MODELING COGNITIVE CONSTRAINTS IN SIGNED-TO-SPOKEN TRANSLATION: A MULTIMODAL NEURAL APPROACH

Bokijonov Beknazar
boqijonovbeknazar2@gmail.com
Asaka District Secondary School No. 42, English teacher
https://doi.org/10.5281/zenodo.17523010

ARTICLE INFO

Qabul qilindi: 25-oktabr 2025 yil
Ma'qullandi: 28-oktabr 2025 yil
Nashr qilindi: 31-oktabr 2025 yil

KEYWORDS

sign language translation, cognitive constraints, multimodal neural networks, visual-manual modality, working memory, low-resource languages, accessibility.

ABSTRACT

In recent years, the field of sign language translation (SLT) has evolved rapidly due to advances in deep learning and multimodal computing. However, despite this progress, translation from signed to spoken languages remains a complex challenge due to fundamental cognitive, linguistic, and modality-specific constraints. This paper investigates how cognitive factors — including working memory limits, multimodal perception, and temporal alignment — affect the process of translating visual-manual languages into spoken forms. The study also explores how multimodal neural networks, especially transformer-based architectures, can model and mitigate these constraints. Using statistical data and comparative analysis from developed countries (the USA, UK, Germany) and emerging contexts, the paper highlights the current state of SLT research, key datasets, and BLEU score benchmarks, providing recommendations for future development of inclusive AI systems that serve deaf and hard-of-hearing communities worldwide.

Introduction

Automatic translation between spoken languages has achieved remarkable success through neural machine translation (NMT) systems. Yet, when it comes to translating sign languages — which are inherently visual-spatial and simultaneous — into spoken or written languages, traditional NMT approaches face fundamental limitations. Sign languages, unlike linear spoken sequences, encode information simultaneously through multiple channels: handshapes, facial expressions, movement trajectories, and spatial positioning.

This creates a cognitive bottleneck for both human interpreters and machine models. While spoken translation typically involves serial auditory processing, sign languages require parallel visual perception and integration, which increases working memory load and cognitive effort. Additionally, most sign languages are low-resource, lacking the large-scale parallel corpora that power spoken-language models.

For example, the PHOENIX-2014T dataset (German Sign Language → German) contains around 80 hours of annotated video, while the CSL-Daily dataset (Chinese Sign Language) offers

approximately 100 hours. In comparison, spoken-language translation datasets such as WMT include millions of sentence pairs. Consequently, the average BLEU score for state-of-the-art SLT systems remains around 25–35, whereas modern NMT systems for spoken languages often exceed 45–50 [1], [5].

In developed countries, government and academic initiatives have expanded resources for sign language research — for example, the ASLLRP corpus in the United States or the BSL Corpus Project in the UK — yet the gap in translation quality persists. This discrepancy stems not only from data scarcity but also from deeper cognitive and linguistic constraints inherent to the modality of sign languages.

Analysis

The process of signed-to-spoken translation involves multiple stages — perception, segmentation, recognition, glossing, and finally, textual or spoken generation. Each stage introduces unique cognitive and computational challenges.

1.Cognitive and Modality Constraints

Sign languages transmit multiple information streams simultaneously. Manual signs, facial gestures, and body movements encode overlapping linguistic, affective, and grammatical data. Translating these into a sequential spoken format requires effective temporal alignment and semantic disambiguation. As De Coster et al. note, "many current approaches are not linguistically grounded and fail to capture the unique simultaneity of sign languages" [1]. Cognitive models show that the human brain processes multimodal communication by synchronizing visual and auditory stimuli in working memory, but artificial systems often lack this adaptive integration mechanism.

2.Working Memory and Processing Load

Translating visual-manual content into linear spoken sentences demands considerable working memory. A signer may express several ideas concurrently — for instance, indicating subject reference with the left hand while using the right hand for a predicate sign, and simultaneously applying facial expressions for adverbial meaning. Such simultaneity must be disassembled into linear units for text generation. Studies in psycholinguistics and cognitive neuroscience suggest that typical human working memory can handle 7 ± 2 informational chunks at a time [3]. Neural architectures inspired by this principle — such as limited-sequence transformers with attention prioritization — can more accurately model human-like cognitive constraints.

