多様な研究分野に利用可能な超高精細・高精度手話言語データベースの開発

多様な研究分野に利用可能な超高精細・高精度手話言語データベースの開発

長嶋祐二 (工学院大学) *, 原大介 (豊田工業大学) , 堀内靖雄 (千葉大学)
酒向慎司 (名古屋工業大学) , 渡辺桂子 (工学院大学) , 菊澤律子 (民族学博物館)
加藤直人 (NHK 放送技術研究所) , 市川熹 (千葉大学/工学院大学)

Development of the Super High-Definition and High-Precision Japanese Sign Language Database Available for Various Research Fields

Yuji Nagashima (Kogakuin University)
Daisuke Hara (Toyota Technological Institute)
Yasuo Horiuchi (Chiba University)
Shinji Sako (Nagoya Institute of Technology)
Keiko Watanabe (Kogakuin University)
Rituko Kikusawa (National Museum of Ethnology)
Naoto Katho (NHK STRL)
Akira Ichikawa(Chiba University/ Kogakuin University)

要旨

手話は言語であるにもかかわらず、音声言語と比べて言語学、工学を含む関連諸分野での研究が進んでいない。本稿では、各個分野における手話研究および学際研究の推進を目的とした、様々な分野の研究者が共通に利用できる汎用的な日本手話の語彙データベース作成について報告する。

言語学者の望むデータ形式と、工学や認知科学の分野で望むデータの形式は異なることが予想される。多分野での利用を可能にするためには、分析や解析内容に応じて手話の多視点の画像、3 次元動作データ、深度画像など様々なデータ形式を含むことが望まれる。さらに、時間軸上で同期したこれらのデータを、各分析者が得意とするデータ形式で解析することを可能にする。データベース上の様々な形式データを同期解析できるアノテーション支援システムも開発する予定である。これにより、様々な視点からの同一手話の解析が可能となり、手話言語に関する新たな知見が得られることが期待できる。

1. はじめに

2006 年 12 月、「障害者の権利に関する条約 」が第 61 回国連総会で採択され、2008 年に 5 月に発効した。この条約の第二条では、「言語とは音声言語及び手話その他の形態の非音声言

* nagasima@cc.kogakuin.ac.jp

語をいう」と定義されている。さらに、第二十一条　表現及び意見の自由並びに情報の利用の機会では、「手話の使用を認め及び促進すること」と、謳われている。日本は、2007 年 9 月 28 日にこの条約に署名し、国内法の整備・改革を行い、2016 年 4 月 1 日に「障害者基本法」の最終改正が施行された。この様な現状の中、全日本ろうあ連盟では「手話言語法」の制定を目指し、活動が行われている。手話言語条例を成立させている自治体は、13 県、75 市、9 町の 97 自治体となっている (2017 年 4 月 1 日現在)。また、国に「手話言語法」の制定を求める意見書は、2016 年 3 月には、全国（47 都道府県・東京 23 区・1,718 市町村）で採択されている [1]。さらに、全国手話言語市区長会には、463 の首長が参加している (2018 年 6 月現在)。手話は、聴覚障害者のコミュニケーション手段の一つであり、音声言語とは異なる文法体系をもつ独立した対話型の自然言語である。手話を構成する要素は、手指動作と非手指動作である。手指動作は、手型、提示される位置、掌方向、および手の大局的な運動により表現される。非手指動作は、視線、頷き、表情、口形など手指動作以外の要素であり、マルチモーダルな機能を表現している。手話の手指動作は、両手で語を構成したり、利き手と非利き手とが独立てそれぞれ語を形成したりする。このように、手話は、音声言語と異なり、複数の調動器官により、線状的にも非線状的にも語を構成する複雑さが存在する。手話は言語であるにもかかわらず、音声言語と比べて言語学、工学を含む関連諸分野での研究が進んでいない。この原因の 1 つは、言語学者や工学者など様々な分野の研究者が共通に利用できる汎用的なデータベースが存在しないためである。しかし、言語学者の望むデータ形式と工学、認知科学の分野で望むデータの形式は異なることが予想される。多くの研究分野で手話を統一的に研究するには、同じ手話動作を各研究者のニーズに合ったデータ形式で提供することが望ましい。同一の手話動作を様々な視点や手法で分析し俯瞰することは、新たな知見を得る機会を増大させる可能性をもつ。

