DEIM Forum 2022 F24-5

プログレッシブトランスフォーマーを用いた 日本手話SLPモデルの性能評価

奥井　優太† 　宮森　恒†

† 京都産業大学情報理工部情報理工学科　〒 603–8555 京都府京都市北区上賀茂本山
E-mail: †{g1853220,miya}@cc.kyoto-su.ac.jp

あらまし　本稿では，プログレッシブトランスフォーマーを用いた手話生成モデルを日本手話に対して適用しその性能を多角的に評価する．手話自動生成 (SLP) は，公共の場などでの手話コミュニケーションの改善を目的に従来から研究されている．従来の SLP モデルは，個別の手話動作の生成結果を連結することで手話全体の系列を出力する必要があったが，近年提案された PTSLP モデルは，End-to-end でテキストから連続的な 3D 手話系列を生成でき，手話コミュニケーションの向上が期待されている．これまでドイツ手話のみに適用されてきた PTSLP モデルを日本手話に対して適用し，その性能を多角的に評価する．

キーワード　手話生成，日本手話，Transformer，PTSLP，3 次元 CG，自然言語処理

1 はじめに

手話とは，耳の聞こえない人 (聾者) 間または聞こえない人と聞こえる人 (聴者) 間に使用される，非音声の，手指の動きを中心とした身振りの一定の体系にもとづいた言語である [1]．聴者と聾者の対話において，聴者は日本語が母語，聾者は手話が母語だとすると，円滑なコミュニケーションを実現するには，手話通訳者が必要となる．しかし，手話通訳者を派遣するための手続きが煩雑であったり，手話通訳者になるための学習費用や学習難度が高いといった問題がある．他に，公共機関の場において，アナウンスなどの音声言語が字幕として表示されることもあるが，手話を母語とする聾者にとっては手話が望ましいという声もみられる [2]．

これらの問題を解決することを目的として，音声言語や文字言語から自動で手話を生成する研究が行われている．手話自動生成 (SLP) によって，人間の手話通訳者が不在でもリアルタイムで自動生成された手話を通して聾者に円滑に情報を提供することができる．

SLP モデルに関する従来研究は，統計的機械翻訳 (SMT) や探索テーブルなどのアプローチがある．統計的機械翻訳 (SMT) によるアプローチ [10] [11] は，テキストやグロス列と手話とのパラレルコーパスに基づき言語モデルとデコーダを構築することで手話を生成するアプローチである．言語モデルへの入力のばらつきを抑えるため，手話翻訳の専門家が作成したテキストを手話に翻訳しやすくするためのルールベース翻訳が組み合わされることもある．探索テーブルによるアプローチ [12] [13] [14] は，事前に用意したフレーズの，アバターによって表現された手話を探索テーブルに登録し，入力テキストからそのテーブルを介してアバターの手話を出力するというアプローチである．しかし，これらの手法では，生成できる手話動作が，エンコードが困難な静的ルールベースの処理や，探索テーブルに依存している．また，手話全体の動作系列は，個別の手話動作の生成結果の並びを連結することで出力する必要があった．

これに対して, 本研究で対象とする Progressive Transformers for End-to-End Sign Language Production [3](PTSLP) のモデルは，テキストと手話動作系列の間のマッピングを End-to-end で直接学習することに焦点を当てており．テキストから連続的な 3D 手話動作系列を生成できる．

しかし, このモデルはドイツ手話データセット PHOWNIX14T [4] のみに適用されており，異言語のデータセットでの性能が報告されていない．本稿では，2021 年に工学院大学から公開された多用途型日本手話言語データセット KoSign [5] に対して PTSLP モデルを適用し，その性能を検証する．

ドイツ手話データセット PHOENIX14T は，ドイツの天気予報から抽出された音声テキストとそのドイツ手話を含む大規模な動画ベースのパラレルコーパスである．このデータセットでは，個別の手話の意味に対応するグロスレベルのアノテーションも提供されている．

一方，日本手話データセット KoSign は，手話動作の多視点カメラ映像に加え，手話動作の高精細かつ高精度の 3 次元データが付与されたデータセットである．数字・単位，あいうえお，アルファベットを含む 3701 ラベル (1 ラベルは，1 語あるいは，異動作同義語を一纏めにした複数語に対応する) の手話動作が収録されている．3 次元動作データ，多視点カメラ画像だけでなく，距離センサによる距離画像の同期収集も行われている．

本稿では，日本手話データセット KoSign から構築される複数のデータセットに対して PTSLP モデルを適用し，その性能を多角的に評価する．具体的には，単語データと対話データに対して，関節位置の 3D 座標を 2D から推定する手法，3D マーカ座標を利用する手法といった異なる方法で 6 種類のデータセットを構築し，各データセットで訓練されたモデルによる手話動作の性能を検証する．

実験では，6 種類の各データセットで訓練されたモデルによる手話動作の性能を定量評価および定性評価で検証する．定性評価では，複数名の手話母語者の協力を得て，生成された手話動作に対して多角的に分析する．

定量評価では，それぞれのデータセットに対して，関節位置の 3D 座標を 2D から推定する手法と 3D マーカ座標を利用する手法を比較したところ，単語データでは前者，対話データでは後者のほうがより優れた性能を示した．定性評価では，単語データでは前者の方が優れており，後者は手話すら生成されない事例も確認された．対話データでは両方とも部分的に良好な結果を示した．評価者のなかには，口形や表情なしで読み取ることが困難というコメントも見られた．

