J-014

マルチモーダル特徴量を用いたストレス推定モデルの構築
Construction of a Stress Estimation Model Using Multimodal Features

高鍋 俊樹 [1] 松本 和幸 [1] 木内 敬太 [2] 柏原 功太郎 [3]
Toshiki Takanabe Kazuyuki Matsumoto Keita Kiuchi Kotaro Kashihara
梅原 英裕 [1] 中瀧 理仁 [1] 沼田 周助 [1] 吉田 稔 [1] 康 鑫 [1]
Hidehiro Umehara Masahito Nakataki Shusuke Numata Minoru Yoshida Xin Kang

1. はじめに

ストレスは心身の健康に大きな影響を及ぼす要因であり、うつ病や不安障害などの精神疾患の発症・悪化と密接に関係している。現代社会では、職場や家庭など様々な環境においてストレスの増加が問題視されており、早期のストレス検知と対処が重要とされている[1]。従来のストレスチェックは、自己記入式のアンケートなど主観的手法に依存しており、被験者の認知的バイアスや記入時の状態に大きく左右される。そのため、より客観的かつ継続的にストレス状態を評価できる手法の開発が急務となっている。

一方で、近年のディープラーニング技術の発展により、人間の表情・声・言語といったマルチモーダルな情報を用いた感情推定が可能となってきた。人は日常的に表情や声のトーン、話し方などから相手の気持ちを推測しているが、これらの情報を統合的に処理してストレス状態を推定する AI の実用化は、依然として課題が多い。

本研究では、オンラインカウンセリングにおいて取得される映像および音声に基づくマルチモーダル情報を活用し、機械学習を用いてストレス状態を高精度に推定する手法の確立を目的とする。複数の特徴抽出モデルおよび学習用データセットを組み合わせることで、ストレス推定の精度向上を図る。本研究により、ストレスチェックの客観化および自動化の実現に貢献し、メンタルヘルス支援における新たな可能性を切り拓くことを本研究の目的とする。

2. 関連研究

近年、マルチモーダル感情認識において、Transformer ベースの手法が注目を集めている[2]。Shayaninasab ら（2023）は、テキスト、音声、映像の 3 モダリティを用いたマルチモーダル感情認識手法を提案している[3]。本研究では、各モダリティに対して事前学習済みの Transformer モデル（BERT、wav2vec 2.0、VideoMAE など）をファインチューニングし、それらから抽出した特徴を融合することで、感情を高精度に認識している。特に、各モダリティから抽出された特徴ベクトルを単純に連結し、サポートベクターマシン（SVM）によって分類を行う手法が、IEMOCAP データセットにおいて 75.42%の精度を達成している。融合手法としては、特徴レベルでの早期融合と意思決定レベルでの後期融合の両方を比較検討し、前者の方が高い性能を示すことが明らかとなっている。この研究は、マルチモーダル感情認識における Transformer モデルと融合戦略の有効性を実証しており、我々の研究におけるモデル設計や融合手法の選定においても重要な示唆を与えている。

1 徳島大学　Tokushima University
2 労働者健康安全機構　Japan Organization of Occupational Health and Safety
3 ワークスアプリケーションズ　Works Applications

3. 提案手法

3.1 使用データ

本研究では、独自に収集した 2 種類の日本語面談データセットを用いて実験を行った。

3.1.1 オンラインカウンセリングデータセット

本データセットは、50 名の日本語話者によるオンラインカウンセリングの映像・音声・文字起こしデータで構成されている[4]。各被験者の面談時間は約 30 分であり、音声データからの文字起こしには、日本語音声認識モデル「ReazonSpeech-NeMo v2」（https://huggingface.co/reazon-research/reazonspeech-nemo-v2）を使用した。

ストレス値は、2 名の産業医による 6 段階評価の主観的スコアの平均を用いて付与した。これらの平均スコアの中央値（3.53）を閾値とし、中央値以上を「高ストレス」、中央値未満を「低ストレス」として 2 値のラベル付けを行った。

3.1.2 うつ面談データセット

補助的な学習データとして、84 名分のうつ病面談データセット[5]を用いた。本データセットには、うつ病のスクリーニング尺度である Patient Health Questionnaire-9（PHQ-9）スコアが付与されており、スコアが 10 点以上のデータを「高ストレス」、10 点未満のデータを「低ストレス」としてラベル付けを行った。

