マルチモーダル分析

大塚和弘（横浜国立大学）
otsuka@ieee.org

1. マルチモーダル分析とは

「マルチモーダル」，あるいは，「マルチモダリティ」とは，辞書的定義によると「複数のモード」を指し，「モード」とは，物の存在や行動の表れ方，様態，流儀などを意味する．計算機科学の分野において，「マルチモーダル」とは，対象となる事象や物体を画像や音声，テキストと言った複数の異なるデータ形式で入力・表現し，処理を行う際の形容詞として用いられることが多い．加えて，「マルチモーダル統合」とは，一つの対象を複数の異なるモダリティから捉えることで，単一モダリティの情報に依存するよりも，より高性能なタスクの実現を狙うアプローチを指す．例えば，動画像中に写っている「鳥」の種類を識別するタスクを考えた場合，画像に写っている鳥の外見的特徴のみから識別するよりも，外見に加えて，鳥の鳴き声の音響特徴を併用した方が，より精度の高い識別を行うことができると想定するアプローチがこれに該当する．

本稿では，マルチモーダル分析を，人のコミュニケーションやインタラクションの文脈に限定し，その取組みの事例を紹介する．コミュニケーションは我々の社会を構成する基本要素であり，中でも二人以上の間で行われる「会話」は，情報共有，意思決定，合意形成，人間関係の形成・維持などの手段として日々，広く行われている．会話場面において生じる人の行動やその意味，そこで生じる成員間の関係性を理解することは，人間のコミュニケーション活動を支援する技術を開発する上でも重要な学術的課題とされる．その知見は，例えば，対話ロボットや人工エージェントの実現に生かされる．また，会話場面のマルチモーダル分析は，人間行動や会話現象の包括的理解を目指す社会学・言語学・心理学・認知科学などの分野とも関係が深く，分析技術の提供による学際的連携が期待されている．

人の会話では，言語情報だけではなく，非言語情報が交わされている．例えば，「以心伝心」や「場の空気を読む」というような高度な社会的知性には，繊細な非言語情報の表出・授受が深く関わっているとされる．よって，会話場面の理解のためには，対話者の言語・非言語双方の機能解明が不可欠である．会話中に表出される非言語行動には，顔の表情，頭の動き，視線の振る舞い，身振り・手振り，身体姿勢，声のトーン，社会的距離などが含まれ，そのモダリティは多岐にわたる．本稿ではこれらをマルチモーダル分析の対象とする．このような対話場面における対話者の非言語行動を分析や認識のタスクとした研究は，2000 年代中頃から，「会話シーン分析」や「社会的信号処理」と呼ばれ取り組まれおり，今日まで盛んに研究が行われている[(1)〜(3)]．

図 1 にマルチモーダル分析の枠組みを示す．まず，複数人の会話場面をカメラやマイクロホンなどを用いて収録し，得られた画像や音声信号から，各対話者の行動が計測・認識される．ここでは，身体姿勢などの物理量の計測に加えて，各モダリティの非言語行動の検出やその種別・機能の認識が行われる．その後，認識された非言語行動などの特徴から，各対話者の印象や感情，能力，グループ内の結束力やリーダーシップなどの推定が行われる．以下，図 1 の前半部分を 2. で，後半部分を 3. で概説する．

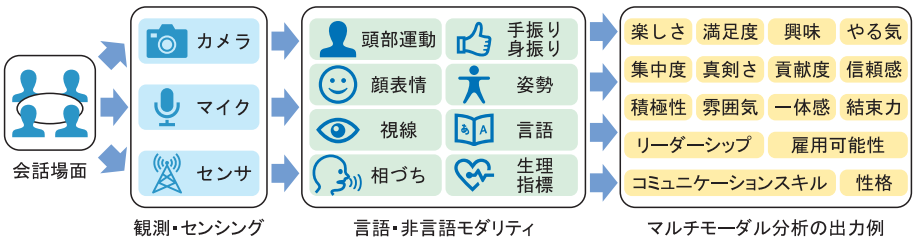

図 1　マルチモーダル分析の枠組み　　会話場面をカメラやマイクロホンなどを用いて収録し，そこで得られたデータから各対話者の言語・非言語行動の認識を行い，更にその結果から対話者の内観・状態などを推定する．

