Towards Realistic Sign Language Animations

Inês Lacerda
ines.lacerda@tecnico.ulisboa.pt
Instituto Superior Técnico, Universidade de Lisboa
Lisbon, Portugal

Hugo Nicolau
hugo.nicolau@tecnico.ulisboa.pt
ITI, LARSYS/Instituto Superior Técnico, Universidade de Lisboa
Lisbon, Portugal

Luísa Coheur
luisa.Coheur@tecnico.ulisboa.pt
INESC-ID/Instituto Superior Técnico, Universidade de Lisboa
Lisbon, Portugal

ABSTRACT

Current signing avatars are often described as unnatural as they cannot accurately reproduce all the subtleties of synchronized body behaviors of a human signer. In this paper, we investigate a new dynamic approach for transitions between signs and the effect of mouthing behaviors. Although native signers preferred animations with dynamic transitions, we did not find significant differences in comprehension and perceived naturalness scores. On the other hand, we show that including mouthing behaviors improved comprehension and perceived naturalness for novice Portuguese sign language learners.

CCS CONCEPTS

• **Human-centered computing → Interactive systems and tools**.

KEYWORDS

Portuguese Sign Language, Synthetic Animation, Computational Linguistics, Natural Language Processing

ACM Reference Format:
Inês Lacerda, Hugo Nicolau, and Luísa Coheur. 2023. Towards Realistic Sign Language Animations. In *ACM International Conference on Intelligent Virtual Agents (IVA '23), September 19–22, 2023, Würzburg, Germany*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3570945.3607354

1 INTRODUCTION

Sign language translators typically require two components: a language translator and a signing avatar. The translator converts written text (or speech) into a sequence of glosses (i.e., lexical units representing each gesture or sign); then, the avatar displays the synthesized glosses and additional linguistic components as signing animations. Planning and scripting a signing avatar's facial and body movements to correctly sign is a difficult task. Minor variations in timing and speed parameters can lead to significant differences in the quality and understandability of sign animations [1, 6]. In the case of sign languages, transitions between signs rely heavily on the phonology of the previous and following signs and determine the movement fluidity that allows sign streams to be intelligible. Therefore, transitions will impact the comprehension and naturalness of sign animations. In this paper, we introduce a new approach for interpolating signs consisting of **dynamic transitions**, which change according to the previous and following signs. We aimed to answer the following research question: *Do dynamic transitions improve linguistic comprehension, naturalness, and preference of sign language animations?* We use an existing text-to-sign language translator [4, 5, 8] to evaluate our animations. Although participants preferred the avatar with dynamic transitions, we did not find significant differences in comprehension, naturalness, and preference. However, dynamic transitions show greater potential for signs that comprise one sole meaning (i.e., composite utterances and negatives) and require faster transitions.

In addition to the previous study, we also investigated the effect of **mouthing** – the production of visual morphemes or syllables that derive from spoken language. Thus, our second research question was: *Does mouthing impact linguistic comprehension, naturalness, and preference of sign language animations?* Figure 1 illustrates an avatar with and without mouthing.

Figure 1: Avatar with (left) and without mouthing (right).

Results show that mouthing improves comprehension and perceived naturalness for novice sign language learners. To the best of our knowledge, research in the field has not yet been published on whether mouthing can improve comprehension. An extended version of this paper can be found in [7].

2 SYNTHESIS

2.1 Dynamic transitions

Since transitions between signs rely heavily on the phonology of the previous and following signs, we propose dynamic transitions, which interpolate signs through transitions that change according to the adjacent signs. In this section, we describe the dynamic transitions algorithm.

We iterate over each gloss in run-time, calculating the differences between hand positions in the last keyframe of the previous sign and the first keyframe of the following sign. Then, the squared magnitude of these vectors is computed. These squared magnitude values are then converted to percentages by defining a scale. To decide this scale, we checked all signs in our database and computed the smallest hand position differences between two signs

