Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology (SLTAT 7), pages 7–12
Language Resources and Evaluation Conference (LREC 2022), Marseille, 20-25 June 2022

Analysis of Torso Movement for Signing Avatar Using Deep Learning

Shatabdi Choudhury
School of Computing, DePaul University, 243 S. Wabash Ave, Chicago, IL 60604, USA,
schoud12@depaul.edu

Abstract

Avatars are virtual or on-screen representations of a human used in various roles for sign language display, including translation and educational tools. Though the ability of avatars to portray acceptable sign language with believable human-like motion has improved in recent years, many still lack the naturalness and supporting motions of human signing. Such details are generally not included in the linguistic annotation. Nevertheless, these motions are highly essential to displaying lifelike and communicative animations. This paper presents a deep learning model for use in a signing avatar. The study focuses on coordinating torso movements and other human body parts. The proposed model will automatically compute the torso rotation based on the avatar's wrist positions. The resulting motion can improve the user experience and engagement with the avatar.

Keywords: Sign Language, Avatar, Neural Network, Deep Learning

1. Introduction

Interactive avatars have grown popular as learning tools for spoken languages. Virtual reality has become a new tool to aid deaf or hard-of-hearing learners with specialized guidance in learning core academic concepts, such as mathematics and science (Zirzow, 2015). A signing avatar has been proposed to assist deaf students in a comprehensive educational environment(De Martino et al., 2017). Avatars are evaluated as a potential communication medium to facilitate language learning in babies (Nasihati Gilani et al., 2019).

Avatars are also increasingly popular in social media, personalizing users' contributions to interacting and representing users and their behaviors. People prefer to have an avatar in their profile to secure their visual anonymity or pseudo-anonymity. Anonymity enables them to express and observe opinions they would not necessarily be comfortable with elsewhere while holding personal characteristics (Vasalou et al., 2008). Historically, people have adopted a pen name or alias to express themselves anonymously for several reasons. Deaf experience in signing online is inherently not anonymous. An avatar would help signers who do not want to reveal their identity.

There is a rise of a new generation of AI avatars for speech interaction, such as Amelia, that serves as virtual cognitive assistant (Davenport et al., 2020). Deep learning capacities support her ability to learn human interaction continuously and create an engaging user experience to drive higher business value.

The use of avatars in signing can be equally exciting and has potential benefits over video recordings of a sign. One can see a sign from a different angle or zoom in, or the pace or rhythm of the signing can be customized based on users' needs. The scene's background can also be adapted based on the context or better clarity of the sign (Jaballah and Jemni, 2014). The modeling and animation of signed contents generated once can automatically become reusable software components, which can be re-purposed for novel utterances.

Though the need to make avatars natural is widely recognized in the animation industry, the current quality of signing avatars is still not satisfactory for producing human-like user experiences, making them less acceptable to the Deaf community (Jancso et al., 2016). Since speech is missing in sign language, supporting movements are essential to engage the communication. This research recognizes that torso movements are critical for direct linguistic communication (McDonald et al., July 8 11 2014). However, incorporating the coordination for each movement is a time-consuming process for animators. The avatars driven by linguistic input become robotic because linguistic descriptions lack the subtleties of human motion. Motion capture can automatically incorporate natural torso support but is inflexible for generating new signing that has not been specifically recorded. Furthermore, multiple processes in signing can affect the torso simultaneously, and the effects can be difficult to separate or isolate in such recordings (Filhol et al., 2017).

This paper introduces a novel application of deep learning to predict the torso movement of a signing avatar. The method will build a deep sequential neural network, implement it in the avatar and test it against the source motion capture data for validation.

2. Importance of Torso Motion in Signing

Analyzing and modeling the supportive motions, such as torso movements, is crucial to make the avatar mimic human movements accurately. The motions supported by the torso include reach, balance, emotion, or the turning of the body to assume participants' positions in reported speech or to indicate a side-facing object. The principles of overlapping movements are essential for the avatar to get a natural and believable feel (Burleigh et al., 2018).

The following figures show three illustrations of torso movements during the signing of a scene description. In Figure 1, the signer is twisting her torso to produce a side-facing sign, and in Figure 2, the signer is leaning her torso to the side to balance. In Figure 3, the signer is bending backward to illustrate a scene.

