社会言語科学　第14巻第1号
2011年9月　82–96ページ

The Japanese Journal of Language in Society
Vol. 14, No. 1, September 2011, pp. 82–96

研究論文

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（公立はこだて未来大学システム情報科学部），
矢野正治・西田豊明（京都大学情報学研究科）

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，3～5人による会話的インタラクションを計測し，会話状況の解釈や支援を目的とした研究基盤の構築を進めている．本稿では，筆者のグループが構築した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，そこで計測されたデータ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．iCorpusStudioの特徴は，映像・音声の閲覧やラベリング作業を可能にするだけでなく，非言語情報間の時間構造分析を行うためのラベル間演算や身体動作，視線移動などの数値データ間計算を容易にし，非言語構造パターン解釈のための仮説検証を支援することである．本稿では，聞き手反応に注目した指さし行為の認定に関する分析例を紹介し，iCorpusStudioの利用価値を示す．また，データから発見的に会話構造を理解する試みの例として，非言語情報から会話参加の積極性を解釈する試みを紹介する．

キーワード：多人数会話，会話構造分析，マルチモーダルデータ，インタラクション・コーパス

Structural Understanding of Multiparty Conversation through Multimodal Data

Yasuyuki Sumi (Faculty of Systems Information Science, Future University-Hakodate)
Masaharu Yano and Toyoaki Nishida (Graduate School of Informatics, Kyoto University)

This paper describes the IMADE (Interaction Measurement, Analysis, and Design Environment) project to provide an environment for recording and analyzing human conversational interactions. The IMADE room is designed to record audio/visual, human-motion, and eye gazing data to build an interaction corpus, which can be used to provide understanding of the nonverbal behavior of humans. In this paper, we describe the notion of interaction corpus and introduce iCorpusStudio, a software environment for browsing and analyzing the interaction corpus. We present the utility value of iCorpusStudio through an analysis of the gesture of 'pointing' in examples focusing on hearers' visual attention. The paper also attempts to show how nonverbal behavior can be used to quantify the activeness of individual participants in consensus-building type conversation.

Key words: multiparty conversation, conversational structure analysis, multimodal data, interaction corpus

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（角・西田・坊農・來嶋，2008; Sumi, Yano, & Nishida, 2010）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，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，これまでにもいくつかなされてきた．その代表的なものであるAMI (Carletta, Ashby, Bourbon, Flynn, Guillemot, Hain, Kadlec, Karaiskos, 2006; Renals et. al., 2007) は，グループミーティングのコーパスを構築し，ミーティングデータの効率的な再利用を目的として，音声認識による自動タギングや会話分析を行った．CHIL (Waibel & Stiefelhagen, 2009) は，オフィスや講義室におけるグループの会話的インタラクションを見守り，グループ活動の記録や個人化サービスの提供を目的とし，知的環境構築のためのフレームワークの構築，音声・映像データからの表情・感情理解，機械学習手法による人の動作自動検出や状況認識の技術開発を行った．VACE (Chen et al., 2006) も主に着座式のミーティングに焦点を当て，映像，音声，動作の自動認識によるミーティングデータのコンテンツ化や，研究推進のためのツール構築を行った．

これらの試みにほぼ共通しているのは，発話に関する映像・音声データに加えて，頭部運動，ジェスチャなども含めたマルチモーダルなデータを扱っていること，また，主に着座式のミーティングを対象とし，グループ内の会話的インタラクションのコンテンツ化を指向した技術開発やツール開発に焦点を当てているということである．グループミーティングは筆者らの興味対象にも含まれており，道具として，カメラ，マイク，モーションキャプチャシステムを利用していることも共通している．

ただし，筆者らの興味は，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）も対象としている．したがって，筆者らは，着座式のミーティングだけでなく，自由に歩き回りながらのおしゃべりや，ポスター発表 (Kawahara, Setoguchi, Takanashi, Ishizuka, & Araki, 2008) 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．また，モダリティとして，特に視線に興味があり，アイマークレコーダを導入し，複数人の視線とジェスチャ，視線と頭部運動の関係なども詳細に計測できる環境を構築した．

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録するために，以下のような様々なセンサやサーバを設置している．

