音声対話システムの対話破綻に対するユーザの反応と個人特性との関連

坪倉和哉[1]　武田拓也[2]　入部百合絵[2]　北岡教英[3]
[1] 愛知県立大学大学院 情報科学研究科　[2] 愛知県立大学 情報科学部
[3] 豊橋技術科学大学 情報・知能工学系
{im212008,is191056}@cis.aichi-pu.ac.jp
iribe@ist.aichi-pu.ac.jp　kitaoka@tut.jp

概要

近年，対話システムが身近な存在になってきたが，同じ内容の質問を繰り返す，事実と異なる発言をする，などの対話破綻が依然として生じている．そのため，対話破綻検出の研究が行われているが，ユーザの対話破綻に対する反応の個人差については考慮されていない．そのような個人差は破綻検出精度を低下させる恐れがあるため，生じる個人差を明らかにし，その要因を追求することは重要である．破綻時には怒りや困惑などの感情変化を引き起こす可能性があるため，本研究では，感情変化と関連の考えられる個人特性に着目し，対話破綻時のユーザの反応の個人差を分析した．分析の結果，開放性が低い人は，破綻時に怒りや嫌悪の表情を強く表出するなど，破綻に対する反応の個人差がどの個人特性に現れるかが明らかとなった．

1　はじめに

近年，自然言語処理や音声認識の技術の発展により，対話システムが身近なものとなった．しかし，現状の対話システムでは，人間同士の対話と同じように自然な対話ができているとは言い難い．特に，ユーザがシステムの発話に対して不適切な発話をしてしまう「対話破綻」が発生している [1]．対話破綻はユーザの対話意欲を低下させ，システムに対する信頼感を損う可能性があるため，対話破綻を検出することで，破綻を事前に回避し，修正する必要がある．これまで，テキストチャットにおける対話破綻を検出することを目的とした「対話破綻検出チャレンジ」 [1] が開催され，関連した研究が行われている．また，マルチモーダル情報を用いた対話破綻検出も研究されており，対話破綻後のユーザのマルチモーダル情報から，対話破綻か否かを識別している [2]．しかし，従来のマルチモーダル情報を用いた対話破綻検出では，ユーザの個人差が考慮されてこなかった．先行研究において，対話破綻に対する非言語特徴には個人差が存在すると報告されている [3, 4]．マルチモーダル情報の表出が個人によって異なれば破綻検出に大きな影響を及ぼす．

システムが破綻すると，ユーザは怒りや困惑の感情を示す可能性がある．そこで本研究では，破綻に対する感情変化と関連の考えられる個人特性に着目し，対話破綻に対するユーザの反応の個人差を分析する．これまで我々は，非言語情報の個人差を性格特性に着目して分析を行ってきた [5]．本研究では，非言語情報に加えて言語情報の個人差も分析した．また，個人差の要因として，性格特性以外にも感情変化と関連の考えられる，性別と社会的スキルも含めた．

以降，2 章では収集したデータについて述べる．次に，3 章で収集したデータから，対話破綻後のユーザのマルチモーダル特徴量の抽出を行う．4 章では，マルチモーダル特徴量と個人特性との関連を分析し，5 章で本稿をまとめる．

2　収集データ

本章では，収集した対話データについて述べる．対話実験の参加者は，愛知県立大学の学生 33 名（男性 19 名，女性 14 名）である．

対話実験では実験参加者 1 名につき 3 セッション，合計 99 セッションの対話を収録した．1 セッションにつき，10 発話以上発話することを実験参加者に指示した．対話は雑談対話である．対話はシステム発話から開始され，実験参加者が対話を終了したいと思ったタイミングで「さようなら」と発話す

