3L3-OS17-7

マルチモーダルデータに基づいた多人数インタラクションの構造理解

Multimodal Data Analysis of Multiparty Conversation

角 康之[*1]
Yasuyuki Sumi

[*1]公立はこだて未来大学
Future University-Hakodate

This paper shows the author's attempt of recording and anlyzing human conversational interactions. We designed an environment to record audio/visual, human-motion, eye gazing data for building interaction corpus mainly focusing on understanding of human nonverbal behaviors. In this paper, we present our attempt to estimate the activeness of each participant with the nonverbal behaviors. We also show a preliminary result of interaction mining; extraction of sequential patterns of nonverbal behaviors from huge set of recorded data on multiparty conversations.

1. はじめに

我々人間は会話において、視線、ジェスチャ、うなずき、あいづちといった非言語行動によって様々な意図を表現する。これらの非言語行動には一定の時間的・空間的なパターンがある。我々は、非言語行動によって、無意識のうちに互いの心的状態を伝え合ったり、会話の流れを制御している。しかし、現在のコンピュータは、そういった人の非言語行動の意味的構造を理解できない。

コンピュータは従来のデスクトップ型の形だけでなく、情報家電、ロボット、センサネットワークなどの形で我々の社会的活動に浸透しつつある。そういったコンピュータを我々の社会的パートナーとして認めるには、言語的な情報だけでなく、我々が何気なく使っている非言語的な情報も、コンピュータに理解してもらう必要がある。近年の Web の発展などに伴う言語的な研究資源が言語情報学の発展に大きく寄与したように、非言語情報の研究は実際の人のインタラクションから得られた非言語データを研究資源とする必要がある。

本稿では、人のインタラクションを記録・分析するための環境構築に関する筆者らの試み [角 08, Sumi 10] を紹介する。まず、実世界インタラクションのセンサ環境について述べ、コーパスに基づいたインタラクション研究の考え方を示す。そして、センサ環境を用いて記録された多人数会話データや、会話構造分析の一部を紹介する。

2. インタラクション計測環境

筆者らは、文部科学省科研費特定研究「情報爆発時代に向けた新しい IT 基盤技術の研究」の一環で、京都大学情報学研究科内の一室（約 80 平方メートル）に、会話的インタラクションを計測するための環境として、IMADE (Interaction Measurement, Analysis, and Design Environment) ルームと呼ばれる環境を構築した。この環境は、人同士のインタラクションに関する様々な種類のマルチモーダルデータ、具体的には、映像、音声、移動、視線、生体反応といったデータを統合的に計測するために設計された。

インタラクションのコーパスを構築することを目的とした研究プロジェクトは、これまでにもいくつかなされてきた。その代表的なものである AMI[Carletta 06] は、グループミーティングのコーパスを構築し、主に会話分析を行った。CHIL[Waibel 09] は、機械学習手法による人の動作自動検出に焦点を当てた。VACE[Chen 06] は、ミーティングの視覚的コンテンツの蓄積と分析を行った。筆者らの目的は、会話の微視的な分析（例えば、発話交替や視線の時間構造分析など）だけではなく、巨視的な分析（つまり、会話グループの生成・分解や移動などのダイナミクスの分析）も目的としている。したがって、筆者らは、着座式のミーティングだけでなく、自由に歩き周りながらのおしゃべりや、ポスター発表などを含む、様々な種類の多人数会話をターゲットとしてきた。

IMADE ルームの構成を図 1 に示す。IMADE ルームでは、インタラクション行動を記録するために、以下のような様々なセンサを設置した。

環境カメラ 複数の映像記録用カメラが室内上部に設置されていて、インタラクション状況を映像として記録できる。AXIS 210A のネットワークカメラを 8 台利用した。

ヘッドウォーンマイク 環境内のすべての人が装着することにより、各人の発話内容を人ごとに分離して収録した。

モーションキャプチャ 環境内の人や物の各部にマーカーを装着し、人の動きや他者との位置関係を 3 次元座標データとして記録することができる。Motion Analysis 社の MAC3D システムを用いた。

アイマークレコーダ 環境内の各人の眼球運動を計測し、頭部に設置した一人称映像とその中の 2 次元座標データとして記録できる。Applied Science Laboratories 社製の Mobile Eye と、NAC イメージテクノロジー社の EMR-9 を利用した。

