手話話者の感情認識

船越孝太郎 [1]　Zhu Yaoxiong[2]
[1] 東京科学大学総合研究院未来産業技術研究所
[2] 東京科学大学工学院情報通信系
{funakoshi,zhuyaoxiong}@lr.first.iir.isct.ac.jp

概要

手話話者の感情認識には，理論的課題と実践的課題が存在する．すなわち，文法的顔表情と情動的顔表情の重なり，およびモデル学習に用いるデータの不足である．本稿はこれら 2 つの課題に対しクロスリンガルな設定で取り組む．具体的には，我々が新しく構築した日本手話話者の感情認識ベンチマークである eJSL solo と，字幕付きの大規模英国手話データセット BOBSL を用いる．実験により，(1) 音声言語側のテキスト感情認識が手話側のデータ不足を緩和し得ること，(2) 時間区間選択が性能に大きく影響すること，(3) ジェスチャー特徴量の導入が手話話者の感情認識を向上させることを示す．既存の音声言語 LLM より強いベースラインも提示する．

1　はじめに

感情認識は，自然言語処理における中核的トピックの一つであり [1]，より自然で共感的なシステムの実現に資する [2, 3]．そのようなシステムは社会的少数派にとっても多数派と同等か，あるいはそれ以上に重要であろうが，少数派は取り残されている状況にある．近年，手話言語処理 [4, 5, 6] にも注目が集まっているが，手話話者の感情認識はこれまでほとんど検討されていない．我々の知る限り，関連して存在するデータセットは，米国手話（ASL）の EmoSign [7] のみである．

そこで我々は，感情認識のための日本手話コーパス eJSL[1)] を構築した．eJSL は独話形式の eJSL solo と対話形式の eJSL dialog を含み，どちらも 2 名の手話話者による台本の演技として収録した．eJSL solo では，78 の発話を各 7 つの異なる感情状態で表出してもらい，計 1,092 本のビデオクリップを収録した．eJSL dialog では，各 4 発話からなる 480 対話について，発話毎に 1 つ指定された感情状態で表出してもらい，計 1,920 本のビデオクリップを収録した（但し dialog で用いた感情状態は 4 つのみ）．

手話話者の感情表現は，顔表情が文法情報と感情情報の双方を担うために，音声言語話者のそれよりも複雑になる [8, 9]．例えば，眉の動きは疑問文 [10] または驚き [11] を表現しうる．非手話話者データで学習した感情認識モデルではこのような多義性を解消することは困難である．

この課題に対処するため，本稿では次の 3 つの仮説を検証する．すなわち，(1) 音声言語におけるテキスト感情認識は，手話感情認識におけるデータ不足を緩和し得る，(2) 文法表現の影響を受けにくい時間区間を選択することは性能に大きく影響する，(3) ジェスチャー特徴量の導入が手話話者の感情認識を向上させる．eJSL solo を用いた実験により，いずれの仮説も支持された．

2　感情認識と手話

前節で述べたように，手話話者の感情認識に固有の課題は，文法的顔表情（GFE）と情動的顔表情（AFE）の重なりにある．手話では，顔表情や視線，頭部動作などを用いて疑問，否定，強調などの文法情報を符号化する．これらの信号は AFE（感情情報）と同時に現れることが多く，両者の分離はコミュニケーション情報を正確に理解する上で重要である [12]．

この課題に対し，da Silva ら [12] は，手話コーパスに対して顔のアクションユニットラベルを付与することで GFE を符号化した．しかしながら，感情ラベルは付与していない．感情ラベルのアノテーションを持つ手話コーパスとしては，前述の EmoSign と我々が構築した eJSL のみが存在し，いずれも小規模なベンチマーク指向のデータセットである．したがって，教師あり学習に利用可能なデータの不足がもう一つの大きな課題となる．

人間のマルチモーダルコミュニケーションでは，

1） https://dataverse.harvard.edu/dataverse/eJSL

言語情報と非言語情報が相補的に用いられるが，場合によっては感情極性が矛盾することもある．そのような矛盾状況では，顔情報がより支配的になることが報告されている [13]．一方で，通常の状況では，マルチモーダル感情認識におけるもっとも強い手掛かりがテキストモダリティであることが繰り返し確認されている（[14, 1] 等）．従って，音声言語の字幕が利用できる手話コーパスでは，手話話者の感情状態を推定するのに音声言語による翻訳テキストを活用できる可能性がある．本稿ではこの可能性を検討する．

