features of the discourse such as posing questions or using constructed dialogue [**6**] [**7**].

As distinct natural languages with individual grammars and structures, sign languages are used as primary languages by the Deaf worldwide. Sign synthesis is a promising method both for Deaf-hearing communication through computer translation, and also for their study and preservation. To address these opportunities, a variety of projects have striven towards the goal of synthesizing animations of sign language.

Such efforts have exploited a range of different animation techniques from key-frame animation [**2**], to motion capture [**5**], to procedural motion synthesis [**8**]. A more complete review of the current literature on sign synthesis can be found in [**9**]. Many center on building sign syntheses from libraries of signs that are either transcribed by artists or captured from live sign performance. Sign language synthesis brings its own unique challenges to animation, as it often imposes rigorous requirements both

on the fine detail of the motion, as in the fingers for handshapes, and on the flexibility of the motion for editing and recombination [B. Remi, Personal Communication, July 12, 2013, London, U.K.].

Both key-frame and capture techniques are expensive and time-consuming. Motion capture requires specialized skills to clean up inherently noisy data and is difficult to edit. Sign synthesis through key-frame animation requires artists who have highly trained eyes for human motion. Both techniques require experts who are also highly versant in the structure of the target sign language. Because of this, some procedural techniques are enormously appealing as they have the potential to augment and speed the work of trained animators.

The rest of this paper is organized as follows: Section 2 introduces the problem of robotic motion as it occurs in signing avatars. Section 3 describes a new automated system of spinal movement to alleviate robotic motion. Section 4 discusses optimization to aid in real-time rendering of an avatar. Section 5 covers an experiment to evaluate the spinal system and avatar optimizations, and section 6 analyzes the results. Section 7 draws conclusions and suggests future directions.

2. CHALLENGES

One of the great challenges of character animation in general is the management of animation keys. Manipulating both the poses of an avatar and the timing of movement are greatly simplified by limiting the number of keys that describe a motion. This is especially true in sign synthesis, where modifying animation keys based on linguistic rules is easier when there are fewer keys. Such sparse animation data correlate well to the segmental structure of the Signed Language Phonetic Annotation (SLPA) parameters described in [**10**].

The resulting ease of manipulation is, however, counterbalanced by a lack of realism in the animation. The very simplicity of a sparse set of keys afforded by the linguistic parameters limits the natural subtlety that one can achieve in the motion. Naturalness has long been a goal in all areas of character animation and sign synthesis is no exception. Without the subtle motions of human signing, an avatar can become highly robotic. Naturalness in motion is extremely important for the acceptability of synthesized sign.

Robotic signing, the equivalent of robotic synthesized speech, can be off-putting, can distract the viewer from the meaning, and can tire viewers in long discourse. In previous studies [**11**], participants had no trouble understanding the generated sentences, but since the fine details that characterize natural motion were missing, the sentences elicited comments such as “She looks stiff,” or “That’s awkward,” or “I wouldn’t sign it like that.”

Ironically, linguistic tags offer little help for this problem, because their purpose is to abstract structure from the fine details of motion. The SLPA model separates sign into postural and transforming segments, analogous to the keys and interpolations of computer animation, but a naïve implementation of the model results in animation which lacks the fine details of human motion necessary for naturalness and believability in an avatar. The same abstraction that is essential for creating signed sentences as animation contributes to the awkwardness of the production.

Unfortunately, basing animation on the abstraction of the SLPA model is only one contributor to awkwardness in sentence production. The following section examines these factors in depth.

2.1 Contributing Factors to Robotic Motion

Animations procedurally created from a library of signs based on linguistic parameters naturally create a sparse distribution of keys, which makes it easy to change the animations based on linguistic context. However that same sparseness of keys can easily lead to robotic motion. The paucity of keys can come from a variety of factors, partly due to the fact that the SLPA model is underdetermined with respect to generating animations. Other factors have to do with ignoring kinematics and animation principles. The following paragraphs examine each factor in turn.

2.1.1 Underdetermination in phonetic models

Since the extralinguistic motions are seldom specified, creating animations by relying exclusively on geometric interpretations of the designations in the SLPA model will create animations where some joints of the avatar's body remain stationary. However, when a human produces the sign, these same joints would be in motion.

For example, the ASL sign THINK only specifies the behavior of the strong hand (index finger extended and thumb and other fingers fully flexed) and contact between the index finger and the ipsilateral temple on the strong side of the body. The contact information determines the behavior of the strong arm. However, because there is no linguistically significant nonmanual signal (NMS) on the spinal column, the torso and neck will not move during the production of the sign. The resulting animation is a signing “pole with arms” which does not have a natural appearance.

2.1.2 Lack of small-scale motion

In real life no part of the human body is ever truly stationary. Perlin addressed this by adding small amounts of random motion to the joints to keep the avatar alive [**12**]. However, random motion added to animations of sign language can interfere with the subtleties in the displayed message. Moreover, the parameters of the noise must be carefully chosen or the avatar will tremble or jerk unrealistically.

2.1.3 Lack of anticipation, follow-through and secondary action

Disney animators developed a set of the core principles of animation in the 1930s that did more than create motion that roughly adhered to the laws of physics [**13**]. These principles also communicated a character’s motivation and action to an audience in a convincing manner. The principles emphasized the most important parts of the animation so that viewers could easily perceive and follow the action of the story and accept the animations as lifelike and natural. Three of these core principles of animation are particularly applicable to the generation of sign, namely anticipation, follow-through and secondary action.

In anticipation, a preparatory motion shifts the animated component slightly in the opposite direction from the intended motion before the intended motion takes place. The function of anticipation is twofold. First, it is a visual cue that alerts viewers to the upcoming motion. Since viewers are expecting to see the motion, they will more likely perceive it. Second, since the anticipatory motion draws the moving component in the opposite direction from the intended one, the overall duration and trajectory

is extended, thus emphasizing and drawing more attention to the motion. Figure 1 shows three frames from an animation of the ASL sign SUCCESS. The main motion of the head is upwards. However, there is a slight anticipatory downward motion.

Figure 1: Neutral pose, anticipatory motion, followed by intended motion in the ASL sign SUCCESS.

The animation principle of follow-through can be thought of as a complement to anticipation; it happens at the end of the movement instead of the beginning and serves a similar function to anticipation. Follow-through extends the motion, overshooting the intended final location and bouncing back. Just as anticipation prepares viewers for the motion they are about to see, follow-through visually reinforces the motion after it has occurred.

Similar to anticipation, the duration and trajectory of the entire motion is extended. Figure 2 is a motion graph depicting wrist height in the ASL sign NOW. The initial and final postures are labeled in the graph. In the follow-through motion, the wrist position drops below the final posture and rises again to settle there. This figure is a trace of the motion of the hand in a sign transcribed by an expert animator based on examples from multiple corpora, and was reviewed for legibility by members of the Deaf community in the Chicago area.

