Authors' Unofficial Copy of:
Hernisa Kacorri and Matt Huenerfauth. 2016. "Continuous Profile Models in ASL Syntactic Facial Expression Synthesis." Proceedings of the 54rd Annual Meeting on Association for Computational Linguistics (ACL '16). Association for Computational Linguistics, Stroudsburg, PA, USA.

Continuous Profile Models in ASL Syntactic Facial Expression Synthesis

Hernisa Kacorri
Carnegie Mellon University
Human-Computer Interaction Institute
5000 Forbes Avenue
Pittsburgh, PA 15213, USA
hkacorri@andrew.cmu.edu

Matt Huenerfauth
Rochester Institute of Technology
B. Thomas Golisano College of
Computing and Information Sciences
152 Lomb Memorial Drive
Rochester, NY 14623, USA
matt.huenerfauth@rit.edu

Abstract

To create accessible content for deaf users, we investigate automatically synthesizing animations of American Sign Language (ASL), including grammatically important facial expressions and head movements. Based on recordings of humans performing various types of syntactic face and head movements (which include idiosyncratic variation), we evaluate the efficacy of Continuous Profile Models (CPMs) at identifying an essential "latent trace" of the performance, for use in producing ASL animations. A metric-based evaluation and a study with deaf users indicated that this approach was more effective than a prior method for producing animations.

1 Introduction and Motivation

While there is much written content online, many people who are deaf have difficulty reading text or may prefer sign language. For example, in the U.S., standardized testing indicates that a majority of deaf high school graduates (age 18+) have a fourth-grade reading level or below (Traxler, 2000) (U.S. fourth-grade students are typically age 9). While it is possible to create video-recordings of a human performing American Sign Language (ASL) for use on websites, updating such material is expensive (i.e., re-recording). Thus, researchers investigate technology to automate the synthesis of animations of a signing virtual human, to make it more cost-effective for organizations to provide sign language content online that is easily updated and maintained. Animations can be automatically synthesized from a symbolic specification of the message authored by a human or perhaps by machine translation, e.g. (Ebling and Glauert, 2013; Filhol et al., 2013; Stein et al., 2012).

1.1 ASL Syntactic Facial Expressions

Facial expressions are essential in ASL, conveying emotion, semantic variations, and syntactic structure. Prior research has verified that ASL animations with missing or poor facial expressions are significantly less understandable for deaf users (Kacorri et al., 2014; Kacorri et al., 2013b; Kacorri et al., 2013a). While artists can produce individual animations with beautiful expressions, such work is time-consuming. For efficiently maintainable online content, we need automatic synthesis of ASL from a sparse script representing the lexical items and basic elements of the sentence.

Specifically, we are studying how to model and generate ASL animations that include syntactic facial expressions, conveying grammatical information during entire phrases and therefore constrained by the timing of the manual signs in a phrase (Baker-Shenk, 1983). Generally speaking, in ASL, upper face movements (examined in this paper) convey syntactic information across entire phrases, with the mouth movements conveying lexical or adverbial information.

The meaning of a sequence of signs performed with the hands depends on the co-occuring facial expression. (While we use the term "facial expressions," these phenomena also include movements of the head.) For instance, the ASL sentence "BOB LIKE CHOCOLATE" (English: "Bob likes chocolate.") becomes a yes/no question (English: "Does Bob like chocolate?"), with the addition of a YesNo facial expression during the sentence. The addition of a Negative facial expression during the verb phrase "LIKE CHOCOLATE" changes the meaning of the sentence to "Bob doesn't like chocolate." (The lexical item NOT may optionally be used.) For interrogative questions, a WhQuestion facial expression must occur during the sentence, e.g., "BOB LIKE

WHAT." The five types of ASL facial expressions investigated in this paper include:

- YesNo: The signer raises his eyebrows while tilting the head forward to indicate that the sentence is a polar question.
- WhQuestion: The signer furrows his eyebrows and tilts his head forward during a sentence to indicate an interrogative question, typically with a "WH" word such as what, who, where, when, how, which, etc.
- Rhetorical: The signer raises his eyebrows and tilts his head backward and to the side to indicate a rhetorical question.
- Topic: The signer raises his eyebrows and tilts his head backward during a clause-initial phrase that should be interpreted as a topic.
- Negative: The signer shakes his head left and right during the verb phrase to indicate negated meaning, often with the sign NOT.

1.2 Prior Work

A survey of recent work of several researchers on producing animations of sign language with facial expressions appears in (Kacorri, 2015). There is recent interest in data-driven approaches using facial motion-capture of human performances to generate sign language animations: For example, (Schmidt et al., 2013) used clustering techniques to select facial expressions that co-occur with individual lexical items, and (Gibet et al., 2011) studied how to map facial motion-capture data to animation controls.

