実インタラクション映像から構築したマルチモーダルモデルを用いた人とロボットのインタラクションにおける異常検出

望月翔太 [1]　山下紗苗 [1]　星牟禮健也 [2]　馬場惇 [2]　窪田智徳 [3]　小川浩平 [3]　東中竜一郎 [1]
[1] 名古屋大学大学院情報学研究科　[2] 株式会社サイバーエージェント
[3] 名古屋大学大学院工学研究科
{mochizuki.shota.k8,yamashita.sanae.w7}@s.mail.nagoya-u.ac.jp
{hoshimure_kenya,baba_jun}@cyberagent.co.jp
{kubota,k-ogawa}@nuee.nagoya-u.ac.jp　higashinaka@i.nagoya-u.ac.jp

概要

本研究では，人とロボットのインタラクションにおいて生じる異常を自動で検出するモデルを構築することを目的とし，データセットの作成および異常検出モデルの構築を行った．具体的には，人とロボットのインタラクションに問題が生じた際に人間が介入する複数人同時対話の枠組みにおいて収集されたインタラクション映像に対して，正常か異常かを人手でアノテーションすることでデータセットを作成した．そして，作成したデータセットを用いて分類モデルを学習することで異常検出モデルを構築した．さらに，複数人同時対話の枠組みを検証する実証実験を実施し，モデルの検出結果をアラートとして提示することがオペレータの介入に有用であることを確認した．

1　はじめに

対話システムの進展に伴い，マルチモーダルなインタラクションが可能なロボットが，雑談や実店舗での商品販促など，様々に活用されている [1, 2]．しかし，対話ロボットのインタラクション性能は未だ完全ではなく，しばしば，対話破綻 [3, 4] を起こす．

そうした中で，人間が対話システムと協力することで効率的な対話サービスを目指す取り組みとして，複数人同時対話の枠組みが提案されている [5, 6, 7, 8]．この枠組みは，人間のオペレータが複数の対話システムの対話を監視し，問題が生じた際にのみ対話に介入することで，複数人に対話サービスを提供するものである．複数人同時対話の枠組みでは，オペレータは対話の様子を注意深く監視し，対話に問題が生じていないか判断する必要がある

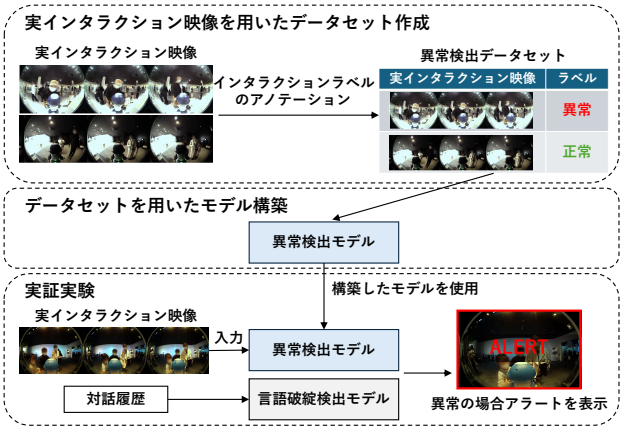

図 1　本研究のアプローチ．

が，これには高い認知的負荷がかかる．そのため，対話中に生じる種々の問題を自動で検出しオペレータにアラートとして提示することで，オペレータの介入判断を補助する手法が強く望まれている．

対話破綻を自動で検出する取り組みとして，テキストベースの対話システムにおいて，BERT [9] などのテキスト分類モデルを用いた破綻検出の研究が進められてきた [10, 11]．さらに，複数人同時対話の枠組みに，対話破綻検出技術を導入した先行研究も存在する [6, 12]．しかし，これらの手法はいずれも言語的な破綻を検出するものであり，人とロボットとのインタラクションにおける重要な要素である，動画や音声などのマルチモーダル情報は利用されていない．なお，マルチモーダル情報を用いた破綻検出の取り組みとして，坪倉ら [13] による手法が挙げられるが，使用している情報はユーザの表情のみと限定的である．

本研究では，人とロボットとのインタラクション中に生じた異常を検出し，オペレータにアラートを提示することを目的とする．図 1 に本研究のアプ

ローチを示す．まず，実際の人とロボットとのインタラクション映像を用いてデータセットを作成する．その後，作成したデータセットを用いて分類モデルを学習することで異常検出モデルを構築し，複数人同時対話の実証実験により，アラートの提示がオペレータの介入に有用であるか検証する．