An arm raised outwards and another arm moved across the body impacts how the torso is twisted and should be positioned to look natural. Shoulder, wrist, and hand movements must be carefully considered, especially when transitioning from one type of composition to another. One must

Figure 1: Twisting the torso to depict objects to the side of the signing space

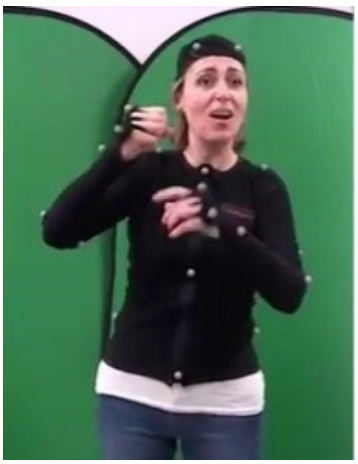

Figure 2: Leaning the torso to the side for balance

consider how much the wrist is bent and how the elbow is raised to orient the palm. All these specific actions can make the avatar more realistic in its movements. Modeling and automating such coordination would result in a practical, accurate, and interactive synthesis. It will elevate the avatar to drive a deeper connection with users.

3. Related Work

Though the natural movements of the spine are captured and can be directly replayed from motion capture data, the segmentation and synthesis of novel discourse from motion capture data is a complicated process that is the focus of ongoing research (Gibet, 2018). Many efforts for sign synthesis focus on describing sign language using a phonetic description called the Hamburg Notation System (Hanke, 2004) (Efthimiou et al., 2010). It subdivides the movements of the signer into a string of individual specifications for the parts of the body. The linguistic descriptions do not encode the motions of the torso unless they have a specific linguistic meaning (Kennaway, 2015).

The Paula avatar uses a heuristic adjustment (McDonald et al., 2016) for the torso position and does not consider other features, such as the neck, shoulder, or wrist orientations. The heuristic model was created based on the artist's proficiency with animating signs and not on data-driven insights. It modeled a precise interaction of the spine and the arms to save the artist time setting up initial poses rather than general movements of the arms and torso. Furthermore, the kind of movements discussed in the last section applies to only the reaching action of the torso. There is no research yet to coordinate torso movement with hand movement in a general way for a signing avatar using a data-driven model. This paper addresses this need by studying a motion capture data set through deep learning.

Figure 3: Bending the torso to depict objects to the front of the signing space

Neural networks (Bishop, 1994) have been used in sign language synthesis. They are employed to combine motion capture sequences for novel utterances for Japanese Sign Language (Brock et al., 2018). It is also used to classify hand positions for signing avatars (Jaballah and Jemni, 2014). Neural Networks have also been explored for their ability to generate continuous 2D skeletal signing motion based on video (Stoll et al., 2018). However, these have not considered direct 3D models of torso postures driven by the positions of the signers' hands.

4. Proposed Solution

This study focuses on the coordination of the torso with other body parts during the signing. The resulting framework predicts the torso movements of a signing avatar. The framework is based on a deep neural network, which learns from large motion capture (Mocap) data sets of human signers. A neural network in this context can learn and detect nonlinear relationships between independent and dependent variables. A sequential neural network model was used since it is the simplest model and can learn without prior application knowledge to find human motions (Baccouche et al., 2011). Implementing the proposed solution on the avatar will produce lifelike natural postures.

Due to the opacity of the neural network for interpretation, a regression model was also trained to compare with the neural network result. This companion model aids the interpretation of the primary relationships computed by the more sophisticated black-box neural network model.

This study is focused on creating a framework that will produce an improved natural movement of the torso in the

avatar, which is based solely on the position and orientation of the signer's wrists. The suggested solution must be accurate compared to the actual human signer.

4.1. Data set

The LIMSI, CNRS laboratory collected human motion data for Langue des Signes Française, (LSF) in BVH format. The data were recorded with a mocap system, video, and annotations of signed descriptions of scenes elicited by a picture from human signers (Benchiheub et al., 2016).
The mocap data is recorded with sensors along the spine, neck, head, shoulders, elbow, and wrist orientations. Since the signing consists of descriptions of scenes, it has very few lexical signs in the data that would differ highly from one sign language to another (Baker et al., 2016). So, even though the recorded signing is in LSF, the body postures captured are applicable across sign languages. However, this should not be generalized to all types of signers.
The 3DS Max software package was used to import the motion capture data, convert the data to match the avatar's coordinate system, resolve data issues, such as outliers, derive new variables required for the ML model, and finally save the data to CSV files. Python scripts tested the data, combined all CSV files into a master file, and performed other intermediate tasks.
The data covered four signers, 25 descriptions, 25 frames per second and roughly 800 frames per description in the study. The final data set has 66644 rows and 34 columns. A specific signer was chosen to train the model to avoid confusion with different signing styles because the specific signer's style is highly consistent, while other signers use more excessive body movements.