本研究では、以下の 2 条件でモデルの検証を行った。

- オンラインカウンセリングデータセットのみを用いて学習を行う条件
- 上記に加え、うつ面談データセットを追加学習データとして利用する条件

3.2 使用特徴量

本研究では、音声・画像・言語の 3 モダリティにまたがる計 10 種類の特徴量を用いて、ストレス状態の推定を行った。以下に各特徴量の詳細を示す。

3.2.1 音声特徴量

(1) GeMAPS

OpenSMILE ツールキットを用いて抽出され、音声感情分析に広く用いられている定義済み特徴セットである GeMAPSv01a（Geneva Minimalistic Acoustic Parameter Set）を採用した。声の高さ、強さ、スペクトル特性などのパラメータが含まれている。

(2) HuBERT

HuBERT は、12 層の Transformer ブロックと各層に 12 個のアテンションヘッドを備えた自己教師あり学習モデルで

ある。本研究では、日本語音声約 19,000 時間を収録した音声コーパス「ReazonSpeech v1」に基づいて事前学習されたモデルを用いた。これにより、音声の潜在的な構造や音響的特徴を効果的に抽出する。

(3) wav2vec 2.0

Meta 社によって提案された wav2vec 2.0 アーキテクチャをベースにした日本語特化モデル「rinna/japanese-wav2vec2-base」を用いた。本モデルも 12 層の Transformer 構造を持ち、ReazonSpeech v1 によって事前学習されている。音声波形から直接特徴を抽出できる点が特徴である。

(4) speechbrain

wav2vec 2.0 ベースの音声分類モデルであり、感情分類タスクに特化している。IEMOCAP データセットで事前学習されており、分類対象の感情は、happy, angry, sad, neutral の 4 種類である。本研究では、speechbrain という音声認識や感情分析の機能が含まれるオープンソースツールキット内の「Emotion Recognition with wav2vec2 base on IEMOCAP」というモデルを使用した。

3.2.2 画像特徴量

(1) AU（アクションユニット）

表情筋の動きを表す AU は、Py-Feat ライブラリを用いて検出した。Py-Feat では、感情表現に関連する 20 種類の AU を、顔が写っているフレームごとに数値として抽出できる。これにより、顔の部位ごとの筋活動を定量的に表現し、微細な表情変化を系統的かつ測定可能な形で捉えることができる。

(2) face_direction（Pitch, Roll, Yaw）

Py-Feat により検出される頭部の向きの情報を利用した。具体的には、以下の 3 次元的な回転成分を特徴量として用いた。

- Pitch：縦方向の首の動き（うなずき）
- Roll：左右に傾ける動き（首をかしげる）
- Yaw：水平方向の回転（左右を向く）

(3) face_emo

発話ユニットにおいて顔画像が検出されたフレームに対し，得られた 7 種類の感情（anger, disgust, fear, happiness, sadness, surprise, neutral）の確率値ベクトルの平均をとったものを特徴量とした。

3.2.3 言語特徴量

(1) BERT

bert-base-multilingual-uncased モデルを、英語、オランダ語、ドイツ語、フランス語、スペイン語、イタリア語の製品レビューに対する感情分類タスクに適応させたモデルを用いた。本モデルは、レビュー内容をもとに 1〜5 の星評価を予測するタスクでファインチューニングされており、評価値を通して発話のポジティブ・ネガティブ傾向を捉えることが可能である。

(2) E5

多言語 C4 コーパスのアノテーション付きサブセットをもとに学習された感情特化型モデルであり、高次元ベクトルとして文の感情的意味を表現し、言語に依存しない感情推定を可能とする点に特徴がある。

(3) Sentiment_ja2

日本語テキストを対象に、6 種類の感情（angry, disgust, fear, happy, sad, surprise）を分類するロジスティック回帰ベースの感情分析モデルである。単純かつ軽量な構造ながら、日本語での感情傾向を捉えるのに有効である。

表 1　特徴量の次元数

特徴量名	次元数
GeMAPS	65
HuBERT	768
wav2vec 2.0	768
speechbrain	65
AU	20
face_direction	3
face_emo	7
BERT	768
E5	768
Sentiment_ja2	6

4. ストレス値の推定実験

4.1 分類器と交差検証

ストレス値を推定する 2 値分類タスクに対し、主分類器として LightGBM（Light Gradient Boosting Machine）を採用した。LightGBM は、勾配ブースティング決定木に基づく高効率な学習アルゴリズムであり、学習速度と精度のバランスが良好である点、また多数の特徴量や欠損値を含むデータにも柔軟に対応できる点から、マルチモーダル情報を扱う本研究に適していると判断した。

