ヘッドマウントディスプレイを用いた一人称視点手話認識と音声合成・認識を組み合わせた対話支援インタフェースの提案

乾 竜躍[1]　辻井 一真[2]　加藤 恒夫[2]　田村 晃裕[2]

概要：手話を主なコミュニケーション手段とする人と手話を知らない人の対面対話を支援するため，筆者らはヘッドマウントディスプレイ (HMD) を用いて一人称視点手話認識と音声合成・音声認識を組み合わせた対話支援インタフェースを構築した．一人称手話認識には，手話の 3 次元モーションキャプチャデータを HMD のローカル座標系と整合するように変換した手骨格系列を用いて，グラフ畳み込みネットワーク ST-GCN を学習して用いた．HMD で記録した手話データを用いてシミュレーション評価を行った結果，20 語の単語認識において認識精度 Top-1: 70.7%を得た．総合評価として，手話を用いる聴覚障碍者 1 名と対話者 1 名に対し 3 種類の対話シナリオを実施した．その結果，手が計測範囲内にある限り推定が更新され続けるため，意図しない動作が誤出力を招く場面が見られた．対話者側では，単語単位の提示では発話意図を文として把握しにくく，誤認識が混入すると，提示された単語列のどの部分が正しく，どの部分が誤りかを対話者が判断できないため，会話に用いる情報の取捨選択が困難であることが事後アンケートの結果わかった．以上から，認識精度の改善と認識可能な語彙の拡張に加えて，手話区間検出の導入による誤出力抑制と，文単位の手話認識の必要性が確認された．

1. はじめに

手話は，聴覚や発話に障碍のある人にとって重要なコミュニケーション手段である．しかし，手話を用いる人と手話を知らない人とのリアルタイムコミュニケーションは容易ではない．そこで筆者らは，手話認識と音声合成・音声認識を組み合わせ，手話を用いる人と手話を知らない人との対話を支援するシステムの構築を目指す．手話は，手指の形状や位置・動きに加えて，腕の動作など上半身の動きで意味を表す視覚言語である．手話認識の研究は，手話を用いる人 (手話者) と手話を知らない人との間の言語的な障壁を低減することを目指して，国内外で盛んに行われている．従来の研究の多くは，手話者を正面から撮影したカメラ映像を前提としているが，日常生活での利用を考えると，撮影者やカメラ設置場所の確保などの課題がある．

この課題に対して，ヘッドマウントディスプレイ (HMD) やスマートグラス等のウェアラブルデバイスを用いれば，手話者がデバイスを装着するだけで一人称視点から手の動きを取得できるため，外部カメラの設置を必要とせず実環境での利用ができる．そこで本研究では，ヘッドマウントディスプレイを用いて一人称視点手話認識と音声合成・音声認識を組み合わせ，手話者とその対話者 (非手話者) との双方向対話支援インタフェースを構築した．本システムでは，手話者が HMD を装着して手話を行うと，手話を自動認識し，その結果を手話を知らない対話者へ音声として提示する．また，対話者の発話は HMD 内蔵マイクで収音し，音声認識によりテキスト化して HMD 上に表示する．図 1 に使用例，図 2 にシステム構成を示す．このように，本システムは「音声→字幕」と「手話→音声」を統合することで，双方がそれぞれ自然なモダリティで情報を受け取

図 1　本システムの使用例．(a) 対話者視点：手話者が HMD を装着して手話を行う様子．(b) 手話者視点：視野内に手話認識結果（上段）と対話者音声の認識字幕（下段）を表示する．

