コーディングと動画を併用した日本手話音節の適格性予測

高藤 朋史　　三輪 誠　　佐々木 裕　　原 大介
豊田工業大学
{sd18421, makoto-miwa, yutaka.sasaki, daisuke}@toyota-ti.ac.jp

1 はじめに

日本手話は語彙が少なく，新たな概念や表現が難しい語彙に対応するため，文献 [8] などにおいて，新しい手話が毎年提案されている．しかし，そのような手話の中には，ろう者が使用しづらい，意味が通じづらいと感じる，日本手話として不適格な音節が含まれている．このため，通訳などの場面で情報伝達が難しい，手話学習者が誤った手話を学習する，といった問題が生じている．原因は，音声言語における音節に対応する手話の表現（手話における音節）が日本手話の表現規則に反しているためだと考えられる．しかし，日本手話の表現規則は，現在明らかになっていない．そのため，音節の適格性の判断はろう者の直感に頼る必要があるが，既に作られた音節・今後新たに作成される音節すべての適格性を人手で判断するのはコストが大きい．加えて，ろう者による適格性の判断は個人差があるため，適格性の判断基準が必要となる．したがって，決まった基準に従った適格性予測を自動的に行うシステムが必要とされている．

日本手話音節の適格性の予測には，八幡ら [6] や高藤らの [7] の研究があるが，研究においても，新日本手話コーディングマニュアル [9] に沿ってコーディングされた手話のデータを用いて適格性の予測を行っており，コーディングデータは含まれる情報やデータ数が乏しいという問題がある．また，これらの研究では適格性ラベルの決定法も曖昧である．

そこで，本研究では，判断の一貫性が高いろう者の適格性ラベルを対象に，コーディングデータと動画データを併用した分類器を学習することで，信頼性の高い高精度な適格性予測を目指す．

2 関連研究

2.1 コーディングによる手話の記述

本研究では，新日本手話コーディングマニュアル [9] に従ってコード化された手話音節を使用する．音節はコード化により，音節タイプ・手型・位置・動き・接触・手のひらの向き・中手骨の向き，の 7 カテゴリにおける全 27 種類によって記述される．日本語-手話辞典 [10]，新しい手話シリーズ [8] に掲載されている音節を対象にコード化を行っている．それぞれの音節は，複数名のろう者により適格な手話音節か不適格な手話音節かの分類情報が付与されている．

2.2 コーディングを用いた手話音節の適格性予測

八幡ら [6] は，コーディングされた手話のデータセット日本手話音節の適格性予測に取り組んだ．この研究では，ある音節が日本手話として適格なのか不適格なのかを予測可能な，ロジスティック回帰モデルを用いて分類器を作成し，日本手話語の適格性予測を行った．手話のデータセットから詳細な特徴を抽出，そこから組合せ特徴を抽出し，学習を行なった．組合せ特徴を利用した場合の分類正解率は，利用しない場合と比べ大きな向上が見られた．この研究において，二つの組合せ特徴のみでなく，より多数の特徴の組合せを見ることでより良い精度で分類が可能ではないかと考察されている．また，深層学習を用いて適格性予測を行った研究 [7] も存在する．この研究では，多層ニューラルネットワークや畳み込みニューラルネットワーク（Convolutional Neural Networks; CNN）モデルの分類器を作成した．

3 提案手法

本研究では，判断の信頼性の高いろう者を選択し，そのろう者の日本手話音節の適格性ラベルに基づいて音節の適格性を予測する深層学習を用いた分類器の作成を行う．ラベルについては，既存手法で考慮していな

