［研究助成（A）］

基盤モデルに基づく手話対話システムの構築

2241011

研究代表者 （助成金受領者）	東京科学大学 （旧：東京工業大学）	准教授	川　上　　玲
共同研究者	東京科学大学	教　授	中　臺　一　博
	東京科学大学	教　授	田　中　正　行
	東京科学大学	准教授	船　越　孝太郎

［研 究 の 目 的］

日本には約６万から７万人の手話ユーザーがおり，彼らへの技術的な支援の充実が求められている。本助成では，基盤モデルを活用して手話の解析を行い，言語的理解を深めることで，手話を用いる方々をサポートする対話システムの構築に向けた基礎技術の確立を目指す。

手話は，主に聴覚障がい者同士や聴覚障がい者とのコミュニケーション手段として発達し，音声言語とは異なる言語体系を持つ自然言語である。手話は聴覚障がい者がありのままで自由にコミュニケーションができる手段であり，聴覚障がいがある子どもの言語的・社会的発達にも重要な役割を果たすため，できるだけ早期での習得が推奨されている。一方で，手話を話せる（健聴）者は限られているため，たとえば，教育現場や公的機関，道案内など手話が必要とされる場面で，工学的技術がサポートできる部分は大きい。ろう学校の減少に伴い日本手話は消滅の危機にあり，環境整備が急務である。

手話認識は主に自然言語処理よりも画像認識の分野で発展してきたが，これまではグロスという手話セグメント（手話における単語に相当）の認識が主であり，手話ビデオからテキストへの対訳がついたコーパスはドイツ手話やアメリカ手話などの小規模なものに限られていた。コーパスが小規模なほど翻訳性能が低くなる傾向にあるが，手話データセットで大規模なものは少なく，事前知識を活用しなければ翻訳の性能を実用段階に近づけられない。

そこで，本研究では，手話のコーパス不足を基盤モデル（主に大規模言語モデル）の知識を活用することで補い，手話と音声言語間の変換における基盤技術の構築を目指す。大規模言語モデルのメカニズムの理解は研究の途上であるが，ある程度言語の意味空間を獲得していると考えられ，意味空間を介したテキスト空間と手話の相互変換が学習しやすくなると予想される。

［研究の内容，成果］

手話研究に関する文献調査

手話翻訳，手話生成，および，それらにおける基盤モデルの活用について，現時点での課題を明確化するため，大規模な文献調査を行った。結果を調査論文としてまとめ，出版した［成果1］。論文の特色は，手話の単語認識，翻訳，生成について網羅的に調査している点と，基盤モデルを用いた研究に関する章を設けている点の二点である。

主な知見として，非言語的な手がかりを捉えきれないグロス（単語的）アプローチの限界や，データセット間の大きなばらつきが，頑健な

SLP（Sign Language Production）システムの開発を妨げていることを示した。さらに，評価指標の一貫性がないことを指摘し，手話と言語の両方の特性を考慮した標準化アプローチの必要性を強調した。最後に，既存のデータセットについて，その関連性と手話研究を進展させる可能性を評価した。

大規模言語モデルについては，BERT-styleとGPT-styleの二種類のアーキテクチャについて調査し長短を論じた。また，言語モデルを利用するものが主流であるが，大規模視覚言語モデルなどマルチモーダルなモデルの利用の可能性についても議論した。手話生成では言語モデルを利用したものはまだ事例がないことから，一般的な動作生成やビデオ生成について調査し，方法論についてまとめた。

大規模言語モデルを用いた手話生成

手話研究の調査［成果1］から，手話翻訳と手話生成が独立して研究されていることや，手話生成は手話翻訳に比べて半分以下の研究事例しかないことが判明した。一方で，相互翻訳のためには手話生成に取り組む必要がある。そこで，大規模言語モデル（LLM）を用いた手話生成に取り組んだ。成果は［成果2］として発表予定である。内容を以下に示す。

これまでの手話生成はグロスを用いるのが一般的であるが，グロスと手話表現が一対一対応でないことから，グロスの正確さに性能が律速されてしまう。また，動作生成では動作を表現するトークンをLLMに推論させ，そのトークンを別のデコーダで動作に変換させるアプローチが性能がよいが［成果1］，どのように手話動作のトークンを学習すればよいかは自明でない。また，1〜2Tパラメータ規模のLLMはある程度の手話知識を有していることが事前調査から分かっていたが，これの手話生成への活用は著者らの知る限り試験されていなかった。

そこで，LLMが持つ手話知識と強力な推論能力を活用する手話生成手法 Teach Me Sign（TEAM-Sign）を提案した［成果2］。TEAM-Signは，LLMを活用し，自己回帰的に手話ポーズ列を生成する。具体的には，1〜2TパラメータLLM（GPT-4o）にプロンプトを与えて，補助シーケンスを生成させる。これを追加学習する数Bパラメータ規模のLLMに補助的に与え，手話とポーズ列の組からなる訓練データから教師あり学習を行う。グロスは用いずにテキストから直接手話を生成するアプローチを選択する。

