

## マルチモーダルデータに基づいた多人数会話の構造理解

角 康 之<sup>†1</sup>

会話中に生ずる言語・非言語情報の構造を理解するために、非言語情報の辞書と文法の構築を目指している。その研究基盤として筆者のグループが構築を進めている、多人数会話のマルチモーダルデータの計測環境 IMADE ルームと、それに基づいて会話の構造分析を行うソフトウェア環境 *iCorpusStudio* を紹介する。非言語情報から会話参加の積極性を解釈する試みや、非言語情報発生の時系列パターンマイニングの試みを紹介する。

### Multimodal Data Analysis of Multiparty Conversation

YASUYUKI SUMI<sup>†1</sup>

This paper shows the IMADE (Interaction Measurement, Analysis, and Design Environment) project to build a recording and analyzing environment of human conversational interactions. The IMADE room is designed 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 show the notion of interaction corpus and *iCorpusStudio*, software environment for browsing and analyzing the interaction corpus. We also present a preliminary experiment 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<sup>1),2)</sup>を紹介する。まず、IMADE ルームと呼ばれる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。次に、インタラクションコーパスを分析するためのソフトウェア環境である *iCorpusStudio* を紹介する。最後に、IMADE ルーム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. IMADE ルーム：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てきた。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、これまでもいくつかなされてきた。その代表的なものである AMI<sup>3)</sup> 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<sup>4)</sup> 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<sup>5)</sup> 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（例えば 6）。

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

<sup>†1</sup> 京都大学 情報学研究科  
Graduate School of Informatics, Kyoto University

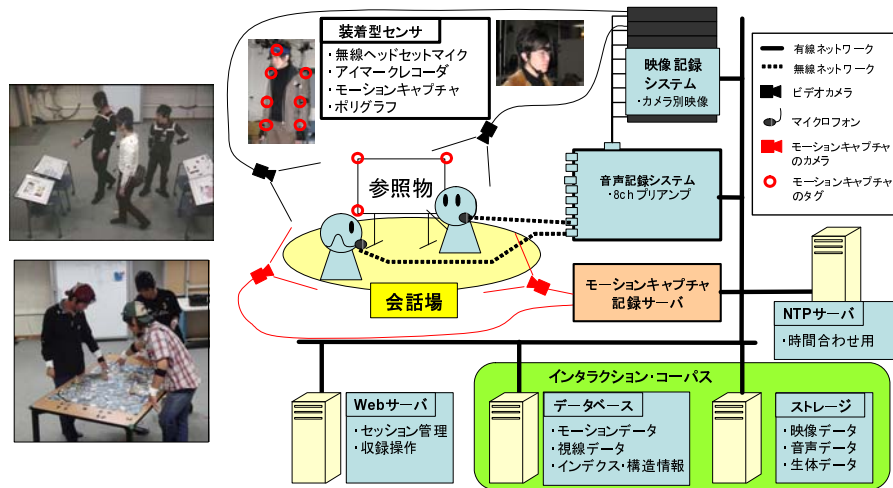


図 1 IMADE ルームの概念図  
 Fig. 1 Configuration of IMADE room

録するために、以下のような様々なセンサを設置している。

**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 台利用している。

**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、各人の会話内容を人ごとに分離して収録できる。

**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 システムを用いている。

**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 を利用している。

**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が必要である。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する必要がある。

**InTrigger の活用** 上記のような大量のデータの一次ストレージとして、また、インタラクションパターンの解析・発見の大規模計算に InTrigger プラットフォーム<sup>7)</sup> を活用している。

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ている。

### 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

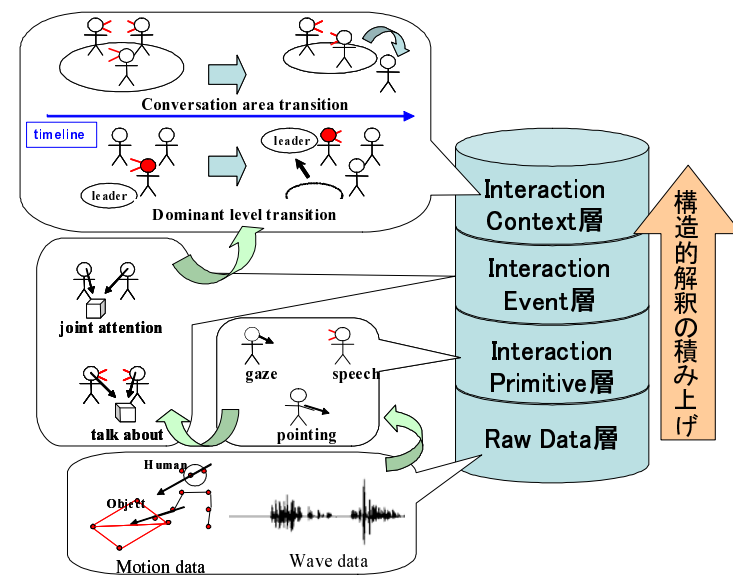


図 2 インタラクションの階層的解釈モデル  
 Fig. 2 Analysis model based on interaction corpus

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。

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. iCorpusStudio: マルチモーダルデータ分析のためのソフトウェア環境