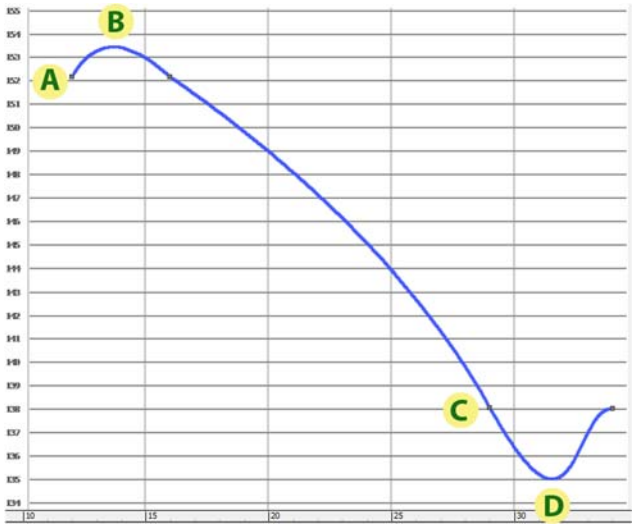

Figure 2: Vertical motion in ASL sign NOW.
Initial posture is at A. B is the anticipation.
Final posture is at C. D is the follow-through.

The last animation principle, secondary action, describes movement that happens in addition to the intended motion. In ASL it occurs in two-handed signs like WORK and AGAIN where there is contact between the hands. This produces movement, however subtle, in the weak hand. Figure 3 shows a motion trace close-up of the wrist in the weak hand during the final posture of the ASL sign AGAIN.

Figure 3: Secondary action in weak hand in the ASL sign AGAIN.
Red boxes show weak hand/wrist motion.

These principles were first applied to computer animation in the late 1980s in computer animated short subjects such as *Luxo, Jr.* when Disney-trained artists began using computer animation packages [**14**]. Although features such as *Cars* and *Ratatouille* are appealing and effective in conveying motion and emotion through computer-assisted hand animation, they require an enormous amount of attention to animation detail and are expensive to produce.

For sign synthesis, we must remember that these principles were generally exaggerated quite extensively for cartoon animation; however, they do describe physical processes that our bodies perform. Therefore, while not exaggerated in the way that the Disney artists originally conceived, these aspects of motion are still critical to the lifelike appearance of the movement.

2.1.4 Overuse of inverse kinematics.

A technique that facilitates quick results when animating by hand is inverse kinematics (IK). IK allows the animator to specify the location of the last component, or end-effector, in a hierarchy. The rotations of each component between the first and last are automatically computed. [**15**]. It is possible to specify various constraints in order to achieve a more natural body pose. While this technique is faster than manually adjusting each component in the hierarchy, problems arise when attempting to use IK for upper body motion. The result is often a marionette-like motion where the trajectory of the end-effector is linear, rather than a natural trajectory of an arc. Figure 4 compares two implementations of a deictic point in ASL. The avatar on the left uses forward kinematics, not IK, for the interpolation, and the hand follows an arced trajectory. The elbow position is relatively stationary. The avatar on the right uses IK to create the motion. To maintain the linear trajectory, the elbow skates backwards in an awkward manner. The difference is marked by the yellow highlight on the elbow.

Figure 4: FK versus IK.

In addition, IK techniques can often force joints into unrealistic configurations due to overly constraining the hand to a fixed position, e.g. vertical in front of the body. Hand positions are always somewhat approximate rather than geometrically perfect, and, especially in natural signing, comfort and efficiency will modify the orientation of the hand enough to make it comfortable but not too much so as to lose the meaning of the sign.

2.1.5 Linear Interpolation of Joint Angles

Linear interpolation causes unnatural and abrupt changes in velocity not found in nature. Linear joint interpolation leaves no room for smooth acceleration or deceleration that are the hallmark of natural human motion derived from muscular forces.

This phenomenon will appear in almost every animated sign, yielding abrupt changes in speed and direction that give a very robotic feel to the motion.

2.1.6 Synchronicity

In the transition after a hold, not all body joints begin moving at the same time [**16**]. When turning to the side, as in a role shift, the eyes move first, followed by the head, neck, shoulders and spinal column. Further, there is a tendency for the hand to complete its transition to a new handshape well before the arm arrives at its new location as can be seen in Figure 5 which depicts the ASL sign INFORM. Unfortunately, the SLPA model does not capture this internal timing in a sign. A naïve interpolation between postures in a sign results in a synchronous motion that is perceived as being stiff.

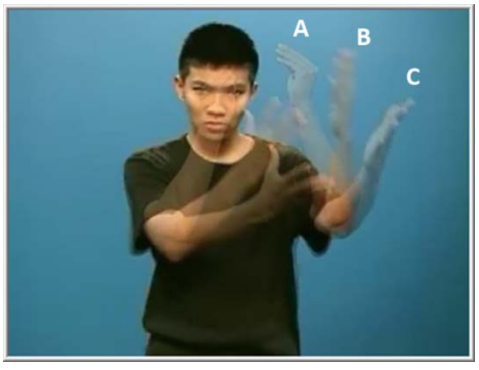

Figure 5: ASL sign INFORM:
Final handshape at B; final position at C [17**].**

2.2 Alternatives to reducing robotic motion

It is possible to overcome these challenges through careful manual animation, but this is a labor intensive process requiring highly skilled artists. Moreover, the animators must have some knowledge of sign language; a rather rare combination. Results from such skilled animators are highly realistic, but are time consuming and very expensive. Moreover, unless the project is structured very carefully, the linguistic components of the discourse can be buried in the kinetic motions produced by the animator as they layer on the realistic variations characteristic of human motion.

The separation of linguistic motion from kinetic motion is absolutely critical if the synthesized sign will be able to change according to morphological rules. It also is important if the animations will be used for linguistic testing, analysis and verification. The problem is that the motion in animated signs is influenced by both linguistics and the natural kinematics of human motion. The linguistic components of a sign and its containing discourse determine the gross movements of the sign, but the kinetic motions of the body determine the naturalness of the sign's production.

This separation can be achieved by using procedural methods to aid animators in creating not only individual signs, but also in combining and modifying them to build longer discourse. Our goal is to infer as much as we can by drawing on the knowledge of human kinematics to produce lifelike, convincing motion that adheres to the SLPA model of American Sign Language.

Using procedural methods shortens the development cycle by aiding animators. We propose a new technique as part of a procedural/hybrid system, which still relies on hand animation for the sparse keys following the segmental structure dictated of the SLPA model, but addresses several problems mentioned in the last section through procedural means. When creating an animation of an individual lexical item for the library, the linguistic components of the sign are recorded for each key to facilitate later modification and recombination. Longer discourse is then built by applying procedures dictated by linguistic tags to drive the automatic recombination of hand-animated signs. Finally, procedural techniques add the kinetic motions.

3. A NEW AUTOMATED SYSTEM FOR SPINAL MOVEMENT

The first part of the proposed solution deals with the lack of motion in the spine as the avatar's arms move according to the SLPA model. The proposed extension will work with any limb IK solution, either iterative or analytic [**18**], and is applied before running the IK solution itself. It assumes that the spine and collar bones of the model are not involved in the IK chain.

While iterative IK systems do allow for the extension of the IK chain through the spine to the hips of the model, these systems become increasingly slower and more difficult to control as the chain becomes longer. Other systems separate the spine, shoulder and arm into separate IK chains that can be coupled to let the spine react to the arms motion more automatically [**19**]. While they do avoid longer individual IK chains, such systems markedly increase the complexity of the computation.

