Comparative Analysis of Sign Language Interpreting Agents Perception: A Study of the Deaf

Alfarabi Imashev[*], **Nurziya Oralbayeva**[†], **Gulmira Baizhanova**[◇], **Anara Sandygulova**[*]
[*] Department of Robotics Engineering, School of Engineering and Digital Sciences, Nazarbayev University
[†] Graduate School of Education, Nazarbayev University
{alfarabi.imashev, nurziya.oralbayeva, anara.sandygulova}@nu.edu.kz
[◇] Public Association "Association of Deaf People "JasNur"
deafjasnur@mail.ru

Abstract

Prior research on sign language recognition has already demonstrated encouraging outcomes in achieving highly accurate and dependable automatic sign language recognition. The use of virtual characters as virtual assistants has significantly increased in the past decade. However, the progress in sign language generation and output that closely resembles physiologically believable human motions is still in its early stages. This assertion explains the lack of progress in virtual intelligent signing generative systems. Aside from the development of signing systems, scholarly research have revealed a significant deficiency in evaluating sign language generation systems by those who are deaf and use sign language. This paper presents the findings of a user study conducted with deaf signers. The study is aimed at comparing a state-of-the-art sign language generation system with a skilled sign language interpreter. The study focused on testing established metrics to gain insights into usability of such metrics for deaf signers and how deaf signers perceive signing agents.

Keywords: sign language, agents, data-driven, subjective evaluation, HCI

1. Introduction

Deaf populations worldwide use sign languages, which are already recognized as developed, comprehensive, and fully fledged natural languages. Like spoken languages, these languages vary in linguistic aspects like word order rules, grammar, lexicon, semantics, morphology, and pragmatics (Pfau et al., 2012). The importance of the simultaneous processing of both manual and non-manual components is of the utmost importance in sign languages, as shown in Pfau and Quer (2010).

In the past 15 years, there has been a burgeoning interest in computer-based systems that generate sign language. This can be explained by the enhanced adaptability of such systems and their capacity to generate various sequences of signs from a collection of individual signs as opposed to using short video clips with human interpreters (Delorme et al., 2009).

Intelligent virtual assistants (IVAs) are becoming more prevalent in daily life, enhancing interactions between humans and technological devices. Intelligent Virtual Assistants (IVAs) are programmed to imitate human behavior and currently offer advantages to individuals who are deaf or use sign language. They achieve this by converting written language into sign language and vice versa (Brock and Nakadai, 2019). As a result, IVAs assist in eliminating barriers in communication (Bragg et al., 2019).

According to the latest research Bragg et al. (2019), Duan et al. (2019), the effort of designing and constructing avatars that are appropriate for deaf individuals while also being intuitive, user-friendly, linguistically and semantically accurate, and easily understandable is a challenging task.

Undoubtedly, the involvement of deaf individuals who utilize sign language is essential when evaluating sign language generation systems. Their viewpoints and observations can aid in detecting flaws in the experimental design (Huenerfauth and Kacorri, 2015; Roelofsen et al., 2021) and indicating areas that necessitate additional refinement (Kipp et al., 2011; Gibet et al., 2011; Schnepp et al., 2012; Roelofsen et al., 2021).

However, conducting user assessment studies with traditional written text questionnaires has been found to be inappropriate for deaf signers. This raises concerns about the reliability of the data obtained from such surveys (Bosch-Baliarda et al., 2019), due to various evident factors. First of all, despite their enough proficiency in multiple languages, individuals primarily communicate using sign language. Consequently, it is not optimal to administer a questionnaire in a written language, as is commonly practiced (Farwell, 1976), so it is imperative to establish effective means of communication with the deaf community in their native language (Gibet et al., 2011; Bosch-Baliarda et al., 2019).

Secondly, scholarly investigations indicate that deaf individuals encounter heightened challenges

LREC-COLING 2024, pages 3603–3609

in effectively comprehending ambiguous and intricate concepts quite often (Parvez et al., 2019). This, in turn, contributes to an increased cognitive burdens (Bosch-Baliarda et al., 2019).

