Vol.2009-UBI-23 No.23
2009/7/17

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

福間良平[†1,*1] 角康之[†1] 西田豊明[†1]

近年のコンピュータの普及と高性能化に伴い，様々な日常生活の支援システムが開発されている．しかし，真に有用な支援システムとなるためには，人と人とのコミュニケーションを理解する能力が必要である．ここで，人と人とのコミュニケーションを考えると，その要素である発話，視線移動などにおいて様々な構造が認められる．これらは従来研究では，様々な観点からインタラクションの分析を試みてきた．しかし，主にそれらの分析は，分析者が興味のあるシーンについて切り出した後に順々に分析を行う，仮説検証がメインであった．また，同時にそのシーンの状態を示すアノテーションの付加を人手で行うため，長時間のデータの分析は困難であった．それら故に，これらの研究では，全体を見渡した統計的な議論は困難であった．

そこで，機械的に生成可能と考えられるデータのみを用い，記録したインタラクションのデータ全体からボトムアップに分析を行う手法の 1 つを本研究では提案する．提案手法では，アノテーション群をインタラクションステートの変化を示す N-gram モデルに乗っ取ったツリーに変換し，その後インタラクションの構造により生成される特徴を抽出する．

Time-Structure Discovery from Multi-modal data of Human-to-Human Interaction

Ryohei Fukuma,[†1,*1] Yasuyuki Sumi[†1] and Toyoaki Nishida[†1]

Recent popularisation and enhancement of computers have been promoting development of daily-life-support-system in many domains. However, a truly valuable support-system requires an ability to understand human-to-human communications. These communications have variety of structures with speech, gaze-movement, etc. , which are elements of communication. Although many researches have been performed to human-to-human interactions, these researches use mostly top-down approach, which analysts clip scenes, and analyse one by one. Moreover, in this approach, analysis of long-time-recording are tough since all scenes are annotated manually. Therefore, making statistical discussion are difficult in it.

Here, we propose a new approach of bottom-up analysis with automatically annotated interaction data in this reserch. In the approach, we convert all annotations into N-gram model that represent changes of interaction-state, and then extract features brought by the interaction structures.

1. はじめに

人と人とのインタラクションには，暗黙の決まりごとが存在する．例えば，話者交代には視線が関係していることが示されている[1)]．このような，時間構造はランダムに発生するものではなく，その発生順序やタイミングに一定のパターンが存在する．そのため，インタラクションにおける暗黙の決まりごと，すなわちプロトコルを説明するにあたって，時間構造を明らかにすることは重要である．

ここで，これまでの人と人とのインタラクションの分析の試みを振り返ると，典型的な手法は，カメラとマイクを用いインタラクションを記録し，手作業でアノテーションを施すことによって分析を行うものであった．また，様々なセンサの発展により，マルチモーダルなデータをインタラクションの分析に生かそうとする試みも行われてきた．例えば，NIST[2)] では，マイクアレイや精密にキャリブレーションされたカメラを用い，ミーティングのキャプチャを行っている．また，AMI[3)] ではそれらに加え，環境に設置されたホワイトボードへの書き込みも記録している．しかし，これらの研究での分析は発話内容の書き起こしを中心としたものであり，センサから記録した情報は補助的に使用されるだけである．一方，VACE[4)] では，発話内容と，体の向き・視線などといった非言語情報を同列に扱いマルチモーダルでのインタラクションの分析を行っている．しかし，VACE でキャプチャするミーティングは被験者が着席した状態であるため，身体行動は限定的である．また，その話題は事前に定められたものに限定されている．

そこで，我々は被験者が自由に移動できる環境下で，かつ話題に関しても比較的自由度の高い会話場で行われたインタラクションを対象として分析を行うこととした（**図 1**）．また，

†1 京都大学
Kyoto University
*1 現在，奈良先端科学技術大学院大学
Presently with Nara Institute of Science and Technology

