Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 101–106
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Sign Language Motion Capture Dataset for Data-driven Synthesis

Pavel Jedlička, Zdeněk Krňoul, Jakub Kanis, Miloš Železný
NTIS - New Technologies for the Information Society,
Faculty of Applied Sciences, University of West Bohemia
Univerzitní 8, 306 14 Pilsen, Czech Republic.
{jedlicka, zdkrnoul, jkanis, zelezny}@ntis.zcu.cz

Abstract

This paper presents a new 3D motion capture dataset of Czech Sign Language (CSE). Its main purpose is to provide the data for further analysis and data-based automatic synthesis of CSE utterances. The content of the data in the given limited domain of weather forecasts was carefully selected by the CSE linguists to provide the necessary utterances needed to produce any new weather forecast. The dataset was recorded using the state-of-the-art motion capture (MoCap) technology to provide the most precise trajectories of the motion. In general, MoCap is a device capable of accurate recording of motion directly in 3D space. The data contains trajectories of body, arms, hands and face markers recorded at once to provide consistent data without the need for the time alignment.

Keywords: Sign Language, Motion Capture, Dataset

1. Introduction

Sign language (SL) is a way of communication that utilizes the movement of a human body. It uses manual, facial, and other body movements to express information. SL is a basic communication system of deaf people and it is often their natural way of communication. According to (Naert et al., 2017), deaf people are often facing problem using written language (based on the spoken language), because it uses the different grammatical rules, and the nature and the spatial organization of linguistic concepts as well. However, most information in the media or the Internet is available in the spoken or the written form. Thus it leads to difficulties for deaf people to access the information.

Computer animation techniques have experienced great improvement recently. There have been developed devices dedicated to the recording of a movement in high precision in 3D space. Animations computed from the data recorded in this way are of high quality and accurate, and their usage is increasingly common outside the film and the computer game industry. An artificial avatar is one possible output of such animation. In public television as an example, they use translation made by a signer which is shown in a window added into the screen. However, the avatar technology is more flexible compared to the real SL signer. It has editable content that can be produced more easily than video (no recording studio with camera) and which also preserves the anonymity of the signer. Using an animated artificial avatar with automatic SL synthesis seems to be a good way to improve the actual way of using CSE on TV.

Recently, some approaches based on key-frame techniques and procedural synthesis have been developed. These approaches provide fine control over the movements of the avatar. These avatars are however poorly accepted by the deaf community because of their lack of human-like motion. There are some works that aim to deal with this problem. In (McDonald et al., 2016) for example, authors added noise measured from MoCap data to the rule-based synthesis to improve the performance of the avatar. Data-driven synthesis, on the other hand, preserves the motion of an original SL signer.

In this paper, we introduce, by our best knowledge, the first MoCap dataset of CSE. This dataset consists of both dictionary items and continuous signing. Manual and non-manual components were recorded simultaneously and the setup includes a high number of markers placed on the face, the body and fingers in order to provide precise and synchronous data. As the main purpose of creating this dataset is to develop an automatic SL synthesis, we also suggest the methods for evaluating the synthesized data.

2. Related Work

Most SL datasets are recorded by an optical camera as they are the most affordable device for this purpose and the recording setup is fast. The difference in data output from the MoCap system and video output is that the MoCap system provides 3D data directly and therefore can be more precise. Although, there are techniques developed for the pose estimation from the image or video, e.g. OpenPose (Cao et al., 2017), the 3D precision is in principle lower than the actual 3D pose measuring provided by the MoCap system.

Some datasets using different motion capture techniques were created in recent years. (Lu and Huenerfauth, 2010) recorded American SL using magnetic-based motion capture for hand and finger tracking. The evolution of motion capture datasets collected in French SL is described in (Gibet, 2018). They recorded three MoCap datasets in the last 15 years. All of them contain manual and non-manual components of SL. The project HuGEx (2005) used Cybergloves for recording finger movements and the Vicon MoCap system for the body and the facial movements. The total recording time was 50 minutes. The next project, SignCom (2011) uses the Vicon MoCap system to record all components and the recording time was 60 minutes, but only 6 markers per hand were used for the hand and finger recording. The most recent project Sign3D (2014) has

all components recorded with the Vicon system and the eye gaze was recorded with a head-mounted oculometer (MocapLab MLab 50-W). It has 10 minutes of recorded data.
There is a continual need for a large amount of data to utilize machine learning techniques. Although the quality and size of datasets are increasing, there is still a lack of such data. The usual size of those datasets is between 10 and 60 minutes of recording time.

