Evaluating a Dynamic Time Warping Based Scoring Algorithm for Facial Expressions in ASL Animations

Hernisa Kacorri[1], *Matt Huenerfauth*[2]

[1]The Graduate Center, CUNY, Doctoral Program in Computer Science, USA
[2]Rochester Institute of Technology, Golisano College of Computing and Information Sciences, USA
hkacorri@gradcenter.cuny.edu, matt.huenerfauth@rit.edu

Abstract

Advancing the automatic synthesis of linguistically accurate and natural-looking American Sign Language (ASL) animations from an easy-to-update script would increase information accessibility for many people who are deaf by facilitating more ASL content to websites and media. We are investigating the production of ASL grammatical facial expressions and head movements coordinated with the manual signs that are crucial for the interpretation of signed sentences. It would be useful for researchers to have an automatic scoring algorithm that could be used to rate the similarity of two animation sequences of ASL facial movements (or an animation sequence and a motion-capture recording of a human signer). We present a novel, sign-language specific similarity scoring algorithm, based on Dynamic Time Warping (DTW), for facial expression performances and the results of a user-study in which the predictions of this algorithm were compared to the judgments of ASL signers. We found that our algorithm had significant correlations with participants' comprehension scores for the animations and the degree to which they reported noticing specific facial expressions.

Index Terms: American Sign Language, accessibility for people who are deaf, animation, natural language generation

1. Introduction

Access to understandable information on websites and other media is necessary for full participation in society. Yet, the vast majority of information content online is in the form of written language text, and there are many users who have difficulty reading this material. For many people who are deaf and hard-of-hearing, there are educational factors that may lead to lower levels of written language literacy. In the U.S., standardized testing has revealed that a majority of deaf high school graduates (students who are age 18 and older) have a fourth-grade English reading level or below [27]. (U.S. students in the fourth grade of school are typically age 10.) While they may have difficulty with written English, many of these users have sophisticated fluency in another language: American Sign Language (ASL).

More than 500,000 people in the U.S. use ASL as a primary means of communication [20]. However, fluency in ASL does not entail fluency in written English since the two are distinct natural languages: with their own word order, linguistic structure, and vocabulary. Thus, information content can be easier to understand for many deaf users if it is presented in ASL. A spontaneous approach to presenting ASL online would be to upload videos of human signers on website and other media, but this is not ideal: re-filming a human performing ASL for frequently updated information is often prohibitively expensive, and the real-time generation of content from a query is not possible. Software is needed that given an easy-to-update script as input can automatically synthesize ASL signing performed by a virtual human character. This software must internally coordinate the movements of the virtual human character such that the animated ASL message is linguistically accurate, understandable, and acceptable among users. The creation of such software is the focus our research.

An ASL utterance consists of the movement of the hands, arms, torso, head, eye-gaze, and facial expressions. In fact, facial expressions are essential to the understandability and meaning of ASL sentences (see section 2). Our research focuses on the automatic synthesis of facial expression movements for an ASL-signing virtual human character such that the resulting animations are judged to be clear and understandable by deaf users. In addition to our ongoing research in this area, other groups have studied issues related to the synthesis of facial expressions for sign language animation, whose methods and contributions we compare and survey in [14]. For researchers like ourselves, who are interested in designing software that generates linguistically-accurate ASL facial expressions performed by virtual human characters, the most comprehensive way to evaluate the quality of the software is to conduct user studies. Typically, we generate animations using the facial expression selection software, set up an experiment in which deaf participants view and evaluate the animations, and compare the scores of animations produced using the software (to some baselines or to prior versions of the software). Of course, conducting such studies with users is time-consuming and resource-intensive; so, these studies cannot be conducted on a frequent basis (e. g., weekly) during the development of ASL facial-expression synthesis software. For this reason, it would be useful to have some automatic method for quickly evaluating whether the facial expression produced by the software for some specific ASL sentence is accurate. In this paper, we present an automatic scoring algorithm that can compare two facial expression performances to rate their similarity. In principle, this automatic scoring tool could be used to quickly evaluate whether the output of facial expression synthesis software is producing a result that is similar to ASL utterances recorded from actual human ASL signers. The proposed algorithm could be incorporated into a data-driven facial expression synthesis architecture, an approach which is also favored by other sign language animation researchers, e. g.: [26] that use computer vision to extract facial features and produce facial expressions that occur during specific signs, and [3] that map facial motion-capture data to animation blend-shapes using machine-learning methods.