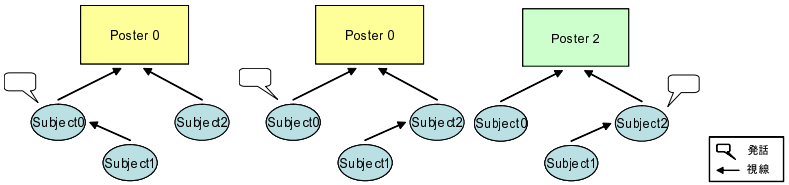
Vol.2009-UBI-23 No.23
2009/7/17

図 1 非拘束的環境でのインタラクション収録
Fig. 1 Interaction recordings under low-restrictive environmet

その分析にあたりセンサから得られるデータにより可能な限り自動で分析することとした．センサから得られるデータのみから分析することで，人の主観に左右されず安定した分析ができると期待できる．また，記録したデータを自動で処理することで，従来研究では困難だった記録全体の分析が可能になり，統計的な扱いが可能になると期待できる．

本研究と同様に，センサから得られたデータを自動で分析を行ったものとして，森田ら[5]の研究が挙げられる．この研究では，ウェアラブルセンサーにより自動付加されたアノテーションからのパターン抽出を試みている．しかし，ウェアラブルセンサーを用いることで注視のモダリティを近似しており，その精度は十分であるとはいえず，また，インタラクションの構造の時間変化については言及していなかった．そこで，本研究ではインタラクションの構造の時間変化に着目するため，自然言語処理で頻繁に用いられる，N-gram[6] を用いた．N-gram を用いた研究としては，長尾ら[7] や Cavnarw[8] らによるものが挙げられる．長尾らは日本語列からの節や単語の抽出を行っており，Cavnarw は N-gram の出力同士の分布の違いから文書分類を行っている．本研究では，これらの研究を参考に，インタラクションを N-gram モデルにより記述することとした．

本論文では，複数のセンサデータから生成したアノテーションから，インタラクションの構造により発生したアノテーションの頻度の差を手がかりにして，インタラクションの構造の抽出を行う手法を提案する．また，その評価として 3 人の被験者からなるポスター会話を記録し，アノテーションの半自動生成と，構造の抽出を行ったうえで，従来研究での知見で抽出された構造の検討を行い，従来研究の枠組みの中で説明可能であることを示す．これにより，本論文で提案する手法の有効性を明らかにする．



(a) ステート 1　(b) ステート 2　(c) ステート 3

図 2 インタラクションのステート
Fig. 2 State of interaction

2. 提 案 手 法

提案手法は変換・抽出の 2 つのプロセスからなる．まず変換プロセスでは，時系列データであるアノテーションを N-gram モデルを用いたツリー構造へ変換を行う．次に抽出プロセスでは，インタラクションの構造が存在しないとしたモデルを考え，このモデルとの差異を用いた抽出を行う．本章では，それぞれのプロセスを詳細に明らかにする．

2.1 変換プロセス

変換プロセスでは，N-gram モデルを用いてアノテーションをツリー構造へと変換する．この時ツリーのルートは，ある瞬間のアノテーションの組から決まるインタラクションステートとする．次に，同じインタラクションステートからの時系列変化を N-gram モデルにて記述し，ツリーとする．これは，アノテーションの変化点にのみ着目することで実行できる．

インタラクションステートの例を**図 2** に示す．この図においては視線を矢印で，発話を吹き出しで表記しているが，これらの場合においては被験者の入れ替えを考慮すると，ステート 1 とステート 3 が同じインタラクションステートであると考えられる．

特に，この段階において被験者の入れ替えや対象物の入れ替えを行うため，今後入れ替え前についてはそれぞれを Subject0，Poster0 のように数字を付加して表記し，入れ替え後については SubjectA，PoseterA のようにアルファベットを付加して表記する．

2.2 抽出プロセス

抽出プロセスでは，まず人と人とのインタラクションに構造は存在しないと仮定したモデ

ルを考える．このモデルでは，会話の参加者は好きなタイミングで任意の行動を取ることができる．このため，インタラクションの時系列変化に偏りが生じない．

ここで，先に示した図 2 のステート 1 を例とする．この場合において，図に示された人が任意のタイミングに任意の行動を取るとすると，たとえば，Subject1 と Subject2 が発話開始する確率は，それぞれ 1/2 であり等しくなる．しかし，現実にはおそらく Subject2 が相槌という形で発話する確率のほうが高くなると考えられる．