2　異常検出データセット

先行研究において我々は，複数人同時対話の枠組みを検証するフィールド実験 [7] において収集された実インタラクション映像に対して，オペレータによる介入が行われる直前の 10 秒間の映像を正例（異常あり），ユーザが発話している途中の区間から抽出した 10 秒間の映像を負例（異常なし）としてラベルを付与することで，異常検出データセットを近似的に作成した [14]．

しかし，このデータセットには，人が映っておらずインタラクションが生じていない映像や，ユーザが発話せずに手を振っているだけのような非言語的なインタラクションが生じている映像，ロボットの前に人は存在するがインタラクションは行われていない映像などが含まれていないため，実環境でロバストに動作するモデルが構築できないという課題があった．また，介入の有無のみを基準としてラベルを付与しているため，ラベルの精度にも問題があると考えられる．

そこで，本研究では，フィールド実験で収集された実インタラクション映像から，多様なインタラクションが含まれるように映像を抽出し，さらに，ラベルの質を向上するために映像に対して人手によるインタラクションラベルのアノテーションを実施することで，実環境に耐えうる異常検出データセットを作成する．

2.1　インタラクション映像の抽出

以下の 2 つの方法によりインタラクション映像を抽出した．

1) 完全にランダムに抽出　フィールド実験で収集された映像から，完全にランダムに 10 秒間の区間を決定し，抽出する．

2) 人が映っていることを条件としてランダムに抽出　ランダムに 10 秒間の区間を決定し，抽出した映像に対して人物検出を行い，映像中に人が映っていればデータセットに加える．人物検出には，YOLOv10 [15] を使用する．

表 1　「インタラクションあり」で一致した 4,517 件の映像に対するインタラクションの分類ラベルの混同行列．

	正常	異常	判定不能	合計
正常	1,049	103	624	1,676
異常	124	475	395	994
判定不能	225	211	1,411	1,847
合計	1,398	789	2,330	4,517

これらの方法により，フィールド実験において収集された約 1,300 時間分の映像から 2,114 件の映像を抽出し，先行研究で抽出した 3,886 件の映像と合わせて，人手でのアノテーションの対象となる 10 秒間の映像を 6,000 件抽出した．

2.2　アノテーションの実施

抽出された 6,000 件の映像に対して，人手でのアノテーションを実施した．アノテーション時には，まず，対象の映像を確認し，その映像内でインタラクションが行われているか否かをラベル付けする．そして，インタラクションが行われている場合，そのインタラクションが「正常」なインタラクションか，「異常」なインタラクションか，あるいは「判定不能」かのいずれかのラベルを付与し，そのラベルを付与した理由を自由記述で回答する．

我々は，クラウドソーシングを通じて 10 名のアノテータを募集した．各アノテータは，対象の映像を目視で確認した後，上述のラベルのいずれかを付与した．アノテーションの質を担保するため，作業は練習と本番の二段階に分けて実施した．まず，各アノテータは，練習として 20 件分のアノテーションを実施した．そして，そのアノテーション結果に対して，著者からフィードバックが行われ，各アノテータはフィードバックを踏まえて本番のアノテーション作業に参加した．

6,000 件の映像に対して，1 映像あたり 2 人がラベルを付与した．結果として，インタラクションが行われているか否かのラベルは，5,823 件で一致し，「インタラクションあり」で一致した映像が 4,517 件，「インタラクションなし」で一致した映像が 1,306 件だった．「インタラクションあり」で一致した 4,517 件について，インタラクションの分類ラベル（正常，異常，判定不能）の一致率は 0.65，Cohen のカッパ係数は 0.45 だった．インタラクションの分類ラベルの混同行列を表 1 に示す．

「正常」と「異常」のように相反するラベルが付与された映像や「判定不能」で一致した映像，音声

表 2　テストデータに対する評価結果．尺度ごとに最も高いスコアを太字で，2 番目に高いスコアを下線で表す．* は，クラス分類モデル，深層距離学習モデルのそれぞれにおいて，他のエンコーダを用いた場合と比較して $p < 0.05$ で有意差が認められたことを示す（Holm 補正をかけた McNemar 検定）．

モデル	エンコーダ	Accuracy	Precision	Recall	F1-score
クラス分類	動画	0.656	0.419	**0.770**	0.542
	音声	0.684	0.439	0.673	0.532
	マルチモーダル	0.701*	0.459	<u>0.701</u>	<u>0.555</u>
深層距離学習	動画	0.724	0.481	0.469	0.474
	音声	<u>0.750</u>	<u>0.529</u>	0.579	0.550
	マルチモーダル	**0.770***	**0.570**	0.623	**0.589**

