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

柏原 功太郎[1]　高鍋 俊樹[1]　木内 敬太[2]
梅原 英裕[3]　入澤 航史[3]　中瀧 理仁[3]　沼田 周助[3]
康 シン[4]　吉田 稔[4]　松本 和幸[4]
[1] 徳島大学大学院 創成科学研究科 [2] 労働者健康安全機構
[3] 徳島大学 医学部 [4] 徳島大学大学院 社会産業理工学研究部
{c612335050, c612435041}@tokushima-u.ac.jp
kiuchi@h.jniosh.johas.go.jp
{umehara.hidehiro, irizawa.koushi, nktk, shu-numata}@tokushima-u.ac.jp
{kang-xin, mino, matumoto}@is.tokushima-u.ac.jp

概要

ストレスや不安によるうつ病が世界的に増加しており, 日本でも深刻な問題となっている. しかし, カウンセラーや精神科医の不足により早期発見が困難である. 本研究では, 我々の研究グループで構築しているカウンセリング面談時の言語・音声・動画・心拍及びアンケートデータから成るデータセットを用いて, 最新の Mamba ベースのマルチモーダルうつ状態検出モデルの学習・評価を行った. その結果, 既存の大規模データセットを用いた事前学習と本データセットでの微調整が, モデル性能の向上に効果的であることが分かった.

1　はじめに

近年, 世界的にストレスや不安によるうつ病が増加しており, 日本でもうつ病などのメンタル不調を原因とした休職者数が増加している [1]. さらに新型コロナウイルス感染症の影響で, 老若男女問わず健康不安や経済的問題からメンタル不調者が増加し, 軽視できない社会問題となっている. しかし, カウンセラーや医師の不足が深刻で早期発見が難しい. また, メンタル不調者の多くは悩みを周囲に相談できず, 自分一人で抱え込んでしまい, 発見が遅れがちである. このことから, メンタル不調者をカウンセリングや診療に誘導することが喫緊の課題である.

そこで我々の研究グループでは, うつ状態の早期発見が可能なマルチモーダル対話システムを構築するために, 専門のカウンセラーが被験者に対し 30 分間程度の面談を行い, その面談時の言語・音声・動画・心拍及びアンケートから成るマルチモーダルデータセットの構築及び, 言語・音声・動画の各特徴量とアンケート結果との相関分析を行っている [2, 3].

本稿ではそのデータセットの構築及び分析を概説し, さらに最新の Mamba ベースのマルチモーダルうつ状態検出モデル DepMamba[4] を用いた学習・評価結果について説明する. 本研究では, 構築したデータセット単体での学習に加え, 既存の大規模データセット DAIC-WOZ[5] を用いた事前学習の後に構築したデータセットでの微調整を行い比較した. その結果, マルチモーダルうつ状態検出についても微調整がモデル性能の向上に効果的であることが分かった.

2　関連研究

我々の構築したデータセットに類似するデータセットとして, DAIC-WOZ[5] がある. DAIC-WOZ は心理的苦痛の有無に関する面談時のマルチモーダル情報と, 心理的苦痛に関する質問票の回答結果を含んだデータセットである. 189 人の面談時の言語・音声・動画データと, ラベルとしてうつ状態を測定する質問票である PHQ-8 のスコアの回答が含まれる.

本研究では, DAIC-WOZ と同様に半構造化面接の形式で面談を行い, その際のマルチモーダル情報を収集するが, 被験者の心理状態を深く理解するため

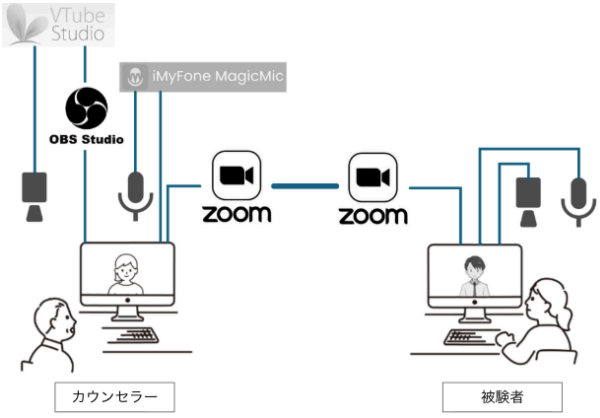

図 1　面談の全体像

に, ラベルとしてうつ状態を測定する質問票である PHQ-9 のほかに, 全般不安や社交不安症に関するスコアの回答も収集した. また, 精神科医などの専門家によるアノテーションも行う予定である. 加えて, DAIC-WOZ ではメンタル不調と診断された被験者のデータを含んでいないが, 本研究では収集を行っている.