In contrast, a sufficient body of research already exists to provide methodological remedies for incorporating deaf signers into the assessment of signing avatars. The aforementioned research examined a range of demographic factors (Kacorri et al., 2017), utilized modified questions and stimuli to evaluate the linguistic elements of avatars (Huenerfauth and Kacorri, 2014), and explored several modalities, including short videos (Kacorri et al., 2013).

This study endeavors to provide a comprehensive account of a pilot user study that evaluated the reliability of assessment measures for the evaluation of sign language avatars. The study specifically focused on deaf signers and adapted with the Godspeed questionnaire (Bartneck et al., 2009).

2. Agents

The deaf community in Kazakhstan has integrated creative and distinct concepts into sign language, encompassing a diverse array of indigenous musical instruments, culinary delights, renowned landmarks, notable personalities, traditional customs, and more. However, the sign language used in Kazakhstan is not indigenous and has close connections with many other sign languages in the Commonwealth of Independent States (CIS), as they originated from the signing system developed in the USSR.

Consequently, individuals from the Commonwealth of Independent States (CIS) and neighboring regions may rely on avatars specifically designed for K-RSL. In this study, we employed human and stick figure agents to perform sentences in the local Kazakh-Russian Sign Language (referred to as K-RSL).

2.1. Hand scripted SignMT - Agent 1

The SignMT project (Moryossef et al., 2021) is chosen as Condition 1. This technology is state-of-the-art, highly distinctive, and inventive. The project capabilities range is wide and includes Sign Language Detection, Identification, Segmentation, Recognition, Translation, and Generation.

Based on our understanding, the text-to-sign process involves converting text into a "SignWriting" scheme, resulting in a pose sequence, similar to Papadogiorgaki et al. (2004), Moemedi and Connan (2010). Subsequently, the pose sequence is utilized as the input for the rendering engine, namely SkeletonViewer. SignMT has three output options: a stick figure, a 3D avatar, and a model, generating a realistic human avatar video (Human GAN).

For the current experiment, we utilized a stick-figure agent as one of multiple avatar options to facilitate translation between written and sign languages. It can be achieved through an interface that bears a resemblance to Google Translate. The software already has built-in functionality to accommodate multiple languages. Given that K-RSL and RSL both derive from the same signing system. For our experiment, we generated phrases for SignMT to perform with signs, which possess identical meanings in both languages. Therefore, the study may be readily replicated in several languages.

One additional benefit of this project is its active and responsive community, which ensures regular updates and the addition of other languages.

2.2. Human Signer - Agent 2

A person proficient in interpreting local sign language has been recruited as a human agent. We selected this interpreter for multiple reasons. Firstly, she is a Child of Deaf Adults (CODA), which means she has grown up in a deaf community and is familiar with its customs and culture. Additionally, she holds a bachelor's degree in defectology and has accumulated 7 years of professional experience as a television news interpreter.

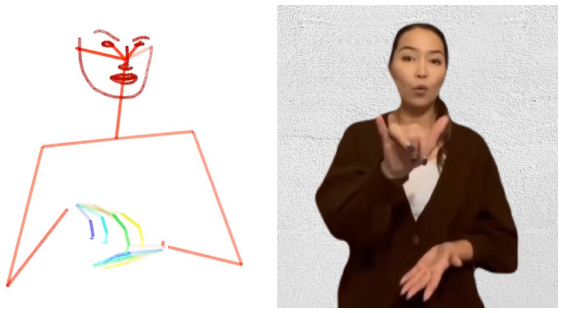

Figure 1: SignMT generated stick figure avatar, and a proficient human interpreter.

3. User Study

The Nazarbayev University Institutional Research Ethics Committee (NU-IREC) granted approval for this research. The consent form, instructions, questions, and tasks were translated into K-RSL, recorded as brief films, and given to participants throughout the experiment. Throughout the entire user study, a sign language interpreter was also in attendance. Participants were remunerated for the duration of their involvement.