本稿の貢献は以下の通りである．

- End-to-end でテキストから連続的な 3D 手話系列を生成できる PTSLP モデルを，日本手話に対して適用し，その性能を検証した．
- 日本手話データセット KoSign から異なる手法で PTSLP モデル用のデータセットを 6 種類構築し，各データセットでの PTSLP の性能を検証した．
- 定量評価だけでなく，複数名の手話母語者の協力を得て定性評価を実施し，生成された手話動作に対して多角的に分析した．

本論文の構成は，以下の通りである．2 節で関連研究について述べ，3 節で本稿で扱う問題を定式化し，PTSLP モデルについて説明する．4 節ではデータセットの内容と構築手順を詳述する．5 節で実験内容と実験結果，考察を示し，6 節で結論と課題をまとめる．

2 関 連 研 究

2.1 ドイツ手話と日本手話

「日本手話」は，耳の聞こえない日本語圏の人の言語である．手や身体の動作を目で見て受け入れるという特徴を最大限に生かして文法体系が組み立られており，日本語とは異なる言語体系をもつ [6]．一方，日本語の文法体系に従い手話単語を並べていく「日本語対応手話 (手指日本語)」と呼ばれるものも存在する．この日本語対応手話には，助詞や自動詞と他動詞の区別等が存在せず，聾者が容易に理解できるものではない [8]．また，日本手話と日本語対応手話の間に「中間型手話」というものも存在する．基本的には日本語の語順に従って表現されるが，日本手話独自の空間使用などの要素が加わっており，日本語の語順に従わない場合もある．日本語圏の手話では中間型手話の話者が最も多いといわれている [7]．

ドイツ手話は，耳の聞こえないドイツ語圏の人の言語である．日本手話と同様の特徴を持ち，ドイツ語とは異なる言語体系をもつ [9]．一方，ドイツ語文法体系に従い手話単語を並べていく「手話ドイツ語 (Signed German)」または「音声付随記号 (LBG：Lautsprachbegleitende Gebärden または Lautbegleitende Gebärden)」と呼ばれるものも存在する．手話ドイツ語では，助詞や接続詞なども含めて全て手話に符号化する．また，手話ドイツ語の個々の手話や抑揚などを省略する「音声サポート手話 (LUG：Lautsprachunterstätzende Gebärden)」と呼ばれるものも存在する．

日本語圏で使われる手話の種類と特徴を表 1 に，ドイツ語圏で使われる手話の種類と特徴を表 2 に示す．

表 1　日本語圏で使われる手話の種類と特徴

日本手話	日本語対応手話	中間型手話
・日本語とは異なる言語体系をもつ	・日本語の文法体系に従い手話単語を並べていく	・日本語対応手話に日本手話独自の空間使用などの要素が加わっている
・CL や NMM など非手指動作をもつ	・助詞や自動詞と他動詞の区別等が存在せず，聾者が容易に理解できるものではない	・この話者が最も多いといわれている

表 2　ドイツ語圏で使われる手話の種類と特徴

ドイツ手話	手話ドイツ語	音声サポート手話
・ドイツ語とは異なる言語体系をもつ	・ドイツ語の文法体系に従い手話単語を並べていく	・手話ドイツ語の個々の手話や抑揚などを省略している
・CL や NMM など非手指動作をもつ	・助詞や接続詞なども含めて全て手話に符号化する	・話し言葉の内容を把握する能力が不十分な聴覚障害者とのコミュニケーションに用いられる

ここで，本研究で用いる日本手話データセット KoSign の手話形は日本手話に該当する．また，PTSLP モデルが適用されたデータセット PHOENIX14T の手話形はドイツ手話に該当する．いずれも手話として手や身体の動作を最大限に活用しているという特徴があり，PTSLP モデルの性能を分析する上で妥当であると考えられる．特に本稿では，PTSLP モデルが日本語圏で用いられている日本手話の動作をどの程度的確に生成できるのかを分析することに焦点を当て，数種類のデータセットを用いるなどして多角的にその性能を検証する．

2.2 個別の手話動作の生成結果を連結する SLP モデル

個別の手話動作の生成結果を連結する SLP モデルの関連研究は，統計的機械翻訳 (SMT) や探索テーブルなどのアプローチがある．

2.2.1 統計的機械翻訳 (SMT) によるアプローチ

統計的機械翻訳 (SMT) によるアプローチ [10] [11] は，テキストやグロス列と手話とのパラレルコーパスに基づき言語モデルとデコーダを構築することで手話を生成するアプローチである．言語モデルへの入力のばらつきを抑えるため，手話翻訳の専門家が作成したテキストを手話に翻訳しやすくするためのルールベース翻訳が組み合わされることもある．このアプローチによって，手話データ取得のコスト削減や性能の向上を実現した．しかし，翻訳結果を手話に出力する対象が必要であった

り，生成する手話動作が，エンコードが困難な静的ルールベースの処理に依存していた．

2.2.2 探索テーブルによるアプローチ

探索テーブルによるアプローチ [12] [13] [14] は，事前に用意したフレーズの，アバターによって表現された手話を探索テーブルに登録し，入力テキストからそのテーブルを介してアバターの手話を出力するというアプローチである．このアプローチによって，テキストではなく手話として伝達することができ，手話を第一言語とするろう者にとって望ましいコミュニケーションを実現した．しかし，事前に登録した探索テーブルに依存していたり，手話全体の動作系列を出力するには個別の手話動作の生成結果の並びを連結する必要があったりした．

2.2.3 深層学習によるアプローチ

深層学習によるアプローチの一つに，Stoll らの，ニューラル機械翻訳 (NMT) と敵対的生成ネットワーク (GAN) を組み合わせた初期の SLP モデルがある [15]．このモデルは，問題を 3 つの別々のプロセスに分割し，それぞれを独立して学習させ，探索テーブルを介して手話のグロスからマッピングされた孤立した 2D スケルトンポーズ [16] の連結を生成する．このスケルトンポーズは，従来研究のアバターアプローチと比較して出力解像度や表現力は劣っているが，最小限のアノテーションを用いて連続的な手話合成が可能となっている．そのため，統計的機械翻訳や探索テーブルによるアプローチの課題点を解決でき，音声言語から手話への，費用対効果の高い翻訳が可能になっている．

