マルチモーダル相互行為研究のための
『日本語日常会話コーパス』へのアノテーションの検討

森大河(千葉大学大学院生)　伝康晴(千葉大学)

1. はじめに

近年マルチモーダル相互行為研究が盛んに行われている．マルチモーダル相互行為研究では姿勢や身体の向きが活動とどのように関わっているか，特定のオブジェクトや動作が相互行為上の資源としてどのように利用されているかが分析の対象となる．例えば，阿部ほか(2018)は会話中に飲み物を飲む行動，鈴木(2023)は演劇稽古場面における演出家の姿勢の変化，千田・伝(2024)は会話中のスマートフォンの使用，居關(2024)はペットに宛てられた発話，細馬(2024)はダイナミックタッチを分析している．

マルチモーダル相互行為研究のために会話コーパスが使用されることも多い．しかし，既存のコーパスには身体動作やオブジェクトのアノテーションが付与されているものはほとんどないため，目的に適した場面を見つけるために個々の研究者が一から対象場面を探さざるを得ず，研究上の大きな負担となっている．また，大規模なコーパスになるほど探索の労力が大きくなるため，実際には一部のデータしか対象としないことも多く，チェリーピッキングの問題が生じる危険性もある．

そこで筆者らは，動画も含めて一般公開され，研究者たちに広く利用されている『日本語日常会話コーパス』(小磯ほか，2023)に画像認識技術を用いてマルチモーダルなアノテーションを付与し，検索システムとして提供する計画を進めている．本発表ではいくつかのデータに対して画像認識技術を用いて試験的にアノテーションを行った結果について発表するとともに，将来的にどのようなアノテーションが有用であるかを議論することを目的としている．

2. データ

データとして，『日本語日常会話コーパス』から，会話環境，活動，話者数，身体の動きなどの観点から6会話を選び，各会話のGoProで撮影された動画から5分程度を切り出して使用した．使用した会話は，K001_003a，K002_004，K004_011，S001_018，T011_005，T015_008aである（以下，セッションIDで呼称する）．K001は3人の女性が飲食店でお茶をしている場面で，会話の途中でスマートフォンが使用されている．K002は自営店舗の外での打ち合わせ場面で，指差しなどのジェスチャーが見られる．K004は母親と子供がキッチンで料理をしている場面で，立ったり座ったりといった姿勢の変化がある．S001はテニスクラブのオーナーが2人の友人に相談をしている場面で，屋外の会話であるため，車や自転車が通過する．T011は家族4人が自宅でリビングで食事をしている場面であり，日常会話コーパスの中で典型的な食事会話である．T015は理容室で散髪を行っている場面で，理容師の移動や運動が多く，また理容室内にテレビがある．

本発表で使用した画像認識技術は，①物体検出・追跡，②表情認識，③姿勢推定，④行動認識の4種類である．物体検出とは，画像中のオブジェクトの種類や位置を特定し，それをバウンディングボックスで示す技術であり，物体追跡はそのオブジェクトの動きをフレーム間で追跡する技術である．物体検出・追跡の目的は，「スマートフォン」や「テレビ」などのオブジェクトの有無から会話や場面を検索できるようにすることと，この後の表情認識や姿勢推定の結果と話者の紐付けを行うことである．表情認識とは，顔の表情を検出し，それを解析して感情や状態を推定する技術である．表情認識の目的は，言語情報だけでは把握できない会話の雰囲気を検索可能にすることである．姿勢推定とは，人物の特定の関節や特徴点の位置を検出し，それを用いて姿勢や構造を推定する技術である．姿勢推定の目的は，姿勢や表情，ジェスチャーを検索できるようにすることである．行動認識とは，画像や動画から人間の動作や行動を識別・分類する技術である．行動認識の目的は，「飲む」や「持つ」など行動を検索できるようにすることである．

本発表では，物体検出・追跡にはultralyticsのYOLOv11(Jocher & Qiu, 2024)，表情推定にはdeepface (Serengil et al. 2021)，姿勢推定にはMMPose(MMPose Contributors, 2020)のRTMW3D(Jiang et al., 2024)，行動認識にはMMAction2(MMAction2 Contributors, 2020)のSlowOnly(Feichtenhofer et al., 2019)を使用した．

