Comparison of Finite-Repertoire and Data-Driven Facial Expressions for Sign Language Avatars

Hernisa Kacorri[1] and Matt Huenerfauth[2]

[1]City University of New York (CUNY)
Doctoral Program in Computer Science, The Graduate Center
365 Fifth Ave, New York, NY 10016 USA
hkacorri@gc.cuny.edu
[2]Rochester Institute of Technology (RIT)
Golisano College of Computing & Information Sciences
152 Lomb Memorial Drive Rochester, NY 14623 USA
matt.huenerfauth@rit.edu

Abstract. To support our research on ASL animation synthesis, we have adopted and enhanced a new virtual human animation platform that provides us with greater fine-grained control of facial movements than our previous platform. To determine whether this new platform is sufficiently expressive to generate understandable ASL animations, we analyzed responses collected from deaf participants who evaluated four types of animations: generated by our old or new animation platform, and with or without facial expressions performed by the character. For animations without facial expressions, our old and new platforms had equivalent comprehension scores; for those with facial expressions, our new platform had higher scores. In addition, this paper demonstrates a methodology by which sign language animation researchers can document transitions in their animation platforms or avatar appearance. Performing such an evaluation enables future readers to compare published results over time, both before and after such a transition in animation technology.

Keywords. American Sign Language · accessibility technology for people who are deaf · facial expression · animation · evaluation · user study

1 Introduction

Many people who are deaf have difficulty reading information content in the form of written language text, due to limitations in spoken language exposure and other educational factors. For example, in the U.S., standardized testing has revealed that many deaf adults graduating from secondary school (age 18) perform at or below fourth-grade English reading level (typically age 10) [10][25]. Thus, if the text on online media is too complex, these adults may not comprehend the message. However, many of these users have sophisticated fluency in American Sign Language (ASL), which is a distinct language from English and is the primary mean of commu-

nication for more than 500,000 people in the U.S. [19]. Technology that can synthesize ASL animations from written text has accessibility benefits for these individuals.

While incorporating videos of real human signers in websites and other media would make information accessible to deaf users, this approach is not ideal: the recordings are difficult and often prohibitively expensive to update, leading to out-of-date information. Further, there is no way to support dynamically generated content from a query. For these reasons, we investigate computer-synthesized animations (from an easy-to-update script as input), which allow for frequent updating, automatic production of messages (via natural language generation or machine translation techniques), wiki-style applications in which multiple authors script a message in ASL collaboratively, or scripting of messages by a single human author.

In ASL, a signer's facial expressions and head movements are essential to the fluency of the performance; these face and head movements convey: emotion, variations in word meaning, and grammatical information during entire sentences or syntactic phrases. This paper focuses on this third use, which is necessary for expressing questions or negation. In fact, a sequence of signs performed on the hands can have different meanings, depending on the syntactic facial expression that co-occurs [20]. E.g., a declarative sentence (ASL: "MARY LIKE BOOK" / English: "Mary likes the book.") can become a Yes-No question (English: "Does Mary like the book?"), with the addition of a Yes-No Question facial expression. This is performed by the signer raising their eyebrows and tilting their head forward during the sentence.

Similarly, the addition of a Negation facial expression (the signer shakes their head left and right while furrowing their eyebrows somewhat) during the verb phrase "LIKE BOOK" can change the meaning of the sentence to "Mary doesn't like the book." It is important to note that the word NOT is actually optional, but the facial expression is required [28]. For interrogative questions (with a WH word like "what, who, where"), a WH-Question facial expression (head tilted forward, eyebrows furrowed) is required during the sentence, e.g., "MARY LIKE WHAT."

There is variation in how these facial expressions are performed during a sentence, based on the length of the phrase when the facial expression occurs, the location of particular words during the phrase (e.g., NOT or WHAT), the facial expressions that precede or follow, the overall speed of signing, and other factors. Thus, in order to build our ASL animation synthesis system, we cannot simply record a single version of this facial expression and replay it whenever needed. We must be able to synthesize the natural variations in the performance of a facial expression, based on these complex linguistic factors, in order to produce understandable results. The production of grammatical facial expressions and head movements, which must be time-coordinated with specific manual signs, is crucial for the interpretation of signed sentences and acceptance of this technology by the users [13][18].

