2007－SLP－67 (6)
2007／7／20

多数のセンサを用いたポスター会話の収録とその分析

瀬戸口 久雄[†] 高梨 克也[‡] 河原 達也[†‡]

[†]京都大学 情報学研究科 知能情報学専攻
[‡]京都大学 学術情報メディアセンター
〒606-8501 京都市左京区吉田二本松町

あらまし 我々は多数の研究者と協力して，会話をマルチモーダル (言語・非言語) な観点で分析できるような実験環境 (IMADE ルーム) を構築している．本研究では，話し手１名，聞き手２名からなるポスター会話を題材として，各話者の音声に加えて，視線やうなずき，ポインティング動作などを様々なセンサで記録した．本稿では，この仕様について述べるとともに，あいづちとうなずきに着目して，発話と視線との関係を分析した結果について報告する．

Multi-Modal Recording of Poster Sessions and Preliminary Analysis

Hisao SETOGUCHI[†] Katsuya TAKANASHI[‡] Tatsuya KAWAHARA[†‡]

[†]Graduate School of Informatics, Kyoto University
[‡]Academic Center for Computing and Media Studies, Kyoto University,
Sakyo-ku, Kyoto 606-8501, Japan

Abstract We are constructing a research environment called "IMADE room" which can capture a variety of multi-modal human interactions. With this setting, we have designed and conducted recordings of poster sessions by one presenter and two listeners. In addition to speech data of individual speakers, gazing, nodding and pointing events are recorded through multiple sensors. In this article, we describe the specification of the experiments and preliminary analysis on the relationship between verbal and non-verbal events.

1　研究の背景・目的

ミーティングや会話などをデジタルアーカイブとして保存・蓄積することが容易にできるようになったが，このようなコンテンツへの効率のよいアクセスを可能にするため，コンテンツが含む音声・映像などの情報に基づいて自動的にインデックスを付与する研究が盛んに行われている．会話ではさまざまなモダリティによるインタラクションがかわされるが，これらのモダリティの情報を複数組み合わせ，そこからメタ情報を抽出して，インデックスの付与を行うことを考える [1],[2]．その上で，どのような情報を組み合わせるか，また得られた情報をどのように利用するかといったことの検討が必要である．

我々は，そのための基礎データを収録するために，音声・映像に加えて，人間の身体・視線の動きを収録することができる実験環境を構築し，ポスター会話の収録を行った [3]．また，収録されたデータに対して，言語・非言語行動に関するラベルを付与し，予備的な分析を行った．特に発話や視線に対するあいづち，うなずきに着目し，それらの分布の相関を調べた．

2　ポスター会話

収録する会話を設計する上で，会話の目的付けの強さ，利用する情報リソース，参加者の役割付けの3点から，様々な会話の形態について検討を行った [4]．会話の目的付け，利用する情報のリソースの2軸によって分類を行った「会話マップ」を図1に示す．

本研究では，この中からポスター会話を選択した．この性質について考察する．

まず目的付けについては，ポスター会話は雑談と比較して，話すべきテーマや内容が決まっている．一方で，チケット予約やマップタスクなどの課題遂行対話と比べると，話の進め方に比較的自由度があるという設定になっている．

次にリソースに関して考える．ミーティングなどは会話を行う上で必ずしも情報リソースを必要としないが，ポスター会話の場合は少なくともポスターという（静的な）情報リソースを必要とし，また実際にポスターの内容を利用して会話が進められる．一方，料理中にかわされる会話などでは，実世界にあるものを動かしたり，形を変化させるなど，物理的なものを動的な情報リソースとして利用しながら会話が進められる．ポスター会話はリソースが全くないものとその場で情報を発生させる動的なリソースがあるものの中間的な形態である．

ゴール＼リソース	なし	情報媒体	実物体
観測可能	チケット予約	マップタスク	料理タスク
達成点あり	議論	講義	
方向性あり	ミーティング	ポスター会話 展示会	
不明確	雑談		

図 1: 会話マップ

参加者の役割付けから見ると，話し手と聞き手が固定されることがないミーティングなどと比較して，ポスター会話は話し手と聞き手という役割付けは存在する．しかし，講演などと異なり，話し手がポスターの内容について話している途中でも，聞き手は必要に応じて比較的容易に質問などをすることができ，またそれによって会話の主導権を握ることができる．つまり，講演とミーティングの中間的な位置づけとなっている．また，ポスター会話では，話し手と聞き手で持っている情報量や役割が非対称になっている．これにより，話し手は聞き手に話を聞かせようとする行動を，聞き手は話し手の話を受け取る行動をより多くとるようになり，それぞれの方向の行動をより明確に分析しやすい形態となっている．