IVA '23, September 19–22, 2023, Würzburg, Germany

ACM ISBN 978-1-4503-9994-4/23/09.
https://doi.org/10.1145/3570945.3607354

(e.g., signs "EU" and "TER") and the largest difference (e.g., signs "ELE" and "TER"). Based on our findings, we defined two scales: one that includes both hands and one hand. Using these scales, the squared magnitude values are converted to percentages that range between 0% and 100%. To find the duration value used in the transition between signs, we use the percentage calculated to linearly interpolate between two duration values, which correspond to the lowest and highest values that the duration of transitions can take. We defined these values by analyzing the lowest and highest transition duration in multiple videos of an Portuguese Sign Language (LGP) corpus[1]. Furthermore, two empirical studies developed by Sedeeq [2, 3] found that ASL signers prefer slower transitions than the timing of human signers and that they prefer animations with an average transition time of 0.5 seconds. Based on the analysis of our LGP corpus, we decided that the duration of transitions would range between 0.3 seconds and 1.1 seconds because this range would include 0.5 seconds as the average transition time, and these are slightly slower than the human signing transitions in our LGP corpus.

Using the calculated duration values, the algorithm creates an interpolation between the current sign and the next sign using dynamic transitions by defining a duration value and an offset value. The first keyframe of every sign in the database starts at 1 second. Using the offset value, we can adjust the timing until the first keyframe matches the transition duration time; therefore, the offset value is 1 second minus the transition duration value. To create more fluid transitions, we defined the offset value as 1.2 seconds minus the transition value.

2.2 Mouthing

We extended the existing translation system to gather all words in Portuguese and, afterwards, combine them into a sentence so that we consider the assimilation between words when executing the phonetic transcription. The phonetic transcription is done by employing the phonemizer tool[2], where the speak backend is used to produce phoneme sequences described based on the International Phonetic Alphabet transcription. After this, normalization is done by encoding non-ASCII to ASCII, words are separated into their corresponding syllables using syllabification rules, and then, each phoneme is mapped into one viseme using the phoneme-viseme mapping we created. Animations for each viseme were created by adjusting the weights of blend shapes. In the translation process, words are translated into phonemes, separated into syllables, and mapped into visemes. In the animation process, mouthing is animated by using an interpolation scheme that concatenates the visemes according to the animated signs. The duration value for the mouthing is defined based on the duration of the sign it is applied to and based on its number of syllables.

3 EVALUATION

3.1 Evaluating Dynamic Transitions

We recruited 11 participants fluent in LGP. Participants had to fill in a questionnaire. In the first 10 sections of the questionnaire, participants had to visualize a one-sentence video, write what they understood about the video, and describe whether the sentence contained an error. The created sentences contained one or more composite utterances. For each sentence, participants also had to evaluate the transitions' speed on a 5-point Likert scale, with one being too slow and five as too fast, and the avatar's naturalness on a 5-point Likert scale, with one being robotic and five as natural. We presented both conditions - dynamic and constant transitions - to all participants in a counterbalanced order. The order of the ten videos was randomized. In the next three sections, participants had to select which video they preferred between two side-by-side videos (one with dynamic transitions and one with constant transitions of 0.5 seconds). We also asked open-ended questions about the avatars' naturalness, and whether transitions between signs affect naturalness and comprehension.

3.1.1 Comprehension. For each sentence in the questionnaire, we measured the percentage of content understood by calculating the number of glosses correctly described with 100% as all glosses correctly understood by a participant (done manually as synonyms of signs also counted as correct). Overall, the **average comprehension scores for all participants with both conditions was 81.56% (**$SD = 23.29$**)**. Seven participants had higher comprehension results in sentences with dynamic transitions, 3 participants had higher comprehension results with constant transitions and one had equal comprehension results in both transitions. According to a Shapiro-Wilk test, we retained the null hypothesis of population normality ($p = 0.901, p = 0.722$); therefore, we conducted a Paired samples T-test to compare differences in comprehension scores between our conditions. Based on the results, **there was no significant difference** ($t(10) = -1.379, p = 0.198$) between dynamic ($M = 82.97, SD = 9.43$) and constant transitions ($M = 80.15, SD = 11.55$).

In almost all cases, participants would either understand a sign, independently of the transition approach, which could be explained by the fact that the difference between transition values of both approaches was not significant. However, there were four cases in two-paired sentences (i.e., eight sentences) where the same sign was only perceived correctly with the dynamic approach. Moreover, there were no cases where a sign was only perceived correctly with a constant approach. Furthermore, seven participants believed transitions between signs indeed impact comprehension, whereas only four believed they do not.

