複数人物の対面会話を対象としたマルコフ切替えモデルに基づく会話構造の確率的推論

大塚和弘[†,†††] 竹前嘉修[††]
大和淳司[†] 村瀬洋[†††]

複数人物による対面会話を対象とし，会話参加者の視線パターン，頭部方向，および，発話の有無に基づき会話の構造の推論を行うための確率的枠組みを提案する．本研究では，まず，会話の構造として，話し手，受け手，傍参与者と呼ばれる参与役割と会話参加者との組合せに着目する．次に，会話中の各人物の行動は，会話の構造によって規定されるという仮説を立て，マルコフ切替えモデルと呼ばれる一種の動的ベイジアンネットを用いた会話モデルを提案する．このモデルは，会話レジームと呼ばれる会話の構造に対応した上位プロセスの状態が，マルコフ過程に従い時間変化しつつ，その会話レジームの状態に依存して，視線パターン，および，発話が確率的に生成され，さらに，各人の視線方向に依存して頭部方向が観測されるという階層的な構造を持つ．このモデルにおいて，会話レジームは，会話中に頻出する視線パターンの特徴的な構造に基づいて仮説的に設定される．また，ギブスサンプリングと呼ばれる一種のマルコフ連鎖モンテカルロ法を用いて，観測された頭部方向と発話の有無の時系列データより，会話レジーム，視線パターン，および，モデルパラメータのベイズ推定を行う方法を提案する．最後に，4人会話を対象とした実験により，視線方向と会話レジームの推定精度を評価し，提案した枠組みの有効性を確認する．

A Probabilistic Inference of Multiparty-conversation Structure Based on Markov-switching Models

Kazuhiro Otsuka,[†,†††] Yoshinao Takemae,[††] Junji Yamato[†] and Hiroshi Murase[†††]

A novel probabilistic framework is proposed for inferring the structure of conversation in face-to-face multiparty communication, based on gaze patterns, head directions, and the presence/absence of utterances. First, as the structure of conversation, this study focuses on the combination of participants and their participation roles. Next, we hypothesize that the structure of conversation governs how people behave during conversation, and propose a conversation model based on the Markov-switching model, a kind of dynamic Bayesian network. In this model, the state of the high-level process, we call it the conversation regime, is assumed to correspond to the conversation structure and that its changes over time exhibit Markov properties. Also, the conversation regime controls the dynamics of utterances and gaze patterns, which stochastically yield measurable head directions. The conversation regimes are hypothetically configured based on typical structures exhibited by gaze patterns among the participants during conversations. Furthermore, a Markov chain Monte Carlo method called the Gibbs sampler is used to realize the Bayesian estimation of conversation regime, gaze pattern, and model parameters from the observed sequential data of head directions and utterances. Finally, experiments on four-person conversations confirm the effectiveness of the proposed framework in estimating gaze directions and conversation regimes.

1. はじめに

複数人物による対面会話は，情報の伝達・共有，他者の意図・感情の理解，グループの意思決定などにおいて，欠かすことができないコミュニケーションの形態の1つである．これまでの対面状況に限定されていた会話の限界を超えて，我々のコミュニケーション能力を拡張するため，遠隔会議システム[1)]，会議映像

† NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories
†† NTT サイバーソリューション研究所
NTT Cyber Solutions Laboratories
††† 名古屋大学大学院情報科学研究科
Graduate School of Information Science, Nagoya University

の自動編集・アーカイブ[2),3)]，人間同士の会話に参加可能なエージェント[4)]/ロボット[5),6)]などを実現する技術の発展に期待が寄せられている．現在，そのための基本技術の1つとして，会話シーンを自動的に認識・理解する技術が求められており，その一環として，著者らは，誰が誰に向かって話しかけているかというような会話の構造の推定を行う技術の確立を目指している．

このような会話シーンの自動的な認識・理解を実現するためには，まず，会話参加者の行動を認識することが必要であると考えられる．従来，人間の行動認識は，コンピュータビジョン分野などにおいて研究が進められており，その対象は，単一の人物の姿勢やジェスチャから，複数の人物間のインタラクションまで拡大している．近年，なかでも人物間のインタラクションが生じる場として，会議や会話シーンに注目が集まりつつあり，隠れマルコフモデルやその拡張であるHMMs[7)]，Layered-HMM[8)]，Coupled-HMM[9)]，動的ベイジアンネットワーク[10)]などを用いて，会議・会話中の各参加者の行動やグループ行動を認識する方法が提案されている．たとえば，McCowanらは，会議中の行動として，プロジェクタ利用の発表，ホワイトボードの使用，メモ書き，ディスカッションなどの認識を目標として，HMMsを用いた手法を提案している[7)]．この方法を含め，これまで提案されている方法では，まず，各人物の行動を認識し，その後，各人の行動の間の直接的な共起関係として人物間のインタラクションをモデル化し，認識を行うというアプローチが主に採用されている．その一方で，会話参加者の行動が生成される背景にある会話現象の性質や構造に着目し，それらと人物の行動との関係についてモデルを構築し，インタラクションの認識を行うというアプローチは，いまだ，ほとんど試みられていない．

従来，複数人物による対面会話は，社会心理学などの分野において研究が行われており，参与枠組みや参与役割と呼ばれる観点から会話の構造をとらえる方法が知られている[11)~15)]．Goffmanの参与枠組み[11)]においては，会話参加者は，承認された参加者と承認されていない参加者（立ち聞きしている者）に分類され，さらに，承認された参加者は，話し手，受け手，および，傍参与者に分類される．話し手，受け手などの役割は参与役割と呼ばれている．ここで，受け手は，話し手に話しかけられている人物を指し，傍参与者は，承認された参加者のうち話し手でも受け手でもない人物を指す．話し手は，発話権（ターン）を保持している者として位置付けられる．また，受け手と傍参与者は聞き手とも呼ばれる．会話の進行とともに参加者間で発話権が移動するにともない，各参加者の参与役割も動的に変化する．このような会話中の各参加者の役割は，会話の構造を表現するための不可欠な要素であると考えられ，その自動的な推定は，様々なアプリケーションにおいて重要な役割を担うものと期待される．たとえば，会議映像の自動編集システムにおいては，参与役割に応じた映像の切替え表示などにより，「誰が誰に話しかけているか」といった情報を視聴者に分かりやすく伝達できるようになると期待される．また，会話ロボットにとってもグループ会話に参加するうえで参与役割の推定は不可欠であることが指摘されている[5),6)]．

本研究では，会話の構造として，このような各参加者の参与役割に着目し，その推定の手がかりとして，以下の観点から参加者の非言語的な行動に着目する．従来，参与役割のうち，話し手については，各参加者の音声信号から各人の発話の有無を判定することで，おおよそ同定が可能であることが知られている．しかし，三者以上の会話においては，聞き手である受け手と傍参与者を音声信号のみから区別することは困難とされている．また，発話者が，ある参加者に話しかける場合，つねに，その相手の名前を発話に含めるとは限らないため，各人の発話に含まれる言語的な情報も，受け手と傍参与者を区別するための手がかりとしては不十分であると考えられる．

