Selecting Exemplar Recordings of American Sign Language Non-Manual Expressions for Animation Synthesis Based on Manual Sign Timing

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 (RIT)
Golisano College of Computing
and Information Sciences
152 Lomb Memorial Dr, Rochester, NY 14623 USA
matt.huenerfauth@rit.edu

Abstract

Animations of sign language can increase the accessibility of information for people who are deaf or hard of hearing (DHH), but prior work has demonstrated that accurate non-manual expressions (NMEs), consisting of face and head movements, are necessary to produce linguistically accurate animations that are easy to understand. When synthesizing animation, given a sequence of signs performed on the hands (and their timing), we must select an NME performance. Given a corpus of facial motion-capture recordings of ASL sentences with annotation of the timing of signs in the recording, we investigate methods (based on word count and on delexicalized sign timing) for selecting the best NME recoding to use as a basis for synthesizing a novel animation. By comparing recordings selected using these methods to a gold-standard recording, we identify the top-performing exemplar selection method for several NME categories.

Index Terms: American Sign Language, non-manual expressions, exemplar selection, animation synthesis

1. Introduction

Being able to access information sources online has become necessary for employment, engaging in commerce, accessing government services, and in various other contexts in modern society. However, the majority of information content on the web is in the form of written-language text. There are many individuals who have difficulty reading text information sources online, including those with low literacy.

What may be less obvious is that even websites without any audio content present accessibility challenges for people who are deaf or hard of hearing (DHH). Due to a variety of factors, e.g., early language exposure or educational background, many DHH users have lower levels of written language literacy. In the U.S. context, standardized educational testing of secondary school graduates (i.e., students age 18+) has indicated that the majority of DHH graduates have English reading levels at the fourth grade or below [1], which would correspond to age 10 U.S. students. Although some DHH individuals may have difficulty reading written English, many have strong fluency in American Sign Language (ASL).

While presenting videos of ASL on websites is a simple solution, it can be difficult to update and maintain information content in the form of video. Therefore, technology to automate the creation of ASL content (in the form of animation) can make it easier and more cost-effective for companies and organizations to provide ASL content on their websites, as discussed in [2].

This paper focuses on methods for generating non-manual expressions (NMEs), i.e. face and head movements, for ASL animation. One method for producing linguistically accurate and natural NMEs is to select a pre-existing recording of a human ASL signer as a basis for the animation, as discussed in [3]. A challenge is selecting which recording in a corpus is the most suitable to serve as the basis for the face and head movements of the animated character, given that a sentence with specific lexical items (and their timings) must be synthesized. In this paper, we define four methods of considering the manual sign similarity between pairs of recordings, and we conduct an evaluation of how effective each technique is for identifying an exemplar human recording that could serve as a basis for synthesizing NMEs for ASL animations.

1.1. Background on American Sign Language and NMEs

As background, this section briefly summarizes ASL linguistics, with a focus on the use of non-manual expressions (NMEs) in the language. Researchers estimate that there are over a half-million people in the U.S. who use ASL as a primary means of communication [4]. As discussed above, many users of ASL are not fluent in written English; the two languages are linguistically distinct, with differences in word order, linguistic structure, and vocabulary. Generally speaking, movements of the hands and arms are used to indicate lexical items (ASL “manual signs”), but a complete production of ASL consists of much more than this, including head movement, facial expressions, eye-gaze, and torso movements, all of which can convey linguistic information. These additional channels of performance are commonly referred to as NMEs.

NMEs can convey a wide variety of information, including emotional connotation, variations in lexical meaning, or prosodic information. In this work, we focus on Syntactic NMEs, which are used to convey syntactic information about sentence structure. These Syntactic NMEs generally consist of movements of the upper face and movements of the head, and they are performed in parallel with phrases containing manual signs. Syntactic NMEs conveying essential grammatical information about individual words or about entire phrases or clauses [5].

