非言語情報の出現パターンによる会話状況の特徴抽出

中田篤志[†]　角　康之[†]　西田豊明[†]

我々が会話を行う際には，発話に伴う視線の変化や指差しといった非言語行動に一定の構造が存在し，それは発話権取得の意図や話題に対する興味，理解度と関連があると考えられる．会話における非言語行動の構造を明らかにすることができれば，我々の会話状況を観察し，適切なタイミングで適切な相手に対して話しかけることができるエージェントやロボットを実現する一助となるであろう．本研究では，参照物を介した 3 人会話において会話参加者間で起こる一定のインタラクションパターンを抽出し，会話構造の特徴の顕在化を試みる．非言語情報の出現パターンを機械的に抽出するインタラクションマイニングと呼ばれる手法を適用し，三者対面時の発話と視線の関係や発話者と聞き手による指さしの使われ方に関するパターンが自動抽出できることを示す．

Sequential Pattern Mining of Nonverbal Behavior in Multiparty Conversations

Atsushi Nakata,[†] Yasuyuki Sumi[†] and Toyoaki Nishida[†]

When we talk, certain structures exist in nonverbal behaviors such as changes of gaze and pointing with speech, and it is considered that the structures are related with intention of speaking next, interest in topics and understanding about those. If we can explain the structures of nonverbal actions in conversations, it helps to realize agents and robots which can understand contexts in our conversations and talk to a proper person at a good timing. In this research, we tried to extract certain interaction patterns in conversations with three people and some objects and reveal characteristics of structure in conversations. We apply the method to extract the patterns of nonverbal actions automatically. We called the method Interaction Mining. Finally we extract automatically relations between speech and gaze in face-to-face conversations with three people and patterns about pointing by speaker or listener.

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文章構造が存在するのと同様に，一定の構造を持って行われていると考えられている．例えば「話している人物が発話を一時中断したとき，聞き手は聞いていることを示すためうなずきを返す」「重要な発話の前には，視線配分やジェスチャを行って他の人の興味を引く」といったものである．

このような会話におけるマルチモーダルな行動の構造を明らかにすることができれば，それ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．

会話的インタラクションの分析において従来使われている手法としては，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，会話の中から分析者が一部のエピソードを取り出し，その中における会話構造を詳細に議論する手法がある．しかし，前者の手法では様々な非言語行動の関連性を調べるのに多大な労力を必要とし，後者の手法では調べたエピソードの一般性などについて議論するのが困難である．

このようなことを踏まえ，本研究ではデータマイニングの手法をインタラクション分析に取り入れることによって，多数の種類における会話構造を数値的根拠に基づいて検証する手法を提案する．これをインタラクションマイニングと呼ぶ．

本論文では，まずインタラクションマイニングの具体的な手順について説明を行う．その後 3 人の被験者

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

による自由会話を収録し，発話・視線・指差しの 3 つの非言語行動を対象として本手法を適用する．この過程でインタラクションマイニングにおける分析者の留意点を例示すると共に，発話・指差し間の関係に関する研究[1]，発話・視線間の関係に関する研究[2] という 2 つの既存研究で議論されてきた会話構造を共に自動処理で抽出できることを示す．これによりインタラクションマイニングの有効性を示す．

2. 関連研究と本研究の位置づけ

1 でも述べたとおり，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数多く行われている[1)2)3)4)5)]．その際の手法として代表的なものとして 2 種類が挙げられ，分析者はこれらの片方，あるいは両方を用いて分析を行う場合が多い．

ひとつは仮説に関わるモダリティを数え上げて検証を行う手法であり，Chen らの論文[3] や榎本らの論文[2] などで用いられている．しかし会話構造には多くの種類の非言語行動が関わっており，それら全てに関して分析を行おうとすると非言語行動の組み合わせの数だけ数え上げを行う必要性がある．

もうひとつは，会話の中から分析者が一部のエピソードを取り出し，その中における会話構造を詳細に議論する手法である．例としては McNeill の論文[4] や細馬らの論文[5] が挙げられる．この手法の場合，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は難しい．