In the most closely related prior work, we had investigated how to generate a face animation based on a set of video recordings of a human signer performing facial expressions (Kacorri et al., 2016), with head and face movement data automatically extracted from the video, and with individual recordings labeled as each of the five syntactic types, as listed in section 1.1. We wanted to identify a single exemplar recording in our dataset, for each of the syntactic types, that could be used as the basis for generating the movements of virtual human character. (In a collection of recordings of face and head movement, there will naturally be non-essential individual variation in the movements; thus, it may be desirable to select a recording that is maximally stereotypical of a set of recordings.) To do so, we made use of a variant of Dynamic Time Warping (DTW) as a distance metric to select the recording with minimal pairwise normalized DTW distance from all of the examples of each syntactic type. We had used this "centroid" recording as the basis for producing a novel animation of the face and head movements for a sign language sentence.

2 Method

In this paper, we present a new methodology for generating face and head movements for sign language animations, given a set of human recordings of various syntactic types of facial expressions. Whereas we had previously selected a single exemplar recording of a human performance to serve as a basis for producing an animation (Kacorri et al., 2016), in this work, we investigate how to construct a model that generalizes across the entire set of recordings, to produce an "average" of the face and head movements, which can serve as a basis for generating an animation. To enable comparison of our new methodology to our prior technique, we make use of an identical training dataset as in (Kacorri et al., 2016) and an identical animation rendering pipeline, described in (Huenerfauth and Kacorri, 2015a). Briefly, the animation pipeline accepts a script of the hand location, hand orientation, and hand-shape information to pose and move the arms of the character over time, and it also accepts a file containing a stream of face movement information in MPEG4 Facial Animation Parameters format (ISO/IEC, 1999) to produce a virtual human animation.

2.1 Dataset and Feature Extraction

ASL is a low-resource language, and it does not have a writing system in common use. Therefore, ASL corpora are generally small in size and in limited supply; they are usually produced through manual annotation of video recordings. Thus, researchers generally work with relatively small datasets. In this work, we make use of two datasets that consist of video recordings of humans performing ASL with annotation labeling the times in the video when each of the five types of syntactic facial expressions listed in section 1.1 occur.

The training dataset used in this study was described in (Kacorri et al., 2016), and consists of 199 examples of facial expressions performed by a female signer recorded at Boston University. While the Training dataset can naturally be partitioned into five subsets, based on each of the five syntactic facial expression types, because adjacent

Type	Subgroup "_A" (Num. of Videos)	Subgroup "_B" (Num. of Videos)
YesNo	Immediately preceded by a facial expression with raised eyebrows, e.g. Topic. (9)	Not immediately preceded by an eyebrow-raising expression. (10)
WhQuestion	Performed during a single word, namely the wh-word (e.g., what, where, when). (4)	Performed during a phrase consisting of multiple words. (8)
Rhetorical	Performed during a single word, namely the wh-word (e.g., what, where, when). (2)	Performed during a phrase consisting of multiple words. (8)
Topic	Performed during a single word. (29)	Performed during a phrase consisting of multiple words. (15)
Negative	Immediately preceded by a facial expression with raised eyebrows, e.g. Topic. (16)	Not immediately preceded by eyebrow-raising expression. (25)

Table 1: Ten subgroups of the training dataset.

facial expressions or phrase durations may affect the performance of ASL facial expressions, in this work, we sub-divide the dataset further, into ten sub-groups, as summarized in Table 1.

The "gold-standard" dataset used in this study was shared with the research community by (Huenerfauth and Kacorri, 2014); we use 10 examples of ASL facial expressions (one for each sub-group listed in Table 1) performed by a male signer who was recorded at the Linguistic and Assistive Technologies laboratory.

To extract face and head movement information from the video, a face-tracker (Visage, 2016) was used to produce a set of MPEG4 facial animation parameters for each frame of video: These values represent face-landmark or head movements of the human appearing in the video, including 14 features used in this study: head_x, head_y, head_z, head_pitch, head_yaw, head_roll, raise_l_i_brow, raise_r_i_brow, raise_l_m_brow, raise_r_m_brow, raise_l_o_brow, raise_r_o_brow, squeeze_l_brow, squeeze_r_brow. The first six values represent head location and orientation. The next six values represent vertical movement of the outer ("o_"), middle ("m_"), or inner ("i_") portion of the right ("r_") or left ("l_") eyebrows. The final values represent horizontal movement of the eyebrows.

2.2 Continuous Profile Models (CPM)

Continuous Profile Model (CPM) aligns a set of related time series data while accounting for changes in amplitude. This model has been previously evaluated on speech signals and on other biological time-series data (Listgarten et al., 2004). With the assumption that a noisy, stochastic process generates the observed time series data, the approach automatically infers the underlying noiseless representation of the data, the so-called "latent trace." Figure 6 (on the last page of this paper) shows an example of multiple time series in unaligned and aligned space, with CPM identifying the the latent trace.

Given a set K of observed time series $\vec{x}^k = (x_1^k, x_2^k, ..., x_N^k)$, CPM assumes there is a latent trace $\vec{z} = (z_1, z_2, ..., z_M)$. While not a requirement of the model, the length of the time series data is assumed to be the same (N) and the length of the latent trace used in practice is $M = (2+\varepsilon)N$, where an ideal M would be large relative to N to allow precise mapping between observed data and an underlying point on the latent trace. Higher temporal resolution of the latent trace also accommodates flexible alignments by allowing an observational series to advance along the latent trace in small or large jumps (Listgarten, 2007).