The movement of the arms and torso have different purposes in the human body, and correspondingly in animation. The motion of the torso supports arm movement and usually precedes it. Thus the arm motion will follow the torso temporally. Long IK chains tend to remove this timing sequence, and in fact

give reverse impression that the arm is pulling on or leading the torso, resulting in the aforementioned marionette effect. Moreover, long IK chains produce a complex collection of joint rotations that cannot be easily staggered temporally to reduce synchronicity.

Our application had three requirements that led to a more direct solution:

1. Our IK system [**18**] is an analytic solution tailored to the human arm. We desired an analytic spine solution as well, so as to avoid an increased computational overhead.
2. We desired a solution tailored to the specific type of spine motion discussed above, rather than a more general spine solution. It was this fact that made requirement 1) easier to satisfy.
3. We wanted a solution that allowed additional independent spine control. The solution serves as a starting point for the artist to speed workflow, but should not inhibit other scripted or automated spine motions such as role shifts.

The last of these is critical. The new extension is intended to be an assist to animators, rather than to assume sole control over the spine. Animators can still control the spine independently of this IK extension. It provides a good first estimate for the spinal movement through an automated system that will enliven the spine, eliminating one of the causes of robotic motion.

With an IK chain that solves for joint angles from the shoulder to the hand, the question becomes, "What kinds of movement will be most natural to use as a sensible default for the spine?" Humans move their spines in a myriad of purposeful ways, and ASL is no exception to this as spinal nonmanual signals are an integral part of the language.

In the absence of purposeful spinal motion, which would normally be animated separately by the artist, one of the main functions of the spine is to extend and enhance the range of motion of the arm. Therefore, this new extension uses the reach of the arm to cue the spinal motion. The greater the distance of the reach, the more the spine will bend to assist the arm. Additionally, sequencing the animation keys involved will eliminate the perception of the arm dragging the spine in the style of a marionette. This will be discussed in detail in Section 3.5.

3.1 Movement of the Spine

When using the motion of the spine to assist the reach of the arm, the most direct solution is to use the position of the end-effector's target from the IK solution. The farther that target is from the trunk of the body the more the spine/shoulder will bend to assist the motion, and the farther the target lies across the body, the more the spine will twist to facilitate the reach. Thus the desired position of the hand will be assisted by both a bending action that increases the range of the hand's reach, and a twisting action which facilitates the lateral movement of the hand across the body.

The magnitude of this effect is highly dependent on the distance of the arm reach, but when the hands are within the more proximal regions of sign space, there will be very little effect on the spine. This displacement should, however, increase rapidly when the hands begin to reach beyond the zone of comfort, and should taper off smoothly as we approach the limit of the spine's rotation.

The rotation of the spine will also depend equally on both hands, so the computation must consider the positions of both the left and right IK targets symmetrically. Moreover, the interaction of the hands with the spine differs for the torso bend and the torso twist. In fact, there are at least three distinct influences which must be blended smoothly:

1. Reaching with one or both hands will bend the torso towards the target(s) and the effect will be greatest when both hands are directed towards the same target. Thus, the bend is additive for the two arms.
2. Lateral motion of a hand across the body will cause the torso to twist and rotate the shoulder towards the direction of motion. This effect is also additive. If the arms go in the same direction, the effect will be amplified, and if in opposite directions, they will cancel each other.
3. In addition to the torso bend, when one hand reaches out from the body, the torso will also twist to move the corresponding shoulder towards the target. This effect is different from the previous two in that equal motion between the two arms will cancel out the effect.

To build the extension, we will assume that the coordinate system on the avatar's torso is oriented with the *x*-axis towards the right, the *y*-axis forward and the *z*-axis up. Consider the right and left articulator targets A_R and A_L, and the corresponding neutral shoulder positions S_R and S_L. We build displacement vectors from each shoulder to the corresponding target as

$$V_R = A_R - S_R$$
$$V_L = A_L - S_L$$

Let V_{reach} be the average of these two vectors:

$$V_{reach} = \frac{(V_R + V_L)}{2}$$

To calculate both the bend angle and direction, we must rotate the torso in the direction of V_{reach}. The bend angle is computed by first normalizing by the arm's length, and then applying a smoothstep function [**20**] to both mute its contribution near the body and smoothly clamp the effect at the end of the arm's range

$$\theta_{\text{bend}} = c_0 * \text{ smoothstep}\left(\frac{|V_{reach}|}{L_{arm}}\right)$$

where c_0 is a tuning parameter that that controls the magnitude of the effect, and

$$\text{smoothstep}(x) = 3x^2 - 2x^3$$

The axis of rotation is given by the cross product of the arm direction with the vertical axis vector Z = <0, 0, 1>.

$$a_{bend} = \frac{Z \times V_{\text{reach}}}{|Z \times V_{\text{reach}}|}$$

The matrix for a rotation of θ_{bend} about the axis a_{bend} is formed in the usual manner, and if the avatar uses Euler angles for joint control, the angles can be extracted as in [**21**].

In the case of a reach up or down, there will be no effect and we will rely on the shoulder to extend the reach. One could extend this easily to arch the back for such a reach, but that is beyond the scope of the present work.

For the twist, we will need a more involved combination of the two wrist displacement vectors since the twist depends on the *x* and *y* positions of the hand differently.

$$V_{twist} = \left\langle V_{R,x} + V_{L,x},\ V_{R,y} - V_{L,y},\ 0 \right\rangle$$

This calculation adds in the *x*-direction and cancels in the *y*.

The twist angle is obtained by normalizing V_{twist} by twice the arm length, since the *x*-term is a sum rather than an average, and taking the dot product with a constant weighting vector V_w

$$V_w = \left\langle X_w, Y_w,\ 0 \right\rangle$$

whose components determine the amount of twist contributed by both the lateral and distal movement of the arms. We then pass this dot product through the same smoothstep function as before to apply the effect gradually when the hands are in close proximity to their respective shoulders, and to smoothly clamp the effect at the far range of motion.

$$\theta_{twist} = c_1 \,\text{smoothstep}\left(V_w \bullet \frac{V_{twist}}{2L_{arm}} \right)$$

Here again, c_1 determines the overall size of the twist.

3.2 Spreading the Effect Over the Spine

The human spine is a chain of 24 articulating vertebrae which are seldom modeled independently in an avatar. Most often the articulation of the spine is approximated by three or more conventional rotational joints, which correspond to the major subdivisions of the spine into the cervical, thoracic and lumbar sections [**22**]. Our avatar has three joints in its back which correspond roughly to the base of the lumbar, the articulation between the lumbar and the thoracic and finally an articulation about 2/3 of the way up the thoracic region, creating lower, middle and upper spine joints. The cervical portion of the spine is subsumed by the neck joints which are not considered here.

For basic bending and twisting motions, the human vertebrae rotate in concert. In the model, we automate the division of the desired rotation among the set of joints in the spine by weighting each joint's rotation according to the relative range of motion in the corresponding subdivision of the spine. For these ranges, see [**23**]. Thus, for example, in the spine bend, we would set

$$\theta_{lower} = w_0 * \theta_{bend}$$
$$\theta_{middle} = w_1 * \theta_{bend}$$
$$\theta_{upper} = w_2 * \theta_{bend}$$

Where the weighting coefficients w_i sum to 1. These coefficients will differ for the bend and twist angles since the ranges of motion in the spine are different for each of these directions. It is also important to note that for the torso bend, these weights are easiest to apply if Euler angles are used for the torso rotation since they nicely correspond to the usual ranges of motion cited in most medical texts such as flexion-extension, bending and rotation (twist).