その一方，対面状況においては参加者の非言語的な行動も，コミュニケーションを遂行するための重要な要素であることが知られている[16)~18)]．このような非言語行動には，視線や顔の表情，頷き，手振り・身振り，姿勢などがあるが，なかでも視線の役割の重要性は古くより指摘されている[15),19)~21)]．たとえば，Kendonは，視線には，他者の行動をモニタリングする機能，自らの態度や意図を表出する機能，会話の流れを調整する機能が具わっていることを示唆している[19)]．また，Goodwinは，話し手は視線を使って，誰に話しかけているのかを他の参加者に提示し，また，受け手は，その視線を話し手に向けることで，話し手の話を聞いていることを話し手に合図しており，このような話し手と受け手の間の相互の注目が会話の成立のうえで不可欠であると主張している[20)]．また，Vertegaalらは，会話参加者は参与役割を理解するための手がかりとして視線を用いていることを実験的に明らかにしている[22)]．これらの知見に基づき，近年では，会話参加者の視線行動を分析することにより，自動的に参与役割を求めることができるものと期待が高まってい

る[23]．

このように会話中の視線と参与役割との関連性は古くより示唆されているが，参与役割の自動的な推定を実現するためには，両者の関係を定量化する必要があり，そのためには，まず会話中の視線行動を実際に計測する必要があると考えられる．人間の視線方向を自動的に計測する方法は，これまでにもいくつか提案がなされているが[24),25)]，複数人の自然な会話を妨げないように視線方向を正確，かつ，安定に計測することは，いまだ困難なタスクとされている．そのため，これまで視線の代用として，より容易に計測が可能な頭部の姿勢・方向を用いるというアプローチが試みられている[6),26)~28)]．このアプローチは，人間には興味の対象をその視野の中央でとらえようとする性質があり，それにより，視線を向けた先の人物との位置関係に応じて，頭部や胴体の姿勢が変化する，という性質に立脚するものである．このアプローチの先駆的な例として，Stiefelhagen らは，4 人会話において，頭部方向に基づいて，誰が誰を見ているか判定可能であることを示している[26),29)]．また，簡単な会話の状況を対象に，話し手の頭部方向から，その発話が向けられている相手（受け手）を判定する試みも，会話ロボット開発の一環として始まっている[5),6)]．

このような背景をふまえ，我々は，N 人物（$N \geq 3$）による対面会話を対象とし，会話の構造として会話参加者とその参与役割の組合せに焦点をあて，その推定のための確率的枠組みの構築を目指している．その一環として，本論文では，従来研究においてとられてきた，人物行動の直接的な認識・解釈に基づくアプローチとは異なる新しいアプローチを提案する．このアプローチとは，会話の各時点における参加者の行動は，会話の構造（以下，会話構造とも呼ぶ）によって規定されるという仮説を立て，会話という現象を，会話の構造に相当する一種の上位プロセスと参加者の行動という下位のプロセスとの間の相互作用によって時間発展する系と見立てて，確率的なモデルを構築し，このモデルに基づいて会話構造を推定するというものである．ここではこのような会話のモデル（以後，会話モデルと呼ぶ）として，マルコフ切替えモデル（Markov-Switching Model）[30),31)] と呼ばれる一種の動的ベイジアンネットに着目する．マルコフ切替えモデルは，レジームと呼ばれる上位にあるプロセスの状態がマルコフ過程に従って変化しつつ，その状態に依存して下位にあるプロセスのダイナミクスが決定されるという階層的な構造を有する．本論文では，このモデルのレジームを「会話レジーム」と呼び☆，会話の構造に相当するものと仮定する．また，下位のプロセスの状態が参加者の行動に対応すると仮定し，会話レジーム（以下，単にレジームとも呼ぶ）の状態に依存して確率的に参加者の行動が生成されるというモデルの構成を試みる．ここでは，参加者の行動として，会話構造との関連性が示唆されている視線行動（視線の方向）と発話状態（発話の有無）を導入する．また，視線方向は直接的に計測が困難であるため，これもレジームと同様に推定すべき隠れ変数と見なし，その推定の手がかりとして頭部方向を計測の対象とし，その観測過程のモデルを会話モデルに組み込む．

本論文では，このような会話モデルを具体的に構成するため，会話中に出現する視線パターン（各参加者の視線方向の集合）の頻度を分析し，頻出する視線パターンに含まれる特徴的な構造に基づいて，会話レジームと会話構造との関係，および，会話レジームと参加者の行動との関係を仮説的に設定する．また，観測された頭部方向と発話状態に基づいて，会話中の各時刻におけるレジーム状態と視線パターン，および，モデルパラメータのベイズ推定を行うために，ギブスサンプリングと呼ばれる一種のマルコフ連鎖モンテカルロ法[32),33)] を用いる方法を提案する．本論文の提案方法は，頭部方向から視線方法を近似的に推定するという問題と，視線方向と発話状態から会話構造を推測するという 2 つの問題を同時に解く方法ともとらえられる．視線方向に対する頭部方向の自由度より，頭部方向からの視線方向の推定には曖昧性がともなうが，会話モデルを用いることで，会話の構造に応じて出現しやすい視線方向が想定できることから，精度の高い視線方向の推測が可能となると期待される．

また，本論文では 4 人会話を対象として実験を行った結果を示す．まず，会話レジームと人物行動の関係が適切にモデル化されていることを確認するために，視線方向の推定精度を評価する．次に，推定された会話レジームが実際の会話構造を反映していることを確認するため，発話の種別・方向性のラベルを用いた評価を行う．これら結果に基づいて，提案した会話モデル，および，会話構造の推定方法の有効性を確認する．

本論文は以下のように構成される．2 章において，視線パターンの分析に基づいて会話レジームの設定を行う．3 章においては，会話モデルの定義を行い，会

☆ レジーム（regime）という用語は，一般的には政治体制，気候変動の型などを指す場合に使われる．本論文では会話構造によって参加者の行動が規定されるという会話モデルの構造を含意させるためにこの用語を用いることとした．

話レジームのベイズ推定を行うアルゴリズムを提案する．4章では実験結果を示し，提案したモデル，および，推定法の有効性を検証する．さらに，5章において議論を行い，6章では本論文のまとめを示す．

2. 会話レジームの設定

本章では，会話モデルに含まれる会話レジームの状態を具体的に設定し，会話レジームと会話構造との関係，および，会話レジームと参加者の行動との関係の仮説的な設定を試みる．1章で述べたように，会話レジームとは会話構造に相当するものであり，また，会話レジームの状態に依存して，参加者各人の行動や参加者間のインタラクションが確率的に生成されるものと仮定する．また，発話権の交替にともなう参与役割の変化など，会話のダイナミクスは，会話レジームの状態遷移として表されるものと考える．

このような会話レジームの設定を行うため，ここではまず，参加者の行動として，会話構造との関連性が示唆されている視線行動に着目する．また，会話構造によって参加者行動が規定されるという仮説より，同じ会話構造が継続する区間においては，同種の視線行動が観測されると考え，参加者の視線方向（視線パターンと呼ぶ）の出現頻度を分析する．続いて，その結果から，出現頻度が高く，かつ，長時間継続するような視線パターンに対応する会話構造を考察することで，会話ジレームの状態の設定を試みる．ここで視線パターンとは，ある時刻における全参加者の視線方向の集合を指す．本研究では，参加者が N 人の場合，各参加者の視線方向は，他の参加者のうちいずれか1人の顔に向けられているか，あるいは，誰の顔からも視線を逸らしているかという N 個の離散的かつ排他的な状態のいずれかにあるものと考える☆．したがって，視線パターンは合計で N^N 通り存在しうる．なお，本論文では，参加者は着席しており，会話中の参加者の増減や移動がないことを前提とする．

なお，本研究では，話し手の話しかけや，聞き手の傾聴に際して，視線を向けるという行為がともなうことを仮定している．ただし，そのような場合，実際の会話では，参加者は対象への凝視と凝視回避を繰り返すことから，同一の会話構造が継続する状況においても，1つの視線パターンが継続することは期待できない．そこで，このような個々の人物の視線方向の変化に影響を受けずに，会話構造を示唆するような安定した視線パターンの特徴を抽出するため，次のような分析を行った．