[1] 同志社大学大学院 理工学研究科
[2] 同志社大学 理工学部

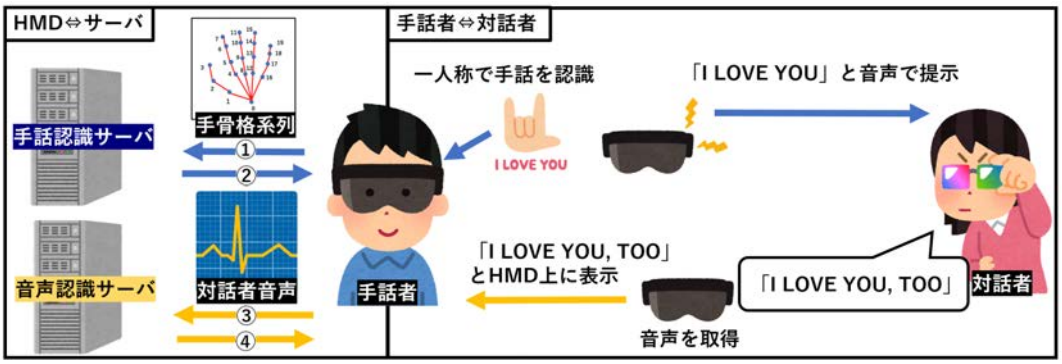

図 2 提案する対話支援システムの構成．(1) 手骨格系列の送信，(2) 手話認識結果の返送，(3) 対話者発話音声の送信，(4) 音声認識結果の返送．

れる対話環境を提供する．

しかし，HMD ／スマートグラスによる一人称視点の大規模日本手話データセットは公開されておらず，深層学習に基づく手話認識を学習するためのデータを新規に収集するには，手話母語者の協力や収録環境の整備など相当のコストがかかる．そこで本システムでは，一人称視点データの代わりに既存の手話モーションキャプチャデータを活用する．具体的には，日本手話の 3 次元モーションキャプチャデータの骨格データを HMD で取得される手骨格系列と幾何的に対応付け，骨格ベースの手話認識モジュールとして ST-GCN[1] を学習し，対話支援システムに組み込む．

本稿では，HMD を用いた手話・音声の双方向対話支援インタフェースの設計と実装を示し，モーションキャプチャデータを活用した学習データ構築手法を述べる．さらに，手話をコミュニケーション手段として利用するユーザが健聴者と対話する場面を対象とした実験を通じて，実際の使われ方や有用性，今後の課題を整理する．

2. 関連研究

手話認識の研究では，従来からカメラ映像に基づく手話認識が盛んに行われている．特に骨格情報を入力とする手法として，Spatial Temporal Graph Convolutional Networks（ST-GCN）[1] などのグラフ畳み込みネットワークが提案され，骨格ベースの動作認識に有効であることが示されている．Nakamura ら [2] は，KoSign[3] のモーションキャプデータから Unity 上で CG 手話動画を生成し，その RGB 動画に MediaPipe[4] を適用して得た骨格系列を入力として，日本手話単語認識における ST-GCN の有効性を少量データ環境で検証している．また，評価では Azure Kinect DK で手話者を正面から撮影した RGB 動画からも骨格を推定している．

一人称視点やウェアラブルデバイスを用いた研究としては，SignGlass[5] がスマートグラスに内蔵したカメラ映像からアメリカ手話を認識・翻訳し，字幕や音声として提示することで，グラス型デバイス上での対話インタフェースを実現している．また，Fujimoto ら [6] は HoloLens 2 のハンドトラッキング機能を用いて一人称視点から手指関節の 3 次元座標を取得し，日本語指文字の認識を行っている．これらは，グラス型デバイスによる一人称視点認識や対話支援の有効性を示している一方で，デバイスから取得する固有のデータを直接入力とする専用データセットを前提としている．

3. 対話支援システム

3.1 HMD インタフェース

図 1 に，本システムを用いた対面対話の様子を示す．手話者は HMD を装着したまま対話者を見て手話を行い（図 1(a)），認識結果と対話者発話を視野内で確認しながら会話を進める．手話者視点では，視野上部に手話認識結果を短いテキストとして表示し，視野下部に対話者音声の認識字幕を表示する（図 1(b)）．例として，上段に直前の手話認識結果「いく」を表示し，下段に対話者の発話「あなたも行きますか？」を字幕として提示する．これにより，手話者は会話を継続しながら，自分の手話がどのように認識されたかを確認できる．手話認識結果は音声合成により読み上げられ，対話者へ提示される．対話者は手話を知らなくても音声として内容を受け取り，通常の音声対話と同様に応答できる．

3.2 システム構成

図 2 に，本システムの構成とデータの流れを示す．本システムは，HMD，手話認識サーバ，音声認識サーバから構成される．

手話者は HMD を装着して手話単語を表現し，HMD は両手の 3 次元骨格情報を取得して時系列の手骨格系列として手話認識サーバへ送信する（図 2 の (1)）．手話認識サーバは受信した手骨格系列から単語クラスを推定し，推定した認識結果を HMD へ返送する（(2)）．HMD は返送された認識結果を視野内に表示し，対話者へは音声合成により提示する．