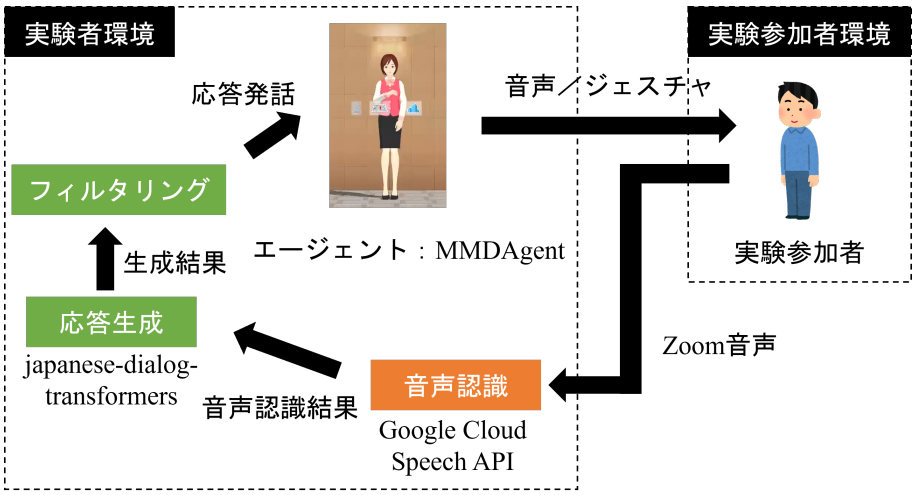

図 1　構築した対話システム

ることで対話を終了した．なお，実験は，愛知県立大学の研究倫理審査委員会の承認を得て実施している．

2.1　実験に用いた対話システム

対話実験のために構築したシステムの概要を図 1 に示す．本実験では，オンライン会議システム Zoom を用いて遠隔で音声対話を行うため，実験参加者の Zoom の音声を入力とする．実験参加者の Zoom の音声を仮想ミキサーにより抽出し，Google Cloud Speech-to-Text API により音声認識を行う．音声認識結果は対話応答生成モデルの入力となる．応答生成モデルには，日本語 Transformer Encoder-decoder 対話モデル「japanese-dialog-transformers」 [6] を用いた．応答生成モデルにより応答候補を生成した後，過去と同じ応答の繰り返しを避けるため，フィルタリングを行った．具体的には，応答候補に含まれる単語の集合とそれ以前にシステムが行った発話のひとつひとつに含まれる単語の集合の Jaccard 係数を計算し，それらの最大値の類似スコアが 0.20 以上となる候補を除外した．閾値 0.20 は先行研究を参考にした [7]．以上のように生成された応答発話テキストは対話エージェント（MMDAgent [8]）に送られ，音声合成が出力される．Zoom では対話エージェントの動きと音声を画面共有により実験参加者に配信している．

2.2　対話の収録

対話実験では，Zoom の録画機能により，実験参加者と対話システムとの対話を録画した．また，実験参加者の発話の音声認識結果とシステムの発話，それらの発話時刻を対話ログとして記録した．さらに，対話終了後に実験参加者本人がシステムの各発話に対して対話破綻か否かのラベルを付与した．アノテーションの基準は，システム発話に対して違和感や不適切さを感じた場合は「破綻」，それ以外は「非破綻」と判断してもらった．

実験では，1,085 発話のシステム発話とユーザ発話のペアを収集した．セッション毎の平均発話数は 11.0 発話であった．全体の破綻割合は，24.3%（264 発話）であったが，個人毎の平均破綻割合は最小で 0.0%，最大で 63.0%であった．従って，実験参加者毎の破綻割合に大きなばらつきがあることがわかる[1)]．どのシステム発話に対しても破綻と感じない実験参加者や半数以上の発話に対して破綻と感じる実験参加者がおり，実験参加者によって破綻の起こりやすさや破綻への感度が異なることが示唆される．

2.3　個人特性に関するアンケート

3 セッションの対話が終了した後，実験参加者の個人特性に関するアンケートを行った．具体的には，BigFive 性格特性と社会的スキルに関するアンケートを行った．

2.3.1　BigFive 性格特性

BigFive[2)] [9] の 5 つの特性を 10 項目で測定する日本語版 Ten Item Personality Inventory（TIPI-J） [10] の

1）　より詳しい分析を付録 A，B に掲載した．
2）　パーソナリティの全体的構造を外向性 (Extroversion)，協調性 (Agreeableness)，勤勉性 (Conscienciousness)，神経症傾向 (Neuroticism)，開放性 (Openness) の 5 つの次元で捉えるモデル．