☆ 各参加者が注意を向ける先を限定するため，ノートや黒板などの道具は使用しないことを想定している．

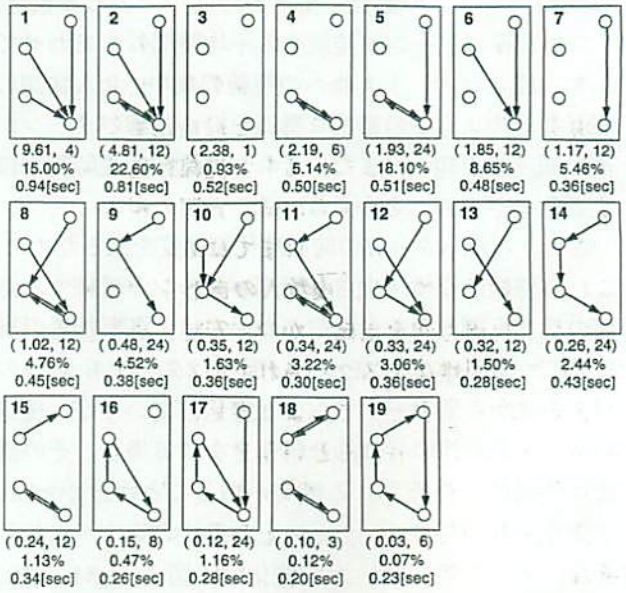


図1 4人会話における視線パターンの出現頻度（会話データ G1-C1 より算出．4.1 節参照）．視線パターンは有向グラフで表現（頂点＝人物，辺＝視線方向，出力辺がない頂点は，視線を逸らしている人物を表す）．同型グラフのクラスに対して相対頻度の大きい順番に 1 から 19 まで番号を付与．各グラフの下部の3行の数値は，上から（チャンスレベルに対する相対頻度，この同型グラフに分類される視線パターン数），全時刻中の（絶対）出現頻度 [%]，平均継続時間長 [秒]

Fig. 1 Frequency of gaze patterns in 4-person conversations (calculated from data G1-C1. See Section 4.1). Gaze pattern is represented as directed graph (Node = person, edge = gaze direction, node without outgoing edge = averted gaze). Gaze patterns are clustered into same isomorphic graph category, which is sorted in descending order of relative frequence. Their statistics are given in the three columns below each graph: (relative frequency to chance level, number of isomorphic graphs), frequency [%], average duration time [sec].

図1には，ある4人会話（4.1 節のデータ G1-C1）を対象として，人手により検出された視線方向に基づいて，視線パターンの出現頻度を分析した結果を示す．図1において，視線パターンは有向グラフとして表されている．ただし，グラフの頂点は各参加者，辺は視線方向を表す．図1に示した19種類のグラフは，人物の入れ替えに不変な視線パターンの構造に着目するため，同型グラフの関係にある視線パターンを1つのグラフとしてまとめたものである．また，これらは，チャンスレベルに対する相対出現頻度の大きい順番に並べられている．ここで，ある2つのグラフが同型であるとは，一方のグラフを他方のグラフに変換するような頂点集合間の1対1写像が存在する場合のことをいう[34]．図1には，各グラフについて含まれる異な

る視線パターンの個数，全体の時間中で占める割合，1つの視線パターンが持続する平均時間長をあわせて示す．図1より，1人物への視線の集中，2人物間の相互凝視のような特徴的な構造を持つ視線パターンが高い頻度で出現し，また，それらは比較的長時間継続する傾向があることが分かった．

また，視線パターンの時間変化の性質を探るために，これら特徴的な構造を含む視線パターンが継続する時間の長さを調べた．会話中においては，各参加者の視線方向が変化するに従って，視線パターンも1つのパターンから別のパターンへと変化していくが，視線パターンが前述の特徴的な構造を有する場合，その構造は時間的に維持される傾向があることが分かった．つまり，ある人物への2人以上の視線の集中が生じる場合，その視線パターンの変化は，図1の#1，#2，#5，#6のような同一人物への視線集中を示すパターン間の遷移にとどまる傾向があり，同様に，2者間での相互凝視の場合，#2，#4，#5，#8のように同じ2者間での相互凝視を持つパターン間の遷移が起こりやすいことが分かった．これら2種類の特徴を持つ構造が継続する時間の長さ（平均継続時間長）を計算したところ，前者については1.76 [sec] となり，後者は1.33 [sec] となった．これは，1つの視線パターンが継続する平均的な時間（=0.52 [sec]）と比較しても長く，これら視線パターンの特徴的な構造は時間的に安定しているといえる．本論文では，このような視線パターンの構造を作り出すメカニズムと，会話の構造との間には強い関連性があるものと予想し，視線パターンの構造に基づいて，以下のように1者集中，2者結合，分散と呼ぶ3つの会話レジームのクラスを仮説的に設定する．

最初に，「1者集中」（convergence）と呼ぶレジームは，図1の#1，#6のように，参加者の視線が1人の人物に集中する視線パターンをもたらすような会話の構造に対応する．このレジームにおいては，この最も多くの視線を受けている人物（中心人物と呼ぶ）が話し手となり，他の人物が受け手になると想定される．そのため，中心人物が，他の参加者に向かって主に発話を行い，受け手はその発話を聞いており，受け手の発話は，相槌などに限定されると考えられる．また，このレジームにおける情報伝達のパターンは，中心人物から他の人物への1対多の1方向性であり，いわゆる，モノローグと呼ばれる会話の型に対応すると考えられる．ここで，このレジームを R_i^C と記す．ただし，i は中心人物を指す．N 人会話の場合，N 通りの1者集中レジームの状態 $\boldsymbol{R}^C = \{R_i^C\}_{i=1}^N$ が存在しうる．このレジームにおいて生じる視線集中のパターンは，Goodwinの主張する，話し手と受け手の間の相互の注目は会話において不可欠な要素である，という性質[20]を反映したものであると考えられる．なお，話し手に対してまったく視線を向けないような人物が存在する状況（考えごとをしている．聞き耳を立てているなど）も考えられるが，そのような状況においても，他の参加者が話し手へ注目している場合には，その会話の場を代表する会話構造として，1者集中レジームが該当するものと考える．

次に，「2者結合」（dyad-link）と呼ぶレジームは，図1の#4，#8のように，2人の人物が互いを見ている，つまり，相互凝視を含む視線パターンが生じるような会話の構造に対応する．このレジームにおいては，この相互凝視の関係にある2者間に限定された情報交換が行われ，この2者が話し手または受け手となり，他の参加者は傍参与者となるものと想定される．そのため，この2者は高い確率で発話を行い，他の参加者の発話は限定的であると仮定する．このレジームは，ダイアローグと呼ばれる会話の型に対応すると考えられる．一方，2者間の相互凝視は，発話交替の合図としても機能することより[15),19),35)]，発話交替時にも瞬間的にこのレジームが出現すると考えられる．このレジームを $R_{(i,j)}^{DL}$ と表記する．ただし，(i,j) は相互凝視の関係にある2者を表す．N 人会話においては，${}_NC_2$ 通りの2者結合レジームの状態 $\boldsymbol{R}^{DL} = \{R_{(i,j)}^{DL} | i = 1, \cdots, N; j = i+1, \cdots, N\}$ が存在しうる．

