DEEP: an automatic bidirectional translator leveraging an ASR for translation of Italian sign language

Nicolas Tagliabue[1], Elisa Colletti[1], Roberto Tedesco[1], Francesco Roberto Dani[1], Alessandro Trivilini[1], Sonia Cenceschi[1],

[1]Scuola Universitaria Professionale della Svizzera Italiana (SUPSI), Switzerland,

Correspondence: nicolas.tagliabue@supsi.ch, elisa.colletti@supsi.ch, roberto.tedesco@supsi.ch, francesco.dani@supsi.ch, alessandro.trivilini@supsi.ch, sonia.cenceschi@supsi.ch

Abstract

DEEP is a bidirectional translation system for the Italian Sign Language, tailored to two specific, common use cases: pharmacies and the registry office of the municipality, for which a custom corpus has been collected. Two independent pipelines permit to create a chat-like interaction style, where the deaf subject just signs in front of a camera, and sees a virtual LIS interpreter, while the hearing subject reads and writes messages into a chat UI. The LIS-to-Italian pipeline leverages, in a novel way, a customized Whisper model (a well-known speech recognition system), by means of "pseudo-spectrograms". The Italian-to-LIS pipeline leverages a virtual avatar created with Viggle.ai. DEEP has been evaluated with LIS signers, obtaining very promising results.

1 Introduction

Sign languages represent a particular challenge for Machine Translation (MT) systems, for various reasons. First of all, sign languages are true languages, with their own lexicon, syntax, and grammar (they are not a "gestural version" of another language); moreover, the signs must be captured, usually by means of a video camera, and the resulting data stream is much more complex than speech or text (usual input of MT systems); in addition, parallel corpora are rare and quite small; finally, sign languages, being *oral languages* (i.e., a standard writing system is not defined), tend to vary a lot among different groups. As a result, MT of sign languages is still an open problem.

In this paper we introduce DEEP, a bidirectional MT system between the Italian Sign Language (LIS) and Italian, designed for two common use cases that can be beneficial to deaf individuals in their daily lives: pharmacies and the registry office of the municipality. DEEP aims to help deaf persons gain more independence when interacting in such cases, without the need of an interpreter.

We collected an ad-hoc, parallel corpus, developed the two MT pipelines, and assembled a preliminary test platform, which will then evolve to a production-ready kiosk. We focused on simplifying the interaction between the deaf user and the system, as an intuitive UI is essential for the system's effectiveness. DEEP is designed for two specific use cases, and aims to provide a pragmatic answer to deaf persons. At the same time, however, the methodology is generic and could be used for creating a full MT (given a proper corpus is collected).

For implementing the two pipelines, we leveraged a couple of neural models and methodologies. In particular, and we argue this is a novel approach, for the LIS-to-Italian pipeline, we leveraged and customized Whisper, an Automatic Speech Recognition (ASR) system from OpenAI. Indeed, we converted the signs in a "pseudo spectrogram", used for training a slightly modified version of Whisper.

The final system is still in an experimental phase, and was built to work with specific webcam settings (optics, frame rate, aperture, etc.). Given the nature of the prototype which requires real-time video streaming and substantial GPU power in order to function, a live demo published on the internet is not feasible to handle for our experimental setup. A demo video is available at https://youtu.be/QWV6mPqhwmE.

2 Related work

Closing the communication gap between hearing and hearing-impaired communities is essential. Achieving seamless, two-way communication relies on the creation of an advanced system capable of performing two key functions: sign language recognition and sign language production.

With the rise of deep learning, many researchers have tried to use neural network methods to deal with sign recognition and generation (Toshpulatov

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 221–229

et al., 2025, Rai et al., 2024). In the following the most interesting works are described.

2.1 Sign Language Recognition

Sign Language Recognition (SLR) systems can be divided into three categories: Isolated Sign Language Recognition (ISLR), Continuous Sign Language Recognition (CSLR) based on glosses, and gloss-free CSLR.

