

# Evaluating Importance of Facial Expression in American Sign Language and Pidgin Signed English Animations

Matt Huenerfauth

The City University of New York  
Queens College and Graduate Center  
Computer Science and Linguistics  
365 Fifth Ave, NY, NY 10016  
+1-718-997-3264

matt@cs.qc.cuny.edu

Pengfei Lu

The City University of New York  
CUNY Graduate Center  
Computer Science  
365 Fifth Ave, NY, NY 10016  
+1-212-817-8190

pengfei.lu@qc.cuny.edu

Andrew Rosenberg

The City University of New York  
Queens College and Graduate Center  
Computer Science  
365 Fifth Ave, NY, NY 10016  
+1-718-997-3562

andrew@cs.qc.cuny.edu

## ABSTRACT

Animations of American Sign Language (ASL) and Pidgin Signed English (PSE) have accessibility benefits for many signers with lower levels of written language literacy. In prior experimental studies we conducted evaluating animations of ASL, native signers gave informal feedback in which they critiqued the insufficient and inaccurate facial expressions of the virtual human character. While face movements are important for conveying grammatical and prosodic information in human ASL signing, no empirical evaluation of their impact on the understandability and perceived quality of ASL animations had previously been conducted. To quantify the suggestions of deaf participants in our prior studies, we experimentally evaluated ASL and PSE animations with and without various types of facial expressions, and we found that their inclusion does lead to measurable benefits for the understandability and perceived quality of the animations. This finding provides motivation for our future work on facial expressions in ASL and PSE animations, and it lays a novel methodological groundwork for evaluating the quality of facial expressions for conveying prosodic or grammatical information.

## Categories and Subject Descriptors

I.2.7 [Artificial Intelligence]: Natural Language Processing – *language generation, machine translation*; K.4.2 [Computers and Society]: Social Issues – *assistive technologies for persons with disabilities*.

## General Terms

Design, Experimentation, Human Factors, Measurement.

## Keywords

Accessibility Technology for People who are Deaf, American Sign Language, Pidgin Signed English, Facial Expression.

## 1. INTRODUCTION

Due to various factors, a majority of deaf high school graduates in the U.S. have a fourth-grade (age 10) English reading level or

below [30]. This means that many deaf adults have difficulty reading English text on websites, captioning, or other media. Over 500,000 people in the U.S. use American Sign Language (ASL), a language with a distinct word order, linguistic structure, and vocabulary than English [23]. Other deaf people in the U.S. use Pidgin Signed English (PSE), a signing system that more closely follows English word order (and can be performed while speaking or mouthing English words simultaneous to hand movements) [21]. Technology for presenting information in the form of computer animations of ASL or PSE can make information and services accessible to deaf people with lower English literacy, as explained in [15]. Animated characters have advantages over video for content that is often modified, is generated or translated automatically, or if the author's anonymity should be preserved. Unfortunately, modern ASL and PSE animation systems require a human to specify when linguistically essential facial expressions should occur and to carefully specify the facial movements to be performed. This is a difficult and time-consuming process.

We have previously conducted studies in which signers evaluate the understandability and naturalness of ASL animations [16, 17]. During these studies focusing on various aspects of ASL (e.g., speed/pauses, use of space around the body, verb movements), there was a trend: in open-ended feedback, participants rarely mentioned the aspect we were studying at the time. The most frequent comments were about the character's facial expressions. Comments included: "face was bland," "too stiff - she needs more facial expression," "eyebrows don't raise enough," "lack of facial expressions affect[ed] some comprehension," "need more head turn and up/down," etc. One participant noted the character attempted a "relationship between the facial movements and the content being said," but felt the timing was off, saying: "Close but no cigar." Another found the overly deadpan facial expression so disturbing that he said: "I'm gonna have nightmares because of that drugged-looking signer." Facial expression seemed to be a key element of ASL animation that signers still feel is incorrect in state-of-the-art systems. Signers were attuned to this aspect of ASL, and it dominated their impression of current technologies. Before investing significant resources in research on facial expression in ASL animation, we wanted to empirically verify and quantify our earlier research participants' comments. So, we conducted studies in which ASL and PSE signers evaluated animations of ASL and PSE with and without various types of facial expressions. Sections 2 and 3 give background on the linguistics of ASL/PSE and the use of facial expression in each. Section 4 surveys current signing animation systems and facial expression. Sections 5, 6, and 7 describe our experimental studies and results, and Section 8 includes conclusions and future work.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ASSETS'11, October 24–26, 2011, Dundee, Scotland, UK.

Copyright 2011 ACM #-####-####-####/\$10.00.

## 2. ASL: USE OF FACIAL EXPRESSION

Mouth shape, eye-brow height, and other face/head movements are a required part of ASL, and identical hand movements may have different meanings depending on the face/head [24]. Facial expressions change the meaning of adjectives (e.g., color *intensity* or distance *magnitude*) or convey adverbial information (e.g., *carelessly* or *with relaxed enjoyment*). The head/face indicates important grammar information about phrases; Fig. 1 shows: YN-question expression occurs during yes/no questions, WH-question expression occurs during interrogative questions (“what, where, how” etc.), negative-expression (and head shaking) negates the meaning of a sentence, and topic-expression indicates that an entity is an important topic for further discussion. A sequence of signs may have different meanings, depending on the head/face; e.g., the ASL sentence “JOHN LOVE MARY” without facial expression means: “John loves Mary.” With a yes/no facial expression, it indicates “Does John love Mary?” With a negative expression and headshake added during “LOVE MARY,” then the same sequence of signs indicates “John doesn’t love Mary.” (Facial expressions are timed to co-occur with hand movements for signs during specific parts of a sentence.) Further, ASL signers also use facial expressions to convey emotional subtext. Thus, facial expressions are essential to the meaning of ASL sentences.



