Real Time English Speech to British Sign Language Translation for Accessible Banking

Abhishek Bharadwaj Varanasi
TCS Research
Tata Consultancy Services Ltd.
Kolkata, India
varanasi.abhishek@tcs.com

Manjira Sinha
TCS Research
Tata Consultancy Services Ltd.
Kolkata, India
sinha.manjira@tcs.com

Tirthankar Dasgupta
TCS Research
Tata Consultancy Services
Kolkata, India
iamtirthankar@gmail.com

Charudatta Jadhav
TCS Research
Tata Consultancy Services Limited
Mumbai, Maharashtra, India
charudatta.jadhav@tcs.com

Abstract

This paper presents a mobile-based framework for real-time English Speech-to-British Sign Language (BSL) translation, focused on accessible banking. Using a transformer based encoder-decoder model, the system converts English speech into BSL gloss HamNoSys for avatar animation, optimized for handheld devices. A parallel dataset was created for experimentation, covering multiple domains. The model achieved a 73.87% ROUGE-L score in speech-to-gloss HamNoSys translation and underwent user-based evaluation with BSL experts, offering valuable insights into its effectiveness and cross-domain performance.

CCS Concepts

• **Human-centered computing** → **Accessibility design and evaluation methods**.

Keywords

Sign Language, Machine Translation, Avatar Animation

ACM Reference Format:
Abhishek Bharadwaj Varanasi, Manjira Sinha, Tirthankar Dasgupta, and Charudatta Jadhav. 2025. Real Time English Speech to British Sign Language Translation for Accessible Banking. In *30th International Conference on Intelligent User Interfaces Companion (IUI Companion '25), March 24–27, 2025, Cagliari, Italy*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3708557.3716351

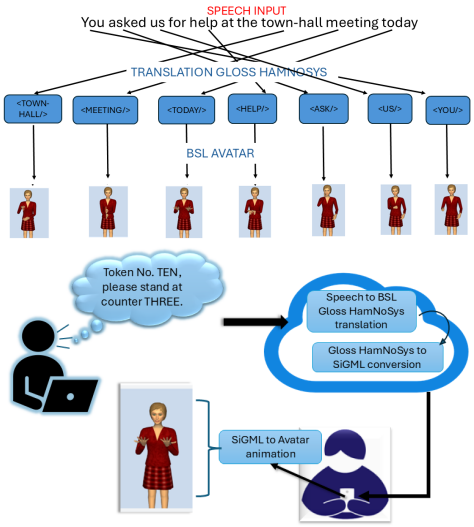


Figure 1: (a) Illustration of speech to British Sign Language (BSL) avatar generation using glosses as intermediate step. (b) Illustration of speech to BSL Avatar application flow for mobile devices.

1 Introduction

This paper presents a mobile-based English Speech-to-BSL translation system designed to improve communication for Deaf and hard-of-hearing individuals, especially in mobile banking. The system translates speech into BSL in real-time using animated avatars through a two-phase process: converting speech to gloss HamNoSys tags and generating signs with avatars (Figure 1(a)). Sign Languages (SLs) are complex, involving both manual and non-manual components, which complicates translation (e.g., facial expressions, hand shapes) [16]. British Sign Language follows a topic-comment structure with object-subject-verb ordering, with the topic placed first, followed by its comment. Machine translation systems for SL-to-text and text-to-SL often rely on glosses [3, 4]. Recent advancements include Neural Machine Translation (NMT) and motion graphs for video generation [13], and text-to-gloss methods for generating SL videos [9]. The system we propose is lightweight and has applications in fields like healthcare and education (Figure 1(b)).

IUI Companion '25, Cagliari, Italy

ACM ISBN 979-8-4007-1409-2/25/03
https://doi.org/10.1145/3708557.3716351

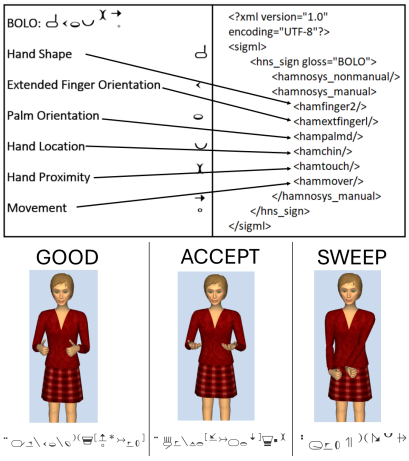


Figure 2: (a) HamNoSys to SiGML Examples [10]. (b) HamNoSys Examples.