3.1.2 Naturalness. For each sentence in the questionnaires, we measured the percentage of naturalness by using the scores submitted on the Likert scale (i.e., one as robotic and five as natural), with 5 being 100%. Overall, the **average naturalness scores for all participants with both conditions was 50.73% (**$SD = 22.78$**)** and the average overall naturalness given at the end of the questionnaire by all participants was 50.91% ($SD = 25.87$). The scores for naturalness were significantly lower than those for the other two measures, which is unsurprising because **naturalness is the most demanding criterion of all**. There were **large discrepancies between naturalness scores throughout our participants** with 20% as the lowest average score and 100% as the highest score. Furthermore, three participants had higher naturalness results in sentences with dynamic transitions, two had higher naturalness results with constant transitions, and six had equal naturalness results in both

[1] https://portallgp.ics.lisboa.ucp.pt/corpus_lgp/
[2] https://github.com/bootphon/phonemizer

conditions. According to a Shapiro-Wilk test, we retained the null hypothesis of population normality ($p = 0.548, p = 0.215$); therefore, we conducted a Paired samples T-test to compare differences in naturalness scores between our conditions. Based on the results, **there was no significant difference** ($t(10) = -0.820, p = 0.432$) between dynamic ($M = 51.27, SD = 22.61$) and constant transitions ($M = 50.18, SD = 22.51$). However, seven participants believed transitions between signs impact naturalness, whereas only four believed they do not.

3.1.3 Preference. We conducted a Chi-Square test to analyze which condition was preferred by participants. We found a **statistically significant difference between transition appriach** ($X^2(1, N = 33) = 6.818, p = .009$), as participants **preferred the dynamic** ($N = 24$) over the constant transitions ($N = 9$).

3.1.4 Transitions Speed. For each sentence in the questionnaires, we measured the percentage of optimal transition speed by using the scores submitted on a 5-point Likert scale (i.e., one as too slow and five as too fast), with three being the optimal speed. We transformed the ordinal data to a 0-100% measure. Overall, the average optimal transition speed scores for all participants in both conditions were 83.64% ($SD = 17.36$), and the average **overall quality of transitions** given at the end of the questionnaire by all participants was 81.82% ($SD = 17.41$). Three participants had higher optimal transition speed results in sentences with dynamic transitions, three participants had higher optimal transition speed results with constant transitions, and five participants had equal optimal transition speed results in both transitions. According to a Shapiro-Wilk test, we retained the null hypothesis of population normality ($p = 0.283, p = 0.064$); therefore, we conducted a Paired samples T-test to compare differences in optimal transition speed scores between our conditions. Based on the results, **there was no significant difference** ($t(10) = -0.319, p = 0.756$) between dynamic ($M = 83.64, SD = 11.68$) and constant transitions ($M = 83.032, SD = 13.45$). However, three participants commented on the importance of faster transitions between signs with one sole meaning. They noted that constant transitions were too slow for composite utterances and, surprisingly, in negatives.

3.2 Mouthing Evaluation

We recruited 20 participants that were learning LGP. Again, we recurred to a questionnaire with thirteen sentences. For this user study, we removed all phonological facial expressions from signs, so that all signs could execute mouthing. The protocol used in this study was similar to the previous one. We also evaluated general quality, signs' quality, and facial expressions' quality using 5-point Likert scales. Additionally, we also asked participants open-ended questions about mouthing and its effect on naturalness and comprehension.

3.2.1 Comprehension. Overall, the **average comprehension scores for all participants with both conditions was 70.94% (**$SD = 37.88$**)** which we found surprisingly high considering that participants were beginners and sentences had a level of complexity and difficulty higher than beginner level with some sentences composed by interrogatives, one composite utterance (i.e., sign "IRMÃ") and dactylology words comprised of numbers with two digits and names with seven letters. There were large discrepancies between comprehension scores among our participants, with 33.33% as the lowest average score and 100% as the highest score. Furthermore, 10 participants had higher comprehension results in sentences with mouthing, three had higher comprehension results without mouthing, and seven had equal comprehension results in both. According to a Shapiro-Wilk test, we rejected the null hypothesis of population normality ($p = 0.012, p = 0.050$); therefore, we conducted a Wilcoxon signed-rank test to compare differences in comprehension scores between our conditions. Based on these results, the **comprehension scores for sentences with mouthing were statistically significantly higher** than for sentences without mouthing ($Z = -2.029, p = 0.043$). Furthermore, 16 participants believed mouthing does indeed have an impact on comprehension, whereas only four participants believed it does not. Additionally, many comments were made by participants noting that mouthing makes it easier to understand the sentences.