Fig. 1. ASL facial expressions: yes/no-question, wh-question, negative face (with left-right headshaking), topicalization.

## 3. PSE: USE OF FACIAL EXPRESSION

Some people who are deaf or hard-of-hearing use forms of signing communication that more closely resemble English; these include: Manually Coded English (MCE) or Pidgin Signed English (PSE).

- MCE includes a variety of communication systems that follow English word order. Unlike ASL, MCE systems are not natural languages but are artificially invented, formally specified systems for conveying English using the hands. Most MCE systems adopt ASL signs to convey content words; some systems use additional signs to indicate English word endings (“-ed,” “-ing”) or function words (“the,” “of”) [21]. MCE can be slower to perform than spoken English or ASL but is used in educational settings to convey English grammar to students.
- Pidgin Signed English (PSE) is less formally specified than the MCE systems. PSE is a hybrid between English and ASL in which signers perform ASL signs for the main content words of an English sentence (in English word order), but they generally do not include extra signs to indicate English word endings or function words [4, 21]. Individual signers vary in how English-like or ASL-like their PSE is, and many ASL signers in the U.S.

can switch into PSE when communicating with someone that has weaker ASL-skills but strong English-skills.

In our prior research [16], we focused on ASL animations. Because PSE uses English word order, knowledge of English is needed to understand PSE; thus, many deaf people with lower English literacy prefer translation to ASL when possible. In this paper, we are studying both PSE and ASL animations – for the following reasons: (1) From a technical perspective, creation of PSE animations from English text is easier than for ASL, because English and PSE are more similar. (2) For low-literacy deaf users, PSE animations may still be somewhat easier to understand than English text, if vocabulary is better understood in the form of signing. (3) Because PSE follows English word order, a character can “mouth” or speak an English sentence simultaneous to PSE performance. Deaf users developing speech-reading (lip-reading) skills may benefit from a system for automatically producing an animated character with accurate lip-movements performing PSE while mouthing/speaking an English sentence provided as input.

While hand movements of PSE convey the content words of an English sentence, this is only part of the message. Spoken English uses prosody (pitch, timing, and volume variations) to convey information in parallel to the words that are spoken. Speech can be divided into *what* is being said (lexical content) and *how* it is said (prosodic content). Prosody can affect the interpretation of a set of words; e.g., rising pitch at the end of a phrase yields a yes/no question (“John and Mary are friends?”) and falling intonation yields a statement (“John and Mary are friends.”). Prosody is used to indicate topicality and focus, where topic words and foci are often produced with intonational prominence [14]. When new concepts are introduced, they are typically made prominent, and concepts introduced previously are typically de-accented, or made less prominent [5]. Prosody can convey the speaker’s state, e.g., anger, frustration [19], or incredulity [32]. Prosody can convey structural information to a listener. Phrases that occur at the start of a segment of discourse are produced with higher pitch, volume, and speed than those at the end of one [18].

PSE can include facial expressions to convey some *prosodic* information. In fact, English speakers (not just PSE signers) tend to perform facial expressions correlated with their vocal prosody:

- The face conveys emotions like anger, frustration, sadness [6] – each of which have characteristic vocal prosody [19].
- Changes in eye-brow height correspond to vocal prominence (emphasizing words or phrases) [7] and changes in pitch [9].
- Head nods can also correspond to vocal prominence [20].
- Use or avoidance of eye-contact during a pause in speaking can differentiate an end of a speaking turn from a temporary pause to think [25] – these correspond to different vocal intonations.
- The face can indicate disbelief/incredulity (e.g., rolling your eyes), and prosody can indicate a speaker’s uncertainty [32].
- Visible facial movements can indicate when speakers are asking questions [29] – which also have characteristic prosody.

Speakers don’t perform these facial expressions as consistently as ASL signers perform grammatical facial expressions (governed by rules of ASL), but these visible facial movements can suggest prosody. Thus, it could be communicatively useful for the virtual character in a PSE animation system to perform these facial expressions to suggest prosodic information. Due to the influence of ASL on PSE, some systematic ASL facial expressions (e.g., for wh-questions) can also appear during PSE signing.

## 4. RELATED WORK

Researchers have studied the selection and synthesis of facial expressions for speaking animated characters, e.g.: in “talking heads” with accurate lip movements [22, 29], rule-based selection of facial expressions for English sentences [3], planning facial expressions to give feedback in conversation [13], and planning of eye-movements and facial expressions with discourse impacts [26]. However, this research has not addressed facial expression during animations of PSE or ASL signing.

In an earlier survey [15], we described current systems for producing animations of sign language or other signing; these include: scripting or generation (e.g., [8, 10, 11]). *Scripting* software allows a human to “word process” an ASL/PSE sentence by arranging signs from a dictionary onto a timeline; the software produces an animation of a virtual character based on this timeline (so the human user does not need to manually control all of the joints of the character’s body). *Generation* software plans an ASL/PSE sentence based on an English input sentence or other information, without a human manually selecting signs. Despite the importance of facial expressions, current ASL/PSE systems include only a small repertoire of hard-coded facial movements. A generation system automatically planning an ASL/PSE sentence must decide which facial expressions are needed and when they begin/end. Due to the subtle prosodic factors that affect the facial expression, making these decisions automatically is beyond the state of the art, except for simple cases (e.g., add a wh-question expression during a sentence if it starts with “who,” “what,” etc.).