これらのことを踏まえ，本研究ではデータマイニングの手法を用いて多量のデータを構造化し分類することで，複数の非言語行動・複数の段階を経た会話構造について回数や発現尤度といった数値的根拠を基に会話構造を検証することを試みる．

次に，多数のカメラやセンサを用いて大量に会話データを収録し，インタラクションの分析を試みている研究として NIST[6]，AMI[7]，VACE[8] といったものがある．中でも，VACE ではジェスチャや音律・語彙といったデータをクラスタリングすることで，会話における文境界を識別することを試みている[9]．このようなデータを基にした自動解釈は我々の研究に通じるものであるが，これらの研究における自動解釈はあくまで文のセグメンテーションや非言語行動の検出にとどまっており，会話の意味構造といった抽象度の高いものについては従来の会話分析の手法が用いられている[4]．

また，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[10] 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．

最後に，本研究で提案しているインタラクションステートの概念，およびインタラクションマイニングの手法は福間らの研究[11] において最初に提案されたものである．本研究ではインタラクションマイニングの手法について整理すると共に，有効性の検証に利用するデータの拡充を行っている．

3. インタラクションマイニングの流れ

インタラクションマイニングは，

- 参加者の一人が指差しを行った
- 参加者の一人が発話した後，別の参加者がが発話者を見た

といった会話状態と遷移に着目し，遷移の出現回数の偏りから会話構造を検証する手法である．インタラクションマイニングは通常のデータマイニングと同様以下のような流れで行われる．

データ収集　人がインタラクションを行っている様子を様々なセンサ・ビデオカメラ・マイクなどで収録し，基礎データとして蓄積する．

データの前処理　収録された基礎データに分析者が前処理をほどこし，インタラクションステート列の形に変換する．

モデル化　変換済みのデータを木構造化し発現遷移に偏りがある部分を抽出する．

結果の検証　得られた結果を評価・解釈し，データマイニング全体の結果を検証する．

以下では，インタラクションマイニングにおけるそれぞれの手順について詳細に説明する．

3.1 データ収集

インタラクションマイニングでは，会話的インタラクションをビデオカメラやマイク，モーションキャプチャや視線計測装置といった様々なセンサによって収録し，これによって生成される自動ラベリングを用いてデータマイニングを行う．

3.2 データの前処理

本論文において提案するモデル化を実行するためには，データ収集によって得られた基礎データをインタラクションステートという状態の列に変換する必要がある．これは，以下の 3 ステップを経て行われる．

ラベリング　得られた基礎データから，被験者の非言語行動の開始点・終了点についてラベリングする

同一性の定義　同一として扱う被験者および非言語行動の対象物を定義する

ステート列への変換　ラベリング結果をインタラクションステートの列に変換する

以下で，これらについて詳細に述べる．

3.2.1　ラベリング

まず通常の会話分析と同様に，注目する非言語行動に対して開始点・終了点に関する情報を付加するラベリングという作業を行う．付加された情報をラベルと呼ぶ．

このとき，多くの会話分析では人が動画や音声を参照しながら手作業でラベリングを行うのが一般的である．しかし，インタラクションマイニングで有用な情報を取得するためには大量のデータを基に行う必要があり，このデータを全て手作業で作成するのは膨大な時間とコストを必要とする．

そのため，インタラクションマイニングではラベリングを自動処理によって行うことを想定している．機械的に得られる情報からラベリングを行う研究は，動画からの注視対象認識[12] など様々なものが研究されている．

3.2.2　同一性の定義

一般的に，会話分析の際は被験者の立場が対等の場合被験者を区別せず，非言語行動を合算して分析する場合がある．例えば 3 人が対面会話を行う様子を分析する際には，発話数や視線を送った回数は「A が発話した回数」「A が B を見た回数」と数えず，被験者全員を同一とみなして「全員の合計発話数」「ある被験者が他の被験者を見た回数」という形で数える場合が多い．一方でプレゼンテーションの様子を分析する場合には，被験者を「発表者」と「その他の聴衆」に分け，発表者の行動は別個に扱う場合もある．