Stoll らに対して，本研究が使用する PTSLP モデルは，問題を 3 つの別々のプロセスに分割し，孤立した手話を連結して生成することに焦点を当てるのではなく，テキストとスケルトンポーズ列間のマッピングを直接学習することに焦点を当てている．他に，新しいカウンターデコーディング技術を導入し，学習・推論時の連続シーケンス生成と，出力手話シーケンスの動的な長さの決定を可能にしている．しかし，モデルの推論時では，手の形が少し表現不足な部分も見られた．また，トレーニングデータには文法的な文脈や例がないため，固有名詞や特定のエンティティの生成が困難であった．

本稿では，PTSLP モデルが日本語圏で用いられている日本手話の動作をどの程度的確に生成できるのかを分析することに焦点を当て，数種類のデータセットを用いるなどして多角的にその性能を検証する．

3 Progressive Transformers for End-to-End Sign Language Production [3]

3.1 概　　要

Progressive Transformers for End-to-End Sign Language Production(PTSLP) は，離散的な話し言葉の文章から連続的な 3D 手話表現のシーケンスに End-to-End で変換する初の SLP モデルである．新しいカウンターデコーディング技術の導入により，学習と推論時に離散記号系列から連続動作系列の生成を可能としている．また，ドリフトの問題を克服し，SLP モデルの性能を飛躍的に向上させるために，いくつかのデータ拡張 (補強) プロセスが含まれている．本モデルの構成は，グロスを介して話し言葉から手話に翻訳する構成 (T2G2P) と，話し言葉から手話への End-to-End の直接翻訳を行う構成 (T2P) の 2 つがある．

3.2 モデルのアイデア

PTSLP モデルは，データから各単語の正しい長さと順序を学習し，カウンターデコーディングを使用して系列生成の終了を決定することで，動的な長さの出力手話系列の生成を可能としている．

3.3 問 題 設 定

PTSLP モデルの問題について定式化する．本モデルの構成は，グロスを介して話し言葉から手話に翻訳する構成 (T2G2P) と，話し言葉から手話への End-to-End の直接翻訳を行う構成 (T2P) の 2 つがある．本稿では，本研究が使用する日本手話データセット KoSign はグロスレベルのアノテーションがされていないため，T2P の構成のみを対象とする．すなわち，図 1b のプログレッシブトランスフォーマー (PT) が T2P に該当する．

T2P は，トランスフォーマーを用いたエンコーダデコーダアーキテクチャで構成されている．エンコーダでは，まず，類似した意味のトークンが互いに近くになるように，話し言葉 x_t の one-hot ベクトルを線形埋め込み層を介して埋め込む (Symbolic Embedding)．さらに，時間的順序を持たせるために，時間的埋め込み層を介して埋め込む (Positional Encoding)．これにより，埋め込まれた $\hat{w}_t$ に対して，シンボリックエンコーダ (E_S) を適用し，自己注意機構により話し言葉の記号系列の文脈ベクトルを学習する．最終的なエンコーダの出力は次のように定式化できる．

$$r_t = E_S(\hat{w}_t|\hat{w}_{1:T}) \tag{1}$$

ここで，r_t は話し言葉の記号系列の文脈表現であり，$\hat{w}_{1:T}$ は，T 単語の埋め込み表現である．

手話を表現する各フレームの関節位置 y_u は，3 次元関節位置の連続ベクトルとして表現される．デコーダでは，類似した内容が密な空間で近くに表現されるように，各フレームの関節位置 y_u を線形埋め込み層を介して埋め込む (Continuous Embedding)．次に，図 1b のように，手話全体の長さにおける相対的な時間位置を示す [0,1] の値をとるカウンター c_u を 3D 関節位置の埋め込み j_u に連結する．これにより，関節位置の順序を扱うことが可能となり，トランスフォーマーにおいても連続動作系列を生成することが可能になる (カウンターデコーディングと呼ばれる)．図 2 に，カウンターデコーディングとプログレッシブデコーダ (D_P) を適用することで，手話生成する様子を示す．最終的なデコーダの出力は次のように定式化できる．

$$[\hat{y}_{u+1}, \hat{c}_{u+1}] = D_P(\hat{j}_u|\hat{j}_{1:u-1}, r_{1:T}) \tag{2}$$

ここで，$\hat{y}_{u+1}$, $\hat{c}_{u+1}$ は，フレーム $u+1$ における手話ポーズを表す 3D 関節位置，カウンタ値をそれぞれ表す．$\hat{j}_u$, $\hat{j}_{1:u-1}$ は，

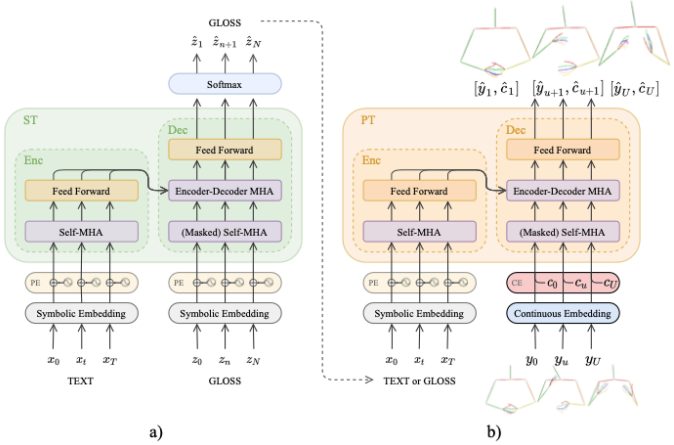

図 1　(a) シンボリックトランスフォーマー，(b) プログレッシブトランスフォーマーのアーキテクチャの詳細 [3]，(PT：Progressive Transformer，ST：Symbolic Transformer，PE：Positional Encoding, CE：Counter Embedding，MHA：Multi-Head Attention)