In a scripting system, the user is typically able to select one facial expression from a list and specify that it occur during a portion of the sentence. For example, in VCom3D Sign Smith Studio [31], a commercially available scripting system for ASL/PSE, the user assembles a sentence on a timeline: with signs on one track and facial expressions on a parallel track. The system’s repertoire of linguistic facial expressions is finite: 11 grammatical expressions (and 10 that function as degree adverbials). Like other current ASL/PSE systems, it cannot overlay one facial expression onto another simultaneously, and when two are performed sequentially, the character interpolates the facial pose from one to the next.

Human signers often perform facial expressions simultaneously (e.g., negative headshake + wh-question, sadness + topicalization, etc.). When humans transition from one grammatical facial expression to the next, complex rules govern how these transitions occur, not simple interpolation. Further, the intensity of facial expressions is often based on the timing of the signs performed by the hands: e.g. intensity of a negative headshake may be strongest during the sign “NOT” in a sentence [24]. Handling these complex aspects of facial expression selection, timing, simultaneous performance, and transitions is beyond the state of the art of current ASL/PSE systems. Animating facial expressions accurately is too difficult for generation systems to handle automatically and time-consuming for users of scripting systems. Thus, we want to explore how to better automate facial expression selection and synthesis in our future ASL/PSE animation research.

## 5. DESIGN OF EVALUATION STUDIES

Before beginning this new line of research, we wanted to know whether improvements in the quality of the facial expression of characters in ASL/PSE animation systems would actually lead to benefits for deaf users. Specifically, would the users’ perception of the naturalness or ease-of-understanding of the animations improve if they included more accurate facial expressions?

Section 1 described how we had some informal evidence from study participants’ comments, but prior research (section 4) has not included a user-based evaluation of the comprehension impact of facial expression in signing animations. The remainder of this paper describes our efforts to obtain quantitative evidence of the importance of facial expressions for ASL and PSE animations.

While some of the information conveyed by facial expressions is categorical (i.e., whether or not the sentence is negated, whether or not the sentence should be interpreted as a question), other information conveyed by facial expressions is more subtle and a matter of interpretation/degree (i.e., what emotional subtext is being conveyed, how much emphasis is the signer placing on a particular word, is the signer conveying a sense of incredulity or doubt). Especially for these non-categorical (matters of degree) cases, there are natural variations in the way in which some facial expressions are performed and are interpreted by viewers. This presents a challenge when designing evaluation studies designed to measure whether the facial expressions in an ASL or PSE animation are accurate – because we can expect some natural variation in responses to questions about these animations.

Further, sometimes the way in which the prosodic information affects the meaning of a sentence is quite subtle. For instance, a sentence like “I didn’t order a pizza” with some vocal prominence on the word “I” could indicate that the speaker believes someone else ordered the pizza. With prominence on the word “pizza,” it could indicate that the speaker placed an order, but for something else. In either case, the basic truth value of the statement is unaffected: the speaker did not order a pizza. What is affected by the prosodic variation is the *implication* that can be inferred. Evaluating subtle implications by users in a study can be difficult.

Fortunately, research on human speech suggests successful methods for measuring the impact of prosodic information on how messages are understood and interpreted by human listeners, e.g., [2, 27]. These researchers designed sets of sentences or short stories that – in the absence of prosodic information – contain some degree of ambiguity in how they can be interpreted (similar to the “I didn’t order a pizza” example above). When prosodic information is added to the sentences, then it is clear that one interpretation is more correct. Participants in the study who listen to audio performances of these sentences are then asked to answer multiple-choice or Likert-scale questions about the meaning of the sentences. These questions are carefully engineered such that someone would answer the question differently – based on which of the alternative possible interpretations of the spoken sentence they had mentally constructed. For example, someone who heard the sentence “I didn’t order a pizza” (with prominence on “I”) may be more likely to respond affirmatively to a question asking: “Does the speaker think that someone else ordered a pizza?” In designing the studies in sections 6 and 7, we have used similar experimental design, stimuli, and comprehension questions.

## 6. EVALUATION OF PSE ANIMATIONS

To evaluate whether facial expressions added to a PSE animation could enable viewers of the animation to identify the prosodic content of the English sentences being performed, we designed a study with 28 sets of 1-2 sentence English passages – inspired by those used in previous speech prosody evaluation projects [2, 27]. Without the prosodic aspect of their spoken performance, these passages contained some degree of ambiguity in how they could be interpreted. Fig. 2 contains examples of some of the passages used in the study as English and PSE transcriptions.

**Original Spoken English Sentence (transcript of audio):**

My brother said he ordered a pizza, but the pizza never arrived.  
(Incredulous emphasis tone during the word “said.”)

**Pidgin Signed English (sequence of signs performed):**

MY BROTHER SAY HE ORDER PIZZA HOWEVER PIZZA  
NEVER ARRIVE

(Incredulous emphasis facial expression during the word “SAY.”)

**Four Comprehension Questions (correct answer in parentheses):**

Does Sue know why the pizza has not arrived? (yes)

Is Sue upset at the pizza restaurant? (no)

Does Sue believe that her brother ordered a pizza? (no)

Did the pizza arrive? (no)

**Original Spoken English Sentence (transcript of audio):**

What did you just tell Charlie about me?

(Angry/accusatory tone during the whole sentence.)

**Pidgin Signed English (sequence of signs performed):**

WHAT DO YOU RECENT TELL #CHARLIE ABOUT ME

(Angry/accusatory facial expression during the whole sentence.)