また、汎化性能を評価するため、被験者単位でデータを分割することで 50 分割交差検証を行った。各分割において、学習と評価を繰り返すことで、バラツキのある少数サンプル環境下でも信頼性の高い精度評価が可能となる。

4.2 特徴量組み合わせ

本実験では、音声、表情、顔の向き、言語といった複数のモダリティから抽出された 10 種類の特徴量を対象に、ストレス値推定性能の評価を行った。評価は、以下の 2 段階に分けて実施した。

（a）各特徴量を単独で用いた分類実験を実施し、どのモダリティがストレス推定において有用な情報を含んでいるかを検証した。これにより、単一特徴量ごとの識別能力を明確にし、後続の統合実験における基礎情報とした。

（b）複数の特徴量を組み合わせた状態での分類実験を設計した。マルチモーダル統合による相乗効果を検証し、音声＋言語、表情＋顔の向きなど、モダリティ間の補完関係に着目した構成も取り入れた。

5. ストレス推定モデルの評価と考察

本節では、第 4 章で述べたストレス値の 2 値分類実験の結果を報告し、推定性能に基づく各特徴量の有効性を検討する。特に、単一特徴量による推定精度と、複数の特徴量を統合した場合の推定精度の変化に注目し、マルチモーダル情報の統合による効果を定量的に評価する。

5.1 単一特徴量による推定性能

まず、各特徴量を単独で用いた推定モデルの性能評価を行った結果を表 2 に示す。本実験では、特徴量ごとに LightGBM を用いたモデルを構築し、50 分割交差検証によって得られた Accuracy（平均値）を算出した。

表 2　単一特徴量におけるストレス推定精度

特徴量名	Accuracy
GeMAPS	0.54
HuBERT	0.56
wav2vec 2.0	0.60
speechbrain	0.52
AU	0.40
face_direction	0.58
face_emo	0.58
BERT	0.38
E5	0.62
Sentiment_ja2	0.46

5.2 複数特徴量の統合による推定性能

続いて、複数の特徴量を組み合わせて推定モデルを構築した場合の推定性能について検討した。複数の有望な特徴量を統合することで、モダリティ間の補完的な情報が得られることが期待される。本研究では、先行研究や単一特徴量の結果に基づき、有望と考えられる特徴量の組合せを複数選定し、同様に LightGBM および 50 分割交差検証により評価を行った。特に高い精度を示した特徴量の組合せを図 1 に示す。

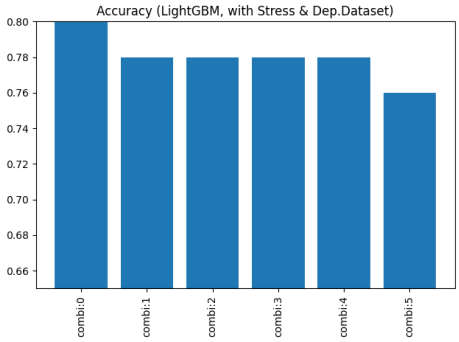

図 1　ストレス値の推定精度

表 3　上記図と特徴量名との対応表

図 1 中の表記名	特徴量の組合せ
combi:0	AU, wav2vec 2.0, HuBERT, E5, BERT, Sentiment_ja2
combi:1	face_direction, wav2vec 2.0, HuBERT, E5, BERT, Sentiment_ja2
combi:2	AU, face_emo, wav2vec 2.0, HuBERT, E5,BERT
combi:3	AU, face_emo, wav2vec 2.0, HuBERT, E5, BERT, speechbrain
combi:4	AU, wav2vec 2.0, HuBERT
combi:5	AU, face_direction, wav2vec 2.0, HuBERT, E5, speechbrain

これらの結果から、音声特徴量（HuBERT、speechbrain）と、顔の向きの変化（Pitch, Roll, Yaw）、および言語的な文脈を捉える BERT などの多様なモダリティを組み合わせることで、高い精度が得られることが示唆された。特に、HuBERT と face_direction の併用は、非言語的ストレス表出の可能性を示している。

6. 考察

本章では、前章において得られたストレス推定モデルの実験結果を踏まえ、各特徴量の寄与やマルチモーダル情報の統合による効果、ならびに本研究における成果について考察を行う。