4.2. Target definition

Three attributes for the spinal movement in the data were the primary targets: the torso's twist, side, and forward motions. The three Twist, Side, and Forward attributes across the spine bones were summed up to a derived variable to simplify the computation. It helped reduce the target or dependent variable set from 12 to only three attributes and gave a better idea of the overall movement of the spine.

Name	Action	Rotation Axis
Twist	Transverse twisting	Z-axis
Side	Lateral bending	X-axis
Forward	Sagittal bending	Y-axis

Table 1: Torso movements

4.3. Linear regression

The primary motivation to start with linear regression is that it is highly interpretable and enables a better understanding of the independent variables' impact on the dependent variable. The linear regression model enabled to match the coordinate systems of the motion capture data, where the data comes from with the signing avatar, where the model is implemented. It helped to calibrate the model. It also served as the baseline model, critical for capturing the evaluation metrics before initiating the deep learning model. The steps followed from start to end are shown in Figure 4.

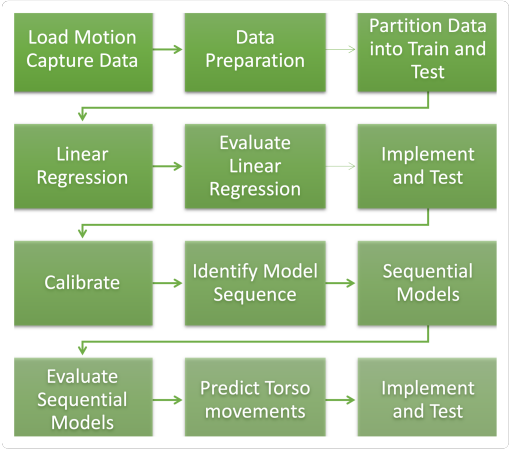

Figure 4: Process diagram

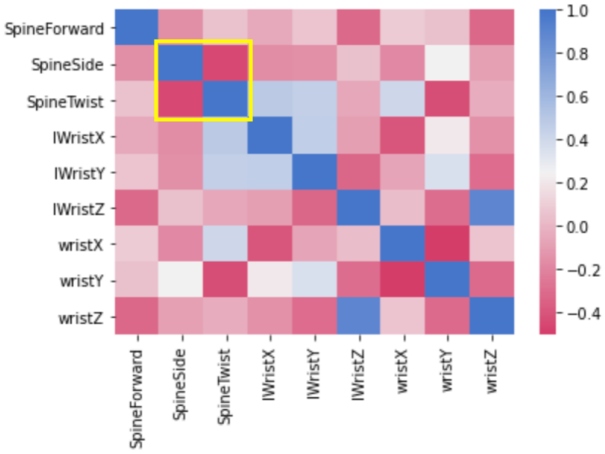

Figure 5: Correlation analysis

The independent variables employed in the study are linear X, Y, and Z positions of the left and right wrists. Based on the exploratory correlation analysis shown in Figure 5 and experiments with linear regression, it was determined that chained regression between the dependent variables was appropriate. The analysis identified a linear sequence to arrange three models. The first model uses all independent variables and predicts spine twist. The second model uses all independent variables and the prediction output from the previous model to predict the spine side rotation, and so on. The chained regression approach increased the predictive power of the model significantly.

High accuracy is the main priority to make the signing avatar natural. However, the evaluation metrics from Linear Regression, such as mean squared error and R-Squared, are not satisfactory. The regression formulas for each of the three movements are displayed in equations (1) - (3).