3.3 Shoulder Movement

The movement of the human shoulder is created by two primary articulations, the acromioclavicular and sternoclavicular joints that connect the clavicle, scapula and the rib cage [**22**]. It is most often approximated in an avatar using a rotational joint placed at the intersection of the spine and the neck, and having two primary axes of rotation. The vertical rotation allows the raising and lowering of the shoulder, while the anteroposterior rotation moves the shoulder forwards and backwards. While the clavicular joints also have a small amount of rotation along the longitudinal axis to twist the shoulder, this is often ignored in favor of incorporating the motion directly in the shoulder joint itself.

The computation for automatically moving the shoulder is somewhat simpler than for the torso, due to the nature of the joint. Reaching out, in the *y*-direction, from the model will rotate the shoulder forward, while reaching up or down will move the shoulder accordingly. The computation follows from the same displacement vectors V_R and V_L as before, except that we consider the *y* and *z* coordinates separately for the two movements. Each is normalized and is passed through a smoothing function

$$\theta_{R,raise} = c_2 * \text{smoothstep}\left(\frac{V_{R,x}}{L_{arm}}\right)$$
$$\theta_{R,lateral} = c_3 * \text{smoothstep}\left(\frac{V_{R,y}}{L_{arm}}\right)$$

where the c_i determine the range of the shoulder motion.

3.4 Results of the Extension

Figure 6 displays an example of Paula, our avatar, reaching across her body. The first image is the raw posture derived from the original IK limb computation. The second displays the more natural posture achieved by the spine and shoulder extensions described above.

Figure 6: Reaching posture of the model. Image at left is without the IK extension. Image at right is with the extension.

3.5 Timing

The last two sections addressed one of the most important aspects of avoiding robotic behavior in the model. Maintaining motion in all of the joints in the spine alleviates the "pole-with-arms" syndrome and gives the avatar more a more life-like appearance. There is, however, one aspect that is made somewhat worse by this solution when used with a set of very sparse keys. All the avatar's joints will start moving at the same time and come to rest synchronously. Asynchronous timing of linguistic components of sign has been investigated [**24**]. The work presented here extends these timing considerations to extralinguistic motions that occur in concert with a signer's arm movements.

The human body, as a complex system of joints, bones and electrical impulses simply never moves with such precision and coordination. The joints begin moving in progression, separated by such short intervals that the effect is not easily noticeable. However, if this progression is missing, the avatar will appear robotic. The progression we use arose from a human motion study with an expert mime [**25**], and is:

1. The eye-gaze shifts to a focus point
2. The neck turns to follow
3. The spine and hips shift
4. The clavicle moves
5. The arm moves

In terms of animation timing, these steps may be separated by only one or two frames, but the difference is subtly noticeable and can make a large difference in the naturalness of the motion. In fact, for IK systems, this progression is one of the main causes of the marionette effect mentioned earlier where the arm appears to drag the entire upper body. We are so used to the progression, that in its absence, we tend to perceive that the arm is in fact leading the rest of the body.

The timing solution itself is a simple heuristic, but requires special handling in an animation system. Coming into or out of any posture we simply need to stagger the spine, shoulder and arm joints with a small amount of time between key frames. At 30fps, we would want at least one or two frames

between each key. For example, while the animator intends that the motion begin, for example, at frame 45, the system may actually start the spine motion at 43, the shoulder at 45 and the arm at 47. This can be done transparently if the keys in the animation are sparse, but is more difficult with many, tightly packed keys. The resulting effect is to break up the synchronicity of the movement and give it a more natural muscular appearance.

Going further, we can break this progression down into a finer granularity. The motion of the spine itself is not completely synchronous either, because the hips tend to move before the upper spine as they initiate the motion. So, in a large spinal motion such as occurs in a role shift for constructed dialogue, the shift will appear much more natural if the spinal joints are staggered slightly. Since there is a much tighter relationship between the spine joints, the corresponding key distribution will also be tighter.

3.6 Handling Held Postures

An additional contribution to robotic behavior from the SLPA model is the representation of postural segments. While the model does not dictate a complete cessation of motion, it does not, and should not, dictate the autonomous motion that occurs while joints are at rest, however momentarily. Likewise, joints that are not involved in the sign, and have no linguistic specification, will remain held. An example of this is the set of joints in the weak hand and arm for one-handed signs. The problem is that the human body is simply incapable of coming completely to rest, unlike a mechanical robot. An avatar perfectly at rest is perceived as a still image and not a 3D animation.

To address this, we add a small component of random movement (noise) to each of the major joints in the avatar as in [**12**]. This causes the avatar's joints to vary slightly over time. The challenge of applying this technique is in determining the amplitude and the frequency of the noise. Noise with high amplitude and/or frequency will give the avatar a very jittery or shaky appearance and will be worse than no noise application at all. The solution is to tune the noise so that its effect is just above the threshold of visibility, both in amplitude and frequency. We used the following guidelines from human anatomy to configure our noise:

1. The amplitude of the noise should decrease for more distal joints since the muscles are generally smaller.
2. The frequency of the noise should increase for more distal joints. This can be seen, for example, in the trembling of an extended hand as one attempts to hold it still.
3. The amplitude of the noise will be greater when a part of the body is not under purposeful control. For example, in a one-handed sign, the strong hand will have random motion with much smaller amplitude than the weak hand.

4. OPTIMIZING THE AVATAR FOR REAL-TIME DISPLAY

ASL synthesis places unique burdens on avatars due to the high fidelity required to communicate the full range of ASL's expressiveness. One of the primary challenges is the range of facial NMS which form key linguistic components in ASL at all levels. In fact, facial NMS can radically change the meaning of individual lexical items as well as whole phrases in addition to their more universal role in human communication to express affect.

Because of this, the fidelity of the avatar's facial model is paramount, and non-realtime ASL rendering systems will often ironically use more polygons for the face, mouth and tongue than they do on the arms and hands of the model. Consider the initial composition of the avatar that has been long used in our project for offline rendering. The head, hair, teeth and eyelashes contained over 70% of the avatar's polygons. In particular, the eyelashes are important for facial NMS as they emphasize the eyes.

Object	Polygon Count
Eyelashes	**17,356**
Tongue and Teeth	**8,929**
Hair	**8,088**
Head and Face	**6,180**
Sweater	8,300
Hands (each)	3,922

Table 1: Polygon counts for main avatar components

Unfortunately, such complexity presents challenges for real-time display, both in the raw number of polygons and in the number of discrete sub-objects contained in the model. Real-time systems are best optimized when large collections of polygons can be pushed to the renderer all at once [**26**].

