C-10

非言語マルチモーダルデータを用いた会話構造分析

―インタラクションマイニングによる会話プロトコルにおける構造抽出―

Analysis of Interaction Structure from Multimodal Nonverbal Data

―Extracting Structure of Conversation Protocol by Interaction Mining

中田　篤志†　　福間　良平†‡　　角　康之†　　西田　豊明†
Atsushi Nakata　Ryohei Fukuma　Yasuyuki Sumi　Toyoaki Nishida

１．はじめに

我々が会話を行う際には，発話や視線，指差しといった非言語的な行動によって会話を制御している．このときの行動は，言語に節や文法といったプロトコルが存在するのと同様に，一定のプロトコルを持って行われている．例えば「返答を期待するときは，その相手の顔を見ながら話す」「指差しを行う前には，他の人々の注目を得てから行う」といったものである．

このような会話における非言語行動のプロトコルを明らかにすることができれば，その構造を基にエージェントやロボットなどの人工物が今よりも自然な形で人とインタラクションを行うことが可能となる．例えばロボットが伝えたい情報がある場合に，それを唐突に伝えるのではなく，うなずきや視線配分を利用して発話の権利を得てから情報を伝えるといったことが期待される．

このような会話的インタラクションにおける構造を明らかにしようとする試みはこれまでにも多く行われている[1] [2]．しかしこれらの多くは，ある仮説を検証するために統制された環境で実験を行うという手法や，自然会話の中から分析者が着目したエピソードのみを取り出して議論するという手法が多い．前者は多数の要因によって生まれる複雑な現象は扱いにくく，後者の手法では得られた知見がどの程度汎用的なのか，どのような状況で発生しやすいのか，といったことについて議論するのが困難である．また，将来的に人工物に応用していくことを想定した場合，機械的に取得できるセンサデータと結びつけた状態で，非言語行動の取得やそれらに関する構造分析を行っていくことが重要となる．

以上のことから，我々の研究ではセンサによって取得された計測データに基づいて解釈を極力自動化することを試み，またそのデータを数理的にモデル化して，人同士の会話構造の「辞書と文法」を構築することを目指している．構築されたものがセンサデータと結びついていることで，人工物への応用もより容易に行えると考えている[3].

このような計測データの自動解釈に関する先行研究としては，森田らの研究[4]が挙げられる．この研究では，ウェアラブルセンサにより自動付加されたラベルからのパターン抽出を試みている．しかし，この研究では構造の時間変化については検討されておらず，また構造の頻出順を正規化し示すのみであった．

以上のことを踏まえ，本論文では会話的インタラクションの文法作りにおける初期の試みとして，計測された会話計測データから発話の有無・視線移動・指差し行為といった非言語情報の発現について生起順序を基に構造化し，さらにその構造の中から発言尤度の高い構造を発見することを試みる．そして，従来研究[1][2]で指摘されてきたような会話構造をデータマイニング的な手法によって発見することで，我々の提案する手法が有用であることを示す．

†京都大学大学院情報学研究科，Graduation School of Informatics,Kyoto University

‡現在は奈良先端技術大学院大学，Nara Institute of Science and Technology

２．構造化および特徴的構造の発見

本研究では，会話的インタラクションの構造化，および発現尤度の高い構造を抽出するための手法として，福間ら[5]が提案したインタラクションマイニングの手法を用いている．

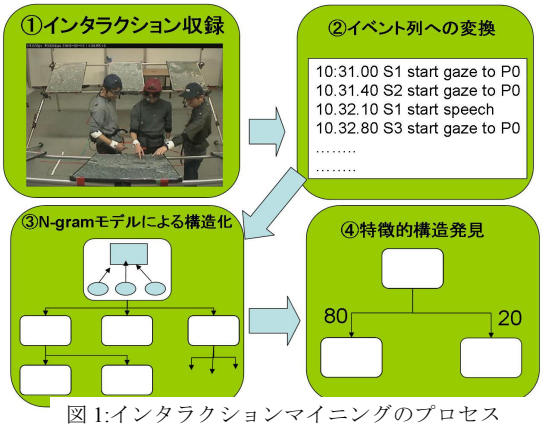

図 1:インタラクションマイニングのプロセス

インタラクションマイニングでは，図１のようなプロセスで会話的インタラクションを構造化し，特徴的な構造を発見する．以下で簡単に説明する．

①　会話的インタラクションの収録

多人数が会話をしている様子を多数のセンサを使って収録する．

②　ラベリングとイベント列への変換

注目する非言語行動に関して，その開始点と終了点・対象に関する情報を付与する．情報の付与をラベリング，作成された情報をラベルと呼ぶ．その後，作成されたアノテーションをイベント列としてひとつにまとめる．