フレーム u およびフレーム区間 1 から $u-1$ の，カウンター連結された関節位置の埋め込みをそれぞれ表す．$r_{1:T}$ は，フレーム区間 1 から T の話し言葉系列の文脈ベクトルである．

4　データセット

本節では，本研究が対象とするデータセットについて述べ，実験用データセット構築について述べる．

4.1　日本手話データセット KoSign

日本手話データセット KoSign は，手話動作の多視点カメラ映像に加え，手話動作の高精細かつ高精度の 3 次元データが付与されたデータセットである．数字・単位，あいうえお，アルファベットを含む 3701 ラベル (1 ラベルは，1 語あるいは，異動作同義語を一纏めにした複数語に対応する) の手話動作が収録されている．3 次元動作データ，多視点カメラ画像だけでなく，距離センサによる距離画像の同期収集も行われている．

4.2　KoSign からの実験用データセット構築

KoSign から構築するデータセットの種類は，単語データ 7402 件 (D_w) と対話データ 171 件 (D_d) のそれぞれに対して，動画 (60fps) 上の 2D 関節点から 3D 関節位置を推定して構築されたデータセット ($D_{x,est60}$, $x \in \{w, d\}$)，および，モーションキャプチャー (MC) で取得された 3D 関節位置で構築された異なるフレームレート (60fps,120fps) のデータセット ($D_{x,raw60}$, $D_{x,raw120}$, $x \in \{w, d\}$ とする) の計 6 種類とした．また，ドイツ手話データセット PHOENIX14T から，PTSLP の原論文 [3] で適用されたデータセット構築と同じ手順で構築されたデータセットを $D_{d,baseline}$ とする．表 3 に，実験用データセットの一覧を示す．

2D 関節点から 3D 関節位置推定による具体的な構築手順は，次のように PTSLP の研究で行われたデータセット構築手順と同じ内容となる．

（1）　日本手話データセット KoSign の各動画 (mp4 ファイル) から，OpenPose [17] を用いて 2D の関節位置を抽出する．

（2）　抽出した 2D の関節位置を，骨格モデル推定の改良方法 [18] を利用して，3D に変換する．

（3）　GAN による手話生成に関する研究 [19] と同様にスケルトンの正規化を行い，3D 関節位置を (x, y, z) 座標で表現する．

モーションキャプチャーで取得された 3D 関節位置による具体的な構築手順としては，3 次元データの関節位置を読み込み，PTSLP の研究で使用された座標データ数 (150 個) と同じ数になるように 3 次元関節位置データを抽出し，上述のスケルトンの正規化を行うという手順となる．KoSign で収集された 3 次元データは 120fps であるため，120fps だけでなく，動画と同じフレームレートである 60fps でも実験用データセットを構築することとした．

なお，対話データについては，3 件の動画に対して，手話母語者である著者がアノテーションを行い，発話の長さや区切りを基準に 171 件のデータに分割した．

表 3　実験用データセットの一覧, (MC: モーションキャプチャ)

手話内容	3D 関節位置の取得方法	フレームレート	データセット記号
単語	2D から推定	60fps	$D_{w,est60}$
単語	MC から取得	60fps	$D_{w,raw60}$
単語	MC から取得	120fps	$D_{w,raw120}$
対話	2D から推定	60fps	$D_{d,est60}$
対話	MC から取得	60fps	$D_{d,raw60}$
対話	MC から取得	120fps	$D_{d,raw120}$

5　実　　験

実験では，表 3 で示す実験用データセットに対してそれぞれ構築された PTSLP モデルの性能を定量評価，定性評価を通して明らかにする．

5.1　比 較 手 法

比較手法としては, 表 3 で示す実験用データセットの $D_{x,est60}$,

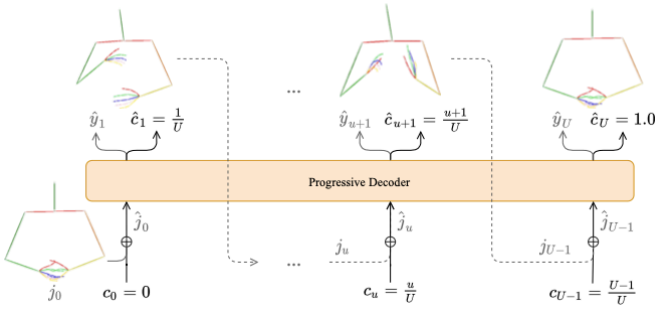

図 2　カウンターデコーディングの例 [3]．手話ポーズ $\hat{y}_u$ とカウンタ $\hat{c}_u \in [0,1]$ を同時に予測する．$\hat{c} = 1.0$ はシーケンスの終わりを示す．

$x \in w, d$ と $D_{x,raw60}$, $D_{x,raw120}$, $x \in w, d$，および，ベースラインである $D_{d,baseline}$ のそれぞれで PTSLP モデルを構築し，定量評価，定性評価を実施する．ベースラインについては，発話レベルで収録されているため，単語データ (D_w) との比較は行わない．モデルの訓練では，PTSLP の原論文 [3] で用いられたパラメータをそのまま用いることとした．

