E-042

音声信号を入力とする日本語から手話への機械翻訳
Machine Translation from Japanese to Sign Language Using Audio Signal Inputs

木下 光太郎† 宮崎 太郎† 金子 浩之†
Kotaro Kinoshita Taro Miyazaki Hiroyuki Kaneko

1. はじめに

手話母語話者への情報保障を目指して、日本語を手話に翻訳し、CG(Computer Graphics)で提示する手話 CG の研究を進めている[1]。手話 CG 生成システムでは、まず日本語テキストを機械翻訳モデルに入力し、手話の意味を表す単語を順番に並べた手話単語列へと翻訳する。この手話単語列に対応する手話単語モーションを CG キャラクターに付与することで、手話 CG の提示を実現している。

従来の手話 CG 生成システムは、日本語のテキストを入力として手話単語列へ翻訳しており、音声信号を入力とする翻訳は未対応であった。音声には、テキストには含まれない間や抑揚などのパラ言語情報が含まれるため、音声から手話単語列へ直接翻訳することが可能になれば、より豊かな言語情報を反映した手話 CG の実現が期待される。そこで本稿では、音声信号を入力とする手話翻訳モデルを構築し、その翻訳精度について検証を行った。

2. 音声を入力とする機械翻訳

従来のテキスト入力による機械翻訳[2]では、Transformer[3]をベースとした Encoder-Decoder モデルが用いられている。この翻訳モデルの概要を図 1 に示す。音声信号を入力とする場合、従来モデルにおける Transformer の Encoder を音声認識モデルの Encoder に置き換えることで、音声入力に対応した機械翻訳モデルを構築できる。音声認識モデルの Encoder には事前学習済みの facebook/wav2vec2-large-xlsr-53[4]を使用した。なお、本稿における音声信号の入力は、逐次的に処理されるものではなく、学習および評価の対象となる音声信号の一文全体を一括で入力・処理するものである。

3. 実験

音声信号を入力とする手話翻訳モデルの学習と評価を行った。学習データには NHK で放送されている「手話ニュース」を基に構築された、手話単語列・音声信号・日本語テキストの対訳データセットである手話ニュースコーパスを使用した。このコーパスを学習データ 8,472 文、開発データ 1,059 文、評価データ 1,059 文となるようにランダムに分割した。まず音声認識モデルのファインチューニングを行い、評価データに対する CER(Character Error Rate) が 6.94 %であることを確認した。この音声認識モデルの Encoder を活用し、以下の 4 条件で手話翻訳モデルの学習を行い、翻訳精度を比較した。

① 音声認識・テキストー手話翻訳

従来手法の組み合わせとして、音声信号をファインチューニング済み音声認識モデルに入力して日本語テキストを出力した後、それを従来のテキストー手話翻訳に入力した。この条件のモデルの概要を図 2 左上に示す。

† NHK

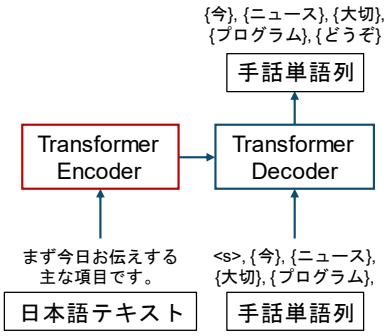

図 1 テキスト入力による手話翻訳モデルの概要

② 音声ー手話翻訳

提案法のベースラインモデルとして、第 2 章で紹介した従来の機械翻訳モデルの Encoder を音声翻訳用の Encoder に置き換える構成を用いた。この条件のモデルの概要を図 2 右上に示す。ここでは音声認識モデルの Encoder 出力を、出力層を介さずに直接 Transformer Decoder に接続することで、音声認識誤差の影響を排除しつつ手話単語列への翻訳を行う。