3.1. Participants

The study included 12 participants - deaf signers, with ages ranging from 27 to 53 years (mean age = 36.3, six males and six females). Please refer to Table 1 for additional information. All individuals were currently resident in Astana during the study.

Gender	Age	Education
M	31	9th grade
M	53	11th grade
F	36	College
F	27	College
M	36	Bachelor
F	33	College
F	34	College
M	44	9th grade
F	44	9th grade
M	37	11th grade
F	28	9th grade
M	33	11th grade

Table 1: Participants' demographics.

3.2. Stimuli

We composed three sentences in K-RSL for each agent. The sentences were evenly distributed based on the handshapes utilized and the complexity of sentence structure and concepts involved in each sentence (refer to Table 2).

Stick Figure Agent (SignMT)
1. BIG AND GREY ELEPHANT
2. I AFRAID WOLF
3. BEAR EAT BERRIES
Human interpreter
1. YOU WATCH YOUTUBE
2. WEATHER TODAY WET
3. WHICH YOUR FAVOURITE MOVIE

Table 2: Sign sentences performed (in GLOSS).

3.3. Measurements

Due to the notable variance typically observed in the backgrounds of participants, as well as their varying levels of comprehension in their native language (L1) and second language (L2), we made the decision to investigate which form of assessment would be most appropriate for these types of studies.

In this section, we introduce a set of standard, established metrics and questionnaires that have been modified to suit the needs of individuals who are deaf.

3.3.1. Godspeed Questionnaire

We utilize the Godspeed questionnaire (Bartneck et al., 2009) that can be commonly applied for the evaluation of avatars and robots in the fields of human-computer interaction (HCI) and human-robot interaction (HRI), respectively. A growing body of literature in virtual avatar design and evaluation recognizes the indices of the Godspeed questionnaire to be uniquely suited to the given domain of application (Takagi and Terada, 2021; Weimann et al., 2022; Imashev et al., 2022). However, it is assumed that some of its concepts (see Figure 2) might be ambiguous and often misinterpreted since acquiring such concepts tends to be quite challenging (Parvez et al., 2019). Envisioning this challenge in the comprehension of the questions by deaf signers; this time, we decided to use vivid artifacts and selected clip art for each section of the questionnaire to make ranges intuitive and minimize interpreters' intervention.

Section	Items		
Anthropomorphism	Fake	-	Natural
	Machinelike	-	Humanlike
	Unconscious	-	Conscious
	Artificial	-	Lifelike
	Moving rigidly	-	Moving elegant
Animacy	Dead	-	Alive
	Stagnant	-	Lively
	Mechanical	-	Organic
	Artificial	-	Lifelike
	Inert	-	Interactive
	Apathetic	-	Responsive
Likeability	Dislike	-	Like
	Unfriendly	-	Friendly
	Unkind	-	Kind
	Unpleasant	-	Pleasant
	Awful	-	Nice
Perceived Intelligence	Incompetent	-	Competent
	Ignorant	-	Knowledgeable
	Irresponsible	-	Responsible
	Unintelligent	-	Intelligent
	Foolish	-	Sensible

Figure 2: Four sections of Godspeed questionnaire used in the study.

3.3.2. The Thermometer Scale

The thermometer scale is a visual scale that enables respondents to express their opinions on a particular issue by ranking it on a scale that ranges from "cold" (indicating complete disapproval) to "hot" (representing acceptance), similar to the temperature range on a real thermometer (Zavala-Rojas, 2014).

The use of "feeling thermometers" has become standard procedure in political research surveys since they were first introduced in the 1964 American National Election Study (ANES). Respondents utilize feeling thermometers to identify attitudinal elements on a continuum that spans from 0 (indicating extreme coldness) to 100 (representing intense warmth) (Wilcox et al., 1989).

3.4. Procedure

At the beginning of a session, the participants viewed videos containing consent forms and descriptions of the tasks translated into K-RSL. Upon completion of the introductory videos, consent form, and demographic information, participants were thereafter directed to the main section of the study.