3　データ収集

徳島大学及び徳島大学病院から被験者を募集した. 集まった被験者に対し専門のカウンセラーが 30 分程度の面談を実施し, その前後でアンケートへ回答してもらい, 面談時の言語・音声・動画・心拍とアンケートのデータを収集した. データ収集に際して, 徳島大学の研究倫理審査を受け承認されている.

面談は Zoom を通じて行われ, 音声及び動画データは Zoom の録画機能を用いて収集した. 言語データは音声データを reazonspeech-nemo-v2[1] に入力し, 誤りを人手で修正し収集した. 心拍データは面談時に被験者の左胸に心拍計を装着し収集した. データ収集の際の面談の全体像を図 1 に示す.

3.1　アバターと音声変換

データ収集の際には複数のカウンセラーが参加したが, それらカウンセラー間の被験者が感じる印象の差異をなくすため, 面談時にはカウンセラーの見た目はリアルタイムでアバターに変換された. 面談中, アバターはカウンセラーの顔表情及び顔の向きをトラッキングした. アバターには一般に販売されている商用利用可能なアイコン風サラリーマン[2] と

図 2　使用したアバター

図 3　アバターとの面談の様子

いう 2D アバターを使用した. 使用したアバターを図 2 に, アバターとの面談の様子を図 3 に示す. アバターのトラッキングには VTube Studio[3] を使用した. また面談時, 被験者にはアバターの名前をハルとして明示し, それ以外の性別や年齢などについては明示しなかった.

また同様の理由から, 面談時にはカウンセラーの音声もリアルタイムで同一の音声へと変換された. 音声の変換には iMyFone MagicMic[4] を用いた.

3.2　アンケート

面談の前後で被験者に対してアンケートを実施した. 面談前には, 被験者の現在の心理状態やパーソナリティを理解するため, うつ病診断の際に用いられる質問票の 1 つである PHQ-9 やパーソナリティ判定のために用いられる BIG5 などについて 11 セクション全 173 問のアンケートを行った. 面談後には, 面談時の気分とアバターへの印象について 2 セクション 34 問のアンケートを行った.

1）https://huggingface.co/reazon-research/reazonspeech-nemo-v2
2）https://nizima.com/Item/DetailItem/87378
3）https://store.steampowered.com/app/1325860/VTube_Studio/
4）https://jp.imyfone.com/voice-changer/

3.3 収集結果

2025 年 1 月現在では, 被験者 101 名のデータを収集することができた. そのうち女性は 46 名, 男性は 55 名となった. また, 被験者のうち 18 歳以下の未成年は 20 名, 成人は 81 名, 平均年齢は 24.65 歳となった. 加えて, PHQ-9 のアンケート回答を数値化し, 10 点未満を非うつ状態, 10 点以上をうつ状態としたとき, 101 名の被験者のうち非うつ状態は 79 名, うつ状態は 22 名となった.

4 データ分析

収集したデータから特徴量を抽出し, アンケートのうつ状態や不安に関する項目との相関分析を行った. 以下では実施した特徴量抽出方法と相関分析の結果を示す.

4.1 言語データ

GiNZA[6] と日本語評価極性辞書 [7, 8] を用いて, 被験者の面談時の言語データ中のポジティブ・ネガティブな名詞の数, 言語データ中の全名詞に対するポジティブ・ネガティブな名詞の割合, 言語データ中のポジティブ・ネガティブな用言の数, 言語データ中の全用言に対するポジティブ・ネガティブな用言の割合を計算した.

相関分析の結果, 被験者の PHQ-9(うつ状態の指標) 及び GAD7(全般不安の指標) のスコアと, ネガティブな名詞・用言の数及び割合との間に強い正の相関が見られた. 未成年の被験者についてはスペンス児童不安尺度 (SCAS) とポジティブな名詞・用言の数及び割合との間に強い正の相関が見られた. これらのことはうつ状態や不安のレベルが高い人ほど, ネガティブな単語を使用しやすいことを示唆している.

4.2 音声データ

OpenSMILE[9] を用いて, 被験者の面談時の音声の特徴量を抽出し分析を行った. 特徴量セットは感情推定に有効とされる eGeMAPsv02 feature set[10] を使用し, 特徴量として音声の大きさ・ピッチ (基本周波数)・ジッターの平均値及び標準偏差を抽出した.

相関分析の結果, LSAS(社会不安の指標) と SCAS のスコアと音声の大きさの平均値との間に強い負の相関が見られた. また音声のピッチの平均値との間にも正の相関が見られた. Galili ら [11] は, 社交不安を持つ人は基本周波数数が大きく声の大きさが小さくなる傾向にあることを示したが, 今回の分析結果は Galili らの結果と同様の結果になった.

4.3 動画データ

被験者の面談時の動画データに対し, OpenFace[12] を用いてフレームごとの各アクションユニットの強さを抽出し, これらの強さの平均値・標準偏差と, 全てのアクションユニットの値の平均値と標準偏差を計算し, 面談時の被験者の顔表情の強さや変化を分析した.