③ 音声ー手話翻訳（日本語音声認識マルチタスク）

音声認識モデルの Encoder の精度向上を図り、②の構成に音声認識 Encoder の出力層を追加し、手話単語列と日本語テキストを同時に出力するマルチタスク学習を行った。この条件のモデルの概要を図 2 左下に示す。この構成は、手話映像から話し言葉への翻訳実験において、同様のマルチタスク学習により翻訳精度が向上することが報告されており[5]、本手法においても手話と日本語の両方の情報を学習することで Encoder の精度向上が期待される。

④ 音声ー手話翻訳（日本語音声翻訳マルチタスク）

③と同様に音声認識モデルの Encoder の精度向上を図って、②の音声認識 Encoder に日本語テキストへ翻訳する Transformer Decoder を接続し、手話単語列と日本語テキストを同時に出力するマルチタスク学習を行った。この条件のモデルの概要を図 2 右下に示す。

以上の 4 条件について、バッチサイズを 16、最大エポック数を 20 とし、音声認識 Encoder の学習率を1×10^{-5}、Transformer の Encoder および Decoder の学習率を 1×10^{-4} に設定して、それぞれランダムシードを変えて 3 回学習を行った。①については日本語テキストを入力とする手話翻訳モデルの部分のみ 3 回学習を行い、最も評価データの BLEU 値が高かったモデルを音声認識モデルに接続した。また、③、④におけるマルチタスク学習では、以下の式に示す手話単語列と日本語テキストの損失関数の重み付き和 $\mathcal{L}(\mathbf{W}, \sigma_1, \sigma_2)$を最小化するように学習を行った[6]。

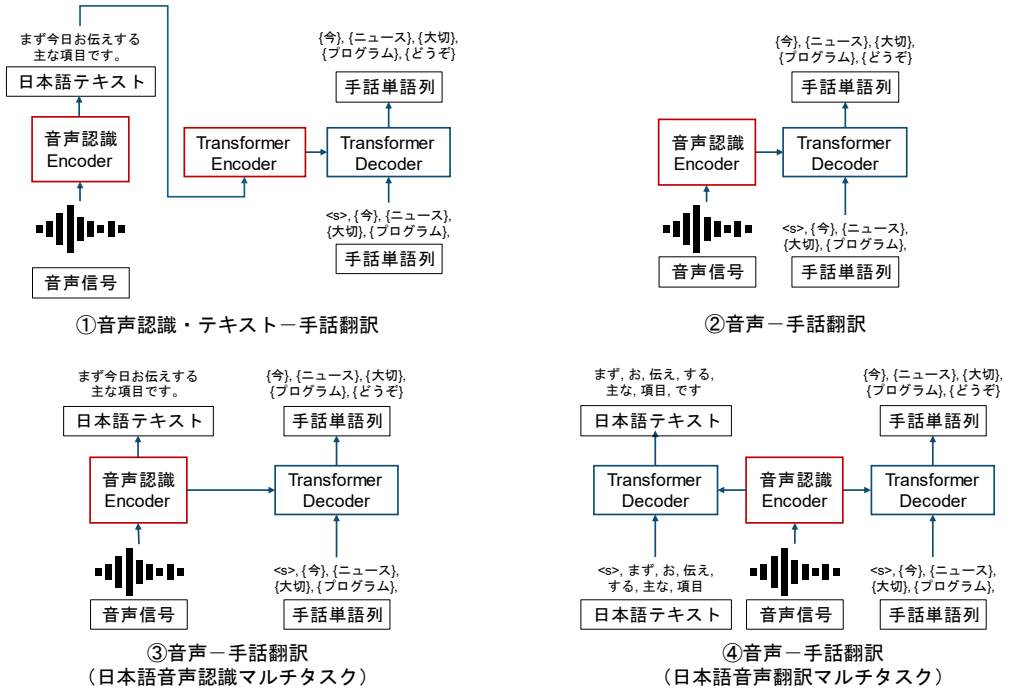

図 2　各実験条件の手話翻訳モデルの概要

$$\mathcal{L}(\mathbf{W},\sigma_1,\sigma_2)=\frac{1}{2\sigma_1^2}\mathcal{L}_1(\mathbf{W})+\frac{1}{2\sigma_2^2}\mathcal{L}_2(\mathbf{W})+\log\sigma_1+\log\sigma_2 \quad (1)$$