Continuous Profile Models (CPMs) build on Hidden Markov Models (HMMs) (Poritz, 1988) and share similarities with Profile HMMs which augment HMMs by two constrained-transition states: 'Insert' and 'Delete' (emitting no observations). Similar to the Profile HMM, the CPM has strict left-to-right transition rules, constrained to only move forward along a sequence. Figure 1 includes a visualization we created, which illustrates the graphical model of a CPM.

2.3 Obtaining the CPM Latent Trace

We applied the CPM model to time align and coherently integrate time series data from multiple ASL facial expression performances of a particular type, e.g., Topic_A as listed in section 2.1, with the goal of using the inferred 'latent traces' to drive ASL animations with facial expressions of that type. This section describes our work to train the CPM and to obtain the latent traces; implementation details appear in Appendix A.

The input time-series data for each CPM model is the face and head movement data extracted from ASL videos of one of the facial expression types,

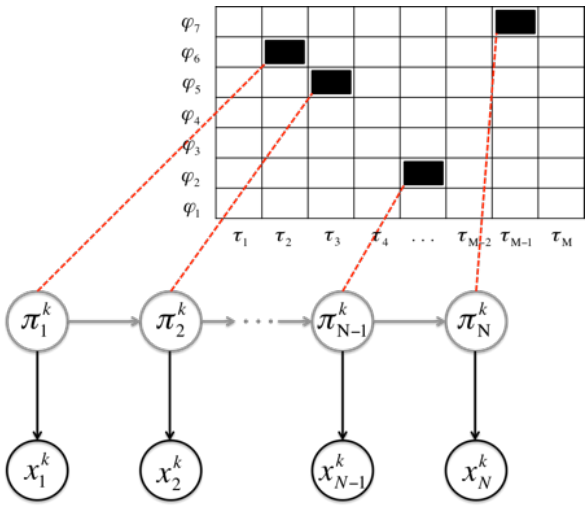


Figure 1: Depiction of a CPM for series x^k, with hidden state variables π_i^k underlying each observation x_i^k. The table illustrates the state-space: time-state/scale-state pairs mapped to the hidden variables, where time states belong to the integer set $(1...M)$ and scale states belong to an ordered set, here with 7 evenly spaced scales in logarithmic space as in (Listgarten et al., 2004).

as shown in Table 2. For each dataset, all the training examples are stretched (resampled using cubic interpolation) to meet the length of the longest example in the set. The length of time series, N, corresponds to the duration in video frames of the longest example in the data set. The recordings in the training set have 14 dimensions, corresponding to the 14 facial features listed in Section 2.1. As discussed above, the latent trace has a time axis of length M, which is approximately double the temporal resolution of the original training examples.

CPM Models	**Training Data** $\#Examples \times N \times \#Features$	**Latent Trace** $M \times \#Features$ where $M = (2+\varepsilon)N$
YesNo_A	9 x 51 x 14	105 x 14
YesNo_B	10 x 78 x 14	160 x 14
WhQuestion_A	4 x 24 x 14	50 x 14
WhQuestion_B	8 x 41 x 14	84 x 14
Rhetorical_A	2 x 16 x 14	33 x 14
Rhetorical_B	8 x 55 x 14	113 x 14
Topic_A	29 x 29 x 14	60 x 14
Topic_B	15 x 45 x 14	93 x 14
Negative_A	16 x 67 x 14	138 x 14
Negative_B	25 x 76 x 14	156 x 14

Table 2: Training data and the obtained latent traces for each of the CPM models on ASL facial expression subcategories.

To demonstrate our experiments, Figure 6 illustrates one of the subcategories, Rhetorical_B. (This figure appears at the end of the paper, due to its large size.) We illustrate the training set, before and after the alignment and amplitude normalization with the CPM, and the obtained latent trace for this subcategory. Figure 6a and Figure 6b illustrate each of the 8 training examples with a subplot extending from $[0, N]$ in the x-axis, which is the observed time axis in video frames. Each of the 14 plots represents one of the head or face features. Figure 6c illustrates the learned latent trace with a subplot extending from $[0, M]$ in the x-axis, which is the latent time axis. While the training set for this subcategory is very small and has high variability, upon visual inspection of Figure 6, we can observe that the learned latent trace shares similarities with most of the time series in the training set without being identical to any of them.

We expect that during the Rhetorical facial expression (Section 2.1), the signer's eyebrows will rise and the head will be tilted back and to the side. In the latent trace, the inner, middle, and outer portions of the left eyebrow rise (Figure 6c, plots 7, 9, 11), and so do the inner, middle, and outer portions of the right eyebrow (Figure 6c, plots 8, 10, 12). Note how the height of the lines in those plots rise, which indicates increased eyebrow height. For the Rhetorical facial expression, we would also expect symmetry in the horizontal displacement of the eyebrows, and we see such mirroring in the latent-trace: In (Figure 6c, plots 13-14), note the tendency for the line in plot 13 (left eyebrow) to increase in height as the line in plot 14 (right eyebrow) decreases in height, and vice versa.