3. Dataset Design

Our aim is to record the SL dataset usable for automatic synthesis and evaluation of new utterances. In order to synthesize any given utterance, the language domain was limited to the terms used in the weather forecast. The weather forecast domain was also selected because of the availability of reference video recordings of daily forecasts in SL from a recent couple of years. The size of the vocabulary is reasonably limited for our purposes.
There are some differences in SL expressions depending on the location due to different dialects of CSE, therefore, we used the video source provided by the Czech national television because the used signs are considered as well understandable and recognizable to most of the audience.
CSE linguist experts selected 36 weather forecasts broadcasted throughout the year in order to provide different expressions needed for weather forecasts in different seasons to provide all the necessary data for further synthesis of any weather forecast in the future.

4. Recording Setup

The Motion capture (MoCap) recording is the process of recording the movements using specialized devices in order to reconstruct motions in the 3D space during the time. There are different approaches for data acquisition using MoCap techniques and there are also devices dedicated to the MoCap recording of different body parts. We did some experimental recordings using a different variation of devices such as Cybergloves2 for finger and VICON Cara for facial recording (Krňoul et al., 2016). The main problem with the usage of such devices was signer's discomfort and limitations to performed movements (e.g. tight gloves reduce free movement of fingers, Cara devices camera placement denies finger-face interactions). Another issue was synchronization and calibration (data alignment in general) of different devices as described in (Huenerfauth et al., 2008) and (Krňoul et al., 2016).
Recording all modalities (arm, hand pose, and facial movement) using one device emerged as the best solution. In our solution using an optical-based MoCap system, the signer is equipped with lightweight markers only, and there is no need for merging data together. The only limitation is that the optical-based approach needs a clear line of view from cameras to markers and, therefore, is sensitive to occlusions of body parts. A large number of cameras are needed as well as their precise placement, for such a complex movement like SL utterances.

4.1. Motion Capture Setup

We used the optical-based MoCap system consisting of 18 VICON cameras (8xT-20, 4xT-10, 6xVero) for dataset recording and one RGB camera as referential and two Kinects v2 for additional data acquisition. MoCap recording frequency was 120Hz. The placement of cameras shown in Figure 1 was developed to cover the place in front of the signer in order to avoid occlusions as much as possible and in order to focus on facial expressions. Camera placement was also adjusted for the particular signer to reduce gaps in trajectories caused by occlusions.

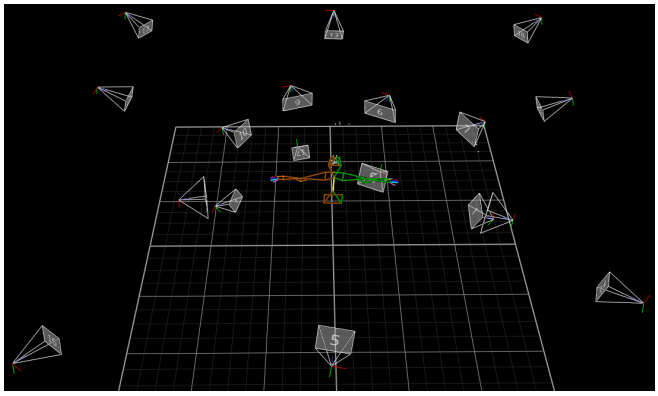

Figure 1: Visualization of MoCap camera layout. View from back and above, the signer is in the middle.

4.2. Subject Setup

The markers placed on the face and fingers were selected to cause minimal disturbance to the signer. We used different marker sizes and shapes for different body parts (see Table 1 and Figure 2). We tracked the upper body and arms by a pair of markers placed on the axis of joints completed by some referential markers. The positions of markers on the face were selected to follow facial muscles and wrinkles. We used 8mm spherical markers around the face, 4 mm hemispherical markers for facial features with the exception of nasolabial folds with 2.5 mm hemispherical markers. The eye gaze and eyelid movement were not tracked by the MoCap device, but it can be obtained from the reference video. Two markers for palm tracking are placed on the index and small finger metacarpals. We tracked fingers using three 4 mm hemispherical markers per finger placed in the middle of each finger phalanx and thumb metacarpals.

	marker diameter [mm]	marker count
Body, arms, hands	8 - 14	33
Fingers	4	30
Face	2.5 - 8	46
Total	2.5 - 14	109

Table 1: Maker sizes and count per segment.