In order to support our research on facial expressions, our laboratory has recently adopted and enhanced a new animation platform (details in section 3), which provides greater control over the face movements of our virtual human character. Since we had conducted several years of research using a previous platform, we needed to compare the new avatar to the old avatar, in regard to their understandability and naturalness. This paper presents the results of experiments with deaf participants evaluating ani-

mations from both of these platforms. This comparison will enable future researchers to compare our published results before and after this platform change, and it will allow us to evaluate whether our new avatar is sufficiently understandable to support our future work. Further, this paper demonstrates a methodology by which sign language animation researchers can empirically evaluate alternative virtual human animation platforms to enable more specific comparisons between systems.

2 Related Work

Sign language avatars have been adopted by researchers that seek to make information accessible to people who are deaf and hard-of-hearing in different settings such as train announcements (e.g. [3]) and education (e.g., [2][6]). While authoring by non-experts is one of the research focuses when designing a new animation platform (e.g. [1][9][26]), typically these platforms are seen as the output medium for machine translation tools that will allow text-to-signing, e.g. [4][7][11]. There has been recent work by several groups (e.g. [5][22][27]) to improve the state-of-the-art of facial expressions and non-manual signals for sign language animation, surveyed in [16].

Other researchers are also studying synthesis of facial expressions for sign language animation, e.g., interrogative (WH-word) questions with co-occurrence of affect [27], using computer-vision data to produce facial expressions during specific words [22], etc. However, few researchers have conducted user studies comparing different avatars with facial expressions. A user study by Smith and Nolan [24] indicated that the addition of emotional facial expressions to a "human-looking" avatar was more successful than a caricature avatar when comparing native signers' comprehension scores. Still, both avatars were created within the same sign language animation platform. Kipp et al. [18] asked native signers' feedback on 6 avatars, created either by an animation synthesis platform or by a 3D artist, with a varying level of facial expressions each. However, the stimuli used in the assessment were not the same, they differ in content and sign language.

3 Finite-Repertoire vs. Data-Driven Facial Expressions

This section explains how our lab has recently made a major change in the avatar platform that is used to synthesize virtual humans for ASL animations. After explaining both platforms, this section will outline our research questions and hypotheses that motivated our comparison of both platforms in experiments with deaf participants.

3.1 Finite-Repertoire Facial Expressions in Our Old Avatar Platform

Our prior animation platform was based on a commercially available American Sign Language authoring tool, VCOM3D Sign Smith Studio [26], which allows users to produce animated ASL sentences by arranging a timeline of animated signs from a prebuilt or user-defined vocabulary. The software includes a library of facial expres-

sions that can be applied over a single sign or multiple manual signs, as shown in Fig. 1. While this finite repertoire covers adverbial, syntactic, and emotional categories of facial expressions, the user cannot modify the intensity of the expressions over time, nor can multiple facial expressions be combined to co-occur. Because such co-occurrences or variations in intensity are necessary for many ASL sentences, we were motivated to investigate alternative animation platforms for our research.

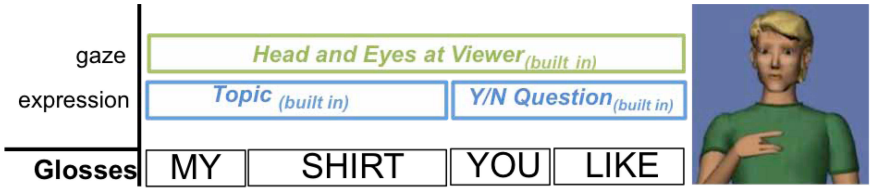

Fig. 1. This graphic depicts a timeline of an ASL sentence consisting of four signs (shown in the "Glosses" row) with co-occurring facial expressions from the software's built-in repertoire as specified by the user (shown in the "expression" row). The creator of this timeline has specified that a "Topic" facial expression should occur during the first two words and a "Yes No Question" facial expression during the final two.