5.2　評価方法

本節では，定量評価と定性評価について説明する．

5.2.1　定量評価

定量評価では，MPJPE(Mean Per Joint Position Error) と 3DPCK(Percentage of Correct 3D Keypoints) を利用する．MPJPE は，関節点の推定座標と正解座標の距離を全ての関節点数およびフレーム数で平均することにより算出される評価指標である．MPJPE は以下の式で計算される．

$$MPJPE = \frac{1}{N_{frame}} \frac{1}{N_{joint}} \sum_{i=1}^{N_{frame}} \sum_{j=1}^{N_{joint}} \| J_j^{(i)} - \hat{J}_j^{(i)} \|_2 \tag{3}$$

N_{frame} は 1 つの手話動画の総フレーム数，N_{joint} は 1 フレームにおいて表現するスケルトンポーズの関節点の数である．J と $\hat{J}$ は，それぞれ正解結果と予測結果の 1 つの関節点の座標 (x, y, z) である．

3DPCK は，関節点の推定座標と正解座標の距離が設定した閾値よりも小さいときにその関節点の推定を正しいものとし，推定が正しく行われた割合をその評価値とする．閾値 α は 0.1 とし，20%の範囲の誤差を許容する．

5.2.2　定性評価

定性評価は，複数名の手話母語者の協力を得て実施する．1 名の評価者が評価する手話動作は，単語データから異なる手法で構築したデータセットによる手話動作の 2 種類と，対話データから異なる手法で構築したデータセットによる手話動作の 2 種類である．1 つのデータセットに対して 5 本の手話動作を評価するため，合計で 5 本× 4 種類の 20 本の手話動作を評価する．評価基準として，予測結果と正解結果を比較したうえで以下の質問に対して 5 段階評価を行うこととする．

- 腕や手の位置が一致しているか
- 指先の位置や方向が一致しているか
- 手話と手話の間の移行が滑らかであるか
- 手話の意味を理解できるか

5.3　実験結果

5.3.1　定量評価

単語データに対する定量評価の結果を表 4 に示す．

$D_{w,raw60}$ によるモデルが最も良好な結果を示した．また，$D_{w,raw120}$ によるモデルは，$D_{w,raw60}$ によるモデルの性能に及ばなかった．この結果は，同フレームレート (60fps) なら 3D 関節位置を直接取得する場合のほうが精度が高いということを示している．また，フレームレートが高くなるとその分誤差が累積していると考えられる．

対話データに対する定量評価の結果を表 5 に示す．

$D_{d,raw120}$ と $D_{d,raw60}$ のモデルが MPJPE と 3DPCK のいずれでも良好な結果を示した．また，$D_{d,est60}$ によるモデルは他のデータセットによるモデルよりも明らかに低い性能を示した．これより，PTSLP は，対話のような長い系列を扱う場合，高フレームレートでも誤差の蓄積があまり見られないと考えられる．また，対話データでは，単語データに比べて一連の素早い手話動作から骨格座標を抽出するため，2D から推定する方法では，モーションキャプチャから取得する方法に比べ，3D 関節位置のデータの質がより劣っていることを示している．

5.3.2　定性評価

$D_{w,est60}$ によるモデルの定性評価の結果を図 3 に示す．

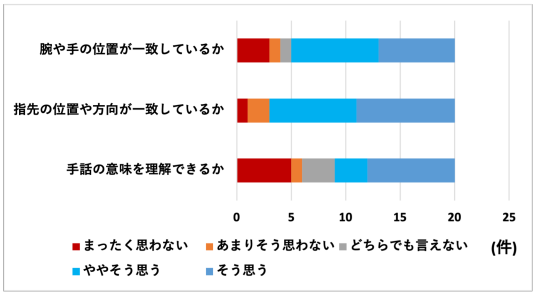

図 3　$D_{w,est60}$ によるモデルの定性評価結果

腕や手の位置，指先の位置や方向が一致しているという声が多く見られたが，手話の意味の理解は半々だった．実際の手話動作においても，図 4 のように腕や手の位置，指先の位置や方向がよく一致していることが確認できた．この結果は，腕の位置や方向など視覚的な評価は高いが，手話の意味までは理解し

表 4　単語データの定量評価の結果

	MPJPE				3DPCK			
	最小値	最大値	平均	標準偏差	最小値	最大値	平均	標準偏差
$D_{w,est60}$	1.6192	7.0606	3.1786	**1.3512**	69.11%	99.26%	87.44%	9.35%
$D_{w,raw120}$	2.1455	10.5153	6.0734	1.9758	62.47%	91.53%	77.73%	**7.28%**
$D_{w,raw60}$	**1.1051**	**6.9350**	**2.9106**	1.6126	**70.22%**	**99.64%**	**88.64%**	9.61%

表 5　対話データの定量評価の結果

	MPJPE				3DPCK			
	最小値	最大値	平均	標準偏差	最小値	最大値	平均	標準偏差
$D_{d,baseline}$	3.8172	**12.0390**	7.7853	**1.8472**	12.02%	74.15%	33.23%	**14.24%**
$D_{d,est60}$	3.0827	14.4016	9.3280	2.5514	11.72%	88.77%	25.15%	15.45%
$D_{d,raw120}$	2.3603	13.1755	**7.4604**	2.2738	**16.05%**	89.60%	36.60%	15.05%
$D_{d,raw60}$	**2.3569**	15.5325	7.6346	2.6482	15.00%	**90.45%**	**36.83%**	17.81%

