Efforts to Improve Avatar Technology for Sign Language Synthesis

Robyn Moncrief
Computing and Digital Media, DePaul University, Chicago, IL.
rmoncrie@depaul.edu

Shatabdi Choudhury
Computing and Digital Media, DePaul University, Chicago, IL., DePaul University
schoud12@depaul.edu

Maria Del Carmen Saenz
Computing and Digital Media, DePaul University, Chicago, IL., DePaul University
msaenz@depaul.edu

ABSTRACT

A promising method for increasing Deaf accessibility is the technology of sign language synthesis, which can be used to automate the translation of spoken or written language to sign language through the manipulation of an avatar. Efforts to automatically translate spoken language to sign have lagged behind spoken-to-spoken translation. Avatars are used in various capacities for sign language display, including translation and educational tools. Though the ability of avatars to portray acceptable sign language producing believable human-like motion has improved in recent years, many still lack the naturalness and supporting motions of human motion. This paper presents current efforts being made at the DePaul University ASL Lab to improve the methods of sign synthesis using avatar technology and create believable human-like motion.

CCS CONCEPTS

• **Software notations and tools**; • **Machine learning**; • **Accessibility**;

KEYWORDS

Avatar Technology, Sign Language Synthesis

ACM Reference Format:
Robyn Moncrief, Shatabdi Choudhury, and Maria Del Carmen Saenz. 2022. Efforts to Improve Avatar Technology for Sign Language Synthesis. In *The15th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '22), June 29–July 01, 2022, Corfu, Greece.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3529190.3534783

1 INTRODUCTION

Access to language is a human right, but one that is often denied to members of the Deaf community. Deafness is a language barrier that is created by more than a lack of sound. For the Deaf community in the U.S. and parts of Canada, American Sign Language (ASL) is the preferred language. ASL and English share some vocabulary, but ASL is an independent language with its own unique grammar and there is no word for word translation between it and English [1]. ASL is composed of handshapes, position, movement, palm orientation, and non-manual signals such as facial expressions, mouthing, and movements of the head and spinal column. For a successful English to ASL synthesizer, all components of the language should be incorporated. Avatar sign syntheses allow for the flexibility for modification of signs and generation of sentences, while respecting ASL's grammatical structure.

2 AUTOMATING ADVERBS OF MANNER IN SIGN LANGUAGE

One of the challenges of synthesizing ASL is that there is not necessarily an independent lexical item to express parts of the language. A case in point is the portrayal of adverbs of manner. Instead, adverbs of manner are often expressed in the quality of the signing. In ASL, adverbs of manner modify the "quality of motion" of a signed verb and are considered non-manual, incorporating gesture [2] [3] [4] [5] [6] [7]. Automating the process of applying an adverb of manner to a sign required building a module that could apply motion modification to the lexical verb and overlay additional non-manual signals. Keyframe animations were modified procedurally based on observations from a selection of signs incorporating adverbs of manner. This modification accounted for motion path adjustments in timing, joint positions along the avatar's arm and spine, and non-manual signal adjustments in facial expression and head movement. This model allows for the application of adverbial modification to any previously animated sign.

3 MOUTHING RECOGNITION WITH OPEN POSE IN AMERICAN SIGN LANGUAGE

Many avatars focus on the hands and how they express sign language. However there has been ongoing interest in how the face can convey information [11]. To have a translation system that will be accepted by the Deaf community, we will need to include non-manual signals such as mouthing and mouth gestures. Just as machine learning is being used on generating hand signs, the work we are focusing on will be doing the same but with mouthing and mouth gestures. We will be using data from The National Center for Sign Language and Gesture Resources. The center has videos of native signers focusing on different areas of signer movement, gesturing, and mouthing, and is annotated specifically for mouthing study as shown in Figure 1. This annotated video data will be running a pre-trained Neural Network model that extracts 2D key face and body points via the application OpenPose [13]. Once those 2D key points are extracted, we will further analyze the data using a Random Forest Classifier, which has been used before for hand gesturing analysis [8].

PETRA '22, June 29–July 01, 2022, Corfu, Greece

ACM ISBN 978-1-4503-9631-8/22/06.
https://doi.org/10.1145/3529190.3534783

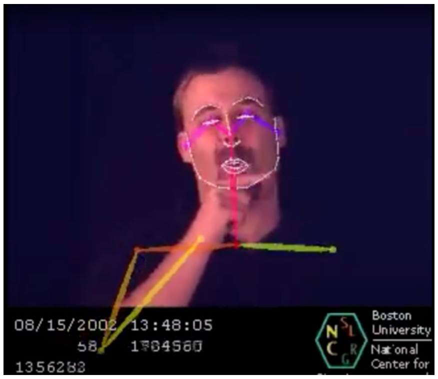

Figure 1: Native signer video via the National Center for Sign Language and Gesture Resources with OpenPose 2D face key points on keyframe. (https://www.bu.edu/asllrp/cslgr/.

Looking over the accuracy samples, we realized we needed more data, and decided to use a resampling method called SMOTE, a Synthetic Minority Oversampling Technique [14]. This assisted us in seeing an increase in accuracy, in identifying the mouthing annotations on the 2D key points from OpenPose. The main objective is to train the algorithms to spot certain mouthing points and output the mouth annotation with a high degree of accuracy, which you can see from Table 1, is a viable option if we have more data to analyze.

