4ZJ-01

読唇を用いた日本手話の映像データにおける口型認識

梅田 唯花　酒向 慎司

名古屋工業大学

1　はじめに

近年，聴覚障がい者と聴者のコミュニケーションの円滑化のため，手話からテキストへの自動翻訳の実現が望まれている．日本手話でも手話文の連続手話認識の研究が行われているが，認識可能な語彙数が限られているため，汎用性に欠けている．このような汎用性の欠如の理由として，手話コーパスの不足が挙げられる．

アメリカ手話(ASL)の大規模なコーパスの一つである How2Sign[1]は手話者 11 名による 79 時間の手話映像であり，語彙数は 16,000 あるとされている．日本手話に関していうと，例えば KoSign[2]は高精度な 3 次元データで記録され語彙数 6,000 と比較的多いが、話者数は 2 名であり、手話者数や時間数では十分とは言えない．

このような問題は，手話のアノテーションが手話の言語的理解を要し，時間のかかる作業であることが主要な原因の一つであると考えられる．そのため，アノテーションの効率化が必要である．

本研究では非手指信号の一種である口型に焦点を当て，日本手話映像における口型認識を行う．口型認識エンジンとして既存の読唇用のモデルを用いて，手話中の口型認識の精度を検証する．

2　手話のアノテーションの困難さ

手話は，手指の動きや位置で構成される手指信号と，顔の表情・口の形・うなずき・視線の動きなどで構成される非手指信号で表現される．手話は表現の構成要素が複数存在するため，複雑なアノテーションが必要となる．

例えば How2Sign には，複数の視点から録画された手話映像，身体・手・顔のキーポイントを含む 2 次元および 3 次元のポーズ情報が記録されており，翻訳されたテキスト，手話が表現する形態素の列（グロス列）がラベルとして付与されている．このときグロスのラベル付け作業は ASL 言語学者が行っていることからもわかるように，手話のアノテーションは手話の言語的な性質を理解した者が行う必要がある．

また，手話は一般的な表記法が普及していないため，アノテーションの結果が統一されないことが挙げられる．このような理由により，日本手話のデータ整備は進みにくい状況にあると考えられる．

3　手話の自動アノテーション研究

手話のアノテーションの効率化を目的とし，自動アノテーションの研究がいくつか行われている．手指認識技術を応用することで，手指信号認識を用いた方法が多く行われているが，手指信号の認識のみでは手話に対してすべてのグロスを付与することが困難であり，アノテーションとして不十分であるといえる．

自動アノテーションに関する研究の一つとして，イギリス手話(BSL)の大規模なデータセットである BOBSL の構築と並行して行われた研究[3]がある．この研究では，動画像を用いた手指信号の認識に加えて，口型認識によるマウジングのグロスラベル推定を行っている．これにより，手指信号だけでは検出が難しい語に対しても頑健にアノテーションできることが示されている．

そこで本研究では，日本手話のおける口型認識を行い，マウジングに対するグロスラベルの推定を試みる．

4　日本手話における口型

日本手話における口型には音声言語由来のマウジングと手話独自のマウスジェスチャがある．

マウジングは音声言語から借用されたもので，名詞，特に固有名詞や同じ動きで異なる意味を表現する同形異義語を表す際に使われることが多い．

一方，マウスジェスチャは音声言語とは関連付いていない口の動きである．動詞と一緒に表れることが多く，副詞的意味の付加や文の開始と終了を示す役割を持つ[4]．

また，前に述べた 2 つ以外にも，手話中では表情に付随して現れる口の動きなども見られる．

5　研究内容

本研究では，日本手話映像を用いて口型認識を行い，手話中のマウジングに対してグロスラベルを推定する．マウジングは名詞を表す際に用いられることが多い傾向があるため，日本語音声を書き起こしたテキストからマウジングの候補となる語と母音列を抽出し，それに該当するパターンを口径認識結果である母音のラベル列から特定することでマウジングが含まれる区間を検出する．本研究の全体図を図 1 に示す．

4 節でも述べたように，日本手話における口型

Recognition of mouth actions in Japanese sign language video using lip reading technique