6.1 単一特徴量の性能に対する考察

単一特徴量による推定実験の結果から、音声特徴量（HuBERT、wav2vec 2.0）および表情特徴量（face_emo、face_direction）は、比較的高い推定性能を示した。これは、音声や表情に含まれるストレスの兆候が、人間の声のトーン、話速、表情などに明確に現れるためと考えられる。特に HuBERT や wav2vec 2.0 は、音声から高次元な表現を抽出可能であり、感情状態の変化に敏感な特徴を反映している可能性がある。

一方、日本語感情特徴量（Sentiment_ja2、BERT）は、単独では高精度の推定には至らなかったが、音声や表情の情報と補完的に用いることで推定精度の向上に寄与する傾向が見られた。

6.2 マルチモーダル統合の効果

複数特徴量の統合による実験では、複数のモダリティを適切に組み合わせることにより、80%以上の推定精度が得られた。特に、音声、言語、顔の向き、顔表情といったモダリティ間の統合が、ストレスの多面的な特徴を捉える上で有効であることが明らかとなった。

この結果は、ストレスが心理的・身体的な複合的要因によって構成される現象であることと整合的であり、単一モダリティで捉えきれない側面を補完するというマルチモーダル統合の意義を支持するものである。たとえば、顔の向きの特徴量は一見単純であるが、HuBERT や BERT と併用することで、非言語的緊張や感情の揺れを含むストレスの検出に寄与したと考えられる。

さらに、speechbrain と HuBERT、BERT といった音声・言語処理に強みを持つ特徴量の併用が安定的な推定精度を示しており、モダリティ間の意味的な関連性を考慮した統合が重要であることも示唆された。

6.3 ストレススコアとの相関に基づく特徴量の分析

本研究では、各種モダリティにおける特徴量とストレススコアとの相関を分析し、ストレス推定に寄与する情報の傾向を明らかにした。本節では、各モダリティに属する特徴量群の相関統計に基づいて、情報の質やモダリティ間の特性について考察を行う。

6.3.1 ストレススコアとモダリティの相関傾向

ストレススコアとの相関分析の結果、特徴量ごとの相関係数の絶対値平均、さらに高相関（相関係数 ≥ 0.2）を持つ特徴量の割合を集計した結果を表 4 に示す。

表 4　ストレススコアとの相関係数

特徴量名	相関係数の絶対値平均	相関≧0.2 の特徴量数	相関≧0.2 の特徴量割合
GeMAPSv01a	0.1101	7	10.8%
HuBERT	0.1274	88	11.5%
wav2vec 2.0	0.1602	130	16.9%
speechbrain	0.1546	0	0.0%
AU	0.0808	0	0.0%
face_direction	0.1695	1	33.3%
face_emo	0.1182	2	28.6%
BERT	0.0739	16	2.1%
E5	0.1909	177	23.0%
Sentiment_ja2	0.1301	0	0.0%

ストレススコアとの相関が高い特徴量を多数含むモダリティとして、e5、wav2vec 2.0、face_direction が挙げられる。特に e5 では、768 次元中 177 次元（23.0%）が相関係数 0.2 以上であり、高密度に有用情報を含むモダリティであると考えられる。

wav2vec 2.0 も同様に 130 次元（16.9%）が高相関を示し、音声の物理的特徴や抑揚、速度などがストレスと関係している可能性が高い。一方で、HuBERT も同じく音声を入力とするが、相関が高い次元の割合は wav2vec 2.0 よりもやや低く、学習の焦点や特徴の抽出傾向に差があることが示唆される。

また、face_direction は 3 次元中 1 次元（33.3%）が高相関を持ち、割合としては最も高い。これは頭の傾きや揺れといった非言語的な身体表現がストレスの兆候と関係していると考えられる。

6.3.2 モダリティ間の相関傾向

モデル精度が高い構成では特定のモダリティが高頻度に使用されていることが確認された。これらのモダリティを、その相関構造と寄与の傾向に基づいて、中核的モダリティと補完的モダリティの 2 つに分類できる。

中核的モダリティとしては、wav2vec 2.0、HuBERT、BERT、E5 が挙げられる。これらは他のモダリティと高い相関を持ち、冗長性があるものの、学習における安定性の向上に寄与していると考えられる。特に wav2vec 2.0 と HuBERT は音声の時系列的情報を、BERT や E5 はテキストの意味情報を豊かに表現しており、感情全体の傾向を捉える上で重要な役割を果たす。