The face and head position of a virtual human character

SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 29–35,
Dresden, Germany, 11 September, 2015.

at any moment in time can be conceptualized as a vector of numbers, specifying joint angles and facial-control parameters at that moment in time. Thus, an animation is a stream of such vectors. While there are a variety of techniques that can be used to measure the similarity between two time-streams of vectors, this paper will specifically explore an approach based on a Dynamic Time Warping (DTW) algorithm. Section 5 describes DTW and discusses how some researchers have begun to use this algorithm to rate the similarity of non-sign-language emotional facial expressions for animated characters [19]; however, no user-study had been performed to verify that such scores actually matched human judgments of similarity – nor has this technique yet been applied to sign-language facial expressions.

This paper presents a novel, sign-language specific scoring algorithm based on DTW, which takes into account the timing of words in the sentence. This paper reflects our first efforts at designing a DTW-based scoring tool, and the goal of this paper is to determine if the technique holds promise – if so, then we intend to investigate further variations of the scoring algorithm, to optimize it for ASL. In order to determine if our scoring tool is useful, we must determine whether the scores it provides actually correlate with the judgments of human ASL signers who evaluate ASL animations in an experiment. This paper presents a user study we conducted in which human ASL signers evaluated animations with facial expressions of different levels of quality (as rated by the automatic scoring tool), and we measure how well our automatic scoring correlates with the human judgments.

The remainder of this paper is organized as follows: Section 2 describes the linguistics of various ASL facial expressions, and section 3 describes how we time-warp a motion-capture recording of a facial expression performance to suit the synthesis of an ASL animation of a sentence with a different time duration. Section 4 describes how the movements of the face of a virtual human character can be parameterized and controlled, and Section 5 defines our new DTW-based automatic scoring algorithm. Section 6 presents our research questions and hypotheses, which were evaluated in a user-study presented in section 7. Finally, section 8 discusses these results and identifies future directions.

2. Syntactic facial expressions

Facial expressions are an essential part of the fluent production of ASL. They can convey emotional information, subtle variations in the meaning of words, and other information, but this paper focuses on a specific use of facial expressions: to convey grammatical information during entire syntactic phrases in an ASL sentence. ASL sentences with identical sequence of signs performed by hands can be interpreted differently based on the accompanying facial expressions. For instance, a declarative sentence (ASL: "ANNA LIKE CHEESECAKE" / English: "Anna likes cheesecake.") can be turned into a Yes-No question (English: "Does Anna like cheesecake?"), with the addition of a Yes-No Question facial expression during the sentence. Similarly, the addition of a Negation facial expression during the verb phrase "LIKE CHEESECAKE" can change the meaning of the sentence to "Anna doesn't like cheesecake." where the signing of the word NOT is optional. For an interrogative question (typically including a "WH" word in English such as where, why, and what), e. g. "ANNA LIKE WHAT", a co-occurring WH-Question facial expression is necessary during the ASL sentence. Instances of these three ASL facial expressions are illustrated in Figure 1.

Figure 1: *Examples of ASL linguistic facial expressions: (a) Yes-No Question, (b) WH-Question, (c) Negation.*

While we use the term "facial expressions," these phenomena also include movements of the head, which we model in this paper. ASL linguistics references contain more detail about each, e. g., [22], but a subset of them is described briefly below:

- Yes-No Question: The signer raises his eyebrows while tilting the head forward during a sentence.
- WH-Question: The signer furrows his eyebrows and tilts his head forward during a sentence.
- Negation: The signer shakes his head left and right during the phrase with some eyebrow furrowing.

An ASL linguistic facial expression varies in the way it is performed during a given sentence based on the overall number of signs, the start and end times for a particular word in the sentence (e. g., WHAT and NOT), preceding and succeeding facial expressions, signing speed, and other factors. Thus, simply playing on a virtual character a pre-recorded human performance of a facial expression to a novel, not previously recorded, sentence is insufficient. For this reason, we are investigating how to model and synchronize to manual movements the performance of a facial expression in various contexts.

3. Time-warping facial expressions

In our research on synthesizing ASL animations, we often need to generate a novel animation by assembling a sequence of individual words from a prebuilt animation dictionary; each word may have its own typical duration, which is used to determine a timeline for the full ASL utterance. We seek to add a facial expression performance to such animations, and in section 4, we discuss how facial features extracted from the recording of a human's face could be used to drive the movements of the animated character. Thus, the time-duration of the recording must be "warped" to match the time duration needed in the animation to be synthesized.

Simplistically, the recording could be linearly stretched or squeezed to suit the target time duration, but animation researchers have investigated a variety of techniques for time-warping motion data to new contexts, e. g., [7, 31]. In many approaches, e. g., [7], key milestones during a recorded action are identified in the timestream (e.g., each footfall during a walking action), and these milestone times are used as parameters to determine how to warp the recording (so that the movements of the human for each "footstep" of the walking action are warped into appropriate footstep actions that meet timing requirements for when the virtual human footsteps must occur in the animation).