3. アノテーション

3.1 物体検出・追跡

図1に各会話で検出されたユニークなオブジェクトの種類と数を示す．ただし，信頼度が80%以上のオブジェクトに限定している．また，追跡が失敗し同一のオブジェクトを多重にカウントしている可能性もあるため，数はあくまで参考値である．紙面の制約上，結果の一部のみを取り上げるが，K001では前述したスマートフォン「cell phone」，T015では「tv」を検出できている．また，屋外であるK002やS001では「car」や「truck」，「bicycle」などが，飲食店と食事場面であるK001とT011では「cup」や「bowl」，「spoon」，「knife」などの食器やカトラリーが検出されている．これらの結果から，物体検出・追跡技術は会話の特徴的なオブジェクトの検出に有用であることが確認された．

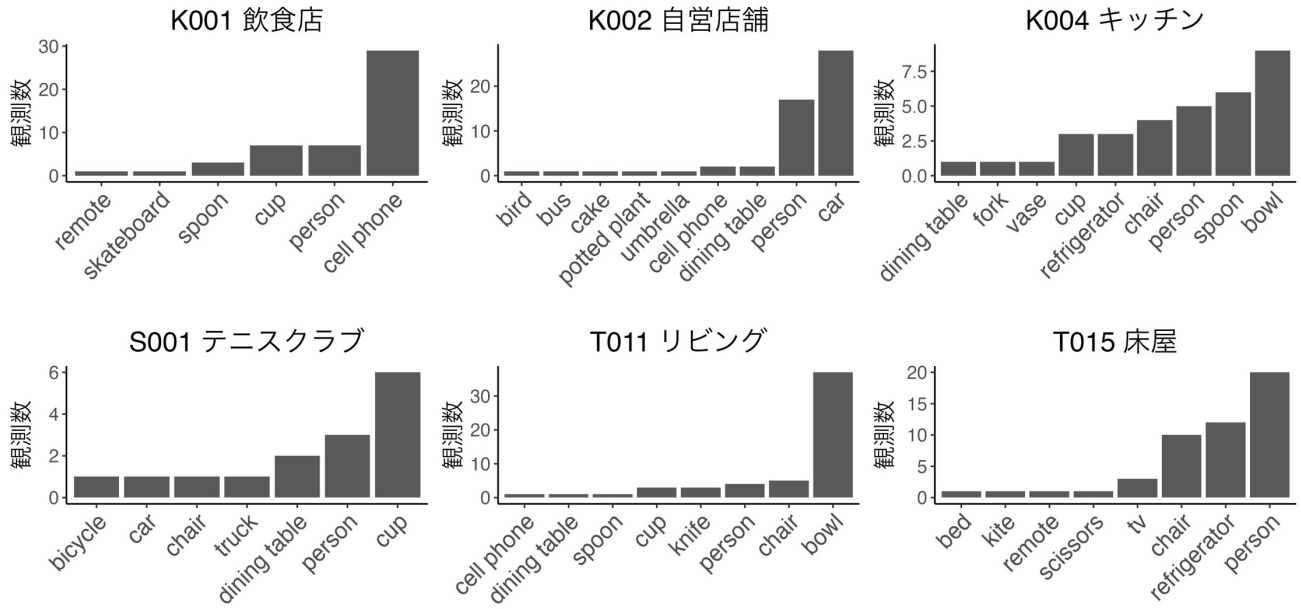

図1. 検出されたオブジェクトの種類と数

3.2 表情認識

図2に各会話で検出された表情の種類と数を示す．感情の種類は全部で6種類である．まず全体的な傾向として，「angry」「fear」「sad」のほとんどは「neutral」の誤認識であると思われる一方，「happy」「surprise」「disgust」はある程度の精度で認識できていた．実際，友人同士の会話であるK001は話者たちが笑っている場面が多い．また，話者が驚いた表情をした場面では「surprise」や「fear」が認識されていた．一方，ビジネスの場面であるK002は笑いが少なく，「fear」（実際