5. Dataset Parameters

We have recorded approximately 30 minutes of continuous signing ($>$ 200000 frames) and 12 minutes of dictionary items. All data were recorded by one expert CSE signer,

Figure 2: Signer marker setup.

who was monitored by another CSE expert during the process. The dataset contains 36 weather forecasts. On average, each such forecast is 30 seconds long and contains 35 glosses. The dictionary contains 318 different glosses. Those dictionary items are single utterances surrounded by the posture with loose hands and arms (a rest pose) in order not to be affected by any context.

Dataset processing is a very demanding work both in terms of time and demands for expert annotation and MoCap data postprocessing. MoCap data have to be processed in order to ensure proper labeling of each marker and to fill eventual gaps in marker trajectories. The next step of MoCap data processing is to solve the marker trajectories (Figure fig:MarkerSetup) to the form of the skeleton model shown in Figure 5. Solving provides data in the angular domain of each body part. Those data can be used directly for the animation.

Another important step in the processing of the dataset is the annotation of content. We used the well-known Elan annotation tool for this purpose, see (Crasborn and Sloetjes, 2008). The reference video of data was used for the annotation as it provides the possibility to annotate the data without need of rendering the MoCap data but it lacks precision because of lower frame-rate (120 fps MoCap vs. 25 fps video). This annotation was made by the CSE native signer. It contains time stamps dividing the data into different signs, transitions between signs and rest pose in one-tier, see Figure 3. The aim of this annotation is to roughly capture those moments of change and it will be used as initialization for a data-driven segmentation/synthesis process. Although annotation is still in progress, almost 80% is already done.

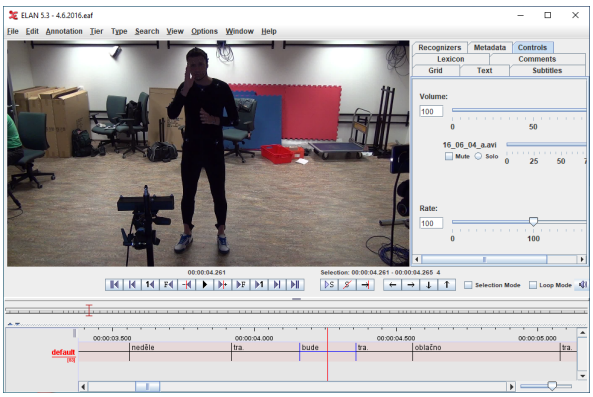

Figure 3: Annotation in ELAN.

6. Data and Synthesis Evaluation

The best way and till now mostly used method for expressing the quality or comparing the similarity of two signs is using subjective evaluating by SL native signers. However, this evaluation is both time and human resources demanding process and moreover usually more than one person is needed for the subjectivity of the evaluation, see (Huenerfauth et al., 2008).

The popularity of automatic and machine learning techniques utilization for data-processing related tasks increased in recent years. An objective criterion in the form of a cost function is crucial for such techniques but it is usually not trivial to choose one. The purpose of such a function is not to replace the human evaluation of the synthesis result, but to provide a proper cost function for machine learning techniques as they need fast evaluation during training process.

The data provided by the MoCap recording are trajectories of all markers. The advantage of such data is direct information of the positions in the 3D space but the human body topology (skeleton) may not be respected in such representation. On the other hand, angular trajectories of bones are bound to the exact human body topology. The topology of a signer is constant during the time. This can improve the consistency of the data if signs from single signer are compared. In both cases, one frame can be considered as a vector of values and the duration of two similar utterances can differ, although the meaning is the same. The signs and utterances are the time-sequences of these vectors.