3.2 Data-Driven Facial Expressions in Our New Avatar Platform

In order to conduct research on synthesizing animations of facial expressions for ASL, our laboratory required an animation platform that exposed greater control of the detailed aspects of the avatar's face movement. Further, we wanted an approach that would allow us to make use of face-movement data recorded from human ASL signers in our research. Our new animation platform is based on the open source tool EMBR [8], which has been previously used for creating sign language animations. Our lab extended its 3D avatar with ASL handshapes and detailed upper-face controls (eyes, eyebrows, and nose) that are compatible with the MPEG-4 Facial Animation standard [14]. As part of our enhancements to EMBR, a professional artist designed a lighting scheme and modified the surface mesh to support skin wrinkling, which is essential to perception of ASL facial movements [27].

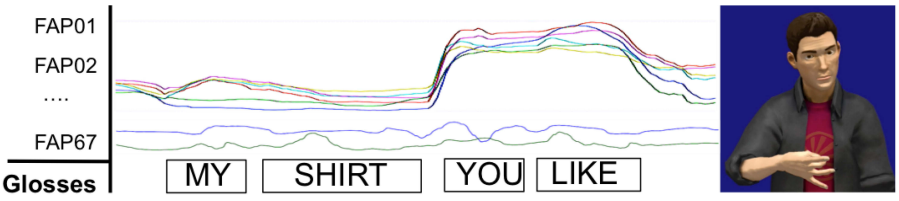

Fig. 2. A timeline is shown that specifies an ASL sentence with four words (shown in the "Glosses" row), with additional curves plotted above, each of which depicts the changing values of a single MPEG-4 parameter that governs the movements of the face/head. For instance, one parameter may govern the height of the inner portion of the signer's left eyebrow.

The MPEG-4 face parameterization scheme allows us to use recordings from multiple human signers, who may have different face proportions, to drive the facial expressions of the avatar. In particular, we implemented an intermediate component that converts MPEG-4 facial data extracted from facial movements in videos of human signers to the script language supported by the EMBR platform. To extract the facial features and head pose of the ASL human signers in the recordings, we use Visage Face Tracker, an automatic face tracking software [21] that provides MPEG-4 compatible output.

3.3 Comparison of New vs. Old Avatar Platform

To compare the naturalness and understandability of the ASL facial expressions synthesized by the two animation platforms, we analyzed data from prior studies [15][17] in which native signers evaluated animations from each. The multi-sentence stimuli shown and the questions asked were identical: Specifically, the hand movements for both avatars in those sentences are nearly identical (differences in avatar body proportion contributes to some hand movement differences). Thus, we present in section 4 data from "Old" vs. "New" animations. Further, for each platform, the stimuli were shown in two versions: with facial expressions ("Expr.") and without facial expressions ("Non"), where the hand movements for both versions were identical. Thus, there were a total of four varieties of animations shown to participants. Participants were asked to report whether they noticed a particular facial expressions being performed by the avatar and to answer comprehension questions about the stimuli.

Clearly, we are interested in comparing results across the two platforms ("old" v. "new"). Further, we are also interested in our ability to see a difference between the animations with facial expressions ("Expr.") and those without facial expressions ("Non") in each case. (If we can't see any difference in "notice" or "comprehension" scores when we animate the face of the character, this suggests that the platform is not producing clear sign language facial expressions.) We hypothesize:

- H1: When comparing our "new" and "old" animation platforms, we expect the "notice" scores will be statistically **equivalent** to the corresponding scores ("Expr." or "Non") between both platforms.
- H2: When comparing "Expr." animations with facial expressions and "Non" animations without facial expressions, we expect that our new platform will re-veal differences in "notice" scores **at least as well as** our earlier platform.