一方、補完的モダリティとしては、AU、face_direction、face_emo、speechbrain、Sentiment_ja2 が挙げられる。これらは相関が低く、独立した情報源として機能し、特に非言語的なストレス兆候や微細な感情の変化を捉える可能性がある。中核的モダリティでは見逃されがちな情報を補完する役割を果たすため、組み合わせて使用することにより、認識性能の向上が期待できる。

以上のことから、安定した情報源（中核的モダリティ）と、独立性の高い非言語的モダリティ（補完的モダリティ）を組み合わせることが、より高精度な感情認識を実現する鍵となると考えられる。

7. 結論と今後の展望

7.1 結論

本研究では、音声・表情・顔の向き・言語といったマルチモーダルな情報を用いて、教師データに基づくストレス値の推定モデルを構築し、その有効性を実証した。特に HuBERT や speechbrain による音声特徴量、ならびに顔の向きの特徴量(face_direction)、言語特徴量（BERT）を組み合わせた場合において高い推定精度（Accuracy ≧ 0.80）を達成しており、非言語的・準言語的情報の統合がストレス推定において重要であることが示唆された。

7.2 今後の展望

今後の展望としては、本研究で開発したストレス検出 AI モデルを、AI ストレスチェッカーアプリケーションへ統合し、日常的なストレスモニタリングに応用することを目指す。本アプリケーションは、アバターAI との対話やスマートフォンアプリを通じて、音声・表情・言語データを継続的に収集し、ストレス値を推定するものである。推定されたストレス値は、個人ごとに設定されたベースライン値と比較され、その逸脱度や持続時間に応じて、ユーザーにフィードバックが返される構造となっている。

また、本システムは個々人のストレス傾向を把握し、メンタルヘルス支援へとつなげることを目的としており、高頻度かつ低負荷なインタラクションによって、利用者の主観的な負担を最小限に抑えながら、客観的なストレス評価を実現する点が特徴である。さらに、今後は Transformer 系モデルや深層学習による時系列データのモデリングの導入も検討し、さらなる推定精度の向上を図る。

将来的には、ストレス推定技術の社会実装に向けて、対象ユーザーを拡大するとともに、ストレス状態の可視化だけでなく、ストレスのケアへの橋渡しとなるようなフィードバック設計、そしてプライバシーへの配慮を踏まえたシステム設計の最適化を行うことで、より実用的かつ倫理的に健全な AI ストレスモニタリングの実現を目指す。

謝辞

本研究の一部は、令和 6 年度徳島大学ものづくり未来共創機構実証研究推進プロジェクト、公益財団法人 JKA 令和 6 年度開発研究補助事業、国立研究開発法人科学技術振興機構（JST）大学発新産業創出基金事業スタートアップ・エコシステム共創プログラム「PSI・GAP ファンド支援プログラム（ステップ 1）」により実施されました。深く謝意を表します。

参考文献

[1] 厚生労働省, ストレスチェック制度の効果的な実施と活用に向けて, 令和 4 年 3 月,（参照 2025 年 6 月 8 日）https://www.mhlw.go.jp/content/000917251.pdf

[2] 立石 修平, 大杉 康仁, 中辻 真 : 非言語モダリティへの言語性付与による感情推定, 人工知能学会論文誌, 2025, 40 巻, 3 号, p. D-O92_1-10, 公開日 2025/05/01

[3] Shayaninasab, Minoo, and Bagher Babaali. "Multi-Modal Emotion Recognition by Text, Speech and Video Using Pretrained Transformers." *arXiv preprint arXiv:2402.07327* (2024).

[4] T. Takanabe, K. Kashihara, K. Matsumoto, K. Kiuchi, X. Kang, R. Nishimura and M. Sasayama: Multimodal Emotion Recognition and Dataset Construction in Online Counseling. *Proc. of the 38th Pacific Asia Conference on Language, Information and Computation*, 1-9, Dec. 2024.

[5] 柏原 功太郎, 髙鍋 俊樹, 木内 敬太, 梅原 英裕, 入澤 航史, 中瀧 理仁, 沼田 周助, 康 鑫, 吉田 稔, 松本 和幸 : 早期うつ状態検出のためのマルチモーダル対話データセットに基づくうつ状態検出モデルの性能評価, 言語処理学会 第 31 回年次大会発表論文集, 1393-1397, 2025 年 3 月.