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．

3. 評価実験

本研究では，提案手法が有効であることを確認するために，評価実験を行い人と人とのインタラクションのマルチモーダルデータを収録した．このデータ収録は京都大学に設置したセンサルーム（11.25m × 7.4m）で行った．また，センサにはモーションキャプチャ・アイマークレコーダ・環境カメラ・ヘッドウォーンマイクを使用した．本章では，実験の詳細と，センサの取り付け等について示す．

3.1 実験設定

本実験での被験者は 3 人である．これは，提案手法では被験者が同じ会話場にいることを前提としており，一方で 2 人では十分に複雑なインタラクションの構造が発生しないと考えたためである．次に，使用するアノテーションを視線・発話・指差しとした．これらは，後述するように自動で生成することが比較的容易に可能であると考えられるためである．最後に，被験者に与えるタスクであるが，指差しが計測しやすいものとして，環境に配置したポスターについて話を行ってもらうポスター会話とした．さらに，ポスターを洛中洛外図屏風を分割し印刷したものとした．これは室町時代に書かれた一種の京都の地図と考えられるため，比較的会話を自由に行えると考えられ，また，現在も京都に残る名所等も書かれているため指差しが多く発生することも期待できる．なお，ポスターに描かれた場所と実際の京都市の配置との対応関係が容易に分かるように，それらの東西南北を一致させた．

3.2 実験環境の配置

環境には人の動き等を計測するモーションキャプチャと，データの確認用の環境カメラを配置した．その配置を**図 3** に示す．赤線で囲まれたエリアがデータを収録するエリアであり，東西 4.3m，南北 3.5m である．エリア内には東側，西側ともに 3 枚のポスターを配置

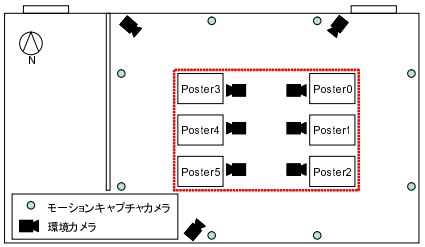


図 3 実験環境の配置
Fig. 3 Arrangement of experiment environment

してあり，それぞれのポスターのサイズは 1034mm x 731mm である．ポスターはモーションキャプチャの記録を阻害しないように斜めに配置する一方，ポスターの下端を十分に高くすることにより無意識のうちに被験者が触れないようにした．これは，アノテーションの自動生成を容易にするための，予備実験から得られた知見である．また，ポスターの 4 隅にはその座標を計測するためのモーションキャプチャのマーカーを取り付けた．

3.3 被験者へのセンサ取り付け

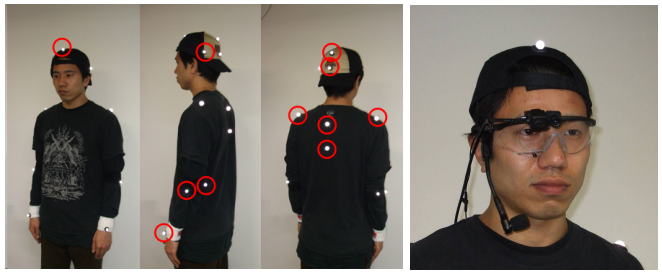
被験者には視線を計測するためのアイマークレコーダ，発話を記録するためのヘッドウォーンマイク，頭部の位置や指差し等を計測するためのモーションキャプチャのマーカーを取り付けた．取り付け方を**図 4** に示す．

4. アノテーション生成

本論文では，アノテーションからのインタラクションの構造の自動抽出を行う手法を提案した．しかし，人手で大量のアノテーションを付加することは困難であり，またセンサによってはアノテーションを付加する人の主観が入るものとなる．これは，提案手法とは相容れないものであり，センサデータへのアノテーションの自動生成は避けて通れないものである．一方，評価実験で得られたセンサデータについてアノテーションの完全自動生成は困難なものであった．そこで，実験データの一部に手動でアノテーションを付加し，そのアノテーションを用いて必要なパラメータを決定したうえで，実験データ全体に対しアノテーションを自動で生成するという手段を用いた．本章では，それぞれのセンサごとにアノテーションの生成の方法を示す．

