2021 Thirteenth International Conference on Quality of Multimedia Experience (QoMEX)

Evaluating the translation of speech to virtually-performed sign language on AR glasses

Lan Thao Nguyen   Florian Schicktanz
Technische Universität Berlin
Berlin, Germany

Aeneas Stankowski   Eleftherios Avramidis
German Research Center for Artificial Intelligence (DFKI)
Berlin, Germany

***Abstract*—This paper describes the proof-of-concept evaluation for a system that provides translation of speech to virtually performed sign language on augmented reality (AR) glasses. The discovery phase via interviews confirmed the idea for a signing avatar displayed within the users field of vision through AR glasses. In the evaluation of the first prototype through a wizard-of-Oz-experiment, the presented AR solution received a high acceptance rate among deaf and hard-of-hearing persons. However, the machine learning based method used to generate sign language from video still lacks the required accuracy for fully preserving comprehensibility. Signed sentences with large recognisable arm movements were understood better than sentences relying mainly on finger movements, where only a small interaction space is visible.**

***Index Terms*—sign language translation, real-time translation, augmented reality, AR glasses, avatar, inclusion**

I. Introduction

An ongoing effort in society is inclusion, as a process of empowering disabled people to participate in areas of social life. About one million people in Europe are deaf and until now depend on other people's help for managing their life beyond their daily routines. This help is provided by trained sign language (SL) interpreters, who charge for their services and have limited availability. Until now, no automatization impact improved the community's needs. The automatic translation and generation of SL still needs continuous research. This work builds the baseline for a new approach by conducting a two-level evaluation aiming to a proof of concept, while including members of the deaf community through the whole iteration process. A translation independent of other people's availability and interpreter services would increase the inclusion of deaf people into public life significantly.

Deaf and hard-of-hearing (DHH) people communicate mainly through sign languages, which consist of hand and arm gestures, body movements, as well as facial expressions. Research and our interviews with deaf people in the *discovery phase* showed that if deaf people can simultaneously see the interpreter's gestures as well as the speaker's mouth and gestures, they can understand the translation even better. Around 30% of what is said can be read from the speaker's lips if the mouth movements are clear [5]. This insight was the center to the chosen approach and choice of using AR glasses.

By using AR glasses, deaf people can follow the situation while translation is being provided in the user's view through e.g. an avatar. An avatar displayed via AR brings a higher benefit for the deaf person by being able to follow the translation as well as the facial expression and gestures. Conversations can be followed with eye contact and facial expressions can be exchanged. To validate the approach, a strategy was designed to provide a fundamental proof of concept.

II. Related Work

The first framework for generating a sign-language avatar from text was specified in 2000 [4]. A SL interpreter in a web application named *WebSign* was presented in 2007 [6]. It contained a word-sign dictionary enabling a real-time text to SL translation.

In 2016, a SL avatar was built that translates German train announcements into Swiss-German Sign Languages [3]. In the same year, [1] developed a holographic avatar that translates English to Signed English, also in real-time and with the help of an animation database. Similar to the present work, the avatar is displayed in AR glasses and the result was tested with people of the deaf community. Their use case is focusing on deaf children during math classes at school, while ours on deaf people independent of age in a doctor appointment.

A lately published paper [2] compared the usability of a head-mounted device with a smartphone while being connected to a live interpreter. The evaluation with hearing people and people from the DHH community showed that the hands-free device was favored. For example, participants liked the possibility to simultaneously have eye contact with the speaker and being able to use their hands. This confirms our underlying assumptions to evaluate the usefulness and acceptance of a virtual signing avatar on AR glasses with DHH people.

Recent studies also exist regarding the automatic generation of SL [8, 10, 11]. The authors of [12] extracted human joint coordinates with Deep Learning and generated realistic sign videos from the resulting 2D skeleton. They evaluated the comprehensibility of both the 2D skeleton presentation and the generated sign videos with four native SL speakers. A preference was found towards the generated videos. Comparable to the current research paper, human body keypoints were identified with machine learning from videos without depth information, which were then used to create virtually

978-1-6654-3589-5/21/$31.00 ©2021 IEEE

performed sign language. In contrast, the current paper examines the clarity of a human-like 3D avatar and does not aim to create realistic depictions of a person.