When synthesizing sign language animations, we have access to information about the underlying timeline of the utterance, which we can use to select useful milestones for time-warping:

- ASL facial expressions occur in relation to the timing of the words during a sentence [22]. Yes-No Question and WH-Question facial expressions typically ex-

tend across entire clauses, and Negation, across an entire verb phrase.

- Signers perform anticipatory head movements so that the main action begins with the clause or phrase [22].
- Many phrases with facial expressions begin with or end with a word that has a special relationship to the facial expression being performed (such that there may be additional intensity of the facial expression during this initial/final word).
  - Negated verb phrases may include the word NOT at the beginning of the phrase, where greatest intensity of the Negation facial expression will occur [22].
  - WH-Question clauses typically end with a WH-word, and in some contexts, the facial expression may occur only (or with greatest intensity) during this word [18].
  - Yes-No Question clauses often end with a right-dislocated pronoun [22] or a "QM-wg" (wiggling finger question mark) sign at the end [1].

For an ASL animation that contains a sequence of words, S, when a facial expression occurs, we define four phases of time based on the intervals between five milestones on the timeline:

M1: The end of the word immediately before S
M2: The beginning of the first word in S
M3: For Negation, M3 is the beginning of the second word in S, otherwise, M3 is the beginning of the last word in S
M4: The end of the final word in S
M5: The beginning of the word that immediately follows S

If S begins or ends an utterance, then M1 and M5 are set to a value 500msec away from S. The rationale for these definitions is:

- Phases M1-M2 and M4-M5 represent the onset and offset of the facial expression, before and after S.
- For a Negation phrase, M2-M3 is the duration of the first word, and M3-M4 is the remainder of the phrase. A Negation phrase may begin with the word NOT, when a particularly intense facial expression may occur. Thus, it is useful to distinguish the time of the first word of the phrase. (If S contains only one word, then these phases are merged.)
- For a Yes-No Question or a WH-Question, M3-M4 is the duration of the final word, and M2-M3 is the remainder of the phrase. There is often additional facial expression intensity during the final word of a question; thus, it is useful to distinguish the time of the final word of the question. (If S contains only one word, then these phases are merged.)

Recall that our goal is to modify the timing of a human's facial movement recording to suit the timeline of a target animation we want to synthesize. For any human recording that we plan to use as a source material for facial movements, we ask an ASL signer to identify these five milestones. When we want to modify the timing of a recording, we perform time-warping for each of these four phases independently. Thus, data from phase M2-M3 of the recorded human utterance is time-warped to fit the duration of phase M2-M3 of the target animation that we are synthesizing. In this way, we can increase the likelihood that the appropriate portion of the human's facial performance coincides with the timing of the appropriate signs in the resulting animation.

The top of Figure 2 shows how a recording of the eyebrow height of a human signer during a Yes-No question might appear during an ASL sentence: "SHE LIVE DC SHE" (English: "Does she live in DC?"). The milestones are marked with vertical lines, and the figure shows how data from each phase of the recording can be linearly time-warped to produce a facial expression for an animation with different word durations. (The graph in Figure 2 is an artist's rendering meant to illustrate the warping technique.)

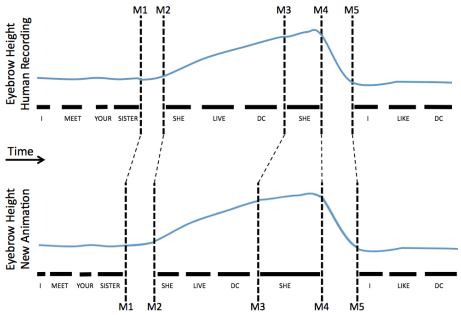

Figure 2: *Phase-based time-warping of a recording of a human's eyebrow movements from a Yes-No Question (above) for an animation with a different timeline (below).*

4. MPEG-4 and ASL animation

In prior work, we constructed a lexicon of ASL signs and a collection of ASL stimuli [9] for use in experiments to evaluate facial expression animation synthesis methods. As part of that project, we recorded videos of a native ASL signer performing the stimuli, and we extracted the facial features and head pose of the human signer in the videos using the Visage Face Tracker (shown in Fig. 3). Visage is an automatic face tracking software [24] that provides a stream of MPEG-4 Facial Action Parameters (FAPs) that represent the facial expression of the human.

The MPEG-4 standard [11] defines a 3D model-based coding for face animation. The facial expression of a human (or an animated character) can be represented by a set of 68 FAPs, representing head motion, eyebrow, nose, mouth, and tongue controls, all of which can be combined for representation of natural facial expressions. For example, "raise_l_i_eyebrow" is one of the FAPs (codename FAP30) in the MPEG4 standard, and it represents the vertical displacement of left inner eyebrow. Larger values for this number would indicate that the eyebrow is raised higher. To specify a changing facial expression over time, a stream of numerical values for all of the FAPs of the face is needed, for each frame of animation.