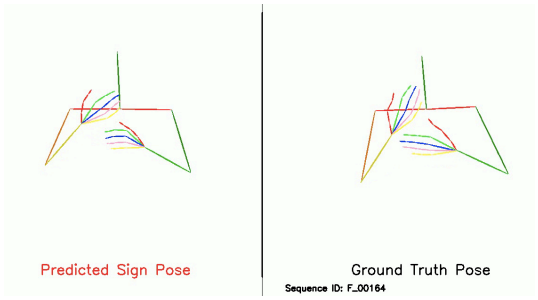

図 4　$D_{w,est60}$ を用いたモデルによる単語「合格」の生成結果，左：予測結果，右：正解結果

にくいということを示している．

$D_{w,raw120}$ によるモデルの定性評価の結果を図 5 に示す．

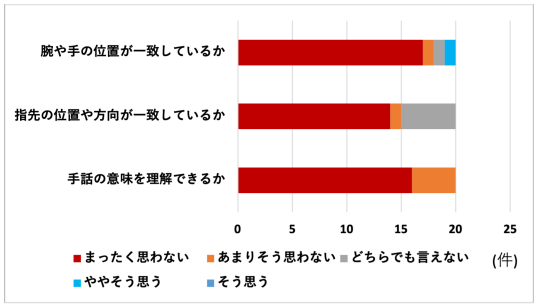

図 5　$D_{w,raw120}$ によるモデルの定性評価結果

全体的に評価が低かった．実際の手話動作では，図 6 のように片手または両手を上げて表現する手話に対して，予測結果はどちらも上げることなく，棒立ちの状態であった．

$D_{w,raw60}$ によるモデルの定性評価の結果を図 7 に示す．

腕や手の位置，指先の位置や方向が一致しているという声が多く見られたが，手話の意味の理解は半々だった．$D_{w,raw60}$ による定性評価結果 (図 7) と比べると，腕や手の位置ではわずかに上回ったが，指先の位置や方向が一致，手話の意味の理解についてはやや及ばなかった．実際の手話動作では，図 8 のように腕や手の位置，指先の位置や方向が概ね一致していた．

フレームレートによって定性評価結果に明らかな差が認め

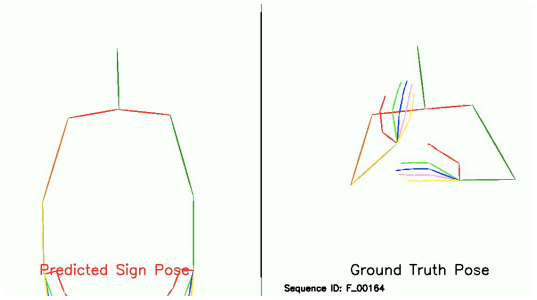

図 6　$D_{w,raw120}$ を用いたモデルによる単語「合格」の生成結果，左：予測結果，右：正解結果

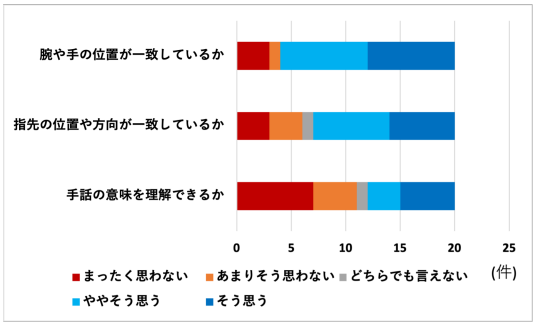

図 7　$D_{w,raw60}$ によるモデルの定性評価結果

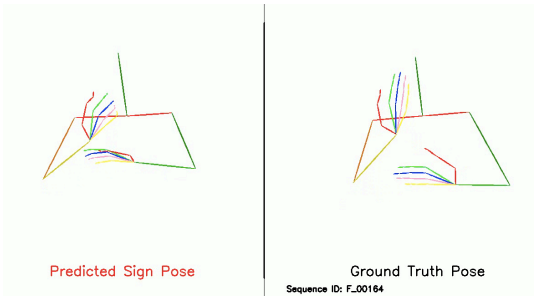

図 8　$D_{w,raw60}$ を用いたモデルによる単語「合格」の生成結果，左：予測結果，右：正解結果

られたことから，定量評価の単語データの考察と同様に，高フレームである $D_{w,raw120}$ のほうが誤差の蓄積が起きやすく，正しく学習できなかったと考えられる．そのため，より適切に学

習を進めるためには，原データの手話以外のシーンが必要以上に含まれないように調整したり，学習回数を増やすことなどが必要である．

また，3D 関節位置の取得方法の違いによる定性評価への影響として，2D から推定する方法と MC から推定する方法とではわずかに前者の方がよい結果となった．しかし，今回の評価では評価した単語数および被験者数が十分でないことが考えられ，今回の結果から明確な優劣を判断することは難しい．今後，評価する単語数や被験者数を増やすことでより明確な影響を判断できると考えられる．

$D_{d,est60}$ によるモデルの定性評価の結果を図 9 に示す．

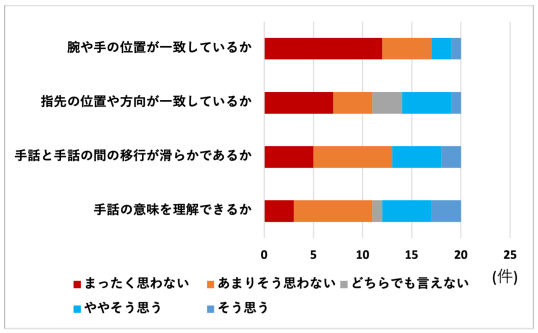

図 9　$D_{d,est60}$ によるモデルの定性評価結果

腕や手の位置の肯定的評価は 1 割程度でかなり低く，指先の位置や方向，手話と手話の間の移行の滑らかさ，手話の意味の理解についても肯定的な評価は 3 割程度にとどまった．実際の手話動作では，図 10 のように腕や手の位置はやや一致しているところもあるが，指先の位置や方向が一致していない事例が多く確認できた．