不良等の問題があった映像を除外した後，正例（異常あり）と，負例（異常なし）を決定し，異常検出データセットを作成した．十分なデータサイズを確保するため，1 つ以上「異常」が付与された映像を正例，1 つ以上「正常」が付与された映像を負例とした．その結果，正例が 1,049 件，負例が 3,088 件抽出され，これを異常検出データセットとした．

3　異常検出モデル

異常検出データセットを用いて，異常検出モデルの構築を行った．

3.1　モデル

先行研究 [14] と同様，(a) softmax 関数の出力に基づいて 2 クラス分類を行うクラス分類モデルと，(b) 深層距離学習 [16] により埋め込みを学習し，推論時には訓練データの埋め込みを用いた k 近傍法によりクラスを決定する深層距離学習モデルの 2 つのモデルを構築した．本研究では，推論時の k として，$k = 10$ を用いた．また各モデルについて，Transformer ベースの動画分類モデルである VideoMAE [17] を用いた動画エンコーダ，ニューラルネットワークベースの音声分類モデルである VGGish [18] を用いた音声エンコーダ，動画エンコーダと音声エンコーダの出力を，Cross-Modal Attention [19] により結合したマルチモーダルエンコーダの 3 つのエンコーダを使用し，2 つのモデルと 3 つのエンコーダの組み合わせから，計 6 つのモデルについて学習及び評価を実施した．

本データセットでは，正例が 1,081 件，負例が 3,088 件とラベルが不均衡である．そのため，クラス分類モデルでは，損失関数の Cross-entropy Loss に対して，クラスごとに異なる重み付けを行う Cost-Sensitive Learning [20] を適用した．重みとしては，データセットにおけるクラスの出現頻度の逆数を用いた [21]．また，深層距離学習モデルでは，ラベルの不均衡を考慮したパラメータの設定が可能な深層距離学習用の損失関数である Ranked List Loss [22] を用いた．

3.2　評価結果

異常検出データセットを訓練データ 70%, 検証データ 15%, テストデータ 15%にランダムに分割し，モデルの学習及び評価に用いた．

表 2 に各モデルの評価結果を示す．評価には，クラス分類モデルの標準的な評価尺度である Accuracy, Precision, Recall, F1-score の 4 つを用いた．

エンコーダの比較では，クラス分類モデル，深層距離学習モデルのいずれにおいても，動画・音声のいずれかのみを使用する場合よりも，マルチモーダルエンコーダにより動画と音声の両方を使用した場合の方が高いスコアを達成しており，Accuracy に対する McNemar 検定（Holm 法で補正）を実施したところ，マルチモーダルエンコーダと他エンコーダとの間に有意差が認められた ($p < 0.05$)．このことから，動画と音声の 2 つのモダリティを活用することが有効だと確かめられた．また，マルチモーダルエンコーダを用いた深層距離学習モデルが最も良い Accuracy, Precision, F1-score を達成したことから，深層距離学習により本データセットの分類に適した埋め込みを獲得することの有用性が示された．

4　実証実験における異常検出

構築した異常検出モデルによりインタラクション中に生じた問題を自動で検出し，オペレータに提示することの有用性を検証するために，我々は，大阪府に位置する施設「ニフレル」に自律対話ロボットの Sota[1)] を複数台配置し，複数人同時対話の枠組みを実証するフィールド実験を，2024 年 11 月 25 日から 12 月 22 日までの 28 日間実施した．施設に設置された 6 台，または 4 台の自律対話ロボットがユー

1）https://www.vstone.co.jp/products/sota

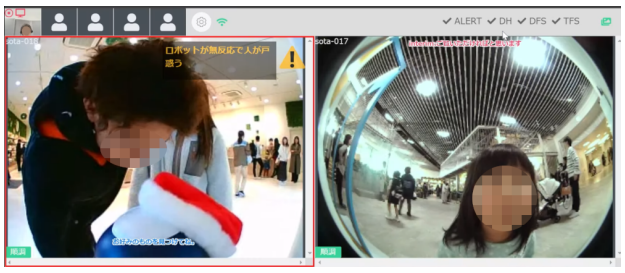

図 2　オペレータに提示されるアラートの例（左）．対象ロボットの画面が赤枠で囲われ，右上に問題発生の理由が 20 文字程度で表示される．

ザに対して施設の案内を実施する様子を，1 人のオペレータが操作用インタフェースを介して監視し，対話に問題が生じた際には介入するという設定で実施した（操作用インタフェースの詳細は [7, 23] を参照のこと）．28 日間のうち 14 日間，異常検出モデルの検出結果をオペレータに提示する条件で行い，14 日間のうち 7 日間は 6 台のロボットを，残りの 7 日間は 4 台のロボットを監視するという設定で実施した．なお，本実験の実施にあたっては，所属機関において倫理審査を経ている（承認番号: R-1-5-17）．