Following the viewing videos featuring an agent, participants completed the task of 22 questions, of which 21 came from the Godspeed questionnaire. Given that deaf signers often perceive proficiency in sign language as a measure of overall intelligence, we have included the 22nd separate question in our survey performed in K-RSL, which asks about the agent's level of proficiency with sign language: "How do you think the agent does not know sign language at all or knows it very well?", response ranges from complete unfamiliarity to high proficiency. In addition, we employed the range of incompetence to competence on two occasions: firstly, in a broad sense, and secondly, specifically, referring to proficiency in sign language.

In this study, we utilized a percentage scale resembling a thermometer, placing it into Godspeed items between extremes to measure participants' evaluations. For instance, the scale used ranged from 0 (representing Inert) to 100 (representing Interactive), without any specific divisions for cold, warm, and hot climates. This lack of clarity could potentially cause confusion among participants. To convey contrasting extremes (concepts), we opted to rely on clip art and artifacts, such as a counterfeit plastic pear, an iron apple, a prop resembling an apple, and an edible apple (see Figure 3).

Figure 3: Vivid artifacts and the Thermometer scale with cliparts.

During the survey, an experienced SL interpreter explained the scale, translated all inquiries and implemented these vivid artifacts and cliparts in order to avoid any misunderstanding that may appear to participants. The procedure was the same for the second agent: participants interacted with both agents, and the order of conditions (agents) was distinct and counterbalanced for each participant.

4. Results

4.1. Performance Comprehension

Table 3 reveals that the Shapiro-Wilk test did not indicate a significant departure from normality for the SignMT, but did indicate a significant departure from normality for the Human agent. The results of the Wilcoxon Signed-Rank test revealed a statistically significant difference between the comprehension of the SignMT and Human agent: $Z = 2.6, p = .009$.

Agent	Mean (SD)	Shapiro-Wilk test
SignMT	37.59 (15.61)	$W = 0.897, p = 0.243$
Human	98.78 (3.67)	$W = 0.39, p < .001$

Table 3: Evaluation on how participants understood agents.

4.2. Godspeed Questionnaire with 0 to 100 Thermometer Scale

We incorporated inquiries from the Godspeed questionnaire, which evaluate Anthropomorphism, Animacy, Perceived Intelligence, and Likeability on a scale of 0 to 100 as it shown in Figure 3.

Calculated mean values and standard deviations of evaluations participants provided on these inquiries refer to Table 4.

Section	SignMT	Human
Anthropomorphism	18.00(12.76)	96.25(4.67)
Animacy	20.49(12.09)	93.26(7.27)
Likeability	40.50 (22.61)	91.17(11.57)
Intelligence	30.13(20.13)	91.00(12.28)

Table 4: Mean values(SD) of evaluation by the Thermometer scale.

Considering the total number of participants and the variety of backgrounds, we utilized the Shapiro-Wilk test to assess normality.

The Shapiro-Wilk test did not indicate a significant departure from normality for SignMT(Agent 1) across all sections: Anthropomorphism ($W = 0.924; p = 0.322$), Animacy ($W = 0.970; p = 0.918$), Likeability ($W = 0.947; p = 0.595$), and Intelligence ($W = 0.945; p = 0.561$). Meanwhile, the Shapiro-Wilk test revealed a significant departure from normality in the Human Agent for all sections: $W = 0.819, p = 0.016$ for Anthropomorphism, $W = 0.822, p = 0.017$ for Animacy, $W = 0.749, p = 0.003$ for Likeability, and $W = 0.750, p = 0.003$ for Intelligence.

The Wilcoxon Signed-Rank test demonstrated a statistically significant difference in participants' perception between SignMT and Human-agent for all sections, as indicated in Table 5).

Section	Wilcoxon Signed-Rank test
Anthropomorphism	$Z = 3, p = .002$
Animacy	$Z = 3, p = .003$
Likeability	$Z = 3.4871, p < .001$
Intelligence	$Z = 3, p = .003$

Table 5: Wilcoxon Signed-Rank test outputs.