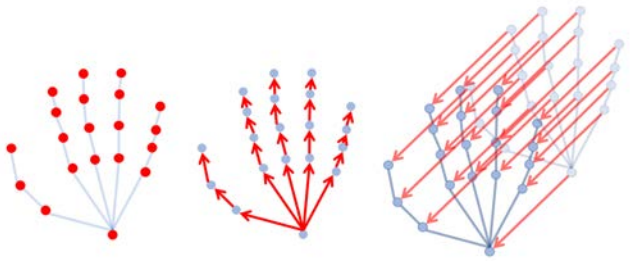

図 3 ST-GCN に入力する 3 種類の特徴量のイメージ．
左：関節位置，
中央：親子関節間ベクトルによる bone 特徴，
右：時刻間の差分ベクトルによる速度特徴．

一方，対話者の発話は HMD 内蔵マイクにより取得され，音声波形信号が音声認識サーバへ送信される（(3)）．音声認識サーバは自動音声認識によりテキストへ変換し，音声認識結果を HMD へ返送する（(4)）．返送されたテキストは HMD 上に字幕として表示される．

3.3 手話認識サーバ

HMD・サーバ間の通信は TCP で行われる．手話認識サーバには，学習済みの骨格ベース手話認識モデルを配置する．サーバは HMD から送信された手骨格系列を受信し，所定の前処理を適用した後にモデルへ入力して，語彙クラスごとの事後確率を推定する．

受信データはフレーム数 T，関節数 V，各関節の座標 C からなる骨格系列であり，各フレームの骨格は V 個の関節座標（C 次元）として表される．前処理では，骨格点の速度や関節間ベクトル（bone 特徴）などの特徴を計算し，モデルに入力する．サーバ側では，事後確率最大のクラス（Top-1）を単語テキストに変換して HMD クライアントに返送する．

3.4 音声認識サーバ

音声認識サーバでは，対話者側の音声を手話者に伝えるため，自動音声認識を行う．HMD 内蔵マイクで取得した音声は 16 kHz で連続収録し，16 bit PCM 信号として，TCP 通信によりサーバへ逐次送信する．送信処理では，マイクのリングバッファに対して HMD の描画フレームレートに同期して前回送信位置からの差分サンプルのみをチャンクとして送信することで，低遅延なストリーミング入力を実現した．また，小音量環境での認識安定化を狙い，送信前に波形へゲインを適用した．

サーバ側では whisper_streaming [7] を用いて逐次的に認識し，認識結果テキストを随時 HMD に返送する．HMD 側では同一 TCP 接続上で音声の上り（PCM 送信）と字幕の下り（認識結果受信）を同時に扱う全二重通信とし，受信したテキストをディスプレイ上に字幕として提示する．

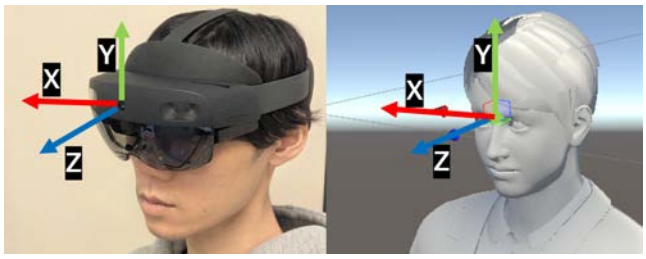

図 4 左：HMD ローカル座標系，右：KoSign モーションキャプチャデータの身体関節座標から構成した頭部ローカル座標系

4. ST-GCN による一人称手話認識

4.1 ST-GCN による手話認識

手骨格系列の認識モデルとして，Yan らの ST-GCN[1] を用いた．ST-GCN は，関節を頂点集合 $\mathcal{V}$，関節間の接続を辺集合 $\mathcal{E}$ とする骨格グラフ $G = (\mathcal{V}, \mathcal{E})$ と，フレーム間の時間的接続を同時に扱うことで，時空間的な動作パターンを学習できる．ここで $(i, j) \in \mathcal{E}$ は，骨格モデル上で関節 i と関節 j が親子関係にあることを表す．

以下，時刻 t における頂点 v の関節位置を $\boldsymbol{p}_t^{(v)} \in \mathbb{R}^3$ と表す．本研究では入力特徴として，位置に加えて時間方向および関節間の関係を明示するため，速度特徴と bone 特徴を用いる．