質問文を利用した．このアンケートにより，BigFive の各項目に対して 2 から 14 点の得点が得られる．

2.3.2 社会的スキル

実験参加者の社会的スキルを測定するために，KiSS-18（Kikuchi's Scale of Social Skill: 18 items）[11] を用いてアンケートを行った．このアンケートでは 18 項目のアンケートにより，社会的スキルを測ることができる。

3 マルチモーダル特徴量の抽出

本章では，対話破綻に対する反応を分析するため，収集した対話データから，対話破綻後の実験参加者のマルチモーダル特徴量の抽出を行う．マルチモーダル特徴量として，言語特徴量，音響特徴量，画像特徴量を抽出する．

3.1 言語特徴量

破綻したシステム発話の直後のユーザ発話から言語特徴量を抽出する．まず，日本語形態素解析器 MeCab[3] により，ユーザ発話を形態素に分解し，品詞毎の割合を求める．品詞には，感動詞と接続詞を選択した．これらの特徴量を選択した理由は，対話破綻時にユーザが話題を切り替えようとしたときに接続詞が増加したり，破綻に対して感情を示す感動詞やフィラーが増加したりすると考えられるためである．

3.2 音響特徴量

対話システム発話終了後から次のシステム発話開始までの音声区間に対して音響特徴量を抽出した．音響特徴量の抽出には OpenSMILE [12] を用いた．本研究では INTERSPEECH2009 Emotion Challenge feature set [13] で使用された音響特徴量セットから，音声の大きさに関する RMSenergy の平均，声の高さに関連する F0 の平均を用いる．これらの特徴量により，破綻に対してユーザが驚いたり怒ったりしたときの韻律の変化を捉える．

3.3 画像特徴量

画像特徴量は，Zoom の機能により録画した動画から抽出を行う．抽出区間は，システム発話開始から次のシステム発話開始までの区間である．画像特徴量の抽出には OpenFace [14] を用い，フレーム毎の ActionUnit 特徴量と頭の動きに関する特徴量を抽出した．ActionUnit（AU）は，顔の表情を記述するための動作単位で，P. Ekman & W.V. Friesen により提案された Facial Action Coding System（FACS）で採用されている特徴量である [15]．本研究では，先行研究 [16] で対話破綻検出に有効とされる AU2（眉の外側を上げる），AU4（眉を下げる），AU6（頬を上げる），AU12（唇両端を上げる）を分析に用い，これらの AU の強度に対して，全フレームの平均値を求めた．AU2 は驚き，AU4 は怒りや嫌悪，AU6 と AU12 は喜びの表出に関わる．また，頭の動きに関する特徴量には，頭のピッチとヨーの全フレームの標準偏差偏差を求めた．これは，頭の動きを用いて対話破綻検出を行っている先行研究 [4] を参考にした．

4 特徴量と個人特性との関連の分析

本章では対話破綻時の特徴量と個人特性との関連を分析する．本研究では，対話における反応の個人差が現れると考えられる，性別，性格特性，社会的スキルの 3 つを個人特性として扱う．

まず，非破綻時のデータを含む全データに対して，個人毎に標準化を行った．これは，非破綻時（すなわち，違和感のない通常の対話時）にも特徴量には個人差が含まれると考えられるが，この影響を取り除き，破綻時における個人差を分析するためである．その後，対話破綻時のマルチモーダル特徴量について各個人特性の得点上位群と下位群（性別は男性群と女性群）に分け，マン=ホイットニーの U 検定を行った．

表 1 に，検定の結果を示す．また，検定の結果，有意差または有意な傾向が確認された特徴量について，各特徴量の上位群と下位群（性別は男性群と女性群）の平均値を表 2 に示す．社会的スキルについては，どの特徴量にも有意差が確認できなかったため，表からは除外している．

言語特徴量では，接続詞において，開放性を除く 4 つ BigFive 性格特性の項目と性別との間に有意差が認められた．接続詞は言い換えたり，話題を切り替えたりする役割を持つ．すなわち，システムが破綻したときに，実験参加者はシステムが前の実験参加者の発話を理解していないと考え，接続詞を用いることで，言い換えたり別の話題に切り替えたのではないかと推察される．よって，この結果は，実験参加者の性格によって，システムの破綻後に実験参加者が取る対話戦略が異なることを示唆するもので