Our goal was thus to optimize the avatar for real-time display, but without any noticeable reduction in fidelity. Our first effort, which focused on the model's hair, drives home the restrictiveness of this goal. The hair seemed a promising candidate for optimization because it is not linguistically significant and contains 12% of model's polygons. We attempted to optimize the polygonal meshes by merging the closest vertices using a standard mesh decimation tool. Although this optimization eliminated 6,000 polygons, the results were not judged aesthetically acceptable. The profile of the hair had a jagged appearance (Figure 7). Since we were unable to find an acceptable balance between polygon reduction and appearance we decided the hair should be left alone and turned our attention to the eyelashes, the smallest objects on the model, but which used the highest percentages of the polygons.

Figure 7: Hair before and after optimization

4.1 Optimizing eyelashes and teeth

Eyelashes are important linguistically, because the eyes are involved in a large number of NMS, and eyelashes emphasize the perimeters of the eye apertures, making them more visible. We initially modeled them as a set of individual extruded 3D tubes. This was satisfactory in an offline rendering environment, but not practical for real-time rendering. In the current avatar, the thousands of eyelash polygons are concentrated in a very small portion of the model that in turn occupies a limited area of the final render.

To optimize the eyelashes we replaced the individually modeled lashes with four curved surfaces and used texture and opacity mapping to convey the appearance of individual lashes. The curved surface followed the contour of the original lashes (Figure 8) and allows the lashes to be seen from all angles. A flat surface would suffer the disadvantage of billboarding where the object will disappear when seen edge-on. To create realistic variation in lash length and thickness, an artist drew an opacity map that determined which parts of the surface would be visible.

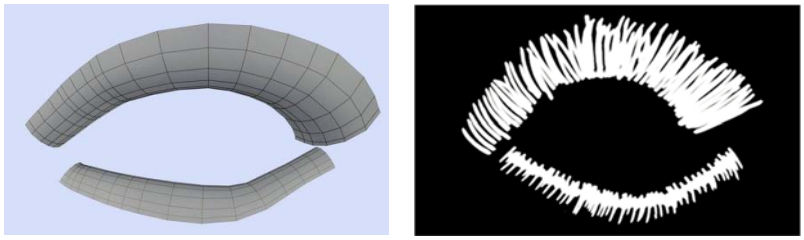

Figure 8: Geometry and opacity map of optimized eyelashes

The revised version of the eyelashes had only 660 polygons in comparison to over 17,000 polygons in the initial version. Additionally, the revised model replaced the somewhat regular lashes with ones that were more natural in appearance (Figure 9).

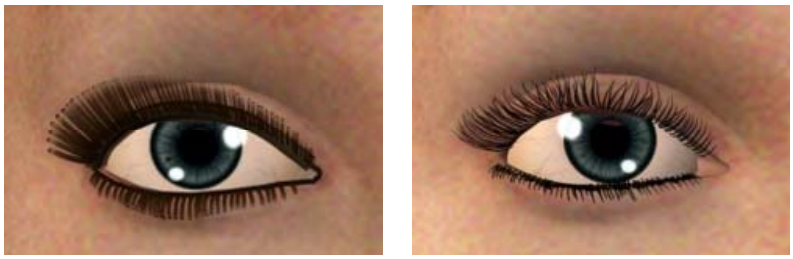

Figure 9: Eyelashes before and after optimization

Similar to the eyelashes, the teeth contained a high number of polygons in a small, hard to see area with a small display area in the final render. The newly optimized version of the teeth is a curved surface mimicking the contours of the original teeth model (Figure 10). Instead of attempting to model each tooth as in the original we created the appearance of teeth through the application of a texture map and an opacity map. Both maps were created from a composite of renders from the original model.

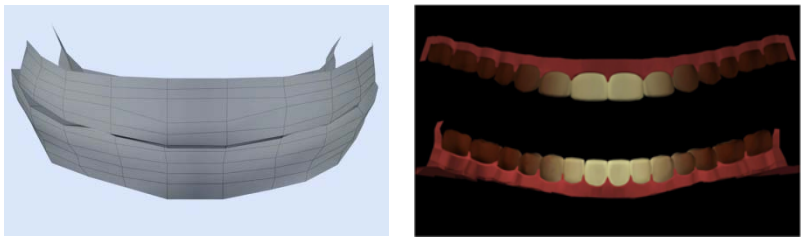

Figure 10: Geometry and texture maps of optimized teeth

Through the optimization process we reduced the number of polygons in the teeth from over 8000 to 440. For normal viewing purposes the optimized teeth appear essentially the same as the originals (Figure 11).

Figure 11: Teeth before and after optimization

4.2 Lighting

Careful choices in lighting techniques can help a viewer perceive the avatar more clearly as well as speed up rendering times. Even lighting can enhance facial features and thus the legibility of facial NMS. Unfortunately, each additional light places increased computational burden on a real-time rendering system. The lighting calculations for our avatar are based on the Phong Illumination Model [**26**] which incurs a modest computational cost that increases linearly with the number of lights. Initially, we used six lights, which provided even lighting, but slowed real-time rendering rates.

To address this, we revisited a technique in practical portrait photography called “three point lighting”, which removes unflattering shadows from the subject’s face, while emphasizing the facial features that are key to NMS. This system employs the use of a fill light and key light to illuminate the subject directly, and a third light to either provide subtle highlights or emphasize the subject’s silhouette. Usually the key light is placed in front of the subject to serve as the primary illumination. A fill light placed to the side of the subject eliminates distracting shadows, and the third light is situated as an accent to open or close specific shadows [**27**].

For our avatar, we found three lights insufficient due to the lack of simulated ambient light in our rendering system. To compensate, we added a fourth light. Figure 12 contains sample renderings of the original and optimized lighting. A nice side effect of the new setup was the removal of the line of symmetry previously cast along the vertical axis of the face.

Figure 12: Comparison of lighting setups, at left initial lighting and at right optimized lighting

We achieved a final optimization by ensuring that only the key light cast shadows, thus removing the time and memory costs of additional shadow computations.

These optimizations provided a reduction of polygons comprising the head, hair, teeth and eyelashes. These originally comprised 70% of the avatar's polygons and now comprise less than 30%. In addition, the complexity of the lighting rig was reduced by a third. The optimized avatar appeared in an evaluation, described in the next section, as a check on its fidelity.

5. EVALUATION

The new spine technique relieved artists from the time-consuming task of manually animating the spinal column. But it remained to be determined if the automated spine movement enhanced either the legibility or acceptability of generated ASL. We also wanted to know if the newly optimized avatar would be received favorably. We conducted a study to address two research questions:

- Does the automated spine system generate movement that is clear, correct, understandable, and natural?
- Does the addition of movement from an automated spine system produce animation that is preferable to animations without it?

The first question evaluates the fidelity of the polygon and lighting optimizations as well as the naturalness of the automated spinal movement system. The second question evaluates the efficacy of the underlying mathematical model for spinal motion. This study was reviewed and approved by three university Institutional Review Boards (DePaul IRB #RW071813CDM, Columbia College IRB #2014-00167, Bowling Green State University IRB #507016-2).

5.1 Stimuli

The test stimuli were sentences from a complete story about a childhood memory as listed in Figure 13. Figure 14 lists the five sentences in the stimuli as glossed ASL. A complete story is necessary to provide context for such linguistic processes as indexing, agreement verbs, and role shifts. Without the proper

context, these lose their meaning. All of the sentences were grammatically correct. All verbs were conjugated and all NMS were present, including syntactic markers. In addition, the sentences contained appropriate pragmatics.