In this paper, we examine five common Syntactic NMEs:

- **Negative:** The signer shakes his head left and right to indicate negated meaning (generally with some eyebrow

10.21437/SLPAT.2016-3

furrowing). For instance, the addition of a Negative NME during the verb phrase "EAT APPLE" in the ASL sentence "TEACHER EAT APPLE" negates the meaning of the clause so that it means "The teacher is not eating the apple." There is a manual sign "NOT" which can optionally be inserted before the verb phrase: While the manual sign is optional, the NME is required.

- **Topic:** The signer raises his eyebrows and tilts his head backward during a clause-initial phrase that should be interpreted as a topic. For instance, a Topic NME would occur during "APPLE" in the sentence "APPLE TEACHER EAT," which translates to English as "As for the apple, the teacher is eating it."
- **Rhetorical:** The signer raises his eyebrows and tilts his head backward and to the side to indicate a rhetorical question. ASL Rhetoricals are immediately answered by signer. For instance, "TEACHER BUY WHAT APPLE" with Rhetorical NME during "WHAT" translates to English as "What is the teacher buying? An apple."
- **Yes-No Question:** The signer raises his eyebrows while tilting the head forward to indicate that the sentence is a yes-or-no question. For instance, the introduction of a Yes-No Question NME during the ASL declarative sentence "TEACHER EAT APPLE" (English translation: "The teacher is eating an apple.") creates a polar question: "Is the teacher eating an apple?"
- **WH Question:** 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. For example, this NME would occur during the ASL sentence "TEACHER EAT WHAT," which translates to English as "What is the teacher eating?"

1.2. Prior Work on NME Animation Synthesis

As discussed in Section 1, posting videos of human signers is not a viable method for providing ASL content on websites. If information must be frequently updated, then re-recording a video of a human signer would be prohibitively expensive; furthermore, a video-based approach would not enable real-time generation of content from a user query. For this reason, "synthesis" software is needed that can convert from a script of an ASL sentence into a full animation of a virtual human performing ASL. This script of the sentence could be generated by a knowledgeable human author or by machine translation software (as the state-of-the-art of machine translation tools for ASL improve in the future). Given the sequence of words in the sentence, the synthesis software must plan the movements of the virtual human character so that the resulting animation is linguistically accurate, understandable, and acceptable by DHH users.

Many researchers have investigated the design of sign language synthesis systems, including research that has specifically focused on the generation of non-manual expressions [6, 7, 8, 9, 10]. Traditionally, researchers select a single recording of how a non-manual expression is performed, and they trigger this movement in parallel to the movements of the virtual human's hands.

In prior work, we have investigated data-driven methods for synthesizing the NMEs of the virtual human. Specifically, our prior work has made use of a small corpus of recordings of a female native signer performing ASL sentences with NMEs. This corpus is relatively small in size, and it has been divided into sub-corpora of sentence recordings for different categories of NMEs (Negation, Rhetorical, Topic, WH Question, Yes-No Question). See Table 1. (We note that sign language corpora are generally small in size, given the resource-intensive nature of obtaining these recordings and the annotation of manual-sign and NME information for individual frames of video.) This corpus was recorded and annotated at Boston University, as described in [3, 11]. The annotations include the timing and identity of manual signs and NMEs, and the videos have been processed by computer vision software [12] to create streams of MPEG4 Facial Action Parameters, which are numerical representations of the movements of various key points on the face [13].

Table 1: *NME corpus characteristics, including the duration of each recoding, in video frames and number of words.*

NME Category (Number of Recordings)	**Video Frames** min - max (mean)	**Num. of Signs** min - max (mean)
Negation (55)	10 - 76 (38.1)	2 - 7 (3.56)
Rhetorical (13)	11 - 46 (28.3)	1 - 4 (3.0)
Topic (96)	5 - 54 (15.5)	1 - 4 (1.43)
WH Question (14)	15 - 69 (31.2)	1 - 5 (2.2)
Yes-No Ques. (21)	9 - 78 (34.6)	2 - 6 (3.6)

Given this resource, our prior work has examined two possible methods for generating animations:

- We have used multidimensional dynamic time warping (DTW) on the MPEG4 FAP values to calculate pairwise similarity between all of the recordings in each sub-corpus, and we calculated the centroid recording in each set, with the minimum pairwise distance to all other members. Assuming that this recording was "most typical" of that category of NME, we used that recording as the basis for synthesizing animations of novel sentences [3].
- We subsequently investigated the use of a generative model of time-series data (Continuous Profile Models) to calculate an underlying "latent-trace" of a set of multiple recordings [11]. We used this latent-trace technique to intelligently "average" across multiple examples of each NME.