筆者らは、会話的インタラクションに関するマルチモーダルなデータを閲覧・ラベリング・分析するための環境として、*iCorpusStudio* と呼ばれるソフトウェアを開発してきた。これまでも、会話分析のためのビデオや音声データのラベリングツールがいくつか存在していた（例えば、Anvil<sup>\*1</sup>や WaveSurfer<sup>\*2</sup>など）。それらに対して *iCorpusStudio* では、映像、音声に加えて、モーションデータ、視線データ、生体データなどのマルチモーダルデータを扱うとともに、複数視点（複数チャンネル）の映像、音声を同時に扱う必要がある。こういったセンサデバイスを利用するかは実験状況によって異なるため、*iCorpusStudio* 本体はデータの読み書き管理とラベリング記述のみを行うコンパクトなシステムにし、各種センサデータを読み込むためのソフトウェアモジュールは、プラグインとして必要に応じてインポートすることとした。また、ラベリングされたデータの分析を支援するために、従来のラベリングツールに比べて、ラベル間の演算やラベルに基づいたシーン検索の機能を強化している。

*iCorpusStudio* は大きく分類してデータ閲覧部と解釈演算部からなる。*iCorpusStudio* を用いることで、分析者は映像・音声・モーションデータなど、収録したデータを同期再生することができる。一方、発話の書き起こしや各モダリティの解釈を時間幅のあるラベルとして表現することができ、ラベル間の演算（AND 検索や OR 検索など）を行うことで、モダリティ間の時間構造解釈のための仮説を即座にプロトタイプし、検証することができる。

図3は *iCorpusStudio* の画面例である。ユーザは、必要に応じてビデオ映像や音声デー

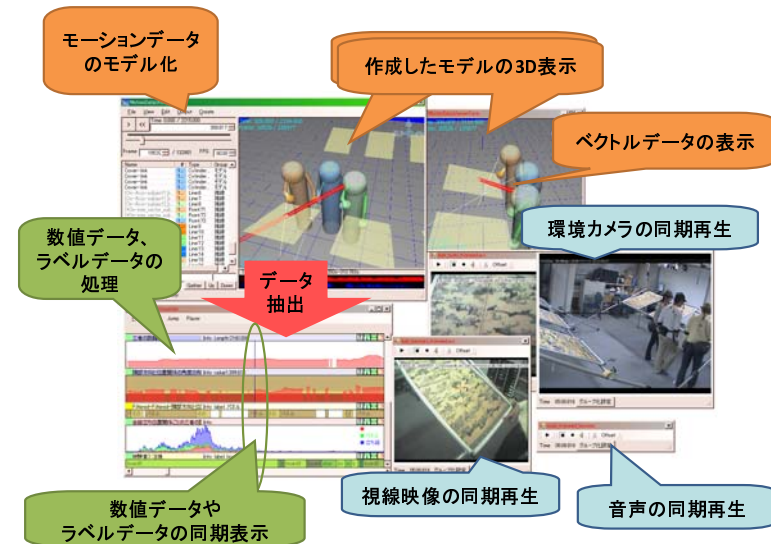


図3 *iCorpusStudio*: インタラクションコーパスの閲覧・ラベリング・分析環境  
Fig.3 Screenshot of *iCorpusStudio*

タを開いて同期させながら閲覧することができる。また、モーションキャプチャで取得された各マーカの3次元座標データから、会話参加者の身体モデルや参照物（ポスターなど）の形状をモデル化し、任意の角度から閲覧することができる。また、モーションデータのビューワの上では、視線や指さしなどのベクトルデータも表示できるので、複数人の共同注視や、指さしと視線の同期など、社会的インタラクションとして興味深い現象を直感的に確認することができる。

左下にあるウィンドウでは、音声波形データや発話書き起こしのラベルデータなどを同期しながら閲覧することができる。また、会話参加者間の立ち位置の距離や任意のベクトル間による角度など、数値データをグラフ表示することが可能である。つまり、分析者であるユーザは、例えば、会話参加者間の立ち位置の距離や角度の変化と話題の関係に注目した分析をしたり、頭部方向が視線をどの程度近似可能かをシーンの文脈に対応させて分析するといったことが、簡単な演算の組み合わせですぐに試すことができる。