A panel of language experts reviewed several versions of the sentences before approving their final form. The panel consisted of three nationally certified interpreters and two native ASL users, including a linguist.

> I remember when I was small, I lost my little toy car. I looked everywhere for it. I was so sad. But, surprise! My brother bought me a new one. It was big and blue. Blue is my favorite color. I was so happy!

Figure 13: Story for Evaluation Study (English)

1. LOOKING-BACK REMEMBER ME SMALL, HAVE LITTLE T-O-Y C-A-R.
2. I LOST WHERE. I SEARCH, SEARCH, CAN'T FIND -- SAD.
3. BUT SURPRISE -- MY BROTHER GET NEW C-A-R, BIG, BLUE, GIVE-ME.
4. INFORM-YOU BLUE MY FAVORITE COLOR.
5. FINALLY HAVE C-A-R AGAIN -- HAPPY!

Figure 14: Story for Evaluation Study (glossed)

We generated two versions of each sentence in the story. The control version did not have the automatically-generated spinal movement and the experimental version did. All of the manually-animated poses were present in both versions. If the artist had included an overt spinal pose, it was included in both versions of the animation. In addition, both versions included noise, so the spine was not entirely stationary, even in the control stimuli. Only the automated aspects of spinal movement were missing from the control stimuli.

5.2 Participants

Participants were all adults that were fluent in ASL. The invitation to participate was posted on several Deaf newsgroups as a link to a video, which contained the fingerspelled URL address for the test.

5.3 Procedure

The tests were performed online, using SignQUOTE [**28**], which facilitates online testing exclusively in ASL. Participants used their own computer, equipped with a web cam, to complete the test.

Participants viewed the informed consent, and indicated their consent by continuing on with the study and filling out a brief pre-test questionnaire. In the first section of the test, they viewed the story, one sentence at a time, and rated each sentence. They could view a sentence as many times as they wished. All of the sentences in this portion of the test included the automated spine movement. See Figure 15 for a screen shot of the SignQUOTE interface.

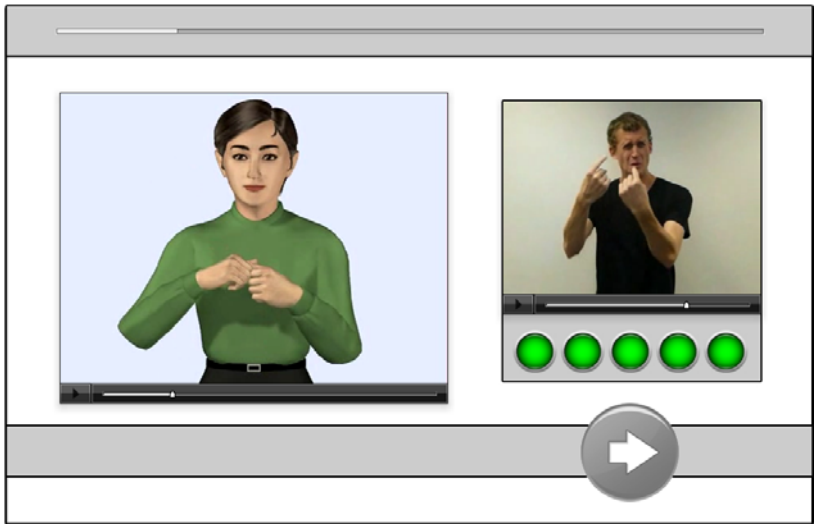

Figure 15: Test stimulus for part 1

In the second section of the test, participants again viewed the story, one sentence at a time, but saw two versions of the sentence, presented side-by-side, see Figure 16. To avoid experimental bias, the positions of the two stimuli were varied, so the control sometimes appeared on the left and sometimes on the right. Participants could view each of the two stimuli as many times as they wished, as both stimuli had their own play/pause/rewind buttons.

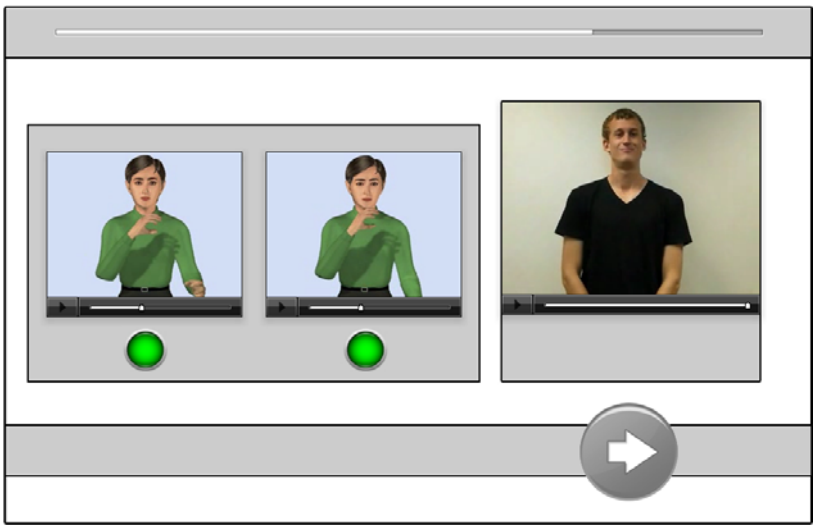

Figure 16: Stimulus for part 2

5.4 Measures

For the first section of the test, participants rated each sentence for clarity, grammaticality, understandability and naturalness on a five-point Likert scale. Table 2 contains a back translation of the questions and the rating scales. For the second section of the test, participants clicked the button corresponding to the animation they preferred.

1. Rate the clarity of the sentence.

Very confusing ● ● ● ● ● Very clear

2. How grammatical was the sentence?

Full of errors, unacceptable ● ● ● ● ● Perfectly correct

3. How understandable was the sentence?

Very hard ● ● ● ● ● Very easy

4. How natural was the motion?

Robotic ● ● ● ● ● Very human-like

Table 2: Back translation of questions and rating scales for part 1

5.5 Results

In total, 22 participants completed the test. The demographic breakdown was as follows

- 19 deaf, one hard-of-hearing, and two hearing participants.
- All participants identified either ASL (16 participants) or Pidgin Signed English (6 participants) as their preferred language.
- 17 had used ASL all their lives, 3 had used ASL between 5 and 15 years, and two between 1 to 2 years.
- 17 of the 22 were born deaf.

For part 1 of the test, participants viewed each of the five sentences in the story in turn and answered the questions in Table 2. The results for the four questions are displayed in Figure 17 as Tukey box-and-whisker plots. The median value in each sentence is displayed as a dark-blue circle, the range between the first (25%) and third (75%) quartiles is displayed as a thick box, and the overall range is given by the whisker lines. The data in this table show that

- With the exception of sentence 3, over 75% of participants rated clarity as a 3 or above, with 50% rating them as clear to very clear.
- With the exception of sentence 3, over 75% of participants rated grammaticality as a 3 or above, with 50% rating them as correct or perfectly correct.
- With the exception of sentence 3, over 75% of participants rated understandability as a 3 or above, with 50% rating them as easy or very easy to understand.
- Over 50% of participants rated every sentence as human-like or very human-like.