速度特徴は隣接フレーム間の差分ベクトル

$$\boldsymbol{v}_t^{(v)} = \boldsymbol{p}_t^{(v)} - \boldsymbol{p}_{t-1}^{(v)} \qquad (t \geq 1)$$

で定義し，$t = 0$ ではゼロベクトルとする．bone 特徴は，スケルトングラフ上の親子関係にある関節ペア $(i, j) \in \mathcal{E}$ に対して

$$\boldsymbol{b}_t^{(i,j)} = \boldsymbol{p}_t^{(j)} - \boldsymbol{p}_t^{(i)}$$

を計算する．図 3 に位置・bone・速度特徴のイメージを示す．

4.2 三人称骨格データを用いた学習

本研究では，一人称視点デバイスから得られる手骨格系列データの欠如を解決するため，既存のモーションキャプチャデータによる手話骨格データを学習に活用する．モーションキャプチャデータはスタジオ等に固定されたワールド座標系で表現される一方，HMD で取得される手骨格系列は，HMD ローカル座標系で表現される．そのため，本研究ではワールド座標系から頭部ローカル座標系への座標変換を行う．

ワールド座標系における時刻 t の頂点 v の関節位置を $\boldsymbol{q}_t^{(v)} \in \mathbb{R}^3$ とする．以下，座標変換後の頭部ローカル座標系における位置を $\boldsymbol{p}_t^{(v)} \in \mathbb{R}^3$ と表す．左右眉の中点を原点 $\boldsymbol{O}$，左右肩の中点を肩中心 $\boldsymbol{s}$ とし，肩中心から頭部へのベクトルを，身体に沿った上方向の単位ベクトル $\boldsymbol{e}_y$ とする．さらに，左右肩の差分ベクトルから $\boldsymbol{e}_y$ 成分を除去して右

表 1 認識対象とする日本手話単語 20 語

00026 家	00029 今	00053 人	00061 見る
00149 皆	00249 行く	00251 言う	00290 違う
00291 良い	00381 時	10052 来る	10098 入る
10134 物	10470 事	10529 自分	20304 話
20374 無い	22206 する	22207 居る	22311 有る

方向単位ベクトル $\boldsymbol{e}_x$ を求め，外積により前方向単位ベクトル $\boldsymbol{e}_z$ を得る．これらを並べた回転行列

$$R = [\boldsymbol{e}_x,\ \boldsymbol{e}_y,\ \boldsymbol{e}_z]$$

を用いて，ワールド座標系における関節位置 $\boldsymbol{q}_t^{(v)}$ を頭部ローカル座標系に写像し，

$$\boldsymbol{p}_t^{(v)} = R^\top(\boldsymbol{q}_t^{(v)} - \boldsymbol{O})$$

として HMD ローカル座標系と整合した座標を得る．図 4 に，HMD ローカル座標系と，モーションキャプチャデータから構成した頭部ローカル座標系の関係を示す．

4.3 シミュレーション評価

本節では，変換処理を施したモーションキャプチャデータから学習した ST-GCN モデルで HMD で収録した一人称視点データを認識したときの単語認識精度を評価する．

4.3.1 ST-GCN の学習データと学習条件

学習には，日本手話モーションキャプチャデータベース KoSign [3] に含まれる FBX 形式の単語モーションを用いた．FBX から各フレームの関節位置を抽出し，前節で述べた頭部基準ローカル座標系への変換を適用することで，HMD ローカル座標系で得られる手骨格系列と同形式の系列を生成した．本研究で扱う KoSign の FBX 骨格座標と HMD から得られる手骨格座標は，いずれも単位がメートルで表現されている．そのため，単位変換や絶対スケール合わせは行わず，KoSign 側の関節間距離（スケール）を保持した．

対象語彙は，中納言日本語日常会話コーパス [8] の頻度情報を参考に候補語を作成し，(1) KoSign に単語モーションが収録されていること，(2) HMD の計測範囲内で表現可能であること，を満たす 20 語を選定した． 表 1 に選定した 20 語を示す．各単語の先頭に付した 5 桁の数値は，KoSign で付与されている単語 ID を表す．ST-GCN の入力は，左右手それぞれ 20 関節（計 $V = 40$）の 手骨格系列データとし，各サンプルの系列長は，元系列から 等間隔にフレームを抽出する間引き処理により $T = 50$ フレームに統一した．また入力特徴として，座標に加えて速度および bone 特徴を用い，座標のみ，座標＋速度，座標＋ bone，座標＋速度＋ bone の 4 設定で学習・評価を行った．対象語彙には片手で表現される単語が含まれるため，非アクティブ側の手は全フレームで $(0, 0, 0)$ に置換した．この処理は KoSign と HMD の双方に適用した．