上記の通り、*iCorpusStudio* は単なるラベリングツールではなく、研究者の仮説を試し、評価し、更に他の仮説を試すというサイクルを支援するラピッドプロトタイピングの環境で

\*1 <http://www.anvil-software.de/>

\*2 <http://www.speech.kth.se/wavesurfer/>

もある。それぞれ異なる観点を持つ様々な研究者が同一のデータを計測・分析・利用することを考えると、iCorpusStudio は彼らの共同作業を支援するグループウェア的役割を果たすことになる。今後、そのための機能、つまり、ラベルの登録・管理の共有支援、プロジェクト管理と作業進捗の共有促進、データ解釈ルールのマクロ化と再利用を促す機能を実現していきたい。

### 5. 視線を用いた指差しジェスチャ検出の精度向上

本稿の残りでは、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。ここでは会話の構造理解の仮説そのものの議論への深入りは避け、IMADE により得られたデータと iCorpusStudio を活用することで、データに基づいた会話分析研究の体系化・効率化が可能になったことを説明したい。

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、指差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、指差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かし、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、指差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差し方向に注目することで初めて指差し行為が成立する、social な現象であると考えた(図4)。つまり、会話参加者が指差し方向の対象物に視線を向ける行為が同期したものを指差しジェスチャとして認定することで、指差しジェスチャの検出精度が向上するのを確かめた<sup>8)</sup>。

以下の手順で指差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、目から指先にのびた方向を用いたものの2種類を準備した。次に、指差しベクトルの先に指差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、指差しジェスチャの候補を広めに抽出した。そして、それらの指差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差し先の対象物と一致しているかを確認した。

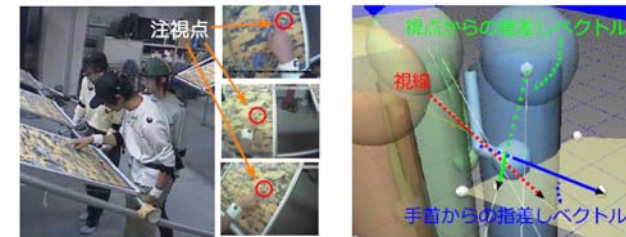


図4 指差しジェスチャの抽出  
Fig.4 Detection of pointing gesture

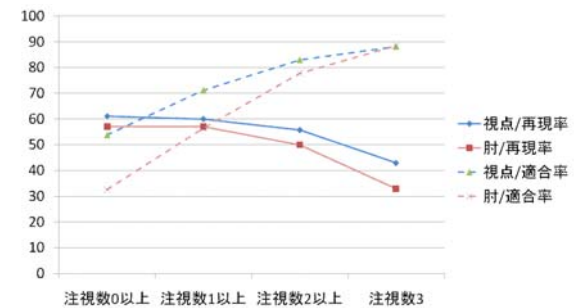


図5 視線獲得による指差しジェスチャ抽出の精度向上  
Fig.5 Detection accuracy improvement of pointing gesture based on hearer's visual attention

結果は図5のようになった。指差しジェスチャの検出精度を求めるために、実際の指差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。指差しベクトルの判定については、全体を通して、目から指先にのびたベクトルの方が精度が高いことが確認された。

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差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差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。

以上のことは、我々の直感に合っているものである。IMADE を使うことで、こういった



仮説を実際に取得したデータを目の前にし、iCorpusStudio の上で、3 次元モデルの生成、ベクトルの定義、複数のモダリティの時空間的共起性の演算、結果のグラフ化といった一連の作業を網羅的・効率的に実施することが可能になった。また、ここで一度作った仮説は、他のデータにも容易に適用できるので、分析研究の効率が上がった。