3.Data Scarcity and Cross-Linguistic Variation

Most sign languages (e.g., Uzbek Sign Language, Kazakh Sign Language, Azerbaijani Sign Language) remain under-documented. In contrast, ASL and BSL enjoy comparatively larger resources and research investments. According to statistical reviews, 80% of global SLT research focuses on only four sign languages: ASL, BSL, DGS (German Sign Language), and CSL [2]. This imbalance limits generalization across linguistic and cultural contexts. Furthermore, each sign language has a distinct grammar and lexicon, meaning that transfer learning from one language to another is only partially effective.

4Multimodal Neural Architectures

Advanced models such as HST-GNN (Hierarchical Spatio-Temporal Graph Neural Network) and SLTUNET [7] attempt to integrate spatial, temporal, and semantic dimensions into unified representations. HST-GNN treats the signer's joints and facial landmarks as graph nodes, enabling fine-grained tracking of body and facial movements. SLTUNET, in turn, performs joint optimization of sign-to-gloss and gloss-to-text translation, improving overall BLEU by up to 20%.

Another recent advancement is context-aware translation, where the system encodes prior discourse and conversational context. For example, Mercanoğlu Sincan et al. [6] demonstrated that including a discourse encoder nearly doubled BLEU-4 scores on large-domain sign translation benchmarks.

5.International Research Trends and Statistics

Statistical data show that the global market for AI-based accessibility technologies is projected to reach USD 12.4 billion by 2030, with sign language recognition and translation comprising nearly 15% of this market [4]. In the U.S., more than 500,000 ASL users benefit indirectly from SLT-based educational tools, while the European Union has funded over 20 multimodal translation projects since 2020. Nevertheless, error rates remain high — particularly in spontaneous, conversational signing — with accuracy between 60–70%, compared to 90–95% for text-based translation systems.

To mitigate these gaps, multimodal neural approaches must go beyond simple video recognition. Incorporating linguistically-informed features (such as sign glosses, classifier structures, and mouthings), attention mechanisms inspired by human cognition, and cross-modal embeddings can improve alignment and reduce semantic loss.

Conclusion

Signed-to-spoken translation represents one of the most interdisciplinary challenges at the intersection of linguistics, cognitive science, and artificial intelligence. Cognitive constraints — particularly those related to working memory, simultaneous modality processing, and context retention — remain the primary barrier to achieving parity with spoken-language translation.

A multimodal neural framework that jointly models visual, spatial, and linguistic information offers a promising path forward. By integrating attention-based temporal alignment, context encoding, and transfer learning from high-resource spoken languages, SLT systems can achieve higher naturalness and semantic fidelity.

In developed countries, such systems are already being piloted in education, broadcasting, and public services. However, expanding their global reach requires not only technical innovation but also inclusive data collection in collaboration with deaf communities. Future research should prioritize large-scale spontaneous sign language corpora, cross-linguistic benchmarking, and cognitive-inspired architectures that mimic human multimodal processing. These advances will bring us closer to equitable communication systems that bridge the gap between signed and spoken worlds.

References:

De Coster, M., Shterionov, D., Van Herreweghe, M., Dambre, J. (2022). Machine Translation from Signed to Spoken Languages: State of the Art and Challenges. arXiv:2202.03086.

Camgöz, N. C. (2020). Neural Sign Language Recognition and Translation. PhD thesis, University of Surrey.

Kita, S. (2023). Gesture links language and cognition for spoken and signed. Frontiers in Psychology.
Mohammed, R., Aljarrah, I., Al-Ayyoub, M., Fadel, A. (2025). Multimodal Multisource Neural Machine Translation. Computation, 13(8):194.
Jiang, Z., Moryossef, A., Müller, M., Ebling, S. (2022). Machine Translation between Spoken and Signed Languages Represented in SignWriting. arXiv:2210.05404.
Mercanoğlu Sincan, Ö., Camgöz, N. C., Bowden, R. (2023). Is Context All You Need? Scaling Neural Sign Language Translation to Large Domains of Discourse. arXiv:2308.09622.
Kan, J., Hu, K., Hagenbuchner, M., Tsoi, A. C., Bennamoun, M., Wang, Z. Y. (2021). Sign Language Translation with Hierarchical Spatio-Temporal Graph Neural Network. arXiv:2111.07258.