表 2 HMD テストセットに対する単語認識精度

入力特徴	Top-1 [%]	Top-3 [%]
座標のみ	42.7	65.5
座標＋速度	43.5	67.5
座標＋ bone 特徴	65.2	**88.5**
座標＋速度＋ bone 特徴	**70.7**	88.0

学習データの多様性を高め，汎化性能を向上させるため，3 次元骨格系列に対するスケール変換や時間方向の拡張が性能向上に寄与することが報告されている [9]．本研究では，骨格全体に一様スケール係数 $s \in \{0.95, 1.0, 1.05, 1.10, 1.15, 1.20, 1.25\}$ を乗じる拡張と，動作速度の変化を模擬する時間方向拡張として，時間伸縮率 $a \in \{0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3\}$ による再サンプリングを適用した．

話者 2 名 × 20 語 × 7 スケール × 7 速度により 1,960 サンプルを生成し，学習用 1,764 サンプル，開発用 196 サンプルに分割した．モデルには ST-GCN[1] を用い，損失関数はクロスエントロピーとした．最適化には AdamW を用い，バッチサイズは 64，学習率は 7×10^{-4} から開始して余弦アニーリングにより減衰させた．ドロップアウト率は 0.5 とした．開発セットの損失がエポック 20 連続して改善しない場合は Early Stopping により学習を打ち切った．

4.3.2 評価方法と評価データ

本研究では HMD として Microsoft HoloLens 2 を用い，HMD で取得した手骨格系列を評価データとした．評価用の HMD テストセットは 4 名から収集した．このうち 2 名は日本手話技能検定 1 級，1 名は 2 級を有する．各話者に対し，表 1 に示す 20 語を各 5 回産出してもらい，収録ミスが発生したサンプルは除外した．合計 400 サンプルのうち，最終的に 382 サンプルを HMD テストセットとして用いた．

テスト時には，HMD で収録した手骨格系列に対して，各サンプルの時系列長に間引き処理を行い $T = 50$ フレームに統一したうえで，モーションキャプチャデータ学習時と同様の手順で位置・速度・bone 特徴を算出し，ST-GCN に入力した．各特徴構成ごとに学習済みモデルを用意し，HMD テストセットに対する Top-1 および Top-3 精度を算出した．

4.3.3 実験結果

HMD テストセット 382 サンプルに対する単語認識精度を表 2 に示す．入力特徴を座標のみとした場合，Top-1 精度は 42.7％，Top-3: 65.5％であった．座標＋速度では Top-1 精度は 43.5％，Top-3 精度は 67.5％に向上した．座標＋ bone では Top-1 精度 65.2％，Top-3 精度 88.5％となり大きく改善した．座標＋速度＋ bone 特徴では Top-1 精度 70.7％が最も高く，Top-3 精度は 88.0％であった．

Q	設問	評価
Q1	短時間の説明だけで理解して使える	3
Q2	相手とのコミュニケーションに役立つ	5
Q3	本システムで対話を続けたい	5
Q4	初対面の相手との会話に使える	5
Q5	手話認識結果は発話とおおむね一致	2
Q6	認識誤りがあっても会話理解の大きな妨げにならない	4
Q7	待ち時間は許容できる	5
Q8	字幕（音声認識結果）は読みやすい	5
Q9	音声認識誤りがあっても推測に必要な情報は得られる	4
Q10	音声認識の遅延は許容できる	5

表 3 手話者による総合評価（5 段階）

Q	設問	評価
Q1	音声出力だけで意図を理解できた	2
Q2	ターン交代（次に話すタイミング）が分かりやすい	3
Q3	待ち時間/テンポは許容範囲	4
Q4	会話している感があった（作業感ではない）	4

表 4 対話者による評価（5 段階）

5. 総合評価

本章では，構築した対話支援システムを実際の手話者に試用してもらい，主観的な使いやすさと課題について評価を行った結果を述べる．ここで扱う評価は少人数による探索的なものであり，統計的に十分なサンプル数を用いた厳密なユーザスタディは今後の課題である．

