## 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1 宮森恒2

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。

#### 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 [3] [4]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 [5] [6] [7]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 [8](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 [9]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 [10]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.



<sup>&</sup>lt;sup>1</sup> 非会員 株式会社インテック yuuta167.9@gmail.com <sup>2</sup> 正会員 京都産業大学 miya@cc.kyoto-su.ac.jp

- 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[11]. 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[12]. 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[13].

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 [14]. 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.

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

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

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

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[3][4]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 [5] [6] [7]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 [8]

#### 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。図1 a)に、T2Pに該当するプログレッシブトランスフォーマー(PT)の構成を示す。

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り、one-hotベクトルとして表すために、ソースとなる話し言葉 $x_t$ を線形埋め込み層を介して埋め込む(Symbolic Embedding). このとき、トランスフォーマーは、すべてのソースに対して位置情報なしで供給されることになり、語順の概念を持たない。そこで、時間的順序を持たせるために、話し言葉 $x_t$ から埋め込まれた $w_t$ を、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(Self-MHA)により話し言葉の記号系列の文脈ベクトルを学習する。この学習の出力は、非線形フィードフォワード(Feed Forward)に投影し、さらに学習を助けるために残差接続と後続層ノルムを用いて正規化を行う。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yu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 手話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, 埋め込まれた3次元関節座標 $j_u$ に対して, 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 $c_u$ を連結する(Temporal Embedding:図1のCE). 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1 b)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. これにより, カウンタ結合された関節埋 め込み $\hat{j}_u$ は、将来の位置へのAttentionを避けるためにターゲット マスキングが行われる((Masked) Self-MHA). そして, その次の 自己注意機構(Encoder-Decoder MHA)により、エンコーダからの 文脈表現 $r_t$ をデコーダの連続領域にマッピングし、話し言葉と手 話の間の重要な順序を学習する. 最後に, エンコーダと同様に, フィードフォワード(Feed Forward)により正規化を行う. 最終的 なデコーダの出力は次のように定式化できる.

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



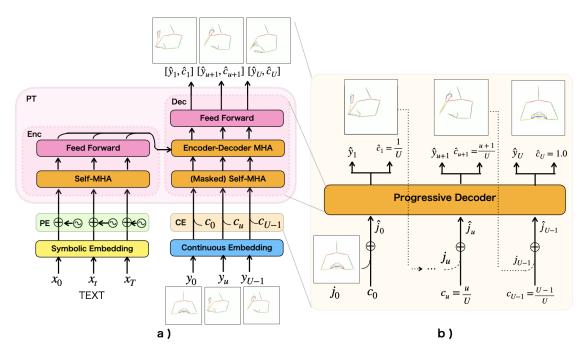


図1 a)プログレッシブトランスフォーマーのアーキテクチャの詳細 [8](PT: Progressive Transformer, PE: Positional Encoding, CE: Counter Embedding, MHA: Multi-Head Attention),

b) カウンターデコーディングの例 [8]. 手話ポーズ $\hat{y}_u$ とカウンタ $\hat{c}_u$ を同時に予測する.  $\hat{c}_u = 1.0$ はシーケンス終了を示す.

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 [8]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。

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	フレーム	データセット
	の取得方法	レート	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	$D_{d,raw120}$

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.

#### 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 $D_{x,est60}$ , $x \in w, d \geq D_{x,raw60}, D_{x,raw120}, x \in w, d$ ,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 [8]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3D PCK(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.

$$MPJPE = \frac{1}{N_{frame}} \frac{1}{N_{joint}} \sum_{i=1}^{N_{frame}} \sum_{j=1}^{N_{joint}} \|J_{j}^{(i)} - \hat{J}_{j}^{(i)}\|_{2} \quad (3)$$

 $N_{frame}$ 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 $\alpha$ 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。 $D_{w,est60}$ によるモデルや $D_{w,raw120}$ 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。これは,同フレームレート(60fps)なら,3D関節位置を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ほうがより正確な手話動作を生成できることを示している。また,3D関節位置を直接取得する際はフレームレートが高いと性能が低下することを示している。これは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ためと考えられる。

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2に示す.