データ統合と閲覧 様々な異種センサによるデータが蓄えられるので、NTP(Network Time Protocol) による時間同期や各データの時間伸縮を吸収するための後処理を行った。また、複数センサデータ間の空間統合、例えば、モーションキャプチャで計測された頭部の位置・方向のデータと、アイマークレコーダによって得られる相対座標系の視線データを統合して、絶対座標系の視線データを生成した。

連絡先: 角 康之，sumi@acm.org

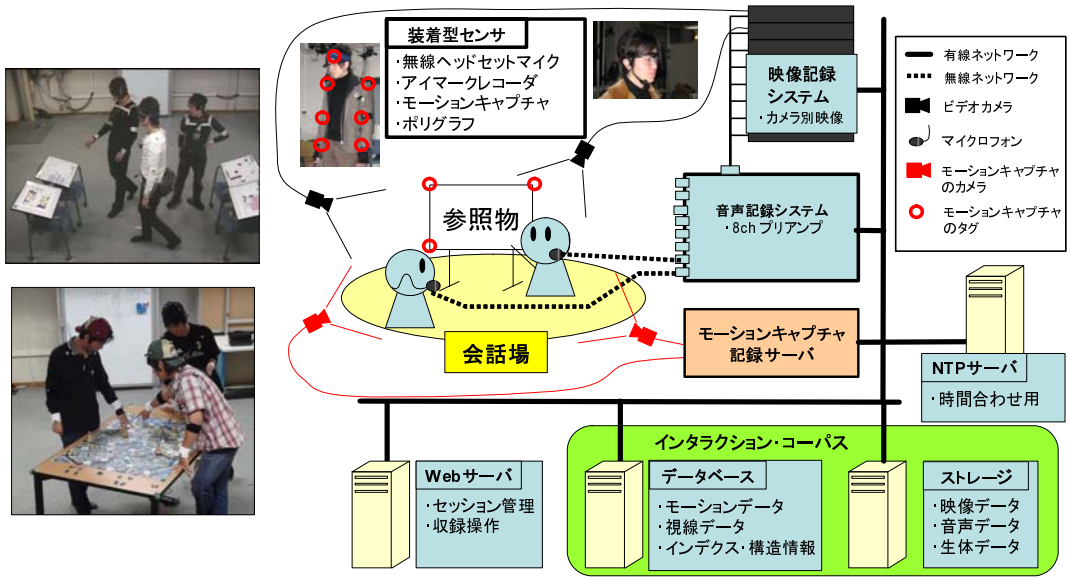

図 1: 実世界インタラクションの計測環境

これらのセンサに加えて、生体反応データ（筋電、脳波、脈拍など）を計測したい場合はポリグラフを一緒に利用したり、頭部のうなずき動作を簡易に計測するためにモーションセンサを利用した。

3. インタラクションコーパスに基づいた会話構造の分析

筆者らは、IMADE ルームで計測されたデータを図 2 に示すような階層的解釈モデルに基づいて蓄積し、これをインタラクションコーパスと呼んでいる。そうすることで、規則されたデータを構造化し、整理された形で解釈を積み上げていくことができる。

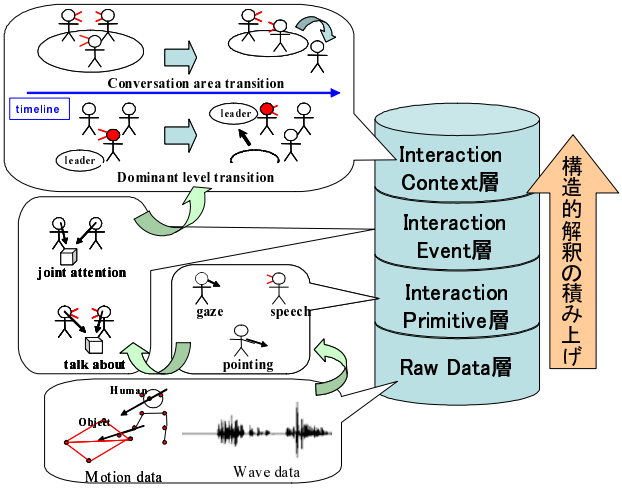

図 2: インタラクションの階層的解釈モデル

インタラクションコーパスは 4 層に分かれている。最下層はセンサにより収録されたデータそのものであり、データの解釈はなされていない。次の層はセンサデータからインタラクション行動を個別に切り出したプリミティブ層で、発話や注視などがここに分類される。3 番目の層は、時空間的に共起するプリミティブを組み合わせたものであり、会話に参加する人同士の社会的インタラクションとして興味深い現象、例えば、共同注視や特定の注目対称を共有しながらの会話などがここで観測される。最上位層はさらにインタラクションの文脈（流れ）を解釈する層であり、例えば、会話場の発生やメンバーの移動や、また、会話のリーダーの交替といった抽象度の高い解釈を試みる層である。