**Four Comprehension Questions (correct answer in parentheses):**

Is Sue accusing you of something? (yes)

Is Sue upset about something? (yes)

Does Sue care about Charlie's opinion about her? (yes)

Did you just tell Charlie something? (yes)

**Fig. 2. Examples of stimuli and comprehension questions.**  
(The “#” symbol indicates a word that was fingerspelled.)



**Fig. 3. Signing character and face images (listed top row first):**  
neutral face, continuing, contrastive emphasis, incredulous  
emphasis, anger, sadness, yes/no question, & wh-question.

The 28 passages used in the study can be divided into several categories, based on the type of prosodic information contained:

- **QUESTION:** Some passages contained a sentence that was a question, but in the absence of prosodic cues could instead be interpreted as a declarative statement or relative clause: “I went to that new restaurant you suggested. It’s Chinese?” (Without prosody, the second sentence could be a declarative statement.) “Last Friday, I saw Metallica. Which is your favorite band?”
- **EMPHASIS:** Some passages contained a single word or phrase that received stronger vocal prominence; this emphasis of the word indicated some form of contrast or incredulity: “Bill bought some shirts yesterday. The *green* ones were nice.” (This suggests the others were not.) “My brother *said* he ordered a pizza, but the pizza never arrived.” (This suggests disbelief.)

- **EMOTION:** Some passages were performed with a strong emotion that affected their meaning: “What did you just tell Charlie about me?” (With an angry tone, this suggests that the speaker is aware of what was said and disapproves.) “Yesterday my sister took me to a concert. It was country music.” (A sad tone during the second sentence suggests dislike of this music.)
- **CONTINUE:** Some passages ended with a slightly rising tone and lack of deceleration so as to convey that the speaker was not yet finished a thought but was only momentarily pausing: “Mary is busy: she plays sports, she goes to school...”

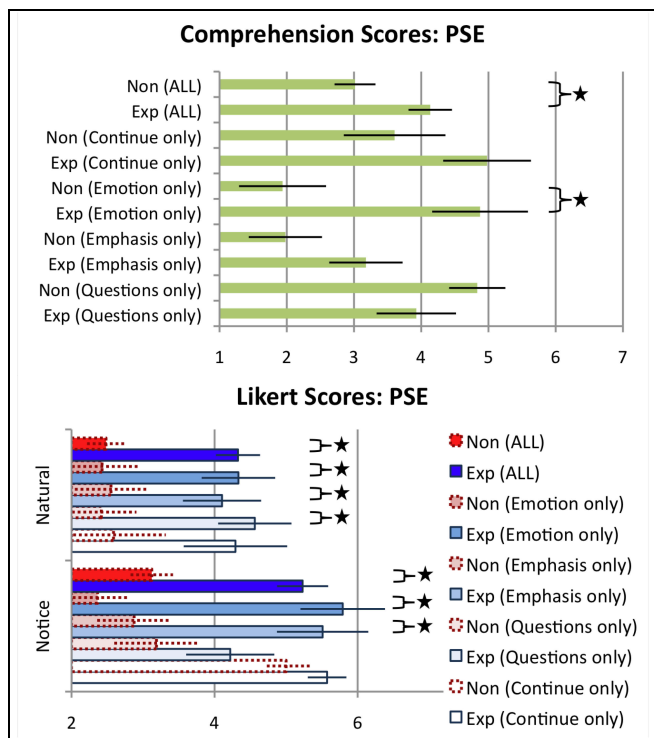
To produce the PSE animations, we began by collecting audio recordings of a native English speaker performing each passage. In that way, we could determine the timing of the hands and the facial expressions (corresponding to the vocal prosody) for our animation based on the timing of the audio. The spoken utterances were recorded using a Sennheiser Mk3 headset microphone in a quiet room. The speaker was a female, non-professional speaker reading from a script with prosodic performance notes. Productions that contained mispronunciations, disfluencies, or insufficiently clear prosodic content were re-recorded.

To determine the precise timing of each phoneme (each consonant/vowel sound), we implemented an automatic speech recognition system. Our 44.1kHz 16bit mono audio recordings were forced-aligned to transcripts of each passage using HTK [33] with tri-phone acoustic models trained on TIMIT [12] and the CMU pronunciation dictionary. The forced-alignment procedure generated phoneme and word start- and end-times with a precision of 10ms. This information can be used to determine the correct shape of the lips for each moment of time in the animations, based on the sound being pronounced. This information also gives us precise timing when each word occurs in the audio so that we can build an accurate timeline of the hand movements for each sign during the PSE animation of the English sentence.

We used VCom3D Vcommunicator [31] to create animations of a character with accurate lip movements speaking each passage; Vcommunicator allows for hand gestures to be added to an animation timeline. We inserted PSE signs for each of the words in the sentence to produce a PSE performance. Although lip movement was included, the animations contained no audio. Vcommunicator includes a finite repertoire of facial expressions that can be added to the timeline, and it allows for head and eye movement controls. Working within this repertoire (and based on observation of a native English speaker performing these passages), we selected a facial expression, head pose, and eye aim for each of the different types of prosodic performance in our 28 passages. Fig. 3 contains images of the facial poses selected. We weren’t entirely satisfied with the wh-question expression, which we felt looked a little angry, but it was the closest approximation we found, working within the face options available in the system.

In prior work, we discussed how to recruit and screen signers for experiments evaluating animations [16, 17]. Ads for the study were posted on Deaf community mailing lists and websites in New York, and participants were asked if they had grown up using PSE at home or attended a school using PSE as a child. The 6 men and 6 women recruited for the study were ages 22-39 (median age 31). Six had used PSE since birth, five began using PSE prior to age 6, and one had learned PSE as an adult (with over 20 years of daily use of PSE in a professional capacity). This included 2 hearing participants, each of which had over 20 years of PSE use; one since birth, growing up with deaf family members.