相関分析の結果, GAD7 や LSAS, SCAS などの不安に関連するスコアが高いほど, 緊張と覚醒を示す「頬を持ち上げる」(AU06) や「瞼を緊張させる」(AU07) のアクションユニットの強さの平均値が大きくなった. また, PHQ-9 のスコアが高いほど「眉の内側を上げる」(AU01) や「唇両端を横に引く」(AU20) といった笑顔に関連するアクションユニットの強さの平均値が小さくなり, うつ状態にある被験者は笑顔が少ないことが示唆される.

5 実験

Mamba をベースとしたマルチモーダルうつ状態検出モデルである DepMamba[4] を利用し構築したデータセットを学習・評価した. このモデルを用い, 構築したデータセット単体での学習に加え, DAIC-WOZ を用いた事前学習の後に構築したデータセットでの微調整を行い, それらの結果を比較した.

5.1 特徴量とラベル

構築したデータセット及び DAIC-WOZ ともに以下の方法で統一して特徴量を抽出し, 学習に用いた.

5.1.1 音声特徴量

被験者の発話区間のみの音声を切り出して結合し, それを VGGish[13] に入力し最終層から得られた 128 次元の特徴量を学習に用いる. VGGish は 0.96 秒ごとに特徴量を出力するため, 最終的な特徴量次元は (ほぼ被験者の発話区間の音声の時間長, 128) となる.

5.1.2 顔表情特徴量

音声と同様に被験者の発話区間の動画フレームのみを切り出して結合し, それを OpenFace に入力して得られた各 68 個の顔ランドマークの x, y 座標を結合し 136 次元の特徴量を学習に用いる. OpenFace は

表 1　DAIC-WOZ と構築したデータセットの DepMamba での学習結果

学習方法	Accuracy(%)	Precision(%)	Recall(%)	F1(%)	平均 (%)
(1) DAIC-WOZ 単体	43.6	32.2	83.3	46.6	51.1
(2) 構築したデータセット単体	74.6	58.3	20.0	28.1	45.2
(3) (1) を構築したデータセットで微調整	**77.8(+3.2)**	**70.1(+11.8)**	**53.3(+33.3)**	**48.7(+20.6)**	**62.7(+17.5)**

フレームごとに特徴量を出力するが, 1 秒毎の特徴量となるようにダウンサンプリングした. そのため最終的な特徴量次元は (被験者の発話区間の動画の時間長, 136) となる.

5.1.3　ラベル

DAIC-WOZ では被験者の PHQ-8 のスコア, 構築したデータセットでは PHQ-9 のスコアが提供されているが, 今回はこのスコアのカットオフポイントを 10 とし, 10 点未満を非うつ状態, 10 点以上をうつ状態とする 2 値分類のためのラベルを作成した. また, このラベルを層化抽出を用い, 学習: 検証: テストデータが 6:2:2 となるよう分割した.

5.2　実験方法

抽出された特徴量を用いて DepMamba モデルを学習した. 構築したデータセットと DAIC-WOZ 両方で, うつ状態のデータの割合が少ないため損失関数に重み付けを実施し学習を行った. 微調整では, 事前学習の検証セットで最高性能を達成した事前学習済みモデルを対象に最終層のみ学習した. またランダム性をなくすためどの学習も 3 回実施し, 平均性能を比較した. その他の学習条件は以下の通りである.

- 学習率: 1e-3
- Weight Decay 率: 1e-4
- 最適化器: AdamW
- バッチサイズ: 16
- エポック数: 120

5.3　実験結果

実験結果を表 1 に示す. DAIC-WOZ で事前学習を行った後, 微調整を行うことで, すべての評価指標において性能が向上していることが分かる. 特に Recall については 33.3 % と大幅に向上しており, モデルが実際にうつ状態の被験者をより正しく検出できるようになったことを示している. この結果はマルチモーダルうつ状態検出のタスクにおいても微調整が有効であることを示唆している.

6　今後の課題

本研究ではうつ状態検出のためのマルチモーダルなデータセットを構築したが, このデータセットに対する客観的なアノテーションは行えていない. 今後は収集したデータに対して精神科医などの専門家が医学的見地から客観的なアノテーションをつけることを検討している. 加えて, 我々の以前の研究 [14] で作成したアノテーションツール[5)] を用い, ラッセルの円環モデル [15] に基づく感情アノテーションをつけることも検討している.

さらに, 今回は PHQ-9 のスコアを 10 をカットオフポイントとして 2 値分類のラベルを作成しモデル学習・評価を行ったが, 複数のカットオフポイントを設けることにより 3 値以上のラベルを作成し多値分類のモデルを構築することも検討している.