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．

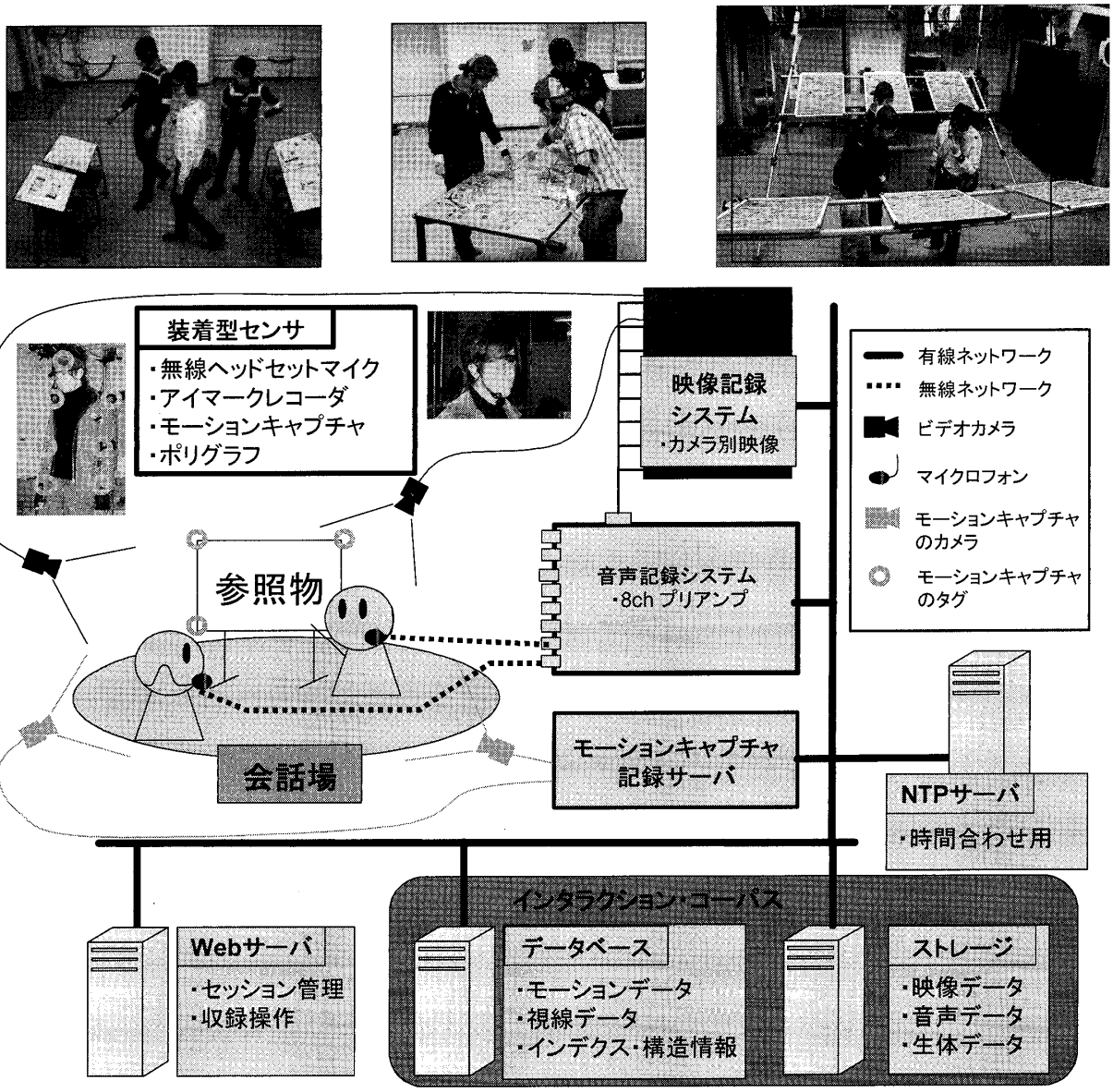

図1　IMADE ルームの概念図

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，各人の発話内容を人ごとに分離して収録できる．ワイヤレスマイクと8チャンネル音声記録機器を導入している．

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 システムを導入し，10台の赤外線カメラを用いている．

アイマークレコーダ　環境内の各人の眼球運動を計測する．頭部に装着した一人称映像と，その中での視線方向を示す2次元座標データが記録できる．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

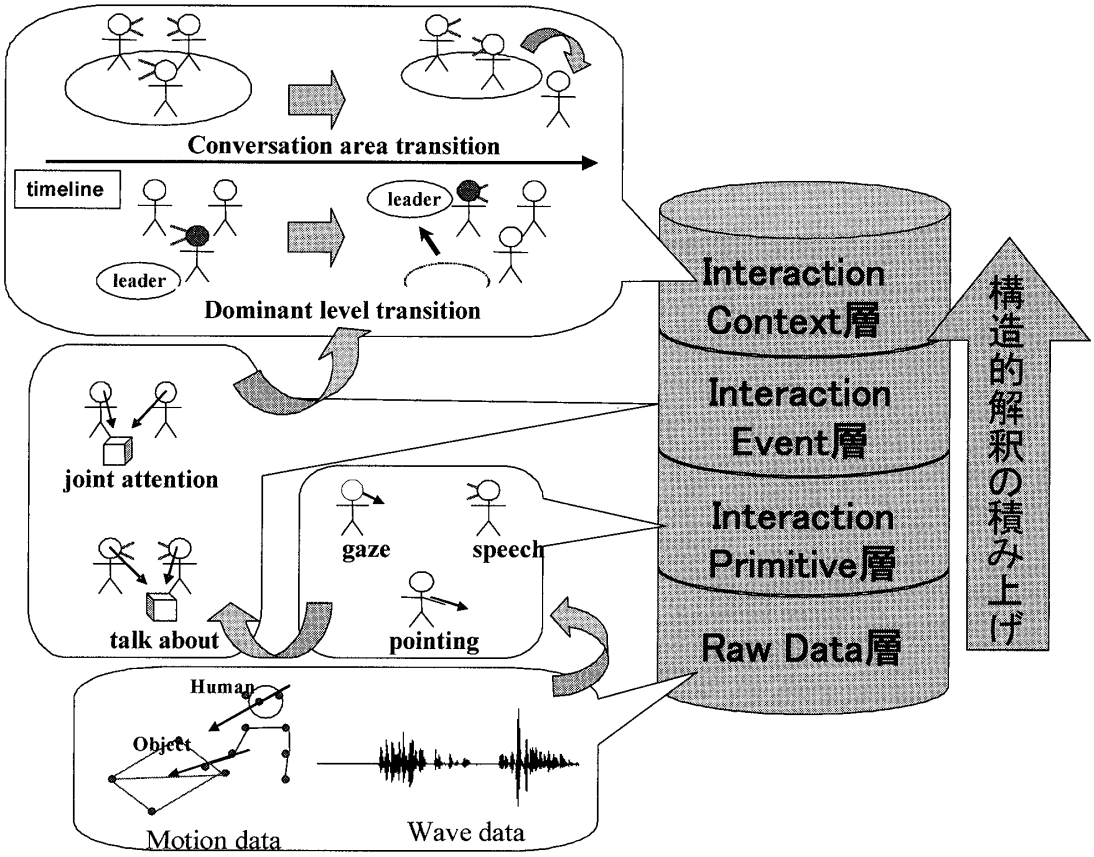

図2　インタラクションの階層的解釈モデル

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（田浦，2008）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

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．計測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を目的としている．

筆者らの目標は，会話の内容や作業対象に関する意味的な理解（高度な音声処理や画像処理）に深入りせずに，会話状況のダイナミクスを理解することである．我々人間は，自分の知らない言語を話している外国人同士の会話を横から観察していても，誰が会話を主導しているか，感情的にどういう状況か，話しかけても良いタイミングか，といった状況を理解することができる．しかし，現在のロボットや知的環境と呼ばれるものは，そういった会話状況を理解することがほとんどできない．筆者らの大きな目標は，人間が当たり前に行っている，非言語情報からの会話状況理解のメカニズムを理解し，コンピュータが利用できるような「辞書と文法」にすることである．その一歩として，インタラクション・コーパスを構築し，その中で非言語インタラクションの「辞書」となりえるようなインタラクションの要素を規定し，さらに，その発生系列のパターンを「文法」として記述していきたい．

当面の目標として，会話参加者の発話交代や身振り手振り，そして視線の移動や共同注意といった非言語的な現象から，話題の転換点を検出したり，各参加者の会話への参与度の変化をとらえたい．解釈の元となるデータは，モーションキャプチャシステムやマイクから得られるが，そういったデータから一足飛びに抽象的な解釈を得ることは難しいので，ここでは4つの階層による体系的な解釈の積み上げを試みる．以下，図2の各層について説明する．

Raw Data 層　ここでは，モーションキャプチャシステムやアイマークレコーダから時々刻々得られる座標データや，音声や生体反応（脈拍，呼吸など）に関する波形データから，身体動作や発話に関する要素を取り出すための準備を行う．例えば，モーションキャプチャデータから頭の向きや腕の向きを得るには，体の形状や動きを単純なモデルに近似し，いくつかのタグから仮想的なベクトルを形成する必要がある．また，会話における参加者間の発話交代の様子を知りたい場合には，各参加者が身に着けた接話マイクのパワー変化から発話の有無を判定する．次節で紹介するマルチモーダルデータの分析ツールiCorpusStudioには，そういった判定ルールをプラグインとして用意してある．