**Fig. 4. Results of the PSE study: “Non” animations lacked facial expressions; “Exp” animations included them.**

Each animation was produced in 2 versions, with and without facial expressions. Each participant in the study viewed 28 PSE stories in a fully-factorial within-subjects design such that: (1) no participant saw the same story twice, (2) the order of presentation was randomized, and (3) each participant saw 28 animations of each version (facial-expression vs. no-facial-expression). After viewing each story, subjects were asked four comprehension questions (see Fig. 2), answers were recorded on a 7-point Likert scale from “definitely no” to “definitely yes.” When tabulating results, the response scale for questions for which the correct answer was “no” was inverted. The top graph in Fig. 4 displays results; the thin line on each bar is the standard error of the mean.

Adding facial expressions led to an increase in comprehension scores. Significant pairwise differences are marked with stars in Fig. 4 (Mann-Whitney test,  $p < 0.05$ ). Overall, comprehension scores are somewhat low; this may be because answering these questions relied on participants drawing subtle inferences based on the prosody. In each set of four comprehension questions for a passage, we only included one question that asked a basic fact (usually the last question); the other three questions relied on the participant making subtle inferences based on the prosodic information – see sample questions in Fig. 2. The EMOTION and CONTINUE categories displayed larger benefits from adding facial expression than the EMPHASIS or QUESTION categories. In fact, adding facial expressions to the QUESTION category *reduced* comprehension scores; this may be due to the “angry” looking facial expression we used for wh-question in this study.

In addition to comprehension questions, we asked participants to respond to a 1-to-10 Likert scale question about how natural the animation appeared. See the “Natural” bars in the lower graph in Fig. 4. There was a significant increase in the scores for the “Exp”

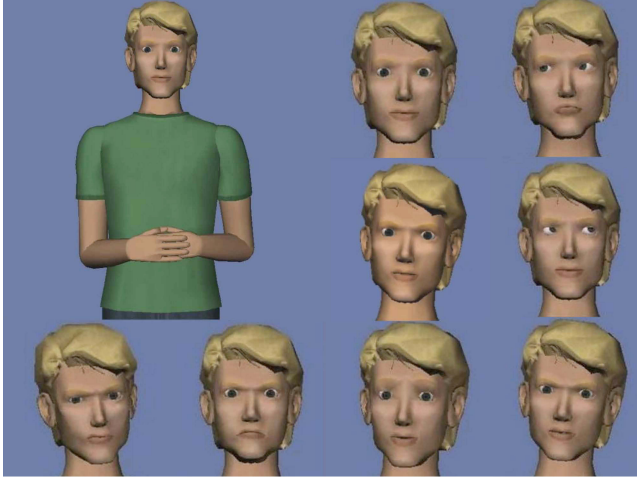
(with facial expression) animations when considering all data and considering data within EMOTION, EMPHASIS, or QUESTION categories only: significant pairwise differences marked with stars in Fig. 4 (Mann-Whitney test,  $p < 0.05$ ). This result suggests that adding facial expressions to the animations led participants to report that the animation was more natural in appearance: this quantitative result confirms the qualitative feedback comments of participants in earlier evaluation studies (discussed in section 1).

For each passage, we also asked participants to answer a question in which they reported on a 1-to-10 Likert scale if they believed that signer was: conveying an emotion, asking a question, emphasizing a word, or still speaking at the end of the animation. An appropriate question was included for each passage based on which category it was. See the “Notice” bars in the lower graph in Fig. 4. The “Exp” animations received higher scores for this question; this result suggests that participants overtly noticed the specific prosodic effect that the facial expression was meant to convey in each of these animations. Of course, if the facial expressions had been completely successful at conveying the intended prosodic meanings, we would have expected scores of “10” for this question. The average response score for the “Exp” animations was 5.2; so, there is still room for improvement in the selection and synthesis of facial expression animations for PSE.

Because it was informal feedback comments collected at the end of a study that prompted us to investigate the issue of facial expression, we continued to collect such comments in this study. Six participants mentioned that they felt that the character could benefit from more facial expressions (but it is unclear whether their comments may be in reaction to half of the animations they viewed during the study lacking any facial expression). Five participants mentioned that the speed of the animations was too fast, especially for words that were fingerspelled. We had set our PSE timing based on the audio of the human, who spoke at a slow but natural pace. In future work, we may study how to systematically slow down the timing from an English audio recording when producing a PSE animation so that there is more time for performing complex signs or fingerspelled words. Two participants noticed that the wh-question facial expression used in this study looked incorrect or somewhat angry. For this study, we were using a pre-existing human animation system [31], and we had merely selected the closest facial expression in the repertoire that matched what we wanted. However, in future facial expression research, we may rely on detailed comments and suggestions from participants about ways to modify subtle aspects of a face movement to achieve a clearer facial expression.

## 7. EVALUATION OF ASL ANIMATIONS

In order to understand whether the addition of facial expressions to ASL animations would also affect deaf users’ perceptions of the naturalness and understandability of the animations, we conducted a second study. In order to make the results of this study more directly comparable to that of the prior PSE study, we used the same set of 28 passages and questions. Each of the passages was translated into ASL by a native ASL signer (who works as a professional interpreter). Next, we used VCom3D Sign Smith Studio [31], a commercially available tool for scripting ASL animations, to produce an animation for each of the 28 passages. The new animations used proper ASL word-order and required grammatical facial expressions. Next, the native signer selected facial expressions to add to portions of each animation to convey the prosodic information that had originally been conveyed by the vocal prosody in the original English passages.