## 6. 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加に対する積極性を、非言語情報のみから推定することを試みた<sup>9)</sup>。

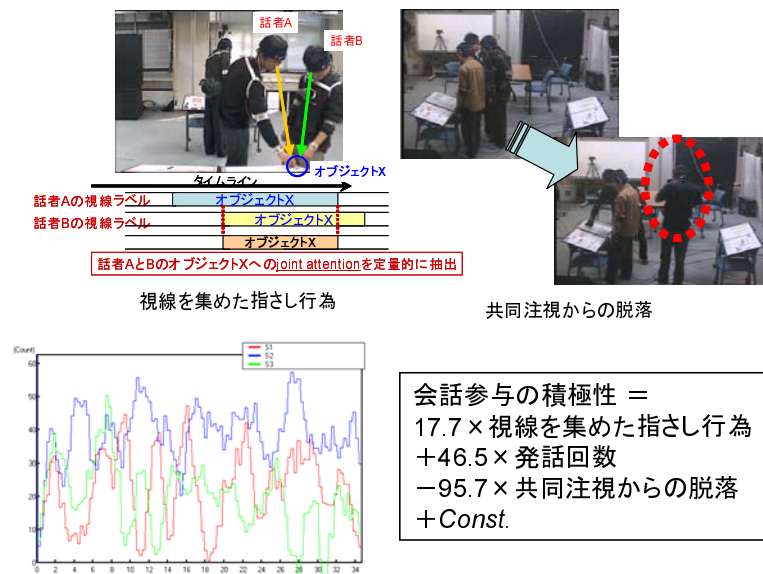


図 6 会話参加の積極性の数値化  
Fig. 6 Evaluation of positive attitude to conversation

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加積

極性を数値化することを試みた(図 6)。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。

その結果、従来研究(例えば 10))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。

## 7. 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<sup>11),12)</sup>。この手法は、発話の有無、指差し、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 7 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差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。

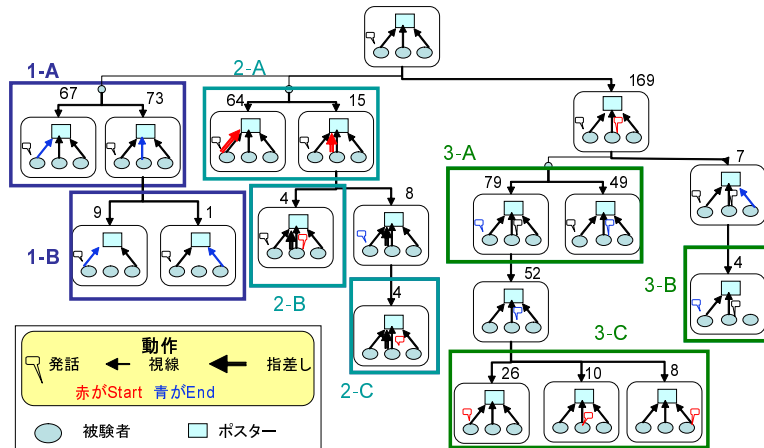


図 7 インタラクションマイニングによって抽出された会話構造の例

Fig. 7 Examples of interaction sequential patterns extracted from multiparty conversation

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。

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」の一環で実施された。IMADE ルームや iCorpusStudio の開発は、京都大学の西田・角研究室を中心に実施され、会話データの収録やラベリング・分析研究は、河原達也、高梨克也、坊農真弓の諸氏を始めとする多くの方と議論・協力しながら進めた。皆様に深く感謝する。

## 参考文献

- 1) 角 康之, 西田豊明, 坊農真弓, 来嶋宏幸: 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).
- 2) Sumi, Y., Yano, M. and Nishida, T.: Analysis environment of conversational structure with nonverbal multimodal data, *12th International Conference on Multimodal Interfaces and 7th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2010)*, ACM (2010).
- 3) Carletta, J., et al.: The AMI meeting corpus: A pre-announcement, *Second International Workshop on Machine Learning for Multimodal Interaction (MLMI 2005)*, LNCS3869, Springer, pp.28-39 (2006).
- 4) Waibel, A. and Stiefelwagen, R.(eds.): *Computers in the Human Interaction Loop*, Springer (2009).
- 5) Chen, L., et al.: VACE multimodal meeting corpus, *Second International Workshop on Machine Learning for Multimodal Interaction (MLMI 2005)*, LNCS3869, Springer, pp.40-51 (2006).
- 6) Kawahara, T., Setoguchi, H., Takanashi, K., Ishizuka, K. and Araki, S.: Multimodal recording, analysis and indexing of poster sessions, *INTERSPEECH-2008*, pp.1622-1625 (2008).
- 7) 田浦健次郎: InTrigger: オープンな情報処理・システム研究プラットフォーム, 情報処理, Vol.49, No.8, pp.939-944 (2008).
- 8) 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(コビキタスコンピューティングシステム), Vol.2009, No.22 (2009).
- 9) 中田篤志, 来嶋宏幸, 角 康之, 西田豊明: 移動・動作に関するセンサデータによる複数人数会話の解釈, 第 22 回人工知能学会全国大会 (2008).
- 10) Rienks, R. and Heylen, D.: Dominance detection in meetings using easily obtainable features, *Second International Workshop on Machine Learning for Multimodal Interaction (MLMI 2005)*, LNCS3869, Springer, pp.76-86 (2006).
- 11) 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(コビキタスコンピューティングシステム), Vol.2009, No.23 (2009).
- 12) 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).