③　N-gram モデルによる構造化

新たに「インタラクションステート」という状態を定義し，インタラクションステートを単語に相当する状態として N-gram によるインタラクションの構造化を行う．

ここで，インタラクションステートはある瞬間のラベル情報の組を表す．例えば「３人が同じパネルを見ていてそのうち一人が発話中である」「３人中２人が互いの顔を見ていて，３人目は２人のうち片方を見ている」といった状態がインタラクションステートに当たる．

④　χ２乗検定による構造の抽出

「異なる状態にある被験者の間で，非言語行動の発生・終了に偏りは存在しない」という帰無仮説を基にχ２乗検定を行い，一定の有意水準より小さい確率で発現している構造を抽出する．

３．インタラクションデータの収録

本章では，評価に利用した多人数インタラクションデータの収録について解説する．

まず，構造化の対象となる非言語行動として，発話・視線・指差しを対象とした．これらの非言語行動は会話の中で重要な役割を果たすことが従来研究で述べられていると共に，自動検出が比較的容易である．そして，これらを検出するため，頭・腕・背中にモーションキャプチャのマーカーを取り付け，また視線計測用のアイマークレコーダと無線マイクを取り付けた．センサを取り付けた被験者の様子を図 2 に示す．

被験者の数は 3 人とした．この理由は 2 つある．まず 3 人以上になると，発話交替における立場の変化，指さしに対する注視・非注視，立ち位置の変化といった，会話に伴う興味深い社会現象が多く発生するといわれている[1][2]．また提案手法ではすべての被験者が同じ会話場にいることを前提としているが，3 人であれば常に一つの会話場を構成しているとして問題ないと考えられる．

会話環境としては，被験者が図 3 のように配置された 6 枚のポスターを見ながら自由に会話を行うという形式をとった．これは，我々が興味の対象としている，話題の発生や発話者の遷移，さらにその前後における話者や聞き手の非言語行動を多く観測するための設定である．また，視線や指差しの対象を増やし，同時に自動検出を容易にすることも目的としている．

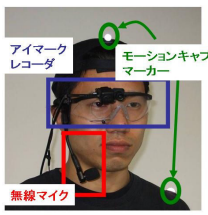

図 2:センサの取り付け

図 3 会話環境の様子

本研究では，提案手法が多くの種類の会話に対し適用できることを示すため，収録ごとにポスターの内容や教示内容を少しずつ変えながら行った．行った収録は下記の 3 回である．

① 室町時代の京都全体を描いた絵画をポスターとし，書かれた人物や歴史的背景について雑談する
② 京都の航空写真をポスターとし，地図上の建物などについて雑談する
③ 京都の航空写真をポスターとし，地図上の寺社仏閣を探すタスクを行う

４．ラベリング

ここでは，収録されたデータに対して行うラベリングについて解説する．

１で解説したとおり，我々はセンサによって自動的に得られたラベルによって分析を行うことを目標としている．しかし，現状では自動認識は精度が悪く多数のノイズが発生する上，構造化を行った際ノイズによる偏りが特徴的な構造として検出される場合が多数見られた．そのため，ここでは自動認識の後，明らかなノイズを人手で修正したものをラベルとして利用した．

以下に，視線・指差し・発話のそれぞれについて自動認識の手法と手作業修正の基準について述べる．なお，文中における時間幅などのパラメータについては，実験の前に行われた数度の予備実験で得られた最も精度よく取れる値を利用している．

４－１．視線

視線ラベルの生成に当たっては，まずアイマークレコーダとモーションキャプチャから視線ベクトルを計算し，その後他の人の頭部をモデル化した球体と，ポスターをモデル化した長方形との衝突判定を行った．さらにセンサからデータが取得できない場合があることを考慮し，250ms 以下の短いラベルに対して補間を行った．