3 Evaluation

This section presents two forms of evaluation of the CPM latent trace model for ASL facial expression synthesis. In Section 3.1, the CPM model will be compared to a "gold-standard" performance of each sub-category of ASL facial expression using a distance-metric-based evaluation, and in Section 3.2, the results of a user-study will be presented, in which ASL signers evaluated animations of ASL based upon the CPM model.

To provide a basis of comparison, in this section, we evaluate the CPM approach in comparison to an alternative approach that we call 'Centroid', which we described in prior work in (Ka-

corri et al., 2016), where we used a multivariate DTW to select one of the time series in the training set as a representative performance of the facial expression. The centroid examples are actual recordings of human ASL signers that are used to drive an animation. Appendix A lists the codenames of the videos from the training dataset selected as centroids and the codenames of the videos used in the gold-standard dataset (Huenerfauth and Kacorri, 2014).

3.1 Metric Evaluation

The gold-standard recordings of a male ASL signer were described in Section 2.1. In addition to the video recordings (which were processed to extract face and head movement data), we have annotation of the timing of the facial expressions and the sequence of signs performed on the hands. To compare the quality of our CPM model and that of the Centroid approach, we used each method to produce a candidate sequence of face and head movements for the sentence performed by the human in the gold-standard recording. Thus, the extracted facial expressions from the human recording can serve as a gold standard for how the face and head should move. In this section, we compare: (a) the distance of the CPM latent trace from the gold standard to (b) the distance of the centroid form the gold standard. It is notable that these gold-standard recordings were previously "unseen" during the creation of the CPM or Centroid models, that is, they were not used in the training data set during the creation of either model.

Since there was variability in the length of the latent trace, centroid, and gold-standard videos, for a fairer comparison, we first resampled these time series, using cubic interpolation, to match the duration (in milliseconds) of the gold-standard ASL sentence, and then we used multivariate DTW to estimate their distance, following the methodology of (Kacorri et al., 2016) and (Kacorri and Huenerfauth, 2015). In prior work (Kacorri and Huenerfauth, 2015), we had shown that a scoring algorithm based on DTW had moderate (yet significant) correlation with scores that participants assigned to ASL animation with facial expressions.

Figure 2 shows an example of a DTW distance scoring between the gold standard and each of the latent trace and the centroid, for one face feature

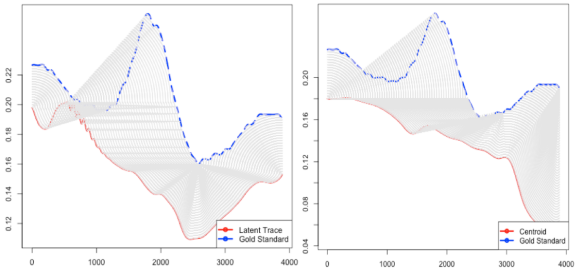


Figure 2: DTW distances on the squeeze_l_brow feature (left eyebrow horizontal movement), during a Negative_A facial expression: (left) between the CPM latent trace and gold standard and (right) between the centroid and gold standard. The timeline is given in milliseconds.

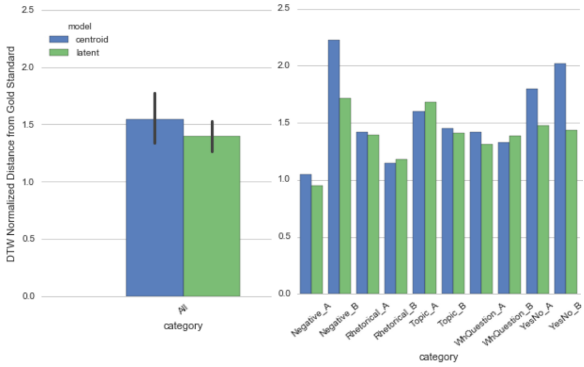


Figure 3: Overall normalized DTW distances for latent trace and centroid (left) and per each subcategory of ASL facial expression (right).

(horizontal movement of the left eyebrow) during a Negative_A facial expression. Given that the centroid and the training data for the latent trace are driven by recordings of a (female) signer and the gold standard is a different (male) signer, there are differences between these facial expressions due to idiosyncratic aspects of individual signers. Thus the metric evaluation in this section is challenging because it is an inter-signer evaluation.

Figure 3 illustrates the overall calculated DTW distances, including a graph with the results broken down per subcategory of ASL facial expression. The results indicate that the CPM latent trace is closer to the gold standard than the centroid is. Note that the distance values are not zero since the latent trace and the centroid are being compared to a recording from a different signer on novel, previously unseen, ASL sentences. The results in these graphs suggest that the latent trace model out-performed the centroid approach.

Figure 4: Screenshots of YesNo_A stimuli of three types: a) neutral, b) centroid, and c) latent trace.

3.2 User Evaluation

To further assess our ASL synthesis approach, we conducted a user study where ASL signers watched short animations of ASL sentences with identical hand movements but differing in their face, head, and torso movements. There were three conditions in this between-subjects study: a) animations with a static neutral face throughout the animation (as a lower baseline), b) animations with facial expressions driven by the centroid human recording, and c) animations with facial expressions driven by the CPM latent trace based on multiple recordings of a human performing that type of facial expression. Figure 4 illustrates screenshots of each stimulus type for a YesNo_A facial expression. The specific sentences used for this study were drawn from a standard test set of stimuli released to the research community by (Huenerfauth and Kacorri, 2014) for evaluating animations of sign language with facial expressions.