A common processing step that is necessary before using either of these two approaches listed above is that we must identify a set of recordings that will serve as the basis for producing a new animation. In prior work, we took the simplistic route of using all of the recordings in our corpus that included the specific category of NME (e.g. Topic) as the "basis set" for calculating our centroid or our latent-trace NME. However, some of those recordings may not have served as good examples of how our virtual human should move, perhaps due to differences between the sentence structure of the corpus recordings and the structure of the sentence we need to synthesize. The premise of this paper is that the selection of a basis set could be determined in a more sophisticated and discerning manner than simply using every recording of that NME category.

1.3. Input to NME Animation Synthesis

To better define the specific task that is the focus of this paper, we list the information and resources that are available during the generation of an ASL NME performance for an animation:

- We assume that the sequence of lexical items has already been determined for the sentence that must be generated. In addition to the identity of each word, we know the timing of when the lexical items begin and end (based partially on the timing information for each sign in the lexicon of our animation system).
- We assume that we already know which spans of lexical items in the sentence need to have an NME performed in parallel. For instance, given an ASL sentence "OLD BOOK I LIKE," we have already selected that a Topic NME should occur during the words "OLD BOOK." In fact, we presented initial research on how to perform this step of the process at SLPAT 2015 [14].
- Finally, we have our corpus of ASL sentence recordings, consisting of videos, the MPEG4 FAP values, and linguistic annotation of manual signs and NMEs (including the video frame numbers when each begins and ends).

2. Basis-Set Selection Techniques

Our task is to determine which of the recordings in our corpus should be included in the basis set for synthesizing an NME performance. Ideally, we would like to select recordings that are similar to the sentence we seek to synthesize. Given the few inputs to our task (listed in the previous section), there are limitations on the types of information that we may consider when defining strategies for selecting items for the basis set: namely, the identity and timing of the manual signs or NMEs. The intuition behind the basis-set selection strategies investigated in this paper is that we may prefer to select sentences with a similar number of words, a similar duration, or similarities in the patterns of the timing of the manual signs. Our selection metric should have the following properties:

1. Two phrases with a similar number of words or with a similar overall time duration should be scored as being similar.
2. Two phrases in which the beginning and ending timings of the words they contain align closely should be scored as similar.
3. Given the small size of our corpus, considering lexically specific information is impractical. Thus, we will consider the timing of manual signs in a "delexicalized" manner; that is, we will replace the sign labels such as "OLD" or "BOOK" in our corpus with a single token, e.g., "SIGN." This, we will not consider the labels of the specific words/glosses – only their timing.
4. A natural unit of time granularity for our analysis is the time duration of a single frame of video, since this is the basis for the linguistic annotation of word and NME timing for the recordings.

2.1. Comparing Temporal Language Signals

Prior to inventing a new metric for scoring the word-timing similarity of recordings of ASL sentences, we first examined the computational linguistic and automatic speech recognition (ASR) literature to examine the methods used to compare language signals with temporal information, specifically those techniques that have been used to evaluate the output of ASR systems against gold-standard annotations of the speech transcript. While there are a variety of metrics used to compare string output, e.g. [15], most techniques are focused on penalizing incorrect string transcription of the speech audio, and thus, scoring techniques rarely incorporate temporal alignment into the score. In our case, we are considering delexicalized word timing similarity.

Researchers focused on ASR temporal alignment accuracy have proposed a variety of metrics, e.g. average word boundary shift [16], and researchers studying speaker-segmentation in recordings of meetings have proposed metrics such as Diarisation Error Rate [17]. However, in both cases, these metrics assume that there will be some word label or speaker-ID correspondences across the two time-annotated transcriptions. Since, for our task, we are focused on delexicalized timing similarity, these previously invented metrics are not well-suited.