モデルには，テストデータに対する評価で最も性能が高かった，マルチモーダルエンコーダを用いた深層距離学習モデル（3.2 節を参照のこと）を使用した．リアルタイムに異常検出を行う流れは以下の通りである．ロボットの前面に設置されたカメラ映像を連続的に取得し，10 秒の長さで区切ってモデルに入力することで，埋め込みを獲得する．得られた埋め込みに対して，訓練データの埋め込みを用いた k 近傍法 ($k = 10$) を適用し，正例の件数が半数以上であれば異常とみなしてオペレータにアラートを提示する．図 2 に，オペレータに提示されるアラートの例を示す．異常が検出された場合，オペレータ操作用インタフェース中の，対象ロボットの画面が赤く囲われ，右上に問題発生の理由が提示される．問題発生の理由は，得られた埋め込みに最も近い正例に対してアノテーション時に付与された問題発生の理由を要約したものである．要約は，GPT-4o[2] を用いて作成される．

なお，我々が構築したマルチモーダル異常検出モデルでは，言語を入力としていないため，ロボットが誤った回答を行うなどの言語的な破綻を検出することができない．異常検出システムとして網羅性を持って動作させるためには，言語的な破綻も検出する必要があるため，GPT-4o を用いて，対話履歴のテキストから破綻を検出するモデルを別途構築し，このモデルが破綻を検出した際も同様にアラートを提示するようにした．なお，アラートの発生頻度は，マルチモーダル異常検出モデルによるものが約 35%，言語破綻検出モデルによるものが約 65%であった．

実験全体を通して，4 人のオペレータが対話の監視・介入を行った．オペレータには各介入終了後，その介入にあたってアラートを活用したかを記録させた．結果，アラートに基づいた介入は計 43 回であった．全体の介入回数（6 台: 137 回，4 台: 120 回）のうち，アラートを使用して行われた介入の割合は，ロボット 6 台の場合が 22.6%，4 台の場合が 10.0%であり，監視するロボットの台数が多いほど，頻繁にアラートが使用されていた．これは，オペレータによる対話の監視の負荷が大きいほど，アラートが効果的に活用されたことを示唆していると考えられる．

また，アラートを使用して介入した際には，アラート・問題発生の理由のそれぞれが介入にどの程度有用であったかを 5 段階のリッカート尺度で回答させた．結果，アラート・問題発生の理由の有用度の平均値は，ロボットの台数によらず約 3.9 と高かった．このことから，異常検出モデルにより問題発生を自動で検出することが，介入に有用であることが確かめられた．

5　おわりに

本研究では，人とロボットの実インタラクション映像に対して人手でアノテーションを行うことで異常検出データセットを作成し，本データセットを用いて既存の分類モデルを学習することで，異常検出モデルを構築した．その結果，マルチモーダル情報を入力とする深層距離学習モデルが最も良いスコアを達成した．さらに，複数人同時対話の実証実験により，異常検出モデルを用いてインタラクション中に生じる問題を自動で検出し，オペレータにアラートを提示することが，オペレータの介入に有用であることを確かめた．また，オペレータによる対話の監視の負荷が大きいほど，異常検出モデルによるアラートが効果的に活用されていたことが確認できた．今後は，アラートが提示された際の対話の様子を具体的に分析することで，アラートが適切なタイミングで提示されていたか，問題発生の理由が妥当なものであったかについて検証を進めたい．

2）　https://platform.openai.com/docs/models#gpt-4o

謝辞

本研究は，JST ムーンショット型研究開発事業，JPMJMS2011 の支援を受けたものです．また，フィールド実験の実施にご協力いただいたニフレルのスタッフの皆様に感謝の意を表します．本研究では名古屋大学のスーパーコンピュータ「不老」を利用しました．

参考文献

[1] Tatsuya Kawahara. Spoken dialogue system for a human-like conversational robot ERICA. In **Proc. IWSDS**, pp. 65–75, 2018.

[2] Takuya Iwamoto, Jun Baba, Kotaro Nishi, Taishi Unokuchi, Daisuke Endo, Junya Nakanishi, Yuichiro Yoshikawa, and Hiroshi Ishiguro. The effectiveness of self-recommending agents in advancing purchase behavior steps in retail marketing. In **Proc. HAI**, pp. 209–217, 2021.