MPEG-4 FAPs have been used by a variety of non-sign-language animation researchers studying, e. g., expressive embodied agents [21], emotional facial expressions during speech in synthetic talking heads [19], or dynamic emotional expressions [30]. A useful property of MPEG-4 is that the FAP values are normalized to the proportion of the character's face as shown in Fig. 3; thus, a stream of FAP values could be used to drive the animation of virtual humans with different face proportions,and the resulting animation would appear to have similar facial expressions, when played on a difference virtual human.

To support our research on ASL facial expressions (especially the development of automatic scoring tools), it was necessary to implement a virtual human animation platform with face-movement control parameters. We decided to use MPEG-4 facial action parameters [11] , and we enhanced the EMBR platform [5, 6, 16] with MPEG-4-based face controls. We also

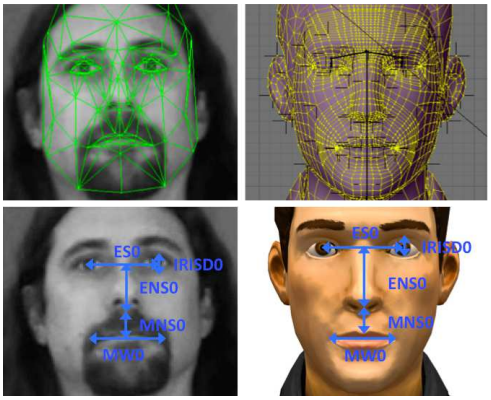

Figure 3: *MPEG-4 facial features and scaling factors on the human signer in Visage (left) and the avatar (right).*

implemented an intermediate component that converts MPEG-4 data to EMBRscript, the script language supported by the EMBR platform. Our script generation component performs the phase-based time-warping approach described in section 3 to align the facial expression with the animated character hand movements. The FAPs that are used to drive our facial expression animations for this paper include the following (additional FAPs may be implemented in future work):

Head orientation (FAP47-FAP49): orientation parameters given in Euler angles defined as pitch, yaw, and roll. In addition to head orientation, the Visage output also includes the head's location in 3D space; we adjust the torso movements of our avatar based on these values.

Vertical displacements of eyebrows (FAP30-FAP35): 6 parameters directly applied the inner, middle, and outer points of the left and right eyebrow to allow for different combinations of raised and lowered eyebrows.

Horizontal displacements of eyebrows (FAP36-FAP37): 2 parameters directly applied in the inner points of the eyebrows that allow for e.g. furrowed eyebrows.

5. The dynamic time warping algorithm

In this paper, we present a novel method for evaluating the quality of synthesized facial expressions for sign-language animations, which is based in the Dynamic Time Warping (DTW) algorithm. DTW arose in the field of speech recognition [25, 28] as a generalization of algorithms for comparing series of values with each other. DTW sums the distance between the individual aligned elements of two time series, which are locally stretched or compressed, to maximize their resemblance. Unlike the Euclidean distance, it can serve as a measure of similarity even for time series of different length. An advantage of DTW over other cross-correlation similarity measures is that it allows for non-linear warping. There are a variety of DTW algorithms, used in several fields, with different global or local constraints (e. g., local slope, endpoints, and windowing), different feature spaces for the time series values, and different local distance metrics between the individual aligned elements (e. g., Euclidean, Manhattan).

DTW has been used as a similarity scoring technique for facial animation, e. g., for the retrieval of facial animation based on a key-pose query [23] and spatio-temporal alignment between face movements recorded from different humans [31]. In prior work, DTW has been also considered as a method for scoring the quality of time series data. Kraljevski et al. [17] found correlation between DTW distance and the measured Perceptual Evaluation of Speech Quality (PESQ) values for test and received speech in a simulated transmission channel with packet loss. (PESQ [12] is a perceptual objective measure typically used for estimating the transmission channel impact in speech. However, it has been also used for synthesized speech quality assessment [2].)

Mana and Pianesi [19] used DTW distance as a quality measure for the quantitative evaluation of synthesized non-sign-language emotional facial expressions in a MPEG-4 compatible avatar. They compared "synthetic" time series of facial markers per frame, with the corresponding "natural" time series performed by a human. While the authors commented that the synthetic animations preferred by DTW appeared (to them) similar to the original human performance, they did not verify that DTW scores related to human judgments of facial expression similarity by conducting a user study (which we have done, as described in section 7).