Interaction Primitive 層　ここでは，Raw Data層で得られたベクトルデータや発話の有無に関するデータから，インタラクション要素の抽出を行う．例えば，頭の向きや腕の向きを表すベクトルが空間内の対象（ポスターや会話相手のジェスチャ空間など）を貫く現象を検出して，ある対象を「見た」，「指さした」というインタラクション要素を推定する．これらのベクトル演算はiCorpusStudioの基本機能として提供されている．ただし，現実的には，人体や動的に発生するジェスチャ空間のモデリングには近似が行われるし，視線ベクトルとその対象物の衝突を検出するには検出範囲のマージンを設定する必要が生じる場合が多い．また，ベクトル衝突や発話有無の検出結果は離散データになるので，それらを時間方向にクラスタリングする必要が生じる場合が多い．このように，モデル化には多くのパラメータが発生し，その調整が分析者ごとの理論に他ならず，そういった仮説や作業プロセスがiCorpusStudioに外在化され，研究コミュニティで再利用されることになる．

Interaction Event 層　この層は，Interaction Primitive層で得られたインタラクション要素を組み合わせて，複数人による社会的現象の検出を試みる層である．例えば，複数人が同一の対象物に視線を向ける（共同注視），複数人がある物を参照しながら会話をする，といった現象を検出する．検出は，iCorpusStudio上でラベル間の重なりを自動検出して行う．ただし，単純なAND検索で満足な検出ができることは稀であり，実際は，重なり検出にマージンを持たせたり，他のモダリティのラベルを考慮した文脈をモデルに組み込んだりする必要があり，その作業こそが研究者の興味の対象となる．

Interaction Context 層　この層は，これまでの層の要素を組み合わせて，より「大胆な」仮説の検証を試みる層である．筆者らの興味の対象は，例えば，複数人の発話，視線方向といったインタラクション要素の時空間的なクラスタリングから，会話場の発生・消滅のダイナミクスを定式化することである（高橋・角・伊藤・間瀬・小暮・西田，2008）．また，発話量の変化，視線移動，対象物への作業行為といった様々なインタラクション要素の組み合わせから，会話に参加するメンバーの会話支配性(dominance)を定量化することも試みている（6節で紹介する）．

こういった分析作業は，大まかに言うと，会話状況のデザイン，会話データの計測，センサデータの手直し，解釈ルールの試作と評価といった一連の作業を行うことになる．通常は一回のサイクルで満足のいく結果が得られることは稀なので，このサイクルを何度か繰り返すことが望まれる．しかし，従来の会話分析はデータの取得や解釈ルールの分析に多くの人的コストと時間をかけてきたため，これらのサイクルを繰り返すことは実際は困難であった．それに対してIMADEでは，計測のためのハードウェア基盤と，分析のためのソフトウェア基盤を整備することで，これらのサイクルの高速化と再利用性の向上を目指している．

表 1　会話分析ツールの基本機能の比較

	Anvil	WaveSurfer	NXT	JFerret	MacVisSTA	ELAN	iCorpusStudio
複数映像	×	×	×	○	○	○	○
音声波形の表示	○	○	×	×	○	○	○
ラベリング作業	○	○	○	×	○	○	○
検索機能	×	×	×	○	×	○	○
データ形式	XML	CSV	XML	XML	XML	XML	CSV

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

4.1 基本的な機能と特徴

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（來嶋・坊農・角・西田, 2007; 矢野・中田・福間・角・西田, 2009）.

これまでにも，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，多くの会話分析者は，Anvil (Kipp, 2001)[1] や WaveSurfer[2] を使ってきた．ほかにも，NXT (Carletta, Evert, Heid, Kilgour, Robertson, & Voormann, 2003)，AMI で JFerret，VACE で MacVisSTA (Rose, Quek, & Shi, 2004) などが開発され，最近では ELAN[3] を使う研究者が増えてきた．また，これらのツール間のデータ共有方法についても議論がなされている[4].

これらのツールと我々が開発した iCorpusStudio の基本機能の比較を表 1 にまとめた．従来の会話分析では，音声発話に注目した分析が多かったため，単一のビデオデータと音声波形を閲覧しながらの分析作業が行われていた．そのため Anvil, WaveSurfer，NXT といった初期のツールは，単一ビデオデータのみの対応となっていた.

しかし，多人数会話の分析では単一のビデオデータだけではなく，複数人数をとらえた複数のセンサから同時に取得されたマルチモーダルなデータを扱う必要性が高まった．そのため，JFerret, MacVisSTA など，多人数会話のコーパスを構築しているプロジェクトにおいて開発されたツールは，多視点映像やそのコーパス特有のデータに対応している．JFerret は AMI の主目的のひとつであるアノテーション（ラベルと読み換えても良い）を用いたミーティングブラウザとして開発された．そのため，可視化されるデータの種類は映像の他に会話内容の書き起こしやミーティングで使用されたスライド，ホワイトボードのストロークと多岐にわたる．また，検索機能を用意することで，分析者の興味のある部分に焦点を当てて閲覧できるという特徴がある．VACE で開発された MacVisSTA も，複数映像やモーションキャプチャの 3 次元座標や音声波形を閲覧しながらラベル付与作業を行うことができる.

筆者らの iCorpusStudio も，基本的な設計方針としては MacVisSTA を手本にして開発が行われ，それに検索機能を加えていった．近年開発が進んでいる ELAN も，ラベリング環境という意味ではほぼ同様の機能を備えており，マルチプラットフォームで動作することやインタフェースの使いやすさから，多くの分析者に利用されている.

しかし筆者らが必要としているのは，従来のシステムが目指してきたような特定シーンへのラベル付与だけではない．大量にあるコーパスデータの中から分析者が閲覧したいシーンを見つけ出したり，社会的関係やシーンの意味的理解といった，より抽象度の高い解釈を見出す方法を検討するための環境である．そのため，従来のラベリング機能に加えて，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ことが，iCorpusStudio の特徴である.