ISLR is too limited so we didn't consider it. Glosses serve as a method for depicting discrete gestures in textual format. Exhibiting a one-to-one correspondence with signs, they can function as a valuable intermediary between manual and oral communication systems. Nevertheless, gloss notations are also regarded as a partial and imprecise portrayal of manual communication systems (Müller et al., 2023). Moreover, the process of creating gloss annotations is a time-consuming and labor-intensive endeavor. Thus, we focus on modern, gloss-free CSLRs.

Hamidullah et al. (2024) introduced a new gloss-free model, sign2(sem+text), that utilizes sentence embeddings for supervision of target sentences during training, effectively replacing the need for glosses. This method significantly narrows the performance gap between gloss-free and gloss-dependent systems, particularly when no additional SLR datasets are used for pretraining. Zhou et al. (2023) achieved notable results by incorporating visual-language pretraining inspired by Contrastive Language-Image Pre-training (Radford et al., 2021). Their two-stage approach integrate this technique with masked self-supervised learning to bridge the semantic gap between visual and text representations.

Lin et al. (2023) introduced the Gloss-Free End-to-end sign language framework (GloFE). This method improves SLR performance by exploiting shared semantics between signs and corresponding spoken translations. Key concepts from text are used as weak intermediate representations. Most recently, Rust et al. (2024) developed a self-supervised model, pretrained on the large-scale YouTube-ASL dataset. This approach led to state-of-the-art performance on the How2Sign dataset, demonstrating the potential of leveraging extensive pretraining data.

Finally, a study by Arib et al. (2025) presented SignFormer-GCN, which utilizes both keypoint and RGB features to capture the pose and configuration of body parts involved in sign language actions. This approach combines transformer and spatio-temporal graph convolutional network (STGCN) architectures to better capture the context and spatial-temporal dependencies of sign language expressions. The method showed competitive performance across multiple datasets.

2.2 Sign Language Production

The field of Sign Language Production (SLP) remains a complex challenge. Nevertheless, the field has witnessed significant advancements in recent years, with studies focusing on developing sophisticated end-to-end models for translating spoken language into continuous sign language sequences.

One of the most notable breakthroughs in this domain has been the development of Progressive Transformers (Saunders et al., 2020). Their model offers a solution for converting spoken language sentences into sign language gestures. The same authors in 2021 introduced a two-stage deep learning method for sign language production: the first stage converts spoken language sentences into a latent sign language representation, while the second stage employs a Mixture of Motion Primitives (MOMP) framework to create expressive sign language sequences from this representation.

Stoll et al. (2020) introduce an innovative method for generating realistic sign language videos from spoken language sentences. The deep learning approach combines neural machine translation with Motion Graph, generative adversarial networks, and motion generation techniques to produce sign video sequences. It achieves this result with minimal reliance on annotated data. Nevertheless, this system relies on gloss representation as an intermediary representation, which can oversimplify sign language. Glosses do not fully capture the richness of sign language, especially non-manual features like facial expressions and body posture, which are crucial for context and meaning.

2.3 Focusing on Italian Sign Language

Most of the works on LIS concerns SLP. Colonna et al. (2024) introduce a model designed to generate accurate LIS gestures from speech. The model uses an iterative framework that integrates textual, audio, and visual data to progressively refine generated gestures, ensuring realism and contextual relevance. Preliminary results show the model effectiveness in producing realistic LIS poses.

About SLR, the LIS2Speech application (Mercurio, 2021) employs advanced technologies such

as neural networks, deep learning, and computer vision to translate isolated LIS signs into Italian text and speech. This approach relies on skeletal features of the hands, body, and face, extracted from videos. However, the application currently translates only one isolated sign at a time, which limits its real-life practicality.

Furthermore, we identified Algho[1], a virtual assistant reported to offer bidirectional translation between spoken language and LIS. However, the commercial availability of this product is uncertain. The interactive demo appears to be limited to SLP. As far as we know, no other recent literature exists for SLR of LIS.