最後に，「分散」（divergence）と呼ぶレジームは，上の2つのレジームに該当しない視線パターンが生じるような会話の構造に対応する．つまり，図1の#3，#9，#13のように，各参加者が，別々の人物を見ていたり，視線を逸らしているような場合に該当する．このレジームにおいては，グループとしての会話は生じていないと想定される．そのため，各参加者が発話を行う確率も低いと考えられる．また，このレジームは，会話の開始前や，会話中の話題の切れ目などにしばし現れると考えられる．このレジームを R^0 と記す．

以上のように，提案する会話モデルにおいては，各レジームについて，対応する会話の構造，および，出現する参加者行動（視線パターン，発話状態）の傾向が仮定される．これらの傾向は，確率分布として会話モデルのパラメータとして表現され，推定の対象となる（3章参照）．また，各レジームが想定する性質を持つようにあらかじめ事前確率分布がモデルパラメータに対して設定される（4.2節参照）．

3. 会話モデルと推定アルゴリズム

本章では，会話中の各時刻における会話レジームの状態を推定するために，マルコフ切替えモデル[30),31)]☆に基づく会話モデルを提案し，続いて，このモデルに基づく推定のアルゴリズムを提案する．

3.1 会話モデルの構造

図 2 は，提案する会話モデルの構造をグラフとして図示したものである．図 2 において，各頂点は状態変数を表し，各辺は状態変数間の依存関係を表す．このモデルには，隠れ状態変数として，会話レジームの状態の時系列 $\boldsymbol{S}_{1:T}=\{S_1,S_2,\cdots,S_T\}$，および，視線パターンの時系列 $\boldsymbol{X}_{1:T}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_T\}$ が含まれる．ただし，本モデルでは，$t=1$ から $t=T$ までの離散時間区間をモデル化の対象とし，この区間内で N 人物による会話が行われているものとする．ある時刻 $t\in\{1,\cdots,T\}$（時間ステップとも呼ぶ）のレジームの状態 S_t は，2 章で定義された $M(=N+{}_NC_2+1)$ 個のレジームのうち，いずれか 1 つの状態 $S_t=R\in\boldsymbol{R}=\boldsymbol{R}^C\cup\boldsymbol{R}^{DL}\cup R^0$ をとる．また，時刻 t の視線パターン $\boldsymbol{X}_t$ は，各人物 $i\in\{1,\cdots,N\}$ の視線方向 $X_{i,t}$ の集合 $\boldsymbol{X}_t=\{X_{1,t},X_{2,t},\cdots,X_{N,t}\}$ として定義される．ここで，人物 i が人物 j $(\neq i)$ の顔を見る場合を $X_{i,t}=j$ と表し，人物 i がいずれの人物の顔からも視線を逸らしている場合を $X_{i,t}=i$ と表す．

また，このモデルには，観測可能なデータとして，頭部方向の時系列 $\boldsymbol{H}_{1:T}=\{H_t\}_{t=1}^T$，および，発話状態の時系列 $\boldsymbol{U}_{1:T}=\{U_t\}_{t=1}^T$ が含まれる．時刻 t の頭部方向 H_t は，各参加者の頭部方向の集合 $H_t=\{h_{i,t}\}_{i=1}^N$ として定義される．ただし，$h_{i,t}$ は，時刻 t における人物 i の頭部方向を表し，図 3 (a) のように世界座標軸 X と顔面正面の方向とのなす方位角として計測できるものと考える☆☆．また，時刻 t の発話状態 U_t は，各参加者の発話状態の集合 $U_t=\{u_{i,t}\}_{i=1}^N$ として定義される．ただし，$u_{i,t}$ は，人物 i の発話状態を表し，時刻 t において人物 i が発話をしている場合は $u_{i,t}=1$，沈黙している場合には $u_{i,t}=0$ の値をとるものとする．

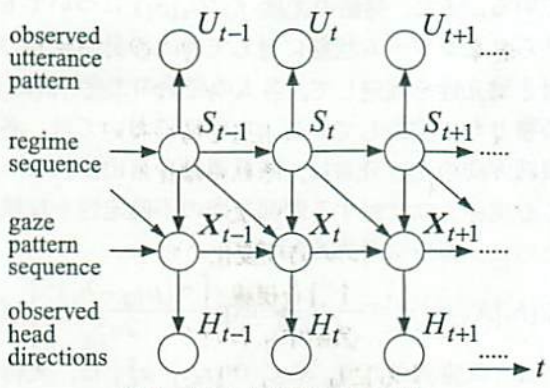
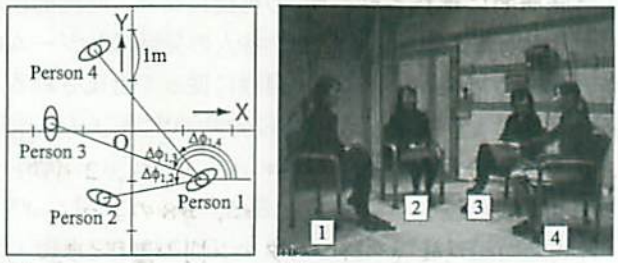


図 2 提案した会話モデルの構造
Fig. 2 Structure of proposed conversation model.



(a) (b)

図 3 会話シーンの概略．(a) 参加者の配置（平面図），(b) 参加者の全体ショット（グループ G1）
Fig. 3 Overview of scene. (a) plan view of participants' location, (b) whole view of participants (Group G1).

このモデルの同時確率分布は，

$$p(\boldsymbol{X}_{1:T},\boldsymbol{S}_{1:T},\boldsymbol{Z}_{1:T},\varphi)\propto f(\boldsymbol{Z}_{1:T}|\boldsymbol{X}_{1:T},\boldsymbol{S}_{1:T},\varphi)\cdot p(\boldsymbol{X}_{1:T}|\boldsymbol{S}_{1:T},\varphi)\cdot p(\boldsymbol{S}_{1:T}|\varphi)\cdot p(\varphi)\quad(1)$$

のように観測データ $\boldsymbol{Z}_{1:T}=\{\boldsymbol{H}_{1:T},\boldsymbol{U}_{1:T}\}$ についての尤度関数 $f(\cdot)$ と事前分布（右辺第 2 項以降）の積として表される．ここで，φ はモデルパラメータの集合を表し（詳細は後述），以降，記述の明瞭性のため必要がない限りこれを省略する．

式 (1) において，観測データ $\boldsymbol{Z}_{1:T}$ についての尤度関数 $f(\cdot)$ は，時刻間の観測データの独立性，および，レジーム状態が与えられたときの頭部方向と発話状態の条件付き独立性を仮定して，

$$\begin{aligned}f(\boldsymbol{Z}_{1:T}|\boldsymbol{X}_{1:T},\boldsymbol{S}_{1:T})&:=\prod_{t=1}^T f_H(H_t|\boldsymbol{X}_t)\cdot f_U(U_t|S_t)\\&:=\prod_{t=1}^T\prod_{i=1}^N f_h(h_{i,t}|X_{i,t})\cdot f_u(u_{i,t}|S_t)\end{aligned}\quad(2)$$

のように定義される．ただし，ここでは，視線パターンが与えられたときの各人の頭部方向の条件付き独立性を仮定し，頭部方向の観測の尤度 $f_H(H_t|\boldsymbol{X}_t)$ を各人の頭部方向の尤度 $f_h(h_{i,t}|X_{i,t})$ の積として定義し

☆ マルコフ切替えモデルは，Interactive Multiple Model, Switching Dynamical System, Jump-Markov Model などとも呼ばれている．また，一般的には，下位の隠れ変数は連続状態をとるが，本論文では，離散状態である視線パターンに拡張して考え，この名称を採用した．