このような会話分析の考え方を考慮し，インタラクションマイニングを行う際には「被験者を同一とみなすかどうか」をあらかじめ定義しておく．これは，後述するステート列への変換に必要となる．

3.2.3　インタラクションステート列への変換

3.2.1 で作成したラベルから，インタラクションステートの列を作成しマイニングのための処理済データとする．

インタラクションステートとは，例えば

- 全員でひとつのボードを見ていて，一人がボードを指差している
- 3 人中 2 人が顔を向かい合わせており，もう一人はそのうちの片方を見ている

といった会話の一場面を表現したものである．インタラクションステートは，分析対象とした非言語行動それぞれについて，その発生の有無（発話をしているか，していないかなど）およびその対象を全て記述したものになる．

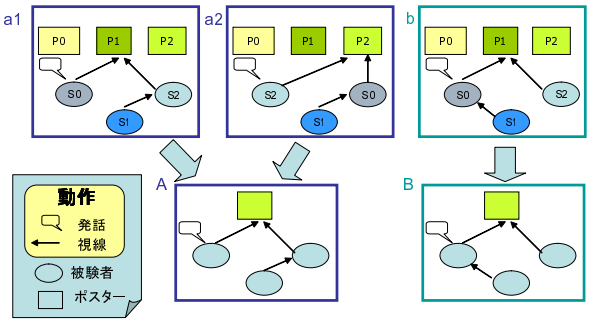

図 1　インタラクションステートの例示
Fig. 1　The samples of Interaction State

ここで,「被験者・対象物の入れ替えによるステートの同一視」について説明する．3.2.2 で述べたような会話分析の考え方を取り入れるため，インタラクションステートでは以下の条件を満たすとき，2 つの会話状態を同じインタラクションステートとして扱う．

(1)　被験者または非言語行動の対象物である a，b が，同一とみなすと分析者によって定義されている

(2)　a,b を入れ替えたときに，各被験者が取っている非言語行動の発生の有無および対象を同一にできる

インタラクションステートの同一視について，例を**図 1** に挙げる．ここでは分析対象は視線・発話の 2 つのモダリティとする．また，全ての被験者，および視線の対象となるポスターは同一とみなされているとする．

ここで，ステート a0 とステート a1 は，被験者 S0 と S2 を入れ替え，ポスター P1 と P2 を入れ替えることで同じ状態となるため，同じインタラクションステート A として扱う．一方，ステート b はステート a0 と S0，S2 の状態は同じであるものの，S1 が発話者である S0 を見ているか，非発話者である S2 を見ているかという点が異なるため，異なるインタラクションステート B として扱う．

なお，図 1 の例はあくまで全員が入れ替え可能である場合の例である．仮にこれがポスター発表の場であり，3.2.2 の段階で分析者が「S0 は発表者であるため S1,S2 と同一とはみなさない」とした場合，a0 と a1 は別のステートとして扱われることになる．

以上のようなインタラクションステートの考え方を基に，分析対象となっている非言語行動のラベルから

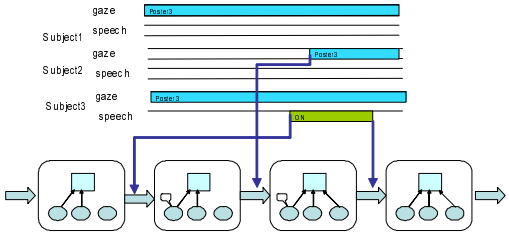

図 2 インタラクションステート列への変換例
Fig. 2 Samples of Interaction State

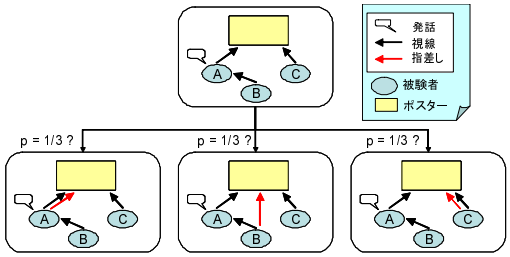

図 3 χ^2 検定による構造抽出
Fig. 3 Extracting structure by χ^2test

インタラクションステートの列を作成する．インタラクションステートは非言語行動の状態を全て記述したものであるため，いずれかの非言語行動の発生の有無・対象が変わるたびにインタラクションステートが変化する．この変換の例が**図 2** である．