**Fig. 5. ASL character and face images (listed top row first): neutral face, continuing, contrastive emphasis, incredulous emphasis, anger, sadness, yes/no question, & wh-question.**

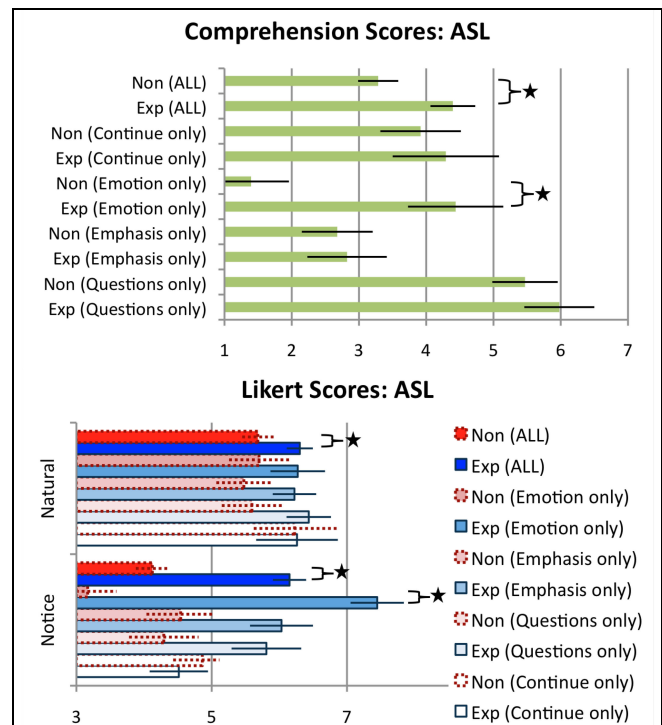
See Fig. 5 for illustrations of the facial expressions used in this study. Note that the character used in the study was different because we were using a different animation tool to produce the animations: one that was specially designed for producing sign language animations, not speaking characters. The character for the ASL animations did not move its lips to speak English words. Also, note that the specific facial expressions used in the animations and their exact timing during the ASL-translated version of each sentence differed slightly from the PSE/English examples in the prior study. For this study, a native signer was asked to produce facial expressions that linguistically or naturally conveyed the prosodic information for ASL sentences. Each animation was produced in two versions: with facial expressions conveying prosodic information and one without. However, even “without” versions may have included some facial expression: any linguistically required facial expressions not directly related to the specific prosodic information that was the focus of that passage still remained in the “without facial expression” animations. For example, the ASL passage “YESTERDAY #BILL BUY SHIRTS. GREEN SHIRTS NICE” (Yesterday, Bill bought some shirts; the green shirts were nice), included a special facial expression during the sign YESTERDAY (used during conditional/when phrases) in both versions of the animation. In the “with facial expression” version of the animation, a special contrastive-emphasis facial expression was added during the word “GREEN.”

To evaluate whether the addition of facial expressions to these ASL animations would affect native ASL signers’ judgments of their naturalness or ability to understand their content, we ran a study similar in design to the PSE study in section 6. We recruited 12 native ASL signers who evaluated ASL computer animations of two types: with facial expressions indicating prosodic information and without such facial expressions. A fully-factorial within-subjects design was used such that: (1) no participant saw the same story twice, (2) the order of presentation was randomized, and (3) each participant saw 28 animations of each version (facial-expression vs. no-facial-expression). In prior studies, we created a set of best-practices to ensure that responses given by participants are as ASL-accurate as possible [16]. The participants should be native ASL signers, and it is important to ask questions to screen for such participants. Further, the study environment should be

ASL-focused with little English influence. All of the instructions and interactions for this study were conducted in ASL by a native signer (a professional interpreter). Advertisements posted on Deaf community websites in New York asked potential participants if they had grown up using ASL at home or attended an ASL-based school as a child. Of the 7 men and 5 women in the study, 8 used ASL since birth, 3 began using ASL prior to age 10 when they began attending a school with instruction in ASL, and 1 learned ASL at age 18. This final participant has used ASL for over 22 years, attended a university with instruction in ASL, and uses ASL daily to communicate with a spouse. Participants were ages 21-46 (median age 32).

In Fig. 6, thin lines indicate standard error of the mean; significant pairwise differences are marked with stars (Mann-Whitney test,  $p < 0.05$ ). The addition of prosodically motivated facial expressions led to an increase in the comprehension scores; the result was significant when considering all data and when considering only the passages in the EMOTION sub-category (Likert 1-to-7 scale data). We see similar results for the questions in which signers were asked to rate the naturalness of the animations or respond to a question in which they indicated if they overtly noticed that the sentence: conveyed an emotion, emphasized a word, asked a question, or indicated that the signer had more to say (see the “Natural” or “Notice” bars in Fig. 6, Likert 1-to-10 scale data).

For ASL (Fig. 6) and for PSE (Fig. 4), it is notable how much of a comprehension benefit was provided by the facial expressions conveying EMOTION, as compared to the results for the other three sub-categories: CONTINUE, QUESTION, and EMPHASIS. From this result, it is clear that future research on facial expression in ASL/PSE animations should continue to explore how to overlay emotional information onto sentences.