$D_{d,raw120}$ によるモデルの定性評価の結果を図 11 に示す．

肯定的な評価は 3 割から 5 割程度であったが，比較した手法の中では最良の評価結果となった．特に，指先の位置や方向，手話と手話の間の移行の滑らかさにおいて他の手法より良好な結果が得られた．実際の手話動作では，図 12 のように腕や手の位置は概ね一致しているものの，指先の位置や方向は必ずしも一致していない場合の方が多かった．ただ，動作全体としては滑らかであり，手話の意味としては比較手法の中では良好なものとなっていることが確認できた．

$D_{d,raw60}$ によるモデルの定性評価の結果を図 13 に示す．

肯定的な評価は 3 割弱にとどまっており，特に，手話と手話の間の移行の滑らかさや手話の意味の理解について肯定的評価が 1 割程度と，比較した手法の中では最も低い結果となった．実際の手話動作では，図 14 のように腕や手の位置は概ね一致しているものの，指先の位置や方向は必ずしも一致していない場合の方が多かった．ただ，動作全体としても滑らかさが感じられず，手話の意味としても比較手法の中では把握が最も困難であることが確認できた．

対話についての定性評価結果としては，$D_{d,raw120}$ によるモデルが最も高い結果を示し，$D_{d,est60}$ は $D_{d,raw60}$ よりもわずかに良好な結果を示した．評価者のなかには，「口形や表情な

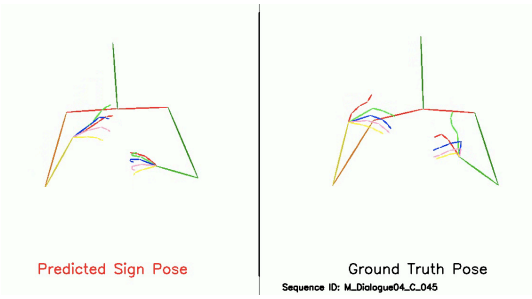

図 10　$D_{d,est60}$ を用いたモデルによる対話「カットかどうかわからない」の生成結果，左：予測結果，右：正解結果

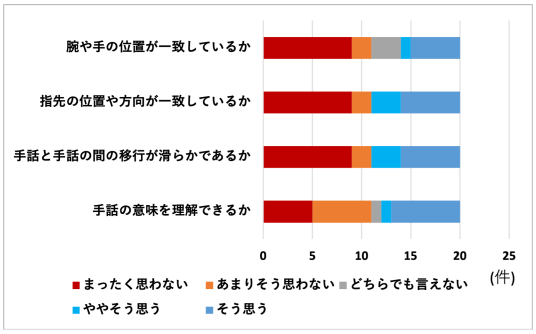

図 11　$D_{d,raw120}$ によるモデルの定性評価結果

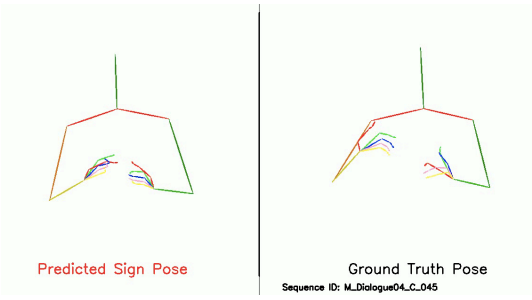

図 12　$D_{d,raw120}$ を用いたモデルによる対話「カットかどうかわからない」の生成結果，左：予測結果，右：正解結果

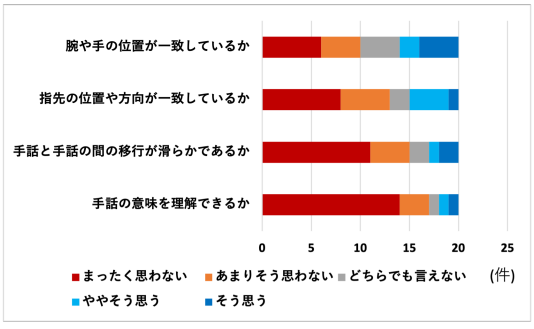

図 13　$D_{d,raw60}$ によるモデルの定性評価結果

しで読み取ることが困難だった」や「正解結果の手話もわかりにくかった」というコメントも見られた．$D_{d,raw120}$ によるモデルのほうがやや高い評価となったのは，$D_{d,raw120}$ の手話動作が $D_{d,est60}$ と $D_{d,raw60}$ の手話動作と比べて 2 倍のフレームレートであるため，再生時に動きがスローとなったため，細かな評価ができていたと考えられる．また，対話の手話の読み取りには，口形や表情など非手指動作の影響もあるため，非手指

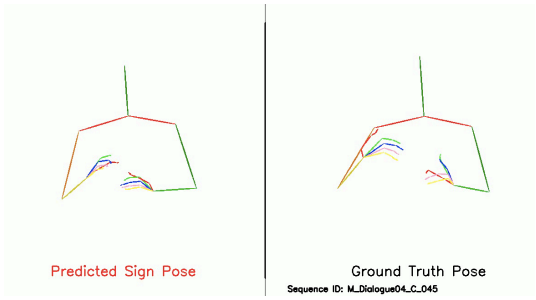

図 14　$D_{d,raw60}$ を用いたモデルによる対話「カットかどうかわからない」の生成結果，左：予測結果，右：正解結果

動作がないスケルトンポーズのみの動きだと手話の読み取りが困難になり，単語データに及ばない評価結果になったと考えられる．以上から，今後は，表情や口形を加えたスケルトンポーズを生成したり，評価動画の再生フレームレートを統一していくことが重要であると考えられる．