4.3. General Feedback

While the participants recognized the smoothness of SignMT's gestures with their hands, finger flexibility, and overall signing, they also encountered difficulties understanding lip expressions and had to observe the agent repeatedly to comprehend the meaning of sentences. Determining the precise facial expression was also challenging; several individuals characterized it as cunning, while others interpreted it as benevolent, playful, or silly.

Initially, the participants displayed a willingness and were open and receptive to the idea of assigning distinct colors to each finger. Nevertheless, as the procedure progressed, the fingers got entangled and difficult to differentiate. In addition, the thumb exhibits an abnormal hypertrophy.

5. Conclusions and Future Scope

The study reveals that despite being a cutting-edge technology that provides a realistic and persuasive way of motion, participants could not comprehend the generated sign language sequences. There is still room for development, and further progress is needed to rival the proficiency of a human sign language interpreter.

The study suggests that the Thermometer scale may offer a broader range for subjective evaluation compared to Likert scales, especially for small groups. Additionally, it enables a more detailed assessment of the distinctions across agents since the research group revealed significant disparities between the agents assessed throughout all questionnaire sections.

The study findings propose to include a new distinct section called **Perceived SL Proficiency** aimed at replacing the present **Perceived Safety** section of the Godspeed questionnaire in cases where sign language avatar evaluation is required. This additional section would entail inquiring about the level of expertise in sign language since, in many instances, participants mistakenly conflate general proficiency and intelligence with sign language expertise quite often. Moreover, the existing two inquiries we have formulated regarding this subject are inadequate to justify the allocation of a distinct section.

It is crucial to note that the present study has a somewhat limited participant number, with only two signing agents to evaluate. Therefore, it is very advisable to carry out a follow-up study including more agents and a larger number of subjects. It is necessary because the results obtained from this study are insufficient to extrapolate to all individuals in the CIS population who use sign language. There are other alternative signing avatar technologies under consideration: SigML(Kaur and Kumar, 2016) is now the most esteemed signing generation system, taking into account comments from deaf individuals. Another modern approach (Kacorri and Huenerfauth, 2016) allows for the appropriate display of facial expressions based on the performed sign. Additionally, a data-driven approach (Imashev et al., 2022) enables the generation of physiologically believable body motions, resulting in more natural-looking performance of signs.

6. Acknowledgements

This work was supported by the Nazarbayev University Faculty Development Competitive Research Grant Program 2022-2024 "Kazakh-Russian Sign Language Processing: Data, Tools, and Interaction". The award number is 11022021FD2902.

We would like to express our gratitude to Aidana Utegenova, a sign language interpreter who portrayed the role of Human Agent in the present user study. We would also like to thank Viktorya Antonishina and Azamat Kenzhekhan for their helpful assistance during the study.

7. Bibliographical References

Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1(1):71–81.

Marta Bosch-Baliarda, Olga Soler Vilageliu, and Pilar Orero. 2019. Toward a sign language-friendly questionnaire design. *The Journal of Deaf Studies and Deaf Education*, 24(4):333–345.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, et al. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 16–31.

H. Brock and K. Nakadai. 2019. Deep jslc: A multimodal corpus collection for data-driven gener-

ation of japanese sign language expressions. pages 4247–4252. European Language Resources Association (ELRA). Cited By 4.

Maxime Delorme, Michael Filhol, and Annelies Braffort. 2009. Animation generation process for sign language synthesis. In *2009 Second International Conferences on Advances in Computer-Human Interactions*, pages 386–390. IEEE.

Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. 2019. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6569–6578.

RM Farwell. 1976. Speechreading, a review of the research. *American Annals of the Deaf*, 121:19–30.

Sylvie Gibet, Nicolas Courty, Kyle Duarte, and Thibaut Le Naour. 2011. The signcom system for data-driven animation of interactive virtual signers: Methodology and evaluation. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 1(1):1–23.

Matt Huenerfauth and Hernisa Kacorri. 2014. Release of experimental stimuli and questions for evaluating facial expressions in animations of american sign language. In *sign-lang@ LREC 2014*, pages 71–76. European Language Resources Association (ELRA).

