仮想エージェントとの認知行動療法中の心理的ストレスの推定に向けたマルチモーダル特徴量の影響の調査

設樂　一碩[†]　田中　宏季[†]　足立　浩祥[††]　金山　大祐[††]　阪上由香子[††]
工藤　喬[††]　中村　哲[†]

† 奈良先端科学技術大学院大学 先端科学技術研究科　〒 630–0192 奈良県生駒市高山町 8916–5
†† 大阪大学 キャンパスライフ健康支援・相談センター　〒 565-0871 大阪府吹田市山田丘 1–1
E-mail: †{shidara.kazuhiro.sc5,hiroki-tan,s-nakamura}@is.naist.jp

あらまし　認知行動療法（CBT）は、精神疾患の治療や日常のメンタルヘルスケアとして確立された手法である。対面での自動対話が可能なインタフェースである仮想エージェントは、CBT をはじめとするヘルスケア支援への応用が期待されている。我々は、CBT 中の心理状態を推定することにより、適切な応答を選択する仮想エージェントの構築を目指している。しかし、先行研究では CBT 中に心理的ストレスを推定するための検討は行われていない。そこで、テキスト、音声、画像のマルチモーダル情報を用いて、CBT 中の心理的ストレスを正確に推定する方法を検討する。本研究では、29 人のユーザとエージェントのインタラクションを収録し、線形回帰により心理的ストレスの尺度である K6 スコアの推定を行った。結果、K6 スコアの実測値と予測値の相関係数で 0.49 を達成した。
キーワード　認知行動療法、仮想エージェント、マルチモーダル、線形回帰、心理的ストレス

Kazuhiro SHIDARA[†], Hiroki TANAKA[†], Hiroyoshi ADACHI[††], Daisuke KANAYAMA[††], Yukako SAKAGAMI[††], Takashi KUDO[††], and Satoshi NAKAMURA[†]

† Nara Institute of Science and Technology,Takayama-cho 8916-5, Ikoma-shi, Nara
†† Health and Counseling Center, Osaka University, Yamadaoka 1-1, Suita-shi, Osaka
E-mail: †{shidara.kazuhiro.sc5,hiroki-tan,s-nakamura}@is.naist.jp

1. はじめに

認知行動療法（Cognitive behavior therapy, CBT）はうつ傾向や不安障害に有効な心理療法の一つであり、日常におけるメンタルヘルスケアの方法としても広く用いられている。治療者が患者の感情を考慮して CBT を行うことは、患者との協働的関係の構築や、より適切な治療方針の判断のために重症視されている [1]。そのため、治療者は患者の感情の認識を心がけている。近年、CBT を理論基盤とした、仮想エージェントによるメンタルヘルスケアに関する研究が多く取り組まれている。これまでの研究では、メッセージアプリの対話様式 [2], [3] や、音声と画像を含むマルチモーダルな対話様式 [4] の仮想エージェントが提案されている。ヘルスケアを行う仮想エージェントは有用性・使いやすさ・信頼性・魅力などの性能向上に向けて多くの取り組みが行われ、そのアプローチの一つに機械学習を用いた心理的ストレスやうつ傾向の推定に基づく応答選択がある [5]。

仮想エージェントとユーザのインタラクション中の行動情報からの心理的ストレスやうつ傾向の重症度の検出は、DAIC-WOZ　(Distress Analysis Interview Corpus-Wizard of Oz) database [6] をはじめとする公開データを用いて、シェアードタスクとして取り組まれている。DAIC-WOZ database はテキスト（ユーザ発話の書き起こし）、音声（ユーザの発話音声）、画像（ユーザの表情を写した画像）の行動情報とうつ傾向尺度のスコアを含んでおり、[7]〜[9] をはじめとして、多くのうつ傾向推定モデルが提案されている。うつ傾向検出においては、感情とうつ傾向の関係の分析も取り組まれている。Wu ら [10] は、転移学習により感情認識モデルをうつ傾向の推定に適用し、推定性能を向上させた。著者らはさらに、セッション内で表現されたネガティブな感情表現の割合と、うつ傾向の関係を調査した。その結果、感情情報はうつ傾向の検出に役立つが、その関係は一義的ではないことが示された。別の観点では、うつ傾向とユーザ応答の定性的な性質の関係を調査したものもある。Muszynski ら [11] はうつ傾向の検出性能とインタラクション中の質問応答の性質の関係を調べた。著者らは、心理的ストレス評価のための質問、会話のようなオープンクエスチョン、肯定