2 Speech to BSL Gloss HamNoSys translation model

We fine-tuned five different transformer based [14] pre-trained speech-to-text models (SEW-D[12], Wav2Vec2-BERT[17], Wav2Vec-mBART[1], SpeechT5[2], and Speech2Text[15]). The HamNoSys tags are considered as English text while fine-tuning the models.

2.1 Transformer based speech-to-text models

The pre-trained models fine-tuned in this research are based on the transformer encoder-decoder architecture, commonly used in speech and language processing. The encoder extracts features from raw audio, while the decoder generates text outputs, utilizing the Connectionist Temporal Classification (CTC) loss function[7] for speech recognition, which aligns audio frames to text sequences without pre-segmented data. The CTC loss is defined as: $\mathcal{L}_{CTC} = -\log \sum_{a \in \mathcal{A}} P(a|x)$ where x is the input feature sequence, a is a valid alignment, and $\mathcal{A}$ represents all possible alignments. Models like SEW-D, Wav2Vec2-BERT, Wav2Vec-mBART, SpeechT5, and Speech2Text integrate enhancements to improve efficiency, multilingual support, and accuracy in speech recognition and translation tasks across various languages and domains.

3 HamNoSys: A sign language notation system

HamNoSys is a standardized system for transcribing sign language gestures, focusing on essential features such as hand shape, location, and movement, using a unique set of symbols. It captures both basic and complex hand shapes, orientations (relative or absolute), and precise locations on the body, along with movement paths and modifiers like repetition or finger-play. While primarily documenting hand actions, HamNoSys also incorporates non-manual elements like facial expressions and upper body movements, providing an effective way to transcribe dynamic and detailed sign language gestures.

4 BSL Avatar Animation

JASigning [6] animates avatars using 3D meshes and SiGML files based on HamNoSys notation, defining hand shape, orientation, location, and movement for precise control. Movement styles like "targeted" and "tense" are modeled using a damped harmonic motion equation $x'' + k'x' + kk'x = kk'x_t$ for realistic gesture transitions. Our application integrates a fine-tuned speech-to-text model that translates English speech into a BSL gloss HamNoSys sequence, which JASigning uses to animate the avatar in real-time (Figure 2(b)).

5 Experiments

5.1 Data, Fine-tuning, and Evaluation

We created a set of 6096 text-BSL gloss pairs, including 3597 banking-related pairs, with input from domain experts (see Appendix A for sample data). These were manually translated into BSL gloss by experts. For fine-tuning, we mapped glosses to HamNoSys tags based on the mappings from [5].During the fine-tuning process, English speech is given as input to generate HamNoSys tag sequences as output. The model was trained on 4877 samples, with 609 for validation and 610 for testing, fine-tuning five transformer model decoders on an A100 GPU (see Appendix B for hyperparameters). Pre-trained speech encoders were used to maintain generalizability. In addition to standard MT metrics like ROUGE-L [8] and BLEU [11], we propose a modified BERTScore [18], where we aggregate word embeddings of HamNoSys tags to compute the cosine similarity for sequence evaluation.

6 Results

6.1 Speech to BSL gloss HamNoSys evaluation

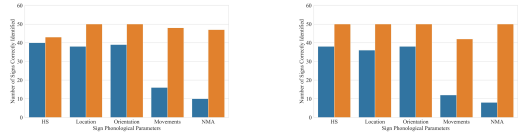
From Table 1, among the models we have tested, the SpeechT5 model gave the best performance with 78.85% ROUGE-L score upon fine-tuning, but several challenges persist. Sometimes the HamNoSys tags that are generated are not valid and hence a few gestures in the final animation might be missed. The gloss ordering is sometimes incorrect which is further lower the BLEU-3 and BLEU-4 scores (refer Appendix C for approach on how we may overcome these issues).