本報告では、手話語彙のデータベースの構築方法について検討い、2017 年度の結果について述べる。

2. データベースへの収録データ形式

日本語の音声や言語データは、国立情報学研究所において、音声資源コンソーシアム (SRC: Speech Resources Consortium) が設立され、日本語の研究の発展に寄与している。音声データは、時間軸方向の 1 次元データであり、様々な解析手法が提案され実用化されている。ビックデータ解析により、新たな音声認識技術が飛躍的に進展している。

一方、手話は複数の調動器官によって語が形成されるとされているが、その音素の構造すらはっきり定義されていない。手話の弁別的特徴や音素、形態素の詳細な分析のためには、手指動作や非手指動作の詳細な分析が必要と考える。音声データが時間軸方向の 1 次元データであったのに対し、手話のデータは時間軸方向の空間的な広がりをもつ 3 次元データである。音声のように、手話の動作データを数値的に解析を行うことで、新たな知見を得られる可能性が高い。しかし、音声に比較して次元数が多くなり、その複雑度はかなり高いと予測される。

さらに、手話のデータは、各研究機関により独自に集められ公開されているものはない。手話を分析するために必要な、3 次元空間上の手指や非手指の構成要素の数値データは公開されていない。このため、手話の数値的な分析には、どの程度の空間・時間分解能のデータが必要

かも不明である。各研究機関によって収録される手話のデータ形式は、動画像映像が多いと考えられる。しかし、利用目的、分析や解析手法が異なるため、カメラ台数や撮影方法、解像度など様々であり、共通に利用することは困難と考えられる。

高精度に手話動作を分析したり高品位な手話 CG 生成したりするには、高精度な動作の 3 次元計測を必要とする。そしてもし、同期して高精度な 3 次元手話動作データと 2 次元の手話映像が存在すれば、非手指動作を含めた手話の認識や動作分析において、3 次元動作がどのように 2 次元に縮退され時間軸方向へ進行しているのかの解析が可能となり、新たな手話理解・認識のための方法論を得ることが可能となると考えられる。

そこで、本データベース構築では、どの程度の空間・時間分解能のデータが必要かも分析できるように、現時点で可能な最高水準の精度の手話動作収録手法とデータ形式について検討を行なう。

2.1 3 次元動作データ

3 次元空間的かつ時間的に高精度にデータを計測する方法は、コストを考えなければ光学式モーションキャプチャ (以下、 MoCap とする) である [2]。文献 [2] の計測では、東映ツークン研究所において、1600 万画素のモーションカメラ 42 台を用いて $2 \times 2 \times 2$ m^3 を計測しているため、空間分解能は 0.5mm を、時間分解能は 120fps を実現している。撮影で用いた再帰性反射マーカは、手型と顔表情を高精度に計測のため直径 3mm を用いている。そこで、本データベースもこの 3 次元計測環境を用いる。これにより、手指動作ならびに非手指 動作の構成要素の詳細な解析を期待できる。

なお、表 1 に、再帰性反射マーカの情報を示す。

表 1　再帰性反射マーカ情報

body Region	Retro-reflective Markers	
	Diameter [mm]	Number
Face	3	33
Hand	3	24×2
Others	10	31
Total Number of Markers		112

2.2 映像データ

手話画像認識や対話分析では、より高解像度のカメラが望まれる。画像計測では、最低 2 台以上のカメラを必要としている。そこで、撮影カメラ構成として、画面解像度はフル HD の 1920×1080pixel、あるいは 4K の 3840×2160 pixel を、時間分解能は 60fps を 3 台用いる。