Algorithm 1 ASL facial expression animations scoring

```
1: function GetDistance(g, c, M, N, c_dur, anim_dur)
2:     G = [g[M1,M2], g[M2,M3], g[M3,M4], g[M4,M5]]
3:     C = [c[T1,T2], c[T2,T3], c[T3,T4], c[T4,T5]]
4:     distance = 0
5:     for ph_g, ph_c in pair (G, C) do
6:         norm_d = DTW(ph_g, ph_c)
7:         distance = distance + norm_d
8:     scale = anim_dur / c_dur
9:     return distance * scale
```

5.1. Our DTW-based scoring algorithm

Our scoring algorithm assumes that we have:

- A timeline of the words for a "target" ASL animation that we want to generate, where the facial expression has a given duration in milliseconds (**anim_dur**). If we are synthesizing an ASL animation using a pre-built animation lexicon of individual ASL signs, then the duration of these items will affect the overall timeline plan for the target animation to be synthesized. Now, a facial expression must be synthesized.
- A "gold standard" (**g**) motion-capture recording of a human's facial expressions for this ASL sentence (or a very high quality animation of a facial expression which is trusted to be of excellent quality) and the list of five milestones on its timeline (**M1, ..., M5**). Notably, the timeline of when the recorded human performed each word of the sentence will be slightly different than the timeline of the target animation. A video recorded performance of ASL grammatical facial expression can be considered as a multivariate time series, a series of detected MPEG-4 FAPs values in each video frame.
- A "candidate" stream (**c**) of MPEG-4 facial expression parameters that has been synthesized by some software (or perhaps another motion-capture recording) that we wish to evaluate, the list of five milestones on its timeline (**T1, ..., T5**), and its duration in milliseconds (**c_dur**).

Our scoring algorithm initially constructs a list of partial streams for the four phases of the facial expressions g and c based on the intervals between the given five milestones on their timeline (Line 2, Line 3). Then it initializes the total distance between the gold standard and the animated candidate with 0 (Line 4). For each pair of steams of the same phase (Line 5) the algorithm calculates the normalized distance based on Dy-

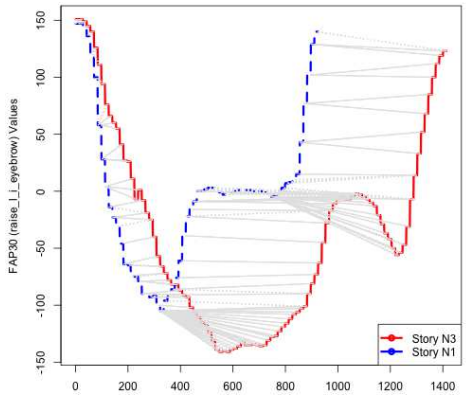

Figure 4: *Example of DTW alignment between the "raise_l_i_eyebrow" values detected in human recordings of two ASL stories containing a Negation facial expression.*

namic Time Warping (Line 6) and adds it to the total distance (Line 7). Since the "candidate" stream and the final animation have different durations, a scaling factor is applied to the distance, based on the stretching or compression of the "candidate" stream (Line 8, Line 9).

To calculate the distance between the two partial streams (Line 6) we used the implementation of multivariate DTW in [4]. It computes a global alignment with minimum distance normalized for path length using Euclidean as a local distance. Computing global alignments means that the time series' heads and tails are constrained to match each other. We further tuned the algorithm by using the asymmetric step pattern and a SakoeChiba warping window of size 10.

Figure 4 illustrates an example of an alignment for the detected values of MPEG-4 control "raise_l_i_eyebrow" with the Visage SDK [24] during a human's performance of two ASL stories containing a Negation facial expression (with codenames N3 and N1 in the stimuli collection [9]). The alignment is preformed with the default multivariate implementation of DTW in the R package, *dtw* [4]. The duration of the facial expression in N3 and N1 is 1414 and 924 frames, respectively and their calculated normalized distance was found to be 8.76.

6. Hypotheses

Our goal for this paper is to evaluate our novel, sign-language specific, DTW-based scoring algorithm for facial expressions. One method would be to conduct a study in which human judges estimate similarity scores between face movements in pairs of ASL recordings (and then compare our algorithm to their scores), but we did not find prior published studies in which human judges were able to provide reliable numerical ratings of facial expression similarity between pairs of ASL animations. On the other hand, in several prior studies [8], human participants have been able to answer comprehension questions about ASL animations and indicate whether they noticed particular facial expressions. Thus, we evaluated our DTW algorithm by: (1) selecting a human ASL recording that serves as a gold-standard face performance, (2) using our similarity scoring algorithm to compare this gold-standard to other candidate recordings, and (3) asking human judges to evaluate the comprehensibility of these candidate ASL performances. If we find that our algorithm?s prediction of the similarity between the candidate and the gold-standard correlates with such human-judgments, then we would posit that our algorithm is a useful tool for researchers who are investigating the synthesis of sign-language facial expressions. Thus, we propose the following two hypotheses:

Hypothesis 1: Our scoring algorithm correlates with participants' implicit understanding of the facial expression, as measured through comprehension questions that probe the participant's understanding of the information content of the animation.