2.4 Comparison against mentioned systems

The SLR works we presented cannot be deployed as actual SLR systems due to the limited corpus they adopt. Our approach is different: we aim at releasing a SLR that can be used in the field. About SLP, our approach is pragmatic, as we depend on pre-signed LIS sentences for the two use cases we focus on: we trade off some flexibility in exchange for highly effective LIS generation.

Moreover, our prototype stands out from existing systems due to its ability to operate in realtime and bidirectionally, without requiring to respect turns. It focuses on two specific scenarios to ensure robustness, non-invasiveness, and usability. Finally, to the best of our knowledge, it is the only system capable of translating complete sentences between LIS and Italian, whereas other systems are limited to individual signs. These features make it a significant step forward, enhancing accessibility and social inclusion for the deaf community.

3 The DEEP Corpus

The DEEP dataset comprises 36 818 samples of LIS video recordings, totaling approximately 62 hours of footage. Such recordings, captured at 1920×1080 resolution and 60 FPS, with carefully calibrated shutter speeds to minimize motion blur, were annotated with corresponding Italian sentences. The dataset was developed to support the DEEP system in two specific scenarios: pharmacies and municipality office interactions. Starting from 3075 commonly used sentences in these contexts, 17 subjects (13 native LIS speakers and 4 LIS interpreters) were asked to sign as many as possible of such sentences. To further enhance the corpus, 56 322 synthetic samples were created by combining recorded sentences with individual words signed in dactylology (person names and surnames, drug names, toponyms, numbers, etc.), adding roughly 128 hours of content. This comprehensive approach resulted in a robust dataset designed to facilitate sign language communication in the two target real-world settings. In total, we had 93 140 samples, corresponding to 190 hours of video. The dataset will be released subsequent to securing consent from all participants.

4 System Design

The DEEP system facilitates bidirectional translation between LIS and Italian. The DEEP system architecture encompasses both translation pipelines: from LIS to Italian and from Italian to LIS. This comprehensive approach enables seamless communication between LIS and Italian speakers, bridging the linguistic gap between these two languages.

Although the current focus is on LIS, the techniques developed in this research could be adapted to other sign languages, broadening the scope and impact of this work.

4.1 LIS-to-Italian Translation Pipeline

This pipeline implements a gloss-free SLR (see Figure 1), where a high-resolution video camera (1920×1080, 60 FPS) continuously captures frames. The Subject Detection module monitors these frames, waiting for a human body to appear in front of the camera. Once detected, a motion detection trigger initiates video recording, which continues until a resting pose is identified. The recorded video then undergoes analysis using Google MediaPipe's Holistic model[2], generating a 3D skeleton for each frame. In this module, we also perform time interpolation to fill in any missing nodes (for frames where MediaPipe lost tracking). Then, for each frame, we reduce the set of 3D nodes into *measures* representing in a compact way the subject's pose of her/his face, torso, arms, and hands.

We utilize three measure typologies: 3D points (e.g., the position of wrists), polar angles (e.g., the direction hands are pointing), and distances (e.g., the distance between lips), which are normalized against invariant body features, like torso height, and further adjusted to fall within a sub-

[1] https://www.alghoncloud.com/funzionalita/artificial-human-lis/

[2] https://github.com/google-ai-edge/mediapipe/blob/master/docs/solutions/holistic.md