3.2.2 Naturalness. Overall, the **average naturalness scores for all participants with both conditions was 78.29% (**$SD = 16.91$**)** and the average overall naturalness given at the end of the questionnaire by all participants was 78.95% ($SD = 15.60$). Eleven participants had higher naturalness results in sentences with mouthing, five participants had higher naturalness results without mouthing, and four participants had equal naturalness results in both. According to a Shapiro-Wilk test, we retained the null hypothesis of population normality ($p = 0.160, p = 0.793$); thus, we conducted a Paired samples T-test to compare differences in naturalness scores between our conditions. Based on the results, the **naturalness scores for sentences with mouthing** ($M = 80.40, SD = 15.24$) **were statistically significantly higher** ($t(19) = -2.094, p = 0.050$) than for sentences without mouthing ($M = 76.10, SD = 13.11$). Furthermore, 18 participants believed mouthing impacts naturalness, whereas only two believed it does not.

3.2.3 Preference. We conducted a Chi-Square test to analyze which animations were preferred on the three trials each participant had (60 trials overall). There was a **statistically significant between conditions** ($X^2(1, N = 60) = 15, p = 0.000108$), as participants **preferred animations with mouthing** ($N = 45$) rather than without ($N = 15$).

4 CONCLUSIONS AND FUTURE WORK

We introduced dynamic transitions and added to the translator's avatar the possibility of performing mouthing. The positive results indicate that the generated animations show great potential in the synthetic animation of signing avatars. However, future research should improve and extend this work by, for example, adding more facial expressions, corporal movements, appropriate pauses, and accelerations between signs.

ACKNOWLEDGEMENTS

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT), projects UIDB/50021/2020, UIDB/50009/2020, and 2022.06596.PTDC, and Plano de Recuperação e Resiliência (PRR), Center for Responsible AI C645008882-00000055.

REFERENCES

[1] Sedeeq Al-khazraji, Larwan Berke, Sushant Kafle, Peter Yeung, and Matt Huenerfauth. 2018. Modeling the speed and timing of American Sign Language to generate realistic animations. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*. Association for Computing Machinery, New York, NY, USA, 259–270.

[2] Sedeeq Al-khazraji, Becca Dingman, and Matt Huenerfauth. 2020. Empirical Investigation of Users' Preferred Timing Parameters for American Sign Language Animations. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–7.

[3] Sedeeq Al-khazraji, Becca Dingman, Sooyeon Lee, and Matt Huenerfauth. 2021. At a Different Pace: Evaluating Whether Users Prefer Timing Parameters in American Sign Language Animations to Differ from Human Signers' Timing, In Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility. *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'21)*, Article 40, 12 pages.

[4] Pedro Cabral, Matilde Gonçalves, Hugo Nicolau, Luísa Coheur, and Ruben Santos. 2020. PE2LGP Animator: A Tool To Animate A Portuguese Sign Language Avatar. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*. European Language Resources Association (ELRA), Marseille, France, 33–38. https://aclanthology.org/2020.signlang-1.6

[5] Matilde Gonçalves, Luisa Coheur, Hugo Nicolau, and Ana Mineiro. 2021. PE2LGP: tradutor de português europeu para língua gestual portuguesa em glosas. *Linguamática* 13, 1 (Jul. 2021), 3–21. https://doi.org/10.21814/lm.13.1.338

[6] Matt Huenerfauth. 2009. A linguistically motivated model for speed and pausing in animations of american sign language. *ACM Transactions on Accessible Computing (TACCESS)* 2, 2 (2009), 1–31.

[7] Inês Lacerda, Hugo Nicolau, and Luisa Coheur. 2023. Enhancing Portuguese Sign Language Animation with Dynamic Timing and Mouthing. arXiv:2307.06124 [cs.CL]

[8] Carolina Neves, Luísa Coheur, and Hugo Nicolau. 2020. HamNoSyS2SiGML: Translating HamNoSys Into SiGML. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 6035–6039. https://aclanthology.org/2020.lrec-1.739