III. Methodology and Experimental Setup

The evaluation consisted of two phases. In the first phase, the so-called *discovery phase*, problem interviews were conducted to gain an understanding for the end-users' needs and to validate the approach of displaying a virtual avatar on AR glasses. In the second phase, the realized approach was evaluated in a user study by seven people from the DHH community and one hearing person. The three female and five male participants had low to mature technical literacy. The aim was to analyze the translations' comprehensibility through the avatar and the general acceptance of the presented technology.

A. Problem Interviews

To gain access to the community, we got in contact with the *Zentrum für Kultur und visuelle Kommunikation der Gehörlosen Berlin und Brandenburg e.V. (ZfK)*, which is the main contact point for deaf people in Berlin and Brandenburg. Interviews were arranged and held with two hearing interpreters, one deaf interpreter who is a native SL speaker, and one hard-of-hearing person.

Nine different mock-ups were presented to the interviewees. They were asked to compare whether they would prefer a real-time translation on a tablet, on a smartphone, or through AR glasses (see figure 1). It was also evaluated if they would rather use a glossary, receive support by a live interpreter or see a virtual avatar. Positive ratings were received regarding the use of AR glasses with an avatar, which validated the approach.

The suggestion to use a SL avatar when an interpreter is not available was met with great approval. Our research has shown that with a little practice, the approach is comparable to the situation of watching news with an interpreter displayed on screen. One of the interviewees proposed to use a HoloLens device and this idea was then adopted for the implementation.

After the four qualitative interviews, a strategy was designed to create a fundamental proof of concept for the chosen approach. The goal of the proof of concept was to validate the practical feasibility of the concept and its acceptance by the DHH community through a user study.

During the interviews, various situations were discovered in which deaf people have difficulties to communicate with hearing people without an interpreter. Besides parent-teacher conferences and going to the citizens' office, visiting the doctor was identified as one of the most difficult situations for deaf people. Its existing challenges can be reduced to a great amount with the help of an interpreter. Therefore, the decision for the study was to simulate a doctor's appointment where the deaf person is accompanied by a virtual SL interpreter.

B. Implementation

For the proof of concept, an AR prototype was built that simulates a live translating sign avatar by displaying the according sign translations into the user's field of view at the correct moment. First, the animations were created using free open source software and no special hardware. Video footage was filmed using a high-quality RGB camera. This was analysed with MediaPipe [7] to track the upper body, hands and face. The XYZ coordinates of the resulting tracking landmarks were processed further to motion capture data in the BVH file format. In the next step, a 3D avatar was created and the motion capture data was applied on it. Given the implementation demands of facial expressions and pose, for this first proof of concept only hand and arm movements were captured, aiming to continue the implementation once the concept is justified. Finally, a HoloLens 1 [9] application was developed and deployed on the AR glasses.

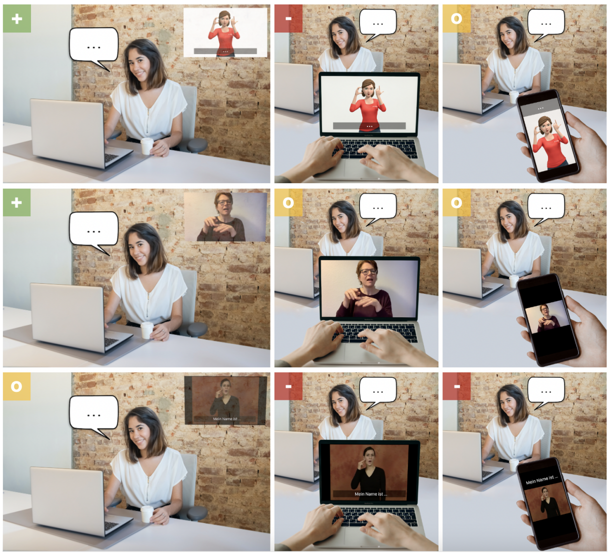

Fig. 1. Mock-ups presented during the problem interview. Green (+) marks the highest, yellow (o) the second highest and red (-) the lowest overall rating. Options from left to right: AR glasses, laptop and smartphone. Options from top to bottom: 3D avatar, remote interpreter and video glossary. Combined with a live stream or an avatar, AR glasses are considered as most useful. Video glossary was the least preferred translation type regardless of the device.