3） https://taku910.github.io/mecab/

表 1　対話破綻に対する反応の個人差（表の数値は p 値である．なお，0.00004 以下は 0.000 と表示している．）

特徴量		性別		外向性		協調性		勤勉性		神経症傾向		開放性	
言語	接続詞の割合	**.0000**	***	**.0000**	***	**.0000**	***	**.0000**	***	**.0009**	***	.7937	n.s.
	感動詞の割合	**.0037**	**	.2558	n.s.	**.0592**	+	.1026	n.s.	.1801	n.s.	**.0068**	**
音響	RMSenergy の平均	.3096	n.s.	.8515	n.s.	.6207	n.s.	.6933	n.s.	.5561	n.s.	.9708	n.s.
	F0 の平均	.4961	n.s.	**.0998**	+	.2523	n.s.	.9096	n.s.	.6885	n.s.	.2481	n.s.
AU	AU2 の平均値	.8557	n.s.	.3047	n.s.	.8637	n.s.	.3837	n.s.	.8444	n.s.	.6160	n.s.
	AU4 の平均値	.5728	n.s.	.4687	n.s.	.4447	n.s.	.6575	n.s.	.1258	n.s.	**.0945**	+
	AU6 の平均値	**.0405**	*	.7616	n.s.	.5762	n.s.	.8864	n.s.	.9461	n.s.	.1744	n.s.
	AU12 の平均値	.1164	n.s.	.7506	n.s.	.8020	n.s.	.5561	n.s.	.7098	n.s.	.1585	n.s.
頭	頭のピッチの標準偏差	.9004	n.s.	.7481	n.s.	.6302	n.s.	.1320	n.s.	.9461	n.s.	.6149	n.s.
	頭のヨーの標準偏差	.4359	n.s.	.5910	n.s.	.5901	n.s.	.9447	n.s.	.7338	n.s.	.1032	n.s.

+: $p < 0.1$, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, n.s.: 有意差なし

表 2　有意差が確認された個人特性の上位群と下位群の平均値（個人特性が性別の場合は，女性群と男性群）

特徴量	個人特性	上位/女性 平均	下位/男性 平均
接続詞	性別	-0.051	0.080
	外向性	0.051	0.013
	協調性	0.021	0.050
	勤勉性	-0.034	0.079
	神経症傾向	0.066	-0.034
感動詞	性別	0.028	0.125
	協調性	0.065	0.134
	開放性	0.109	0.074
F0	外向性	-0.102	0.075
AU4	開放性	-0.060	0.125
AU6	性別	0.356	0.116

ある．また，感動詞では，性別，協調性，開放性について有意差または有意傾向が確認された．性別では男性が，協調性では下位群が，開放性では上位群が破綻時に，より感動詞を用いる傾向にある．女性や協調性が高い人は，破綻してしまったシステムに対して気を使って感情的な言葉を使わない傾向にあると考えられる．また，開放性が高い人は好奇心が高い可能性があるが，システムの破綻に対して感情的な語彙を用いて応答すると推察される．

音響特徴量では，声の高さを表す F0 と外向性の間に有意傾向が確認された．外向性上位群の方が F0 の値が低くなっており，外向性が低い人は，システムが破綻したときに声が上ずることが示唆される．

ActionUnit 特徴量では，怒りや嫌悪を表す AU4 と開放性との間に有意傾向が確認された．開放性上位群の方が AU4 の表出が小さいことから，好奇心のある人は破綻に対して不快感を抱きにくいと考えられる．また，喜びの表出にかかわる AU6 と性別との間に有意差が確認された．システムの破綻時に，女性は男性より喜びの表出を強める傾向にある．

頭の動きに関する特徴量については，どの個人特性についても有意差が認められなかった．高齢者を対象とした破綻検出において，頭の動きの周波数に個人差があることが報告されている [4]．本研究では大学生を対象としているが，本研究で対象とした個人特性以外に実験参加者の年齢が破綻に対する反応の違いを引き起こす可能性がある．