視　線

視線アノテーション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



(a) モーションキャプチャマーカー　　(b) アイマークレコーダ・ワイヤレスマイク

図 4 被験者へのセンサー類の取り付け
Fig. 4 Arrangement of sensors to subject

チャから視線ベクトルを計算した．次に，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．その後，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，適当な時間スレッショルドを用い補間を行った．

発　　話

発話アノテーションの生成は，まず各被験者が装着したヘッドウォーンマイクから得られた音声波形を 50msec ごとに分割し，FFT（Fast Fourier transform）を用いパワーを計算したうえで適当なスレッショルドを用い 2 値化を行った．次に，隣接するアノテーション同士を適当な時間スレッショルドを用い補間を行った．これは，発話中の声の大きさのムラにより，発話中にも関わらず発話していないと判定されることがあるためである．最後に，短いアノテーションを適当な時間スレッショルドを用い削除した．これは，他の被験者の声が回りこんだ場合には，特に声の大きな部分しか拾われず，長さの短いアノテーションとして現れる傾向があったためである．

指　差　し

指差しアノテーションの生成には，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．最後に，衝突判定から得られたアノテーションを適当な時間スレッショルドを用い補間し，その後，短いアノテーションについては削除を行った．この補間は，センサからデータが取得できない場合と，指差しの仕方によってはアノテーションが生成されない場合があるためである．一方，削除は，ポスター前で腕の移動を伴ったジェスチャーを行ったときに発生する誤検出を削除するためである．

表 1 アイマークレコーダと頭部マーカーの取得率
Table 1 Capture rate of eyemark-recorders and motion capture markers on heads

	Subject0	Subject1	Subject2
アイマークレコーダ	90.64%	90.75%	69.24%
頭部マーカー	99.56%	99.90%	99.99%
アイマークレコーダ&頭部マーカー	90.23%	90.66%	69.23%

表 2 腕のモーションキャプチャのマーカーの取得率
Table 2 Capture rate of motion capture markers on arms

	Subject0	Subject1	Subject2
左肘外	96.62%	91.83%	91.39%
左手首	90.14%	79.20%	84.55%
右肘外	95.98%	88.02%	76.77%
右手首	86.33%	68.13%	90.10%

5. 収 録 結 果

評価実験は 2008 年 11 月 14 日（金曜日）に 2 セッション行い，それぞれのセッションの長さは約 22 分と約 35 分であった．1 つ目のセッションは Subject2 のアイマークレコーダの記録精度が悪かったため，本論文では 2 つ目のセッションを分析に用いた．このセッションでの使用したデータの時間に対する取得率について以下に示す．

まず，視線アノテーションの付加にはアイマークレコーダで視線が取得できており，同時に頭部の 4 点のモーションキャプチャのマーカーのうち 3 点以上が取得できている必要がある．時間に対する取得率を**表 1** に示す．

次に，指差しアノテーションの計算に使用する，腕に付けたモーションキャプチャのマーカーの時間に対する取得率を**表 2** に示す．

6. アノテーション生成の結果

それぞれの，センサデータから半自動でアノテーションを生成した．理想的には，すべて

Vol.2009-UBI-23 No.23
2009/7/17

表 3 生成されたアノテーション数
Table 3 Number of generated annotations

	Subject0	Subject1	Subject2
発話アノテーション	612	577	336
視線アノテーション（ポスター）	479	484	471
視線アノテーション（被験者）	858	344	539
指差しアノテーション	36	23	22

の被験者のアノテーションの誤検出，見落としが同じ頻度で発生するため，提案手法にさほど影響しない．しかし，実際はセンサの取り付けやキャリブレーション，その人の動きや話し方といったものに，アノテーションの誤検出と見落としが依存し，さらには実験被験者の組み合わせによってその会話場に参加する立場に偏りが出ることも相まって，提案手法に影響をもたらす．そこで，本章では生成されたアノテーションの傾向と，原因を示し，7 で述べる抽出結果を解釈するにあたっての準備とする．ここで，生成されたアノテーションの数を，被験者ごとに**表 3** に示す．