To explain our reasoning for H1 and H2: For the "Non" case, there is no reason to think that the change in virtual human platform should affect the scores since the face does not move during these animations. For the "Expr." case, while the new platform may have more detailed movements, there is no reason to think that people would notice face movements more in our new character, even if they were more detailed.

We also hypothesize the following, in regard to the "comprehension" scores:

- H3: When comparing our "old" and "new" animation platforms, comprehension scores assigned to "Non" animations without facial expressions will be statistically **equivalent** between both platforms.
- H4: When comparing "old" and "new" platforms, comprehension scores assigned to "Expr." animations with facial expressions in our new platform will be statistically **higher** than those for the old platform.
- H5: When comparing "Expr." animations with facial expressions and "Non" animations without facial expressions, we expect our new platform to reveal differences in comprehension scores **at least as well as** our old platform.

To explain our reasoning: When comparing the "Non" versions (no face movements), we expect the comprehension scores to be similar between both platforms because the hand movements are similar. However, we expect the animations with facial expressions created in the new platform to be more comprehensible, given that the new platform should be able to reproduce subtle movements from human signers.

4 Experiment Setup and Results

While different animation platforms were used to generate the animations shown to participants, the "script" of words in the stimuli was identical. We previously published a set of stimuli for use in evaluation studies of ASL animations [12], and we used stimuli codenamed N1, N2, N3, W2, W3, W5, Y3, Y4, and Y5 from the set in this study. These nine multi-sentence stimuli included three categories of ASL facial expressions: yes/no questions, negation, and WH-word questions.

A fully-factorial design was used such that: (1) no participant saw the same story twice, (2) order of presentation was randomized, and (3) each participant saw half of the animations in each version: i) without facial expressions ("Non") or ii) with facial expressions ("Expr"). All of the instructions and interactions for both studies were conducted in ASL by a deaf native signer, who is a professional interpreter. Part of the introduction, included in the beginning of the experiment, and the comprehension questions of both studies were presented by a video recording of the interpreter.

Animations generated using our old animation platform were shown to 16 native signers [17]. Of 16 participants, 10 learned ASL prior to age 5, and 6 attended residential schools using ASL since early childhood. The remaining 10 participants had been using ASL for over 9 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 11 men and 5 women of ages 20-41 (average age 31.2). Similarly, animations generated using our new animation platform were shown to 18 native signers [15], with the following characteristics: 15 participants learned ASL prior to age 9, The remaining 3 participants 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).

After viewing each animation stimulus one time, the participant answered a 1-to-10 Likert scale question as to whether they noticed a facial expression during the

animation; next, they answered four comprehension questions about the information content in the animation (using a 1-to-7 scale from "Definitely Yes" to "Definitely No").

Fig. 3 and 4 display the distribution of the Notice and Comprehension scores for the "Expr." and "Non" types of stimuli in the studies. (Box indicates quartiles, centerline indicates median, star indicates mean, whiskers indicate 1.5 inter-quartile ranges, crosses indicate outliers, and asterisks indicate statistical significance. To aid the comparison, mean values are added as labels at the top of each boxplot.) Labels with the subscript "(OLD)" indicate animations produced using our prior animation platform, and labels with the subscript "(NEW)" indicate animations produced using our new animation platform.

Since hypotheses H1 and H3 require us to determine if pairs of values are statistically equivalent, we performed "equivalence testing" using the two one-sided test (TOST) procedure [23], which consists of: (1) selecting an equivalence margin theta, (2) calculating appropriate confidence intervals from the observed data, and (3) determining whether the entire confidence interval falls within the interval (-theta, +theta). If it falls within this interval, then the two values are deemed equivalent. We selected equivalence margin intervals for the "notice" and comprehension scores based on their scale unit as the minimum meaningful difference. This results intervals of (-0.1, +0.1) for the 1-to-10 scale "notice" scores and (-0.14, +0.14) for the 1-to-7 scale comprehension scores. Having selected an alpha-value of 0.05, confidence intervals for TOST were evaluated using Mann-Whitney U-tests for Likert-scale data and t-tests for comprehension-question data. (Non-parametric tests were used for the Likert-scale data because it was not normally distributed.)

