1F2-5

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

Interpretion of Multiparty Conversation Situation based on Bodily Movement and Gestures

中田篤志*1 Atsushi Nakata　來嶋宏幸*1*2 Kijima Hiroyuki　角康之*1 Sumi Yasuyuki　西田豊明*1 Nishida Toyoaki

*1京都大学 大学院情報学研究科
Graduate School of Informatics,Kyoto University

This paper proposes a method of automatic interpretation of situations in human interactions from body movements. This method needs three steps. In first step we obtain basic body movements, pointing gesture and gazing, by automatic extraction algorithm, and in next step we find events in conversation from the basic body movements,and last we calculate authority of participants in conversation with a linear expression from basic movements and events.In automatic extraction, we get pointing gesture in 55% precision and 83% recall, and gazing in 77% precision and 54% recall from our automatic extraction algorithm. In addition, we obtained findings of the algorithm. In calculating authority, we discovered that major factors of the authority are "leaving from joint attention", "pointing gesture in joint attention" and "frequency of speech". It was carried out by a multiple regression analysis and confirmed by t-test with 95% confidence.

1. はじめに

多人数会話における話題の遷移や主導者などの会話の状況を機械的に理解することは，話題に応じた情報を提示するシステムや主導者を理解した上で会話に参加するエージェントなどに応用でき大変有用である．近年，このような会話の状況を非言語情報も援用して理解しようとする研究が多く行われている [Chen 06, Rienks 06, 角 03]．これらの研究は，座った状態でのミーティングといった会話環境や目標とするアプリケーションに強く依存しているため，もっと体系的に会話状況を解釈するための枠組みが求められている．

そこで我々は，会話状況をセンサデータから抽象的な状況理解までボトムアップに解釈していくための枠組みとして階層的解釈手法を提案している [來嶋 07, 高橋 04] ．この手法では，４つの段階を経て解釈を行う．最下層の RawData 層では，環境や人間に設置したセンサによってデータを取得する．次の Interaction Primitive 層 (以後 Primitive 層) では，センサデータから視線・発話といった基本的な動作を抽出する．Interaction Event 層（以後 Event 層） では，Primitive 層のデータを組み合わせて「二人以上が同時に同じものを見ている」(共同注視) といった会話中の重要な場面を発見する．Interaction Context 層（以後 Context 層）では，Primitive 層，Event 層のデータから会話の状況を推定する．これらの階層構造をとることによって,「A が B を見ている」という物理的な情報と「A が会話の主導権をとっている」という意味的な情報を別個のものとして扱うことができる．また，環境に応じて一部の階層の処理のみを変更し残りの階層で共通の処理を行うことで過去の分析を有効に利用していくことができる．

本研究では，この階層的解釈手法に基づいて意味的な情報を取得することを目指す．将来的には取得したセンサデータから自動的に意味的情報を取得することが目的であるが，現在の環境でははセンサから動作を自動抽出する際に見落としや誤認識が生じることが多く，意味的情報を取得するには精度が不十分である．そこで，本研究では指差し・視線といった Primitive 層のデータに対し正解を手作業で用意し，センサデータからの自動抽出を正解データに近づけることと，手作業で作成した正しいデータから重要な場面の発見・状況の解釈を行うことの２つに取り組んだ．