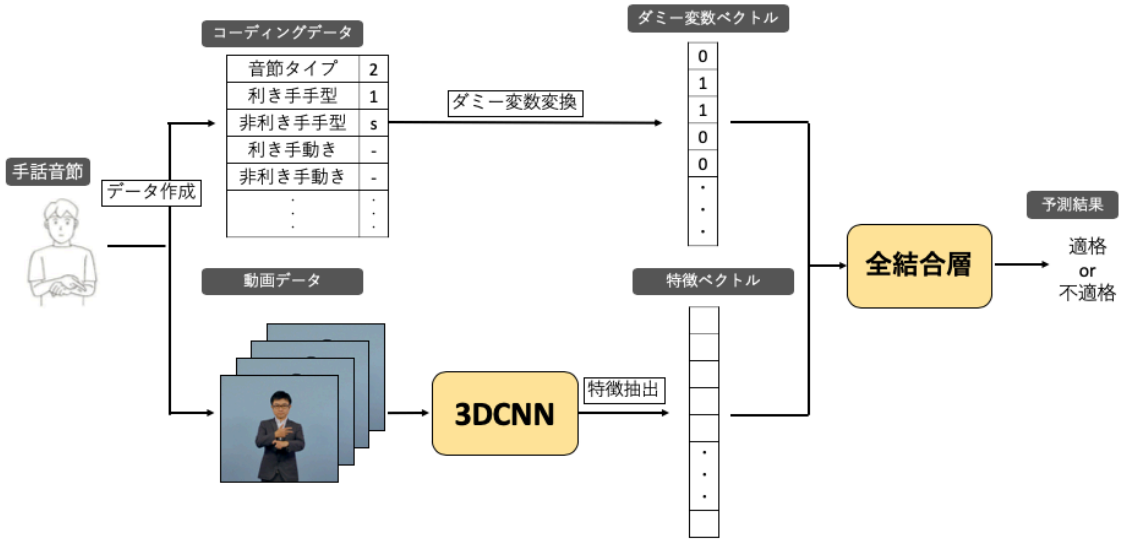

図1　提案する分類器モデルの概要

かったろう者の判断の一貫性に基づく信頼度を定義し，信頼度の高いろう者を選択し，このろう者の適格性ラベルを用いて日本手話音節の適格性予測実験に取り組む．分類器は 2 つの形式で表現された手話音節を入力とするモデルを提案する．具体的には手型や動きなどの構成要素を記述したコーディングデータ，ろう者による音節の表現を撮影した動画データの 2 形式を用いる．以降，適格性ラベルの決定法と分類器モデルについて説明を行う．

3.1　適格性ラベルの決定

これまで日本手話音節の適格性ラベルは複数名のろう者による判断をもとにした多数決により決定されてきた．しかし，この調査では一般的にデータセット作成で考慮すべきラベル判定者（以下，ワーカ）の信頼度を考慮しておらず，ワーカに正しいラベル判定能力があるか検証されていない．加えて，日本手話音節適格性の判断には大きく個人差があり，個人の中でも判断の揺れが大きい．このようなことから，これまでの適格性ラベル決定方法ではラベル信頼度が低いと判断し，新たにラベルの決定方法を検討した．

検討の結果，適格性判断の一貫性があるろう者を選択し，そのラベルを使用することとした．これは，データセット作成時，大量のデータに対してラベル付けを行うワーカは，判断基準が動かないように判断を下す必要があるためである．クラウドソーシングによるラベル付け作業においても，同じデータに対して複数回ラベル付けを行うことで判断の一貫性を保証している．一方で，これまで行われてきた日本手話音節データセット作成時のろう者による適格性判断は，ろう者の判断一貫性を保証する調査方法が取られていない．

具体的には，これまでに適格性判断を行った経験のあるろう者に対して，再び同じ音節の適格性判断の依頼を行い，ろう者の判断の一貫性を調査することとした．そして，回答結果より，2 回の判断を比較し，各ラベルの判断に対する F 値，分類正解率を求めることで信頼できるろう者を選定し，そのろう者のラベルを使用する．

3.2　分類器

これまでの日本手話音節の適格性予測には，コーディングデータが用いられてきた．しかしコーディングデータの問題点として，手の左右位置や速度などの情報が欠如している点，手話動画と異なり類似データセットが存在せず，拡張が困難である点などが挙げられる．近年の手話認識に関する研究では動画を用いた研究が中心となっている [1, 2, 4, 5]．このため拡張が容易であり，類似データセットの活用が見込めるため，本研究では日本手話の適格性解析に動画データを取り入れる．

提案する分類器の概要を図 1に示す．この分類器は，日本手話音節のコーディングデータと動画データの 2 形式を分類器への入力としている．データ形式による予

測精度の違いを検討するため，各モデルは基礎的なモデルとし，コーディングモデルは多層ニューラルネットワーク，動画データは 3DCNN によって予測を行う．コーディングデータから 2 値のダミー変数へ変換した特徴ベクトルを，日本手話動画から 3D 畳み込みニューラルネットワーク（3D Convolutional Neural Networks; 3DCNN）で動画の特徴ベクトルを，それぞれ抽出し，結合し，全結合層に入力することで音節の適格性を予測する．

4　適格性ラベルの決定

4.1　実験設定

ろう者の適格性判断の一貫性を調査するため，5 人のろう者に，十分な期間を空けて，2 度に渡って，同じ手話音節への適格性判断を行ってもらった．調査は 2 つの動画データセットについて行われた．対象のデータセットはろう者が単独の手話音節を表現している様子が単視点で撮影されている．データ数は 1 つ目のデータセットで 216 件，2 つ目のデータセットで 114 件となっている．ここでは，ろう者による判断を適格と不適格の二値として判断を行うこととした．判断の一貫性は一度目の判断を正解とし，二度目の判断で同じ回答ができているかを測定した．評価指標は適格と不適格における F 値とした．

4.2　結果と考察

調査結果として，各データセットにおける評価を表 1，表 2に，それをまとめたもの表 3に示す．各データセットの結果より，適格・不適格の両方において，ろう者 B と E の判断が高い一貫性を示した．その中でも，全データにおける F 値において適格と不適格どちらも上回ったろう者 B の適格性判断が最も信頼できると判断した．このため，本研究での適格性予測に使用する適格性ラベルはこのろう者 B の判断を採用することとした[*1]．

5　コーディングと動画を併用した適格性予測性能の評価

5.1　実験設定

コーディングデータと動画データを併用することの有効性を示すため，提案手法において，片方の形式のベクトルのみを利用し，もう片方を取り除いた，コーディングデータのみを扱うモデル（コーディングモデル）と動画データのみを扱うモデル（動画モデル）を比較対象として用意した．最適化手法は Adam[3] とした．また，評価指標は分類正解率とする．