4. 視線を用いた指差しジェスチャ検出の精度向上

以下、IMADE を利用した研究事例紹介として、筆者らによる会話構造分析の試みを紹介する。まずここでは、会話参加者によるジェスチャの自動検出について紹介する。会話の中では、形状を表すためのハンドジェスチャや、指さしジェスチャが頻繁に行われる。指差しは会話の中で参照している対象物を示す行為であり、会話の内容の理解や、会話参加者の参加積極性を計るのに役立つ。

モーションキャプチャを利用すれば、腕が伸びた状態で指が指し示している方向に存在する対象物を特定することで、指さしジェスチャとその対象物を判定することは容易であると考えられがちである。しかし実際は、指差し行為以外にも人の腕は頻繁に動くし、指差し対象物を特定することもそれほど容易ではない。

そこで筆者らは、指さし行為というものを行為者のみの ego-centric な行為とは考えず、会話のパートナーが存在すること、もっと正確に言うと、パートナーが行為者の指さし方向に注目することで初めて指さし行為が成立する、social な現象であると考えた（図 3）。つまり、会話参加者が指さし方向の対象物に視線を向ける行為が同期したものだけを指さしジェスチャとして認定することで、指さしジェスチャの検出精度が向上するかを確かめた [矢野 09]。

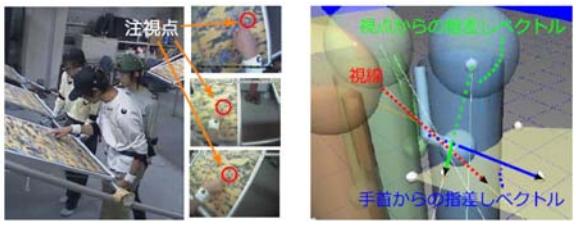

図 3: 指さしジェスチャの抽出

以下の手順で指さしジェスチャの自動抽出を試みた。まず、指差しベクトルを定義した。指差しベクトルは、腕の伸びた方向（つまり、肘と掌をつないだ方向）を用いたものと、目から指先にのばした方向を用いたものの 2 種類を準備した。次に、指さしベクトルの先に指さし対象となり得る対象物（つまり、会話参照物となるポスターや他の会話者の身体）があるかどうかを網羅的に検索し、指さしジェスチャの候補を広めに抽出した。そして、それらの指さしジェスチャ候補それぞれの発生と同期して起きている各会話参加者の視線データを参照し、それらの視線ターゲットが指さし先の対象物と一致しているかを確認した。

結果は図 4 のようになった。指さしジェスチャの検出精度を求めるために、実際の指さし行為と思われるものをハンドラベリングし、それを正解データとして、各手法の再現率・適合率を比較した。指さしベクトルの判定については、全体を通して、目から指先にのばしたベクトルの方が精度が高いことが確認された。

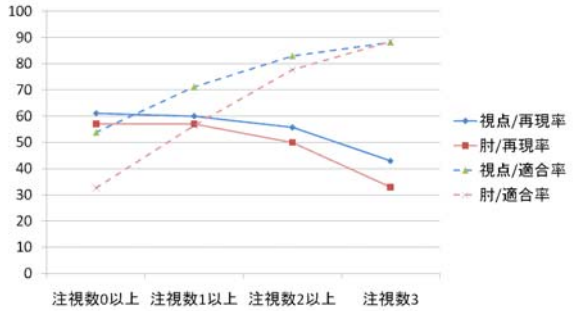

図 4: 視線獲得による指さしジェスチャ抽出の精度向上

会話参加者の視線獲得の影響については、以下のことが観察された。視線獲得に関係なく指さしベクトルが何らかの対象物に衝突しているものをすべて指さしジェスチャと認定してしまうと（グラフの一番左）当然再現率は良いが適合率が極めて低い。一方、会話参加者全員（3 人）すべての視線を獲得していることを条件としてしまうと（グラフの一番右）条件が厳しすぎるのか、再現率が急激に下がってしまう。グラフからは、3 人中 2 人以上の視線を獲得しているくらいが、再現率・適合率のバランスがとれていることがわかる。