**Fig. 6. Results of the ASL study: “Non” animations lacked facial expressions; “Exp” animations included them.**

Comparing the results for PSE (Fig. 4) to ASL (Fig. 6), the PSE animations seemed to benefit more from the addition of facial expression in the CONTINUE and EMPHASIS categories. In future work, we want to study further whether human ASL signers do use facial expression to convey this type of information in ASL (and which facial expressions they really use). It is possible that ASL signers instead use other mechanisms to emphasize words or to indicate that they are not yet at the end of a conversational turn, e.g., changes in timing, speed, size of movements, selection of different word-order or phrasing. It is interesting is that signers did appear to *notice* that some emphasis had occurred when facial expressions were added; note the larger (though not significantly so) “Notice” bar for “Exp (EMPHASIS only)” in Fig. 6. Yet, this didn’t lead to a change in how they answered the comprehension questions; this may indicate that while signers noticed that some emphasis had been added to a word/sign in the ASL sentence, it did not lead them to draw the same inferences about the meaning. This is an issue we plan to examine further in future work; it will be important for us to distinguish whether we merely had the wrong facial expression in our animations or whether this is not information that is primarily conveyed on the face in ASL.

## 8. CONCLUSIONS AND FUTURE WORK

The research described in this paper is an example of the benefits of including actual users with disabilities in experimental evaluations of accessibility technology – and encouraging them to offer open-ended feedback and suggestions. This line of research began when looking through quotes from participants in past studies in which they critiqued aspects of our ASL animation technology that we were not explicitly focusing on at the time. We noticed a trend in their comments and decided to conduct formal studies to examine whether the effect they were suggesting could be measured quantitatively. In this case, we did observe that adding facial expressions to animations of ASL and PSE led to significant improvements in users’ subjective evaluation of the naturalness of the animations and in scores on comprehension questions about the information conveyed. This is the first study on ASL or PSE animations to measure such an effect, and it provides motivation for future research on automating the insertion of accurate facial expressions to ASL/PSE animations.

Further, this paper has laid a methodological foundation for future research on facial expressions or conveying prosodic information in ASL and PSE animations. By adapting experimental techniques used by speech researchers studying prosody, we have designed sets of stimuli and questions for measuring the impact of facial expressions in ASL and PSE. We can use these methods as we examine detailed aspects of facial expression in future work; e.g., comparative studies of variations in face movements can determine precisely which facial expressions are most successful at conveying particular prosodic information in these animations. The methodologies we have used in this paper could be adapted for use by researchers studying other sign languages used internationally or studying other ways of conveying prosody in ASL/PSE animations (e.g., speed/timing variations).

The ultimate goal of our future research on facial expressions in ASL and PSE is to construct computational models of when to perform facial expressions during ASL/PSE animations and how to articulate the character’s face so as to convey these facial expressions most clearly (including under complex conditions when multiple facial expressions are performed sequentially or simultaneously). Determining when to insert facial expressions and how to set all of the parameters of the face correctly is

difficult for generation systems to do automatically and is very time-consuming for users of ASL/PSE scripting systems. We want to study how to automate the synthesis of these animations.

A challenge in synthesizing ASL/PSE facial expressions is that users seem sensitive to minor errors in face articulation, leading them to misidentify an expression. In the PSE study, 2 participants commented they were misled in their interpretation of some sentences because they felt the character had an angry face, not a wh-question face. While these facial expressions are similar, there was other evidence in the sentence that it’s a question (including “wh” words like “who”). Even still, users thought the character was angry, not questioning, because it didn’t have a perfect wh-question face. This is an interesting contrast to the robustness we observed in deaf users in prior studies who tolerated errors in the location or orientation of the character’s hands for certain signs; they were often able to use contextual information to guess the correct sign. If there is a difference in the degree to which users are able to tolerate errors in facial expression, then this would add a new layer of challenge to this line of research. There is linguistic precedent: humans listening to speech performed by non-native speakers are more sensitive to errors in the prosody/intonation than to errors in the pronunciation of individual consonant/vowel phonemes [1]. In future work, we will examine if there is analogous sensitivity to prosody errors in ASL/PSE animations.

Another new methodological foundation laid by this current study is that, as the first time our lab studied PSE animations, we had to develop speech recognition software for determining lip/hand timing for PSE animations and to recruit a new type of participant (with PSE experience). Because ASL animations have greater accessibility benefits for many deaf users with low English literacy, we will continue to study ASL animations at our lab. However, English-to-PSE systems may benefit users with less ASL skill or who prefer a character that speaks/mouths an English sentence while it signs. In fact, several participants in our PSE study wrote feedback comments that expressed excitement at having a PSE animation system with lip movements and signing, especially for English instruction applications in educational settings for deaf children. We anticipate interesting challenges in balancing the timing demands of English speech audio with the movements of the hands in PSE animations, and we are excited about examining these issues further in our future work.

Finally, our future work will include collaboration between animation researchers and speech-recognition researchers who study how to automatically detect prosody and intonation in speech. The AuToBI toolkit [28] provides a mechanism to generate hypothesized prosodic phrases and pitch accents; this “automatic prosodic analysis” technology could be used to *automatically* select and set the timing of facial expressions for PSE animation characters. Our future work will investigate robust techniques to convey this acoustic/prosodic information visually.

## 9. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 0746556. This work was supported by the Research Enhancement Committee at Queens College, PSC-CUNY Research Award Program, Siemens A&D UGS PLM Software (Go PLM Grant Program), and a free academic license for character animation software from Visage Technologies AB. Jonathan Lamberton recruited participants and collected response-data during the user-based evaluation study.