的・否定的な言葉の概念の説明を促す質問の 3 つの質問リストを用い、音声と画像情報から抽出した特徴を入力としたモデルを作成した。その結果、うつ傾向の分類に重要な行動特性は、使用する質問リストによって異なることがわかった。マルチモダリティ、質問内容、感情の関係性の調査はまだ途上であり、近年もさまざまな取り組みがなされている。

一方で、これまで CBT 中の行動情報からユーザの心理的ストレスを推定した研究はない。本研究では、CBT 中での心理的ストレス推定における、各モダリティの行動情報がモデルの性能向上に及ぼす影響を調査する。さらに、推定モデルに感情の主観値も特徴量として含め、推定精度に与える影響を行動情報と比較する。

2. 方　　法

本研究では、仮想エージェントとユーザ認知行動療法のデータを収録し、収録直前に取得した心理的ストレスの傾向を推定する機械学習モデルを構築した。機械学習モデルの入力として用いる特徴量の組み合わせごとの推定精度を比較することで、各特徴量が与える影響を調査した。

2.1　心理的ストレスの評価

推定モデルの目的変数である心理的ストレスとして、Kesller Psychological Distress Scale（K6） [12] のスコアを用いた。K6 は日本をはじめとする多くの地域で医療機関、学校、企業などで広くメンタルヘルスの状態を測定するために使用されている。K6 スコアの範囲は 0〜24 で、高いほど深刻である。

2.2　仮想エージェントの設計

我々の以前の研究 [13] で構築した仮想エージェントを使用した。この仮想エージェントは、仮想エージェントツールキットである GRETA [14] を利用している。エージェントのアニメーションの外観は、Tanaka ら [15] により調査された、日本人に受け入れられやすいものを使用している。また、参加者の表情はディプレイ上部設置したカメラ、音声はヘッドマイクを使用して収録した。

2.3　認知行動療法に基づく対話シナリオ

CBT はうつ傾向や不安障害などの精神疾患の治療法であり、日常的なヘルスケアの方法としても広く利用されている。CBT は、新たな思考を発見することで患者の気分や思考を改善するためのメンタルケアの方法である。CBT においては、状況に対する思い込みや解釈のことを自動思考と呼び、自動思考が最良の考え方を客観視することで新たな考えを取り入れる。問題に適応した思考を得ることで、落ち込みの改善が期待される。

データ収集では、我々の先行研究 [13] で既に気分の改善が示されている質問シナリオの一部を使用した。シナリオは、CBT セラピストのためのガイドブック [1] に基づき、精神科医が監修したものである。モデルの入力には、図 1 に示す状況・気分・自動思考の同定の三つの質問項目に対する回答区間のみを行動情報として用いた。これらの質問は、CBT の中でもユーザの心理状態を確認するための質問である。本来の CBT ではこの後にユーザが抱えている問題を改善するために、自動思考を修正することを目的とした質問が続くが、自動思考の修正を目的とした質問以降のユーザ応答は使用しなかった。理由は、分析の要因をモダリティの比較に絞るためである。自動思考を修正することを目的とした質問の最中は、シナリオが進むにつれ心理状態が変化することが想定される。しかし本研究におけるデータ量は少数であり、モダリティの比較に加えインタラクション中の心理状態の変化まで要因に含むことは困難だと判断した。

2.4　データ収集の手続き

データ収集は、奈良先端科学技術大学院大学の研究倫理委員会の審査・承認のもとで行なった（資料番号：2019-I-24-2）。実験前に全参加者から書面での同意を得た。実験前に参加者に同意書と実験説明書を提示し、参加者全員が同意した。その後、参加者は K6 を記入した。その後、参加者は CBT を説明するリーフレット（https://www.cbtjp.net/downloads/skillup/）を読んだ。リーフレットは精神科医が作成したもので、一般の人に CBT を説明するために無料で公開されている。参加者はリーフレットを読み終えると、説明を理解したことを実験担当者に伝え、実験を開始した。実験は、以前の我々の研究 [13] で作成した仮想エージェントの対話シナリオと同じ質問で構成された。すべての実験は、インタラクションの時間を 30 分以内と想定して行われた。