The equations helped a more intuitive knowledge of the relationship between the independent and dependent variables, such as if wrist X increases, the twist also increases, and so on.

$$
\begin{aligned}
Twist = 0.02 &+ 0.12 * wristX \\
&- 0.20 * wristY \\
&+ 0.04 * wristZ \\
&+ 0.16 * lWristX \\
&+ 0.16 * lWristY \\
&- 0.05 * lWristZ
\end{aligned} \tag{1}
$$

$$
\begin{aligned}
Side = 1.56 &- 0.02 * SpineTwist(Predicted) \\
&- 0.03 * wristX \\
&+ 0.08 * wristY \\
&- 0.14 * wristZ \\
&- 0.05 * lWristX \\
&- 0.04 * lWristY \\
&+ 0.15 * lWristZ
\end{aligned} \tag{2}
$$

$$
\begin{aligned}
Forward = 5.78 &+ 0.45 * SpineTwist(Predicted) \\
&+ 8.57 * SpineSide(Predicted) \\
&+ 0.18 * wristX \\
&- 0.58 * wristY \\
&+ 1.09 * wristZ \\
&+ 0.32 * lWristX \\
&+ 0.25 * lWristY \\
&- 1.22 * lWristZ
\end{aligned} \tag{3}
$$

4.4. Applying the neural network

Deep learning techniques can train the nonlinear representation of data through multiple hidden layers. The deep learning structure can perform feature extraction and transformation without prior knowledge. Keras, an open-source neural network library (Chollet and others, 2015) was used. Keras runs on top of the TensorFlow platform (Bisong, 2019), used to run computations requiring tensors. A tensor can be considered a machine that accepts vectors as inputs and produces another vector as output. The most straightforward way to build a deep learning model in Keras is a sequential model. The sequential model is suitable for a typical stack of layers where each layer has precisely one input tensor and one output tensor. Figure 6 shows the sequential model summary used in the study.

The model used a simple multi-layer perceptron with three layers with the shape of the independent variables (predictors) as a parameter. The first and second layers contain 64 units with rectified linear activation function (ReLU), and the output layer contains just one unit. The network used the "Adam" optimizer, a stochastic gradient descent method for the training model. The 'Mean Squared Error' served as the regression loss function that the model minimized during training. The model was trained on movement information from the descriptions of the first 80% of scenes and held out the rest as a test set. The training sample was used to build a deep learning model and the test sample to evaluate the model. The regression metrics reported are loss, root mean squared error (RMSE), and R-Squared. The optimal model was chosen based on test RMSE using the smallest value as it also indicates the overall expected error in the predictions.

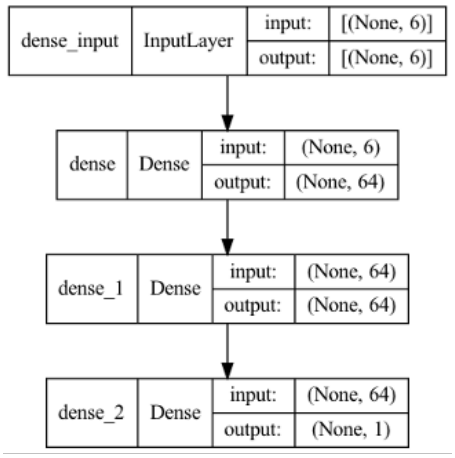

Figure 6: Model summary

4.5. Chained regression approach

As indicated by the regression models, the neural network also followed a chained approach. The first model predicts spine twist by using all independent variables. The second model uses all independent variables and the prediction output from the spine twist model to predict the spine side. The third model uses all independent variables, predicted output from the previous two models, and predicts spine forward.

5. Results

The models are evaluated based on the resulting predictive performance on the holdout test data using loss, RMSE, and R- Squared of the test set with 100 epochs. Table 2 shows the performance metrics of three dependent variables.

	Twist	**Side**	**Forward**
MSE	1.68	4.82	6.89
RMSE	1.30	2.20	2.63
R^2	0.95	0.70	0.48

Table 2: Performance of the neural network models

The results show that the proposed application significantly improves accuracy over linear regression, the baseline, and the companion model. It also improves accuracy over the heuristic model from (McDonald et al., July 8 11 2014), which currently used on the avatar. The RMSE using the neural network is 1.3, while the RMSE using the heuristic model using identical predictors is 4.60 in spine twist movement. Compared to the heuristic methods, the proposed model resulted in a 72% reduction of RMSE for the twist. Tables 3 to 5 compares the models based on RMSE of the predicted spine angles in degrees. The comparison includes the performance of the regression and heuristic methods.