☆☆ 各参加者の頭部は，床面に平行な面上に位置すると考え，各人の視線方向に応じて変化する頭部方向は，この面上の成分のみで十分に表現できるものと考えた．なお，本論文で提案する会話モデルは，この 1 次元の頭部方向に限定されるものではなく，3 次元頭部方向にも容易に適用可能である．

ている．また，発話の尤度 $f_U(U_t|S_t)$ についても，与えられたレジーム状態に対して各人の発話状態の条件付き独立性を仮定して，各人の発話の尤度 $f_u(u_{i,t}|S_t)$ の積として定義している．本論文においては，各人の頭部方向の尤度分布は，眼球運動の自由度によって生じる視線方向に対する頭部方向の不確定性を反映するために，ガウス分布を用いて，

$$f_h(h_i|X_i=j) := \frac{1}{\sqrt{2\pi\sigma_{i,j}^2}}\exp\left[-\frac{(\mu_{i,j}-h_i)^2}{2\sigma_{i,j}^2}\right] \quad (3)$$

のように定義される．ここで $\mu_{i,j}$，$\sigma_{i,j}^2$ は，人物 i が方向 j を見るときの頭部方向の尤度分布の平均と分散を表すパラメータである．これらパラメータは会話中，参加者の座席移動はないとの前提により，時間的に不変であると仮定する．また，各人の発話はレジームの状態に依存したベルヌーイ過程に従って生成されると仮定し，式 (2) 中の各人物の発話の尤度 $f_u(u_{i,t}|S_t)$ は，$f_u(u_{i,t}=1|S_t=R)=\eta_{R,i}, f_u(u_{i,t}=0|S_t=R)=1-\eta_{R,i}$ のように定義される．ただし，パラメータ $\eta_{R,i}$ は，レジーム $S_t=R$ において人物 i が発話を行う確率を表す．

式 (1) 中の右辺第 2 項 $p(\boldsymbol{X}_{1:T}|\boldsymbol{S}_{1:T},\varphi)$ は，レジームの状態に依存して視線パターンが生成される条件付き確率分布を表しており，与えられたレジーム状態のもと各人の視線方向の条件付き独立性を仮定し，

$$\begin{aligned} &p(\boldsymbol{X}_{1:T}|\boldsymbol{S}_{1:T}) \quad (4)\\ &:= p(\boldsymbol{X}_1|S_1)\cdot\textstyle\prod_{t=2}^{T} p(\boldsymbol{X}_t|\boldsymbol{X}_{t-1},S_t,S_{t-1})\\ &:= \textstyle\prod_{i=1}^{N}\left[g(X_{i,1}|S_1)\cdot\prod_{t=2}^{T} p(X_{i,t}|X_{i,t-1},S_t,S_{t-1})\right]\\ &\propto \textstyle\prod_{R\in\boldsymbol{R}}\prod_{i=1}^{N}\prod_{j=1}^{N}\left[\theta_{R,i,0,j}^{m_{R,i,0,j}}\times\prod_{k=1}^{N}\theta_{R,i,k,j}^{m_{R,i,k,j}}\right] \end{aligned}$$

のように定義する．式 (4) において，$p(X_{i,t}|X_{i,t-1},S_t,S_{t-1})$ は人物 i の視線方向が前時刻 $t-1$ において $X_{i,t-1}$ であった場合，現時刻 t における視線方向が $X_{i,t}$ となる確率分布を表す．本論文では，これを

$$\begin{aligned} &p(X_{i,t}|X_{i,t-1},S_t,S_{t-1})\\ &\quad\propto g(X_{i,t}|S_t)\cdot w(X_{i,t}|X_{i,t-1},S_{t-1}) \quad (5)\end{aligned}$$

のように生成重み $g(X_{i,t}=j|S_t=R)=\theta_{R,i,0,j}$ と遷移重み $w(X_{i,t}=j|X_{i,t-1}=k,S_{t-1}=R)=\theta_{R,i,k,j}$ の積に比例するものとして定義する．ここで，生成重み $\theta_{R,i,0,j}$ は，あるレジーム R において，人物 i が方向 j を見る傾向の強さを表し，遷移重み $\theta_{R,i,k,j}$ は，人物 i がその視線を方向 k から j に変化させる傾向の強さを表すパラメータである．ただし，$\sum_{j=1}^{N}\theta_{R,i,k,j}=1,\ \forall k\in\{0,1,\cdots,N\}$ とする．また，式 (4) 中の $m_{R,i,k,j}$ $(k\neq 0)$ は，レジーム状態 R において，人物 i が視線を方向 k から j に変化させた回数を表し，$m_{R,i,0,j}$ は，レジーム状態が R の場合に，人物 i が方向 j を見た時間ステップの総数を表す．以上の視線パターンに関するパラメータをまとめて，$\Theta=\{\boldsymbol{\theta}_{R,i,k}|R\in\boldsymbol{R};i=1,\cdots,N;k=0,\cdots,N\}$，$\boldsymbol{\theta}_{R,i,k}=\{\theta_{R,i,k,j}\}_{j=1}^{N}$ と記す．

また，式 (1) 中のレジーム系列の事前分布 $p(\boldsymbol{S}_{1:T}|\varphi)$ は，レジームのダイナミクスが 1 次のマルコフ過程であると仮定することにより，

$$\begin{aligned} &p(\boldsymbol{S}_{1:T}) := p(S_1)\cdot\textstyle\prod_{t=2}^{T} p(S_t|S_{t-1}) \quad (6)\\ &:= \textstyle\prod_{R\in\boldsymbol{R}}\left[{\pi_{0,R}}^{\delta_R(S_1)}\times\prod_{R'\in\boldsymbol{R}}{\pi_{R,R'}}^{n_{R,R'}}\right]\end{aligned}$$

のように定義することができる．式 (6) において，レジーム状態の初期確率は，$p(S_1=R)=\pi_{0,R}, R\in\boldsymbol{R}$ として表される．また，時刻 $t-1$ から t において，レジーム状態が R から R' に変化する遷移確率は $p(S_t=R'|S_{t-1}=R)=\pi_{R,R'}$ として表される．ただし，$\sum_{R'\in\boldsymbol{R}}\pi_{R,R'}=1,\ \forall R\in 0\cup\boldsymbol{R}$ とする．式 (6) において，$\delta_R(S)$ は，$S=R$ のときに 1 をとり，その他の場合 0 をとる関数である．また，$n_{R,R'}$ は，レジームが R から R' に変化する回数を表す．これらのレジームに関するモデルパラメータをまとめて，$\Pi=\boldsymbol{\pi}_0\cup\{\boldsymbol{\pi}_R|R\in\boldsymbol{R}\}$，$\boldsymbol{\pi}_R=\{\pi_{R,R'}|R'\in\boldsymbol{R}\}$ のように記す．

以上で説明したモデルパラメータをまとめて，$\varphi=\{\Pi,\Theta,\{\mu_{i,j}\}_{i,j},\{\sigma_{i,j}^2\}_{i,j},\{\eta_{R,i}\}_{R,i}\}$ と表す．本論文では，これらのパラメータは時間的に変化しないものと仮定する．また，式 (1) 中のパラメータの事前分布 $p(\varphi)$ は，それぞれのパラメータの独立性を仮定し，各々のパラメータの事前分布の積として定義する (3.2.1 項参照)．