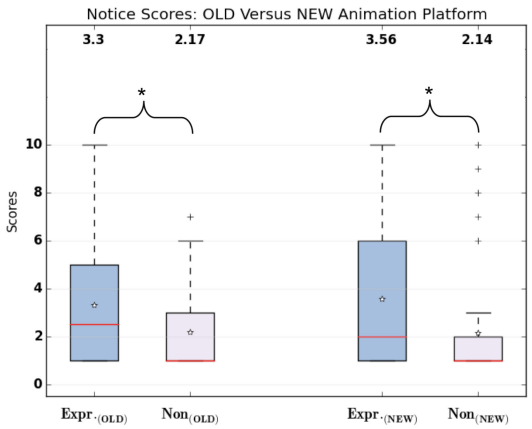

Fig. 3. Notice Scores for OLD and NEW Animation Platform.

Hypothesis H1 would predict that the "notice" scores for both "Non" and "Expr." stimuli would be unaffected by changing our animation platform. The following confidence intervals were calculated for TOST equivalence testing: (-0.00002, +0.00003) for $\text{Non}_{(\text{OLD})}$ vs. $\text{Non}_{(\text{NEW})}$ and (-0.000008, 0.00006) for $\text{Expr.}_{(\text{OLD})}$ vs. $\text{Expr.}_{(\text{NEW})}$. Giv-

en that these intervals are entirely within our equivalence margin interval of (-0.1, +0.1), we determine that the pairs are equivalent. Thus, hypothesis H1 is supported.

Hypothesis H2 would predict that evaluations conducted with our new animation platform are able to reveal with-vs.-without facial expressions differences in "notice" scores at least as well as our old animation platform. Thus, the statistical test is as follows: If there is a pairwise significant difference between $Expr._{(OLD)}$-$Non_{(OLD)}$, then there must be a statistically significant difference between $Expr._{(NEW)}$-$Non_{(NEW)}$. In support of H2, Fig. 3 illustrates significant difference between both pairs on the basis of Kruskal-Wallis and post hoc tests ($p<0.05$). We also observed that the magnitude of this difference is bigger in our new platform (d: 33, p-value: 0.001) than it is in our prior animation platform (d: 24, p-value: 0.02). Thus, hypothesis H2 is supported.

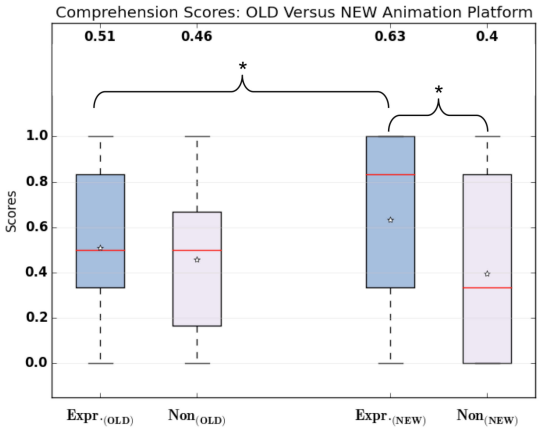

Fig. 4. Comprehension Scores for OLD and NEW Animation Platform.

Hypothesis H3 would predict that the comprehension scores for "Non" stimuli would be unaffected by changing our animation platform. The following confidence intervals were calculated for TOST equivalence testing: (+0.002, +0.119) for $Non_{(OLD)}$ vs. $Non_{(NEW)}$. Given that these intervals are within our equivalence margin interval of (-0.14, +0.14), we determine that the pairs are equivalent. Thus, H3 is supported.