本研究では、学生 29 名が仮想エージェントと認知行動療法に基づくインタラクションを行った際のユーザ行動から抽出した特徴量を使用した。データセットには、31 人の卒業生（女性 9 人、男性 22 人）のインタラクションの記録が含まれている [13] で収集されたデータと新たに収集されたデータを使用した。31 人中 20 人（女性 6 人、男性 14 人）が、以前の我々の研究 [13] で収集されたものである。全データのうち 31 人中 11 人（女性 3 人、男性 8 人）が新たに収集された。 今回新たに収集したデータ収集の流れは、 [13] と同じである。我々のこれまでの研究で、質問内容とユーザの主観値の関係を分析した結果、本シナリオが気分の改善に役立つことが示されている [13]。本研究は新たにマルチモーダル行動情報の分析を行った点が以前の研究と異なる。

推定モデルの構築の際、K6 スコアが 13 未満のユーザのデータは除いた。除いたデータ数は 2 で、29 人分のデータを使用した。K6 スコアの 13 は、重度の精神的ストレスを持つ人を分類するための閾値の 1 つである。今回の実験条件では、13 を超える参加者は少数であり、含めた場合の推定精度に大きな影響を与えるため、再現性のある結果を報告することが困難であった。そのため、本研究では推定対象を閾値以下の参加者のみとし、高い信頼性を確保するために 13 以上のスコアを持つサンプルは除外した。

2.5　マルチモーダル特徴量

テキスト、音声、画像の行動特徴値と、ユーザがインタラクション中に述べた気分の強さを入力とした。収集データから抽出した特徴量について述べる。

2.5.1　テキスト特徴量

テキスト特徴量として TF-IDF(Term frequency-inverse document frequency) を用いた。。TF-IDF は文章のベクトルかの

項目名	システムの質問例	ユーザの応答例
状況	どのようなことが起こりましたか?	友人に送ったメールの返信がこない
気分 （気分の強さ: 0〜100%）	どのような気分ですか？	落ち込み（80%）
自動思考の同定	どのような考えが浮かびましたか？	彼は私のことが嫌いなんだ

図 1　仮想エージェントが行った質問と想定されるユーザ応答

手法の一つであり、一つの文書における単語の使用頻度と全ての文書における単語の使用頻度を用いて、各文書を代表する単語を算出する。ここでは、1 つの対話セッションを 1 文書として TF-IDF を計算し、計算における単語の単位を uni-gram とした。

2.5.2　音声特徴量

音声特徴量の抽出には、オープンソースの音声解析ツールである OpenSmile [16] を用いた。OpenSmile の設定ファイルには eGeMaps [17] を用いた。この設定ファイルにより、抽出する特徴量に対応するセットを選択することができる。音声特徴量は、フレーム幅 60ms で、10ms ごとにスライドして計算される。計算結果の統計量は、フレームごとに計算される。1 フレームから抽出される特徴量の数は 88 次元である。全てのフレームから抽出した各特徴量の平均値を特徴量として用いた。

2.5.3　画像特徴量

オープンソースの表情解析ツールである OpenFace [18] を用いて、画像特徴を抽出した。画像特徴は Facial Action Coding System [19] に基づいた Action Units である。特徴量はフレームごとに抽出した。全てのフレームから抽出した各特徴量の平均値を特徴量として用いた。最終的な画像特徴量の次元数 18 次元である。

2.6　気分の強さ

CBT の一般的な質問構成では、否定的な気分の種類を問う際に、その気分の強さについても問う（図 1）。気分の強さは 0 から 100 の百分率で表され、高いほどその気分が強いことを示す。本研究では、マルチモーダル情報に加え、ユーザの主観値である気分の強さを特徴量として扱った。

2.7　推定モデルと評価指標

図 2 に実験の概要図を示す。本研究では、各特徴量をモデル入力前に融合する、early fusion を用いた。分類器には Lasso 回帰モデル (l1 正則化を用いた線形回帰モデル) を使用した。Lasso 回帰の実装には Scikit-learn（version 1.0.2）を使用した。モデルの評価には Nested-Leave-One-Out-Cross-Validation を実施した。正則化項の係数 α をハイパーパラメータとし、自動選択を行った。評価指標として、一致度相関係数（concordance correlation coefficient, CCC）と二乗平均平方根誤差（root mean squared error, RMSE）を用いた。CCC と RMSE の算出方法を式 1 と式 2 に示す。

$$CCC = \frac{2\rho\hat{\sigma}\sigma}{(\hat{\mu}-\mu)^2 + \hat{\sigma}^2 + \sigma^2} \tag{1}$$