3 データセット

本研究では，手話データセットとして eJSL，EmoSign，BOBSL の 3 つを用いる．eJSL と EmoSign は評価に用い，BOBSL は評価に加えて弱ラベル付き学習データの構築にも用いる．これらのデータセットは異なる手話言語によるものであるが，本稿では感情認識という課題に対する当面の影響は限定的であると一旦仮定する[2]．

3.1 eJSL

eJSL（emotional Japanese Sign Language）は，我々が新たに収集したデータセットである．前述のように eJSL コーパスは eJSL solo と eJSL dialog を含むが，以降では eJSL solo のみを扱い，断りなく eJSL と書くときは eJSL solo を指す．

eJSL solo では 7 つの基本感情（中立を含む）を対象とし，2 名の協力者が 78 の発話で演じた．合計で 1,092 クリップから構成される（図 1 に例を示す）．協力者は 2 名とも職業的なろう俳優として活動する JSL のネイティブ話者である．両名とも，日本語の読み書きにも堪能であり，収録時の指示や発話文は日本語テキストで提示した．収録は，文書による同意を得た上で 2025 年 2 月に実施し，適切な謝金を支払った．

各クリップは，単一の感情指定を持つ完結した JSL 発話である．78 の発話文は公開スクリプト[3] を基に，手話通訳者との相談の上で，ろう者向けに改変して使用した（例：固有名詞の代名詞化，擬音語の回避など）．

3.2 EmoSign

EmoSign [7] は，既存の ASL コーパスを元にして感情認識のために編纂・ラベル付与されたデータセットであり，200 クリップからなる．本稿では，単一の支配的感情でラベル付けされた *Single Expression Set*（140 クリップ）を用いる．EmoSign は 10 種類の感情カテゴリを含むため，eJSL の 7 感情のラベル集合に対応付けて用いた（対応付けは付録 A を参照）．[7] は画像入力が可能な大規模言語モデルを用いたベースライン性能も報告しており，本稿ではその再現および比較も行う．

3.3 BOBSL

BOBSL [15] は，BBC 番組に対する 39 名の手話通訳による BSL 手話映像を含み，10 万個以上のビデオクリップで構成される．本稿では，BOBSL に含まれる字幕データに対してテキスト感情認識（TER）モデルを適用し，7 つの基本感情の弱ラベルを付与した大規模訓練データを構築する．

まず，公式の区分に従い，2 つの部分集合に分けた．自動字幕アラインメントに基づく **BOBSL-A**（113,826 クリップ）と，手動字幕アラインメントに基づく **BOBSL-M**（34,046 クリップ）である．

次に BOBSL-M から一部（1,438 クリップ）を抽出し，対応する字幕のみをみて 2 名のアノテータが感情ラベルを付与した．さらに 2 名のラベルが一致した部分集合を **BOBSL-M_C** として抽出した（付録 B 参照）．

最後に，事前学習済み TER モデル[4] を BOBSL-A に適用してラベルを自動付与した．TER モデルには，複数の候補の中から BOBSL-M_C に対して最も良好であったモデルを採用した．得られた感情ラベル付きデータを **BOBSL-A_TEA** と呼ぶ．

4 実験

本節では，3 つの仮説を検証する．すなわち，(1) 字幕に対するテキスト感情認識（TER）に基づく弱ラベル付与は，手話感情認識のデータ不足を緩和する，(2) 文法表現の影響を受けにくい時間区間を選択することは性能を改善する，(3) ジェスチャー特徴量の導入が手話話者に対する感情認識性能を向上

2） これはかなり強い仮定であり，実際には手話言語間・文化間の差異も当然影響するであろうが，それらは研究がもっと進呈して認識性能が高まってから課題として現れてくるものと考える．そのような影響がいつどのように現れてくるのかを明らかにすることも，今後の研究課題の 1 つである．

3） https://github.com/mmorise/ita-corpus/blob/main/emotion_transcript_utf8.txt

4） https://huggingface.co/michellejieli/emotion_text_classifier

図 1　発話「えっ，絶対嘘でしょ？早く嘘って言って.」を 6 つの感情で表現した例.