All three types of stimuli (neutral, centroid and latent trace), shared identical animation-control scripts specifying the hand and arm movements; these scripts were hand-crafted by ASL signers in a pose-by-pose manner. For the neutral animations, we did not specify any torso, head, nor face movements; rather, we left them in their neutral pose throughout the sentences. As for the centroid and latent trace animations, we applied the head and face movements (as specified by the centroid model or by the latent trace model) only to the portion of the animation where the facial expression of interest occurs, leaving the head and face for the rest of the animation to a neutral pose. For instance, during a stimulus that contains a Wh-question, the face and head are animated only during the Wh-question, but they are left in a neutral pose for the rest of the stimulus (which may include other sentences). The period of time when the facial expression occurred was time-aligned with the subset of words (the sequence of signs performed on the hands) for the appropriate syntactic domain; the phrase-beginning and phrase-ending was aligned with the performance of the facial expression. Thus, the difference in appearance between our animation stimuli was subtle: The only portion of the animations that differed between the three conditions (neutral, centroid, and latent-trace) was the face and the head movements during the span of time when the syntactic facial expression should occur (e.g., during the Wh-question).

We resampled the centroid and CPM time series, using cubic interpolation, to match the duration (in milliseconds) of the animation they would be applied to. To convert the centroid and latent trace time series into the input for the animation-generation system, we used the MPEG4-features-to-animation pipeline described in (Kacorri et al., 2016). That platform is based upon the open-source EMBR animation system for producing human animation (Heloir and Kipp, 2009); specifically, the facial expressions were represented as an EMBR PoseSequence with a pose defined every 133 milliseconds.

In prior work (Huenerfauth and Kacorri, 2015b), we investigated key methodological considerations in conducting a study to evaluate sign language animations with deaf users, including the use of appropriate baselines for comparison, appropriate presentation of questions and instructions, demographic and technology experience factors influencing acceptance of signing avatars, and other factors that we have considered in the design of this current study. Our recent work (Kacorri et al., 2015) has established a set of demographic and technology experience questions which can be used to screen for the most critical participants in a user study of ASL signers to evaluate animation. Specifically, we screened for participants that identified themselves as "deaf/Deaf" or "hard-of-hearing," who had grown up using ASL at home or had attended an ASL-based school as a young child, such as a residential or daytime school.

Deaf researchers (all fluent ASL signers) recruited and collected data from participants, during meetings conducted in ASL. Initial advertise-

ments were sent to local email distribution lists and Facebook groups. A total of 17 participants met the above criteria, where 14 participants self-identified as deaf/Deaf and 3 as hard-of-hearing. Of our participants in the study, 10 had attended a residential school for deaf students, and 7, a daytime school for deaf students. 14 participants had learned ASL prior to age 5, and the remaining 3 had been using ASL for over 7 years. There were 8 men and 9 women of ages 19-29 (average age 22.8). In prior work, we (Kacorri et al., 2015) have advocated that participants in studies evaluating sign language animation complete a two standardized surveys about their technology experience (MediaSharing and AnimationAttitude) and that researchers report these values for participants, to enable comparison across studies. In our study, participant scores for MediaSharing varied between 3 and 6, with a mean score of 4.3, and scores for AnimationAttitude varied from 2 to 6, with a mean score of 3.8.

At the beginning of the study, participants viewed a sample animation, to familiarize them with the experiment and the questions they would be asked about each animation. (This sample used a different stimulus than the other ten animations shown in the study.) Next, they responded to a set of questions that measured their subjective impression of each animation, using a 1-to-10 scalar response. Each question was conveyed using ASL through an onscreen video, and the following English question text was shown on the questionnaire: (a) Good ASL grammar? (10=Perfect, 1=Bad); (b) Easy to understand? (10=Clear, 1=Confusing); (c) Natural? (10=Moves like person, 1=Like robot). These questions have been used in many prior experimental studies to evaluate animations of ASL, e.g. (Kacorri and Huenerfauth, 2015), and were shared with research community as a standard evaluation tool in (Huenerfauth and Kacorri, 2014). To calculate a single score for each animation, the scalar response scores for the three questions were averaged.

Figure 5 shows distributions of subjective scores as boxplots with a 1.5 interquartile range (IQR). For comparison, means are denoted with a star and their values are labeled above each boxplot. When comparing the subjective scores that participants assigned to the animations in Figure 5, we found a significant difference (Kruskal-Wallis test used since the data was not normally

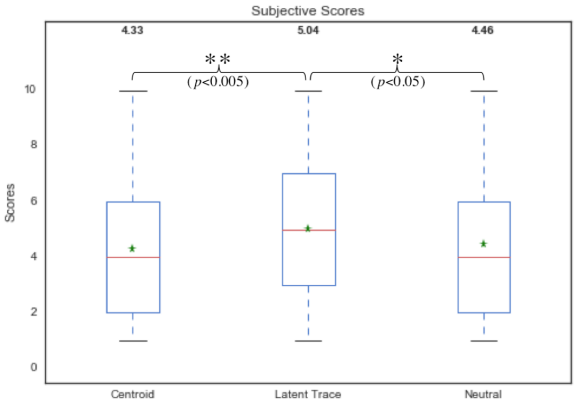


Figure 5: Subjective scores for centroid, latent trace, and neutral animations.

distributed) between the latent trace and centroid ($p < 0.005$) and between the latent trace and neutral ($p < 0.05$).