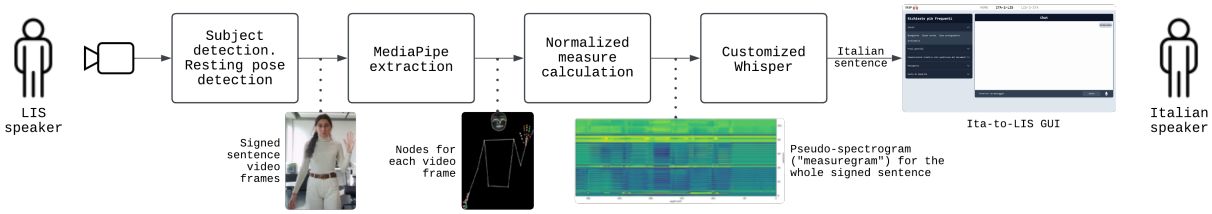

Figure 1: LIS to Italian Translation Pipeline

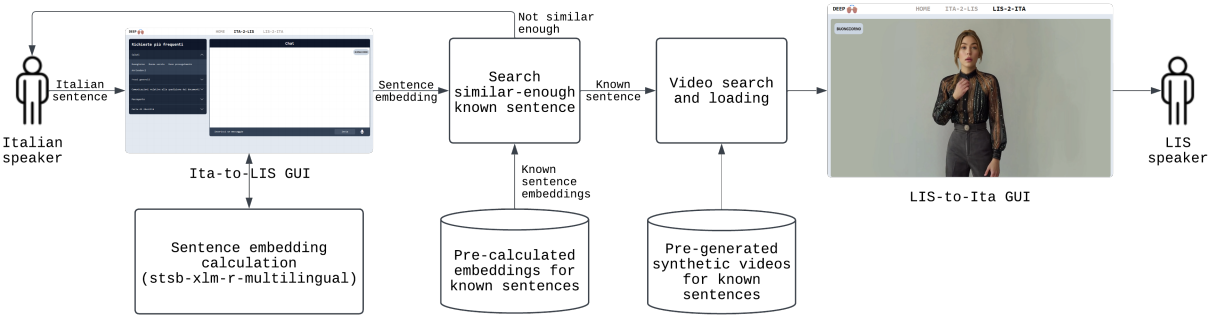

Figure 2: Italian to LIS Translation Pipeline.

set of the (-1, 1) range. The result is a vector of 110 numbers, for each frame; placing these vectors side by side we create a "measuregram", in analogy with the spectrogram often used by ASRs (see Appendix A). This measuregram is then processed by a customized version of OpenAI's Whisper model, adapted from HuggingFace, to generate Italian transcriptions. An autoencoder, built using the Whisper's encoder and an ad-hoc decoder, was trained on DEEP videos; the autoencoder was fed with measuregram and its goal was to reconstruct them. Then, the autoencoder's encoder was copied to the Whisper's encoder, and the whole Whisper was refined on the DEEP parallel corpus (see Appendix C).

4.2 Italian-to-LIS Translation Pipeline

This pipeline permits to translate Italian into LIS (see Figure 2). The Italian user inserts a sentence, which is converted into an embedding using the stsb-xlm-r-multilingual model from HuggingFace[3]. This embedding is then compared against a set of Italian sentences pre-calculated embeddings (from the DEEP dataset). If a sufficiently similar sentence is found (i.e., the Euclidean distance is below a given threshold), it serves as a key to retrieve the corresponding pre-generated synthetic video. Such videos are created using the DEEP collection of recorded LIS sentences, a photo of one of the researchers, and the Viggle.ia[4] online service. This method ensures privacy protection for individuals in the DEEP dataset while still producing high-quality synthetic sign language videos.

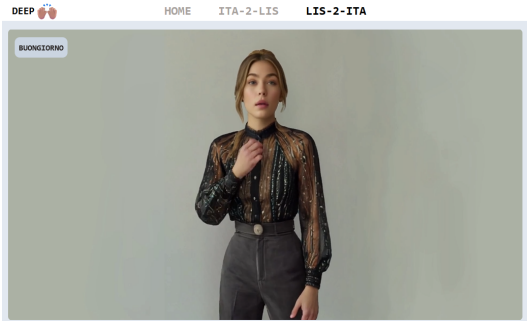

Figure 3: LIS-2-ITA page for deaf subject's UI.

Figure 4: "In Pose" and "Recording" triggers.

[3] https://huggingface.co/sentence-transformers/stsb-xlm-r-multilingual

[4] https://viggle.ai/

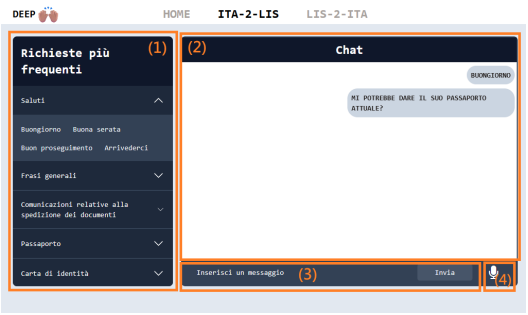

Figure 5: ITA-2-LIS page for hearing subject's UI.