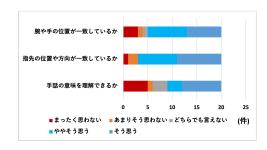


図2  $D_{w,est60}$ によるモデルの定性評価結果

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3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4に示す.

全体的に評価が低かった.実際の手話動作では,図5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6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



	МРЈРЕ				3D	PCK		
	最小値	最大値	平均	標準偏差	最小值	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%	<b>7.28</b> %
$D_{w,raw60}$	1.1051	6.9350	2.9106	1.6126	70.22%	99.64%	88.64%	9.61%

表4 単語データの定量評価の結果

表5 対話データの定量評価の結果

	МРЈРЕ				3D	PCK		
	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%

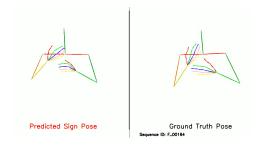


図3  $D_{w,est60}$ を用いたモデルによる単語「合格」の生成結果, 左:予測結果,右:正解結果

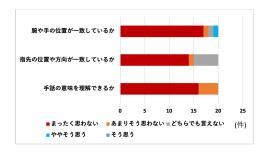


図4  $D_{w,raw120}$ によるモデルの定性評価結果

Predicted Sign Pose Sequence 10: F\_00164

D 100を用いたモデルによる単語 [全核] の生

図5  $D_{w,raw120}$ を用いたモデルによる単語「合格」の生成結果,E:予測結果,右:正解結果

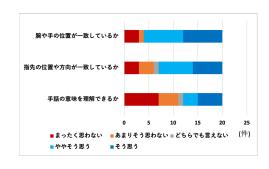


図6  $D_{w,raw60}$ によるモデルの定性評価結果

る定性評価結果(図6)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7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8に示す.

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



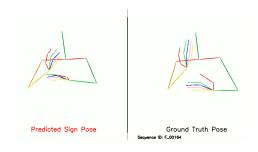


図7  $D_{w,raw60}$ を用いたモデルによる単語「合格」の生成結果, $E: \mathbb{R}$ 期結果,右:正解結果

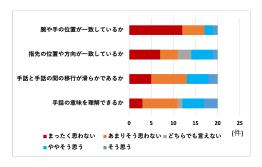


図8  $D_{d.est60}$ によるモデルの定性評価結果

作では、図9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.

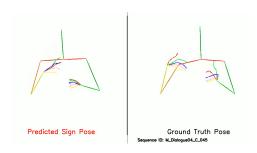


図9  $D_{d,est60}$ を用いたモデルによる対話「カットかどうかわからない」の生成結果、左:予測結果、右:正解結果

 $D_{d,raw120}$ によるモデルの定性評価の結果を図10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11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12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

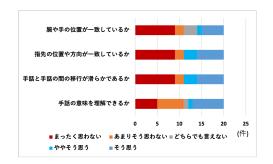


図10  $D_{d,raw120}$ によるモデルの定性評価結果

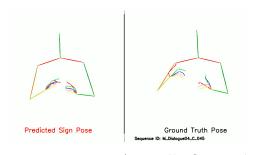


図11  $D_{d,raw120}$ を用いたモデルによる対話「カットかどうかわからない」の生成結果、左:予測結果、右:正解結果

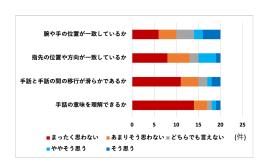


図12  $D_{d,raw60}$ によるモデルの定性評価結果

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13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,非手指動作がないス



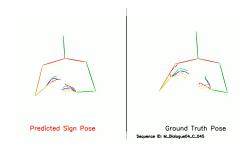


図13  $D_{d,raw60}$ を用いたモデルによる対話「カットかどうかわからない」の生成結果、左:予測結果、右:正解結果

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、 $D_{w,raw60}$  > $D_{w,raw120}$  > $D_{w,est60}$ , $D_{w,est60}$  =  $D_{w,raw60}$  > $D_{w,raw120}$ の順に それぞれ精度・評価が高かった。対話データでは、 $D_{d,raw120}$  > $D_{d,raw60}$  > $D_{d,est60}$ , $D_{d,raw120}$  > $D_{d,est60}$  =  $D_{d,raw60}$ 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.