今回，話し手1人対聞き手2人のセッティングとした．これは，聞き手間でのインタラクションの発生を含んでいるデータとすること，また個人間での行動パターンの比較を行えるデータとすることを意図したものである．

3　収録環境の構築

3.1　IMADE ルーム

様々なマルチモーダルインタラクションを収録するための環境として，科研費特定領域研究「情報爆発ＩＴ基盤」[5],[6] の一環で，京都大学工学部10号館2階に「IMADE ルーム」が構築されている．音声，映像の他，人間の動きや生態信号を記録できるように設計が行われている．

3.2　収録に用いたセンサと配置

現在，IMADE ルーム内では音声，映像，人間の動作と視線を収録できるセンサがある．そのうち，今

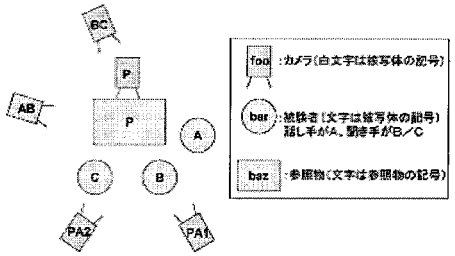

図 2: カメラの設置位置（概略図）

図 3: ポスターの設置例

回の収録に用いたものに関して述べる．

まず音声に関しては，各話者の音声をそれぞれ独立に収録でき，かつ部屋内を自由に動き回れる程度の行動の自由度を持たせるため，話者数分のワイヤレスマイクを使用した．マイクは SHURE 社製 WH30-XLR ヘッドウォーンマイクである．音声は，ATI 社製オーディオプリアンプで増幅した後，48kHz,16bit の PCM データとして各チャネル独立に記録した．

これに加えて，ポスターの上部に 8 本のマイクからなるマイクロフォンアレイを設置し，ヘッドウォーンマイクよりも距離を置いた形で発話を収録した．

映像を収録するために，SONY 製据付型カメラを IMADE ルーム内に 8 基設置している．これにより，最大で 8 視点の映像を同時に収録することができる．映像は Mpeg-2 形式でサーバに蓄積される．

今回のポスター会話の収録に際しては，全ての話者の行動を 8 台のカメラのうち少なくとも 1 台で捕捉できる必要がある．特に，ポスターボードによって死角にならないよう留意する必要がある．そのために，図 2 に示すような配置とした．図中ではポスターを P，話し手を A，話し手に近い聞き手を B，遠い聞き手を C と表記している．以下，この表記を利用する．カメラの視点を，PA，AB，BC 間のそれぞれのインタラクションを捉えられる位置・方向として，これらの視点の映像を組み合わせることで，P,A,B,C の任意の組み合わせの間で発生しているインタラクションの情報を復元可能な形とした．さらにある時点の発話に対応して，ポスターへのポインティングがどのように発生しているかを詳細に記録するため，ポスターのみを拡大して撮影するカメラを一台追加した．

人間の動作を収録するセンサとしてモーションキャプチャシステムを設置している．これは PhaseSpace 社製のもので，赤外線を発する LED を取り付け，CCD センサを搭載したカメラが LED からの光を受け取って LED の 3 次元座標を計測する方式である．3 次元座標のデータは 30Hz の時系列データの形で記録される．

モーションキャプチャの配置についても，ポスターが遮蔽にならず，かつ参照する人間の動きが不自然に下を向くといった状態にならないよう，図 3 に示すように，ポスター台を設計し水平に対して持ち上げた形で設置するようにした．今回の収録では，持ち上げた角度は約 22° となっている．

視線を収録するセンサとして，アイマークレコーダも用いた．ASL 社製 MobileEye で，デジタルビデオカメラに接続されたゴーグルを装着し，赤外線の反射を利用して装着者の視線の動きを追跡するシステムである．ゴーグル上に取り付けられたカメラから収録された一人称視点の映像上に，装着者の視線の動きから計算された注視点をトラックしたマーカを重ね合わせた映像が，デジタルビデオとして収録され，映像上のマーカ位置の時系列データとして，装着者の視線の動きの数値データが記録される．