最後に, 我々の最終的な目標は, 早期うつ状態検出のためのマルチモーダル対話システムの構築であるが, この構築のためにマルチモーダル対話システムツールキット Remdis[16] を拡張する予定である. Remdis には音声入力/出力, LLM による対話応答, ターンテイキング, アバター応答などの機能が既に実装されているため, これを利用し, システム利用者の動画・音声をマルチモーダルうつ状態検出モデルに入力し, アバターとの対話時のメンタル状態を判定・記録するシステムを構築する予定である.

7　まとめ

本研究では, 早期うつ状態検出を目的としたマルチモーダル対話システムの構築に向け, 専門家による面談時のデータセットを構築し, その分析とモデル学習を行った. 収集したデータセットは, 言語・音声・動画・心拍及びアンケートデータを統合したもので, 既存の大規模データセットと併用することで, 検出モデルの性能向上が確認された. 今後は, データに客観的アノテーションを付与し, モデル性能をさらに向上させ, マルチモーダル対話システムへ組み込むことを目指す.

5)　https://github.com/A2TokushimaUniv/russell_emotion_annotation

7.1 謝辞

本研究は, 令和 5 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および, 公益財団法人 JKA 令和 6 年度開発研究補助事業により実施されました. 深く謝意を表します.

参考文献

[1] Job 総研. Job 総研による『2023 年 メンタルケアの意識調査』を実施 我慢こそ美徳？弱音吐露”歓迎”も 7 割が不調”言えない. https://prtimes.jp/main/html/rd/p/000000183.000013597.html, 2023. Accessed 21 June 2024.

[2] 柏原功太郎, 俊樹, 木内敬太, 梅原英裕, 入澤航史, 中瀧理仁, 沼田周助, 康シン, 吉田稔, 松本和幸. マルチモーダルなカウンセリングデータセットの構築と特徴量の分析. IT ヘルスケア誌 第 17 回年次学術大会抄録集, 第 19 巻, pp. 94–99, 8 2024.

[3] Kotaro Kashihara, Toshiki Takanabe, Keita Kiuchi, Hidehiro Umehara, Koushi Irizawa, Masahito Nakataki, Shusuke Numata, Xin Kang, Minoru Yoshida, and Kazuyuki Matsumoto. Constructing multimodal counseling dataset for depressive states and feature analysis. In **Proceedings of the 2024 8th International Conference on Natural Language Processing and Information Retrieval**, 12 2024.

[4] Jiaxin Ye, Junping Zhang, and Hongming Shan. Depmamba: Progressive fusion mamba for multimodal depression detection. In **ICASSP 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, April 06-11, 2025**, pp. 1–5, 2025.

[5] Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. Avec 2017: Real-life depression, and affect recognition workshop and challenge. In **Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge**, AVEC '17, p. 3–9, New York, NY, USA, 2017. Association for Computing Machinery.

[6] 松田寛. Ginza - universal dependencies による実用的日本語解析. 自然言語処理, Vol. 27, , 2020.

[7] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一. 意見抽出のための評価表現の収集. 自然言語処理, Vol. 12, pp. 203–222, 2005.

[8] 東山昌彦, 乾健太郎, 松本裕治. 述語の選択選好性に着目した名詞評価極性の獲得. 言語処理学会第 14 回年次大会論文集, pp. 584–587, 2008.

[9] Florian Eyben, Martin Wollmer, and Bjorn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. **Proceedings of the 18th ACM international conference on Multimedia**, 2010.

[10] Florian Eyben, Klaus R. Scherer, Bjorn Schuller, Johan Sundberg, Elisabeth Andre, Carlos Busso, Laurence Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. **IEEE Transactions on Affective Computing**, Vol. 7, pp. 190–202, 2016.

[11] Ofer Amir, lior Galili, and Eva Gilboa-Schechtman. Acoustic properties of dominance and request utterances in social anxiety. **Journal of Social and Clinical Psychology**, Vol. 32, pp. 651–673, 06 2013.

[12] Tadas Baltruvsaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. **2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)**, pp. 59–66, 2018.

[13] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin W. Wilson. Cnn architectures for large-scale audio classification. **2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, pp. 131–135, 2016.

[14] Toshiki Takanabe, Kotaro Kashihara, Kazuyuki Matsumoto, Keita Kiuchi, Xin Kang, Ryota Nishimura, and Manabu Sasayama. Multimodal emotion recognition and dataset construction in online counseling. In **The 38th Pacific Asia Conference on Language, Information and Computation**, 2024.

[15] James Russell. A circumplex model of affect. **Journal of Personality and Social Psychology**, Vol. 39, pp. 1161–1178, 12 1980.

[16] 千葉祐弥, 光田航, 李晃伸, 東中竜一郎. Remdis: リアルタイムマルチモーダル対話システム構築ツールキット. 人工知能学会 言語・音声理解と対話処理研究会（第 99 回）, pp. 25–30, 2023.