C. User Study

The final user test to observe and confirm the assumptions took place in a conference room at the ZfK in Babelsberg (see figure 2). To capture all reactions during the test, the scene was filmed from three different perspectives. Due to the short reaction time from the participants after each sentence, the recording from multiple angles became imperative. One camera filmed the person performing the doctor and reading aloud prepared questions, one the study participants, and a third one the whole scene from a side view. All involved participants showed a high level of interest to test the new technology stack.

The designed test resembles a Wizard-of-Oz-experiment, where the user does not know about the early stage of the product development. In the prepared HoloLens 1 application the 3D avatar can be dragged around in space at the beginning to enable the project team to position the sign avatar next to the impersonated doctor. The subjects should be able to watch the avatar's signs while having the possibility to observe the speaker's gestures, facial expressions and lip movement. The

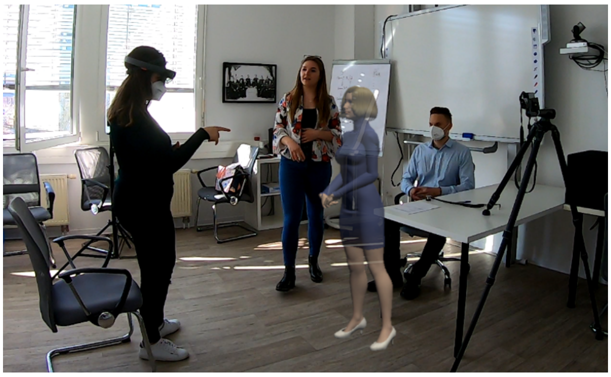

Fig. 2. Final user test setup. A participant (left) is wearing a HoloLens 1 displaying a virtual avatar in front of the table by the impersonated doctor (right), while a sign interpreter (middle) is prepared to translate the participant's feedback.

avatar signs 14 sentences and questions in a defined order that form together with the subject's reactions a dialog for a common medical examination. Each phrase can be started manually by one of the experiment supervisors.

Two sign interpreters supported the user study and made it possible to communicate and comprehend the immediate reactions of the study participants, helping us to check if the sentences were understood properly.

When the study participants and the setup were ready, the person playing the doctor read the sentences aloud. In the same moment, the performance of the avatar was triggered in the HoloLens via the external clicker. Intuitively, most of the study participants responded to the performance, making a statement to the interpreter. The statements of the study participants were translated to spoken language, so it was clear, whether or not the sentence has been understood.

Additionally to the recorded reactions, the participants answered a 5-point scale feedback questionnaire with the questions *"How did it feel to wear a HoloLens?"* (Q1), *"Would you make use of this kind of translation if it was more mature?"* (Q2), *"How well could you understand the avatar's signing?"* (Q3) and *"How much did the missing avatar's facial expressions affect you?"* (Q4).

IV. Results

It was observed that phrases with prominent arm movements were understood better than most others. For example the questions *"Do you have pain?"* or *"Do you need a doctor's certificate?"* include striking, recognisable arm movements and were thus understood very well. On the other hand, sentences signed mainly with finger movements had a lower comprehensibility rate. Since fingers are harder to track by MediaPipe, motions could not be transferred fully accurately, but conducting a sign with the proper hand shape and movement is important for preserving its meaning. The participants gave the feedback that they would be interested in a more advanced and more user-friendly AR solution. One participant commented that they would not consider wearing a HoloLens in public. An AR device resembling usual eyeglasses might increase attractiveness, while using a lighter device with a greater field of view could improve usability. A fully automatic translation system still has to be realized. More than 50% indicated or tended to the statement, that they would be glad to use this translation method if it was more mature (see table I). Over a third of the participants stated to understand the avatar's translation at least moderately. Following the verbal feedback during the study, about 30% of the translation was understood. However, this seemingly low rate is related to the insufficient accuracy of the avatar's movements, as well as the missing facial expressions which adopt grammatical functions and lip movements that can be part of a sign. Three quarters of the participants felt affected severely by these missing expressions.