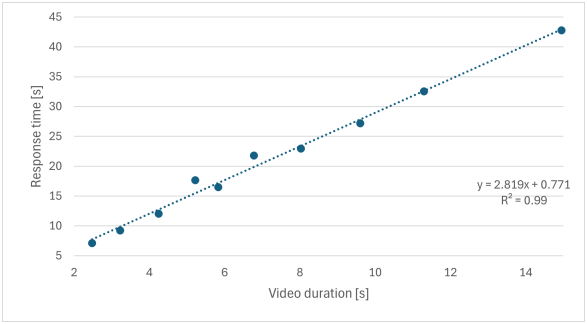

Figure 6: LIS-to-Italian response time experiment results; linear relationship between video duration and system response time. The dotted line shows the best-fit linear regression ($R^2 = 0.99$).

5 The Web Application

The DEEP experimental system is structured around two web applications, LIS-to-Italian and Italian-to-LIS, which implement a chat-like user experience. For the sake of simplicity the demo version shown in the pictures merges the two web applications into two pages of a single web app.

The two translation pages show the deaf subject and the hearing subject UIs. Note that the two communication "channels" between the two parties are independent: any of the two subjects can insert a sentence at any moment, thus ensuring a natural interaction between the two parties.

5.1 The LIS-2-ITA page

The LIS-2-ITA page is dedicated to the deaf subject, and its layout is straightforward, featuring a single section that prominently displays an avatar. When communication commences, a chat overlay appears in the top-left corner of the avatar section. This overlay has a transparent background, allowing for unobstructed viewing of the avatar while maintaining visibility of the conversation. The design, as illustrated in Figure 3, ensures a seamless and intuitive user experience for the deaf subject.

When a LIS speaker approaches the system, it activates an "In Pose" trigger; as a second trigger detects the start of a sign the system goes to the "Recording" mode, until a final "Resting" trigger detects a resting pose. During this phase, the system interprets the subject's LIS signing (Figure 4). Once the LIS phrase is completed, the system stops capturing the video and goes back to the "In Pose" mode until the subject moves out of the trigger zone. The LIS-to-Italian pipeline translates the recorded video, and the text is sent to the hearing person, who will see it on its ITA-2-LIS page, and to the chat overlay in the top-left corner of the avatar.

5.2 The ITA-2-LIS page

The ITA-2-LIS page is designed for hearing individuals. This interface consists of four primary sections (see Figure 5): A collection of predefined text messages featuring frequently used expressions, situated on the left side of the page, a textual conversation display on the right side of the page, a text entry field positioned at the lower edge of the chat display, and microphone activation button for speech input functionality.

All communication modalities transmit a textual message to both the LIS-to-Italian and Italian-to-LIS chat interfaces, which show it. Furthermore, the Italian-to-LIS translation pipeline generates the video of the avatar signing the Italian sentence, in LIS; such video is then sent to the LIS-2-ITA page.

6 Experiments and results

We conducted two experiments, to evaluate the effectiveness of our prototype, on a system equipped with an Intel Core i9-11900K, 64 GB DDR4 RAM, NVIDIA RTX 4090 (see Appendix B for the experimental setting).

6.1 Response time

For the SLR pipeline, the average translation time (measured from the moment the signer begins signing) was 2.9× the duration of the LIS sentences.

Although our customized Whisper model successfully utilized GPU acceleration, we faced challenges in adapting and recompiling MediaPipe to run on GPU; on our system MediaPipe ran exclusively on CPU.