#### 7 参考文献

#### 参考文献

- [1] 明彦米川. これから手話を学ぶ人のために. 言語, Vol. 27, No. 4, pp. 20–25, 04 1998.
- [2] 総務省. 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. 2006.
- [3] Dilek Kayahan and Tunga Güngör. A hybrid translation system from turkish spoken language to turkish sign language. In 2019 IEEE International Symposium on INnovations in Intelligent SysTems and Applications (INISTA), pp. 1–6, 2019.
- [4] Dimitris Kouremenos, Klimis S Ntalianis, Giorgos Siolas, and Andreas Stafylopatis. Statistical machine translation for greek to greek sign language using parallel corpora produced via rule-based machine translation. In CIMA@ ICTAI, pp. 28–42, 2018.
- [5] J.R.W. Glauerta, R. Elliott, S.J. Cox, and M. Sheard. Vanessa a system for communication between deaf and hearing people. *Technology and Disability*, Vol. 18, pp. 207–216, 12 2006.
- [6] K. Karpouzis, G. Caridakis, S.-E. Fotinea, and E. Efthimiou. Educational resources and implementation of a greek sign language synthesis architecture. *Computers & Education*, Vol. 49, No. 1, pp. 54–74, 2007. Web3D Technologies in Learning, Education and Training.
- [7] John Mcdonald, Rosalee Wolfe, Jerry Schnepp, Julie Hochgesang, Diana Gorman Jamrozik, Marie Stumbo, Larwan Berke, Melissa Bialek, and Farah Thomas. An automated technique for real-time production of lifelike animations of american sign language. *Univers. Access Inf. Soc.*, Vol. 15, No. 4, p. 551–566, nov 2016.
- [8] Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. Progressive transformers for end-to-end sign language production. *CoRR*, Vol. abs/2004.14874, , 2020.
- [9] Oscar Koller, Jens Forster, and Hermann Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, Vol. 141, pp. 108–125, 2015. Pose & Gesture.
- [10] 工学院大学, University Kogakuin. 工学院大学多用途型日本手話言語データベース(KoSign), 06 2021.
- [11] 泰弘市田. 手話の言語学(1) (12). 言語, Vol. 34, No. 1, pp. 90–97, 01 2005.
- [12] 晴美木村. 日本手話と日本語対応手話(手指日本語): 間にある「深 い谷」. 生活書院, 2011.
- [13] 和美松岡. 日本手話で学ぶ手話言語学の基礎. くろしお出版, 2015.
- [14] Fabian Bross and Daniel Hole. Scope-taking strategies and the order of clausal categories in german sign language. Glossa: a journal of general linguistics, Vol. 2, pp. 1–30, 08 2017.
- [15] Stephanie Stoll, Necati Camgoz, Simon Hadfield, and Richard Bowden. Text2sign: Towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision*, 04 2020.
- [16] Sarah Ebling, Necati Cihan Camgöz, Penny Boyes Braem, Katja Tissi, Sandra Sidler-Miserez, Stephanie Stoll, Simon Hadfield, Tobias Haug, Richard Bowden, Sandrine Tornay, et al. Smile:swiss german sign language dataset. In Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC) 2018. The European Language Resources Association (ELRA), 2018.
- [17] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *CoRR*, Vol. abs/1812.08008, , 2018.
- [18] Jan Zelinka and Jakub Kanis. Neural sign language synthesis: Words are our glosses. In 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 3384–3392, 2020.
- [19] Stephanie Stoll, Necati Camgoz, Simon Hadfield, and Richard Bowden. Sign language production using neural machine translation and generative adversarial networks. In *British Machine Vision Confer*ence (BMVC), 08 2018.