本会ハンドブック「知識の森」
https://www.ieice-hbkb.org/portal/doc_index.html

© 2025 電子情報通信学会

2. 非言語行動のモダリティ

2.1 頭部運動

頷（うなず）き，首振り，傾（かし）げといった頭部運動は，頭部ジェスチャとも呼ばれ，従来，多クラス分類の問題として捉えることが常であった．これは，頷きであれば同意を意味するというように，ジェスチャを単純な記号としてみなすことを意味し，HCIの文脈においては一定の妥当性が認められていた．しかし，人と人との会話において，頭部運動は，話し手，聞き手ともに多数の重要な役割を果たすことが指摘されている[4),(5)]．話し手の機能としては，発話の強調や聞き手に対する反応確認，発話終了のサイン，発話権譲渡などが知られ，聞き手の機能としては，傾聴のサイン，思考，理解，同意，発話権受諾などの機能が上げられ，更に感情表出の一端も担うとされる．筆者らは，これらの機能が，同時に複数出現し得ること，及び，受け手の解釈が曖昧かつ多様であるという性質に注目し，これらを考慮するため，機能認識の問題を，非排他的な検出問題として捉え，頭部姿勢などを入力とする認識モデルを提案している[6]．

2.2 顔表情

従来，顔表情の機能としては，感情表出の側面が主に注目されており，中でも「怒り，嫌悪，恐怖，幸福，悲しみ，驚き」の基本6感情のクラスがよく知られており，顔画像からこれら感情のクラス分類を行う研究が多数行われている．一方，顔の表情には，感情表出以外にもコミュニケーション上の機能を担うことが指摘されている．そのような表情は「顔ジェスチャ」と呼ばれ，眉毛を上げることによる発話の強調や，笑顔が話し手への相づちや応答として機能することが示唆されている[7]．その観点から，今村らは，顔表情の機能44種を定義し，そのうち頻出5種と頭部運動機能10種との共起性を調査し，顔表情の肯定と頭部の相づちの機能が同時に現れやすく，互いの機能を強化する関係にあることなどを考察し，それらの同時出現を捉えるモデルを提案している[8]．

2.3 相づち

会話は，話し手と聞き手による共同行為であり，話し手の発話に対する聞き手のフィードバックも重要な会話の構成要素である．その典型例として，「うんうん」と言った，相づちと呼ばれる短い発話が挙げられる．相づちの機能としては，話し手の発話継続を支持するシグナル，内容理解の提示，話し手の判断の支持，賛成の意思表示，感情表出，情報の追加・訂正・要求などが知られている[5]．従来，聞き手の相づちとその種別の検出には，聞き手音声の統計的，韻律的性質（継続時間，強弱，基本周波数など）に基づくモデルが提案されている[9]．また，対話エージェントの相づち生成のため，適切な相づちの種別を予測するモデルも提案されている[10]．また，相づちは頭部の頷きと同時に現れる傾向があるが，その関係を明らかにするため，飯塚らは29種の相づち機能を定義し，その内，頻出6機能と頭部運動機能の頻出10種との共起関係を分析し，話し手発話の継続を支持するサインである相づちcontinuerについて，約6割の時間割合で頭部運動の同機能が共起することから，これら2モダリティが密接に連携していることを確認している[11]．

2.4 視線

従来，視線には，観察，調整，表出の機能があることが知られている[12]．また，対話者の視線方向（視覚的な注意の焦点とも呼ぶ）の計測・推定の問題が中心的な研究課題とされてきた．加えて，自閉症者が特有のアイコンタクトのパターンを示すことから，視線の分析は，発達障害者の診断・支援のためにも有用とされてきた．一方，視線の持つコミュニケーション上の機能をより広域に捉えるため，田代らは，発話者の注意・視線回避に関わる機能，聞き手による注視・視線回避に関わる機能など，計43種の機能を定義し，顔の表情や頭の動きなどのマルチモーダルな非言語行動から視線の機能を認識するモデルを提案している[13]．

3. マルチモーダル分析の事例

対話者のマルチモーダルな非言語行動から対話者の心的状態や対話者間の関係性を推測・認識する研究が様々行われている．例えば，Itoらは，4人対話において，対話者本人が事後に報告した印象スコアを，頭部運動の機能と顔表情の機能から予測するモデルを提案している[14]．彼らは頭部運動と顔表情それぞれの機能の出現頻度や機能間の比率，及び，顔と頭部の機能の共起率などを特徴量として定義し，特徴選択により有効なモダリティの組合せを検証した．結果，貢献度や興味，楽しさ，難しさなどの印象項目において，頭部と顔の双方の特徴量を用いた場合に最も高い推定性能が達成されることを確認した．また，Katadaらは，エージェント対話におけるユーザの正負の本人心象を認識するため，言語，音響，視覚，生理指標などのマルチモーダル・データを統合した認識モデルを提案している[15]．また，Walochaらは，3人一組のグループタスクにおける成員間の結束力（Cohesion）を予測するため，対人距離や歩行距離，身体姿勢などの物理情報を入力とするモデルを提案している[16]．