Figure 6 shows a clear linear relationship between the duration of the videos and the system's response time. In this experiment, we used 10

Subject	LIS sentences			Dactylology signs			Class
	Identical	Clear enough	Obscure	Identical	Clear enough	Obscure	
Subject 1	**17**	13	1	**9**	7	0	LIS L1
Subject 2	**22**	8	1	4	**12**	0	LIS L1
Subject 3	**17**	14	0	0	**16**	0	LIS L2
Subject 4	**19**	11	1	**14**	0	2	LIS L1
Subject 5	**22**	8	1	**13**	3	0	LIS L2
Subject 6	**28**	3	0	**14**	2	0	LIS L2
Subject 7	**21**	8	2	2	**9**	5	Interpret.

Table 1: Italian-to-LIS experiment results.

videos of LIS sentence signed, with durations ranging from 2.48 to 14.95 seconds. The system's response time increased proportionally, from 7.13 seconds for the shortest video to 42.75 seconds for the longest.

The majority of the processing time, approximately 99%, is consumed within the MediaPipe model, taking between 7.03 and 42.24 seconds depending on the video length. The inference processing time was much shorter, ranging from 0.10 to 0.51 seconds, which is only about 1% of the total response time.

Both the Mediapipe processing and inference processing times showed a linear increase with video duration, indicating consistent performance scaling as video length grows.

For the SLP pipeline, the translation time was nearly instantaneous.

6.2 Italian-to-LIS Experimental Setting

We conducted a study with 7 LIS signers: 6 deaf individuals and 1 interpreter. The participants were asked to evaluate 31 LIS sentences from the DEEP corpus, signed by our avatar. Out of such sentences, 16 included dactylology signs, while 15 did not.

Each participant watched the LIS sentences and completed a form in which they assessed the alignment between the correct Italian sentences and the meaning conveyed by the avatar, answering two questions: "Do the signs of the avatar convey the same meaning as the Italian sentence?" and "If the LIS sentence contains a dactylology sign, is this sign comprehensible?".

For the first question the options were: the meanings in LIS and Italian are identical; the overall meaning is clear enough; and the meaning is obscure. For the question about dactylology signs, we used similar options.

Each participant was classified based on their LIS fluency: Individuals who learned LIS during early childhood as their first language ("LIS L1"), individuals who learned LIS later in life (after childhood; "LIS L2"), and interpreters.

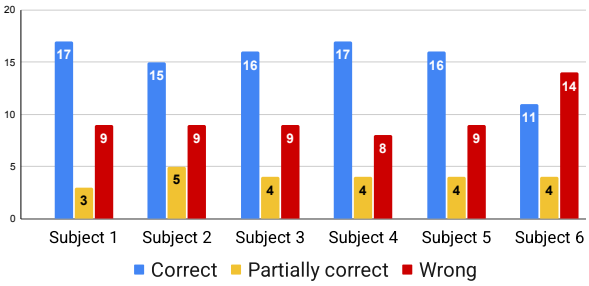

Figure 7: LIS-to-Italian experiment results. Partially correct means that the semantic was mostly conveyed.

Uncut	Cut at the beginning	Cut at the end	No sign
154	16	4	0

Table 2: Errors "Recording" & "Resting" triggers.

We summed the values for each answer option across all participants (Table 1; the most selected option is highlighted in bold). Based on results, the meaning conveyed by the virtual avatar was identical or clear enough to the Italian sentences (98.2%). About dactylology signs, the meaning was identical or clear enough (93.8%). The quality of the virtual avatar was thus quite satisfying.

6.3 LIS-to-Italian Experimental Setting

For this experiment, 6 deaf subjects performed 29 LIS sentences. Our goal here was twofold: calculating the accuracy of the LIS-to-Italian pipeline and testing the effectiveness of the triggers (which could result in truncated videos, if not working properly). The video of LIS signs was then passed to the LIS-to-Italian pipeline. The results are shown in Figure 7: 66.7% of correct or partially correct

sentences (69.7% not considering Subject 6).