The usual metrics (among the others) for evaluating difference/similarity between two single vectors $p = (p_0, p_1, ..., p_i, ...p_N)$ and $q = (q_0, q_1, ..., q_i, ...q_N)$ of the same length N are:

- Euclidean distance:

$$d = \sqrt{\sum_{i=0}^{N} (q_i - p_i)^2}, \qquad (1)$$

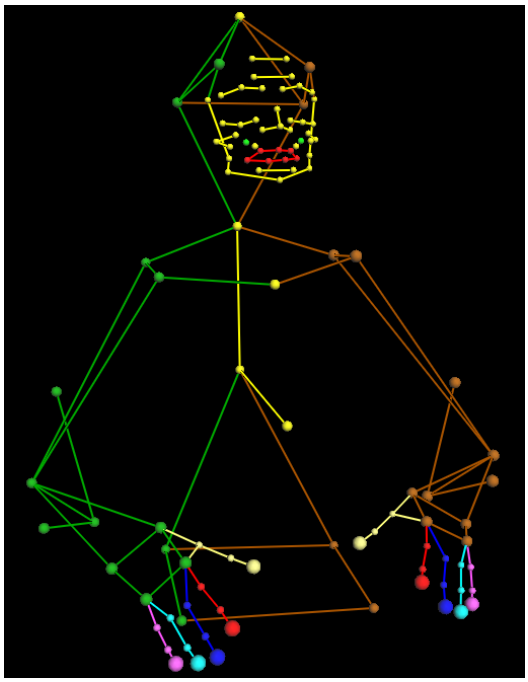

Figure 4: Marker setup (data visualization).

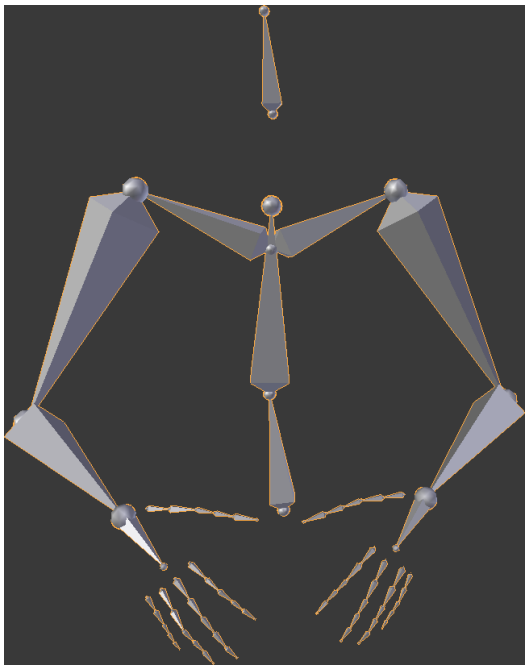

Figure 5: Model visualization.

- Root mean square error (RMSE):

$$d = \sqrt{\frac{\sum_{i=0}^{N}(p_i - q_i)^2}{N}}, \tag{2}$$

- Correlation coefficients (Corr):

$$d = \frac{\sum_{i=0}^{N}(p_i - \overline{p})(q_i - \overline{q})}{\sqrt{\sum_{i=0}^{N}(p_i - \overline{p})^2 \sum_{i=0}^{N}(q_i - \overline{q})^2}}, \tag{3}$$

where $\overline{p}$ and $\overline{q}$ are mean values of p and q respectively.

The time component of the data (the time-sequence of the vectors) can be addressed by the following approaches. One of them is a time alignment in the form of re-sampling the time-sequence of two compared components to the same length and then measure the distance. In (Sedmidubsky et al., 2018) they used normalization for motion data comparison for query purposes in the form of the time axis movement sequence normalization and Euclidean distance for each motion.
Dynamic time warping (Berndt and Clifford, 1994) (DTW) is commonly used algorithm for the time-series comparison. This method computes the best per frame alignment in terms of the chosen distance. It provides us a possibility to get minimal distance of two time-sequence with different lengths, for example two utterances with different signing pace. The computed DTW distance d_DTW is a minimal distance with the optimal time alignment of sequences p and q, $path$ describes the alignment of the vectors:

$$d_{DTW}, path = DTW(p, q). \tag{4}$$

We tested the DTW algorithm with the Euclidean distance (1) for measuring a distance between two different signs and between different instances of the same sign. We limited this test for the signs with meanings ”one”, ”two”, ”three”, ”four”, and ”five”, both from the dictionary and the continuous signing and compared measured distances between signs with the same meaning (different instance of the same sign) and different signs (all instances of other signs from the same test-set). The DTW distance was measured between two signs, the distance was normalized to the vector size and the length of the DTW path, so the distance is independent on the skeleton complexity and duration of the sequence. The normalized DTW $d_{normDTW}$ distance is defined as:

$$d_{normDTW} = \frac{d_{DTW}}{M \cdot N}, \tag{5}$$

where M is the length of the $path$ from DTW algorithm and N is the number of channels of the data.