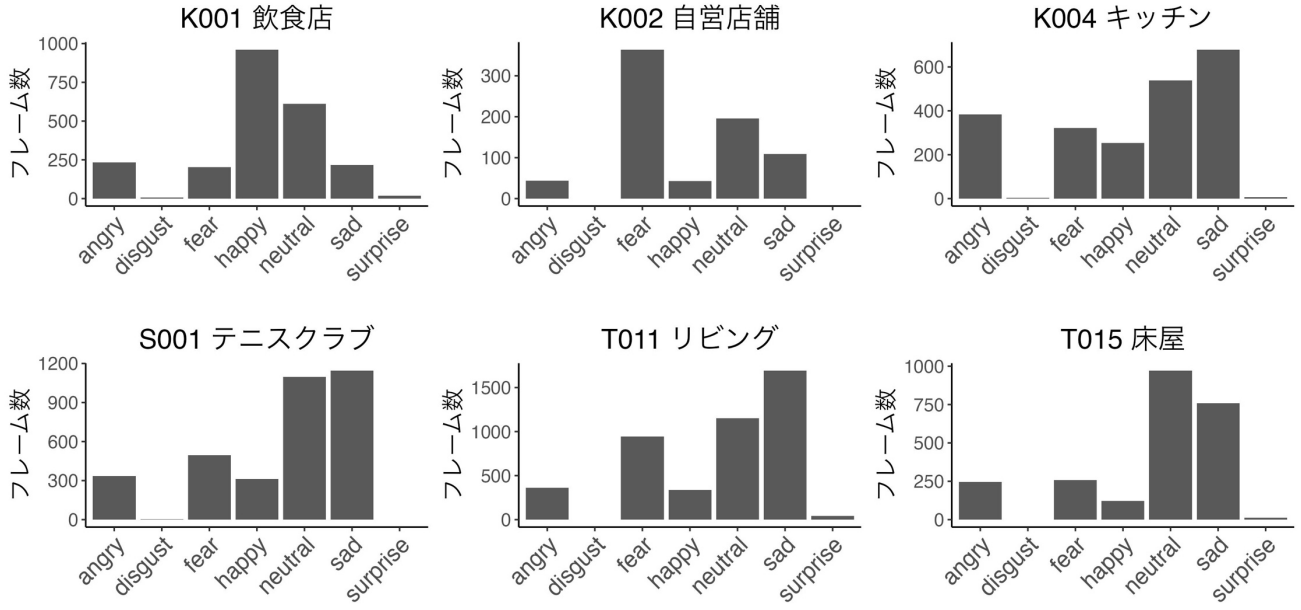

図2. 検出された表情の種類と数

には「neutral」に見える）が多い結果となった．これらの結果から，表情認識技術は会話の雰囲気を検出するのにある程度有用であることが確かめられたが，表情の種類によっては精度に問題があることがわかった．

3.3 姿勢推定

図 3 に各参与者の姿勢推定の結果の一部を示す．ただし，検出の信頼度が 50%以上のキーポイントのみ描画している．まず，前述の通り K002 では IC02 が指差しをしており，画像はその場面のものである．同時刻の推定結果の図を見ると，指差しの形が正確に推定できていることがわかる．次に，S001 は IC01 が座って前方に手差しをしている．図を見ると，伸ばした左腕や右手の握った形が推定できている．しかし，全体として画像に写っていないキーポイントの推定は精度が悪く，人間であれば常識的に推測できる情報でも現在の技術ではまだ困難であることがわかった．これらの結果から，姿勢推定技術はジェスチャーの検出などには有用であるものの，姿勢の推定には限界もあることがわかった．