TABLE I: Feedback questionnaire results

		Votes for scale points in %					
		1	*2*	*3*	*4*	*5*	
Q1	*comfortable*	0	62.5	12.5	25	0	*uncomfortable*
Q2	*I'd be glad to*	25	37.5	25	12.5	0	*not at all*
Q3	*very well*	0	12.5	25	50	12.5	*not at all*
Q4	*severely*	75	12.5	0	0	12.5	*not at all*

V. Conclusion

To validate the concept and understand the DHH community needs, a *discovery phase* was conducted before the implementation, suggesting the generation of a SL avatar in a HoloLens. A wizard-of-Oz-evaluation was performed to draw conclusions about how promising the approach could be. By monitoring the test persons' responses, it was possible to analyze to which extent the actor's questions were understood.

However, not only through the participants' reactions during the test, but also through a final questionnaire, it was possible to validate what percentage of the conversation was comprehended. It should be noted that translations requiring stronger, remarkable arm movements were perceived better than others due to the larger visible interaction space and higher recognition factor. Overall, the participants were able to understand around 30% of the translation via the avatar, phrases with stronger arms movements were understood by over 80%, based on the reactions of the participants. For a higher comprehensibility rate, the presented machine learning based SL generation method needs further improvement. Five out of eight participants indicated they understood the avatar's signing lower than on a moderate level.

We conclude that a signing avatar is a promising solution to enable deaf people the access to the hearing world. AR glasses enable deaf people to monitor the translation while following the speaker's facial expressions, gestures, and body language side by side. Eye contact and the exchange of expressions become possible at the same time, improving the quality of a conversation. In further work, we will repeat the evaluation along the further implementation of a more complete avatar, following our described findings.

Acknowledgements

This research was supported by Uwe Schönfeld, Mathias Schäfer, Helene Sudermann, Lukas Gesche, Johann Carl Reischies, Marco Richardson, Willi Reuter, Fabrizio Nunnari, Esther Zahn and the QU Lab of TU Berlin, incl. funding by the BMBF (German Federal Ministry of Education and Research) in the project SocialWear (Socially Interactive Smart Fashion).

References

[1] N. Adamo-Villani and Saikiran Anasingaraju. Holographic Signing Avatars for Deaf Education. In *eLEOT*, 2016.

[2] Larwan Berke, William Thies, and Danielle Bragg. Chat in the Hat: A Portable Interpreter for Sign Language Users. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '20, pages 1–11, New York, NY, USA, October 2020. Association for Computing Machinery.

[3] Sarah Ebling and John Glauert. Building a Swiss German Sign Language avatar with JASigning and evaluating it among the Deaf community. *Universal Access in the Information Society*, 15(4):577–587, November 2016.

[4] R. Elliott, J. R. W. Glauert, J. R. Kennaway, and I. Marshall. The development of language processing support for the visicast project. In *Proceedings of the Fourth International ACM Conference on Assistive Technologies*, Assets '00, page 101–108, New York, NY, USA, 2000. Association for Computing Machinery.

[5] Deutscher Gehörlosen-Bund e.V. Gehörlosigkeit. https://www.gehoerlosen-bund.de/faq/geh%C3%B6rlosigkeit, 2021.

[6] Mohamed Jemni and Oussama Elghoul. An Avatar Based Approach for Automatic Interpretation of Text to Sign Language. *9th European Conference for the Advancement of the Assistive Technologies in Europe*, pages 3–5, January 2007.

[7] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A Framework for Perceiving and Processing Reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*, 2019.

[8] Matthew Mcconnell, Mary Ellen Foster, and Mary Ellen. Two Dimensional Sign Language Agent. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, New York, NY, USA, 2020. ACM.

[9] Microsoft. Hololens-Hardware (1. Generation). https://docs.microsoft.com/de-de/hololens/hololens1-hardware, 2019.

[10] Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. Text2Sign: Towards Sign Language Production Using Neural Machine Translation and Generative Adversarial Networks. *International Journal of Computer Vision*, 128(4):891–908, apr 2020.

[11] Sugandhi, Parteek Kumar, and Sanmeet Kaur. Sign Language Generation System Based on Indian Sign Language Grammar. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 19(4):1–26, jul 2020.

[12] Lucas Ventura, Amanda Duarte, and Xavier Giró i Nieto. Can Everybody Sign Now? Exploring Sign Language Video Generation from 2D Poses. In *ECCV 2020 Workshop on Sign Language recognition, Production and Translation (SLRTP)*, 08/2020 2020.