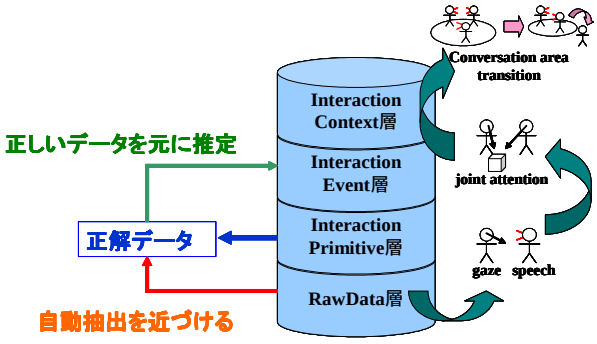

図 1: 階層的解釈手法

2. 多人数会話記録とアノテーション

分析を行うにあたって，身体動作を頻繁に行う協調作業をセンサデータで記録するための実験を行った．

センサデータの取得は，我々が構築しているインタラクション記録環境「IMADE ルーム」[來嶋 07] にて行った．この部屋は日常的に利用する研究室の一室を利用しており，日常的な空間にセンサを配置する形になっている．IMADE ルームには人間のインタラクションを記録するためのカメラ，マイク，モーションキャプチャ，アイマークレコーダなどのセンサが多数設置できるようになっており，それぞれのデータを同期させてデータベースに記録することができる．

実験では，3 人の被験者が「来年の研究室の備品・人物配置をどのようにするか」について話し合う様子を記録した．センサは，環境側に設置した複数台のカメラ，被験者の音声を記録するマイク，身体動作を取得するモーションキャプチャを使用した (図 2 左)．また，身体動作を頻繁に行ってもらうため，研究室で利用している 4 つの部屋を表すボードと，備品や人

連絡先: 中田篤志, 京都大学情報学研究科, 京都市左京区吉田本町,075-753-5391,nakata@ii.ist.i.kyoto-u.ac.jp

*2　來嶋宏幸：現在は KDDI 所属

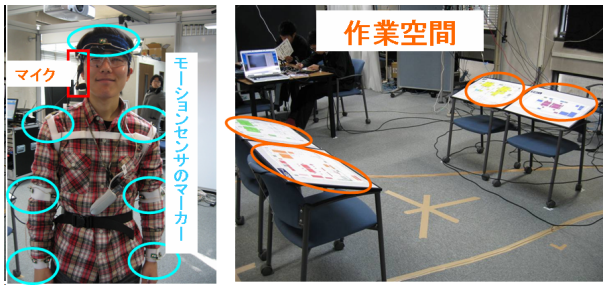

図 2: 実験の被験者と作業空間

物を表すマグネットを用意し，ボード上にマグネットを配置してもらった (図 2 右)．この実験は被験者を変えて計 5 回行われ，本研究ではそのうちの 1 回分を研究対象とした．

その後，実験で得られたデータに対して，指差し動作およびマグネットの操作，視線の遷移，発話のそれぞれに関して，筆者が映像と音声を閲覧して時刻との対応付けを行い，正解データとした．

3. 腕動作・視線の自動抽出

本研究では，モーションキャプチャのデータから，被験者のボードに対する指差し・マグネットの操作，視線の遷移を自動抽出することを試みた．ここでは手作業で作成したラベルを正解データとして利用した．

3.1 腕動作の自動抽出

指差し・マグネット操作の自動抽出では，まず肩と手のマーカーからなるベクトルがボードと交差している場面を抽出し，誤認識・見落としが生じている場面を網羅的に閲覧した．これにより，指差しを行っている際にマーカーが人の影に入るため，受信機でデータ取得を出来ない場面が多数見られた．また，この際受信失敗の前後では正確に自動抽出が行えることが分かった．そこで，マーカーの取得失敗の際に前後の自動抽出の結果から補間するヒューリスティクスを導入し，自動抽出の際の見落としを減らした (図 3)．結果として，適合率約 55%・再現率約 83% で自動抽出を行った．

図 3: マーカー取得失敗場面の分析環境 [來嶋 07] 上での画面．上の 2 つの画面がカメラ映像．下がその場面に対する正解データと自動抽出データ

3.2 視線の自動抽出

視線の自動抽出は，頭部方向による近似で行うものとし，頭の正面と真後ろに取り付けたマーカーを基準とした．

２つの頭部センサから得られたベクトルと動画から見える実際の視線方向を比較したところ，本実験では作業対象が腰の高さにあるため，視線方向が頭部方向に比べ鉛直下方向に向いていた．また，特に遠くのボードを見る際には視線がボードに比べ水平方向でずれていた．そこで，交差を取る平面をより広く取り，また頭部の 2 つのマーカーから取得したベクトルを鉛直下方向に回転させることでより正確な自動抽出を試みた．最も良い精度で自動抽出ができる回転角を検討した結果，最終的に鉛直方向の回転角は 25° を採用し，またその際の適合率は約 77%・再現率は約 54% となった．