3.2 ギブスサンプリングによるベイズ推定

前節において定義されたモデル構造に基づき，観測データ $\boldsymbol{Z}_{1:T}$ が与えられたときの，レジームの状態系列 $\boldsymbol{S}_{1:T}$，視線パターンの系列 $\boldsymbol{X}_{1:T}$，および，モデルパラメータ φ の推定を行う．本論文ではベイズ流のアプローチ[36)~38)] を採用し，すべての未知変数についての同時事後確率分布 $p(\boldsymbol{S}_{1:T},\boldsymbol{X}_{1:T},\varphi|\boldsymbol{Z}_{1:T})$ の推定を行う．しかしながら，この事後分布を厳密に計算することは，モデルの複雑性の観点から困難であるため，本論文では，ギブスサンプリング[30)~33),39)] と呼ばれる一種のマルコフ連鎖モンテカルロ法を用いる．ギブスサンプリングとは，各々の未知変数について全条件付き事後分布からのランダムサンプリングを繰り返し行い，その過程で生成されるサンプル値の系列を用いて同時事後分布を近似的に求める手法である．これまで，ギブスサンプリングを隠れマルコフモデルのベイズ推定へ適用する方法が提案されており[31),39)]，本論

文ではそれに倣い以下のように推定を行う．

3.2.1 事前分布

まず，モデルに関する事前知識を，モデルパラメータの事前分布として設定し，各々のパラメータについて，それぞれの事前分布からのサンプリングにより初期値を設定する．本論文では，その事前分布として自然共役事前分布を採用する．その理由としては，i) 事前分布と事後分布とが同じ分布族に属するため，事後分布の導出が容易である，ii) 多様な分布が表現できる，iii) ハイパーパラメータの解釈が容易である，といった点[36)~38)]があげられる．

レジーム状態の初期確率 π_0，および，遷移確率 π_R の事前分布としては，隠れマルコフモデルの推定において一般的に用いられているディリクレー分布 $\mathcal{D}(\pi_R|\alpha_R)$ をそれぞれ採用する．ここで，ディリクレー分布は，$\mathcal{D}(\pi_R|\alpha_R) := c\cdot\prod_{R'\in\boldsymbol{R}}\pi_{R,R'}^{\alpha_{R,R'}-1}$ のように定義される．ただし，$\alpha_R=\{\alpha_{R,R'}\}_{R'\in\boldsymbol{R}}$ はディリクレー分布のパラメータであり，条件 $\alpha_{R,R'}>0$ を満たす．また，c は，正規化のための係数を表す．また，視線パターンの生成重み/遷移重み $\theta_{R,i,k}$ の事前分布についても，それぞれ独立なディリクレー分布 $\mathcal{D}(\theta_{R,i,k}|\beta_{R,i,k})$ を用いる．ただし，$\beta_{R,i,k}=\{\beta_{R,i,k,j}\}_{j=1}^{N}$ はそのパラメータを表す．また，発話確率 $\eta_{R,i}$ は，ベルヌーイ過程の仮定により，その事前分布として知られている[36)~38)]ベータ分布 $\mathcal{B}e(\gamma_{R,i,0},\gamma_{R,i,1})$ を用いる．ただし，$\gamma_{R,i,0},\gamma_{R,i,1}$ はベータ分布のパラメータを表す．

頭部方向の尤度分布のパラメータの事前分布は，一般的なガウス分布のベイズ推定[36)~38)]に倣い，平均値 $\mu_{i,j}$ についてはガウス分布 $N(\phi_{i,j},\xi_{i,j}^2)$ を仮定し，分散 $\sigma_{i,j}^2$ については逆カイ二乗分布 $\chi^{-2}(\nu_{i,j},\lambda_{i,j})$ を採用する．ここで，$\phi_{i,j}$，$\xi_{i,j}^2$ は，ガウス分布の平均と分散をそれぞれ表し，$\nu_{i,j}$ と $\lambda_{i,j}$ は，逆カイ二乗分布の自由度と尺度パラメータをそれぞれ表す．

これらのモデルパラメータの初期値は，それぞれの事前分布からの乱数発生により設定される．また，レジームの状態系列，および，視線パターンの系列の初期値は，各モデルパラメータの初期値を用いて，式 (6)，式 (4) で表されるモデル構造に従い生成される．

3.2.2 全条件付き事後分布

ギブスサンプリングでは，各未知変数について，各々の全条件付き事後分布からのサンプリングを行い，逐次的に各変数の値を新しいサンプル値に置き換えるという処理を反復的に実行する．各変数についての全条件付き事後分布は，その変数以外のすべての変数の値が与えられたという条件のもとでの，その変数についての事後分布であり，ベイズ則に基づき各変数に関連した事前分布と尤度の積から以下のように導出できる．

まず，レジーム状態の遷移確率 π_R の全条件付き事後分布は，式 (1) に表されるモデルの構造を用いて，

$$\begin{aligned}&p(\pi_R|\boldsymbol{S}_{1:T},\boldsymbol{X}_{1:T},\varphi\setminus\pi_R,Z_{1:T})\\&\propto p(\boldsymbol{S}_{1:T}|\varphi)\cdot p(\varphi)\\&\propto\textstyle\prod_{R'\in\boldsymbol{R}}(\pi_{R,R'})^{n_{R,R'}}\cdot\mathcal{D}(\pi_R|\alpha_R)\\&\propto\textstyle\prod_{R'\in\boldsymbol{R}}(\pi_{R,R'})^{n_{R,R'}+\alpha_{R,R'}-1}\\&\propto\mathcal{D}(\pi_R|\alpha_R^*)\end{aligned}\tag{7}$$

のように導出される．ただし，$\alpha_R^*=\{\alpha_{R,R'}^*\}_{R'\in\boldsymbol{R}}$ は，$\alpha_{R,R'}^*=\alpha_{R,R'}+n_{R,R'}$ により与えられる．また，レジーム状態の初期確率に関する全条件付き事後分布としては，$\mathcal{D}(\pi_0|\alpha_0^*)$，$\alpha_{0,R'}^*=\alpha_{0,R'}+n_{0,R'}$ を用いる．ただし，$n_{0,R'}$ は，レジーム状態が R' であった総時間ステップ数を表す．

また，視線パターンの生成重み・遷移重み $\theta_{R,i,k}$ の全条件付き事後分布は，レジームの場合と同様に導出でき，

$$\begin{aligned}&p(\theta_{R,i,k}|\boldsymbol{S}_{1:T},\boldsymbol{X}_{1:T},\varphi\setminus\theta_{R,i,k},Z_{1:T})\\&\propto\mathcal{D}(\theta_{R,i,k}|\beta_{R,i,k}^*)\end{aligned}\tag{8}$$

のように定義できる．ただし，パラメータ $\beta_{R,i,k}^*=\{\beta_{R,i,k,j}^*\}_{j=1}^{N}$ は，$\beta_{R,i,k,j}^*=\beta_{R,i,k,j}+m_{R,i,k,j}$ として与えられる．

発話確率 $\eta_{R,i}$ の全条件付事後分布は，その事前分布であるベータ分布が，2 変数の場合のディリクレー分布と等価であることから，レジームの場合と同様にして，$p(\eta_{R,i}|\boldsymbol{S}_{1:T},\boldsymbol{X}_{1:T},\varphi\setminus\eta_{R,i},\boldsymbol{Z}_{1:T})\sim\mathcal{B}e(\gamma_{R,i,0}^*,\gamma_{R,i,1}^*)$ のように導出できる．ただし，$\gamma_{R,i,0}^*=\gamma_{R,i,0}+y_{R,i}$，$\gamma_{R,i,1}^*=\gamma_{R,i,1}+n_{0,R}-y_{R,i}$ である．ここで $y_{R,i}$ は，レジーム状態が R のときに，人物 i が発話をしていた時間ステップの総数を指す．