6 ま と め

テキストと手話動作系列の間のマッピングを End-to-end で直接学習する Progressive Transformers for End-to-End Sign Language Production(PTSLP) モデルに対し，日本手話言語データセット KoSign から構築されたさまざまなデータセットを適用し，その性能を検証した．

実験では，定量評価だけでなく，複数名の手話母語者の協力を得て定性評価を実施し，生成された手話動作に対して多角的に分析した．

定量評価と定性評価を行った結果，単語データでは，$D_{w,raw60}$ $>D_{w,raw120}$ $>D_{w,est60}$，$D_{w,est60}$ $\fallingdotseq$ $D_{w,raw60}$ $>D_{w,raw120}$ の順にそれぞれ精度・評価が高かった．対話データでは，$D_{d,raw120}$ $>D_{d,raw60}$ $>D_{d,est60}$，$D_{d,raw120}$ $>D_{d,est60}$ $\fallingdotseq$ $D_{d,raw60}$ の順にそれぞれ精度・評価が高かった．評価者の中には，口形や表情なしでは読み取ることが困難というコメントが見られた．

同じフレームレートなら，2D 関節点から推定するよりも，3D 関節位置を直接取得する場合の方が単語，対話のいずれでも常に高い結果を出す傾向が確認できる．関節位置の取得方法が同じなら，単語の場合，フレームレートが高くなるとその分誤差が累積しており，フレームレートが低い方がより高い結果を出している．一方，対話の場合，フレームレートが高い場合でも誤差の累積がそれほど起こらず，より高い結果を出している．この結果は，PTSLP が対話のようにある程度長い系列を扱う場合，より優れた性能を発揮することを示唆している．

今後は，特に対話を主とする大量かつ良質な手話データを収集することと，日本手話の特性となる非手指動作を考慮したデータセットを構築することが課題である．

謝　　辞

本研究の一部は科研費 18K11557 の助成を受けたものです．ここに記して感謝の意を表します．

文　　献

[1] 米川明彦. (1998). 「これから手話を学ぶ人のために」. 言語. 27:4, 20-25. 大修館書店.
[2] 総務省. (2006). 「国内における視聴覚障害者のテレビ利用状況等に関する現状調査」.
[3] Ben Saunders, Necati Cihan Camgoz, Richard Bowden. (2020). ”Progressive Transformers for End-to-End Sign Language Production”.
[4] O. Koller, J. Forster, and H. Ney. (2015). ”Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers”. Computer Vision and Image Understanding, volume 141, pages 108-125.
[5] 長嶋祐二. (2021). 「工学院大学 多用途型日本手話言語データベース コサイン (KoSign)」. 工学院大学.
[6] 市田泰弘. (2005). 「手話の言語学 (1) (12)」. 『月刊言語』第 34 巻第 1 号 12 号. 大修館書店.*連載記事.
[7] 松岡和美. (2005). 「日本手話で学ぶ 手話言語学の基礎」. 東京: くろしお出版.
[8] 木村晴美. (2011). 「日本手話と日本語対応手話 (手指日本語)」. 生活書院.
[9] Fabian Bross ,Daniel Hole. (2017). ”Scope-taking strategies in German Sign Language”. Glossa, A Journal of General Linguistics 2(1): 76. 1-30.
[10] Kayahan, D., Güngör, T. (2019). ”A Hybrid Translation System from Turkish Spoken Language to Turkish Sign Language”. IEEE International Symposium on INovations in Intelligent SysTems and Applications (INISTA).
[11] Kouremenos, D., Ntalianis, K.S., Siolas, G., Stafylopatis, (2018). ”Statistical Machine Translation for Greek to Greek Sign Language Using Parallel Corpora Produced via Rule-Based Machine Translation”. IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI).
[12] Glauert, J., Elliott, R., Cox, S., Tryggvason, J., Sheard, M. (2006). ”VANESSA-A System for Communication between Deaf and Hearing People”. Technology and Disability.
[13] Karpouzis, K., Caridakis, G., Fotinea, S.E., Efthimiou, E. (2007). ”Educational Resources and Implementation of a Greek Sign Language Synthesis Architecture”. Computers & Education (CAEO).
[14] McDonald, J., Wolfe, R., Schnepp, J., Hochgesang, J., Jamrozik, D.G., Stumbo, M., Berke, L., Bialek, M., Thomas, F. (2016). ”Automated Technique for Real-Time Production of Lifelike Animations of American Sign Language”. Universal Access in the Information Society (UAIS).
[15] Stoll,S.,Camgoz,N.C.,Hadfield,S.,Bowden,R.(2020). ”Text2Sign: Towards Sign Language Production using Neural Machine Translation and Generative Adversarial Networks”. International Journal of Computer Vision (IJCV).
[16] Ebling, S., Camgöz, N.C., Braem, P.B., Tissi, K., Sidler-Miserez, S., Stoll, S., Hadfield, S., Haug, T., Bowden, R., Tornay, S., et al. (2018). ”SMILE: Swiss German Sign Language Dataset”. Proceedings of the International Conference on Language Resources and Evaluation (LREC).
[17] Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y. (2017). ”OpenPose: Realtime Multi- Person 2D Pose Estimation using Part Affinity Fields”. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR).
[18] Zelinka, J., Kanis, J. (2020). ”Neural Sign Language Synthesis: Words Are Our Glosses”. The IEEE Winter Conference on Applications of Computer Vision(WACV).
[19] Stoll, S., Camgoz, N.C., Hadfield, S., Bowden, R. (2018). ”Sign Language Production using Neural Machine Translation and Generative Adversarial Networks”. Proceedings of the British Machine Vision Conference(BMVC).