Hypothesis 2: Our scoring algorithm correlates with participants' explicit recognition of the facial expression, as measured through a question that asks participants whether they noticed a particular facial expression during the animation.

7. User study

To evaluate our hypotheses, we conducted a user study, where participants viewed animations of short stories in ASL and then answered comprehension and scalar-response questions.

Stimuli. To produce animated stimuli, we selected 6 recordings of a human ASL signer performing ASL stories for each of the 3 categories of ASL grammatical facial expressions (Negation, WH-Question, or Yes-No Question). This is a total of 18 stimuli. We describe our collection of recordings in [9], and the codenames of the selected stories used in this paper were N1-N6, W1-W6, and Y2-Y7, respectively. To obtain the facial expression data that would drive the animations we run Visage Face Tracker [24] on the video recordings of a native ASL signer performing each of the stories. Then we extracted the head position, head orientation, and MPEG-4 FAPs values for the portion of the story where the facial expression occurs.

Next, to generate our stimuli, we rendered an ASL animation of each story in two different versions:

min-distance: Face, head, and torso movements are driven by the recorded performance of the story with the smallest DTW distance from the 5 stories available in the same category. That is, to synthesize an animation of story N1, we used the face and head movements from the story in the set N2-N6 that had the minimum distance from the N1 recording, based on our new scoring algorithm (section 5.1). Notably, stories N2-N6 had different words, but were all Negation stories.

max-distance: Face, head, and torso movements are driven by the recorded performance of the story with the largest DTW distance from the 5 stories available in the same category.

Figure 5 illustrates the two versions of a Yes-No Question story (codename Y3). The video size, resolution, and framerate for all stimuli were identical. The hand movements in each version were identical and were created by native ASL signers using our laboratory's animation platform [5]. The facial movements were added during the portion of the story where the facial expression of interest should occur; the rest of the story had a static neutral face. The recorded head and facial movements were warped based on the timing of the words in the target animation, as described in section 3. Example stimuli animations from our study are available here: http://latlab.ist.rit.edu/2015slpat.

Experiment Setup. We conducted an evaluation study in which native ASL signers viewed animations of a virtual human character telling a short story in ASL. Each story included instances of one of the facial expressions of interest: Negation, WH-Question, or Yes-No Question. After watching each story

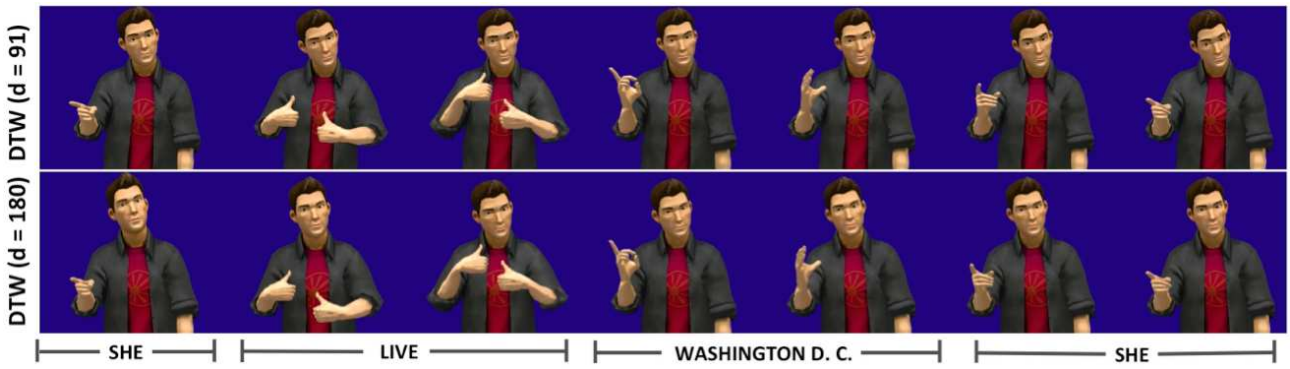

Figure 5: *Screenshots from a min-distance and max-distance version of a Yes-No Question stimulus in the study.*

animation (with facial expressions of one of two types: min-distance or max-distance) one time, participants responded to a "Notice" question (1-to-10 from "yes" to "no" in relation to how much they noticed an emotional, negative, questions, and topic facial expression during the story). Participants were asked to watch the story once more and answer four comprehension questions [9] on a 7-point scale from "definitely no" to "definitely yes." Participants could choose "I'm not sure" instead of answering. As discussed in [15], these stories and comprehension questions were engineered in such a way that the wrong answers to the comprehension questions would indicate that the participants had misunderstood the facial expression displayed [15]. E.g. the comprehension-question responses would indicate whether a participant had noticed a "yes/no question" facial expression or instead had considered the story to be a declarative statement.