Table 1: Model Evaluation upon fine-tuning averaged over two different train-test splits with same test ratio

Model	SEW-D	Wav2Vec2BERT	Wav2Vec-mBART	FairseqS2T	SpeechT5
ROUGE-L	12.35%	32.79%	60.92%	70.82%	**73.87%**
BLEU-1	8.23%	29.21%	59.35%	71.22%	**78.34%**
BLEU-2	5.45%	20.13%	45.32%	57.18%	**62.30%**
BLEU-3	3.98%	14.87%	36.67%	47.45%	**52.97%**
BLEU-4	2.75%	10.86%	30.64%	40.83%	**45.97%**
Modified BERTScore	38.49%	53.86%	72.22%	73.13%	**77.14%**

6.2 Subjective Evaluation

In addition to automatic evaluation, we conducted a user-based evaluation with BSL experts to assess the proposed machine translation



(a) Avatar Based Signs (b) Video Based Signs

Figure 3: Comparing the sign understandability between avatar based and video signs for both one-handed and two-handed signs.

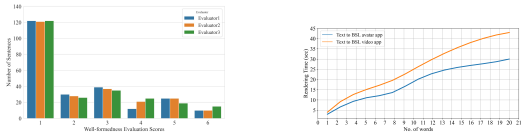


Figure 4: (a) Well-formedness scores of the output generated sentences. (b) Sign rendering time (in sec) vs Number of words in the input text.

engine's performance. The evaluation focused on two criteria: sign understandability and well-formedness of the output BSL sentences.

6.2.1 Sign Understandability Test. The sign understandability test evaluates the accuracy of avatar-based BSL. We selected 100 signs (50 single-handed, 50 double-handed) in both avatar and video formats, assessing hand shapes, orientation, movements, and non-manual components. As shown in Figure 3, video-based signs generally perform better, especially for complex gestures, while avatar-based signs excel in dynamic directional movements. The second evaluation level assesses sentence well-formedness, including grammar, word usage, and morphological accuracy with categories like no corrections, minor errors, word order issues, tense/number errors, and missing phrases or subjects.

Figure 4(a) summarizes the well-formedness scores assigned by each of the experts. The X-axis shows the evaluation score range(1-6) for wellformedness and the Y-axis shows the number of sentences that received the score. The graph in Figure 3 shows that the score range of above 4 is given to the least number of sentences while most sentences scored below the range of 4, which is desirable.

6.3 Analyzing memory requirements and rendering time

We compare the memory requirements and rendering times of two applications: speech-to-BSL video and speech-to-BSL avatar. The speech-to-BSL video application requires about 7 GB of disk space, which increases with the size of the gloss-video dictionary, while the speech-to-BSL avatar application requires only 910 MB, making it more lightweight and mobile-friendly. As shown in Figure 4(b), the rendering time for the text-to-BSL video app is higher than that for the avatar app, and both rendering times increase almost linearly with the number of words in the input text, with the gap between the two applications widening as the text lengthens.

7 Conclusion

This paper presents a framework that uses pre-trained transformer-based speech-to-text models for real-time translation of English speech into British Sign Language (BSL) where the best-performing speech-to-gloss HamNoSys model achieved a 73.87% ROUGE-L score. The system converts speech into BSL gloss sequences, which animate avatars for communication via mobile devices and it was subjected to evaluation by BSL experts. The technology, integrated into a lightweight app, allows Deaf users to receive real-time banking notifications in BSL, improving accessibility in areas like banking, healthcare, and education, and promoting greater independence for the Deaf community.

References

[1] Meta AI. 2022. facebook/wav2vec2-xls-r-300m-en-to-15. https://huggingface.co/facebook/wav2vec2-xls-r-300m-en-to-15.

[2] Junyi Ao et al. 2022. SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing. In *Proceedings of the ACL*.

[3] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7784–7793.

[4] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. 2017. Subunets: End-to-end hand shape and continuous sign language recognition. In *Proceedings of the IEEE international conference on computer vision*. 3056–3065.