発話アノテーション

Subject2 の発話アノテーションが少なくなっているが，これは音声と生成された発話アノテーションとを比較した結果，発話の際の声が極めて小さく，他の被験者の声が回り込んでいためにパワーのスレッショルドを高めに設定したためであった．

視線アノテーション

Subject0 の他の被験者を対象にした視線アノテーションが 858 と大きくなっているが，これは Subject0 の後頭部のマーカーが，マーカー同士でのオクリュージョンにより計測された座標がずれ，何度も間違ったアノテーションと正常なアノテーションを交互に生成したからである．また，Subject2 の他の被験者を対象にした視線アノテーションも 539 あるが，これは Subject2 のアイマークレコーダが下向き気味にキャリブレーションされていたために，他の被験者を見るときに断片的な短いアノテーションを何度も生成したためである．

指差しアノテーション

指差しアノテーションについては，誤検出や見落としは多少あったものの，被験者ごとの大きな偏りはなかった．

7. 抽 出 結 果

本章では評価実験で得られたアノテーションを提案手法を用いて分析し，その結果を示す．

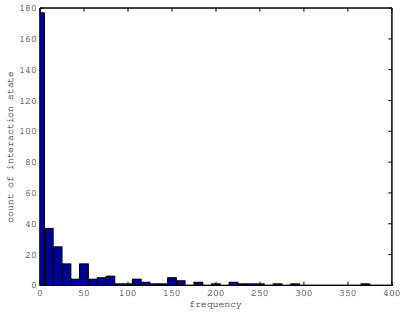


図 5 インタラクションステートの出現回数
Fig. 5 Frequency of interaction-state

まず，変換プロセス後には 315 種類のインタラクションステートをルートとした N-gram ツリーが生成された．これらのインタラクションステートの出現回数のヒストグラムを**図 5** に示す．この図より明らかに，数回しか出現しないインタラクションステートが大半を占めている．これらの低頻度のインタラクションステートは，被験者のうち 1 人以上が指差しをしている場合が多かった．これは，インタラクションステートがそれぞれの被験者の状態の組み合わせで決まる一方，被験者が指差しをしている時間が短いためであると考えられる．また，これらの 315 種類のインタラクションステートのうち，有意水準 5%の χ^2 検定で有意な差が出たとして抽出された時系列変化を持つインタラクションステートは 59 種類であった．

7.1 抽 出 例

抽出したインタラクションの時間変化の 1 例を**図 6** に示す．これは，最も多く観測されたインタラクションステートからの時間変化を示したものである．赤い枠で囲まれた部分が，モデルに対して有意水準 5%の χ^2 検定で検出された部分である．なお，このツリーのルートであるインタラクションステートは 366 回観測された．本節では，この最も多く観測されたインタラクションステートから開始する時間構造を例として挙げ，抽出プロセスの手続きをより具体的に示す．

例えば，図 6 内の c の状況は，開始時のインタラクションステートから SubjectA が発話を終了した後に SubjectA・SubjectB・SubjectC が次に話し出す場合である．この場合，モデルではそれぞれ 1/3 の確率であると考えられるが，観測された回数はそれぞれ 12 回・