As discussed in the next section, some of our proposed metrics make use of Inside-Outside-Beginning (IOB) labelling. For this reason, we also considered comparison metrics in the named entity detection and information extraction literature. While the output of many systems consists of IOB labelling of the tokens in a string, the traditional evaluation metrics in this field are based on per-token precision, recall, or F-score [18]. Such metrics are ill-suited to evaluating fine-grained IOB similarity at the video-frame level, as in our situation. Some authors propose metrics to support evaluation of partial-matches (in which a system's named-entity tagging partially overlaps with the true gold-standard labeling) [19]. However, even these metrics do not consider the temporal dimension at a fine-enough granularity for our task.

2.2. Techniques Examined in This Paper

Since we did not find a suitable pre-existing metric for comparison of the delexicalized timing similarity of the manual component of two ASL sentences, we invented four sets of basis-set selection approaches (and a simplistic baseline), which we will investigate and compare in this paper:

- **Baseline Method.** This simplistic method for defining the basis set was used in our prior work [3]: We filter the corpus, leaving only those recordings containing the specific category of NME that we seek to generate (e.g., Topic). For this baseline approach (and in all of the other approaches listed below), we select and extract the portion of each recording that coincides with the span of time when the NME is occurring in that sentence (based on the linguistic annotation). Thus, if a Topic NME occurs during the first two words of some recording, then we extract the portion of the recording corresponding to this period of time for inclusion in the basis set.
- **Word Count.** This technique is based upon the intuition that an NME that occurs during a portion of a sentence with a large number of words may differ from an NME that occurs during a portion of a sentence containing few words. For instance, facial expressions with periodic movements, such as the head shaking that occurs during Negation, may consist of a larger number of individual movements when it occurs during a longer verb phrase. In this technique, we first filter for only those recordings that contain the category of NME we need to generate (e.g., Negation), as in the baseline approach above. Next, we count how many words co-occur with the NME in each recording, and we select items for the basis set that have a similar number of words within the timespan of the NME. Thus, if we must generate an ASL sentence with a Negation during a verb phrase consisting of five words, then we would prefer to select recordings

from our corpus that contain Negation performances during an identical (or similar) number of words.

- **Frame Count.** This technique is similar to the above, except we use the time duration of the NME (measured according to the number of video frames) as the similarity metric. This, if we needed to generate an ASL animation with a Topic facial expression that must last for 25 frames, then we would prefer to select Topic NME recordings of a similar duration from our corpus for inclusion in the basis set. (Our video recordings have a frame-rate of 30 frames per second.)
- **Levenshtein IOB.** In this technique, we pre-process each of the sentence recordings to generate a string consisting of the letters "I," "O," or "B," representing Inside, Outside, or Beginning, in the following manner: For each frame of video, we add one character to the string, based on whether this frame of video is the Beginning of a manual sign (the single video frame where this word begins), Inside (during) a manual sign, or Outside of a manual sign (i.e. during a period of time in-between signs or before/after all signs in the recording). Thus, an ASL sentence recording of duration 20 frames containing the words "BOOK" (frame 3 to 8) and "LOST" (frame 10 to 15) appears as: `OOBIIIIIOBIIIIIOOOOO`. We select all of the recordings in the corpus that contain the same category of NME (e.g., Topic) as the one we need to generate, and we focus on the IOB substring that corresponds to the time duration of each NME. To calculate similarity between pairs of substrings, we calculate the Levenshtein distance (with equal penalty for insertion, deletion, and substitution, with normalization based on the length of the shorter substring). The intuition behind this technique is that it may capture the temporal structure of a recording in a delexicalized manner such that we would prefer to include recordings in the basis set that consist of NME recordings with a similar number of words with similar word durations and timing.
- **Bigram IOB.** This technique uses a similar IOB string representation as above. After extracting the substrings that correspond to all examples of the category of NME we must generate (e.g. Rhetorical), then we count all character bigrams in each IOB substring. These counts are stored in a vector corresponding to each string; to calculate the similarity between a pair of recordings, we use the cosine similarity between their vectors. The intuition behind this approach is that it may capture some information about both word count (based on the number of n-grams containing the "B" character), and it would also indicate overall time duration (with longer recordings having higher counts in the vector cells).