[3] Bilyana Martinovsky and David Traum. The error is the clue: Breakdown in human-machine interaction. In **Proc. ISCA Workshop on Error Handling in Spoken Dialogue Systems**, pp. 11–16, 2003.

[4] Ryuichiro Higashinaka, Masahiro Araki, Hiroshi Tsukahara, and Masahiro Mizukami. Integrated taxonomy of errors in chat-oriented dialogue systems. In **Proc. SIGDIAL**, pp. 89–98, 2021.

[5] Dylan F Glas, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. Teleoperation of multiple social robots. **IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans**, Vol. 42, No. 3, pp. 530–644, 2012.

[6] Tatsuya Kawahara, Naoyuki Muramatsu, Kenta Yamamoto, Divesh Lala, and Koji Inoue. Semi-autonomous avatar enabling unconstrained parallel conversations – seamless hybrid of woz and autonomous dialogue systems–. **Advanced Robotics**, Vol. 35, No. 11, pp. 657–663, 2021.

[7] Shota Mochizuki, Sanae Yamashita, Kazuyoshi Kawasaki, Reiko Yuasa, Tomonori Kubota, Kohei Ogawa, Jun Baba, and Ryuichiro Higashinaka. Investigating the intervention in parallel conversations. In **Proc. HAI**, pp. 30–38, 2023.

[8] Tatsuya Kawahara, Hiroshi Saruwatari, Ryuichiro Higashinaka, Kazunori Komatani, and Akinobu Lee. Spoken dialogue technology for semi-autonomous cybernetic avatars. In Hiroshi Ishiguro, Fuki Ueno, and Eiki Tachibana, editors, **Cybernetic Avatar**, pp. 71–105. 2024.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proc. NAACL-HLT**, pp. 4171–4186, 2019.

[10] Ryuichiro Higashinaka, Kotaro Funakoshi, Yuka Kobayashi, and Michimasa Inaba. The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In **Proc. LREC**, pp. 3146–3150, 2016.

[11] Hiroaki Sugiyama. Dialogue breakdown detection using BERT with traditional dialogue features. In **Proc. IWSDS**, pp. 419–427, 2019.

[12] Haruki Kawai, Yusuke Muraki, Kenta Yamamoto, Divesh Lala, Koji Inoue, and Tatsuya Kawahara. Simultaneous job interview system using multiple semi-autonomous agents. In **Proc. SIGDIAL**, pp. 107–110, 2022.

[13] Kazuya Tsubokura, Yurie Iribe, and Norihide Kitaoka. Dialog breakdown detection using multimodal features for non-task-oriented dialog systems. In **Proc. GCCE**, pp. 352–356, 2022.

[14] Shota Mochizuki, Sanae Yamashita, Kazuyoshi Kawasaki, Tomonori Kubota, Kohei Ogawa, and Ryuichiro Higashinaka. Learning anomaly detection models for human-robot interaction. In **Proc. RO-MAN**, pp. 1720–1725, 2024.

[15] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. YOLOv10: Real-time end-to-end object detection. **arXiv preprint arXiv:2405.14458**, 2024.

[16] Mahmut Kaya and Hasan Şakir Bilge. Deep metric learning: A survey. **Symmetry**, Vol. 11, p. 1066, 2019.

[17] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In **Proc. NeurIPS**, Vol. 35, pp. 10078–10093, 2022.

[18] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, Ron J Weiss, and Kevin Wilson. CNN architectures for large-scale audio classification. In **Proc. ICASSP**, pp. 131–135, 2017.

[19] Ying Cheng, Ruize Wang, Zhihao Pan, Rui Feng, and Yuejie Zhang. Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning. In **Proc. ACM MM**, pp. 3884–3892, 2020.

[20] Charles Elkan. The foundations of cost-sensitive learning. In **Proc. IJCAI**, Vol. 2, pp. 973–978, 2001.

[21] Chen Change Loy Chen Huang, Yining Li and Xiaoou Tang. Learning deep representation for imbalanced classification. In **Proc. CVPR**, pp. 5375–5384, 2016.

[22] Xinshao Wang, Yang Hua, Elyor Kodirov, Guosheng Hu, Romain Garnier, and Neil M Robertson. Ranked list loss for deep metric learning. In **Proc. CVPR**, pp. 5207–5216, 2019.

[23] Sanae Yamashita, Shota Mochizuki, Kazuyoshi Kawasaki, Tomonori Kubota, Kohei Ogawa, Jun Baba, and Ryuichiro Higashinaka. Investigating the effects of dialogue summarization on intervention in human-system collaborative dialogue. In **Proc. HAI**, pp. 316–324, 2023.