補助シーケンスには，手話の語順や各単語を表現するのにかかる時間といった情報が含まれており，追加学習用LLMに正確なステップを教える役割を果たす。手話トークンは，VQVAE（Vector-Quantized Variational Auto Encoder）を用いて，ポーズの再構成を学習させることで獲得する。VQVAEによるトークン学習は動作生成の既存研究に倣っている。

図1に手法の概要を示す。事前に，動画から抽出されたポーズ列の再構成をVQVAEで学習しておき，ポーズからトークンのコードを生成するエンコーダ（z_e），および，コードからポーズを生成するデコーダ（z_d）を得ておく。追加学習LLM（図1のLLaMA）への入力は，翻訳テキスト（図1のText部分），補助シーケンス（図1のAssistance部分），プロンプト（You are a sign language expert. Generate a sequence of number tokens to express the following sentence in {American/German} sign language.），の三つである。

追加学習用LLMは，正解ポーズ列に対応するコードを予測するように交差エントロピー（CE）損失で学習を行う。学習にはLoRA（Low-rank adaptation）を用い，LLMの重みを凍結させつつ，全体の数％程度の学習可能パラメータを追加してそれのみを学習させることで学習を効率化する。この学習の際は，VQVAEのエンコーダ，デコーダ，コードは凍結されている。

Phoenix14T（ドイツ手話），および，How2Sign

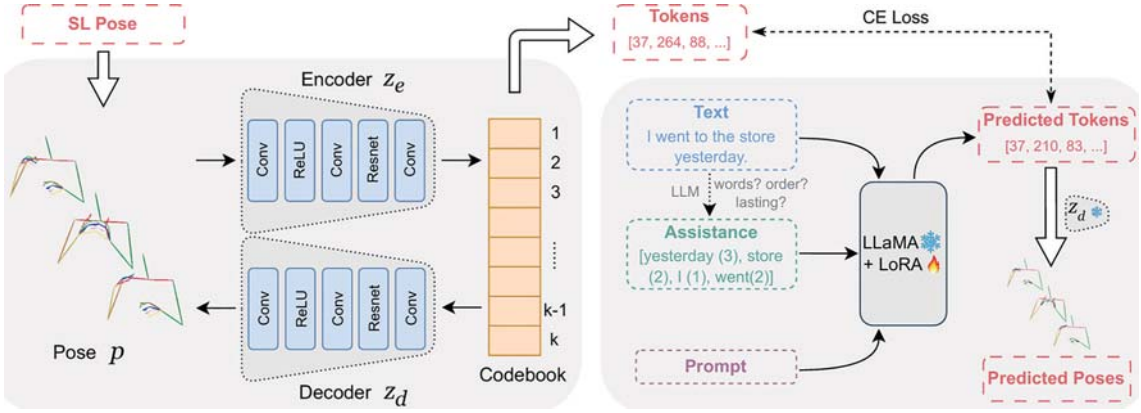

図1　LLM を活用した手話生成手法の概要

Table 1　Quantitative results on Phoenix14T dataset

	DTW-MJE	BLEU-1	BLEU-2	BLEU-3	BLEU-4
PT（w/o data augmentation）	0.1383	9.583	3.787	1.642	0.692
PT	0.1276	11.388	5.782	3.569	2.077
Proposed - LLaMA3（w/o assistance）	0.1204	12.861	6.105	3.719	2.636
Proposed - LLaMA3	0.1056	**13.366**	**6.462**	**4.239**	**3.151**
Proposed - Qwen2（w/o assistance）	0.1051	12.950	6.129	3.865	2.813
Proposed - Qwen2	**0.1038**	13.022	6.131	4.136	3.107
Groundtruth	0.0000	30.730	20.995	15.522	12.298

Table 2　Quantitative results on How2Sign dataset

	DTW-MJE	BLEU-1	BLEU-2	BLEU-3	BLEU-4
PT（w/o data augmentation）	0.1905	4.609	2.351	1.108	0.561
PT	0.1733	5.490	2.744	1.262	0.628
Proposed - LLaMA3（w/o assistance）	0.1406	9.963	5.044	**2.476**	1.058
Proposed - LLaMA3	**0.1371**	10.419	5.264	2.348	1.116
Proposed - Qwen2（w/o assistance）	0.1422	9.971	5.041	2.329	1.097
Proposed - Qwen2	0.1397	**10.532**	**5.278**	2.415	**1.140**
Groundtruth	0.0000	11.230	5.469	2.500	1.228