5. 非言語情報による会話状況の構造理解

ここでは、発話内容の意味的な解釈を伴わず、非言語情報のみから会話状況の解釈を試みた例を示す。具体的には、タスク遂行型の 3 人会話において、会話参加者ごとの会話参与に対する積極性を、非言語情報のみから推定することを試みた [中田 08]。

我々の興味は、抽象度の低い非言語行動の要素の組み合わせから、会話の意味的な状況やシーンの転換点などを見つけることである。その最初の試みとして、発話、視線変化、指差し行為といったインタラクション・プリミティブの組み合わせから会話参加者の会話参与積極性を数値化することを試みた（図 5）。具体的には 3 人によるボード上の作業を伴う合意形成型の会話状況を設定し、35 分間の会話データを計測した。そのデータから、分析者によって分けられた 16 のシーンについて、9 人の被験者からの主観的評価（各シーンにおいて 3 人の参加者の積極性を順位付ける）を平均化し、正解データとした。その正解データと、我々の計測環境から得られたインタラクション・プリミティブ、インタラクション・イベントを説明変数とした重回帰分析を行った。

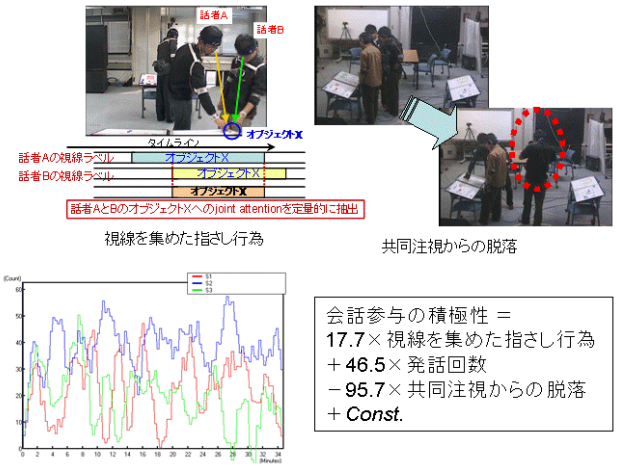

図 5: 会話参加の積極性の数値化

その結果、従来研究（例えば [Rienks 06]）と同様に発話量は積極性に対して強い正の相関を示した。それ以外に、「視線を集めた指差し行為」に正の相関が見られ、逆に、「共同注視からの脱落」に大きな負の相関が見られた。これらは我々の直感と合い、つまり、データ分析的なアプローチで、人の直感にあう社会的インタラクションの解釈を見出すことの可能性を示すことができたと考えている。

6. インタラクションマイニングによる会話構造抽出

我々は会話中に、視線、指差し、頷きといった様々な非言語情報を無意識のうちに用いながら、発話内容の補完をしたり会話の制御を行っており、それらの出現パターン（会話構造）には一定の構造がある。前節までは、そういった会話構造について、先に仮説を立て、データに基づいてその検証を行うというアプローチをとってきた。

その一方で、IMADE を用いて多くの会話的インタラクションのデータをコーパス化することの意義のひとつは、データの中からボトムアップ的に新しい会話構造（会話プロトコルと呼んでも良いであろう）を見つけられる可能性があることである。また、会話構造は、会話の状況、会話参加者の個性、会話内容によって大きく変わると思われるので、会話構造を単独で議論するのではなく、そういった周辺的状況とあわせて会話構造の発生パターンを理解したい。

そこで筆者らは、データマイニング的手法を用いて、会話中に発生する非言語行動の出現パターンを抽出する手法をインタラクションマイニングと名付け、会話状況を特徴づける会話構造発見を試みている [福間 09, 中田 11]。この手法は、発話の有無、指さし、視線、うなずき、相槌といった非言語インタラクションが同時に出現する状態（インタラクションステート）の時間変化パターンを N-gram で表現し、$\chi 2$ 乗検定によって機械的に有意なパターンを抽出する方法である。図 6 はその一例であり、頻出するインタラクションステート（この例では、3 人がポスターに共同注視している状態）から続く一連の状態変化を示している。この例からは、共同注視しているときには、発話者が聞き手よりも指さしをすることが多く、発話が重なったときには元の発話者が発話を続けることが多い、といったものが読み取れる。このことは、我々が普段行っている無意識の常識的な会話プロトコルがインタラクションコーパスから機械的に抽出できることを示している。