We also examined the impact of the "In Pose" trigger and found only one false positive (unnecessary activation) out of the 174 sentences. Moreover, Table 2 highlights the errors where the LIS sentence may be cut at the beginning if the "Recording" trigger activates too late, at the end if the "Resting" trigger activates too early, or the entire LIS sentence may not be saved if the "Recording" trigger does not activate at all. We found 88.5% of LIS sentences were correctly treated, while 9.2% were cut at the beginning (all of which belonging to Subject 6, who obtained the worst results in Figure 7).

7 Discussion and Conclusions

This paper introduced DEEP, a bidirectional translator for LIS. The system, tailored on two common use cases, aims to help deaf persons gain more independence when interacting in such cases, without the need of an interpreter.

The system UI is particularly easy to use and permits a fluid interaction between deaf person and hearing person. The LIS-to-Italian pipeline is based on a customized, well-known ASR, demonstrating the feasibility of such models for sign language translation.

The recognition accuracy is still not optimal but we obtained very promising results. The sign generation was highly appreciated by the testers.

8 Limitations

Currently, the use cases are limited to pharmacies and municipality's registry office; this is due to the difficulties (and costs) of collecting corpora for sign languages. Moreover, the Italian-to-LIS pipeline (which selects a LIS video among a list of pre-recorded videos), although effective in vertical use cases, is less scalable than the LIS-to-Italian pipeline (which implements a true translation system). Finally, the recognition accuracy of the LIS-to-Italian pipeline should be further improved.

9 Ethical Considerations

For the corpus creation, LIS signers were compensated fairly. All LIS signers involved in the experiments were voluntary. All LIS signers were trained and informed about the task before participating. We guaranteed privacy of personal information, for all LIS signers. In particular, the experimental web application implemented strict data protection principles, refraining from storing any personal information.

Special thanks to Daniele Raffa of Handy Systems for making this project possible. This research was funded by Innosuisse, the Swiss Innovation Agency (grant[5] 100.656 IP-ICT). DEEP would not have been possible without the experience, feedback, and active contributions to data collection from members of the deaf community.

References

Safaeid Hossain Arib, Rabeya Akter, Sejuti Rahman, and Shafin Rahman. 2025. Signformer-gcn: Continuous sign language translation using spatio-temporal graph convolutional networks. *PLOS ONE*, 20(2):1–19.

Emanuele Colonna, Alessandro Arezzo, Domenico Roberto, David Landi, Felice Vitulano, Gennaro Vessio, Giovanna Castellano, et al. 2024. Towards italian sign language generation for digital humans. In *Proceedings of the Eighth Workshop on Natural Language for Artificial Intelligence (NL4AI 2024) co-located with 23th International Conference of the Italian Association for Artificial Intelligence (AI* IA 2024), CEUR-WS. org*.

Yasser Hamidullah, Josef van Genabith, and Cristina España-Bonet. 2024. Sign language translation with sentence embedding supervision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 425–434, Bangkok, Thailand. Association for Computational Linguistics.

Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Bang Zhang, and Yi Yang. 2023. Gloss-free end-to-end sign language translation. *Preprint*, arXiv:2305.12876.

Giuseppe Mercurio. 2021. LIS2SPEECH LIS translation in written text and spoken language. Master's thesis, Politecnico di Torino, Torino, Italy.

Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2023. Considerations for meaningful sign language machine translation based on glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 682–693, Toronto, Canada. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

[5] www.aramis.admin.ch/Grunddaten/?ProjectID=50430

Deepak Rai, Niharika Rana, Naman Kotak, and Manya Sharma. 2024. Real-time speech to sign language translation using machine and deep learning. In *2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, pages 1–5.

Phillip Rust, Bowen Shi, Skyler Wang, Necati Cihan Camgöz, and Jean Maillard. 2024. Towards privacy-aware sign language translation at scale. *Preprint*, arXiv:2402.09611.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020. Progressive transformers for end-to-end sign language production. In *Computer Vision – ECCV 2020*, pages 687–705, Cham. Springer International Publishing.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021. Mixed signals: Sign language production via a mixture of motion primitives. *Preprint*, arXiv:2107.11317.

Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. 2020. Text2sign: Towards sign language production using neural machine translation and generative adversarial networks. *Int. J. Comput. Vision*, 128(4):891–908.

Mukhiddin Toshpulatov, Wookey Lee, Jaesung Jun, and Suan Lee. 2025. Deep learning pathways for automatic sign language processing. *Pattern Recognition*, 164:111475.

Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. 2023. Gloss-free sign language translation: Improving from visual-language pretraining. *Preprint*, arXiv:2307.14768.

A Measuregram

A spectrogram is a visual representation of how the frequency content of a signal changes over time. The x axis shows the time while the y axis reports frequency bins; color is used to indicate the amplitude (loudness) of each frequency at each time instant. It's commonly used in audio analysis, speech recognition, and music visualization to show which frequencies are present in a signal and how they vary.

Figure 8 shows our *measuregram*, a representation inspired by spectrograms, where the 110 measures are shown on the y axis, the frame number on the x axis and the color indicates the measure value in the (-1,1) range. Notice that, differently from spectrograms, values in measuregrams can be negative; thus, in our customized Whisper the GeLU function has been substituted with the tanh function.

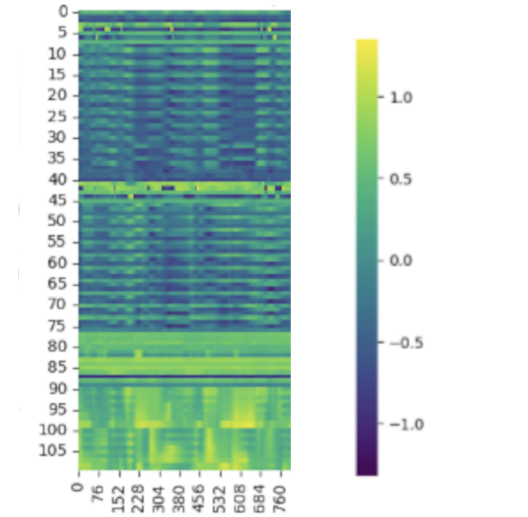

Figure 8: An example of a measuregram

Figure 9: Experimental setting for deaf subject interacting with hearing subject.

Figure 10: Experimental setting for hearing subject interacting with deaf subject.

B Experimental Settings

In Figure 9 is depicted a deaf subject signing in front of the camera, while the monitor shows the avatar. Figure 10 shows the hearing subject using the chat.

Hyperparam.	Autoencoder	Whisper phase 1	Whisper phase 2
per_device_train_batch_size	4	16	16
gradient_accumulation_steps	4	4	4
learning_rate	1e-5	1e-4	1e-4
warmup_steps	100	500	500
max_steps	100000	100000	100000
gradient_checkpointing	false	true	true
fp16	true	true	true
eval_accumulation_steps	4	4	4
evaluation_strategy	steps	steps	steps
per_device_eval_batch_size	4	16	16
predict_with_generate	true	true	true
eval_steps	100	500	500

Table 3: Hyperparameters.

C Hyperparameters

Table 3 shows the most relevant hyperparameters used by the HuggingFace framework, which we adopted for defining and training our models. In particular: the autoencoder, the Whisper model during phase 1 (frozen encoder weights), and the Whisper model during phase 2 (all weights unfrozen).

All trains were split across four 4090 GPUs. The autoencoder train was stopped when the train loss ceased to improve (no overfitting was detected). All other trains were stopped when the evaluation BLEU index started decreasing.

D Kiosk

Figure 11 shows the envisioned kiosk setup we are going to build for field testing. The deaf subject will use this kiosk to interact with the hearing subject.

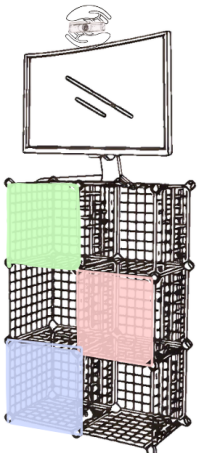

Figure 11: Kiosk for deaf subject for on-the-field test.