また，筆者らの環境（IMADE ルーム）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（複数チャネル）の映像，音声を同時に扱う必要がある．どういったセンサデバ

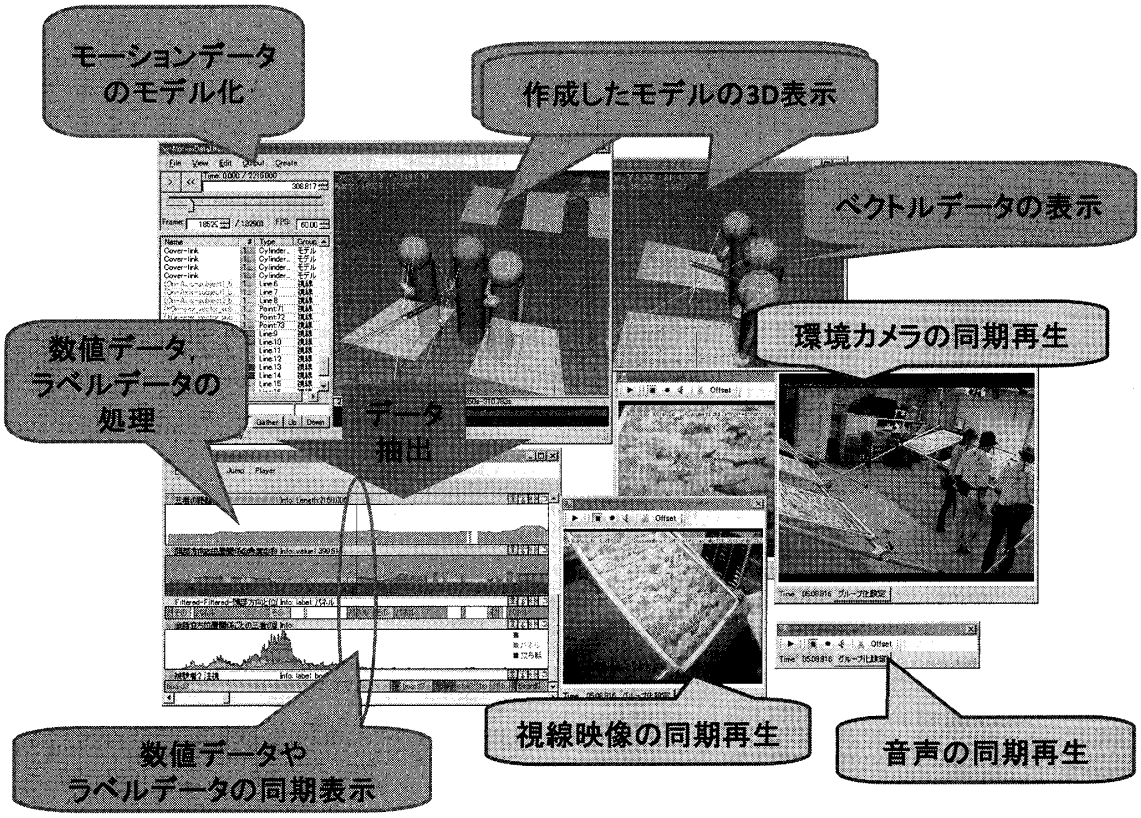

図3　iCorpusStudio：インタラクション・コーパスの閲覧・ラベリング・分析環境

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．

以上のように，iCorpusStudioの特徴は分析作業の支援にあるので，必ずしもラベリング作業までiCorpusStudioで行うことにこだわっていない．実際，筆者ら自身も，音声時区分の切り出しは，よりインタフェースが成熟したWaveSurferを用いることが多い．また，他の分析者がAnvilやELANを用いてラベリングしたデータをiCorpusStudioに読み込んで分析を行う，ということにも対応できるよう，iCorpusStudioでは，ラベルデータ形式としてより汎用的なCSV (comma separated value) 形式を採用している．

4.2　ユーザインタフェース

iCorpusStudioのユーザインタフェースを概説する．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[5]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

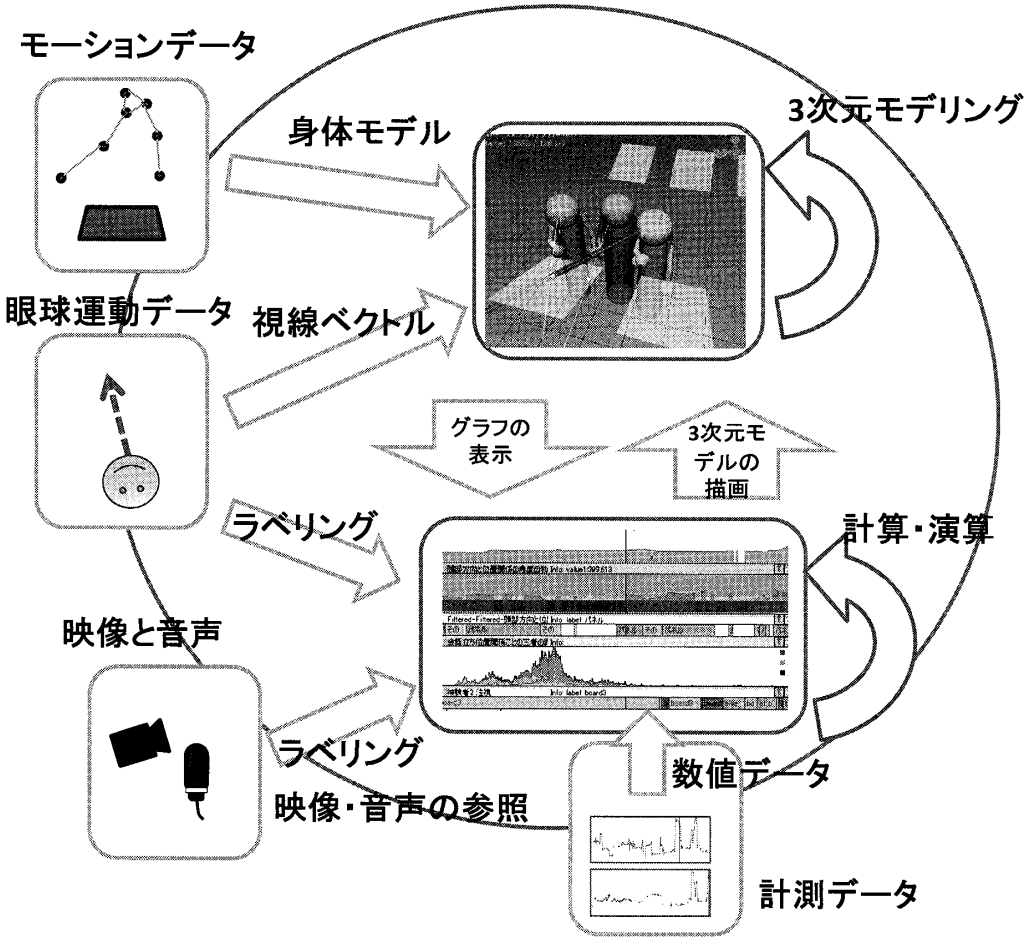

図4　iCorpusStudio を利用したマルチモーダルデータの解釈プロセス

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て分析したり，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．