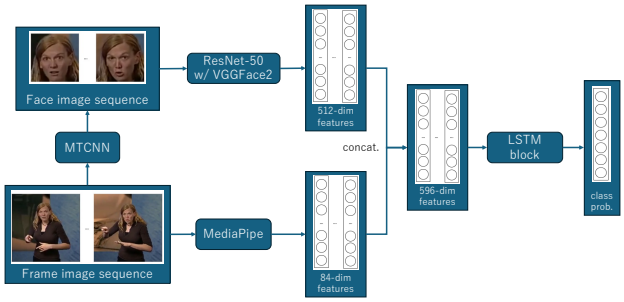

図 2　顔特徴と手特徴の双方を用いた EANwH のモデル構成.

させる，の 3 つである.

4.1　感情認識モデルと評価指標

本稿の実験では，手話話者の感情認識モデルとして EMO-AffectNet [16]（以降，EAN とよぶ）を採用する．EAN は顔表情認識（FER）を基盤とするフレームワークであり，本稿では手指ジェスチャ特徴を追加する小規模な拡張を行う．評価指標としては，文献に倣い weighted accuracy（wAcc）と macro F1 を用いる.

EAN を開発した Ryumin ら [16] は 8 つの感情データセットに跨る包括的なクロスコーパス評価を報告している．同フレームワークは，VGGFace2 で事前学習した ResNet-50 の FER バックボーンに，時間モデリングモジュールを組み合わせ，複数のデータ拡張とラベルバランシングを用いている．本稿では非手話話者データで学習された公開されているモデル重み[5]を用い，ベースラインとする．また EAN を拡張し，図 2 に示すように手の動き特徴を導入したモデル EANwH を構成する（実装の詳細は付録 C を参照）．実験では EAN および EANwH を，3.3 で用意した BOBSL-A_TEA で訓練する.

4.2　TER に基づく自動データラベリング

テキスト感情認識（TER）に基づく自動ラベル付与の有効性を検証するため，3.3 節で構築した BOBSL-A_TEA を用いて EAN を追加訓練し，BOBSL-M_C および eJSL で評価した.

表 1　自動付与した弱ラベルを用いてファインチューニングしたモデルの BOBSL-M_C に対する性能.

method	wAcc (%)	macro F1 (%)
EAN w/ non-signers data	15.54	12.12
EAN w/ BOBSL-A_TEA	**27.85**	**17.75**

表 2　自動付与した弱ラベルを用いてファインチューニングしたモデルの eJSL に対する性能

method	wAcc (%)	macro F1 (%)
EAN w/ non-signers data	7.41	9.25
EAN w/ BOBSL-A-TEA	**15.11**	**12.11**

表 1，2 に示すように，BOBSL-A_TEA による追加訓練は両データセットにおける認識性能を向上させた．TER に基づく弱ラベル付与が手話感情認識データの不足を緩和し得るという仮説は支持された.

4.3　時間区間選択

GFE が情動的手掛かりを覆い隠すのであれば，FER に用いる時間区間を手話発話を含まない時間区間に限定することで性能が改善するはずである．特に発話単位での演技データである eJSL を観察すると，手話発話後の区間（post-signing 区間）が感情的に顕著であるように見える．そこで次の 3 戦略を比較する．すなわち，(1) クリップ全体を入力（表 1, 2 の設定と同等），(2) 各クリップからランダムに 2 秒区間を選択，(3) 各クリップの手話発話後の 2 秒区間を使用[6].

表 3 の結果から，感情的に顕著な非手話発話区間を選択的に使用することは手話話者の感情認識に実際に有効であることが示された.

5）https://github.com/ElenaRyumina/EMO-AffectNetModel

6）手話発話末の検出は，動作強度の振幅値をもとに自動で行った．eJSL における手話録画に際しては，手話話者には各発話の後に 3 秒間静止するように指示していた.