（アメリカ手話）という二つのデータセットで定量評価を行った。先行研究に倣い，DTW-MJE，BLUE-n を評価指標として用いた。前者は，二つの時系列をできる限り揃えたあとに，真値のポーズとの関節間の差の総和を計算する。BLUE-n は生成したポーズを機械翻訳に通し，翻訳された文章の正確さを評価する。生成したポーズ列を翻訳するモデルが公開されていなかったため，これも著者らで実装した。すべてのコードは研究室のウェブサイトで公開予定である。

定量評価の結果を Table 1，および Table 2 に示す。PT（Progressive Transformer）の行は先行研究の結果を示しており，提案法では，追加学習 LLM に LLaMA3 を用いた場合，Qwen2 を用いた場合の双方で，PT を上回る結果を得た。図 2 は，Phoenix14T と How2Sign での定性評価を示す。視覚的にも，提案法の方が，PT よりも真値の動きに近い。

一方で，翻訳結果の BLUE-n スコアは BLUE-4 で最大で 3 程度であり，実用には不足している。これは，翻訳機の性能によるところもあるが，データセットの性能によって指の表現の正確さが不足している点や，表情が利用できない点も起因している。今後，学習をより大規模化させる必要がある。

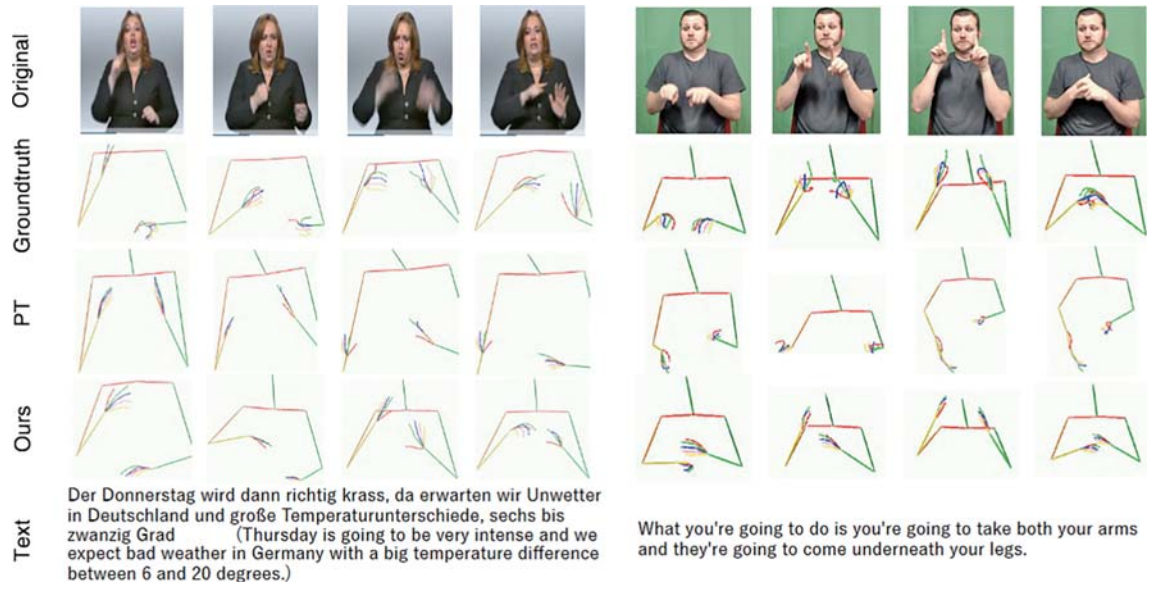

図 2　Phoenix14T（左）と How2Sign（右）　データセットにおける定性評価

その他周辺技術の確立

多言語手話翻訳において，Sign2（LID＋Text）という新しい多言語グロスフリーモデルを提案した［成果 5］。トークンレベルの手話言語識別（Sign2LID）と手話からテキストへの CTC アラインメント（Sign2Text）を組み合わせることで，異なる手話言語間の衝突やアラインメントの困難さを克服している。このモデルは，一対一，多対一，多対多の様々な翻訳シナリオをサポートし，10 種類の手話言語に対応している。

Table 3 に示す通り，Phoenix14T，CSL-Daily で既存の SOTA 手法を上回る性能を示した。また，SP-10 では SOTA 手法に匹敵する性能を示した。この研究は統一された手話基盤モデルの構築に向けた重要な一歩と位置づけられる。

［成果 5］は，CTC アラインメント［成果 6］に関する知見が活かされている。［成果 6］では，手話ビデオと音声言語間のアラインメントを CTC と Attention の組み合わせによって達成する。CTC と Attention の利用は，一般的な機械翻訳からヒントを得ている。具体的には，エンコーディングの際，グロスの CTC 損失をエンコーダの出力で計算し，テキストの CTC 損失をテキストに翻訳するためのエンコーダの出力で再度計算する。このように段階的に CTC 損失を計算することで，手話における異