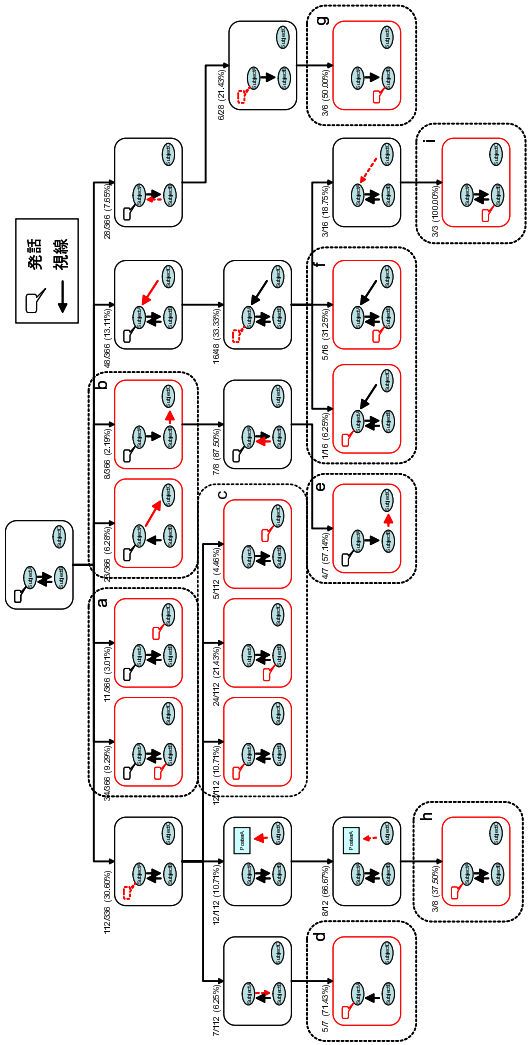


図 6 抽出されたインタラクションの時間変化例
Fig. 6 Example of extracted time varying interaction structure

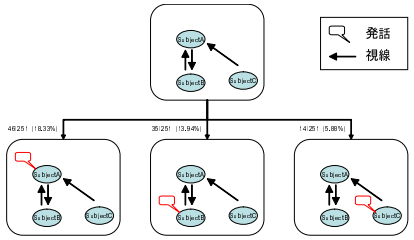


図 7 立場の違いによる発話確率の違い A
Fig. 7 Difference of speech frequency from interaction position: A

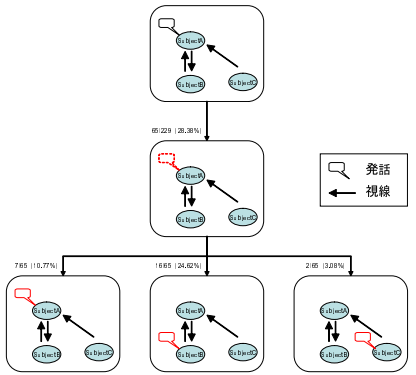


図 8 立場の違いによる発話確率の違い B
Fig. 8 Difference of speech frequency from interaction position: B

24 回・5 回であったため抽出された．

同様に，図 6 内の g の状況を考えると，開始時のインタラクションステートから，SubjectB が SubjectA に視線を向けるのを終了し，その後 SubjectA が発話を終了した場合に，SubjectA・SubjectB・SubjectC が次に話し出す場合である．これらの確率はそれぞれ 1/3 であると考えられるが，観測された回数はそれぞれ 0 回・3 回・0 回であった．

7.2 抽出結果と従来研究での解釈

7.1 では，最も多く検出されたインタラクションステートからの時間構造を例に示した．しかし，先述したように自動生成したアノテーションには偏ったノイズが含まれており，7.1 のインタラクションステートも比較的不自然なものであった．そこで，提案手法の評価を行うために，ノイズにあまり影響されていないと考えられる部分のみを頻繁に起こったものから調べていくこととした．具体的には，すべての Subject に視線アノテーションが付加されたインタラクションステートから始まる，視線アノテーションの変化を含まない時間構造のみに着目した．これは，視線アノテーションのノイズが抽出結果に対し最も大きな影響を与えていたからである．

7.2.1 インタラクションでの立場の違いによる発話確率の違い

ノイズにあまり影響されていないと考えられるうち，最も多く検出されたインタラクションの時間構造を**図 7** に示す．

このインタラクションステートは実験を通して，全部で 251 回観測された．この後，視線

Vol.2009-UBI-23 No.23
2009/7/17

に関するイベント 156 回，発話に関するイベントが 95 回であった．特に発話に関するイベントの内訳は，SubjectA の発話開始が 41 回，SubjectB の発話開始が 23 回，SubjectC の発話開始が 14 回であった．実際に，ここで検出されたシーンをワイヤレスマイクの音声で確認したところ，相手の体や顔の少し横を見て，視線をはずしている場合に，視線アノテーションが付加されることにノイズが，SubjectA が発話開始した場合に 18 回，SubjectB が発話を開始した場合に 12 回，SubjectC が発話を開始した場合に 3 回あった．これらを取り除いた後も，有意水準 5%の χ^2 検定で有意であり，これは 3 人のインタラクションでの立場における発話確率の差であると考えられる．