4 会話の収録とアノテーション

4.1 収録したデータ

現在までに，計 5 セッションのポスター会話を収録した．1 回のポスター会話の長さは約 20 分である．以下，セッションを取った時系列順にセッション 1,2…と番号を用いて各セッションを参照する．

全セッションについて音声，映像のデータが収録されており，セッション 3 を除き，モーションキャプチャのデータが，またセッション 1,2 のみアイマークレコーダのデータが収録されている．セッション

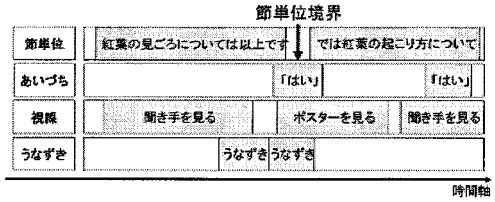

図 4: アノテーションの例

1,2,3 と 4,5 では話し手が異なり，使用したポスターも異なる．また全てのセッションについて，聞き手も異なる．ポスターは 4 つに分割したそれぞれの領域に 1 つずつ異なる話題についての情報を記載したものである．

4.2 付与したアノテーション

図 4 のような言語・非言語のアノテーションを施した．言語的なアノテーションは日本語話し言葉コーパス (CSJ)[7] の基準に準拠して，音声を書き起こし，IPU，節単位境界のラベルを付与した．これに加えて，今回作成した基準に従って，あいづちを認定した．非言語ラベリングは映像を元にうなずき・視線のラベルを付けた．今回は，モーションキャプチャのデータを利用していない．これはモーションキャプチャのデータは自動処理を行う際の基礎データとなるものであるが，自動処理を行う際の正解データを与える作業としてハンドラベリングを優先したためである．

今回のアノテーションはセッション 4 のデータの先頭 2 つの話題の部分，411.196 秒の会話に対して行い，「話し手の視線」「聞き手のうなずき」「話し手の発話における節単位境界」「聞き手のあいづち」について，それぞれ開始・終了時刻をマークした．

会話中に発生している言語・非言語の情報はこの他にも多数あるが，今回このようなアノテーションを選んだのは，話し手に関しては視線と発話という発信者としての行動を，聞き手に関してはうなずきやあいづちという受け手としての行動を捉えることで，最低限の分析を行うことができると考えたためである．

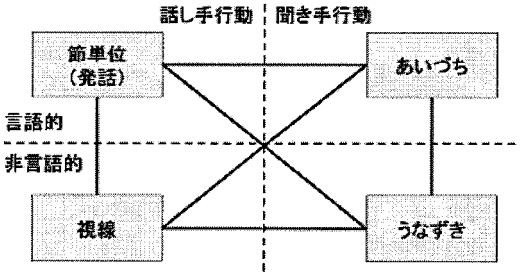

図 5: 話し手・聞き手，言語・非言語の各行動の結びつき

5 言語・非言語行動の相関についての分析

5.1 分析の着眼点

前節で述べたアノテーションを元に，話し手と聞き手の間で起こっている現象の時間的な相関について分析を行った．

図 5 に，アノテーションした各行動が話し手行動か聞き手行動か，また言語的な行動か非言語的な行動かを示す．この図の中で線で結ばれている行動の組について，それぞれ相関を求めることができる．

話し手の視線や発話に聞き手が何らかの反応を返すということは我々の会話に関する経験から自然に想定される．今回の分析では特に，視線の切り替わりや発話の境界において，聞き手の反応がどのように分布しているかに着目した．また，話し手と聞き手の言語側どうしの行動の相関を調べるために話し手の発話と聞き手のあいづち，非言語側どうしの相関を調べるために話し手の視線と聞き手のうなずきの関係にそれぞれ注目した．さらに，話し手の発話と視線の重複部分と聞き手行動の相関に注目することで，会話における言語・非言語の行動の相関を調べる．

なおポスター会話は参加者の役割が非対称な形態であるため，特に分析対象区間においては話し手と聞き手の発話やあいづち，うなずきの発生頻度のバランスも非対称であり，聞き手の発話，話し手のあいづち，うなずきの発生頻度はともに小さい．