Table 3　Experimental results on PHOENIX14T and CSL-Daily dataset for gloss-free SLT（one-to-one SLT）

Methods	PHOENIX14T				CSL-Daily			
	Dev		Test		Dev		Test	
	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE
Gloss-free								
NSLT＋Luong (Camgoz et al., 2018)	10.00	32.60	9.00	30.70	7.96	34.28	7.56	34.54
CSGCR (Zhao et al., 2022)	15.08	38.96	15.18	38.85	–	–	–	–
GFSLT-VLP (Zhou et al., 2023)	22.12	43.72	21.44	42.49	11.07	36.07	11.00	36.44
Sign2GPT (Wong et al., 2024)	–	–	22.52	48.90	–	–	15.40	42.36
Fla-LLM (Chen et al., 2024)	–	–	23.09	45.27	–	–	14.20	37.25
SignLLM (Gong et al., 2024)	**25.25**	47.23	23.40	44.49	12.23	39.18	**15.73**	39.91
Baseline	22.59	49.88	22.52	49.85	12.23	36.39	11.76	36.25
Ours w TxtCTC	24.18	**51.74**	**24.23**	**50.60**	**13.66**	**39.33**	14.18	**40.00**

なる長さの単語や語順の変更に対応できる。定量評価では，特に CSL-Daily で高い性能を示し，SOTA を達成した。

ほかにも，手話翻訳のためのデータ拡張［成果 7］，手話生成における，Multimodal Gated Attention の利用［成果 10］や，実時間で手話生成を行う手法［成果 9］，動きのフレーム補間を行う手法［成果 4］を開発した。ろう者を招いて，会話のデータ撮影も行い，データセット構築の準備を進めている。学習における基礎技術としては，損失関数が平坦である局所解を探すオプティマイザ向けの交差エントロピー損失を代替する損失関数を提案した［成果 11］。

［成果の発表，論文など］

［1］ Tan, S., Khan, N., An, Z., Ando, Y., Kawakami, R., & Nakadai, K.（2024）. A Review of Deep Learning-based Approaches to Sign Language Processing. Advanced Robotics, 38(23), 1649-1667

［2］ An, Z. Kawakami, R.（2025）. Teach Me Sign: Stepwise Prompting LLM for Sign Language Production. In Proc. IEEE International Conference on Image Processing（ICIP）. To appear, Oct. 2025, Anchorage, USA.

［3］ S. Tan, K. Itoyama, & K. Nakadai.（2024）. Advancing Human-Computer Interaction: End-to-End Sign Language Translation. ヒューマンインタフェース学会論文誌，26(4) pp. 391-398.

［4］ N. Khan, S. Tan, K. Itoyama, K. Nakadai.（2024）. Motion Inbetweening Based on Body Parts Integration for Sign Language Generation. ヒューマンインタフェース学会論文誌，26(4) pp. 431-442.

［5］ Tan, S., Miyazaki, T., Nakadai, K.（2025）. Multilingual Gloss-free Sign Language Translation: Towards Building a Sign Language Foundation Model. In Proc. of the 63rd Annual Meeting of the Association for Computational Linguistics（ACL）（Acceptance rate ~21%）

［6］ Tan, S., Miyazaki, T., Kahn, N., & Nakadai, K.（2025）. Improvement in Sign Language Translation Using Text CTC Alignment. In Proc. of the 31st International Conference on Computational Linguistics（COLING）, pp. 3255-3266, Abu Dhabi, UAE, Jan. 2025.（Acceptance rate ~28%）

［7］ S. Tan, T. Miyazaki, K. Itoyama, and K. Nakadai.（2024）. SEDA: Simple and Effective Data Augmentation for Sign Language Understanding. In Proc. LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources, pp. 370-375, Torino, Italia. ELRA and ICCL.

［8］ N. Khan, S. Tan, K. Nakadai.（2025）. Towards Online Sign Language Expression for Real-Time Human-Robot Interaction, IEEE International Conference on Robot and Human Interactive Communication（RO-MAN）, to appear（Aug. 2025）.

［9］ N. Khan, B. Wu, C. T. Ishi, and K. Nakadai,（2025）. MultiGAU: Real Time Sign Language Generation using Multimodal Gated Attention, The 38th International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems（IEA/AIE 2025）, to appear（July 2025）.

［10］ Khan, N. & Nakadai, K.（2025）. End to End Text to Sign Language Generation using MultiGAU, IEEE International Conference on Multimedia & Expo（ICME 2025）, To appear, July 2025.

［11］ Ratchatorn, T., Tanaka, M.（2025）. Adaptive Adversarial Cross-Entropy Loss for Sharpness-Aware Minimization. In Proc. IEEE International Conference on Image Processing（ICIP）. To appear, Oct. 2025, Anchorage, USA.