5.1 評価方法

評価には，聴覚障碍を有し手話を主なコミュニケーション手段とする手話者 1 名（手話歴 15 年）と，対話者として健聴者 1 名（手話経験 2 年強）に協力してもらい，本システムを用いた対話タスクを実施した．手話者には HMD を装着してもらい，対話はすべて本システム経由で行うよう依頼した．

対話タスクは 3 種類のシナリオで構成し，各シナリオの冒頭に短い練習セッションを設けた．シナリオ (1) は対話者主導の一問一答を 10 個，シナリオ (2) は手話者主導の一問一答 10 個，シナリオ (3) は複数ターンの対話ラリーを 5 個とした．参加者それぞれには自分の発話部分のみを記載した台本を提示し，相手側の台詞は事前に知らせない条件で実施した．

全シナリオ終了後にアンケートを実施した．手話者にはシステム全体 4 項目・手話認識 3 項目・音声認識 3 項目の計 10 項目，対話者には計 4 項目の主観評価を実施した（設問内容は表 3，表 4 に示す）．自由記述では，本システムのよかったところ，悪かったところ，またそれに対する改善点といった率直な意見を記述してもらった．

5.2 評価結果

5.2.1 手話者による評価

手話者の 5 段階評価（表 3）では，導入容易性 (Q1=3) は突出して高い評価ではなかった一方，システム全体の有用性に関する項目（Q2–Q4）が高評価（5/5）であり，対話継続意向（Q3=5）も示された．一方，手話認識結果と実際の発話の一致度（Q5=2）は低く，認識精度に問題があることが示唆される．また，非動作時にも出力が発生する場合があるとのコメントが得られた．これは，手が HMD の計測範囲内にある限りシステムが常に手話の計測を行うため，意図しない動作が単語として出力されることに起因する．

音声認識については，字幕の読みやすさ（Q8=5）や遅延の許容性（Q10=5）が高く，音声誤認識も内容を理解できる範囲で生じたとコメントがあった．

自由記述では，良かった点として音声認識がある程度の変換が可能である点が肯定的に評価された．悪かった点と改善点としては，手話認識の精度向上と語彙数が挙げられた．

5.2.2 対話者（健聴者）による評価

対話者の 5 段階評価（表 4）では，待ち時間/テンポ（Q3=4）および会話している感（Q4=4）は比較的高かった一方，音声出力だけで意図を理解できたか（Q1=2）は低かった．また，ターン交代（次に話すタイミング）の分かりやすさは中程度（Q2=3）であった．この点については，本システムが手話認識結果を単語単位で逐次読み上げる構成であるため，発話の区切りや発話の終了が明確になりにくいことが一因として考えられる．

自由記述では，対面で会話できることで表情やリアクション等の非言語情報を同時に得られ，チャット等と比べて「相手の言葉を感じられる」点が肯定的に評価された．また，手話認識が不十分な場面でも，対面で相手の様子が見えるため，返答が遅い理由を推測できる点が利点として挙げられた．

一方で課題として，(i) 手話認識の誤りにより出力音声と実際の表現が乖離し，意図推定が難しくなる点，(ii) 音声話者は自分の発話がどの程度伝わっているかを判断しづらく，応答が返らない場合に「待つべきか，言い直すべきか」を決めにくい点，(iii) 音声合成で提示される単語列を保持しつつ文意を構成する負担がある点，が指摘された．特に (iii) では，誤った単語が繰り返し提示されると，どの情報を採用すべきか分かりにくく，手話を知らない場合ほど判断が困難になるとの指摘を受けた．

6. おわりに

本研究では，ヘッドマウントディスプレイを用いた一人称視点日本手話認識と，音声認識・音声合成を組み合わせた双方向対話インタフェースのプロトタイプを構築した．三人称視点で収録された KoSign のモーションキャプチャデータを HMD の手骨格形式に変換し，位置・速度・bone 特徴を入力とする ST-GCN を学習することで，HMD 上で取得した手骨格系列に対して 20 語の日本手話単語を認識できることを示した．シミュレーション評価では，HMD テストセット 382 サンプルに対して Top-1 精度 70.7％を確認し，三人称視点データのみで学習したモデルでも一人称視点データに一定程度適用可能であることが分かった．また，音声認識による対話者の文字起こしと，手話認識結果の音声読み上げを組み合わせることで，手話者とその対話者がそれぞれ自然なモダリティで情報を受け取れる対話支援インタフェースを実現した．