5　おわりに

本研究では，対話破綻に対する実験参加者の反応を個人特性の観点から分析を行った．まず，対話データの収集を行い，収集したデータからマルチモーダル特徴量を抽出した．マルチモーダル特徴量と個人特性（性別，性格特性，社会的スキル）との関連を分析した結果，破綻に対する反応の個人差がどの個人特性に現れるかが明らかとなった．今後は，個人差の要因として，年齢や文化による違いについても新たにデータを収集し分析を行う．また，個人特性を考慮することで対話破綻検出精度の向上を目指す．

参考文献

[1] Ryuichiro Higashinaka, Luis F. D' Haro, Bayan Abu Shawar, Rafael E. Banchs, Kotaro Funakoshi, Michi-

masa Inaba, Yuiko Tsunomori, Tetsuro Takahashi, and João Sedoc. Overview of the Dialogue Breakdown Detection Challenge 4. In Erik Marchi, Sabato Marco Siniscalchi, Sandro Cumani, Valerio Mario Salerno, and Haizhou Li, editors, *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, Lecture Notes in Electrical Engineering, pp. 403–417. Springer, Singapore, 2021.

[2] Tsubokura Kazuya, Iribe Yurie, and Kitaoka Norihide. Dialog breakdown detection using multimodal features for non-task-oriented dialog systems,. In *Proceedings of the IEEE GCCE 2022*, October 2022.

[3] 阿部元樹, 栂井良太, 綱川隆司, 西田昌史, 西村雅史. 個人差と対話行為を考慮した対話破綻検出に関する検討. 第 16 回情報科学技術フォーラム, E-002, 2017.

[4] Ayaka Kawamoto, Kazuyoshi Wada, Tomohiko Kitamura, and Kaoto Kuroki. Preliminary Study on Detection of Behavioral Features at Conversation Breakdown in Human-Robot Interactions. In *2019 IEEE/SICE International Symposium on System Integration (SII)*, pp. 375–378, Paris, France, January 2019. IEEE.

[5] 坪倉和哉, 入部百合絵, 北岡教英. マルチモーダル対話システムにおける対話破綻時のユーザの個人差. 日本音響学会第 148 回 (2022 年秋季) 研究発表会, 3-Q-13, 2022.

[6] Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. Empirical analysis of training strategies of transformer-based japanese chit-chat systems. *arXiv:2109.05217*, 2021.

[7] 藤原吏生, 岸波洋介, 今野颯人, 佐藤志貴, 佐藤汰亮, 宮脇峻平, 加藤拓真, 鈴木潤, 乾健太郎. Ilys aoba bot: 大規模ニューラル応答生成モデルとルールベースを統合した雑談対話システム. 人工知能学会研究会資料 言語・音声理解と対話処理研究会, Vol. 90, p. 25, 2020.

[8] Akinobu Lee, Keiichiro Oura, and Keiichi Tokuda. Mmdagent—a fully open-source toolkit for voice interaction systems. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8382–8385. IEEE, 2013.

[9] Lewis R Goldberg. An alternative" description of personality": the big-five factor structure. *Journal of personality and social psychology*, Vol. 59, No. 6, p. 1216, 1990.

[10] 小塩真司, 阿部晋吾, Pino Cutrone. 日本語版 ten item personality inventory（tipi-j）作成の試み. パーソナリティ研究, Vol. 21, No. 1, pp. 40–52, 2012.

[11] 菊池章夫. 思いやりを科学する. 向社会的行動の心理とスキル, 1988.

[12] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, p. 1459–1462, New York, NY, USA, 2010. Association for Computing Machinery.

[13] Björn Schuller, Stefan Steidl, and Anton Batliner. The INTERSPEECH 2009 emotion challenge. In *Proc. Interspeech 2009*, pp. 312–315, 2009.

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

[15] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

[16] Dimosthenis Kontogiorgos, Minh Tran, Joakim Gustafson, and Mohammad Soleymani. A Systematic Cross-Corpus Analysis of Human Reactions to Robot Conversational Failures. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pp. 112–120, Montréal QC Canada, October 2021. ACM.