ここで、$\hat{\sigma}$ と σ は K6 スコアの実測値と予測値の標準偏差の分散、$\hat{\mu}$ と μ は K6 スコアの実測値と予測値の平均値、ρ は K6 スコアの実測値と予測値の間のピアソンの相関係数を示す。

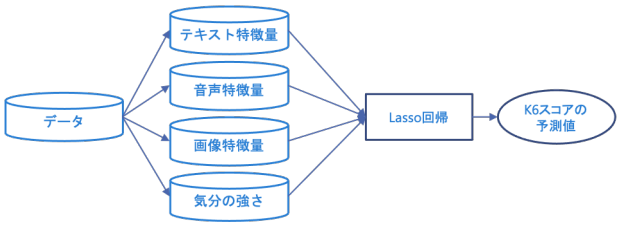

図 2　Early fusion による各モダリティの影響の比較

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y_i} - y_i)^2} \tag{2}$$

ここで、n はデータのサンプルサイズ、$\hat{y_i}$ と y_i は K6 スコアの実測値と予測値を示す。

3.　結果と考察

K6 のスコアの推定結果を表 1 に示す。テキスト、音声、画像の特徴量を含めた際に最も高い推定性能を示し、CCC において 0.49 を達成した（図 3）。さらに、気分の強さを含めた場合と含めない場合では、含めない場合の方が高い推定性能を示した。この結果から、本人の主観と比較して客観的に現れる行動特徴がより正確に K6 スコアを推定できることが示唆された。

行動特徴のみを用いた場合の方が高い推定性能を発揮するという結果は、気分の強さがその時の一時的なものであるということが一因だと考えられる。Wu ら [10] による分析は一時的な感情と長期的なうつ傾向が必ずしも一致しないことを示唆している。本研究においても同様の結果となり、標準化された尺度で計測されるうつ傾向やストレスを客観的な行動特徴に推定することの重要性が示された。

また、公開データーセットを用いたインタラクション中の心理状態推定の関連研究 [7]〜[9] と同様に、マルチモーダル情報を用いることの重要性が示された。CBT はメッセージアプリを用いたテキストの対話様式のシステムへの適用が多く提案されているが [2], [3]、本研究によりマルチモダリティを介した対話が可能な仮想エージェントの優位性が示唆された。

4.　ま　と　め

本研究の目的は、CBT 中の心理的ストレス推定の可能性と、モダリティが推定精度に与える影響を調査することであった。本研究では、解釈のしやすさと少量のデータへの適用のために、Lasso 回帰を使用した。本研究の推定実験は、テキスト、画像、音声を含むマルチモーダル情報を用いて行われた。その結果、これら 3 つ全てを使用することで 0.49 という高い精度を達成

表 1　推 定 結 果

	Feature	CCC	RMSE
気分の強さ	-	0.18	2.53
行動情報	テキスト	0.26	3.10
＋ 気分の強さ	テキスト＋音声＋画像	0.41	2.53
行動情報	テキスト	0.10	3.02
	音声	0.31	2.67
	画像	0.17	3.65
	テキスト＋音声＋画像	**0.49**	**2.22**

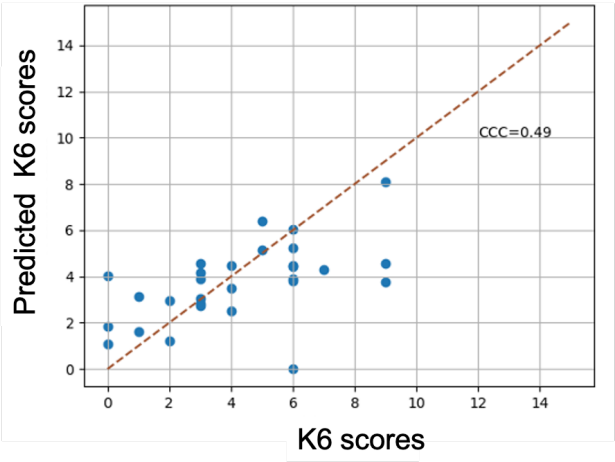

図 3　テキスト＋音声＋画像（気分の強さなし）を特徴量として用いた条件における K6 スコアの予測結果

した。

本研究の限界は、比較的心理的ストレスの大きい人の推定ができないことと、データ量が小さいことである。本データセットは、K6 スコアが 0 から 12 の間の一般の人々のデータからから構成されている。今後は、より広い範囲の被験者を募集し、モデルの検証に含める必要がある。今後はデータ量の拡張とともに、データの時系列を考慮した推定が可能な深層学習技術の適用による精度向上も検討する。

5. 謝　　辞

本研究は CREST（グラント番号: JPMJCR19A5）の支援によって行われた。

文　　献

[1] J.S. Beck, Cognitive behavior therapy: Basics and beyond, Guilford Publications, 2020.