また，頭部方向の尤度分布の平均値 $\mu_{i,j}$ の全条件付き事後分布は，一般的な分散既知の場合のガウス分布のベイズ推定[36)~38)]に倣い，

$$\begin{aligned}&p(\mu_{i,j}|\boldsymbol{S}_{1:T},\boldsymbol{X}_{1:T},\varphi\setminus\mu_{i,j},Z_{1:T})=N\left(\phi_{i,j}^*,\xi_{i,j}^{*2}\right)\\&\phi_{i,j}^*=\xi_{i,j}^{*2}\cdot(l_{i,j}\cdot\overline{h}_{i,j}/c_{i,j}^2+\phi_{i,j}/\xi_{i,j}^2)\\&\xi_{i,j}^{*2}=(l_{i,j}/c_{i,j}^2+1/\xi_{i,j}^2)^{-1}\end{aligned}\tag{9}$$

のように平均 $\phi_{i,j}^*$，分散 $\xi_{i,j}^{*2}$ を持つガウス分布として与えられる．ただし，$l_{i,j}$ は，人物 i が方向 j を見た総時間ステップ数を表し，$\overline{h}_{i,j}$，$c_{i,j}^2$ は，そのときの頭部方向の標本平均，標本分散をそれぞれ表す．ただし，この平均値 $\mu_{i,j}$ のサンプリングにおいては，参加者の位置関係が既知であるという仮定のもと，ある人物から他の人物を見るときの頭部方向について制約条件を導入する．たとえば，図 3 (a) のように参加者

が位置する場合，人物 1 の頭部方向の尤度分布について，$\mu_{1,2} > \mu_{1,3} > \mu_{1,4}$ を満たすような平均値の組が得られるまで繰り返しサンプリングを行う．また，頭部方向尤度分布の分散 $\sigma^2_{i,j}$ の全条件付き事後分布は，一般的な平均既知の場合のガウス分布のベイズ推定[36)~38)] に倣い，

$$p(\sigma^2_{i,j}|S_{1:T}, X_{1:T}, \varphi \setminus \sigma^2_{i,j}, Z_{1:T}) = \chi^{-2}(\nu^*_{i,j}, \lambda^*_{i,j})$$
$$\nu^*_{i,j} = \nu_{i,j} + l_{i,j}$$
$$\lambda^*_{i,j} = \lambda_{i,j} + l_{i,j} \cdot c^2_{i,j} \quad (10)$$

のように，自由度 $\nu^*_{i,j}$，尺度パラメータ $\lambda^*_{i,j}$ の逆カイ二乗分布として与えられる．

さらに，各時刻 $t \in \{1, \cdots, T\}$ におけるレジーム状態 S_t の全条件付き事後分布は，式 (1)，式 (2) から，

$$p(S_t|S_{1:T} \setminus S_t, X_{1:T}, \varphi, Z_{1:T})$$
$$\propto p(S_t|S_{t-1}) \cdot p(S_{t+1}|S_t) \cdot p(X_t|X_{t-1}, S_t, S_{t-1})$$
$$\cdot p(X_{t+1}|X_t, S_{t+1}, S_t) \cdot f_U(U_t|S_t) \quad (11)$$

のように求めることができる．また，各時刻 t における視線パターン X_t の全条件付き事後分布は，

$$p(X_t|S_{1:T}, X_{1:T} \setminus X_t, \varphi, Z_{1:T})$$
$$\propto p(X_t|X_{t-1}, S_t, S_{t-1})$$
$$\cdot p(X_{t+1}|X_t, S_{t+1}, S_t) \cdot f_H(H_t|X_t) \quad (12)$$

として与えられる．

3.2.3 事後分布からの統計量の計算

すべての未知変数についてひととおりサンプリングと値の更新を行う処理を 1 単位とし，それを $\mathcal{N}$ 回，繰り返し実行する．反復終了後，反復回 $q = \mathcal{N}'$ から $\mathcal{N}$ において得られたサンプル集合 $\{S^{(q)}_{1:T}, X^{(q)}_{1:T}, \varphi^{(q)}\}^{\mathcal{N}}_{q=\mathcal{N}'}$ から未知変数についての統計量が計算される．ただし，$\mathcal{N}'$ はサンプル値の系列が十分に収束したと見なせる反復回とする．また，上付き添え字 (q) は，q 回目の反復回で得られた各変数の値を表す．レジーム系列，および，視線パターン系列については，最大事後確率推定値を $\hat{S}_t = \arg\max_{R \in \boldsymbol{R}} \sum^{\mathcal{N}}_{q=\mathcal{N}'} \delta_R(S^{(q)}_t)$ のように計算する．他の変数については，推定値を $\hat{\mu} = (\mathcal{N} - \mathcal{N}' + 1)^{-1} \sum^{\mathcal{N}}_{q=\mathcal{N}'} \mu^{(q)}$ のように計算する．

4. 実　　験

提案した会話モデル，および，会話構造推定法の有効性を検証するため，4 人会話を対象とした実験を行った．本章では，まず，実験に使用したデータについて述べ，続いて，モデルのハイパーパラメータの設定について説明する．その後，会話レジームと人物行動の関係が適切にモデル化されていることを確認するために，視線方向の推定精度を評価する（評価 1 と呼ぶ）．次に，推定された会話レジームが実際の会話の構造を反映したものであることを確認するために，発話の種別・方向性のラベルに基づいた評価法を提案し，他手法との比較を交えて，推定されたレジームの評価を行う（評価 2 と呼ぶ）．最後にこれらの結果をあわせて，提案した会話モデル，および，会話構造推定法の有効性を確認する．

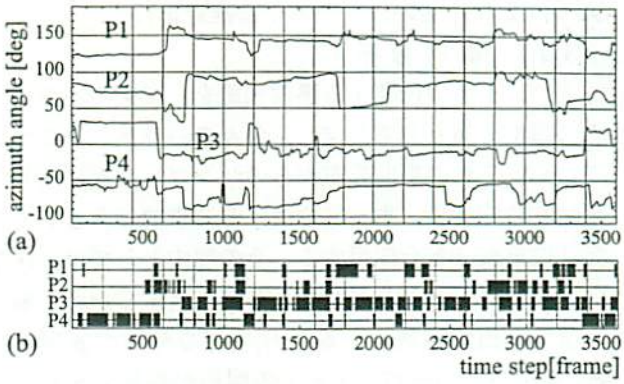


図 4 観測されたデータ（会話 G1-C1），(a) 頭部方向，(b) 発話区間．(Pi は人物 i を示す)

Fig. 4 Observed data for 2 [min] from G1-C1, (a) head azimuth, (b) temporal intervals with utterance (Pi denotes person i).