4.3　ラベリングによるインタラクション解釈の蓄積

以下，iCorpusStudio を使ってマルチモーダルデータを解釈する一連の作業例を，図4を用いて説明する．iCorpusStudio ユーザは，インタラクションの要素やシーンの解釈を時間幅のあるラベルとして書き下すことができる．ビデオや音声データに対してラベル付与する機能自体は，これまでの分析ツールにも存在していた．それらに対して iCorpusStudio の特徴は，計測された数値データから数値計算を行い，半自動的にラベルを付与するための環境を用意していることである．例えば，発話区間をラベリングするには，多くの場合，音声パワーの閾値を設けて発話の有無を判別し，時間方向の細かすぎる空白やノイズを取り除くことで平滑化を施す作業を行う（図2の Interaction Primitive の抽出）．こういった作業では，複数存在するパラメータの調整を繰り返しながら，データ毎に同じことを繰り返す必要がある．そういった一連の作業を支援するために，iCorpusStudio では，閾値を超えた時間区域を抽出してラベルを自動生成したり，ラベル列からのノイ

ズ除去や平滑化といった作業をツールキットとして提供している．

さらにiCorpusStudioの重要な機能として，ラベル間の演算がある．ラベル間の演算を行うことで，図2における，Interaction Primitiveの集合からInteraction Event，さらにはInteraction Eventの集合からInteraction Contextといった，より抽象度の高い解釈を行うことが可能になる．例えば，複数人が同一の対象物に視線を向ける（共同注視），複数人がある物を参照しながら会話をする，といった現象を検出することを考える．検出は，iCorpusStudio上でラベル間の重なりを網羅的に検出することで行う．

しかし，実際にラベル同士が重なっている部分，つまり，異なる非言語行動や別々の会話参加者の行動が同時に発生した部分の抽出だけでは不十分な場合が多い．異なるラベルが連続して発生するパターン（例えば，誰かの発話の後に他の人のあいづちが続く等）を網羅的に抽出したいことがある．ラベル間の重なりにある程度の時間差を許して検索することで，特定のモダリティ間の時間遷移パターンを抽出することが重要なことが多い．iCorpusStudioでは，ラベル間の重なり検出の際に，時間遅れや時間差を設定することができる．

また，他のモダリティのラベルを用いた文脈を考慮し，ある条件の元で発生するラベルだけを抽出したいことがあり，そういったモダリティ間の発生ルールを見いだすことが多くの分析研究者の興味の対象となる．iCorpuStudioでは，そういった仮説の検証作業を支援するために，ラベル間の共起関係の検索結果をルールとして設定し，それらのルールを段階的に組み合わせることで，多層的な条件検索を可能にする機能を提供している．

4.4　モーションデータからの身体動作の可視化

映像・音声は，人の動作や表情などのちょっとした動きから発話時のパラ言語情報など，あらゆる情報を含んでいる．その一方で，多くの情報の中に本質的な情報が埋もれてしまうのも事実である．分析には，ある観点に着目し，それ以外の情報を捨象したモデル化が必要である．

iCorpusStudioには，モーションデータを読み込み，コンピュータグラフィクスによる3次元モデルを可視化するビューワを用意してある．モーションキャプチャシステムから得られるデータ自体は，身体に貼りつけたタグの3次元座標でしかないので，iCorpusStudioではそれらのタグを基点とした球，円柱，円錐といった基本的なモデル定義と，それらの組み合わせによる身体モデル生成を可能にする環境を用意した．

また，アイマークレコーダのデータを読み込んで視線ベクトルを表示したり，上記で生成されたモデル上の任意の2点をつなぐことで，頭部方向，指さし方向，身体の向きなどのベクトルを定義し，半直線として表示することが可能である．

このようにして可視化された3次元モデルを参照することで，任意の方向から身体動作を閲覧することが可能になり，映像と情報を補完し合いながら，データを閲覧することが可能になる．

4.5　数値データを用いた前記号的なインタラクション解釈

上述したような3次元モデルが一旦手に入ると，モデル上の任意の2点間の距離やベクトル間の角度など，元々観測していなかったデータを新たに生成することができる．そのことで，インタラクションの解釈をよる抽象的なレベルに積み上げていくことが可能になる．iCorpusStudioでは，そういった新たに生成された数値データをグラフ表示するビューワを用意することで，解釈の試行錯誤を容易にすることを目指した．

数値データに基づいた動作解釈の一例として，モーションデータからジェスチャ時区間を推定することを考える．図5は，人の胴体と掌の間の距離をグラフ表示している例である．値が高いところは，掌が胴体から離れたことを意味しているので，何らかのジェスチャや指さし行為をしている可能性が高い[6]．閾値を設定してそれよりも高い値の時区間を抽出すれば，ジェスチャを行っている可能性のあるシーンを網羅的かつ機械的に確認することが可能になる．

また，3次元モデル上から計測されたデータをグ

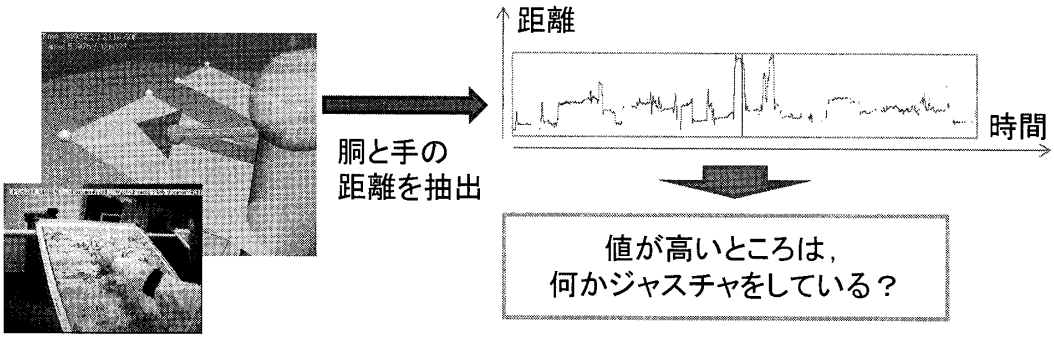

図5　数値データの抽出と仮説生成の例

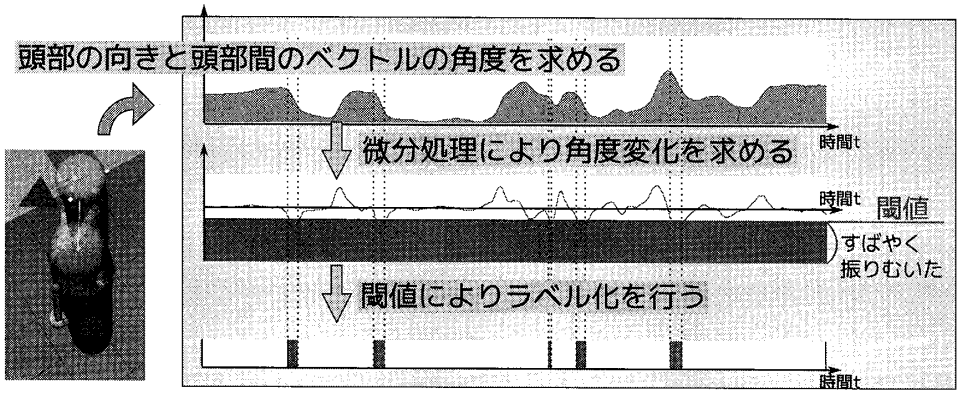

図6　数値データ処理によるラベル生成

ラフ表示するだけでなく，数値データ間の演算をし，それをグラフ表示やラベリングに使うことも可能である．図6はその例である．この例では，他者に対する身体方向に対する頭部方向の角度をグラフ化し，それをさらに微分演算して角速度を求めている．その値が特に低い時間領域を求めることで，素早くその人の方を振り向いた現象を抽出することができる．