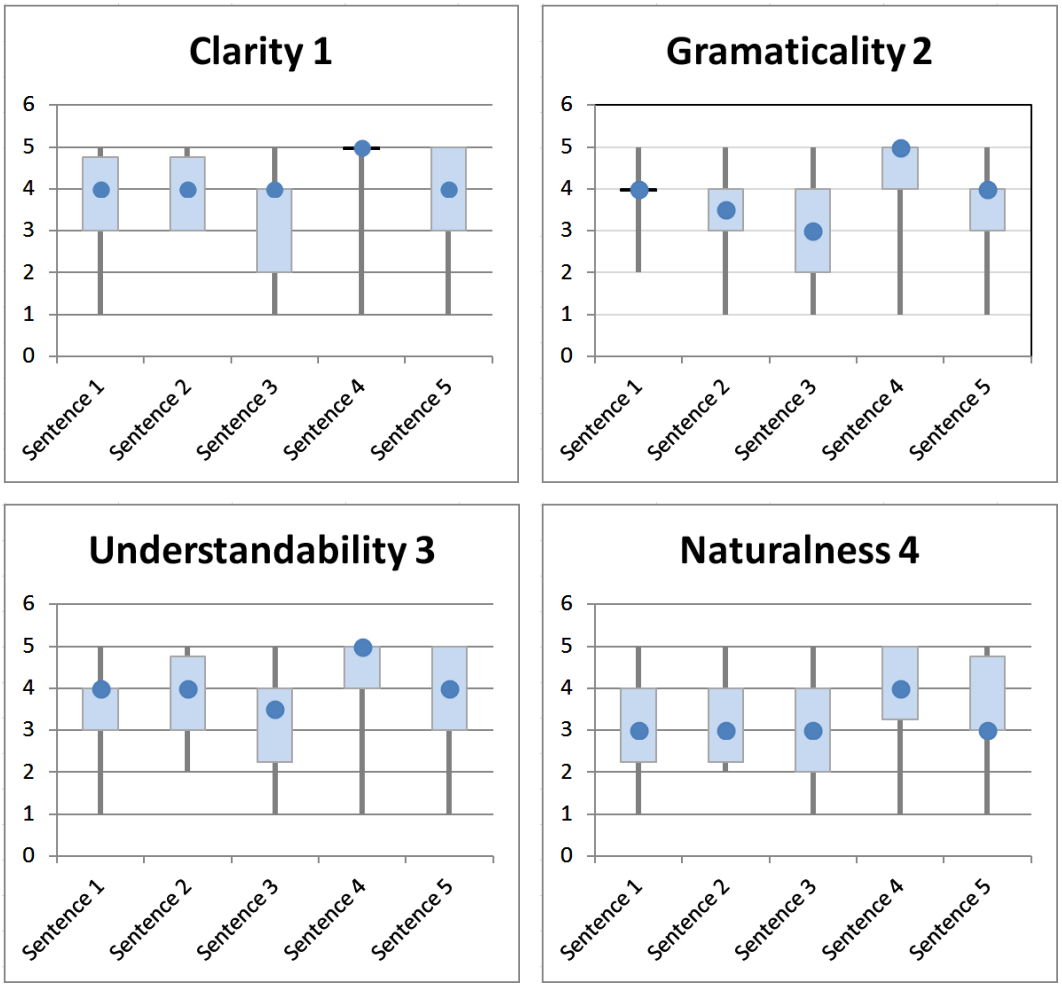

Figure 17: Results from part 1 of test

Table 3 displays the responses from part 2 of the test as A/B preferences. In the test instrument, the A/B order of the control stimulus varied, however the responses given below have been reordered to make the data more readable. The last row of the table contains the results of a binomial test expressing the probability that the results could be due to random chance. In all cases, participants preferred the spine movement at least as well as the control, and in three of the sentences (2, 3, 4) their preference for the spine movement was statistically significant ($p < 0.05$).

	S1	S2	S3	S4	S5
Control	11	4	16	3	13
Spine Movement	11	18	6	19	9
Binomial p-value	0.5840	0.0021	0.0262	0.0004	0.2617

Table 3: Preferences from Part 2 of test

6. DISCUSSION AND CONCLUSIONS

The data from the first part of the test address the research question, "Does the automated spine system generate movement that is clear, correct, understandable, and natural?" Participants perceived the majority of the sentences in a positive light. The results were particularly strong for clarity and understandability where over half of the participants rated all sentences as either clear or very clear and easy or very easy to understand[1]. Also, sentences received favorable ratings in grammaticality with the exception of sentence 3. This may be due in part to the fact that the structure of this sentence was more complex. Sentence 3 was the only one where the signer was not the subject of the sentence, and the only one that contained both direct and indirect objects. These results compare favorably to results from similar user studies of generated ASL [**29**][**30**].

English: My brother bought me a new one. It was big and blue.
ASL: BUT SURPRISE -- MY BROTHER GET NEW C-A-R, BIG, BLUE, GIVE-ME.

Figure 18: Sentence 3 of the test stimuli

Overall, the ratings for naturalness were lower than for the other three measures with 50 percent of the respondents rating all of the sentences as neutral or above. It may be that naturalness is the most demanding criterion of all. This is analogous to synthesized speech. It is possible to *understand* a *grammatical* synthetic voice message *clearly,* such as an automated airline flight notification sent to a person's phone*,* but it would not be perceived as *natural.* Although lower than the other three, this result still compares favorably with prior tests of naturalness in key-frame and procedurally generated animations of sign language avatars.

Data from the second part of the test yield insights into the second research question, "Does the addition of movement from an automated spine system produce animation that is preferable to animations without it?" In the majority of cases, the answer appears to be "yes." Responses from participants showed a statistically significant preference ($p < 0.05$) for the added spinal movement sentences 2, 3 and 4.

For sentences 1 and 5, however, the responses showed no statistically significant preference for either version, although a majority of participants preferred the added spinal movement in sentence 5. This may be due to the fact that, in these sentences, the spinal motion caused by the system was largely in the forward-back direction. In contrast, the other three sentences had more lateral and twisting motion, which is more apparent from a front camera view. Another contributing factor in sentence 1 may have been its placement as the first video in the sequence. Participants may have needed acclimation to the side-by-side video comparison. More study is indicated, and perhaps in future tests a warm-up exercise may be appropriate.

It is notable that, in all the cases that involved lateral motion of the body, a significant majority of participants judged the automated spine algorithm favorably. Recall that the automated spine system is not intended to be a sole solution for spine animation, but as an aid to animators, or as a supplement to

[1] Recall that a median of 3.5 on an integer 5-point Likert scale indicates that 50% were either a 4 or a 5.

other procedurally generated techniques. This user test indicates that the system would be highly beneficial in both of these roles.

7. SUMMARY AND FUTURE WORK

The techniques described in the present work address two of the many causes of robotic motion that can arise from implementing the SLPA model of ASL in procedurally-based synthesis systems.