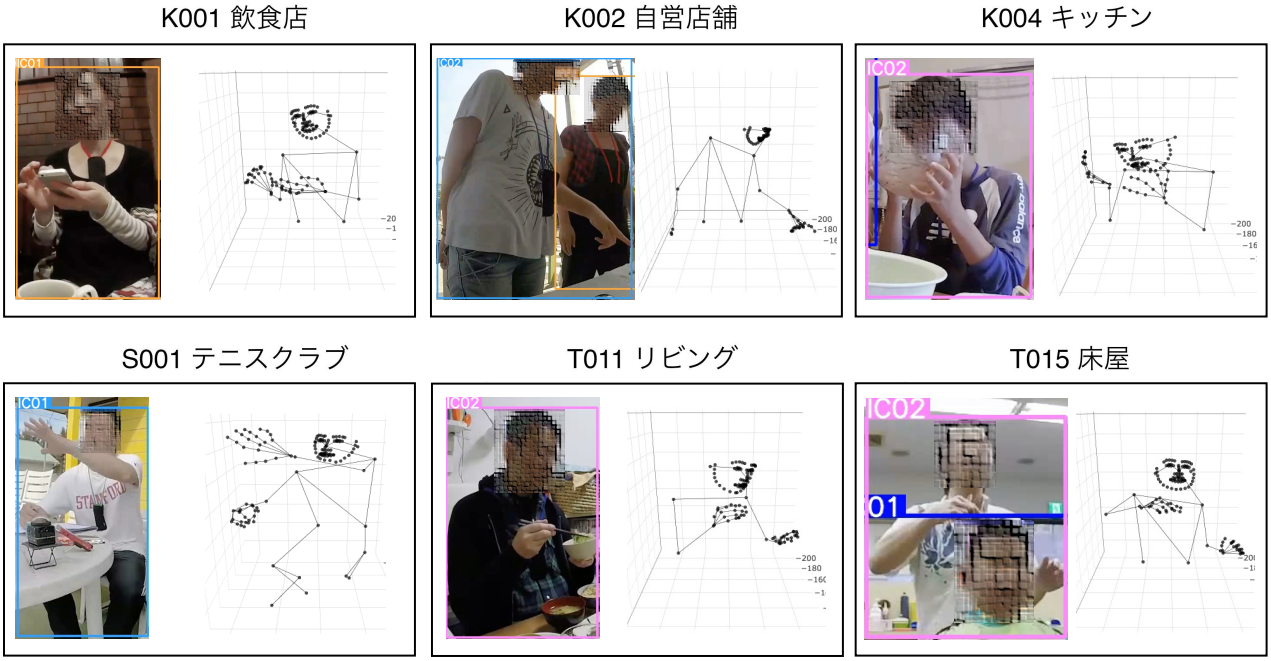

図 3. 姿勢推定の結果

3.4 行動認識

図 4 に各会話で検出された行動の種類とフレーム数を示す．ただし，信頼度が 80%以上の行動のみに限定している．図を見ると，まず，座りながらの会話である K001，S001，T011 は「sit」が多く，姿勢の変化がある K002 と K004，座っている話者と立っている話者が混在している T015 は「stand」も検出されている．また，話者の移動がある K004 と T015 では「walk」も検出できている．食事場面である T011 では「eat」，「drink」が検出されているが，想定よりも少ない．認識結果を調査したところ，距離が遠く小さく写っている参与者は精度が悪い傾向が見られた．これらの結果から，行動認識技術は会話中の行動を検出するのにある程度有用であることが確認された．

4. 応用

最後に，作成したアノテーションの有用性を評価するため，これらを活用して飲み物を飲む場面（阿部ほか，2018）をどの程度正確に検索できるかを検証した．行動認識の結果，「drink」が検出された場面は全体で 12 件あった．このうち，参与者が実際に飲み物や汁物を飲んでいた場面は 3 件であった．一方，検出されなかったものの，実際には飲み物や汁物を飲んでいた場面が 6 件確認された．これらの結果から，現在のモデルには偽陽性および偽陰性が多く含まれていることがわかる．しかし，コーパス全体を一から探索する場合と比較すれば，一定の有用性が認められると考えられる．

5. おわりに

本発表ではマルチモーダル相互行為研究の促進のために，『日本語日常会話コーパス』に物体検出・追跡，表情認識，姿勢推定，行動認識の 4 種類の画像認識技術を使って試験的にアノテーションを行った結果について発表した．今後の展望としては，まず，研究上有用なオブジェクトや行動を選定する．また，姿勢推定結果から姿勢やジェスチャーのアノテーションを行う．最後に，それらの情報を言語情報などと合わせて検索できる検索システムを開発する．