At the beginning of the study, participants viewed a sample animation, to familiarize them with the experiment. A native ASL signer conducted all of the experiments in ASL. In prior work [9], we developed methods to ensure that responses given by participants are as ASL-accurate as possible.

Participants. In [10], 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 this 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, 15 participants learned ASL prior to age 9, The remaining 3 participants had been using ASL for over 11 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 10 men and 8 women of ages 22-42 (average age 29.8).

8. Results

Our hypotheses considered whether our new scoring algorithm would correlate with participants' implicit understanding of the facial expression (Hypothesis 1) or explicit recognition of the facial expression (Hypothesis 2).

To examine Hypothesis 1, we calculate the correlation between the distance score from the new algorithm and the score from comprehension questions in the user study. We found a significant correlation (Spearman's rho -0.38, $p - value < 0.001$): Hypothesis 1 was supported.

To examine Hypothesis 2, we consider the correlation between the distance score from the new algorithm and the score from the "Notice" question in the study. We found a significant correlation (Spearman's rho -0.33, $p - value < 0.001$): Hypothesis 2 was supported.

9. Conclusions and future work

While we believe that studies with ASL signers are the most conclusive way to evaluate the understandability and naturalness of animations of ASL, our positive results for hypotheses 1 and 2 suggest that sign-language animation researchers could use our new scoring algorithm to evaluate the facial expressions produced by their software. Having a rapid, repeatable method of evaluating the output of facial expression synthesis software is useful for monitoring the development of software, and this evaluation can be performed more frequently than user-based evaluations.

We believe that the time-warping algorithm (section 3) and our scoring algorithm (section 5.1) are a first-attempt at developing an automatic scoring approach, and now that we have observed some moderate though significant correlations in this study, we plan on investigating further variations of these techniques that might prove even more effective. For example, we may investigate the use of Longest Common Subsequence [29] instead of Dynamic Time Warping – or other probabilistic approaches to similarity – and compare them to our findings. We noticed that some of the phases (e. g., M4-M5) of the facial expressions had higher correlations with the participants' scores compared to other phases. This might indicate the need for further tuning of the coefficients for the partial distances calculated on each of the 4 phases.

In future work, we are interested in designing learning-based models for the synthesis of ASL facial expressions, including: topic, rhetorical questions, and emotional affect [13].

10. Acknowledgments

This material is based upon work supported by the National Science Foundation under award number 1506786. We are grateful for assistance from Andy Cocksey, Alexis Heloir, and student researchers, including Christine Singh, Evans Seraphin, Kaushik Pillapakkam, Jennifer Marfino, Fang Yang, and Priscilla Diaz. We would like to thank Miriam Morrow and Jonathan Lamberton for their ASL linguistic expertise and assistance in recruiting participants and conducting studies.

11. References

[1] C. L. BakerShenk, "American Sign Language: A teacher's resource text on grammar and culture," *Gallaudet University*, 1991.

[2] M. Cernak and M. Rusko, "An evaluation of synthetic speech using the PESQ measure," *In Proc. European Congress on Acoustics*, pp. 2725–2728, 2005.

[3] S. Gibet, N. Courty, K. Duarte, and T. L. Naour, "The SignCom system for data-driven animation of interactive virtual signers: Methodology and Evaluation," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 1, no. 1, pp. 6, 2011.

[4] T. Giorgino, "Computing and visualizing dynamic time warping alignments in R: the dtw package," *Journal of statistical Software*, vol. 31, no. 7, pp. 1–24, 2009.

[5] A. Heloir and M. Kipp, "Real-time animation of interactive agents: Specification and realization," *Applied Artificial Intelligence*, vol. 24, no. 6, pp. 510–529, 2010.

[6] A. Heloir, Q. Nguyen, and M. Kipp, "Signing Avatars: a Feasibility Study," *The Second International Workshop on Sign Language Translation and Avatar Technology (SLTAT)*, Dundee, Scotland, United Kingdom, 2011.

[7] E. Hsu, M. da Silva, and J. Popovi?, "Guided time warping for motion editing," *In Proc. of the 2007 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pp. 45–52, Eurographics Association, 2007.

[8] M. Huenerfauth and H. Kacorri, "Best practices for conducting evaluations of sign language animation," *In Proc. of the 30th Annual International Technology and Persons with Disabilities Conference (CSUN 2015), Scientific/Research Track*, San Diego, California, USA, 2015.