A　個人毎の破綻割合

実験参加者毎の破綻割合を図 2 に示す．図より，実験参加者毎の破綻割合に大きなばらつきがあることがわかる．また，破綻割合は最小で 0.0%，最大で 63.0%であった．どのシステム発話に対しても破綻と感じない実験参加者や半数以上の発話に対して破綻と感じる実験参加者がおり，実験参加者によって破綻の起こりやすさや破綻への感度が異なることが示唆される．

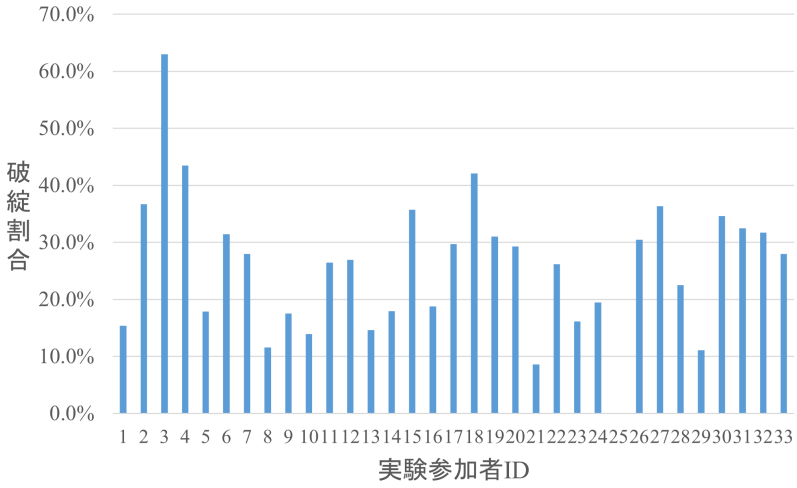

図 2　実験参加者毎の破綻割合

B　破綻割合と個人特性との関連

個人特性毎に得点の上位群（性別の場合は女性），下位群（性別の場合は男性）の 2 群に実験参加者を分割し，群毎の破綻割合を求めた（図 3）．図 3 の縦軸は破綻割合を表している．各個人特性の左側（赤色）の箱ひげ図は得点上位群（性別の場合は女性），右側（青色）の箱ひげ図は得点下位群（性別の場合は男性）である．各個人特性に対して 2 群間でマン=ホイットニーの U 検定を行った結果，性別に対して有意な傾向がみられた（$p = 0.089$）．また，神経症傾向に対して有意差が確認された（$p = 0.012$）．性別では，男性の方が破綻割合が高い傾向にあることがわかった．神経症傾向では，神経症傾向が高いグループの方が統計的に有意に破綻しやすいことがわかった．

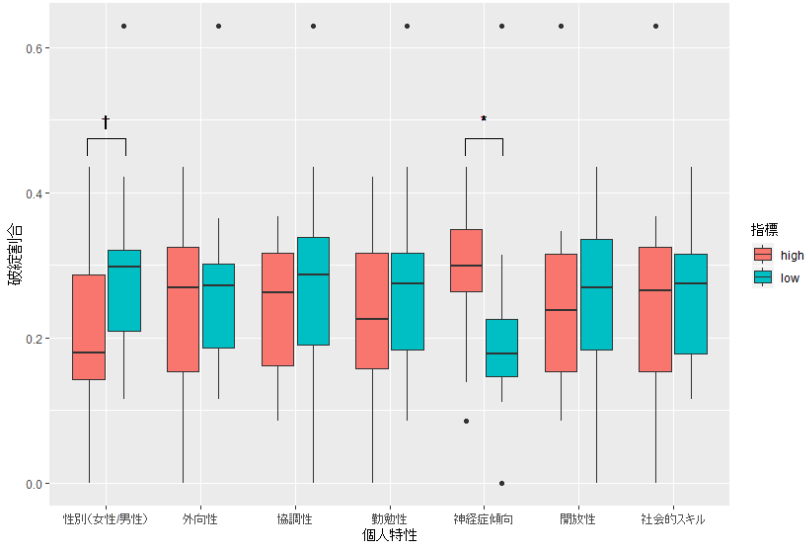

図 3　破綻割合と個人特性の関係（†：$p < 0.1$, ∗：$p < 0.05$）