Model	**R-Squared**	**RMSE**
Neural Network	0.95	1.30
Linear Regression	0.86	3.62
Heuristic Model	-	4.60

Figure 7: Signing avatar twisting the torso to depict objects

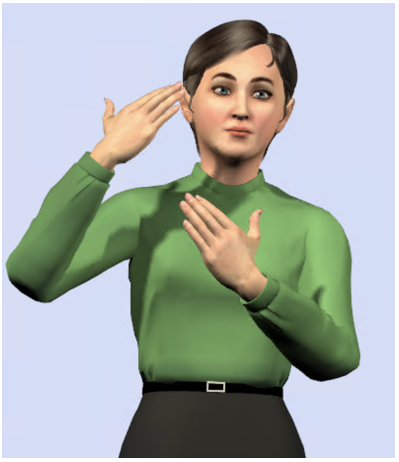

Figure 8: Signing avatar leaning the torso to the side for balance

Table 3: Spine Twist Performance Comparison

Model	R-Squared	RMSE
Neural Network	0.70	2.20
Linear Regression	0.24	11.29
Heuristic Model	-	3.5

Table 4: Spine Side Performance Comparison

Model	R-Squared	RMSE
Neural Network	0.48	2.63
Linear Regression	0.17	21.73
Heuristic Model	-	3.2

Table 5: Spine Forward Performance Comparison

6. Implementation

The model is successfully implemented in the avatar using Python and tested against the original mocap positions. We scaled the torso movements to adapt the morphology of the avatar to that of the skeleton of the captured data. Examples for each of the three key spine movements are displayed in Figures 7 to 9. Naturalness is a piece of subjective information, and there is an effort to figure out how to measure it. A

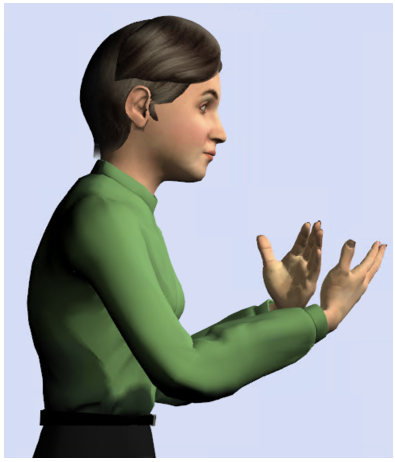

Figure 9: Signing avatar leaning forward the torso to depict objects

user survey from the ASL community, which combines the Deaf community and the experts in the ASL domain, will be requested to compare the avatar with and without the proposed solution. The outcomes of the survey will serve as a measure of naturalness. Currently, the performance is fast enough for the avatar to respond to user interaction in real-time. This framework will be updated once future data is collected, so the model will learn using new data.

7. Conclusions and Future Work

This paper describes the potential power of the proposed model to compute the torso positions of the avatar, which will improve the interaction and engagement of users with the avatar. The proposed model is implemented on an avatar using motion capture data. The initial testing and validation produce satisfactory results. In sign language, signing style varies from person to person. Different signers use the torso in very distinct ways. Some signers like to move more than others. The future effort has started incorporating personal signing styles and refining the models to include additional independent variables and data. Additionally, work is in progress to create a multi-target neural network model to combine the current implementation's three models. The unified model will streamline the implementation process and may deliver better predictions than individual models. Deep neural networks have many parameters, and it is usually prone to overfitting. Since the model will soon include more independent variables, it may have overfitting issues. The companion linear regression model will be leveraged to prevent the overfit. There is a plan to handle overfitting by applying regularization techniques or tuning the neural network parameters. Though the primary focus of the study is sign language avatars, the model can be implemented in any other human animation. There are plans to apply this framework to other sign languages, such as German or Mexican.

8. Acknowledgement

This paper and the study would not have been possible without the outstanding support of Dr. John McDonald. The assistance provided by Dr. McDonald, Dr. Wolfe and

the entire ASL team at DePaul University is deeply appreciated.

9. Bibliographical References

Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., and Baskurt, A. (2011). Sequential deep learning for human action recognition. In *International workshop on human behavior understanding*, pages 29–39. Springer.

Baker, A., van den Bogaerde, B., Pfau, R., and Schermer, T. (2016). *The linguistics of sign languages: An introduction*. John Benjamins Publishing Company.

Benchiheub, M., Berret, B., and Braffort, A. (2016). Collecting and analysing a motion-capture corpus of french sign language. In *10th LREC Workshop on the Representation and Processing of Sign Languages: Corpus Mining, ELRA*.

Bishop, C. M. (1994). Neural networks and their applications. *Review of scientific instruments*, 65(6):1803–1832.