その後，人手での訂正が必要な箇所については，それぞれのアイマークレコーダの映像を参考にして訂正を行った．

ただし，3 回目の実験における被験者の 1 人の視線データは，アイマークレコーダのキャリブレーション用データに重大な欠損があったため，完全手作業でラベルを作成した．

４－２．指差し

指差しラベルの生成のため，まず指差しベクトルを計算した．このベクトルの始点は頭部のモーションキャプチャのマーカー位置から推定した眼の位置であり，その方向は手首につけられたモーションキャプチャのマーカーの位置である．次に，このベクトルとポスターをモデル化した長方形との衝突判定を行った．その後衝突判定から得られたラベルのうち間が 250ms 以下のものを補間し，さらに発生区間が 150ms より短いラベルについては削除を行った．

また，人手での訂正は，基本的に誤認識したラベルを削除することで行ったが，モーションキャプチャのマーカーが隠れている等の理由でラベルが断続的に生成されている場合は補間を行った．また，ラベルの開始点や終了点に問題があり，ずれている場合は，指先もしくは手のひらがボードに向いているかどうかを判断基準として調整を行った．

４－３．発話

発話ラベルの生成は，まず各被験者が装着したマイクから得られた音声波形を 50msec ごとに分割し，FFT を用い音量を計算したうえで適当な閾値で 2 値化を行った．閾値は実験の最初の部分に人手でラベルを付加したうえで，それらの部分で再現率 90%を超え適合率を最大とする値とした．

次に，隣接するラベル同士の間が 250ms 以下のものを補間した．最後に，発生区間が 150ms より短いラベルを削除した．

また，人手での訂正では，基本的にラベルの追加は行わず誤認識したラベルを削除することで行った．ラベルの開始点や終了点の調整が必要な場合には，それらの点は音の立ち上がり・立下りの点にあわせた．

５． 特徴的なインタラクション構造の抽出結果

作成された３回分のラベルをまとめて，N-gram による構造化と $\chi 2$ 乗検定による特徴的な構造の抽出を行った．このとき，$\chi 2$ 乗検定の有意水準は 5%とした．

本論文では，得られた構造を下記の 2 通りの手法で検証する．

1. 最も多く出現したインタラクションステートを初めとした構造に関して，４段階の遷移までを検証する
2. 1 段階の遷移で有意差が見られた構造のうち，総ステート数が多いもの上位 5 つに関して検証する．

5-1. 最頻インタラクションステートを基点とした構造

図４が，最も多く出現した（出現回数 972 回）インタラクションステートを頂点とする構造である．このインタラクションステートは，「3 人がボードを注視し，かつそのうちの一人が発話中である」という状態である．図では発現頻度に有意差が見られた構造，およびそれに関連する構造のみを記述し，他は省略してある．矢印の側の数字はそれぞれの遷移の回数を表している．

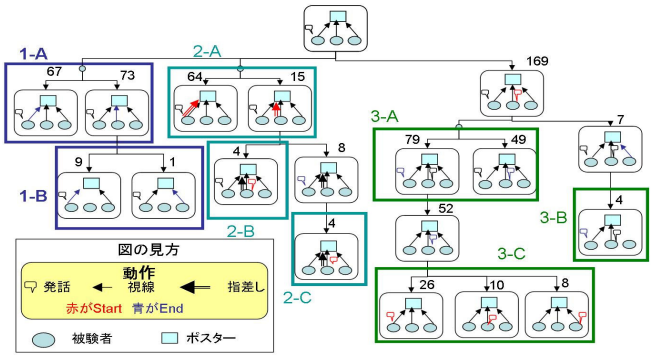

図 4:最頻インタラクションステートを頂点とする構造

以下は得られた構造を３つの着眼点で検証する．

① 視線の遷移に関する構造

1-A，1-B は，いずれも被験者の視線の遷移に関する特徴的な構造を示したものである．

1-A に関してだが，初期の「全員がボードを注視」という状態から，発話中の人物が視線を外す回数が 67 回，発話していない人物が 73 回となっている．仮に会話におけるプロトコルが存在しない場合，発話中の人物と発話中でない人物が視線を外す確率はそれぞれ 1/3，2/3 となるはずであり，ここには何らかのプロトコルが存在していると考えられる．

これらの構造が発生している場面を実験の映像から確認したところ，発話者・非発話者に関わらず，他の発話者に視線を送る場合や別のポスターに視線を向けて会話場を変化させる場面が多く見られた．

以上のようなことから，発話者は非発話者よりも頻繁に視線を配分し，会話をコントロールしているのではないかと考えることができる．このような現象は従来研究でも論じられている[1][2].