ここで、$\mathbf{W}$はモデルのパラメータ、$\mathcal{L}_1(\mathbf{W})$、$\mathcal{L}_2(\mathbf{W})$は各タスクの損失関数、$\sigma_1$、$\sigma_2$は各タスクにおける不確実性を表すパラメータである。σ_1、σ_2は学習率1×10^{-5}として最適化を図った。なお、③では音声認識タスクの CTC (Connectional Temporal Classification) 損失と手話翻訳タスクのクロスエントロピー誤差の値の大小を考慮して、各σの初期値を$\sigma_1=2$、$\sigma_2=1$と設定した。④では両方とも翻訳タスクでクロスエントロピー誤差を最小化することからσの初期値をともに 1 に設定した。

4. 実験結果

各条件の評価データに対する BLEU 値の平均値および標準偏差を表 1 に示す。また、①で使用したテキスト－手話翻訳の結果も併せて示す。また、③のマルチタスクである日本語音声認識の CER は平均 4.29%、④のマルチタスクである日本語音声翻訳の BLEU は平均 18.57 であった。

①はテキスト－手話翻訳と比較して BLEU 値が下がっており、音声認識結果の誤りが翻訳精度の劣化に影響していることがわかる。一方、音声を直接手話へ翻訳するモデル(②~④)では、テキスト－手話翻訳に及ばないものの、①より高い BLEU 値を示した。また、マルチタスク学習を行った③、④では②と比較して BLEU 値が向上しており、Encoder の表現力が向上したことが示唆された。

また、今回の翻訳結果から、音声特有のパラ言語情報が反映された翻訳結果は確認できなかった。これは今回の学習データが比較的単調に読み上げられるニュース番組をドメインとしたことに起因すると考えられる。

5. まとめ

本稿では、音声信号を入力とする手話翻訳モデルについて翻訳精度の検証を行った。その結果、従来手法の組み合わせとして音声認識のテキスト出力結果をテキスト－手話翻訳にかけるのではなく、音声信号から直接手話へ翻訳することで翻訳精度の改善が示唆された。今後は、モデルの改良による翻訳精度の更なる向上を図るとともに、ニュース以外のドメインに着目してパラ言語情報を活用した手話翻訳モデルの構築についても検討していく。

表 1　各翻訳モデルの BLEU 値

モデル	平均値	標準偏差
①音声認識・テキスト－手話翻訳	15.00	－
②音声－手話翻訳	15.31	0.52
③音声－手話翻訳 （日本語音声認識マルチタスク）	15.53	0.18
④音声－手話翻訳 （日本語音声翻訳マルチタスク）	15.72	0.27
テキスト－手話翻訳	15.81	－

参考文献

[1] 内田翼, “情報保障サービス拡充に向けた手話 CG 生成技術”, 映情学誌, 79(1), pp.15-20 (2025)
[2] T. Miyazaki, S. Tan, T. Uchida, H. Kaneko, “Sign Language Translation with Gloss Pair Encoding”, Proc. Workshop on the Representation and Processing of Sign Languages, pp. 262-268 (2024)
[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, “Attention Is All You Need”, NeurIPS (2017)
[4] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, M. Auli, “Unsupervised Cross-lingual Representation Learning for Speech Recognition”, arXiv preprint (2020)
[5] S. Tan, T. Miyazaki, N. Khan, K. Nakadai, “Improvement in Sign Language Translation Using Text CTC Alignment”, Proc. COLING, pp. 3255–3266 (2025)
[6] A. Kendall, Y. Gal, R. Cipolla, “Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics”, CVPR, pp.7482-7491 (2018)