4. 会話参与に関する積極性の数値化

基本的な身体動作から会話状況を取得する一例として，会話の参加に関する積極性に関して数値化することを試みた．ここでの積極性は場面ごとに変化していくものと考える．

数値化は手作業で作成したラベルを基にして，重回帰分析による線形式の構築によって行った．

4.1 目的変数

目的変数はアンケートによる評価を正解データとみなして用いた．

まず実験を 16 の場面に区切った．場面は，３秒以上の無音区間を場面転換の候補として，その前後を筆者が閲覧し場面が変わっていないと判断した部分をつなげる形で切り分けを行った．

このようにして切り分けられた実験風景を元に，9 人の評価者がそれぞれの場面における 3 人の被験者の会話参与に対する積極性に関して順位をつけた．「積極性」とは何かについては評価者に説明せず，評価者の主観に任せる形とした上で，実験後に採点の上で重要視した要素を尋ねた．その後，1 位＝ 3，2 位＝ 2，3 位＝ 1 という数値に置き換えることで積極性の値とし，評価者 9 人分の値を合計してその場面の参加者の積極性の値とした．

4.2 説明変数の候補

説明変数は，まず会話中で重要と思われる要素を Interaction Primitive 層，Interaction Event 層からそれぞれ選び出し，それらの 1 秒あたりの平均回数を変数の候補とした．その上で，変数増減法によって最終的な説明変数を選択した．

説明変数の候補として挙げたものは以下のものである．

Interaction Primitive 層

- 指差し・マグネット操作を行った回数
- 発話回数

Interaction Event 層

- 300msec 以上の沈黙の後に発話を行った回数
- 他の被験者の発話にかぶせて発話を行った回数
- 他の人に注視された回数
- 他の被験者が注視している中で指差し・マグネット操作を行った回数 (図 4)
- 共同注視から視線をはずした回数 (図 5)

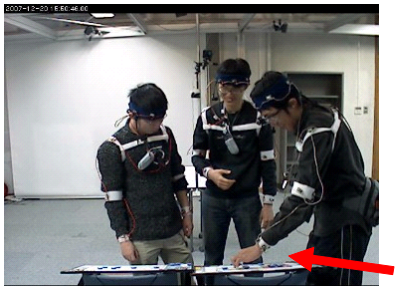

図 4: 注視中のマグネット操作の例

図 5: 共同注視から視線を外した場面の例

4.3 構築した式と統計データ

重回帰分析により，以下のような線形式が得られた．

$$Auth = 17.7x_1 + 46.4x_2 - 95.7x_3 + 13.7 \qquad (1)$$

$Auth$:会話に対する積極性の値
x_1:共同注視の中での指差し・マグネット操作の回数
x_2:発話回数
x_3:共同注視から外れた回数

重相関 R	0.790445
重決定 R2	0.624804
補正 R2	0.599222
観測数	48
有意 F	1.85E-09

	t	P-値
切片	10.6395	9.54E-14
x_1	2.608171	0.012386
x_2	5.800226	6.63E-07
x_3	-3.52978	0.000988

表 1: 積極性算出式の統計データ

(1) のそれぞれの係数に着目すると，共同注視の中での指差しの回数，発話回数は正の相関が，共同注視から外れた回数は負の相関があることが分かった．

この線形式の統計データを示したものが表 1 である．まず，重相関 R, 補正重決定 R2 に注目すると，この算出式は目的変数に対して十分相関が高いと考えられる．また，ここで得られた t 値を基に t 検定を行ったところ，x_2, x_3 に関しては 1% 水準で，x_1 に関しては 5% の水準で有意性を見ることが出来た．したがって，これらの身体動作は会話参与における積極性を見る上で大きく関わりがあるのではないかと考えられる．