4. 今後の展望

本稿では，人の会話場面に的を絞り，マルチモーダル分析の事例を紹介した．近年，深層学習分野においてすう勢をなすEnd-to-endアプローチを用いることで，多様なモダリティを統合することは比較的容易となっている．また，「説明可能AI」の手法を用いて，学習済みモデルから認識に寄与した特徴量を特定し，対話者の主観と行動との関連性を分析するという試みも始まっている．非言語行動は，言語や文化，性別，対話のタスクなどの要因に強く依存するため，その全容を理解するためには更に多くの事例で検証を続けることが必要である．以上を含め，今後，マルチモーダル分析は，計算機科学分野のみならず，幅広い分野において有用なツールとして発展が期待される．

文　献

（1） A. Vinciarelli, M. Pantic, and H. Bourlard, “Social signal processing : Survey of an emerging domain,” Image Vis. Comput., vol. 27, no. 12, pp. 1743-1759, Nov. 2009.

（2） D. Gatica-Perez, “Automatic nonverbal analysis of social interaction in small groups : A review,” Image Vis. Comput., vol. 27, no. 12, pp. 1775-1787, Nov. 2009.

（3） K. Otsuka, “Conversation scene analysis,” IEEE Signal Processing Magazine, vol. 28, no. 4, pp. 127-131, 2011.

（4） D. Heylen, “Head gestures, gaze and the principles of conversational

structure,” Int. J. Humanoid Robot., vol. 03, no. 03, pp. 241-267, 2006.
（5） S.K. Maynard,“On back-channel behavior in Japanese and English casual conversation,” Linguistics, vol. 24, no. 6, pp. 1079-1108, 1986.
（6） K. Otsuka and M. Tsumori,“Analyzing multifunctionality of head movements in face-to-face conversations using deep convolutional neural networks,” IEEE Access, vol. 8, pp. 217169-217195, 2020.
（7） J. Bavelas, J. Gerwing, and S. Healing,“Including facial gestures in gesture-speech ensembles,” From Gesture in Conversation to Visible Action as Utterance, pp. 15-34, John Benjamins, 2014.
（8） 今村まい，武田一輝，熊野史朗，大塚和弘，“対話中の頭部運動と顔表情の相乗機能の認識，”信学論（A），vol. J106-A, no. 3, pp. 70-87, March 2023.
（9） D. Lala, K. Inoue, P. Milhorat, and T. Kawahara,“Detection of social signals for recognizing engagement in human-robot interaction,” Proc. AAAI Fall Symposium on Natural Communication for Human-Robot Collaboration 2017, 2017.
（10） D. Ortega, C.-Y. Li, and N.T. Vu,“OH, JEEZ! or UH-HUH? A listener-aware backchannel predictor on ASR transcriptions,” Proc. ICASSP 2020, pp. 8064-8068, 2020.
（11） 飯塚海斗，大塚和弘，“対話中における聞き手の頭部運動と相槌の相乗機能の解析，”人工知能論文誌，vol. 38, no. 3, 2023.
（12） A. Kendon,“Some functions of gaze-direction in social interaction,” Acta Psychologica, vol. 26, pp. 22-63, 1967.
（13） 田代絢子，今村まい，熊野史朗，大塚和弘，“マルチモーダル非言語行動に基づく対話者の視線機能の認識，”信学論（A），vol. J106-A, no. 12, pp. 296-311, Dec. 2023.
（14） K. Ito, Y. Ishii, R. Ishii, S. Eitoku and K. Otsuka,“Exploring multimodal nonverbal functional features for predicting the subjective impressions of interlocutors,” IEEE Access, vol. 12, pp. 96769-96782, 2024.
（15） S. Katada, S. Okada, and K. Komatani,“Transformer-based physiological feature learning for multimodal analysis of self-reported sentiment,” Proc. 2022 Int. Conf. Multimodal Interaction, pp. 349-358, 2022.
（16） F. Walocha, L. Maman, M. Chetouani, and G. Varni,“Modeling dynamics of task and social cohesion from the group perspective using nonverbal motion capture-based features,” Proc. Companion Publication of 2020 Int. Conf. Multimodal Interaction, pp. 182-190, 2020.

（2024 年 10 月 21 日受付）