[5] Eleni Efthimiou, Stavroula-Evita Fontinea, Thomas Hanke, John Glauert, Rihard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos, and François Goudenove. 2010. Dicta-sign–sign language recognition, generation and modelling: a research effort with applications in deaf communication. In *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*. 80–83.

[6] Ralph Elliott, John RW Glauert, JR Kennaway, Ian Marshall, and Eva Safar. 2008. Linguistic modelling and language-processing technologies for Avatar-based sign language presentation. *Universal access in the information society* 6 (2008), 375–391.

[7] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*. 369–376.

[8] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[9] Amit Moryossef, Mathias Müller, Anne Göhring, Zifan Jiang, Yoav Goldberg, and Sarah Ebling. 2023. An open-source gloss-based baseline for spoken to signed language translation. *arXiv preprint arXiv:2305.17714* (2023).

[10] Carolina Neves, Luísa Coheur, and Hugo Nicolau. 2020. HamNoSyS2SiGML: translating HamNoSys into SiGML. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 6035–6039.

[11] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

[12] Sangjun Seo et al. 2021. Performance-Efficiency Trade-offs in Unsupervised Pre-training for Speech Recognition. In *ICASSP*.

[13] Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. 2020. Text2Sign: towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision* 128, 4 (2020), 891–908.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[15] Changhan Wang et al. 2020. Fairseq S2T: Fast Speech-to-Text Modeling with Fairseq. In *Proceedings of the AACL*.

[16] Ulrike Zeshan. 2003. Indo-Pakistani Sign Language grammar: a typological outline. *Sign Language Studies* (2003), 157–212.

[17] Changhan Zhang et al. 2023. Seamless: Multilingual Expressive and Streaming Speech Translation. *arXiv preprint arXiv:2303.04161* (2023).

[18] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).

[19] Zijun Zhang. 2018. Improved adam optimizer for deep neural networks. In *2018 IEEE/ACM 26th international symposium on quality of service (IWQoS)*. Ieee, 1–2.

A Sample Data

Table 2 has sample data points from the data we have created

Table 2: Sample text to BSL gloss data. The English texts are text versions of the English speech. Each gloss in the gloss sequence is mapped to its corresponding HamNoSys tags for machine translation from speech.

English Text	BSL Gloss Sequence
Would you like to schedule an appointment to meet with a financial advisor?	APPOINTMENT SCHEDULE MEET FINANCIAL ADVISOR YOU WANT?
A basic savings account it is. I'll need you to fill out this form with your personal details.	SAVINGS ACCOUNT BASIC, IT. FORM THIS FILL-OUT NEED YOU, YOUR PERSONAL DETAILS.
Can I deposit a check via mobile banking?	DEPOSIT CHECK MOBILE BANKING CAN I?
How can I assist you in updating your contact information for your account?	HOW CAN I ASSIST YOU IN UPDATING YOUR CONTACT INFORMATION FOR YOUR ACCOUNT ?
Your credit check came back clear, so we can proceed with finalizing your account.	CREDIT CHECK YOUR FINISH, CLEAR. ACCOUNT YOUR FINALIZE CAN PROCEED WE.

B Hyperparameters

Table 3 has the hyperparameters used for model fine-tuning.

Table 3: Hyperparameters for the model.

Hyperparameter	Value
Training epochs	10
Maximum learning rate	5×10^{-5}
Weight Decay	1×10^{-6}
Batch size	1
Gradient accumulation steps	2
Attention Dropout	0.1
Optimizer	Adam [19]

C Approaches to improve the model predictions

- Instead of considering the HamNoSys gloss sequence as English text, we can try by replacing vocabulary of the language decoder with all the existing HamNoSys tags and then fine-tune the model so that the HamNoSys tags won't be invalid.
- To improve the correct ordering of the predicted sequence, we may consider this as a two step process where first the speech is converted to gloss sequences and then the glosses are mapped to their corresponding HamNoSys tags. But for this we need to have a gloss-HamNoSys dictionary with us which is not limited to just 1000 glosses like we currently have [5].