2.3 深度 (距離) データ

最近、ToF(Time of Flight) 方式により、安価でかつ比較的高精度に距離を計測可能なセンサが普及している。手話認識でもこの方式を用いる研究機関が多くなっている [3],[4]。しかし、時間分解能は最大 $30fps$ となっている。この深度計測システムでは、赤外線映像と通常の映像を同時記録できるメリットがあるので、データベースに収録する。

2.4 異種データの同期収録

本データベースでは、様々な分野での利用とその解析データを統一的に扱うため、2.1, 2.2, 2.3 で得られる手話 データの同期収録を目指す。同一の手話動作を様々な時間分解能や空間分解能で観察したり、分析したりすることの意義は非常に大きいと考える。

3. 語彙の選定方法と言語資料提供者

3.1 語彙の選定方法

紙媒体で出版されている手話の辞書は多く存在する。この中で、最も収録語彙数の多い辞書は、全日本ろうあ連盟から発行されている「日本語－手話辞典」である [5]。この辞書は、日本語語彙数にして約 6、000 語を収録している。また、比較的規模の大きい手話文データベースには、NHK の E テレの手話ニュースからの手話文データベースがある。この手話文データベースには、約 130,000 文、総単語数約 3,036,000 語 (異なり語数で約 76,000 語) となっている (2018 年 3 月時点)[6]。

そこで、提案する DB への収録語彙の選定では、NTT データベースシリーズ「日本語の語彙特性」第 9 巻「 単語親密度　増補版 」[7] と、国立国語研究所・ 情報通信研究機構（旧通信総合研究所）・ 東京工業大学 が共同開発した日本語の自発音声を大量にあつめて多くの研究用情報を付加した話し言葉研究用のデータベース「日本語話し言葉コーパス」[8]、および NHK の E テレの手話ニュースからの手話文データベース用いた。単語親密度では音声親密度の高いもの、日本語話し言葉コーパスと手話文データベースでは出現頻度の高い語彙から、選定候補語彙としている。そして、選定語彙候補の中から、「日本語-手話辞典」に掲載されている語彙を最終的にデータベース収録語彙と決める。ただし、「日本語-手話辞典」に掲載されていなくても、日常よく使われる手話単語と判断される場合には、その語彙も収録候補とする。

2020 年までに、約 5〜6 千語彙の抽出を目指す。2017 年には、テスト的に日本語ラベル数で 400 語彙の抽出作業を行った。そして、この 400 語彙に対して、異動作同義語を含めて 525 語彙の手話動作の確定作業を行った。収録する手話動作形の確定作業は、筆者と研究協力者のろう者の手話母語者、CODA の手話母語者との既存の辞書分析作業と話し合いで決定した。

3.2 言語資料提供者

構築を目指している DB は、最終年度に公開を予定している。そこで、手話語彙の収録は、DB のより広い応用を考えて、男性と女性の各 1 名で行う。言語資料提供者を選定するための条件は、

- 手話母語者の家系の手話母語者
- 撮影にある程度慣れている
- 手話の読み取りがしやすい
- 撮影した映像の公開を許諾する

とした。この DB を構築するプロジェクトの研究協力者の手話母語者の面接などを行うことで、男性 M(38 歳) と女性 K(39 歳) に決定した。

4. 手話語彙の収録方法

構築する DB では、2.1, 2.2, 2.3 での検討に従って、MoCap による 120 fps の超高精度な 3 次元動作、60 fps の Full HD の高精細の映像、30 fps の深度センサからの映像の 3 種類の異なるフレームレートのデータを 2.4 の目的により、同期させて収録することを目指す。この目的達成のため撮影は、専門の知識とノウハウを持つ東映のツークン研究所のモーションキャプチャスタジオで行うことにした。