また，非発話者が視線を外した場合でも，その後発話者が視線を外す回数はもう一人の非発話者が視線を外す場合に比べて有意に多い（1-B）．この事実も，1-A で見られた会話構造を補強するものだと考えることができる．

② 指差しに関わる構造

2-A，2-B，2-C は，いずれも発話と指差しの関係についての構造である．

2-A に関してだが，初期のステートから注視対象を指差す回数は，発話者が 65 回，非発話者が 15 回となっており，発話者の指差し回数は非常に多いといえる．

また，非発話者が指差しを行った場合に注目すると，指差しを行った被験者が直後に発話する場面は見られたが，行っていない被験者が発話する場面は見られなかった（2-B）．さらに，初期状態で発話していた被験者が発話を終了した後，指差しを行った被験者が発話する場面はみられたが他の被験者が発話する場面は見られなかった(2-C).

これらは生起回数が少ないので必ずしも会話のプロトコルをあらわしているとは言い難いが，これが会話のプロトコルである場合は「指差しと発話は強い共起関係があり，前後関係にはあまり大きな意味がない」と考えることができる．事実，これらのステートの発生箇所を確認したところ，被験者が発話をしながら発話内容に関わる指差しを行っている場面が多かった．

③ 同時発話時の発話権の遷移に関する構造

3-A，3-B，3-C は，いずれも 2 人が同時発話を行ったときの発話の遷移に関わる構造である．

3-A に関してだが，初期のステートから「2 人が同時発話」の状態に移った後，先に話していたほうが発話をやめる回数が 79 回，後から話し始めた方が発話をやめる回数が 49 回であった．プロトコルがない場合の発話終了確率はどちらも 1/2 であり，ここには何らかのプロトコルがあると考えられる．

また回数は少ないものの，3-B の構造から，このプロトコルは間に何らかのイベントを挟んでも成り立つのではないかと考えられる．

さらに特徴的な場面として 3-C が挙げられる．この構造は，発話が重なって双方が発話をやめた後，初めに発話をしていた被験者が発話をする回数が，他の被験者よりも有意に多いことを示している．これは，「一般的な会話では，片方の話者が一方的に話すよりもある程度の発話のやりとりをしながら話すほうが自然である」ということを表していると考えられる．

５－２．１段階遷移における特徴的構造

図 5 の 5 つの構造は，1 段階のみの遷移において特徴的な構造を持つもののうち，総生起回数が多い順に５つを取り出したものである．

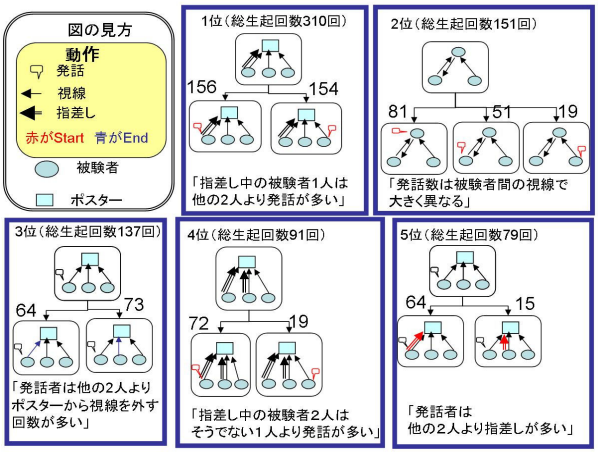

図 5: 1 段階遷移における発現尤度の高い構造

これらの構造のうち，1 位・4 位・5 位は，指差しと発話の共起関係に関する構造であると考えられる．まず，1 位と 4 位の構造から，「指差しを行っている被験者は，その直後に発話を行う場合が他の被験者に比べて多い」ということがわかる．さらに，5 位では「発話者は非発話者に比べて指差しをする場合が多い」ということが分かる．