3.3 モデル化：木構造化と χ^2 検定による構造抽出

ここではインタラクションステートの列を基に，会話状態遷移における特徴的な構造を見つけるためのモデル化手法について説明する．

まず前処理で得られたインタラクションステート列を一定の深さで切り分け，木構造によって構造化する．これによって，それぞれのインタラクションステートごとに木構造およびそれぞれの構造の出現回数を得ることができる．その後,「インタラクションステートの遷移において，被験者間の偏りは存在しない」という帰無仮説を基に χ^2 検定を行い，仮説が棄却される木構造を抽出する．これによって抽出された木構造が，会話的インタラクションにおける特徴的な構造となる．

具体例を**図 3** に示す．最初の状態ではどの被験者も指差しを行っていない．この図では，この状態からインタラクションステートが変化した場合のうち，変化がポスターへの指差しによるものであった場合のみを示している．

ここで前記の帰無仮説を採用すると，どの被験者も一様に行動の開始・終了を決定することになる．すると A，B，C の誰もが指差しを行っていないため，次の行動として指差しを開始する可能性があり，その確率は 1/3 で等しい．すなわち，図 3 の 2 段目のインタラクションステートへの遷移が同じ回数ずつ起きるということになる．しかし，インタラクションの持つ会話構造を考えた場合，ポスターを見ていない B が指差しを行うことは非常に考えづらい．また，非発話者である C が指差しを行うという遷移の発生回数も A が行う遷移よりも少ないのではないかと考えられる．

この抽出プロセスでは，このような偏りを χ^2 検定で検出する．これにより，変換プロセスで生成された非常に大きなツリーから，インタラクションの構造によるものであると考えられる部分を検出し，抽出することができる．

このモデル化において，何段階の木構造を作るかは重要な問題である．段数の深い木構造を作った場合，複雑な会話構造を検証することができる一方で，データ量が十分でない場合は階層の深い部分では発生回数が非常に少ないものとなるため，信頼性が低く結果の検証の手間を増やすだけとなる可能性もある．したがって，何段階の木構造を作るかについては最下段での発生回数を参照しつつ調整することが重要である．

3.4 結果の検証

3.3 で得られる特徴的構造は大量である．また通常のデータマイニングと同様に，ラベリングやステートの定義といった前処理によって得られる結果は大きく変化し，結果の中にはラベリングにおけるノイズが原因であるものなど会話プロトコルに起因する偏り以外も含まれる可能性がある．

そのため，分析者は得られた特徴的構造に関して，実験データや従来研究などと照らし合わせて検証を行う必要性がある．

4. 3 人自由会話へのインタラクションマイニング実行

本論文では，3 人の被験者による自由会話に対して提案したインタラクションマイニングを適用し，得られる結果を従来研究と比較することでインタラクションマイニングの有用性を示す．本章ではデータの収録からモデル化までに分析者が行った手順を説明し，次章で得られた結果の検証を行う．

4.1 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説明する．

まず，被験者の数は 3 人とした．理由は，発話交替における立場の変化，指差しに対する注視・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)] ためである．

前処理の対象とするデータとしては，音声データ・

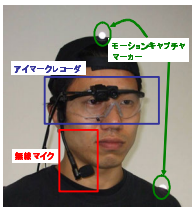

図 4 被験者
Fig. 4 Subject

図 5 収録環境
Fig. 5 Experimentation environment

3 次元位置情報・視線方向を採用した．これらを計測するため，被験者には頭・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 4** に示す．またインタラクションマイニングの結果を検証するため，実験環境に多数のカメラを設置し，マイクによる音声データと合わせて実験の様子を記録した．