数値データに対する演算としては，ほかに四則演算や平滑化などがある．また，複数の数値データ間で各時点の最大の値を持つデータがどれであるかを求めたり，ラベルの値を参照して特定のラベルが付与されている時区間だけを取り出し，その頻度を求めたりすることも可能である．

5.　視線を用いた指さ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．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，指さしジェスチャが頻繁に行われる．指さ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測るのに役立つ．

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，指さ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，指さし行為以外にも人の腕は頻繁に動くし，指さ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，指さ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さし方向に注目することで初めて指さし行為が成立するsocialな現象であると考えた（図7）．つまり，会話参加者が指さし方向の対

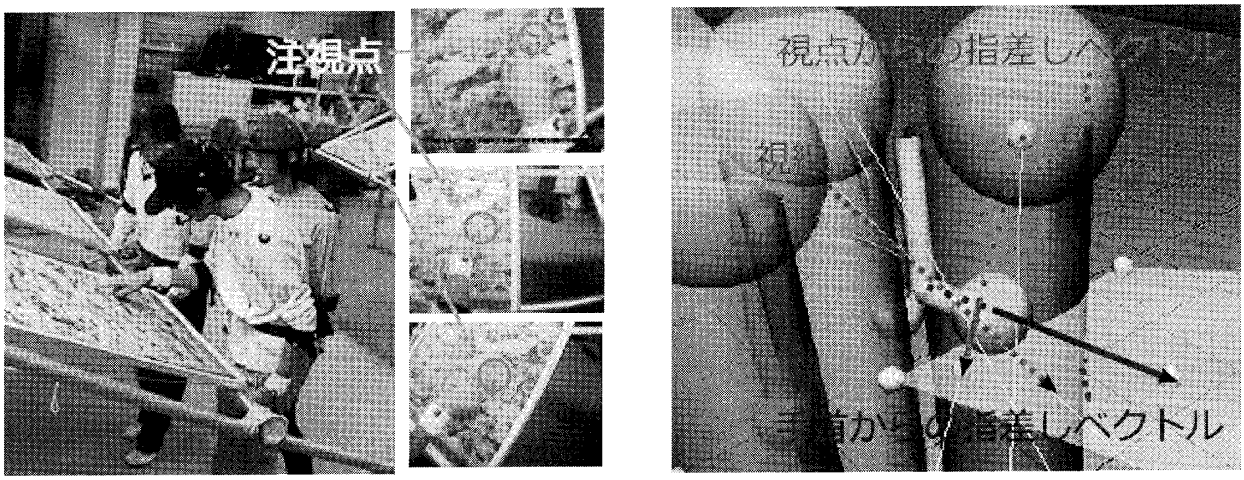

図7　指さしジェスチャの抽出

象物に視線を向ける行為が同期したものだけを指さしジェスチャとして認定することで，指さしジェスチャの検出精度が向上するかを確かめた（矢野ほか，2009）．このことは，近年注目されている，聞き手行動からコミュニケーションを理解しようとする考え方（高梨・榎本，2009）に通じるものである．

以下の手順で指さしジェスチャの自動抽出を試みた．まず，指さしベクトルを定義した．指さしベクトルは，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，目から指先にのばした方向を用いたものの2種類を準備した．次に，指さしベクトルの先に指さ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，指さしジェスチャの候補を広めに抽出した．そして，それらの指さ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さし先の対象物と一致しているかを確認した．

より詳細な分析のプロセスを，図8に基づいて説明する．この例では，指さしの対象は空間内に設置した6枚のパネルに限定した．指さし行為の検出は，モーションセンサから得られた会話参加者の目の中心位置もしくは肘と手首の位置をむすぶベクトルが，パネルが成す面に衝突するかどうかで判定した（図8①，②）．ベクトルの開始点は手首のマーカーの位置とした．注視検出は，アイマークレコーダから得られた視線ベクトルと，指さし検出から得られた指さし先との距離を測り（図8③），500mm以下である場合に指さし対象を注視しているものとし（図8④），1回の指さし行為の中で各会話参加者が50ms以上注視している場合にその指さしがその会話参加者に注視されているものとして計算した(図8⑤)．検出精度の算出は，ハンドラベリングによる指さしジェスチャの正解ラベルと，検出された指さしのラベルの一致・不一致の時間の多寡を用いた（図8⑥）．ある一点を指す指さし行為の場合は手の動きが止まっている時間帯を，手を動かしてある範囲を指している場合にはパネル上で手を動かしている時間帯を指さし行為としてラベル付けした．

結果は図9のようになった．指さしジェスチャの検出精度を求めるために，実際の指さ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．指さしベクトルの判定については，全体を通して，目から指先にのばし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さ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さ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