Hypothesis H4 predicted that when considering the comprehension questions in evaluations conducted with our new animation platform the "Expr." stimuli would receive higher scores than the "Expr" scores for our old platform. As illustrated in Fig. 4, we observed a significant difference ($p<0.05$) between $Expr_{(OLD)}$-$Expr_{(NEW)}$ comprehension scores by performing one-way ANOVA. Thus, H4 is supported.

Hypothesis H5 predicted that evaluations conducted with our new animation platform would reveal with-vs.-without facial expressions differences in comprehension scores at least as well as our old animation platform. Fig. 4 illustrates a significant difference when comparing $Expr._{(NEW)}$-vs.-$Non_{(5EW)}$ comprehension scores for our new animation platform but not for the prior platform. Significance testing was based on one-way ANOVA and post hoc tests ($p<0.05$). Thus, H5 is supported.

5 Discussion and Future Work

This paper has demonstrated a methodology by which sign language animation researchers can directly compare the understandability and expressiveness of alternative animation platforms or avatars, through the use of experimental studies with deaf participants evaluating stimuli of identical ASL messages and responding to identical sets of comprehension questions. Such a comparison is valuable for ensuring that a new animation platform is able to produce human animations that are sufficiently expressive, and it also allows readers to understand how the results and benchmark baselines published in prior work would compare to results that are published using the new platform. In this case, we found that our new platform was able to produce animations that achieve similar scores to our old platform (when no facial expressions are included) or higher scores (when facial expressions are included). We also found that our new platform was able to produce animations with facial expressions that achieved significantly higher scores than animations without facial expressions.

Now that we have determined that this new animation platform is suitable for our research, in future work, we will investigate models for automatically synthesizing facial expressions for ASL animations, to convey essential grammatical information.

Acknowledgments. This material is based upon work supported by the National Science Foundation under award numbers 1506786 and 1065009. We acknowledge support from Visage Technologies AB. We are grateful for assistance from Andy Cocksey, Alexis Heloir, Jonathan Lamberton, Miriam Morrow, and student researchers, including Dhananjai Hariharan, Kasmira Patel, Christine Singh, Evans Seraphin, Kaushik Pillapakkam, Jennifer Marfino, Fang Yang, and Priscilla Diaz.

References

1. Adamo-Villani, N., Popescu, V., and Lestina, J.: A non-expert-user interface for posing signing avatars. *Disability and Rehabilitation: Assistive Technology*, 8(3), 238-248 (2013)
2. Adamo-Villani, N., and Wilbur, R.: Software for math and science education for the deaf. *Disability and Rehabilitation: Assistive Technology*, 5(2), pp. 115-124 (2010)
3. Ebling, S., and Glauert, J.: Exploiting the full potential of JASigning to build an avatar signing train announcements. In *3rd Int'l Symp on Sign Language Translation and Avatar Technology* (2013)
4. Elliott, R., Glauert, J., Kennaway, J., Marshall, I., and Safar, E.: Linguistic modeling and language-processing technologies for avatar-based sign language presentation. *Univ Access Inf Soc* 6(4), 375-391 Berlin: Springer (2008)
5. Filhol, M., Hadjadj, M.N, and Choisier, A.: Non-manual features: the right to indifference. In *6th Workshop on the Representation and Processing of Sign Language (LREC)* (2014)
6. Fotinea, S.E., Efthimiou, E., and Dimou, A. L. Sign language computer-aided education: Exploiting GSL resources and technologies for web deaf communication, pp. 237-244, Berlin: Springer (2012)
7. Gibet, S., Courty, N., Duarte, K., and Naour, T.L.: The SignCom system for data-driven animation of interactive virtual signers: Methodology and Evaluation. *ACM Transactions on Interactive Intelligent Systems*, 1(1), pp. 6 (2011)