会話環境としては，被験者が**図 5** 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．

以上のような条件で 3 回の実験を行い，インタラクションマイニングの基礎データとした．

4.2 対象とする非言語行動の選択

これまでの会話分析において，発話と視線，および発話とジェスチャの関連性は非常に強いものとして注目されてきた．例えば榎本らの研究[2] では，発話交代における視線の役割を分析するため，発話の終わり付近とそのときの視線対象について数え上げを行っている．また，坊農の研究[1] では発話とジェスチャのタイミングについて関連性が議論されている．

以上の理由から，今回のインタラクションマイニングでは，分析の対象とする情報として発話の有無，視線の対象，さらにジェスチャの中でも本実験の環境で頻繁に行われる指差しの 3 種類を選択した．これらの会話構造をインタラクションマイニングで検証し，既存研究で議論されている関連性が自動処理で容易に抽出できることから本手法の有効性を示す．

4.3 ラベリング

3.2.1 で説明したとおり，インタラクションマイニングではセンサから自動処理によって得られたラベルを用いて分析を行うことを想定している．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．

4.3.1 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一人分の視線ラベルに関しては，アイマークレコーダの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．

4.3.2 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．

4.3.3 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．

4.4 被験者・対象物における同一性の定義

今回収録したデータは 3 人による自由会話であり，被験者の役割に差異は存在しない．また，ポスターに関しても明らかに注目されやすいものなどは想定していない．

以上のような前提から，今回のインタラクションステートにおいては被験者・ポスターを共に入れ替え可能であると定めた．

4.5 モデル化における値の設定

今回のインタラクションマイニングでは，木構造の深さを「4 段階」,「2 段階」の 2 種類に設定してモデル化を行うと共に，得られた結果のうち以下のような一部分のみに着目して結果の検証を行った．

(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 段階の深さまで検証する

(2) 2 段階の深さ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．

また，いずれにおいても χ^2 検定の有意水準は 5% とした．

5. 特徴的なインタラクション構造の抽出結果

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 6 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 6 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．

以下は得られた構造を 3 つの着眼点で検証する．

5.1.1 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．仮に一定の会話構造が存在しない場合，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，ここには一定の会話構造が存在していると考えられる．

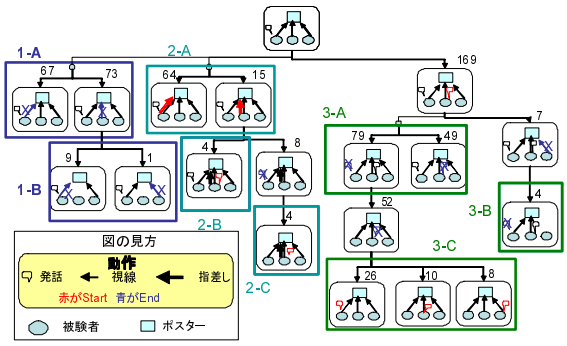

図 6 最頻インタラクションステートを基点とした構造
Fig. 6 The structure based on the most frequent interaction state

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，他の被験者やポスターに視線を向けて会話をコントロールしているのではないかと考えるられる．このような現象は従来研究でも論じられている[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．

5.1.2 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言える．

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)．

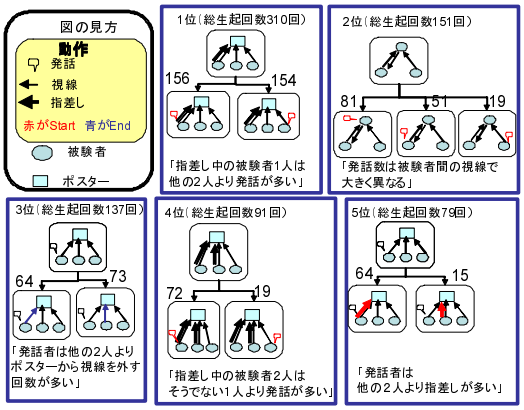

図 7 1 段階遷移における発現尤度の高い構造
Fig. 7 Structure with higher presentation likelihood in one stage transition