In summary, our CPM modeling approach for generating an animation out-performed an animation produced from an actual recording of a single human performance (the "centroid" approach). In prior methodological studies, we demonstrated that it is valid to use either videos of humans or animations (driven by a human performance) as the baseline for comparison in a study of ASL animation (Kacorri et al., 2013a). As suggested by Figure 4, the differences in face and head movements between the Centroid and CPM conditions were subtle, yet fluent ASL signers rated the CPM animations higher in this study.

4 Conclusion and Future Work

To facilitate the creation of ASL content that can easily be updated or maintained, we have investigated technologies for automating the synthesis of ASL animations from a sparse representation of the message. Specifically, this paper has focused on the synthesis of syntactic ASL facial expressions, which are essential to sentence meaning, using a data-driven methodology in which recordings of human ASL signers are used as a basis for generating face and head movements for animation. To avoid idiosyncratic aspects of a single performance, we have modeled a facial expression based on the underlying trace of the movement trained on multiple recordings of different sentences where this type of facial expression occurs. We obtain the latent trace with Continuous Profile Model (CPM), a probabilistic generative model that relies on Hidden Markov Models. We

assessed our modeling approach through comparison to an alternative centroid approach, where a single performance was selected as a representative. Through both a metric evaluation and an experimental user study, we found that the facial expressions driven by our CPM models produce high-quality facial expressions that are more similar to human performance of novel sentences.

While this work used the latent trace as the basis for animation, in future work, we also plan to explore methods for sampling from the model to produce variations in face and head movement. In addition, to aid CPM convergence to a good local optimum, in future work we will investigate dimensionality reduction approaches that are reversible such as Principal Component Analysis (Pearson, 1901) and other pre-processing approaches similar to (Listgarten, 2007), where the training data set is coarsely pre-aligned and pre-scaled based on the center of mass of the time series. In addition we plan to further investigate how to fine-tune some of the hyper parameters of the CPM such as spline scaling, single global scaling factor, convergence tolerance, and initialization of the latent trace with a centroid. In subsequent work, we would like to explore alternatives for enhancing CPMs by incorporating contextual features in the training data set such as timing of hand movements, and preceding, succeeding, and co-occurring facial expressions.

Acknowledgments

This material is based upon work supported by the National Science Foundation under award number 1065009 and 1506786. This material is also based upon work supported by the Science Fellowship and Dissertation Fellowship programs of The Graduate Center, CUNY. We are grateful for support and resources provided by Ali Raza Syed at The Graduate Center, CUNY, and by Carol Neidle at Boston University.

References

Charlotte Baker-Shenk. 1983. A microanalysis of the nonmanual components of questions in american sign language.

Sarah Ebling and John Glauert. 2013. Exploiting the full potential of jasigning to build an avatar signing train announcements. In *Proceedings of the Third International Symposium on Sign Language Translation and Avatar Technology (SLTAT), Chicago, USA, October*, volume 18, page 19.

Michael Filhol, Mohamed N Hadjadj, and Benoît Testu. 2013. A rule triggering system for automatic text-to-sign translation. *Universal Access in the Information Society*, pages 1–12.

Sylvie Gibet, Nicolas Courty, Kyle Duarte, and Thibaut Le Naour. 2011. The signcom system for data-driven animation of interactive virtual signers: Methodology and evaluation. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 1(1):6.

Alexis Heloir and Michael Kipp. 2009. Embr–a realtime animation engine for interactive embodied agents. In *Intelligent Virtual Agents*, pages 393–404. Springer.

Matt Huenerfauth and Hernisa Kacorri. 2014. Release of experimental stimuli and questions for evaluating facial expressions in animations of american sign language. In *Proceedings of the 6th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel, The 9th International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland*.

Matt Huenerfauth and Hernisa Kacorri. 2015a. Augmenting embr virtual human animation system with mpeg-4 controls for producing asl facial expressions. In *International Symposium on Sign Language Translation and Avatar Technology*, volume 3.

Matt Huenerfauth and Hernisa Kacorri. 2015b. Best practices for conducting evaluations of sign language animation. *Journal on Technology and Persons with Disabilities*, 3.

ISO/IEC. 1999. Information technology—Coding of audio-visual objects—Part 2: Visual. ISO 14496-2:1999, International Organization for Standardization, Geneva, Switzerland.

Hernisa Kacorri and Matt Huenerfauth. 2015. Evaluating a dynamic time warping based scoring algorithm for facial expressions in asl animations. In *6th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, page 29.

Hernisa Kacorri, Pengfei Lu, and Matt Huenerfauth. 2013a. Effect of displaying human videos during an evaluation study of american sign language animation. *ACM Transactions on Accessible Computing (TACCESS)*, 5(2):4.