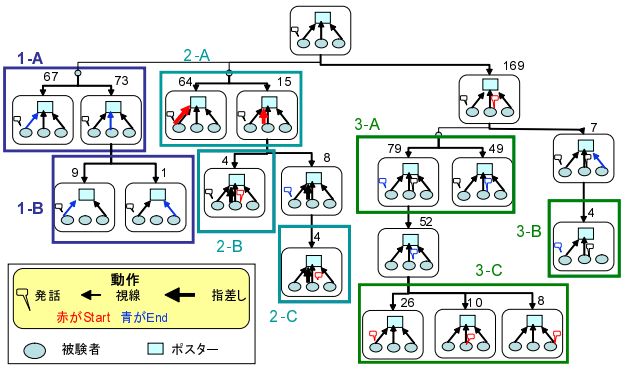

図 6: インタラクションマイニングによって抽出された会話構造の例

この手法を用いてポスター発表会話とポスター環境自由会話という 2 種類の会話状況における会話構造の自動抽出を試みた。その結果、発話者は非発話者より指差しが多い、とか、頷きの後に相槌を行うことが多いといった会話構造は 2 つの会話状況に共通して見られる一方で、沈黙の後には元の発話者が発話を続ける傾向が高いという会話構造はポスター発表会話特有のものであるといったことを確認することができた。このことより、インタラクションマイニングが会話状況の差異を可視化する道具として使える可能性を示すことができたと考える。

7. おわりに

会話時の非言語的行動を詳細に計測し、それらの出現パターンを分析するとこで、会話積極性の定量的な推定や、会話プロトコルの半自動的な抽出が可能になることを示した。我々の大きな目標は、非言語行動に関する「辞書と文法」を構築することであるが、本稿で示したのは、それに対するトップダウンな方法（仮説を立てて検証する）とボトムアップな方法（データから構造発見する）の初期的な試みであると位置づけられよう。今後、これらの融合を強め、多人数会話の構造理解を促進するための研究資源の構築を進めていきたい。

謝辞

本稿の内容は主に、著者が 2011 年 3 月まで所属していた京都大学情報学研究科において、西田豊明、河原達也、高梨克也、坊農真弓の諸氏や学生諸氏と議論・協力しながら進めたものである。

参考文献

[Carletta 06] Carletta, J., et al.: The AMI meeting corpus: A pre-announcement, in *Second International Workshop on Machine Learning for Multimodal Interaction (MLMI 2005)*, Vol. 3869 of *Lecture Notes in Computer Science*, pp. 28–39, Springer (2006)

[Chen 06] Chen, L., et al.: VACE multimodal meeting corpus, in *Second International Workshop on Machine Learning for Multimodal Interaction (MLMI 2005)*, Vol. 3869 of *Lecture Notes in Computer Science*, pp. 40–51, Springer (2006)

[Rienks 06] Rienks, R. and Heylen, D.: Dominance detection in meetings using easily obtainable features, in *Second International Workshop on Machine Learning for Multimodal Interaction (MLMI 2005)*, Vol. 3869 of *Lecture Notes in Computer Science*, pp. 76–86, Springer (2006)

[Sumi 10] Sumi, Y., Yano, M., and Nishida, T.: Analysis environment of conversational structure with nonverbal multimodal data, in *12th International Conference on Multimodal Interfaces and 7th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2010)*ACM (2010)

[Waibel 09] Waibel, A. and Stiefelhagen, R. eds.: *Computers in the Human Interaction Loop*, Springer (2009)

[角 08] 角 康之, 西田 豊明, 坊農 真弓, 來嶋 宏幸：IMADE: 会話の構造理解とコンテンツ化のための実世界インタラクション研究基盤, 情報処理学会誌, Vol. 49, No. 8, pp. 945–949 (2008)

[中田 08] 中田 篤志, 來嶋 宏幸, 角 康之, 西田 豊明：移動・動作に関するセンサデータによる多人数会話の解釈, 第 22 回人工知能学会全国大会 (2008)

[中田 11] 中田 篤志, 角 康之, 西田 豊明：非言語行動の出現パターンによる会話構造抽出, 電子情報通信学会論文誌, Vol. J94-D, No. 1, pp. 113–123 (2011)

[福間 09] 福間 良平, 角 康之, 西田 豊明：人のインタラクションに関するマルチモーダルデータからの時間構造発見,, 情報処理学会研究報告（ユビキタスコンピューティングシステム）, Vol. 2009, No. 23 (2009)

[矢野 09] 矢野 正治, 中田 篤志, 福間 良平, 角 康之, 西田 豊明：非言語マルチモーダルデータを用いた会話構造の分析のための環境構築, 情報処理学会研究報告（ユビキタスコンピューティングシステム）, Vol. 2009, No. 22 (2009)