これらは生起回数が少ないので必ずしも一定の会話構造であるとは言い難いが，これが会話構造である場合は「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．

5.1.3 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．一定の会話構造がないとした場合の発話終了確率はどちらも 1/2 であるから，ここには一定の会話構造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，この構造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 7 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．

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」ということが分かる．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」という会話構造が読み取れる．この構造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[1]．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 (1)(2) 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[2]，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」ということが分かる．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，マルチモーダルな多人数会話データから特徴的構造を効率的に抽出するための手法として，インタラクションマイニングを提案した．また提案手法の有用性を示すため，実際に 3 人の被験者による自由会話に対してインタラクションマイニングの手法を適用し，提案手法において分析者が留意すべき点について具体的に述べると共に，得られたインタラクションの特徴的構造について検証を行った．

結果，指差しと発話の共起性，3 者会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会話構造を自動処理によって抽出することができた．また，発話のオーバーラップに関わ

る会話構造を検証した．これらのような，3 種類の非言語行動に関する様々な会話構造を数値的根拠を持って得られたことは，インタラクションマイニングという手法の有用性を示唆していると考えている．

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，既存手法では困難な多くの非言語行動が関わる構造の検証が期待できる．

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，ごく一般的な会話構造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マイニングを行うことで，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会話構造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，本稿で述べた内容の初期の研究開発に携わった福間良平氏（現在，奈良先端科学技術大学院大学所属）に感謝いたします．最後に，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) 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).
2) 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，　Vol.SIG-SLUD-A301, pp.25 – 30 (2003).
3) Chen, L., Harper, M., Franklin, A., R.Rose, T., Kimbara, I., Huang, Z. and Quek, F.: A Multimodal Analysis of Floor Control in Meetings, *Machine Learning for Multimodal Interaction*, Vol.3869, pp.36–49 (2006).
4) McNeill, D.: Gesture, gaze, and ground, *Machine Learning for Multimodal Interaction*, pp. 1–14 (2005).
5) 細馬宏通，石津香菜，繁松麻衣子，中村智代，矢野雅人：身体を示し合う会話－自分の身体で相手の身体を語ること－，社会言語科学会第 14 回大会，pp.67 – 70 (2004).
6) Michel, M., Ajot, J. and Fiscus, J. G.: The NIST Meeting Room Corpus 2 Phase 1, *Machine Learning for Multimodal Interaction*, pp. 13–23 (2006).
7) Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D. and Wellner, P.: The AMI Meeting Corpus: A Pre-announcement, *Machine Learning for Multimodal Interaction*, pp.28–39 (2005).
8) Chen, L., Rose, R., Qiao, Y., Kimbara, I., Parrill, F., Welji, H., Han, T. X., Tu, J., Huang, Z., Harper, M. P., Quek, F. K. H., Xiong, Y., McNeill, D., Tuttle, R. and Huang, T. S.: VACE Multimodal Meeting Corpus, *Machine Learning for Multimodal Interaction*, pp.40–51 (2005).
9) Chen, L., Liu, Y., Harper, M. and Shriberg, E.: Multimodal model integration for sentence unit detection, *Proc. of Int. Conf. on Multimodal Interface (ICMI), University Park, PA* (2004).
10) 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).
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 (ユビキタスコンピューティングシステム)，No.2009-UBI-23 (2009).
12) Otsuka, K., Sawada, H. and Yamato, J.: Automatic inference of cross-modal nonverbal interactions in multiparty conversations:" who responds to whom, when, and how?" from gaze, head gestures, and utterances, *Proceedings of the 9th international conference on Multimodal interfaces*, ACM New York, NY, USA, pp.255–262 (2007).