4 USING MOTION CAPTURE TO EXTEND AN ANALYTICAL MODEL OF TORSO MOVEMENTS

Sign languages are communicated through more than just hand movements. Torso movements are critical for direct linguistic communication and supporting hand and arm motions. These supporting motions are highly essential to displaying lifelike and communicative animations. However, such details are generally not included in the linguistic annotation. Our study focused on building a data-driven analytic model for use in a signing avatar that will automatically compute the torso position based on other avatar body parts. We expect it to improve the user experience and engagement with the avatar.

The following figures show two illustrations of torso movements. In Figure 2, the signer is twisting her torso to depict objects to the side of the signing space, and in Figure 3, the signer is bending backward to depict objects to the front of the signing space.

We analyzed a corpus of human motion data recorded with a motion capture system from four signers [12]. The combined dataset has 34 columns and approximately 66600 rows with information about torso rotations, wrist positions in space (X, Y, and Z), and palm orientations. The torso's spinal twist and side and forward motions are the three parameters of interest. We created regression models using manual and automatic model-building techniques. We selected the linear regression model to implement in the avatar because it has a closed-form and is more adaptable to inclusion in linguistic models in the future. It is also intuitive and easier to implement in the python code in the avatar.

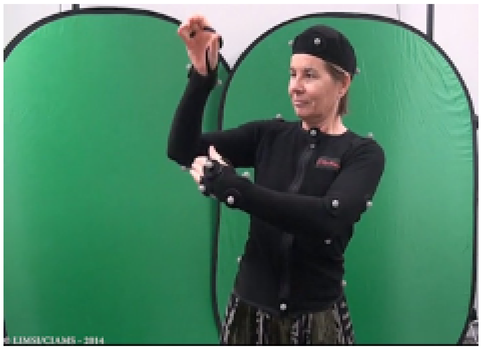

Figure 2: Twisting the torso.

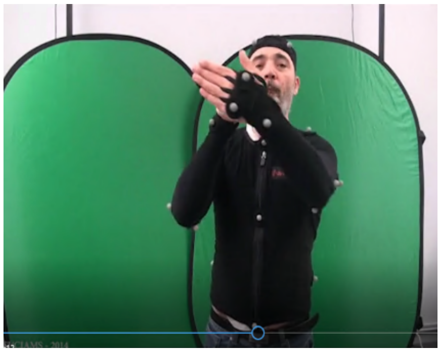

Figure 3: Bending the torso.

REFERENCES

[1] W. Stokoe, "Sign language structure," Studies in Linguistics, Occasional Papers 8, Buffalo, New York, (1960)

[2] A. J. Thomson and A. V. Martinet, A practical English grammar, Oxford University Press, 1980.

[3] C. L. Baker-Shenk and D. Cokely, American Sign Language: A teacher's resource text on grammar and culture, Gallaudet University Press, 1991.

[4] T. N. Kluwin, "A rationale for modifying classroom signing systems," Sign Language Studies, vol. 31, p. 179-187, Gallaudet University Press, 1981.

[5] Jerry Schnepp, Rosalee Wolfe, John McDonald and J. A. Toro, "Combining emotion and facial nonmanual signals in synthesized american sign language," in Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility, 2012.

Table 1: Accuracy of algorithms for spotting mouthing points.

Dataset	Validation Accuracy	Test Balance Accuracy
With Resampling	0.96 (+/-0.01)	0.67
Without Resampling	0.43 (+/-0.03)	0.44

[6] T. Johnston and L. De Beuzeville, “Auslan corpus annotation guidelines”, Auslan Corpus, 2016.

[7] C. A. Padden, Interaction of morphology and syntax in American Sign Language, Routledge, 2016.

[8] Ruiliang Su, Chen Xiang, Shuai Cao, & Xu Zhang. (2016). Random Forest-Based Recognition of Isolated Sign Language Subwords Using Data from Accelerometers and Surface Electromyographic Sensors. Sensors. 16. 100. 10.3390/s16010100.

[9] Oscar Koller, Hermann Ney, & Richard Bowden. (2015). Deep Learning of Mouth Shapes for Sign Language. 10.1109/ICCVW.2015.69.

[10] Z. Cao, G. Hidalgo, T. Simon, S. -E. Wei and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 172-186, 1 Jan. 2021, doi: 10.1109/TPAMI.2019.2929257.

[11] John McDonald, Rosalee Wolfe, Robyn Moncrief, Souad Baowidan, & Jerry Schnepp, Kinematic Model for Constructed Dialog in American Sign Language. 6th Conference of the International Society for Gesture Studies, San Diego, CA, July 8-11, 2014.

[12] Mohamed-El-Fatah Benchiheub, Bastien Berret, & Annelies Braffort,. (2016). Collecting and Analysing a Motion-Capture Corpus of French Sign Language.

[13] Zhe Cao, Tomas Simon, Shih-En Wei, Yaser Sheikh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7291-7299

[14] Chawla, Nitesh V., *et al.* "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16 (2002): 321-357.