Sign	distances [deg] (same meaning)	distances [deg] (different meaning)
”one”	0.84 - 1.79	2.49 - 8.67
”two”	0.45 - 1.29	2.49 - 7.08
”three”	1.18	2.54 - 5.80
”four”	0.33 - 0.85	3.24 - 8.67
”five”	0.33 - 0.85	2.49 - 7.78

Table 2: Normalized DTW distances between signs (handshapes only).

The Euclidean distances of angular trajectories computed using DTW are summarized in Tables 2 and 3 for handshape only and for the whole body (hand included) respectively. The tested signs (numbers from 1 to 5) were chosen because they are very similar and differs only in the handshape. The signs are compared to other instances with the same meaning and to all instances of all different signs (e.g.

Sign	distances [deg] (same meaning)	distances [deg] (different meaning)
"one"	2.30 - 3.25	3.08 - 6.90
"two"	1.04 - 3.59	2.74 - 6.37
"three"	2.58	2.94 - 5.42
"four"	0.89 - 3.28	2.73 - 6.90
"five"	1.10 - 2.07	3.13 - 5.57

Table 3: Normalized DTW distances between signs (whole body without face).

all instances with the meaning "one" are compared to all other instances with the same meaning and to all instances with different meanings such as "two", "three", ...). According to the results in Table 3, using normalized DTW distance for raw trajectories of the angular representation seems to have the ability to objectively measure the difference between signs, because the distance is generally lower for the signs with the same meaning than others.
In case of the hand-shapes (Table 2, there seems to be the ability to not only measure the distances between signs with the same meanings but also to distinct different signs completely.
We suggest some approaches to improve the evaluation of distances calculated by DTW. We can use different weights for the distance measure for different bones based on its corresponding importance for the signs distinction. We can also use trajectories of different body parts to compare signs components separately. For example, compare hand-shapes, palm orientation and location with their counterparts respectively to enable more precise modeling of SL grammar such as classifiers, the co-occurrence of manual and non-manual, etc.

7. Experiments

7.1. Methods

We propose the following baseline technique for the SL utterance synthesis. The purpose of this baseline is not to solve the synthesis problem itself but to provide a reference algorithm and performance for further developed and more sophisticated techniques. We assemble the utterance from dictionary item trajectories for each sign. Then we compute trajectories of transition movement between these signs. We set the fixed length for all transitions as the average length of all transitions in our dataset. We interpolated the transition trajectory for each joint by the cubic spline. For evaluation, we compared the synthesized utterance with the utterance captured in the continuous signing by the normalized DTW with Euclidean distance.

7.2. Results

We selected a pair of utterances that have more appearances in the dataset in order to provide a comparison with a reference.

- Utterance 1: "zima-hory-kolem" (literal translation: cold-hills-approximately). Confusion matrix is shown in Table 4
- Utterance 2: "pocasi-zitra-bude" (literal translation: weather-tomorrow-will be). Confusion matrix is shown in Table 5

In confusion matrices (Tables 4 and 5), we can see the normalized DTW distances of the synthesized utterance compared to utterances with the same meaning that appear in continuous signing. For reference, we added a comparison with the utterance with other meaning.

	synth	*appear1*	*appear2*	*appear3*	*other*
synth	0	2.58	2.69	2.82	5.28
appear1	2.58	0	1.03	1.27	6.14
appear2	2.69	1.03	0	1.41	6.19
appear3	2.82	1.27	1.41	0	6.62
other	5.28	6.14	6.19	6.62	0

Table 4: Confusion matrix of normalized DTW distances for utterance 1. Synthesised data (*synth*), compared with real data (*appear1-3*) and *other* utterance with different meaning.

	synth	*appear1*	*appear2*	*appear3*	*other*
synth	0	1.51	1.43	1.61	5.28
appear1	1.51	0	0.62	0.71	4.69
appear2	1.43	0.62	0	0.82	4.84
appear3	1.61	0.71	0.82	0	4.60
other	5.28	4.69	4.84	4.60	0

Table 5: Confusion matrix of normalized DTW distances for utterance 2. Synthesised data (*synth*), compared with real data (*appear1-3*) and *other* utterance with different meaning.