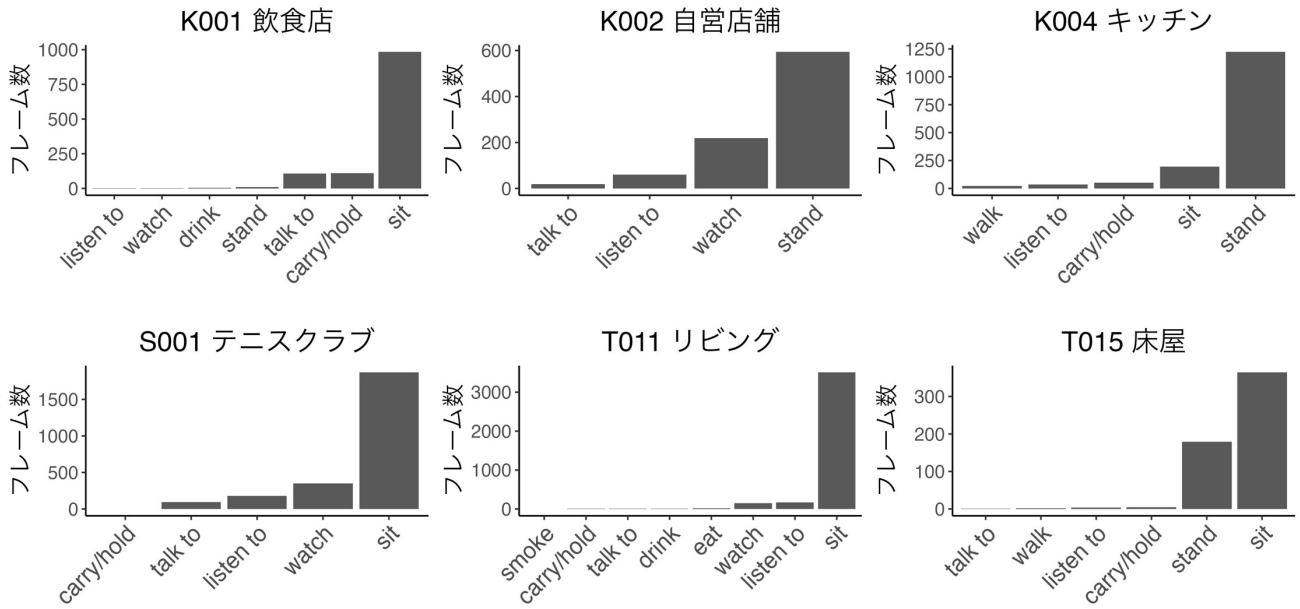

図 4. 検出された行動の種類とフレーム数

参考文献

阿部廣二・牧野遼作・門田圭祐・山本敦・古山宣洋 (2018). いつなら飲んでも良い？雑談場面における"飲むこと"の相互行為的調整 日本認知科学会第 35 回大会論文集, 685-694.

千田真緒・伝康晴 (2024). 日常会話場面におけるスマホの持ち替え 日本認知科学会第 41 回大会論文集, 146-149.

Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). Slowfast networks for video recognition. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 6202-6211).

細馬宏通 (2024). 会話におけるものの重さの表出：個人の認知から相互行為の認知へ 日本認知科学会第 41 回大会論文集, 15-18.

居關友里子 (2024). 動物のいる場における人と人とのコミュニケーション―伴侶動物に発話を宛てるやり取りに着目して― 社会言語科学会第 48 回大会発表論文集, 287-290.

Jiang, T., Xie, X., & Li, Y. (2024). RTMW: Real-time multi-person 2D and 3D whole-body pose estimation. *arXiv preprint arXiv:2407.08634*.

Jocher, G., & Qiu, J. (2024). Ultralytics YOLO11 (Version 11.0.0) [Software]. Retrieved from https://github.com/ultralytics/ultralytics.

小磯花絵・天谷晴香・居關友里子・臼田泰如・柏野和佳子・川端良子・田中弥生・伝康晴・西川賢哉・渡邊友香 (2023). 『日本語日常会話コーパス』設計と構築 国立国語研究所論集, 24, 153-168.

MMAction2 Contributors (2020). OpenMMLab's Next Generation Video Understanding Toolbox and Benchmark [Software]. Retrieved from https://github.com/open-mmlab/mmaction2.

MMPose Contributors (2020). OpenMMLab Pose Estimation Toolbox and Benchmark [Software]. Retrieved from https://github.com/open-mmlab/mmpose.

Serengil, S. I., & Ozpinar, A. (2021). HyperExtended LightFace: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)* (pp. 1-4). IEEE. https://doi.org/10.1109/ICEET53442.2021.9659697

鈴木南音 (2023). 演劇の稽古場面における演技の相互行為的構成――現代演劇の会話分析―― 年報社会学論集, 36, 52-63.