一方，本稿で扱った語彙は 20 語に限られており，連続手話認識による文レベルの表現には対応していない．また，モーションキャプチャデータのワールド座標系から HMD ローカル座標系への変換は線形な幾何変換にとどまり，HMD で取得される骨格との厳密な整合を保証するものではないため，依然としてデバイス間の差異は残っている．

総合評価では，手話者は本システムがコミュニケーションに役立つと感じ，対話継続意向も示した一方で，手話認識の誤りが生じた際に意図した単語が得られるまで言い直しを要することが課題として得られた．また，手が計測範囲内にある限り推定が更新され続けるため，非意図動作が誤出力を招く場面が見られ，手話の開始・終了の区間推定や発話終了の明示による出力制御の必要性が示唆された．対話者側では，対面により表情等の非言語情報を共有できる点は会話感の維持に寄与する一方，単語単位の音声提示では発話意図を文として把握しにくく，誤認識が混入すると，提示された単語列のどの部分が正しく，どの部分が誤りかを対話者が判断できないため，会話に用いる情報の取捨選択が困難であることことが確認された．さらに，音声認識結果が対話者に提示されないため，応答が遅れた際に「待機／言い直し」の判断が難しくなることが指摘された．

今後は，中納言日本語日常会話コーパスの頻度情報を用いて語彙を段階的に拡張しつつ，認識精度の向上を進める．あわせて，非意図動作の混入を抑えるため，手の動きの速度や加速度に基づく手話区間検出と，短区間の棄却や信頼度に基づく確定処理等による出力の安定化を検討する．さらに，連続手話認識により発話単位を推定し，推定結果を言語モデル等で文章化して音声提示することで，手話者の動作負担と対話者の解釈負担を低減し，円滑なターン交代を支援する．加えて，複数参加者によるユーザスタディを実施し，今回の評価で得られた知見に対する対応の妥当性を検証する．

謝辞

本研究では，国立情報学研究所 情報学研究データリポジトリ（IDR）データセットサービスを通じて工学院大学より提供されている「工学院大学 多用途型日本手話言語データベース KoSign」を利用した．

参考文献

[1] Yan, S., Xiong, Y., and Lin, D.: “Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition”, Proc. AAAI (AAAI-18), pp. 7444–7452 (2018).
[2] Nakamura, Y. and Jing, L.: “Skeleton-Based Sign Language Recognition with Graph Convolutional Networks on Small Data”, HCI International 2022 - Late Breaking Papers, LNCS 13519, pp. 134–142, Springer (2022).
[3] 長嶋祐二：工学院大学 多用途型日本手話言語データベース KoSign 概要説明書, 国立情報学研究所 情報学研究データリポジトリ，2023.
[4] Lugaresi, C., Tang, J., Nash, H., et al.: “MediaPipe: A Framework for Building Perception Pipelines”, arXiv:1906.08172 (2019).
[5] Cai, Y., Lu, T., Li, Z., Zhou, H., et al.: “SignGlass: First-Person View Comprehensive and Generalizable ASL Translation Using Wearable Glass”, Proc. UIST 2025, pp. 204:1–204:17 (2025).
[6] Fujimoto, T., Kawamura, T., Zempo, K., and Puentes, S.: “First-person View Hand Posture Estimation and Fingerspelling Recognition Using HoloLens”, Proc. IEEE GCCE 2022, pp. 323–327 (2022).
[7] Macháček, D., Dabre, R., and Bojar, O.: Turning Whisper into Real-Time Transcription System, Proceedings of the 13th IJCNLP and the 3rd AACL: System Demonstrations, pp. 17–24, 2023.
[8] 小磯花絵・天谷晴香・居關友里子・臼田泰如・柏野和佳子・川端良子・田中弥生・伝康晴・西川賢哉・渡邊友香：「『日本語日常会話コーパス』設計と特徴」，国立国語研究所論集，24，pp. 153–168，2023.
[9] Xin, C., Kim, S., Cho, Y., and Park, K.S.: “Enhancing Human Action Recognition with 3D Skeleton Data: A Comprehensive Study of Deep Learning and Data Augmentation”, Electronics, Vol. 13, No. 4, 747 (2024). doi:10.3390/electronics13040747.