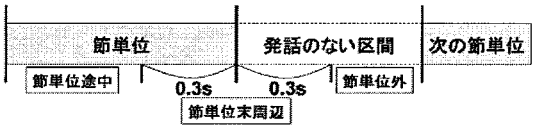

図 6: 節単位に関する区間の認定

5.2 話し手の発話と聞き手のあいづち

話し手の発話の単位として節単位を用い，図 6 に示すように，会話全体を節単位途中・節単位末周辺・節単位外時間という 3 種類に分類した．今回の分析では，原則として節単位終了直前，直後の時間は各 0.3 秒とし，節単位末周辺の時間は 0.6 秒とした．

ここで分類した 3 種類の時間の長さと，節単位途中と節単位末周辺について，その時間中に発生した聞き手行動の数を元に，節単位途中と節単位末周辺での単位時間当たりの聞き手反応の数を求めた．これらの数を表 1，表 2 にまとめる．分析した会話に含まれる節単位の数は 59 個，節単位途中の累計時間は 340.2 秒，節単位末周辺の累計時間は 35.4 秒である．

聞き手の節単位途中でのあいづちの程度に対する節単位末周辺での程度は，話者 B の場合約 3.7 倍，話者 C の場合約 4.9 倍である．同様の計算をうなずきについて行うと話者 B の場合約 2.0 倍，話者 C の場合約 1.9 倍となり，節単位途中より節単位末周辺の方がより聞き手行動が発生しやすい傾向が見られる．また，節単位末でのあいづちとうなずきの単位時間当たり数の比の比較から，節単位末ではうなずきよりあいづちのほうが聞き手行動として多い傾向があることもわかる．

表 1: 発話に対する聞き手行動の数（聞き手 B）

うなずき	度数 [回]	時間平均 [回/s]
節単位途中	82	0.24
節単位末	16	0.45
あいづち	**度数 [回]**	**時間平均 [回/s]**
節単位途中	43	0.13
節単位末	17	0.48

表 2: 発話に対する聞き手行動の数（聞き手 C）

うなずき	度数 [回]	時間平均 [回/s]
節単位途中	53	0.16
節単位末	11	0.31
あいづち	**度数 [回]**	**時間平均 [回/s]**
節単位途中	42	0.12
節単位末	21	0.59

表 3: 視線の有無と聞き手行動の数（聞き手 B）

うなずき	度数 [回]	時間平均 [回/s]
視線あり	44	0.39
視線なし	56	0.19
あいづち	**度数 [回]**	**時間平均 [回/s]**
視線あり	22	0.20
視線なし	40	0.13

表 4: 視線の有無と聞き手行動の数（聞き手 C）

うなずき	度数 [回]	時間平均 [回/s]
視線あり	36	0.23
視線なし	31	0.12
あいづち	**度数 [回]**	**時間平均 [回/s]**
視線あり	38	0.24
視線なし	28	0.11

5.3 話し手の視線と聞き手のうなずき

前節と同様の分析を，次は話し手の視線と聞き手行動についても行った．この分析においては，全会話時間を視線が向けられている時間とそうでない時間の 2 種類に分割する．視線については発話と異なり，2 人いる聞き手のどちらに向けられているかが分かれるため，それぞれの聞き手について分析を行う必要がある．話者 B,C に対して，視線を向けられていた時間とそうでない時間，それぞれの時間での聞き手行動の数，単位時間当たりの聞き手行動の数をまとめた表をそれぞれ，表 3，表 4 に示す．視線が向いている累計時間は話者 B で 112.6 秒，話者 C で 156.9 秒であり，視線が向いていない累計時間は話者 B で 298.6 秒，話者 C で 254.3 秒である．

話し手の視線が向けられている時間における聞き手行動の単位時間当たり数と，視線がない時間の聞き手行動の単位時間当たり数に，約 1.5 倍から 2.2 倍の開きが見られ，話し手の視線が向けられていると，あいづち・うなずきの両方の聞き手行動がより多く起こる傾向が見られる．

話者 B に関しては，視線の向けられている時間でのあいづちとうなずきの単位時間あたり数に有意な差が見られるが，話者 C についてはその差は見られない．視線に対する聞き手の反応として，あいづちかうなずきのいずれを行うかが個人差によるものな