†Yuika Umeda・Nagoya Institute of Technology

‡Shinji Sako・Nagoya Institute of Technology

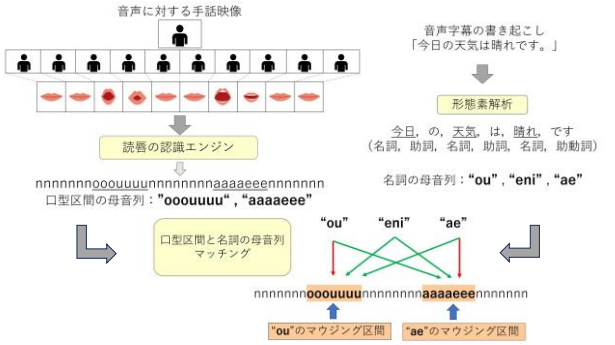

図 1. 提案手法の全体図

にはマウジング以外にもマウスジェスチャや表情に付随したものが存在する．そのため本来であれば，それらを分類してアノテーションする必要があるが，本研究では検出対象外とした．マウスジェスチャの表出は個人差も大きく、マウスジェスチャに該当しない口の動きとの弁別は容易ではないが、これらのアノテーションと分類手法の検討は今後の課題とした．

5.1 実験データについて

検証に使用するデータとして，手話者が日本語の音声に対応する日本手話で表現している映像を収集した．上半身を正面から撮影しており，フレームレートは 30fps である．手話者は 10 名で全員がネイティブサイナーである．収集したデータは文単位に分割し，日本語音声のテキストとの紐付けを行った．また，名詞を表すマウジングが現れたフレーム区間の開始・終了時点と母音列の正解ラベルを付与した．

5.2 口型認識によるマウジング検出

マウジングの認識には，口の形から発話内容を推定する読唇の技術を用いて検証を行う．前処理として，読唇モデルの入力画像となる，顔の下半分を切り出した LFROI (Lower half Face ROI)を手話映像から抽出している．顔検出には MediaPipe の FaceMesh を用いて検出されたランドマークをもとに，顔の傾きを補正する画像の回転と顔の大きさを補正するスケール変更の処理をした後，LFROI の切り出しを行う．

抽出された LFROI の画像から推定された母音列の中で，一定時間認識結果に変化がない区間では口型が現れていないとし，それ以外を口型区間とする．この口型区間の中から，マウジングが表出したとみられる区間を検出する．

マウジングは名詞を表現する際に使われることが多いため，今回は検出する対象を名詞に限定する．音声から書き起こしたテキストを形態素解析にかけ，テキスト内の名詞の母音列と口型区間の母音列に対してマッチングを行うことで，対象の名詞に対応する口型区間を検出する．

6 検証

検証として，認識結果から口型区間を検出し，口型区間の母音列と音声テキスト内の名詞の母音列の DP マッチングによって編集距離（Levenshtein 距離）を類似度として算出した．

この検証では，認識結果のノイズ処理として，認識結果の確率が 0.8 未満の場合に，次に確率が高い結果が次の口型の認識結果と同じであれば前の文字が続いているとみなす．同じ結果が 20 フレーム続いた場合に，口型が現れていない区間としている．表 1 に口型区間の検出結果と母音列の類似度を示す．

表 1. 口型区間の検出結果と母音列の類似度

マウジング単語（フレーム数）	検出結果（フレーム数）	類似度
被爆者(24)	nneeaaaannnaa…(212)	0.0173
昭和(8)		0.0115
三十一年(23)		0.0289
アウンサンスーチー(230)	aaaaeennnnnnneea…(212)	0.0432

1 つ目の口型区間は 3 つのマウジングを含み，2 つ目の口型区間は 1 つのマウジングに対して検出された．マウジング以外の口型も含まれて検出されてしまうことがあることが分かった．

また，マウジングのラベリング作業を通じて、マウジングは必ずしもテキスト内と同じ形で出てくることとは限らないことがわかった．例えば、数字（3 桁ぐらいの数字）であれば、100 に該当する語だけマウジングを行う場合や、××だけマウジングをするなど、同じ語でも内容や手話者によって揺れがあることが確認された．以上のことから、マウジングの候補と口型ラベル列とのマッチング方法を改善する必要がある．

7 まとめ

本研究では，手話中の口型認識によるグロスラベルの推定を行う．検証では口型区間の検出と，母音列のマッチングを行った．今後認識結果に対するマッチングとノイズ処理の方法を検討する．

謝辞

本研究は科研費 23747929，23K11197，22509579 ならびに日比科学技術振興財団の支援を受けた．

参考文献

[1] Amanda, D., et al.: How2Sign: A large scale multimodal dataset for continuous American Sign Language. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),2021.

[2] 工学院大学."工学院大学 多用途型 日本手話言語データベース(KoSign)". 国 立情報学研究所 情報学研究データリポジトリ(2021).

[3] Albanie, S., et al.: BOBSL: BBC-Oxford British Sign Language Dataset. arXiv:preprint arXiv:2111.03635 (2021)

[4] 直人加藤，太郎宮崎．手話ニュースコーパスの拡張；ニュースに出現する口型の分析． 電子情報通信学会技術研究報告，Vol.115，No. 193，pp. 47-52, 08.2015.