3. Evaluation of Selection Techniques

Given the inputs described in section 1.3, a good basis-set selection technique would identify a subset of ASL recordings in our corpus that contain similar face and head movements to what a human would perform for the ASL sentence that we seek to synthesize.

3.1. Scoring Metric Used in This Evaluation

In prior work presented at SLPAT 2015, we demonstrated that multidimensional dynamic time warping (DTW) operating in the space of MPEG4 Facial Action Parameters can assign similarity scores to pairs of ASL NME recordings that correlate with the judgements of native ASL signers [20], and we defined a refined version of this scoring algorithm in [3]. This scoring algorithm provides a numerical score of the similarity in the face and head movements (specifically the eyebrows and head displacement/orientation) between any pair of ASL recordings. In the evaluation presented below, we use this multidimensional DTW scoring algorithm to evaluate how well each of the selection techniques is able to chose a basis set with recordings that are similar to gold-standard human performances.

3.2. Evaluation Methodology

We compared the efficacy of each of the five techniques listed in section 2.2, for each of the categories or NME in our corpus (Negation, Rhetorical, Topic, WH Question, Yes-No Question), using a leave-one-out evaluation paradigm, described below. To explain the process more clearly, we will discuss, by way of example, how the process occurs for the WH Question recordings.

1. We extracted a set of all the recordings in our corpus for this category of NME (e.g., there were 14 recordings of WH Question in our corpus). We iteratively held-out each of the recordings in this set (i.e., we repeated this process for all 14 items in the set of WH Question recordings), and we consider the held-out recording to be a gold-standard of how a human should move his face and head when performing the NME for the given sequence (and timing) of manual signs in this sentence. The remaining 13 recordings are used as the superset from which the basis set must be drawn, for this held-out recording.
2. For each of the basis-set selection techniques, we identify a subset of the recordings that are predicted to yield NME movements that are similar to the gold-standard held-out recording. We use each of the five selection techniques to identify a (potentially) different basis set.
   (a) For the baseline method, this is trivial: In the case of WH Question, we would simply use all 13 of our non-held-out recordings in the superset in order to form our basis set.
   (b) For the remaining four selection techniques, the similarity scoring methods defined in section 2.2 enable us to assign a score to each of the 13 WH Question recordings. For each of the selection techniques, we select the top 5 most similar recordings to form a basis set. Thus, each of the four selection techniques will be used to produce its own basis set (with cardinality 5), and each basis set may have different membership, as determined by that selection technique.
3. To evaluate the quality of the basis set chosen by each selection technique, we must compare how well the face and head movements of each of the recordings in the set matches the face and head movements of the held-out recording (considered as a gold standard). Using the DTW metric from [3] mentioned above, we calculate the distance between each of the recordings in the basis set and the held-out gold-standard recording. To produce a single score for each basis set, the individual distance-to-gold-standard scores for the members of the set are averaged to produce a single score.