この例では発話確率が開始時のインタラクションステートのみに依存すると考えられる例であるため，より複雑な構造として，出現頻度が 2 番目であったインタラクションステートからの時間変化から検出された例を**図 8** に示す．この抽出された発話に関する変化の内訳は，SubjectA が発話を開始した場合は 7 回，SubjectB が発話を開始したのが 16 回，SubjectC が発話を開始したのが 2 回であった．この 25 回のシーンも環境カメラの映像とワイヤレスマイクの音声で確認したところ，自動で生成したアノテーションの誤検出と検出失敗によるものは，SubjectA が発話開始した場合に 1 回，SubjectB が発話を開始した場合に 4 回あった．これらを取り除いた後も，有意水準 5%の χ^2 検定で有意であり，これも 3 人のインタラクションでの立場における発話確率の差であると考えられる．

SubjectA の発話終了後，SubjectA がいったん間をおいて再度発話した場合は，よどみながら発言する（4 回），短時間息を止め再度発話を開始する（2 回）であった．また，SubjectC が発話を開始した場合は，話を始めた場合（1 回），相槌を行った場合（1 回）であった．

特に SubjectA の発話終了後，SubjectB が発話を開始した場合では，SubjectA への相槌（6 回），SubjectA への返答（1 回），話を開始する（6 回）という振る舞いが観察された．SubjectB が話を開始する振る舞いは，会話分析の研究では SubjectA が発話を終了することで，発話権を他の参与者である SubjectB や SubjectC に譲ろうとし，これを聞き手が発話することで，その発話権を取得したものであると解釈される[9]．

以上の結果を従来研究の枠組みで考えると，参与構造を支える話者交代の基本的な構造と一致している．SubjectA はそ SubjectB に視線配分することにより，SubjectB を次の発話者に選んでいると説明できる[10]．

7.2.2 発話中の指差し頻度

ノイズにあまり影響されていないと考えられるインタラクション構造から，立ち話以外であると考えられるインタラクションステートの中で最も多く検出された時間構造を**図 9** に示す．

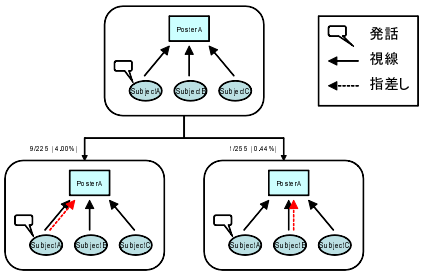


図 9 発話中の指差し頻度
Fig. 9 Pointing frequency during speech

この図では，視線を実線で，指差しを点線で，発話をフキダシで表記している．この開始時のインタラクションステートは実験を通して，全部で 225 回観測され，そのうちの 10 回を占める指差しに関する時間構造が抽出された．ここでは，被験者がインタラクションのコンテキストを無視しランダムに行動を起こすモデルでは，SubjectB と SubjectC が同じ立場であり，入れ替え済みであることを考慮し，SubjectA と SubjctB のそれぞれの発話開始確率は 1/3 と 2/3 となることに留意する必要がある．これらのシーンを特に詳しく，環境カメラの映像とワイヤレスマイクの音声で確認したところ，SubjectB が指差しを開始した場合として抽出されたイベントは，SubjectA の発話が途切れたときに，SubjectB が相槌を行い，その相槌中に指差しが検出された場合であった．一方，SubjectA の発話中に SubjectA が指差しを開始した場合は，9 回すべての場合において相槌中に行う指差しでは無かった．この 1 件の例外を無視しても，有意水準 5%の χ^2 検定で有意であり，同じ対象物を見ているときの指差しの開始と会話場での発話状態に強い関連性があることを示唆するものである．