## 10. REFERENCES

- [1] Anderson-Hsieh, J., Johnson, R., Koehler, K. 1992. The relationship between native speaker judgments of non native pronunciation and deviance in segmentals, prosody and syllable structure. *Language Learning*, 42: 529-555.
- [2] Allbritton, D.W., Mckoon, G., Ratcliff, R. 1996. Reliability of prosodic cues for resolving syntactic ambiguity. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 22: 714-735.
- [3] Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., Stone, M. 1994. Animated conversation: Rule based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In *Computer Graphics Annual Conference Series (SIGGRAPH'94)*, 413-420.
- [4] Cokely, D.R. 1983. When is a Pidgin not a Pidgin? An alternate analysis of the ASL-English contact situation. *Sign Language Studies*, 12(38): 1-24.
- [5] Dahan, D., Tanenhaus, M., Chambers, C. 2002. Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language*, 47: 292-314.
- [6] Ekman, P. 1982. *Emotion in the human face*. Cambridge, England: Cambridge University Press.
- [7] Grandstrom, B., House, D., Lundeborg, M. 1999. Prosodic cues in multimodal speech perception. In *Proc. Int'l Congress of Phonetic Sciences (ICPhS 99)*, 655-658.
- [8] Elliott, R., Glauert, J., Kennaway, J., Marshall, I., Safar, E. 2008. Linguistic modeling and language-processing technologies for avatar-based sign language presentation. *Univ Access Inf Soc* 6(4), 375-391. Berlin: Springer.
- [9] Flecha-Garcia, M.L. 2009. Eyebrow raises in dialogue and their relation to discourse structure, utterance function and pitch accents in English. *Speech Communication*, 52:542-554.
- [10] Filhol, M., Delorme, M., Braffort, A. 2010. Combining constraint-based models for Sign Language synthesis. In *Proc. 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, Language Resources and Evaluation Conference (LREC), Valetta, Malta*.
- [11] Fotinea, S.E., E. Efthimiou, G. Caridakis, K. Karpouzis. 2008. A knowledge-based sign synthesis architecture. *Univ Access Inf Soc* 6(4):405-418. Berlin: Springer.
- [12] Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L., Zue, V. 1993. *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Philadelphia, PA: Linguistic Data Consortium.
- [13] Granström, B., House, D., Swerts, M. 2002. Multimodal feedback cues in human-machine interactions. In *Proc. of Speech Prosody (SP-2002)*, 347-350.
- [14] Hedberg, N., Sosa, J. 2007. The prosody of topic and focus in spontaneous English dialogue. In: *Topic and Focus: Cross-Linguistic Perspectives on Meaning and Intonation*. Berlin: Springer.
- [15] Huenerfauth, M., Hanson, V. 2009. Sign language in the interface: access for deaf signers. In C. Stephanidis (ed.), *Universal Access Handbook*. NJ: Erlbaum. 38.1-38.18.
- [16] Huenerfauth, M., L. Zhao, E. Gu, J. Allbeck. 2008. Evaluation of American sign language generation by native ASL signers. *ACM Trans Access Comput* 1(1):1-27.
- [17] Huenerfauth, M. 2009. A Linguistically Motivated Model for Speed and Pausing in Animations of American Sign Language. *ACM Trans. Access. Comput.* 2, 2, Article 9 (June 2009), 31 pages.
- [18] Hirschberg, J., Nakatani, C. 1996. A prosodic analysis of discourse segments in direction-giving monologues. In *Proceedings of the 34th conference on Association for Computational Linguistics*, 286-293.
- [19] Juslin, P.N., Laukka, P. 2003. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin* 5.
- [20] Krahmer, E., Swerts, M. 2007. The effect of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, 57(3): 396-414.
- [21] Lucas, C. 2001. *The Sociolinguistics of Sign Languages*. Washington, DC: Gallaudet University Press.
- [22] Massaro, D., Beskow, J. 2002. Multimodal speech perception: A paradigm for speech science. In B. Granstrom, D. House, & I. Karlsson (eds.), *Multimodality in language and speech systems*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 45-71.
- [23] Mitchell, R., Young, T., Bachleda, B., & Karchmer, M. 2006. How many people use ASL in the United States? Why estimates need updating. *Sign Lang Studies*, 6(3):306-335.
- [24] Neidle, C., D. Kegl, D. MacLaughlin, B. Bahan, R.G. Lee. 2000. *The syntax of ASL: functional categories and hierarchical structure*. Cambridge: MIT Press.
- [25] Novick, D., Hansen, B., & Ward, K. 1996. Coordinating turn-taking with gaze. In *Proceedings of ICSLP-96, Philadelphia, PA*, 3, 1888-91.
- [26] Pelachaud, C., Badler, N. I., Steedman, M. 1996. Generating Facial Expressions for Speech. *Cognitive Science*, 20:1-46.
- [27] Price, P., Ostendorf, M., Shattuck-Hufnagel, S., Fong, C. 1991. The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America*.
- [28] Rosenberg, A. 2010. AuToBI - A Tool for Automatic ToBI Annotation. In *Proc. 11th Annual Conference of the International Speech Communication Association INTERSPEECH 2010*.
- [29] Srinivasan, R., Massaro, D. 2003. Perceiving prosody from the face and voice: distinguishing statements from echoic questions in English. *Language and Speech*, 46(1): 1-22.
- [30] Traxler, C. 2000. The Stanford achievement test, 9<sup>th</sup> edition: national norming and performance standards for deaf & hard-of-hearing students. *J Deaf Stud & Deaf Educ* 5(4):337-348.
- [31] VCom3D. 2011. Homepage. <http://www.vcom3d.com/>
- [32] Ward, G., Hirschberg, J. 1985. Implicating uncertainty: The pragmatics of fall-rise intonation. *Language*, 61: 747-776.
- [33] Young, S.J. 1994. The HTK Hidden Markov Model Toolkit: Design and Philosophy. *Entropic Cambridge Research Laboratory, Ltd.* 2: 2-44.