2017 年に撮影した時の撮影機材の構成と同期の概念図を図 1 に示す。MoCap カメラには、Vicon の 1,600 万画素の T160 と V16 の合計 42 台を用いている。DB に収録するデータ形式は、C3D データ、BVH ファイル形式である。C3D データは、MoCap 用の再帰性反射マーカ点の座標と角度データである。BVH ファイル形式は、Biovision 社によって開発された BioVision Hierarchy によるフォーマット形式である。BVH ファイル内には再帰性反射マーカの点の位置情報は無く、モデル情報が記録されており、その内容は主に HIERARCHY 部と MOTION 部に分けられる。HIERARCHY 部には、キャラクタのスケルトン階層構造が定義されている。MOTION 部には、階層構造中の関節 (JOINT) に対しての位置や回転の値がオイラー角表示で記述されている。

Full HD の映像は、camcorder#01～#03 のカメラ 3 台により正面・左側・右側から撮影した。さらに、図 1 の camcorder #04 は、参考映像として 120 fps の Full HD の映像も収録している。この映像の再生速度は、30 fps となりスーパースローとなっている。Kinect #01 の深度センサには、Kinect v2 を用い赤外線画像と深度画像の収録を行っている。カメラやセン

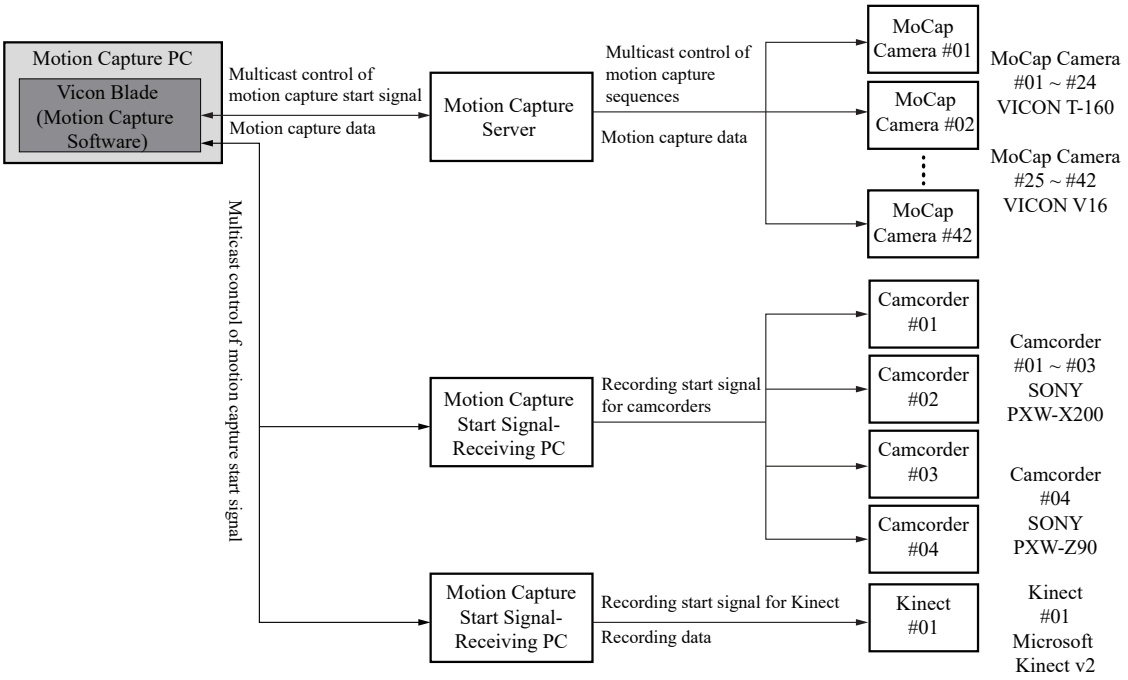

図 1　撮影機材の構成と同期の概念図

サの配置図を図 2 に示す。図 2 の長さと角度の詳細な値を表 2 に示す。また、2017 年撮影時の各カメラとセンサの仕様を表 3 に示す。図 3 に、東映東京撮影所モーションキャプチャスタジオでの DB 収録のためのスタジオの外観を示す。