さらに，これらのシーンの発話中で「この」「これ」という指示語は 4 回に現れ，その対象を特定しており，「このへん」「こっち」という範囲を特定する指示語がそれぞれ 1 回現れた．この発話が意味を持つためには，指差し等で対象や範囲を明確にする必要がある．さらに前者の指示語は，空間や視野を共有していないと効力を持たないものである．そのため，これは従来研究の枠組みでは，3 人が同じ対象物を見ているジョイントアテンションと呼ばれる状況下で指差しが起きたために発生した会話であると説明される．

Vol.2009-UBI-23 No.23
2009/7/17

8. おわりに

本論文では，記録したデータ全体から分析を行う，ボトムアップでのインタラクション分析の手法を提案した．また，評価実験を行い，計測したデータから半自動でアノテーションを生成した．その後，これらのアノテーションに提案した手法を適用し，時間構造の抽出を行った．最後に，アノテーションのノイズによるものと考えられる部分を除去したうえで，抽出された構造を従来研究の枠組みで検討した．これにより，この手法の将来的な可能性を示した．

一方で，この手法の問題点として，大きな時間構造を抽出するためには大量のデータが必要となることである．本論文では，大量のデータを必要とすることを考慮し，半自動アノテーション生成を行いある程度の有効性は示したと考えられるものの，同時にセンサのノイズや取り付け方・キャリブレーションの行い方によって自動アノテーションは結果が大きく影響されるという欠点も明らかにしたと考えられる．しかし，アノテーションの生成の問題に関しては，センサ機器の発展により近く解決すると考えている．また，人と人とのインタラクションの分析を行っていくに当たり，分析において統計的概念は欠かせないものであろう．

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．本研究における実験とデータ整理に多大な助力を頂いた，勝木弘氏，中田篤志氏，矢野正治氏に感謝します．

参 考 文 献

1) Chen, L., Harper, M., Franklin, A., R.Rose, T., Kimbara, I., Huang, Z. and Quek, F.: A Multimodal Analysis of Floor Control in Meetings, *MLMI*, Vol.3869, pp.36–49 (2006).
2) Michel, M., Ajot, J. and Fiscus, J.: The NIST Meeting Room Corpus 2 Phase 1, *Machine Learning for Multimodal Interaction*, Lecture Notes in Computer Science, Vol.4299/2006, Springer Berlin / Heidelberg, pp.13–23 (2006).
3) Carletta, J., Ashby, S., Bourban, S., Flynn, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A. and Reidsma, M. W. P.D.: The AMI meeting corpus: A pre-announcement, *In Proc. MLMI*, pp. 28–39 (2005).
4) Chen, L., Rose, R.T., Qiao, Y., Kimbara, I., Parrill, F., Welji, H., Han, T.X., Tu, J., Huang, Z., Harper, M., Quek, F., Xiong, Y., McNeill, D., Tuttle, R. and Huang, T.: VACE Multimodal Meeting Corpus, *Machine Learning for Multimodal Interaction*, Lecture Notes in Computer Science, Vol.3869/2006, Springer Berlin / Heidelberg, pp.40–51 (2006).
5) 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 (グループインタラクション支援, 特集ユビキタス社会におけるコラボレーションサービス)，情報処理学会論文誌，Vol.47, No.1, pp.121–130 (2006).
6) Shannon, C.: Prediction and entropy of printed English, *Bell System Technical Journal*, Vol.30, No.1, pp.50–64 (1951).
7) Nagao, M. and Mori, S.: A new method of N-gram statistics for large number of n and automatic extraction of words and phrases from large text data of Japanese, *Proceedings of the 15th conference on Computational linguistics-Volume 1*, Association for Computational Linguistics Morristown, NJ, USA, pp.611–615 (1994).
8) Cavnar, W. and Trenkle, J.: N-Gram-Based Text Categorization, *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pp.161–175 (1994).
9) Sacks, H., Schegloff, E. and Jefferson, G.: A simplest systematics for the organization of turn-taking for conversation, *Language*, Vol.50, No.4, pp.696–735 (1974).
10) 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行動の分析 (テーマ:一般)，言語・音声理解と対話処理研究会，Vol.38, pp.25–30 (2003).