表 3　eJSL における時間区間選択戦略の比較.

method	wAcc (%)	macro F1 (%)
Full Clip Input	15.11	12.11
Random 2s Segment	15.20	12.29
Post-Signing 2s Segment	**23.17**	**19.26**

表 4　eJSL における EANwH の性能.

method	wAcc (%)	macro F1 (%)
EAN (full clip)	27.85	17.75
EANwH (full clip)	**32.72**	**20.03**

表 5　BOBSL-M_C における EANwH の性能.

method	wAcc (%)	macro F1 (%)
EAN (full clip)	15.11	12.11
EAN (post-signing 2s)	23.17	19.26
EANwH (full clip)	**24.63**	**21.09**

4.4　ジェスチャー特徴量の導入

手の動きそのものがが感情表現となりうることに加え，手の動きは手話発話区間の手掛かりとなり得るため，手特徴を導入することで，モデルが非手話発話区間の顔特徴へ注意を向けるよう学習できる可能性がある．

表 5 および表 4 が示すように，手特徴の導入（EANwH）は BOBSL と eJSL の双方で有効である．特に eJSL では，EANwH（全クリップ入力）が，EAN による post-signing 区間選択と同等かそれよりも良好と思われる性能を示した．

4.5　画像入力可能な LLM との比較

EmoSign [7] と eJSL を用い，EANwH を画像入力が可能な LLM（Qwen 2.5，GPT-4o）と比較する．手順は [7] に従い，温度パラメータは 0 に設定した．

EmoSign を用いた表 6 の結果[7]，比較した LLM（Qwen 2.5，GPT-4o）よりも EANwH が良好であることを示唆する．特に Neutral において EANwH が大幅に優位である．BOBSL-M_C のクラス分布（付録表 8）からも分かるように，Neutral クラスは一般に多数を占めるため，それに対する性能は実アプリケーションにおいてユーザが感じる性能に大きく影響し得る．同一手順による eJSL での評価（表 7 ）でも，EANwH が最良の結果を得た．

7）　Qwen2.5 および GPT-4o については [7] で報告されている結果を 1 事例の誤差のみで再現できていた．

表 6　EmoSign における LLM と EANwH のクラス別 F1 および macro F1

Model	Joy	Sad.	Ang.	Dis.	Fear	Sur.	Neu.	Total
Qwen2.5	39.18	4.26	**28.57**	0.00	0.00	17.65	10.17	14.26
GPT-4o	38.38	**27.27**	0.00	**28.57**	8.33	0.00	0.00	14.65
EANwH	30.99	16.67	26.67	8.33	**10.53**	0.00	**25.00**	**16.88**

表 7　eJSL における LLM と EANwH のクラス別 F1 および macro F1

Model	Joy	Sad.	Ang.	Dis.	Fear	Sur.	Neu.	Total
Qwen2.5	20.91	**11.98**	2.53	12.10	9.57	1.27	19.84	11.17
GPT-4o	7.38	4.64	**15.93**	**23.79**	8.61	11.00	6.67	11.15
EANwH	**35.91**	10.64	15.55	14.29	**9.65**	**21.10**	**40.49**	**21.09**

4.6　考察

最も単純なベースラインからの改善は観察されたものの，全体の性能は依然として非常に限定的である[8]．ただし EANwH は手の動きを導入したとはいえ素朴な拡張であり，技術的改良の余地は大きい．

既存の別資源の活用も性能向上に寄与するだろう．本稿はクロスリンガル設定で注釈付きデータを利用したが，同一手話言語のデータを学習に加えることで改善が期待できる．

根本的には，EAN も EANwH も手話言語そのものを理解していない．通常状況ではテキスト内容が強い感情指標となり得ることを踏まえると，手話理解との統合も今後探索すべきである．

5　結論

手話話者の感情認識研究を前進させるため，新たなベンチマークデータセット eJSL を提出した．なお本稿では eJSL solo だけを用いた．eJSL dialog の利用は今後の課題である．また話者を 2 名しか含まないため，その拡充も重要な課題である．

eJSL に加え，BOBSL と EmoSign を用いた実験により，テキスト感情認識（TER），時間区間選択，およびジェスチャー特徴量の有効性を実証した．eJSL が手話話者および手話における感情認識研究に貢献し，手話話者のための情動対応型支援技術の基盤となることを期待する．