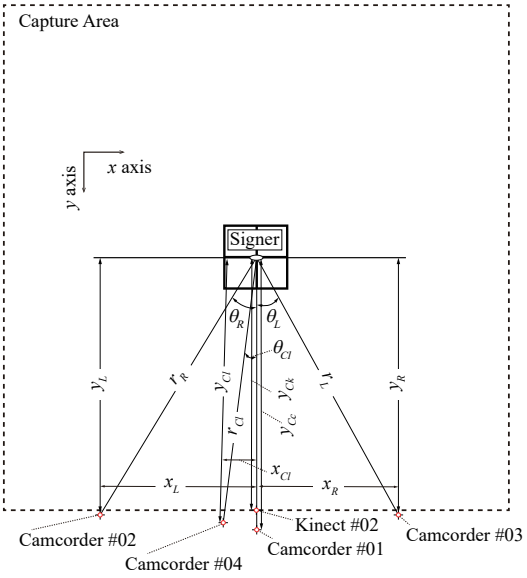

図 2　カメラとセンサの配置図

表 2　図 2 の長さと角度の詳細な値

Equipment	angle [degree]	distance from signer [m]			
		direct distance	y axis	x axis	height
Camcorder#01	0.0	3.27	$y_{Cc} = 3.27$	0.00	1.21
Camcorder#02	$\theta_L = 30.5$	$r_L = 3.41$	$y_L = 2.94$	$x_L = 1.73$	1.20
Camcorder#03	$\theta_R = 28.8$	$r_R = 3.40$	$y_R = 2.98$	$x_L = 1.63$	1.20
Camcorder#04	$\theta_{Cl} = 7.4$	$r_{Cl} = 3.17$	$y_{Cl} = 3.14$	$x_{Cl} = 0.41$	1.28
Kinect #01	0.0	2.51	$y_{Ck} = 2.51$	0.00	1.08

図 3　データ収録のスタジオの外観

表 3　2017 年撮影時の各カメラとセンサの仕様

Optical Motion Capture	
Model Number	VICON T160 (V16)
Frame rate	120 fps
Number of effective pixels	4,704 × 3,456
Number of camera	42
Number of markers	112
Camcorder	
Model Number	SONY PWX-X200
Frame rate	60 fps
Number of effective pixels	1,920 × 1,080
Format	MPEG-4 AVC/H.264
Number of camera	3
Super slow camcorder	
Model Number	SONY PWX-Z90
Frame rate	120 fps
Number of effective pixels	1,920 × 1,080
Format	XAVC
Number of camera	1
Depth sensor	
Model Number	Kinect One Sensor
Resolution	512 × 512
Horizontal field of view	70 degrees
Vertical field of view	60 degrees
Frame rate	30 fps
Number of sensor	1

5.　ビュワーの開発

前節で同期撮影されたデータは異なるフレームレートが混在している。3 次元データや 3 次元アニメーションなど複数の素材を描画することから、広くゲームの世界で利用されている Unity[9] により、同期再生可能なビュアーの開発を行う。ビュアーの開発方針は、任意の 4 種類のデータを同期して再生することである。2017 年度に開発した、ビュアーの主な機能を以下に示す。

- 画面を最大 4 分割して収録されている任意のデータを同期再生
- BVH ファイル形式データによる 3DCG の描画
- C3D データによるマーカ点の描画
- MoCap データは任意の視点と視野角で描画
- MoCap データの描画背景は任意データで可能