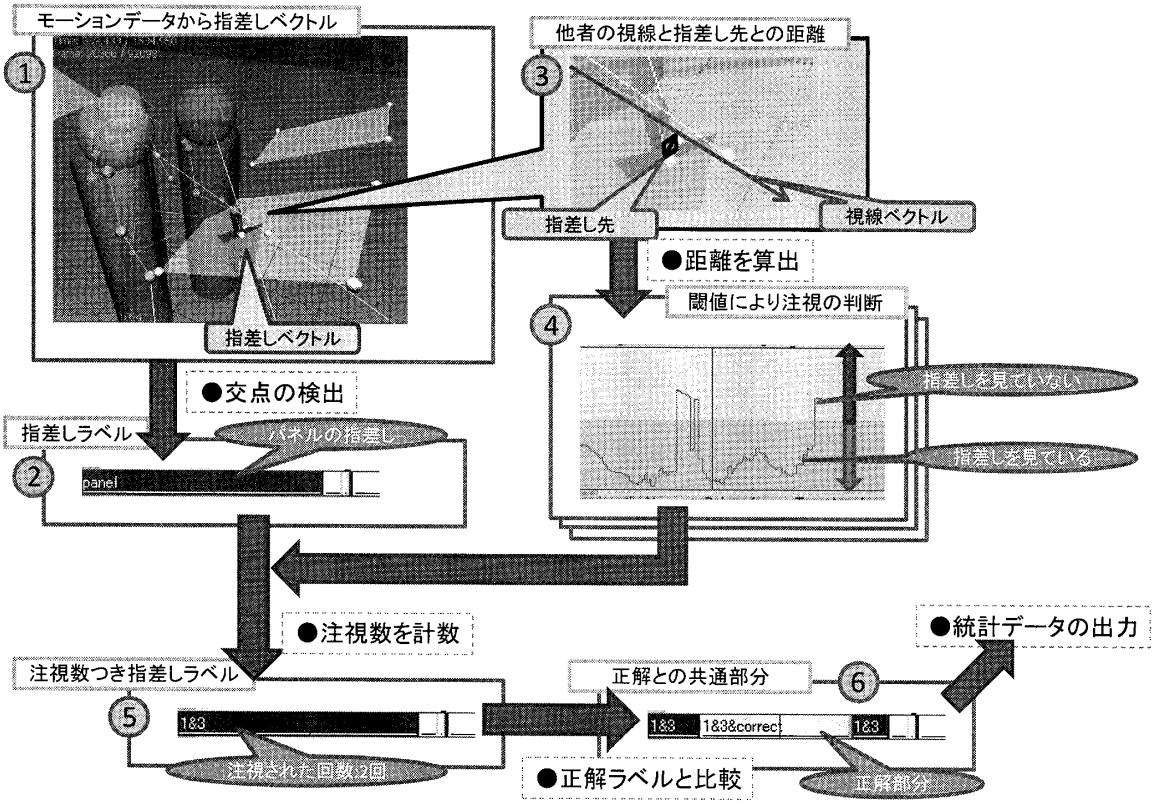

図８　指さしジェスチャ検出の被注視数の計測プロセス

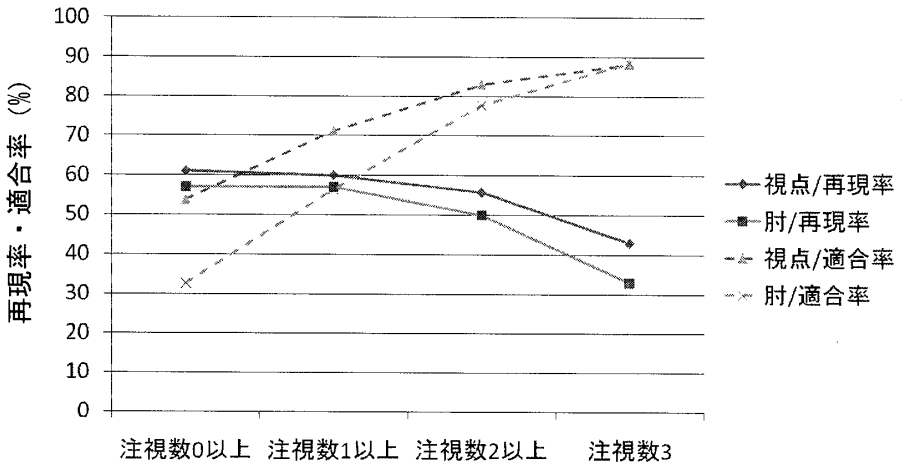

図９　視線獲得による指さしジェスチャ抽出の精度向上

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，実際に取得したデータを目の前にしながら，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

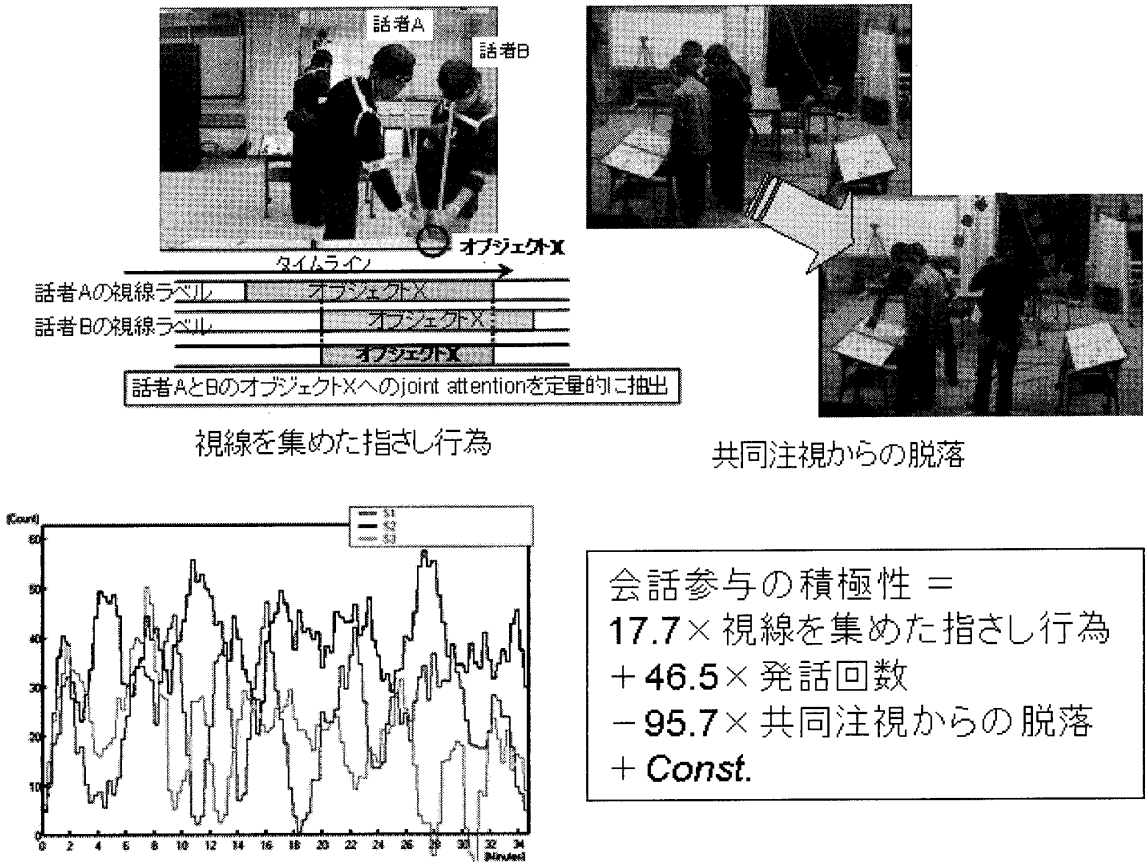

図10　会話参与の積極性の数値化

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ることが期待される．

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与に対する積極性を，非言語情報のみから推定することを試みた（中田・來嶋・角・西田，2008）．

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．その試みのひとつとして，発話，視線変化，指さし行為といったInteraction Primitiveの組み合わせから会話参加者の会話参与積極性を数値化することを試みた（図10）．

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．その会話データを，分析者（筆者ら）が16のシーンに分け，9人の実験協力者に閲覧してもらい，それぞれのシーンについて主観的に評価してもらった値（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，会話参与積極性の正解データとした．