Hernisa Kacorri, Pengfei Lu, and Matt Huenerfauth. 2013b. Evaluating facial expressions in american sign language animations for accessible online information. In *Universal Access in Human-Computer Interaction. Design Methods, Tools, and Interaction Techniques for eInclusion*, pages 510–519. Springer.

Hernisa Kacorri, Allen Harper, and Matt Huenerfauth. 2014. Measuring the perception of facial expressions in american sign language animations with eye tracking. In *Universal Access in Human-Computer Interaction. Design for All and Accessibility Practice*, pages 553–563. Springer.

Hernisa Kacorri, Matt Huenerfauth, Sarah Ebling, Kasmira Patel, and Mackenzie Willard. 2015. Demographic and experiential factors influencing acceptance of sign language animation by deaf users. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*, pages 147–154. ACM.

Hernisa Kacorri, Ali Raza Syed, Matt Huenerfauth, and Carol Neidle. 2016. Centroid-based exemplar selection of asl non-manual expressions using multidimensional dynamic time warping and mpeg4 features. In *Proceedings of the 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining, The 10th International Conference on Language Resources and Evaluation (LREC 2016), Portoroz, Slovenia*. http://huenerfauth.ist.rit.edu/pubs/lrec2016.pdf.

Hernisa Kacorri. 2015. Tr-2015001: A survey and critique of facial expression synthesis in sign language animation. Technical report, The Graduate Center, CUNY. http://academicworks.cuny.edu/gc_cs_tr/403.

Jennifer Listgarten, Radford M Neal, Sam T Roweis, and Andrew Emili. 2004. Multiple alignment of continuous time series. In *Advances in neural information processing systems*, pages 817–824.

Jennifer Listgarten. 2007. *Analysis of sibling time series data: alignment and difference detection*. Ph.D. thesis, University of Toronto.

Carol Neidle, Jingjing Liu, Bo Liu, Xi Peng, Christian Vogler, and Dimitris Metaxas. 2014. Computer-based tracking, analysis, and visualization of linguistically significant nonmanual events in american sign language (asl). In *LREC Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel*. Citeseer.

Karl Pearson. 1901. Principal components analysis. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 6(2):559.

Alan B Poritz. 1988. Hidden markov models: A guided tour. In *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, pages 7–13. IEEE.

Christoph Schmidt, Oscar Koller, Hermann Ney, Thomas Hoyoux, and Justus Piater. 2013. Enhancing gloss-based corpora with facial features using active appearance models. In *International Symposium on Sign Language Translation and Avatar Technology*, volume 2.

Daniel Stein, Christoph Schmidt, and Hermann Ney. 2012. Analysis, preparation, and optimization of statistical sign language machine translation. *Machine Translation*, 26(4):325–357.

Carol Bloomquist Traxler. 2000. The stanford achievement test: National norming and performance standards for deaf and hard-of-hearing students. *Journal of deaf studies and deaf education*, 5(4):337–348.

Technologies Visage. 2016. Face tracking. https://visagetechnologies.com/products-and-services/visagesdk/facetrack. Accessed: 2016-03-10.

A Appendix: Supplemental Material

In Section 2.3, we made use of a freely available CPM implementation available from http://www.cs.toronto.edu/~jenn/CPM/ in MATLAB, Version 8.5.0.197613 (R2015a).

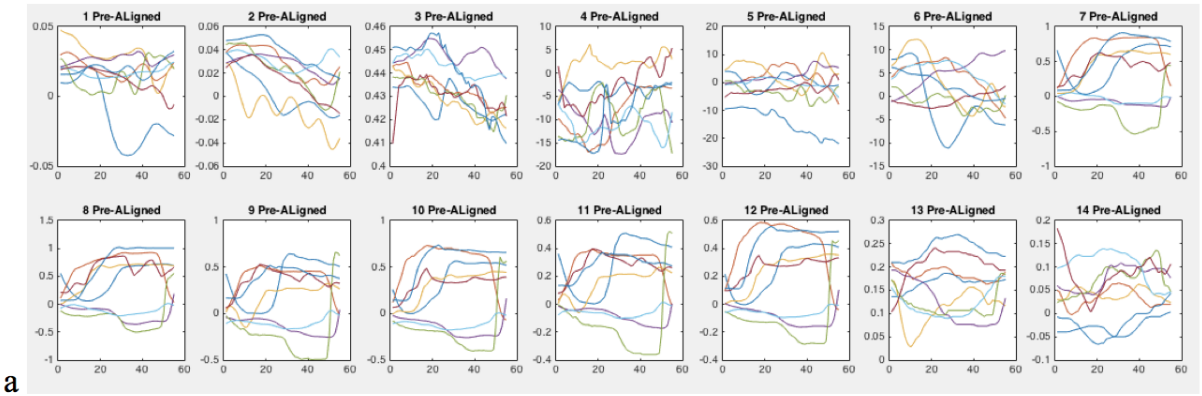
One parameter for regularizing the latent trace (Listgarten, 2007) is a smoothing parameter (λ), with values being dataset-dependent. To select a good λ, we experimented with held-out data and found that $\lambda = 4$ and $NumberOfIterations = 3$ resulted in a latent trace curve that captures the shape of the ASL features well. Other CPM parameters were:

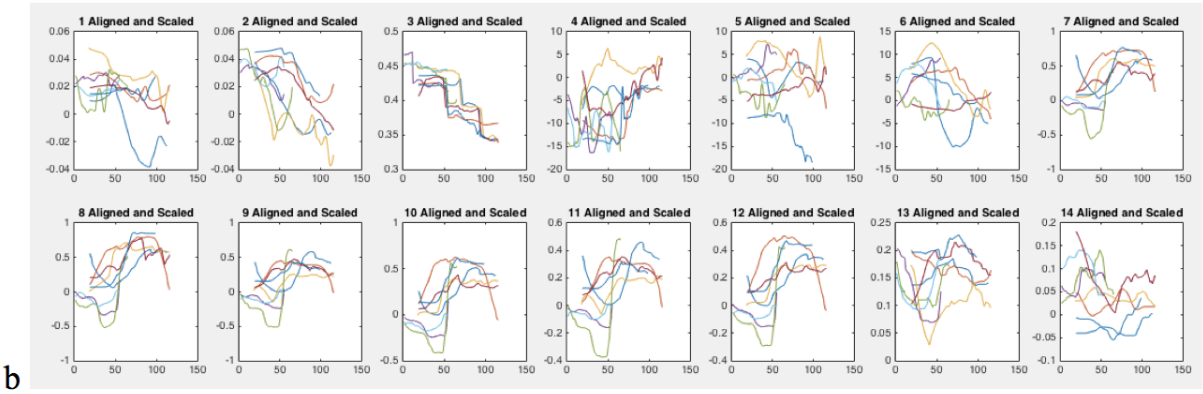
- $USE\_SPLINE = 0$: if set to 1, uses spline scaling rather than HMM scale states
- $oneScaleOnly = 0$: no HMM scale states (only a single global scaling factor is applied to each time series.)
- $extraPercent(\varepsilon)$ = 0.05: slack on the length of the latent trace M, where $M = (2 + \varepsilon)N$.
- $learnStateTransitions = 0$: whether to learn the HMM state-transition probabilities
- $learnGlobalScaleFactor = 1$: learn single global scale factor for each time series

Section 3.1 described how the centroids were selected from among videos in the Boston University dataset (Neidle et al., 2014), and the gold standard videos were selected from among videos in a different dataset (Huenerfauth and Kacorri, 2014). Table 3 lists the code names of the selected videos, using the nomenclature of each dataset.

Subcategory	Centroid Codename	Gold-Standard Codename
YesNo_A	2011-12-01_0037-cam2-05	Y4
YesNo_B	2011-12-01_0037-cam2-09	Y3
WhQuestion_A	2011-12-01_0038-cam2-05	W1
WhQuestion_B	2011-12-01_0038-cam2-07	W2
Rhetorical_A	2011-12-01_0041-cam2-04	R3
Rhetorical_B	2011-12-01_0041-cam2-02	R9
Topic_A	2012-01-27_0050-cam2-05	T4
Topic_B	2012-01-27_0051-cam2-09	T3
Negative_A	2012-01-27_0051-cam2-03	N2
Negative_B	2012-01-27_0051-cam2-30	N5

Table 3: Codenames of videos selected as centoids and gold standards for comparison in section 3.1.





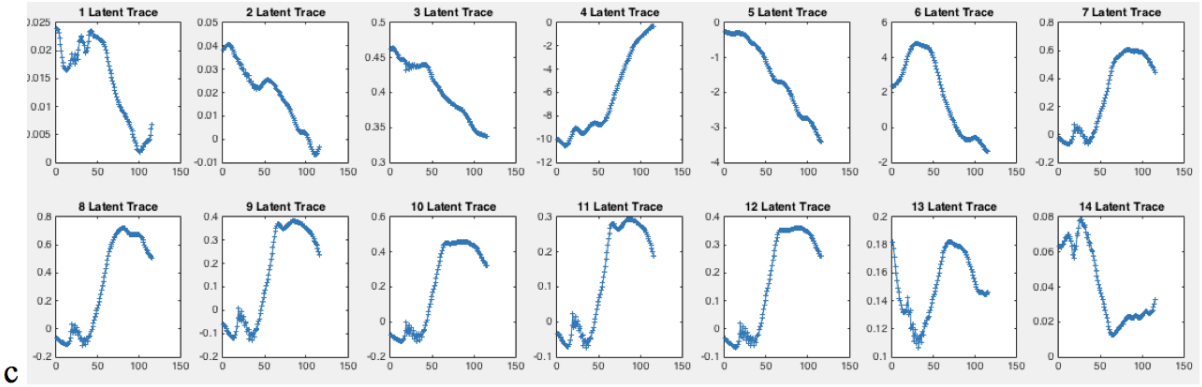


Figure 6: Example of CPM modeling for Rhetorical_B: (a) training examples before CPM (each plot shows one of the 14 face features over time, with 8 colored lines in each plot showing each of the 8 training examples), (b) after CPM time-alignment and rescaling, and (c) the final latent trace based upon all 8 examples.