8）　eJSL から 1 名の手話話者の 70 クリップ（各クラス 10）を抽出し，もう 1 名の手話話者と，データ収録に関与していない非手話話者 1 名に感情分類をさせたところ，F1 評価で，手話話者は 77.78，非手話話者は 57.85 を達成した．前者は実用における上限，後者は下限とみなせる．

6 謝辞

eJSL コーパスのデータ収集は，公益財団法人 立石科学技術振興財団の支援を受けた（研究代表：東京科学大学 川上玲）．協力者の推薦・依頼に尽力いただいた早稲田大学の牧野遼作氏，国立情報学研究所の岡田智裕氏には深く謝意を申し上げる．またデータ収集に用いた収録ソフトウェアの作成に協力いただいた東京科学大学の小尾賢生氏に感謝する．

参考文献

[1] Taeyang Yun, Hyunkuk Lim, Jeonghwan Lee, and Min Song. Telme: Teacher-leading multimodal fusion network for emotion recognition in conversation. In **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 82–95, Mexico City, Mexico, 2024. Association for Computational Linguistics.

[2] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. **IEEE transactions on pattern analysis and machine intelligence**, Vol. 31, No. 1, pp. 39–58, 2009.

[3] Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. **Pattern recognition**, Vol. 44, No. 3, pp. 572–587, 2011.

[4] Zhiyu Long, Xingyou Liu, Jiaqi Qiao, and Zhi Li. Sign language recognition based on facial expression and hand skeleton. In **Proceedings of the International Conference on Automation and Artificial Intelligence**. Southeast University, 2024. Available at: https://arxiv.org/abs/2407.02241.

[5] Kayo Yin, Terry Regier, and Dan Klein. American sign language handshapes reflect pressures for communicative efficiency. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 15715–15724, Bangkok, Thailand, 2024. Association for Computational Linguistics.

[6] Zhen Wang, Dongyuan Li, Renhe Jiang, and Manabu Okumura. Continuous sign language recognition with multi-scale spatial-temporal feature enhancement. **IEEE Access**, Vol. 13, pp. 5491–5506, 2025.

[7] Phoebe Chua, Cathy Mengying Fang, Takehiko Ohkawa, Raja Kushalnagar, Suranga Nanayakkara, and Pattie Maes. EmoSign: A multimodal dataset for understanding emotions in american sign language, 2025.

[8] Diane Brentari. **A Prosodic Model of Sign Language Phonology**. The MIT Press, 02 1999.

[9] Ronnie Wilbur. Phonological and prosodic layering of nonmanuals in american sign language. In **Sign Language & Linguistics**, 2000.

[10] Roland Pfau and Josep Quer. **Nonmanuals: their grammatical and prosodic roles**, p. 381–402. Cambridge Language Surveys. Cambridge University Press, 2010.

[11] Clayton Valli and Ceil Lucas. **Linguistics of American Sign Language: An Introduction**. Gallaudet University Press, Washington, D.C., 2000.

[12] Emely Pujolli da Silva, Paula Dornhofer Paro Costa, Kate Mamhy Oliveira Kumada, José Mario De Martino, and Gabriela Araújo Florentino. Recognition of affective and grammatical facial expressions: A study for brazilian sign language. In **ECCV 2020 Workshops**, pp. 218–236. Springer, 2020.

[13] Albert Mehrabian. **Silent Messages**. Wadsworth, Belmont, CA, 1971.

[14] Dongyuan Li, Yusong Wang, Kotaro Funakoshi, and Manabu Okumura. Joyful: Joint modality fusion and graph contrastive learning for multimodal emotion recognition. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 16051–16069, Singapore, 2023. Association for Computational Linguistics.

[15] Samuel Albanie, Gül Varol, Lida Momeni, Triantafyllos Afouras, Xuankai Ma, Yuting Wang, Joon Son Chung, Helen Bear, Thomas Hain, Stephen Cox, Patrick Buehler, and Andrew Zisserman. BBC-Oxford British sign language dataset. **arXiv preprint arXiv:2111.03635**, 2021.