その正解データを用いて，我々の計測環境から得られたInteraction Primitive，Interaction Eventを説明変数とした重回帰分析を行った．具体的な説明変数は，Interaction Primitiveとして，

- 指さし，あるいは，マグネット操作を行った回数
- 発話回数

を，Interaction Eventとして，

- 300ms以上の沈黙の後に発話を行った回数
- ほかの会話参加者の発話にかぶせて発話を行っ

た回数
- ほかの会話参加者に注視された回数
- ほかの会話参加者が注視された状況で指さし・マグネット操作を行った回数
- 共同注視から視線をはずした回数

を採用した．

その結果，従来研究（例えばRienks & Heylen, 2006）と同様に発話回数は積極性に対して強い正の相関を示した．それ以外に，「視線を集めた指さし・マグネット操作」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う．つまり，データ分析的なアプローチで，人の直感に合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. おわりに

多人数会話を記録・分析するための環境構築に関する筆者らの試みを紹介した．発話，視線，身振り手振りといったマルチモーダルなデータを複数人数について同時計測するためのセンサ環境IMADEを紹介した．IMADEで計測されたデータからインタラクションの要素を抽出し，それらの時空間的な解釈を積み上げることでインタラクション・コーパスを構築する方法論を示し，それに基づいた会話構造分析の例を紹介した．

インタラクション・コーパスを用いた分析のためのソフトウェア環境としてiCorpusStudioを紹介し，会話分析に適用した例を示した．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，更にほか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．

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ルームの構築にあたっては河原達也，高梨克也，坊農真弓の諸氏をはじめとする多くの方と議論・協力しながら進めた．iCorpusStudioの初期バージョンは，來嶋宏幸，中田篤志の両氏によって開発された．以上の皆様に深く感謝する．

注

1) http://www.anvil-software.de/
2) http://www.speech.kth.se/wavesurfer/
3) http://www.lat-mpi.eu/tools/elan/
4) http://multimodal-annotation.org/
5) 例えば，しばらく聞き手に回っていた会話参加者が発話者交代によって新しく発話を始める前には，ほかの会話参加者からの視線を獲得していることが多い，といったような仮説．
6) もちろん，この解釈が常に正しいとは限らない．胴体に近い場所でジェスチャを行うこともあるであろうことを考えると，例えば，手のホームポジションを定義し，そこから掌への距離を計るなどの工夫が必要である．しかし，そのためには，そもそも手のホームポジションはどこか，という新しい課題を生む．ここではあくまでもiCorpusStudioの動作例として単純な解釈を例にして紹介している．

【参考文献】

Carletta, Jean, Ashby, Simone, Bourban, Sebastien, Flynn, Mike, Guillemot, Mael, Hain, Thomas, Kadlec, Jaroslav, Karaiskos, Vasilis, Kraaij, Wessel, Kronenthal, Melissa, Lathoud, Guillaume, Lincoln, Mike, Lisowska, Agnes, Mc-Cowan, Iain, Post, Wilfried, Reidsma, Dennis, & Wellner, Pierre (2006). The AMI meeting corpus: A pre-announcement. Second International Workshop on Machine Learning for Multimodal Interaction (MLMI 2005), Vol. 3869. *Lecture Notes in Computer Science*, 28–39. Springer.

Carletta, Jean, Evert, Stefan, Heid, Ulrich, Kilgour, Jonathan, Robertson, Judy, & Voormann, Holger (2003). The NITE XML Toolkit: Flexible annotation for multi-modal language data. *Behavior Research Methods, Instruments, and Computers*, **35**(3) 353–363.

Chen, Lei, Rose, Travis R., Qiao, Ying, Kimbara, Irene, Parrill, Fey, Welji, Haleema, Han, Tony, Tu, Jilin, Huang, Zhongqiang, Harper, Mary, Quek, Francis, Xiong, Yingen, McNeill, David, Tuttle, Ronald, & Huang, Thomas (2006).

VACE multimodal meeting corpus. Second International Workshop on Machine Learning for Multimodal Interaction (MLMI 2005), Vol. 3869. *Lecture Notes in Computer Science*, 40–51. Springer.

Kawahara, Tatsuya, Setoguchi, Hisao, Takanashi, Katsuya, Ishizuka, Kentaro, & Araki, Shoko (2008). Multi-modal recording, analysis and indexing of poster sessions. *Interspeech-2008*. 1622–1625.

來嶋宏幸・坊農真弓・角康之・西田豊明　(2007).　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（ヒューマンコンピュータインタラクション），**2007**(99), 63–70.

Kipp, Michael (2001). ANVIL: A generic annotation tool for multimodal dialogue. *Eurospeech 2001*, 1367–1370.

中田篤志・來嶋宏幸・角　康之・西田豊明　(2008).　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　第22回人工知能学会全国大会

Renals, Steve, Hain, Thomas, & Bourlard, Hervé (2007). Recognition and understanding of meetings the AMI and AMIDA projects. IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU 2007), 238–247.

Rienks, Rutger, & Heylen, Dirk (2006). Dominance detection in meetings using easily obtainable features. Second International Workshop on Machine Learning for Multimodal Interaction (MLMI 2005), Vol. 3869. *Lecture Notes in Computer Science*. 76–86. Springer.

Rose, Travis R., Quek, Francis, & Shi, Yang (2004). MacVisSTA: A system for multimodal analysis. 6th International Conference on Multimodal Interfaces (ICMI 2004). 259–264. ACM.

角康之・西田豊明・坊農真弓・來嶋宏幸　(2008). 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，**49**, 945–949.

Sumi, Yasuyuki, Yano, Masaharu, & Nishida, Toyoaki (2010). Analysis environment of conversational structure with nonverbal multimodal data. In 12th International Conference on Multimodal Interfaces and 7th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2010). ACM.

高橋昌史・角康之・伊藤禎宣・間瀬健二・小暮潔・西田豊明　(2008).　時系列イベント発見のためのグラフクラスタリング手法の提案　情報処理学会論文誌，**49**, 1942–1963.

高梨克也・榎本美香　(2009).　特集「聞き手行動から見たコミュニケーション」の編集にあたって　日本認知科学会誌，**16**(1), 5–11.

田浦健次郎　(2008).　InTrigger：オープンな情報処理・システム研究プラットフォーム　情報処理学会誌，**49**, 939–944.

Waibel, Alexander, & Stiefelhagen, Rainer (Eds.) (2009). *Computers in the human interaction loop*. London: Springer.

矢野正治・中田篤志・福間良平・角　康之・西田豊明　(2009).　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（ユビキタスコンピューティングシステム），**2009**(22).

(2010年12月20日受付)
(2011年7月8日修正版受付)
(2011年7月13日掲載決定)