Bisong, E. (2019). Tensorflow 2.0 and keras. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, pages 347–399. Springer.

Brock, H., Nishina, S., and Nakadai, K. (2018). To animate or anime-te? investigating sign avatar comprehensibility. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 331–332.

Burleigh, T. L., Stavropoulos, V., Liew, L. W., Adams, B. L., and Griffiths, M. D. (2018). Depression, internet gaming disorder, and the moderating effect of the gamer-avatar relationship: An exploratory longitudinal study. *International Journal of Mental Health and Addiction*, 16(1):102–124.

Chollet, F. et al. (2015). Keras. `https://github.com/fchollet/keras`.

Davenport, T., Guha, A., Grewal, D., and Bressgott, T. (2020). How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science*, 48(1):24–42.

De Martino, J. M., Silva, I. R., Bolognini, C. Z., Costa, P. D. P., Kumada, K. M. O., Coradine, L. C., Brito, P. H. d. S., do Amaral, W. M., Benetti, Â. B., Poeta, E. T., et al. (2017). Signing avatars: making education more inclusive. *Universal Access in the Information Society*, 16(3):793–808.

Efthimiou, E., Fontinea, S.-E., Hanke, T., Glauert, J., Bowden, R., Braffort, A., Collet, C., Maragos, P., and Goudenove, F. (2010). Dicta-sign–sign language recognition, generation and modelling: a research effort with applications in deaf communication. In *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 80–83.

Filhol, M., McDonald, J., and Wolfe, R. (2017). Synthesizing Sign Language by connecting linguistically structured descriptions to a multi-track animation system. In *11th International Conference on Universal Access in Human-Computer Interaction (UAHCI 2017) Held as Part of HCI International 2017*, volume 10278 of *Universal Access in Human–Computer Interaction. Designing Novel Interactions*, Vancouver, Canada, july. Springer.

Gibet, S. (2018). Building french sign language motion capture corpora for signing avatars. In *Workshop on the Representation and Processing of Sign Languages: Involving the Language Community, LREC 2018*.

Hanke, T. (2004). Lexical sign language resources: synergies between empirical work and automatic language generation. In *Fourth International Conference on Language Resources and Evaluation, LREC*.

Jaballah, K. and Jemni, M. (2014). Hand location classification from 3d signing virtual avatars using neural networks. In *International Conference on Computers for Handicapped Persons*, pages 439–445. Springer.

Jancso, A., Rao, X., Graën, J., and Ebling, S. (2016). A web application for geolocalized signs in synthesized swiss german sign language. In *International Conference on Computers Helping People with Special Needs*, pages 438–445. Springer.

Kennaway, R. (2015). Avatar-independent scripting for real-time gesture animation. *arXiv preprint arXiv:1502.02961*.

McDonald, J., Wolfe, R., Schnepp, J., Hochgesang, J., Jamrozik, D. G., Stumbo, M., Berke, L., Bialek, M., and Thomas, F. (2016). An automated technique for real-time production of lifelike animations of american sign language. *Universal Access in the Information Society*, 15(4):551–566.

McDonald, J. C., Wolfe, R., Moncrief, R., Baowidan, S., and Schnepp, J. (July 8-11, 2014). A kinematic model for constructed dialog in american sign language john c. mcdonald1, rosalee wolfe1, robyn moncrief1, souad baowidan1, jerry schnepp2. In *6th Conference of the International Society for Gesture Studies*. San Diego, CA.

Nasihati Gilani, S., Traum, D., Sortino, R., Gallagher, G., Aaron-Lozano, K., Padilla, C., Shapiro, A., Lamberton, J., and Petitto, L.-A. (2019). Can a signing virtual human engage a baby's attention? In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 162–169.

Stoll, S., Camgöz, N. C., Hadfield, S., and Bowden, R. (2018). Sign language production using neural machine translation and generative adversarial networks. In *Proceedings of the 29th British Machine Vision Conference (BMVC 2018)*. University of Surrey.

Vasalou, A., Joinson, A., Bänziger, T., Goldie, P., and Pitt, J. (2008). Avatars in social media: Balancing accuracy, playfulness and embodied messages. *International Journal of Human-Computer Studies*, 66(11):801–811.

Zirzow, N. K. (2015). Signing avatars: Using virtual reality to support students with hearing loss. *Rural Special Education Quarterly*, 34(3):33–36.