5.2　結果と考察

実験結果を表 4に示す．結果より，動画データを用いた分類器はコーディグデータを用いた分類器よりも分類正解率が 0.005 下回った．また，コーディングデータと動画データを用いる分類器は，コーディングデータを用いた分類器よりも分類正解率が 0.04 上回った．

まず，コーディングデータを用いた分類器による結果と，動画データを用いた分類器による結果の比較を

表1　二値分類としたときの再調査結果（データセット 1）

ろう者	適格			不適格		
	適合率	再現率	F 値	適合率	再現率	F 値
A	0.64	0.83	0.72	0.50	0.26	0.34
B	0.86	0.80	0.83	0.78	0.84	0.81
C	0.87	0.94	0.91	0.42	0.24	0.31
D	0.42	0.72	0.53	0.94	0.80	0.86
E	0.71	0.88	0.79	0.87	0.70	0.78

表2　二値分類としたときの再調査結果（データセット 2）

ろう者	適格			不適格		
	適合率	再現率	F 値	適合率	再現率	F 値
A	0.80	0.83	0.81	0.38	0.33	0.35
B	0.93	0.86	0.89	0.57	0.74	0.64
C	0.99	0.96	0.98	0	0	0
D	0.29	0.77	0.43	0.96	0.76	0.85
E	0.88	0.99	0.93	0.93	0.54	0.68

表3　二値分類としたときの再調査結果（全データ）

ろう者	適格			不適格		
	適合率	再現率	F 値	適合率	再現率	F 値
A	0.69	0.83	0.76	0.46	0.28	0.35
B	0.89	0.83	0.86	0.74	0.82	0.78
C	0.92	0.95	0.93	0.35	0.24	0.28
D	0.38	0.73	0.50	0.94	0.79	0.86
E	0.79	0.93	0.85	0.88	0.67	0.76

[*1] ろう者内での判断のずれやろう者間の判断の違いに関する考察については，今後の課題とする．

行う．分類正解率において動画データが 0.5 下回る，ほぼ同程度の結果となった．コーディングデータは専門家による高コストなコーディング作業が必要となるデータであり，適格性に寄与するであろう要素を抽出したものである．それに対して動画データは撮影するのみの低コストで得られるデータである．データ作成のコストを考えると，動画データのみでコーディングデータとほぼ同程度の精度が得られることは動画データの有用性を示していると考える．今回，動画データは単純なモデルによって分類を行ったが，近年の手話認識で用いられる手法によって，より高精度な分類ができる可能性がある．

次に提案手法であるコーディングデータと動画データの併用は，どちらかのデータしか用いない分類器と比較すると分類精度が向上することがわかった．これにより，動画とコーディングは異なる特徴を捉えていること，提案手法である複数データの併用が日本手話音節の適格性分類に有効であることがわかった．

6　おわりに

本研究では，ろう者の判断の一貫性をもとに決定した適格性ラベルを用いた，コーディングデータと動画データの併用した高精度な日本手話音節適格性予測の提案を行った．動画データを用いることでコーディングデータを用いる場合とほぼ同程度の適格性分類が可能であることが示された．また，コーディングデータと動画データの併用によって，どちらか一方を利用した分類器より高精度で適格性分類ができることがわかった．

謝辞

本研究は JSPS 科研費 JP18H00671 の助成を受けたものである．

表4　分類器による適格性予測結果

用いたデータの形式	分類正解率
コーディングデータ	0.730
動画データ	0.725
コーディングデータ・動画データ	0.770

参考文献

[1] Necati Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. Subunets: End-to-end hand shape and continuous sign language recognition. In *Proceedings of ICCV*, pp. 3075–3084, 2017.

[2] Cleison Correia de Amorim, David Macêdo, and Cleber Zanchettin. Spatial-temporal graph convolutional networks for sign language recognition, 2019.

[3] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

[4] Junfu Pu, Wengang Zhou, and Houqiang Li. Dilated convolutional network with iterative optimization for continuous sign language recognition. In *Proceedings of IJCAI*, pp. 885–891. International Joint Conferences on Artificial Intelligence Organization, 2018.

[5] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of CVPR*, pp. 1145–1153, 2017.

[6] Satoshi Yawata, Makoto Miwa, Yutaka Sasaki, , and Daisuke Hara. Analyzing well-formedness of syllables in Japanese sign language. In *Proceedings of IJCNLP*, pp. 26–30, 2017.

[7] 高藤朋史, 三輪誠, 佐々木裕, 原大介．深層学習を用いた日本手話音節の適格性解析．言語処理学会第 25 回年次大会発表論文集, pp. 486–489, 2019.

[8] 全日本ろうあ連盟, 日本手話研究所「日本手話確定普及研究部」．わたしたちの手話　新しい手話シリーズ．全日本ろうあ連盟出版局.

[9] 原大介. 新日本手話コーディングマニュアル, 2016（12/11 更新）.

[10] 日本手話研究所．日本語-手話辞典．全日本ろうあ連盟出版局.