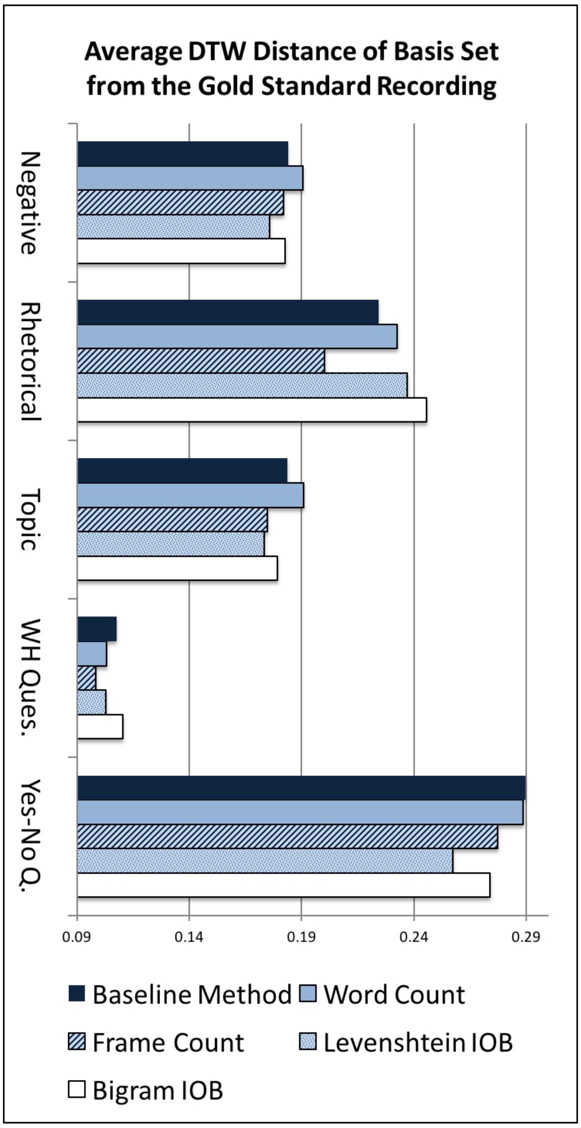

Figure 1: *Average DTW distance between basis set members and gold-standard sentences, for each NME category, for each selection technique. Note that smaller bars are better.*

4. Discussion of Results

As shown in Figure 1, at the end of the evaluation process, for each NME category, for each of the five selection techniques, we have a single score that represents how well that selection technique was able to identify a basis set of recordings from our corpus that were similar to human performance of that NME for the held-out gold-standard sentences.

For three of the NME categories (Negative, Topic, and Yes-No Question), the best performing selection technique was Levenshtein IOB. For the Rhetorical and WH Question categories, the best performing selection technique was Frame Count. (For WH Question, the performance of all of the selection algorithms was quite close, with Levenshtein IOB in second place.)

Our corpus contains relatively few recordings of Rhetorical (13) and WH Question (14), and due to the nature of how these NMEs are used in ASL, many of these recording examples occur during phrases consisting of a single word (e.g., often a single WH-word). We speculate that the difference in efficacy of the selection techniques for these two categories may relate to the relatively low cardinality of examples in our dataset and the relatively short duration of these NMEs.

No selection algorithm obtained the best (lowest) distance scores across all five categories of NME, and in principle, it is reasonable that a different selection technique could be best suited to each of the NME categories. This could be due to the way in which the lexical timing of manual signs may influence how that particular NME is performed by ASL signers.

5. Conclusions

This paper has investigated techniques for selecting a subset of recordings from a corpus that can be used as a basis for synthesizing the Syntactic NMEs for a sentence to be generated, based only on information about the delexicalized manual sign timing of the sentence. By identifying a set of similar recordings for inclusion in this basis set, various approaches can be used to select a single recording [11] or to identify a latent-trace of the set [11], in order to plan the face and head movements of a virtual human in the ASL animation. Ultimately, the goal of this work is to improve the state of the art of sign language animation synthesis technologies, especially since prior studies have demonstrated that the understandability of such animations is affected by the quality of the synthesized NMEs. Such technology has potential to make it easier for organizations to provide sign language content on websites in a manner that is more efficient and easier to maintain, which may increase the prevalence of such content online.

In future work, we plan to evaluate the efficacy of these basis-set selection techniques within the context of a full animation synthesis pipeline. By performing final animation production step, we can generate stimuli for display in user-based evaluation studies, in which native ASL signers could view animations generated using these selection algorithms as an intermediate pipeline stage. In this way, we can determine the degree to which the differences in efficacy identified in this study may influence DHH users' perception of the linguistic accuracy and understandability of the resulting animations.