8. Heloir. A, Nguyen, Q., and Kipp, M.: Signing Avatars: a Feasibility Study. In *2nd Int'l Workshop on Sign Language Translation and Avatar Technology* (2011)
9. Heloir, A. and Nunnari, F.: Towards an intuitive sign language animation authoring environment for the Deaf. In *Proc. of the 2nd Workshop in Sign Language Translation and Avatar Technology* (2013)
10. Holt, J.A.: Stanford Achievement Test - 8th Edition: Reading comprehension subgroup results. *American Annals of the Deaf* 138, 172–175 (1993)
11. Huenerfauth, M.: Spatial and Planning Models of ASL Classifier Predicates for Machine Translation. In *the 10th Int'l Conf on Theoret and Methodol Issues in Mach Transl* (2004)
12. Huenerfauth, M., and Kacorri, H.: Release of experimental stimuli and questions for evaluating facial expressions in animations of American Sign Language. In *Proc. of the 6th Workshop on the Representation and Processing of Sign Languages (LREC)* (2014)
13. Huenerfauth, M., Lu, P., and Rosenberg, A.: Evaluating importance of facial expression in American Sign Language and Pidgin Signed English animations. In the *Proc. of the 13th Int'l ACM SIGACCESS Conf on Computers and Accessibility*, pp. 99-106 (2011).
14. ISO/IECIS 14496-2 Visual (1999)
15. Kacorri, H., and Huenerfauth, M.: Implementation and evaluation of animation controls sufficient for conveying ASL facial expressions. In *Proc. of the 16th Int'l ACM SIGACCESS Conf on Computers and Accessibility*, pp. 261-262 (2014).
16. Kacorri, H.: TR-2015001: A Survey and Critique of Facial Expression Synthesis in Sign Language Animation. *Computer Science Technical Reports*. Paper 403 (2015)
17. Kacorri, H., Lu, P., and Huenerfauth, M.: Effect of Displaying Human Videos During an Evaluation Study of American Sign Language Animation. *ACM Transactions on Accessible Computing*, 5(2), pp. 4 (2013)
18. Kipp, M., Nguyen, Q., Heloir, A., and Matthes, S.: Assessing the deaf user perspective on sign language avatars. In the *Proc. of the 13th Int'l ACM SIGACCESS Conf on Computers and Accessibility*, pp. 107-114. New York: ACM Press (2011)
19. Mitchell, R., Young, T., Bachleda, B., and Karchmer, M.: How many people use ASL in the United States? Why estimates need updating. *Sign Lang Studies*, 6(3): 306-335, (2006)
20. Neidle, C., Kegl, D., MacLaughlin, D., Bahan, B., and Lee, R.G.: The syntax of ASL: functional categories and hierarchical structure. Cambridge: MIT Press (2000)
21. Pejsa, T., and Pandzic, I.S.: Architecture of an animation system for human characters. In *Proc. of the 10th Int'l Conf on Telecommunications*, pp. 171-176 (2009)
22. Schmidt, C., Koller, O., Ney, H., Hoyoux, T., and Piater, J.: Enhancing gloss-based corpora with facial features using active appearance models. In *Proc. of the 2nd Workshop in Sign Language Translation and Avatar Technology* (2013)
23. Schuirmann, D.J.: A comparison of the two one-sided tests procedure and the power approach for assessing equivalence of average bioavailability. *J Pharmacokin Biopharm*, pp. 15:657–680 (1987)
24. Smith, R. and Nolan, B.: Manual evaluation of synthesised sign language avatars. In Proc. of the 15th Int'l ACM SIGACCESS Conf on Computers and Accessibility, pp. 57 (2013)
25. Traxler, C.: The Stanford achievement test, 9th edition: national norming and performance standards for deaf & hard-of-hearing students. *J Deaf Stud & Deaf Educ*, 5:4, pp. 337-348 (2000)
26. VCOM3D.: Homepage. http://www.vcom3d.com/ (2015)
27. Wolfe, R., Cook, P., McDonald, J.C., and Schnepp, J.: Linguistics as structure in computer animation: Toward a more effective synthesis of brow motion in American Sign Language. *Sign Language & Linguistics*, 14(1), 179-199 (2011)
28. Zeshan, U.: Interrogative and negative constructions in sign languages, (2006)