First, our new approach can serve as an extension to any limb IK system, and aids artists by automatically rotating the torso and spine of an avatar to support the specified arm motions from the linguistic model. The approach staggers the timing of the keys to reduce the marionette effect arising from completely synchronous joint movement. The second technique addresses the complete lack of motion in held joints. Careful applications of Perlin noise to rotational joints can enliven the avatar even when the linguistic model specifies that an avatar hold a pose, or a joint's motion is unspecified. These techniques have the potential for both aiding manual animation and also for supplementing procedurally generated avatar movements systems such as [**1**] [**2**].

Both of these techniques still present opportunities for further study. For example, it is possible that basing the shoulder and torso extension on the position of the avatar's elbow rather than the IK end-effector would yield more realistic motions. The elbow's position is more closely related to the shoulder, having only a single bone segment between them.

Also, both the torso and the noise methods presented here would benefit greatly from a detailed study of motion-capture data and application of the comfort model presented in [**31**]. The present model was formed through close collaboration between computer scientists and animators, but could benefit from both improvement and validation though a data-driven study.

Finally, we are studying the other causes of robotic motion mentioned here, and intend to address them in the future.

8. REFERENCES

1 Hanke, Thomas, Matthes, Silke, Regen, Anja et al. Using Timing Information To Improve the Performance of Avatars. (Dundee, Scotland, UK 2011), Second International Workshop on Sign Language Translation and Avatar Technology (SLTAT).

2 Delorme, Maxime, Filhol, Michael, and Braffort, Annelies. Animation Genreation Process for Sign Language Synthesis. (2009), Advances in Computer-Human Interactions, ACHI '09, pp 386-390.

3 Efthimiou, Eleni, Stavroula-Evita, Fotinea, Vogler, Christian et al. Sign Language Recognition, Generation, and Modelling: A Research Effort with Applications in Deaf Communication. *Universal Access in Human-Computer Interaction, Addressing Diversity. Lecture Notes in Computer Science.*, Volume 5614 (2009), pp 21-30.

4 Wolfe, Rosalee, McDonald, John, and Schnepp, Jerry. An Avatar to Depict Sign Language: Building from Reusable Hand Animation. (Berlin, Germany, January 10-11, 2011), International Workshop on Sign Language Translation and Avatar Technology (SLTAT).

5 Gibet, Sylvie, Courty, Nicolas, Duarte, Kyle, and Le Naour, Thibaut. The SignCom system for data-driven animation of interactive virtual signers: Methodology and evaluation. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 1, 1 (2011), 1-26.

6 Metzger, M. Constructed Dialogue and Constructed Action in American Sign Language. In *The Sociolinguistics in Deaf Communities*. Gaullaudet University Press, Washington, D.C., 1995.

7 Wilbur, Ronnie. Phonological and Prosodic Layering of Nonmanuals in American Sign Language. In Emmorey, K et al., eds., *The Signs of Language Revisited: Festscrift for Ursula Bellugi and Edward Klima*. 2000.

8 Wolfe, Rosalee, Cook, Peter, McDonald, John, and Schnepp, Jerry. Linguistics as structure in computer animation: Toward a more effective synthesis of brow motion in American Sign Language. *Nonmanuals in Sign Language Special issue of Sign Language & Linguistics*, 14, 1 (2011), 179-199.

9 Courty, Nicolas and Gibet, Sylvie. Why Is the Creation of a Virtual Signer Challenging Computer. (Utrecht, The Netherlands 2010), Springer, 290-300.

10 Johnson, Robert E and Liddell, Scott K. A Segmental Framework for Representing Signs Phonetically. *Sign Language Studies*, 11, 3 (2011), 408-463.

11 Schnepp, J. *A representation of selected nonmanual signals in American Sign Language*. ProQuest Dissertations and Theses, DePaul University, 2012.

12 Perlin, Ken. A system for scripting interactive actors in virtual worlds. In *Proceedings of ACM SIGGRAPH 96* (New Orleans 1996), Association for Computing Machinery, 205-216.

13 Johnston, Ollie and Thomas, Frank. *The Illusion of Life: Disney Animation*. Random House (Disney Press), New York, 1995.

14 Lasseter, John. Priniples of traditional animation applied to 3D computer animation. In *SIGGRAPH '87 Proceedings of the 14th annual conference on Computer graphics and interactive techniques* (Anaheim, California 1987), Association of Computing Machinery, 35-44.

15 Tolani, Deepak, Goswami, Ambarish, and Badler, Norman. Real-Time Inverse Kinematics Techniques for Anthropomorphic Limbs. *Graphical Models*, 62 (2000), 353-388.

16 Whitaker, Harold and Halas, John. *Timing for Animation*. Focal Press, Burlington, Massachusetts, 2008.

17 Poor, G. *ASL Video Dictionary and Inflection Guide*. 2008.

18 McDonald, John, Alkoby, Karen, Carter, Roymieco et al. A Direct Method for Positioning the Arms of a Human Model. In *Proceedings of Graphics Interface* (Calgary, Alberta Canada 2002), 99-106.

19 Paolo, Baerlocher and Ronan, Boulic. An Inverse Kinematics Architecture Enforcing an Arbitrary Number of Strict Priority Levels. *The Visual Computer*, 20, 6 (2004), 402-417.

20 Wyvill, Brian, McPheeters, Craig, and Wyvill, Geoff. Animating soft objects. *The Visual Computer*, 2 (1986), 235-242.

21 Shoemake, Ken. Euler angle conversion. In *Graphics Gems IV*. Academic Press, San Diego, CA, 1994.

22 Saladin, Kenneth. *Human Anatomy*. McGraw-Hill, New York, NY, 2007.

23 Reese, Nancy Berryman. *Joint Range of Motion and Muscle Length Testing*. Elsevier, St. Louis, MO, 2010.

24 Filhol, Michael. A combination of two synchronisation methods to formalise sign language animation. *Proceedings of the 9th international Gesture Workshop* (2011).

25 Sozio, Syl. *The Mastery of Mimodrame: An In-Depth Study of Mime Techniques*. Destiny Image Publishers, Shippensburg, PA, 1989.

26 Phong, Tuong Bui. Illumination for computer generated pictures. *Communications of the ACM*, 18, 6 (June 1975), 311-317.

27 Napoli, Rob and Gloman, Chuck. *Scene Design and Lighting Techniques: A Basic Guide for Theatre*. Focal Press, Burlington, MA, 2007.

28 Schnepp, J., Wolfe, R., Shiver, B., McDonald, J., and Toro, J. SignQUOTE: A Remote Testing Facility for Eliciting Signed Qualitative Feedback. In *Second International Workshop on Sign Language Translation and Avatar Technology (SLTAT)* (2011).

29 Kacorri, Hernisa, Lu, Pengfei, and Huenerfauth, Matt. Effect of Displaying Human Videos During an Evaluation Study of American Sign Language Animation. *ACM Transactions on Accessible Computing (TACCESS)*, 5, 2 (October 2013), 4.

30 Huenerfauth, Matt, Lu, Pengfei, and Rosenberg, Andrew. Evaluating importance of facial expression in American Sign Language and pidgin signed English animations. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility* (New York 2011), Association For Computing Machinery, 99-106.

31 Delorme, Maxime. *Modelisation du squelette pour la generation realiste de postures de la lange de signes francaise*. Ph.D.

Dissertation, Laboratoire d'Informatique pour la Mechanique et les Sciences de L'Ingenieur (LIMSI). 2011.