以上のことから，「発話と指差しには強い共起関係がある」という会話のプロトコルが読み取れる．このプロトコルは普段の日常会話でも納得のできるものであり，また発話とジェスチャの共起関係については先行研究でも議論がなされている[2]．このような場面が自動的に発見できたことは，我々の提案手法を支持する結果であると考えられる．

次に 2 位の構造に着目する．この構造は 3 人が互いの顔を見ながら話している場面で，発話の回数が下記の順に多くなっていることを示している．

① 他の被験者の 1 人と視線を交し合っており，かつもう一人の被験者から視線を受けている人物
② ①と視線を交し合っている人物
③ ①を注視しているが①②のどちらからも注視されていない人物

3 人会話において視線が非常に大きな役割を果たしていることはこれまでに研究がなされており[1]，それらの研究において議論されている構造が自動的に発見できたことは提案手法を支持する結果であると考えられる．

最後に 3 位の構造に着目する．この構造から，「発話者は全員が見ているポスターの対象から視線を外す場合が非発話者よりも多い」ということがわかる．この構造から，発話者は他の被験者を見て様子を伺う，他のポスターを見て会話場の移動を促すといった，会話をコントロールする行動を多くとっているのではないかと考えられる．

6． おわりに

本論文では，収録したインタラクションのラベルの組み合わせを網羅的に調べ，会話参加者間での差の発生する部分を抽出ことで，会話における発現尤度の高い場面を自動的に抽出する手法を提案した．また，人と人とのインタラクションを実際に収録し，そこから生成したラベルに対し，この手法を適用し抽出された構造について検討した．

結果，指差しと発話の共起性，3 者対話における視線と発話の関係など，従来研究でも議論されてきたインタラクションのプロトコルを自動的に抽出することができた．また，発話のオーバーラップ時の遷移に関わる構造という複雑なインタラクションのプロトコルを自動発見することができた．

今後の展望としては，まず対象とする非言語行動を増やしていくことが考えられる．今回は発話・視線・指差しという 3 つの行動のみに注目して分析を行ったが，このほかにうなずきや相槌などといった会話における重要な行動を追加することで，より複雑な構造の発見が期待できる．

また，今回は抽出された構造の中でも頻度の多いものに関して議論を行ったため，ごく一般的なプロトコルを抽出するのみにとどまった．しかし，より多くのデータを基にして N-gram モデルによる構造化を行えば，中程度の頻度を持つ部分ではこれまで議論がされていなかった新たなプロトコルが見つかる可能性がある．

最後に，完全自動で作成されたラベルからの分析が課題として挙げられる．前述した取り組みを行っていくためには大量のインタラクションデータが不可欠だが，それらのデータに対しラベルの作成や修正を手作業で行うには限界がある．今後はより自動認識の精度を高める実験デザイン・認識手法を用いて，自動ラベリングによる構造化・構造抽出に取り組んでいきたい．

謝辞 本研究は，文部科学省科学研究費補助金「情報爆発時代に向けた新しいIT 基盤技術の研究」の一環で実施されました．また，本研究における実験協力や論文へのアドバイスなど多大な助力を頂いた，西田・角研究室の皆様に感謝いたします．

参考文献

[1]榎本 美香，伝 康晴：.3 人会話における参与役割の交替に関わる非言語的行動の分析，人工知能学会研究会資料，pp.25 – 30(2003)

[2]坊農 真弓：日本語会話における言語・非言語表現の動的構造に関する研 究，ひつじ書房(2008)

[3]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，情報処理学会論文誌，Vol.49，No.8，pp.945 – 949(2008)

[4]森田 友幸，平野 靖，角 康之，梶田 将司，間瀬 健二，萩田 紀博：マルチモーダルインタラクション記録からのパターン発見手法，情報処理学会論文誌，Vol.47，No.1，pp.121-130(2006)

[5]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，情報処理学会第 23 回ユビキタスコンピューティングシステム研究会発表会論文誌(2009)