In this study, we found that the Levenshtein IOB metric was most effective at selecting basis set recordings for three of the five NME categories in this study, and the number of recordings in our corpus for the remaining two categories (Rhetorical and WH Question) was relatively small. In future work, we are interested in acquiring additional ASL recordings of these NMEs from multiple signers so that we may repeat this analysis on a larger testing set.

6. Acknowledgements

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.

7. References

[1] C. B. Traxler, "The stanford achievement test: National norming and performance standards for deaf and hard-of-hearing students," *Journal of deaf studies and deaf education*, vol. 5, no. 4, pp. 337–348, 2000.

[2] M. Huenerfauth, "Generating american sign language animation: overcoming misconceptions and technical challenges," *Universal Access in the Information Society*, vol. 6, no. 4, pp. 419–434, 2008.

[3] H. Kacorri, A. Syed Raza, M. Huenerfauth, and C. Neidle, "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*, 2016.

[4] R. E. Mitchell, T. A. Young, B. Bachleda, and M. A. Karchmer, "How many people use asl in the united states? why estimates need updating," *Sign Language Studies*, vol. 6, no. 3, pp. 306–335, 2006.

[5] C. Neidle, D. Kegl, D. MacLaughlin, B. Bahan, and R. Lee, "The syntax of asl: functional categories and hierarchical structure," 2000.

[6] C. Schmidt, O. Koller, H. Ney, T. Hoyoux, and J. Piater, "Enhancing gloss-based corpora with facial features using active appearance models," in *International Symposium on Sign Language Translation and Avatar Technology*, vol. 2, 2013.

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

[8] N. Adamo-Villani and R. B. Wilbur, "Asl-pro: American sign language animation with prosodic elements," in *International Conference on Universal Access in Human-Computer Interaction*. Springer, 2015, pp. 307–318.

[9] M. Huenerfauth and P. Lu, "Modeling and synthesizing spatially inflected verbs for american sign language animations," in *Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility*. ACM, 2010, pp. 99–106.

[10] M. Huenerfauth, L. Zhao, E. Gu, and J. Allbeck, "Evaluation of american sign language generation by native asl signers," *ACM Transactions on Accessible Computing (TACCESS)*, vol. 1, no. 1, p. 3, 2008.

[11] H. Kacorri and M. Huenerfauth, "Continuous profile models in asl syntactic facial expression synthesis," in *ACL 2016: the 54rd Annual Meeting of the Association for Computational Linguistics*. Curran Proceedings, 2016.

[12] T. Visage, "Face tracking," https://visagetechnologies.com/products-and-services/visagesdk/facetrack, 2016, accessed: 2016-03-10.

[13] I. S. Pandzic and R. Forchheimer, *MPEG-4 facial animation: The standard, implementation and applications*. Wiley, 2003.

[14] S. Ebling and M. Huenerfauth, "Bridging the gap between sign language machine translation and sign language animation using sequence classification," in *Proceedings of the 6th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 2015.

[15] S. Dobrišek and F. Mihelic, "Criteria for the evaluation of automated speech-recognition scoring algorithms," *Electrotechnical Review*, vol. 75, no. 4, pp. 229–234, 2008.

[16] L. Chen, Y. Liu, M. P. Harper, E. Maia, and S. McRoy, "Evaluating factors impacting the accuracy of forced alignments in a multimodal corpus." in *LREC*, 2004.

[17] S. Tranter, K. Yu, G. Everinann, and P. C. Woodland, "Generating and evaluating segmentations for automatic speech recognition of conversational telephone speech," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1. IEEE, 2004, pp. I–753.

[18] M. Marrero, S. Sánchez-Cuadrado, J. M. Lara, and G. Andreadakis, "Evaluation of named entity extraction systems," *Advances in Computational Linguistics, Research in Computing Science*, vol. 41, pp. 47–58, 2009.

[19] S. Atdağ and V. Labatut, "A comparison of named entity recognition tools applied to biographical texts," in *Systems and Computer Science (ICSCS), 2013 2nd International Conference on*. IEEE, 2013, pp. 228–233.

[20] H. Kacorri and M. Huenerfauth, "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)*, 2015, p. 29.