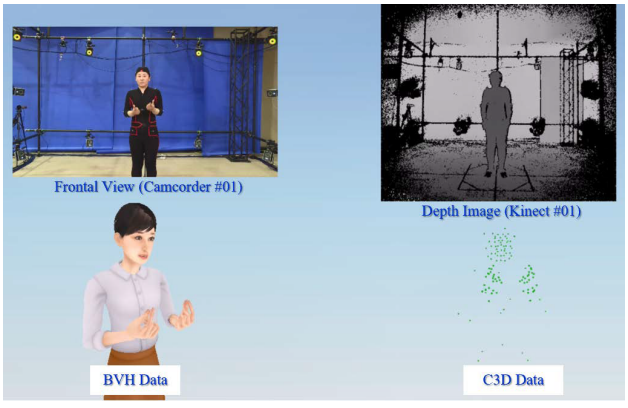

図 4　ビュアーによる各種データの同期再生画面 (手話単語 oNAJI(SAME))

- BVH ファイル形式描画には男性 2 モデル、女性 3 モデルからの選択
- 再生画面の録画 (動画キャプチャ) 機能

図 4 に、手話単語 oNAJI(SAME) の BVH データ (3DCG)、C3D データ (ドットデータ)、正面映像、Kinect センサからの深度データのビュアーによる同期された再生画面を示す。

しかし、Unity の機能の限界から、Full HD 動画と 3 次元動作データとの同期再生では、動画の再生が遅延する問題が発生した。そこで、この問題を解決するため、新たなビュアーの開発を開始している。

6. 今後の課題

本報告では、構築している多用途 DB の構築概念を述べた。構築する DB では、同期撮影されているため異なるフレームレートのデータを時間軸上で分析者の得意とするデータ形式で同期解析を可能としている。言語学、工学などの学際分野で利用可能な手話の言語、認識や生成など工学テータとして、コミュニケーションのモダリティ解析、ビッグデータ解析の礎になるなど様々な分野で利用可能なデータとする。これにより、様々な研究者の視点から同一手話が解析可能となり新たな知見が得られることが期待できる。

今後は、DB の拡張に向けた収録語彙の選定並びに、DB 上の様々な形式データを同期解析できるアノテーション支援システムも開発する予定である。構築する DB は、国立情報学研究所の情報学研究データリポジトリへの手話言語資料の登録も視野に入れて行う。

謝　辞

本研究の一部は JSPS 科研費 17H06114 の助成を受けたものです。

参考文献

[1] 長嶋祐二, 加藤直人, 山内結子, 河野純大：手話コミュニケーションのための情報保障技術, 電子情報通信学誌, Vol.101, No,1, pp.66–72, 2017.
[2] 渡辺桂子, 長嶋祐二：手話形態素辞書作成のための情報入力支援システム, 電子情報通信学会論文誌 D, Vol.J100-D, No.3, pp.298–309, 2017.
[3] Mika Hatano, Shinji Sako and Tadashi Kitamura：Contour-based Hand Pose Recognition for Sign Language Recognition, Proc. of 6th Workshop on Speech and Language Processing for Assistive Technologies, Sep. 2015.
[4] 古谷佳大, 堀内靖雄, 川本一彦, 下元正義, 眞崎浩一, 黒岩眞吾, 鈴木広一：“手話認識における位置・動き特徴量の検討”, 電子情報通信学会論文誌 D, Vol.J99-D, No.1, pp.90–92, 2016.
[5] 日本手話研究所 (編), 日本語–手話辞典, 全日本ろうあ連盟, 1999.
[6] 加藤直人, 内田翼, 東真希子, 梅田修一：ニュースを対象にした手話マルチメディアコーパスの構築, 言語資源活用ワークショップ (LRW2018), 2018.
[7] 天野成昭, 笠原 要, 近藤公久 (編著)：日本語の語彙特性 第 9 巻–単語親密度 増補版–, 三省堂, 2008.
[8] 日本語話し言葉コーパス, http://pj.ninjal.ac.jp/corpus_center/csj/ (2018 年 7 月 27 日参照).
[9] https://unity3d.com/ (2018 年 7 月 27 日参照).