[16] Elena Ryumina, Denis Dresvyanskiy, and Alexey Karpov. In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study. **Neurocomputing**, Vol. 514, pp. 435–450, 2022.

A　EmoSign と eJSL との対応付け

eJSL は 7 種類の感情ラベル，EmoSign は 10 種類の感情ラベルを持つ．そのため表 8 のように，両者の感情ラベルを対応付けた．対応付けに加え，EmoSign のラベル毎のデータ件数も示す．eJSL と異なり，既存の手話コーパスにアノテーションしており，感情ラベルの数は不均等である．

表 8　EmoSign（Single Expression Set, N=140）の感情分布と Ekman の基本感情への対応付け．

eJSL (Ekman)	EmoSign	Count
Joy	Happyness	54
Sadness	Sadness	10
	Frustration	19
Anger	Anger	3
Disgust	Disgust	10
Fear	Fear	7
	Worry	14
Surprise	Surprise_pos	5
	Surprise_neg	7
Neutral	Neutral	11

B　BOBSL の字幕データに対する手動感情アノテーション

BOBSL-M の 1,438 発話の部分集合に対して，英語字幕だけを見て 2 名のアノテータ（A1, A2）が手動で感情ラベルを付与した．2 名のアノテータのラベルが一致した部分集合をさらに **BOBSL-M_C** として抽出した．（表 9）．Gwet の AC1 を用いた A1 と A2 の間の一致度は 0.6176 で，中-高程度の一致率と評価された．

表 9　BOBSL-M 各部分集合における感情ラベル数．

Emotion	M_A1	M_A2	M_C
Joy	59	251	48
Sadness	37	110	25
Anger	35	92	26
Disgust	19	55	10
Fear	21	33	5
Surprise	34	47	8
Neutral	1233	850	808
Total	1438	1438	930

アノテーション指示

Task description: Use 7 emotion labels to annotate sentences in several text document. Each sentence can only correspond to exactly 1 emotion label. Use the annotation tool to label sentences.

Emotion category: Anger, Disgust, Fear, Joy, Neutral, Sadness, Surprise. For the “Neutral” label, it is used for the sentence that does not have an obvious emotion.

How to use the labeling tool: The tool shows the sentence to be annotated and its context, determine the emotion of the sentence to be annotated with its context. When labeling, each emotion maps to a key, just press a key to do the corresponding labeling: ’a’: ’Anger’, ’d’: ’Disgust’, ’f’: ’Fear’, ’j’: ’Joy’, ’n’:’Neutral’, ’s’: ’Sadness’, ’u’: ’Surprise.

Examples for each emotion category:

** The quoted sentences are from the internet.

** The unquoted sentences are from the dataset.

1. Anger:

"I can’t believe you did that! How could you be so careless?"

What the fuck is wrong with you?!

2. Disgust:

"The way they treated those poor animals is revolting."

Oh, horrible.

3. Fear:

（後略）

C　EANwH のアーキテクチャ

EMO-AffectNet [16] を拡張し，図 2 に示すようにジェスチャー（手の動き）特徴を導入したモデル（EANwH）を構成する．以下に，EANwH における特徴抽出と情報融合の概略を示す．

C.1　特徴抽出

EANwH では，顔画像と手の骨格データの双方から，モダリティ固有の特徴を抽出する．

顔特徴の抽出：　顔特徴については，[16] と同様の方法を用い，各フレームから 512 次元のベクトルを得て，下流の系列モデリングに与える．

手特徴の抽出：　MeidaPipe[9] によって各フレームの手の 2D キーポイントから 42 × 2 の特徴行列を構成し，これをフラット化して 84 次元のベクトルとし，時間系列入力の一部として用いる．

C.2　特徴の同期と融合

顔由来特徴と手由来特徴を統合するため，両者のベクトルを時間的に同期させた上で結合し，顔表情と手ジェスチャの高次の相互作用をモデル化する．

モダリティ融合したベクトル列を 2 層の LSTM（隠れユニット数 512 および 256）モジュールに入力し，ビデオ全体の系列依存性と情動ダイナミクスを捉える．パラメータ数は約 300M である．

9）https://github.com/google-ai-edge/mediapipe