[9] M. Huenerfauth and H. Kacorri, "Release of experimental stimuli and questions for evaluating facial expressions in animations of American Sign Language," *In Proc. of the Workshop on the Representation and Processing of Signed Languages (LREC 2014)*, Rekjavik, Iceland, 2014.

[10] M. Huenerfauth, L. Zhao, E. Gu, and J. Allbeck, "Evaluation of American sign language generation by native ASL signers," *ACM Trans Access Comput*, vol. 1, no. 1, pp. 1–27, 2008.

[11] ISO/IECIS14496-2Visual, 1999.

[12] ITU-T Rec. P.862, "Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs," *ITU Geneva*, 2011.

[13] H. Kacorri, "Models of linguistic facial expressions for American Sign Language animation," *ACM SIGACCESS Accessibility and Computing*, vol. 105, pp. 19–23, 2013.

[14] H. Kacorri, "TR-2015001: A survey and critique of facial expression synthesis in sign language animation," *Computer Science Technical Reports. Paper 403*, 2015.

[15] H. Kacorri, P. Lu, and M. Huenerfauth, "Evaluating facial expressions in American Sign Language animations for accessible online information," *In Proc. of the International Conference on Universal Access in Human-Computer Interaction (UAHCI)*, Las Vegas, NV, USA, 2013.

[16] M. Kipp, A. Heloir, and Q. Nguyen, "Sign language avatars: Animation and comprehensibility," *In Intelligent Virtual Agents*, pp. 113–126, Springer, 2011.

[17] I. Kraljevski, S. Chungurski, Z. Gacovski, and S. Arsenovski, "Perceived speech quality estimation using DTW algorithm," *In 16th TELFOR*, Belgrade, Serbia, 2008.

[18] D. Lillo-Martin, "Aspects of the syntax and acquisition of WH-questions in American Sign Language," *In K. Emmorey & H. Lane (Eds.), The Signs of Language Revisited*, pp. 401–413, Mahwah, NJ: Lawrence Erlbaum, 2000.

[19] N. Mana and F. Pianesi, "HMM-based synthesis of emotional facial expressions during speech in synthetic talking heads," *In Proc. of the 8th international conference on Multimodal Interfaces*, pp. 380–387, ACM, 2006.

[20] R. Mitchell, T. Young, B. Bachleda, and M. Karchmer, "How many people use ASL in the United States? Why estimates need updating," *Sign Lang Studies*, vol. 6, no. 3, pp. 306–335, 2006.

[21] I. Mlakar, and M. Rojc, "Towards ECA?s animation of expressive complex behaviour," *In Analysis of Verbal and Nonverbal Communication and Enactment, The Processing Issues*, pp. 185–198, Springer Berlin Heidelberg, 2011.

[22] C. Neidle, D. Kegl, D. MacLaughlin, B. Bahan, and R. G. Lee, "The syntax of ASL: functional categories and hierarchical structure," *Cambridge: MIT Press*, 2000.

[23] M. Ouhyoung, H. S. Lin, Y. T. Wu, Y. S. Cheng, and D. Seifert, "Unconventional approaches for facial animation and tracking," *In SIGGRAPH Asia*, pp. 24, 2012.

[24] T. Pejsa and I. S. Pandzic, "Architecture of an animation system for human characters," *In Proc. 10th Int?l Conf on Telecommunications (ConTEL)*, pp. 171–176, IEEE, 2009.

[25] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *Acoustics, Speech and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[26] C. Schmidt, O. Koller, H. Ney, T. Hoyoux, and J. Piater, "Enhancing gloss-based corpora with facial features using active appearance models," *Third International Symposium on Sign Language Translation and Avatar Technology (SLTAT)*, 2013.

[27] C. Traxler, "The Stanford achievement test, 9th edition: national norming and performance standards for deaf and hard-of-hearing students," *J Deaf Stud & Deaf Educ*, vol. 5, no. 4, pp. 337–348, 2000.

[28] V. M. Velichko and N. G. Zagoruyko, "Automatic recognition of 200 words," *International Journal of Man-Machine Studies*, vol. 2, no. 3, pp. 223–234, 1970.

[29] M. Vlachos, M. Hadjieleftheriou, D. Gunopulos, and E. Keogh, "Indexing multidimensional time-series," *The VLDB Journal?The International Journal on Very Large Data Bases*, vol. 15, no. 1, pp. 1–20, 2006.

[30] Y. Zhang, Q. Ji, Z. Zhu, and B. Yi, "Dynamic facial expression analysis and synthesis with MPEG-4 facial animation parameters," *Circuits and Systems for Video Technology*, vol. 18, no. 10, pp. 1383–1396, 2008.

[31] F. Zhou, F. De la Torre, "Canonical time warping for alignment of human behavior," *In NIPS*, pp. 2286–2294.