The comparison of the normalized DTW distances shows larger differences between synthesized utterance and examples from continuous data then among the continuous data. We can also distinct different utterances from each other. The difference between synthesized data and examples from continuous data can be caused by various reasons. We try to explain some of those in the following discussion.

8. Discussion

There is a difference in the pacing and the method of signing for signs in the dictionary and the same signs in the continuous signing. On average, the dictionary signs are more than twice longer than signs from continuous signing. The average duration of signs in our dataset is 0.81/0.38 seconds in dictionary/continuous signing. There are also differences in signs that consist of repetitive moves. Usually, more repetitions are made in dictionary items than in continuous signing. Those differences are insignificant in human understanding of the sign but enlarge the measured distance.
The transitions are synthesized with a constant length and such an approximation does not correspond with the observed reality. The cubic spline interpolation is also heavily dependant on the annotation's precise selection of the start

and the end point and also does not respect the nature of the human movement.

9. Conclusion

We presented a new 3D motion capture dataset of Czech Sign Language (CSE), which we would like to share with the community. Its main purpose is to provide the data for further analysis and data-based automatic synthesis of CSE utterances. The dataset was recorded using the state-of-the-art motion capture technology to provide the most precise trajectories of the motion. The size of the dataset and the precision of tracked components are comparable to the best existing datasets for other SLs. The dataset contains trajectories of body, arms, hands and face markers recorded at once in order to provide consistent data without the need for the time alignment.

We introduced a baseline for the data-driven synthesis of SL utterances and suggested a method for objective data evaluation in the form of normalized DTW algorithm and Euclidean distance.

In future work, we will focus on improving the quality of the synthesis by using machine learning techniques and the normalized DTW distance as an objective function. We would also like to verify the correlation between objective and subjective evaluations.

We also would like to further improve synthesis by adding a non-manual property as well as other more complex SL grammar concepts. This will require annotations in more than one-tier. The additional annotation can be done in semi-automatic or fully automatic mode. It will also be beneficial to use multiple annotators on the same task to eliminate human errors and improve the precision of an annotation.

10. Acknowledgement

This work was supported by the Ministry of Education of the Czech Republic, project No. LTARF18017. This paper was supported by the Ministry of Education, Youth and Sports of the Czech Republic project No. LO1506. This work was supported by the European Regional Development Fund under the project AI&Reasoning (reg. no. CZ.02.1.01/0.0/0.0/15 003/0000466).

11. Bibliographical References

Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA.

Cao, Z., Simon, T., Wei, S., and Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1302–1310.

Crasborn, O. and Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. In *6th International Conference on Language Resources and Evaluation (LREC 2008) 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, pages 39–43.

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*, Miyazaki, Japan, May.

Huenerfauth, M., Zhao, L., Gu, E., and Allbeck, J. (2008). Evaluation of american sign language generation by native asl signers. *ACM Transactions on Accessible Computing (TACCESS)*, 1(1):1–27.

Krňoul, Z., Kanis, J., Železnỳ, M., and Müller, L. (2016). Semiautomatic data glove calibration for sign language corpora building. In *7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining, LREC*.

Krňoul, Z., Jedlička, P., Kanis, J., and Železný, M. (2016). Toward sign language motion capture dataset building. In *Speech and Computer*, pages 706–713, Cham, 08. Springer International Publishing.

Lu, P. and Huenerfauth, M. (2010). Collecting a motion-capture corpus of american sign language for data-driven generation research. In *Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies*, pages 89–97. Association for Computational Linguistics.

McDonald, J., Wolfe, R., Wilbur, R. B., Moncrief, R., Malaia, E., Fujimoto, S., Baowidan, S., and Stec, J. (2016). A new tool to facilitate prosodic analysis of motion capture data and a data- driven technique for the improvement of avatar motion.

Naert, L., Larboulette, C., and Gibet, S. (2017). Coarticulation analysis for sign language synthesis. In Margherita Antona et al., editors, *Universal Access in Human–Computer Interaction. Designing Novel Interactions*, pages 55–75, Cham. Springer International Publishing.

Sedmidubsky, J., Elias, P., and Zezula, P. (2018). Effective and efficient similarity searching in motion capture data. *Multimedia Tools Appl.*, 77(10):12073–12094, May.