表 5: 話し手の発話・視線とあいづち（参与者平均）

発話	視線	累計時間 [s]	度数 [回]	時間平均
単位末	あり	9.98	8.5	0.85
単位末	なし	22.18	10.5	0.47
途中	あり	121.36	21.5	0.18
途中	なし	219.18	21	0.10
単位外	あり	3.42	0	0.00
単位外	なし	35.09	2.5	0.07

表 6: 話し手の発話・視線とうなずき（参与者平均）

発話	視線	累計時間 [s]	度数 [回]	時間平均
単位末	あり	9.98	6	0.60
単位末	なし	22.18	7	0.32
途中	あり	121.36	34.5	0.28
途中	なし	219.18	34.5	0.16
単位外	あり	3.42	0.5	0.15
単位外	なし	35.09	1.5	0.04

のか，一般的にどちらかが多い傾向にあるのかを判定するにはさらなるデータの分析が必要である．

5.4 発話と視線の組み合わせと聞き手行動

話し手の発話と視線の組み合わせに対する聞き手行動の相関を分析し，話し手の言語・非言語の行動の組み合わせが聞き手行動に与える影響について調べる．

話し手の発話は節単位途中・節単位末・節単位外，視線は視線を向けているか向けていないかに分けられ，これらの組み合わせにより分析対象区間は 6 種類に分類される．このように分類された行動に対する聞き手行動の頻度の分布を表 5，表 6 に示す．表中の「累計時間」は話し手行動の累計時間，「度数」は聞き手行動の回数を表す．「時間平均」は度数を累計時間で割ったものである．節単位途中・節単位末周辺・節単位外をそれぞれ「途中」「単位末」「単位外」と略記している．ここでは聞き手の平均を取っているが，これは聞き手ごとに傾向に顕著な差が見られなかったことと，聞き手ごとのサンプルサイズが小さかったことによる．

「単位末・視線あり」のときのあいづち・うなずきの頻度は「単位末・視線なし」「途中・視線あり」の場合と比べて約 1.8 倍から約 4.7 倍となっており，特に視線が向けられている状態で発話が節単位末に入ったときの差がかなり大きい．このことは発話末であることと視線が向いているということの組み合わせが，聞き手行動を促す上で相乗的な作用を持っていることを示唆している．これは言語・非言語のモダリティ間の相関に意味があることを示している．

6 今後の課題

今回の分析にあたり行ったアノテーションは非常に限定された範囲に限られたものであった．これら以外に，聞き手の視線，ポインティング，話し手のうなずき，ハンドジェスチャ等は会話を分析する上で重要な意味を持つものと考えられる．これらにより，話し手と聞き手の視線が合っている場合とそうでない場合の比較，話し手のポインティング・ジェスチャが聞き手の行動に与える影響の分析，話し手のうなずきの意味など，会話の興味深い側面を探ることが可能になると考えている．また，モーションキャプチャ・アイマークレコーダを用いてこのようなアノテーションを（半）自動化する可能性についても今後探っていきたい．

謝辞

データ収録のための設備を提供していただきました京都大学情報学研究科西田・角研究室，工学研究科中村研究室の皆様と，(株) NTT コミュニケーション科学基礎研究所の荒木章子氏，石塚健太郎氏，またデータ収録にご参加いただいた皆様に心より感謝いたします．

参考文献

[1] 角康之，熊谷賢，瀬戸口久雄，西田豊明：非言語情報を利用した会話シーンの抽出と意味的インデキシング，情報処理学会研究報告（ヒューマンインタフェース），No.2006-HI-119，pp.87-94(2006)

[2] 長谷川将宏，河原達也，奥乃博：談話標識の抽出に基づいた討論音声への MPEG-7 タグの自動付与，人工知能学会研究会資料，SIG-SLUD-A203，pp.15-20

[3] 坊農真弓，鈴木紀子，片桐恭弘：多人数会話における参与構造分析—インタラクション行動から興味対象を抽出する，認知科学，No.11，Vol.3，pp.214-227.(2004)

[4] AMI Project：
http://www.amiproject.org/

[5] 松山隆司，西田豊明，國吉康夫：情報爆発時代におけるヒューマンコミュニケーション基盤，人工知能学会誌，No.22，Vol.6，pp.229-234.(2006)

[6] i-explosion 情報爆発：
http://i-explosion.ex.nii.ac.jp/i-explosion/

[7] 日本語話し言葉コーパス：
http://www2.kokken.go.jp/%7Ecsj/public/members_only/manuals/