これを元に，今回の実験である時刻の前後 50 秒間の各説明変数の回数を線形式 (1) に適用してその時刻の積極性の値とし，グラフとしたものがが図 6 である．

4.4 考察

線形式 (1) で発話数や共同注視の中での指差し・マグネット操作が積極性に対して強い正の相関があることは非常に直感的で納得のいく結果であるといえる．特に，通常の指差し・マグネット操作を説明変数に使った場合より共同注視の中でのマグネット操作を説明変数とするほうが良い式が得られたことは現実を反映していると考えられる．

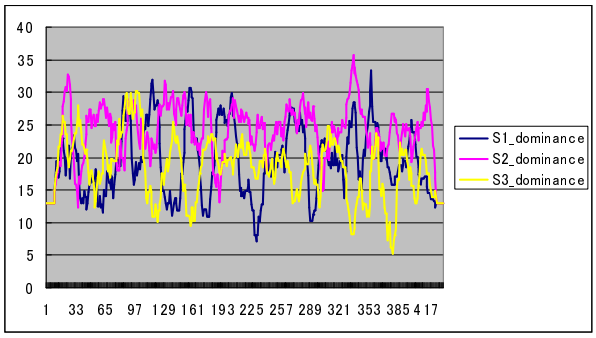

図 6: 会話参与への積極性を表すグラフ

一方，共同注視から外れるという行動に関しては他者を自身の興味対象に引き込むという意図と，話題に興味を失い別の対象に興味を向けているという 2 通りが考えられ，前者は会話参与への積極性に関し正の相関が，後者は負の相関があることが考えられる．線形式 (1) から負の相関が現れたことから，共同注視からの離脱は後者の場面が多いのではないかと推測される．今後，別の被験者による同様の実験データで確かめていきたい．

また，発話・300msec 以上の沈黙後の発話・他者の発話に割り込んだ発話という発話に関する要素を説明変数にした場合の線形式を重回帰分析で構築し，今回の身体動作を含む線形式の場合と比較した (図 7，比較しやすいように被験者のうち 2 人の積極性を示した)．すると，図中 A のような一人が話しながら作業を行い一人がそれを傍観するという場面では，どちらの算出式でも作業者が積極性の値が高くなり傍観者が低くなるという結果が得られた．それに対し図中 B のような一人が黙って作業したものに対し他の人がコメントをするという場面では，発話情報ではコメントをしたものが積極性の値が高くなったのに対し身体動作を使った場合は作業者のほうが高くなり，また評価者による採点も作業者のほうが高かった．

評価実験後に尋ねた積極性に関する採点の基準としては，課題への貢献度といった会話の意味内容に踏み込んで評価している例も見られ，非言語情報に着目して評価をした評価者はいなかった．意味内容を考慮せず取得した，非言語情報から構築した積極性の算出式がこの採点に対し有意性が見られたことは，音声認識や会話の意味内容の抽出といったコストのかかる手法をとらない会話状況の解釈の可能性を示唆している．

5. 終わりに

本研究では．センサデータからの身体動作の自動抽出を試みるとともに会話への参与における積極性を数値化する式の構築に取り組んだ．結果として，センサ情報から身体動作の抽出を行う際のヒューリスティクスをいくつか取得し，それによってある程度の精度で身体動作の自動抽出を行うことができた．また非言語情報を取り入れた，統計的に有意性を持つ積極性の算出式の構築を行った．

今後の課題としては，まず今回の身体動作の自動抽出の際に得られた知識を元に，加速度センサやアイマークレコーダの導入・適切な実験環境の設定などを行い，より精度の高い自動抽