[2] K.K. Fitzpatrick, A. Darcy, and M. Vierhile, “Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial,” JMIR mental health, vol.4, no.2, p.e19, 2017.

[3] B. Inkster, S. Sarda, and V. Subramanian, “An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: real-world data evaluation mixed-methods study,” JMIR mHealth and uHealth, vol.6, no.11, p.e12106, 2018.

[4] E. Kimani, T. Bickmore, H. Trinh, and P. Pedrelli, “You’ll be great: Virtual agent-based cognitive restructuring to reduce public speaking anxiety,” 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), pp.641–647, IEEE, 2019.

[5] A.A. Abd-Alrazaq, M. Alajlani, N. Ali, K. Denecke, B.M. Bewick, and M. Househ, “Perceptions and opinions of patients about mental health chatbots: scoping review,” Journal of medical Internet research, vol.23, no.1, p.e17828, 2021.

[6] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, *et al.*, “The distress analysis interview corpus of human and computer interviews,” tech. rep., UNIVERSITY OF SOUTHERN CALIFORNIA LOS ANGELES, 2014.

[7] A. Ray, S. Kumar, R. Reddy, P. Mukherjee, and R. Garg, “Multi-level attention network using text, audio and video for depression prediction,” Proceedings of the 9th international on audio/visual emotion challenge and workshop, pp.81–88, 2019.

[8] D. Xezonaki, G. Paraskevopoulos, A. Potamianos, and S. Narayanan, “Affective conditioning on hierarchical networks applied to depression detection from transcribed clinical interviews,” arXiv preprint arXiv:2006.08336, 2020.

[9] K. Mao, W. Zhang, D.B. Wang, A. Li, R. Jiao, Y. Zhu, B. Wu, T. Zheng, L. Qian, W. Lyu, *et al.*, “Prediction of depression severity based on the prosodic and semantic features with bidirectional lstm and time distributed cnn,” IEEE Transactions on Affective Computing, 2022.

[10] W. Wu, M. Wu, and K. Yu, “Climate and weather: Inspecting depression detection via emotion recognition,” ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.6262–6266, IEEE, 2022.

[11] M. Muszynski, J. Zelazny, J.M. Girard, and L.P. Morency, “Depression severity assessment for adolescents at high risk of mental disorders,” Proceedings of the 2020 International Conference on Multimodal Interaction, pp.70–78, 2020.

[12] T.A. Furukawa, R.C. Kessler, T. Slade, and G. Andrews, “The performance of the k6 and k10 screening scales for psychological distress in the australian national survey of mental health and well-being,” Psychological medicine, vol.33, no.2, pp.357–362, 2003.

[13] K. Shidara, H. Tanaka, H. Adachi, D. Kanayama, Y. Sakagami, T. Kudo, and S. Nakamura, “Automatic thoughts and facial expressions in cognitive restructuring with virtual agents,” 2022.

[14] R. Niewiadomski, E. Bevacqua, M. Mancini, and C. Pelachaud, “Greta: an interactive expressive eca system,” Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2, pp.1399–1400, Citeseer, 2009.

[15] H. Tanaka, S. Nakamura, *et al.*, “The acceptability of virtual characters as social skills trainers: Usability study,” JMIR human factors, vol.9, no.1, p.e35358, 2022.

[16] F. Eyben, F. Weninger, F. Gross, and B. Schuller, “Recent developments in opensmile, the munich open-source multimedia feature extractor,” Proceedings of the 21st ACM international conference on Multimedia, pp.835–838, 2013.

[17] F. Eyben, K.R. Scherer, B.W. Schuller, J. Sundberg, E. André, C. Busso, L.Y. Devillers, J. Epps, P. Laukka, S.S. Narayanan, *et al.*, “The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing,” IEEE transactions on affective computing, vol.7, no.2, pp.190–202, 2015.

[18] T. Baltrusaitis, A. Zadeh, Y.C. Lim, and L.P. Morency, “Openface 2.0: Facial behavior analysis toolkit,” 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018), pp.59–66, IEEE, 2018.

[19] P. Ekman and W.V. Friesen, “Facial action coding system,” Environmental Psychology & Nonverbal Behavior, 1978.