Matt Huenerfauth and Hernisa Kacorri. 2015. Best practices for conducting evaluations of sign language animation. *Journal on Technology and Persons with Disabilities*, 3:20–32.

Alfarabi Imashev, Nurziya Oralbayeva, Vadim Kimmelman, and Anara Sandygulova. 2022. A user-centered evaluation of the data-driven sign language avatar system: A pilot study. In *Proceedings of the 10th International Conference on Human-Agent Interaction*, HAI '22, page 194–202, New York, NY, USA. Association for Computing Machinery.

Hernisa Kacorri and Matt Huenerfauth. 2016. Continuous profile models in asl syntactic facial expression synthesis. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2084–2093.

Hernisa Kacorri, Matt Huenerfauth, Sarah Ebling, Kasmira Patel, Kellie Menzies, and Mackenzie Willard. 2017. Regression analysis of demographic and technology-experience factors influencing acceptance of sign language animation. *ACM Transactions on Accessible Computing (TACCESS)*, 10(1):1–33.

Hernisa Kacorri, Pengfei Lu, and Matt Huenerfauth. 2013. Effect of displaying human videos during an evaluation study of american sign language animation. *ACM Transactions on Accessible Computing (TACCESS)*, 5(2):1–31.

Khushdeep Kaur and Parteek Kumar. 2016. Hamnosys to sigml conversion system for sign language automation. *Procedia Computer Science*, 89:794–803.

Michael Kipp, Quan Nguyen, Alexis Heloir, and Silke Matthes. 2011. Assessing the deaf user perspective on sign language avatars. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, pages 107–114.

Kgatlhego Moemedi and James Connan. 2010. Rendering an animated avatar from signwriting notation. In *Proceedings of the Southern Africa Telecommunication Networks and Applications Conference*.

Amit Moryossef, Kayo Yin, Graham Neubig, and Yoav Goldberg. 2021. Data augmentation for sign language gloss translation. In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 1–11, Virtual. Association for Machine Translation in the Americas.

Maria Papadogiorgaki, N Grammalidis, and N Sarris. 2004. Vsigns–a virtual sign synthesis web tool. In *Workshop on Information and Knowledge Management for Integrated Media Communication*, pages 25–31. Citeseer.

Komal Parvez, Muzafar Khan, Javed Iqbal, Muhammad Tahir, Ahmed Alghamdi, Mohammed Alqarni, Amer Awad Alzaidi, and Nadeem Javaid. 2019. Measuring effectiveness of mobile application in learning basic mathematical concepts using sign language. *Sustainability*, 11(11):3064.

R. Pfau and J. Quer. 2010. *Nonmanuals: Their grammatical and prosodic roles*. Cambridge University Press. Cited By 86.

Roland Pfau, Markus Steinbach, and Bencie Woll. 2012. *Sign language: An international handbook*, volume 37. Walter de Gruyter.

Floris Roelofsen, Lyke Esselink, Shani Mende-Gillings, Maartje De Meulder, Nienke Sijm, and Anika Smeijers. 2021. Online evaluation of text-to-sign translation by deaf end users: Some methodological recommendations (short paper). In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 82–87.

Jerry C Schnepp, Rosalee J Wolfe, John C McDonald, and Jorge A Toro. 2012. Combining emotion and facial nonmanual signals in synthesized american sign language. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*, pages 249–250.

Hisashi Takagi and Kazunori Terada. 2021. The effect of anime character's facial expressions and eye blinking on donation behavior. *Scientific reports*, 11(1):1–8.

Thure Weimann, Martin Fischer, and Hannes Schlieter. 2022. Peer buddy or expert?-on the avatar design of a virtual coach for obesity patients. In *HICSS*, pages 1–10.

Clyde Wilcox, Lee Sigelman, and Elizabeth Cook. 1989. Some like it hot: Individual differences in responses to group feeling thermometers. *Public Opinion Quarterly*, 53(2):246–257.

Diana Zavala-Rojas. 2014. Thermometer scale (feeling thermometer). *Encyclopedia of quality of life and well-being research, Springer Dordrecht (Netherlands)*, pages 6633–6634.