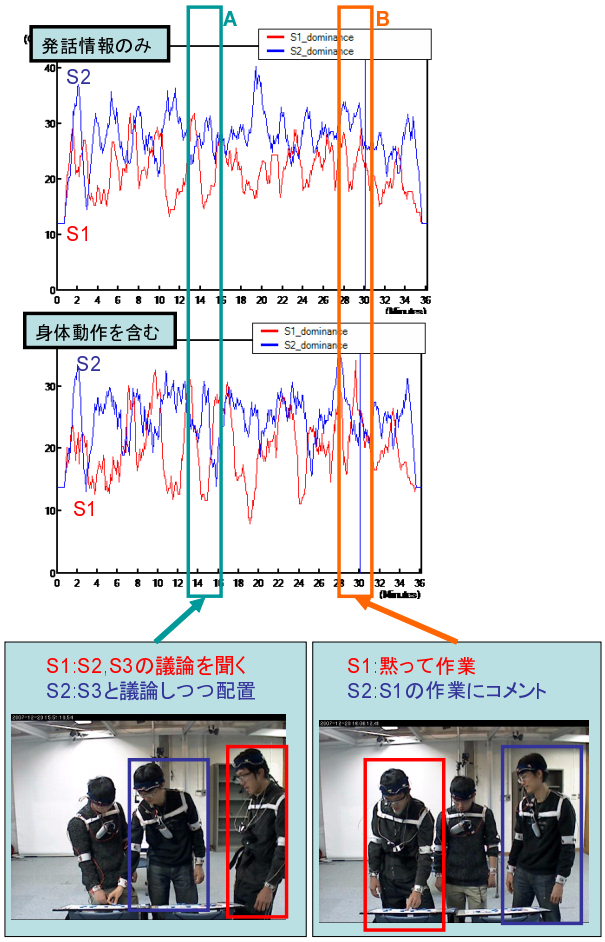

図 7: 本研究の積極性算出式と発話要素のみによる算出式との比較

出を行っていくことが挙げられる．また，今回得られた積極性算出式を他の被験者・会話環境に適用しその評価を行うこと，実際に積極性算出式を利用したアプリケーションの作成なども実践していきたい．

謝辞

本研究を進めるにあたり，研究の上でさまざまな面でのご助言・ご助力をいただきました坊農真弓氏に深く感謝いたします．

また，本論文の実験に協力してくださった，京都大学院情報学研究科知能情報学専攻河原研究室の常志強氏，京都大学大学院工学研究科電気工学専攻中村研究室の前田俊一氏，および西田・角研究室の福間良平氏，中沢拓磨氏，勝木弘氏に感謝いたします．

最後に，本論文に関して有益な議論をしていただき，実験にも協力してくださった西田・角研究室の皆様に感謝いたします．

参考文献

[Chen 06] 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 of Multimodal Interaction*, Vol. 3869, pp. 36–49 (2006)

[Rienks 06] Rienks, R. and Heylen, D.: Dominance Detection in Meetings Easily Obtainable Features, *Machine Learning of Multimodal Interaction*, Vol. 3869, pp. 76–86 (2006)

[角 03] 角 康之, 伊藤 禎宣, Fels, S., 松口 哲也, 間瀬 健二：協調的なインタラクションの記録と解釈, 情報処理学会論文誌, Vol. 44, No. 11, pp. 2628–2637 (2003)

[高橋 04] 高橋 昌史, 伊藤 禎宣, 土川 仁, 角 康之, 間瀬 健二, 小暮 潔：インタラクション解釈における階層構造の検討, 人工知能学会全国大会 (第 18 回), Vol. 2B1-02, pp. 1–4 (2004)

[來嶋 07] 來嶋 宏幸, 坊農 真弓, 角 康之, 西田 豊明：マルチモーダルインタラクション分析のためのコーパス環境構築, 情報処理学会研究報告, Vol. 2007, No. 99, pp. 63–70 (2007)