4.1 会話データの収集と準備

本論文では，参加者数 4 人によるグループ会話を対象とした．参加者は女性 8 人とし，4 人ずつの 2 グループ G1，G2 について，それぞれ会話を収録した．各グループの年齢構成は，グループ G1 は，23 歳～28 歳（平均 25.8 歳），グループ G2 は，28 歳～29 歳（平均 28.5 歳）であった．参加者に対しては，与えられた議題に対して議論を行い，5 分を目安にグループとして 1 つの結論を出すよう指示が与えれた．1 つの議題として「恋愛と結婚は一緒か別か？」という議題を G1，G2 に与えた．この議題に対して，2 つのグループが行った会話を G1-C1，G2-C1 とそれぞれ表記する．また，別の議題として，G1 に対して「安楽死は法的に認めるべきか？」という議題（G1-C2）を，また，G2 に対して「専業主婦に対して法的優遇措置をとるべきか？」という議題（G2-C2）をそれぞれ与えた．なお，これらの議題の選択に際し，会話収録前にいくつかの議題に対してアンケートを行い，同一グループの参加者間で意見の異なる議題を選択した．

頭部方向の計測は，磁気式の 6 自由度センサ（POLHEMUS Fastrak™）を採用し，各参加者の頭部にヘアバンドによりセンサを装着し，30 [Hz] で計測を行った．図 4 (a) には，会話 G1-C1 において計測された頭部の方向（水平方向の方位角）の時系列の一部（3,600 ステップ = 2 [min]）を示す．また，音声データを，各参加者に装着したピンマイクによって収録し，各人物

の発話区間を，音声波形編集ソフトを用いて人手により抽出した．ここでは1つの発話区間を300 [ms] 以上の無音区間に挟まれた区間として定義した．図4 (b) には，図4 (a) と同一の時間区間について得られた発話区間の様子を示す．さらに，会話の様子を映像として記録するため，全参加者をとらえた全体ショット（図3 (b)），各参加者のバストショット（図8 (a)），および，各参加者の顔領域をとらえたアップショットを撮影した．フレームレートは30 [frame/sec] とした．これらのデータは，単位時間ステップ1/30 [sec] にあわせて同期された．分析に用いる会話データ G1-C1，G1-C2，G2-C1，G2-C2 の時間長は，それぞれ 10,000，9,300，9,100，10,000 [frame] とした（5.1分から5.6分に相当）．

4.2 ハイパーパラメータの設定

会話モデルの各パラメータの事前分布について，その分布の形状を規定するハイパーパラメータの設定を行った．この設定は，会話モデルを特徴付けるうえで重要であり，ここでは，次のような方針に従って経験的にハイパーパラメータの値を決めた．なお，すべての会話データに対して同一の値を用いた．

以下では，事前分布としてディリクレー分布を用いるパラメータについては，その分布のパラメータ $\boldsymbol{\alpha}_R$（レジームの初期・遷移確率の場合）を，$\boldsymbol{\alpha}_R = \{\alpha_{R,R'}\}_{R'\in\boldsymbol{R}} = \{\alpha'_R \cdot \alpha'_{R,R'}\}_{R'\in\boldsymbol{R}}$ のように，事前平均 $\alpha'_{R,R'} = E[\pi_{R,R'}]$ と，仮設的な事前分布の標本サイズ α'_R の積に分解し，それぞれを設定することとした．ただし，$\sum_{R'\in\boldsymbol{R}} \alpha'_{R,R'} = 1$ である．なお，視線パターンの生成重み/遷移重み，および，発話確率についても同様にパラメータを設定した．

4.2.1 レジーム

レジームの初期確率 π_0 の事前平均は，1者集中レジームの各々，および，分散レジームに対して均等な値 $\alpha'_{0,R} = 0.19$，$R \in R^0 \cup \boldsymbol{R}^C$ を与え，残りの2者結合レジームの各々についても均等な値を設定した．また，レジーム間の遷移確率の事前平均については，分散レジームから1者集中レジーム（$R^0 \to R \in \boldsymbol{R}^C$）については均等な値を設定し，分散レジームから2者結合レジーム（$R^0 \to R \in \boldsymbol{R}^{DL}$）への遷移はないものと仮定した．1者集中からの遷移，$R_i^C \to R_j^C$ $(j \neq i)$，$R_i^C \to R^0$，については均等な値を設定した．また，1者集中 R_i^C から2者結合への遷移は，1者集中の中心人物 i が2者結合のペアに含まれるもの（$R_i^C \to R_{(k,j)}^{DL}$ $(i = k \text{ or } i = j)$）のみ生じるものと仮定した．2者結合 $R_{(i,j)}^{DL}$ からの遷移は，$R_{(i,j)}^{DL} \to R_i^C$，R_j^C，R^0 に限り均等な事前平均を与え，それ以外の2者結合からの遷移はないものと仮定した．また，仮設的な事前分布の標本サイズは $\alpha'_R = 5{,}000$，$\forall R \in \boldsymbol{R}$ とした．

4.2.2 視線パターンと発話

1者集中レジーム $R = R_i^C$ においては，中心人物 i の視線方向分布は一様であると仮定し，視線パターンの生成重みの事前平均 $\beta'_{R,i,0,j}$ は均等とした．一方，他の人物 j $(\neq i)$ は，高い確率で話し手である中心人物 i の方を向くと仮定した（$\beta'_{R,j,0,i} = 0.7$）．なお，視線パターンの遷移重みの事前平均は，聞き手の視線が中心人物を向いていない場合，中心人物へと遷移しやすく，また，すでに中心人物を見ている場合，見続ける傾向が強くなるように設定した．さらに，発話確率の事前平均は，中心人物 i が主に発話を行い（$\gamma'_{R,i,0} = 0.97$），その他の人物の発話は相槌などに限定されるよう設定した（$\gamma'_{R,j,0} = 0.03$，$j \neq i$）．

2者結合レジーム $R = R_{(i,j)}^{DL}$ においては，ペアをなす人物 i，j は高い確率で互いを見るものとの仮定をおき，生成重みの事前平均を $\beta'_{R,i,0,j} = \beta'_{R,j,0,i} = 0.95$ と設定した．一方，他の人物については各視線方向に均等な事前平均を設定した．遷移重みについては，ペアをなす人物は，互いの方向を見続けるとし（$\beta'_{R,i,j,j} = \beta'_{R,j,i,i} = 1.0$），他の人物については，ランダムな方向に視線が変化するよう均等な事前平均を設定した．また，ペアをなす人物が主に発話を行い（$\gamma'_{R,i,0} = \gamma'_{R,j,0} = 0.7$），他の人物の発話はほとんどないものと仮定した（$\gamma'_{R,k,0} = 0.03$）．

分散レジーム $R = R^0$ においては，各参加者 $i \in \{1, \cdots, N\}$ の視線方向はランダムであると仮定し，各々の視線方向の生成重みの事前平均は均等とした（$\beta'_{R,i,0,j} = 1/N, j = 1, \cdots, N$）．また，視線方向の変化もランダムであると仮定し，均等な事前平均を設定した．また，このレジームにおいては，発話は少ないものと仮定した（$\gamma'_{R,i,0} = 0.01$）．

なお，仮設的な事前の標本サイズについては，視線パターンに関しては，$\beta'_{R,i,0} = 10{,}000$，$\beta'_{R,i,k} = 500$ $(k \neq 0)$ とし，発話確率については $\gamma'_{R,i} = 500$ とした．

4.2.3 頭部方向の尤度分布

頭部方向の尤度分布の平均値の事前分布 $p(\mu_{i,j})$ については，その平均値 $\phi_{i,j}$ を，図3 (a) のように，ある人物 i から他の人物 j を見込んだ方位角 $\Delta\phi_{i,j}$ として与えた（$i \neq j$ の場合）．また，視線を逸らしている状態に関しては，$\phi_{i,j}$ は，他の人物に関する $\phi_{i,j}$ $(j \neq i)$ の平均値を与えた．また，その他の頭部方向に関するパラメータは，経験的に $\xi_{i,j}^2 = 0.2$，$\nu_{i,j} = 1{,}000$ $(i \neq j)$，$\nu_{i,i} = 3{,}000$，$\lambda_{i,j} = 100$ のように与えた．

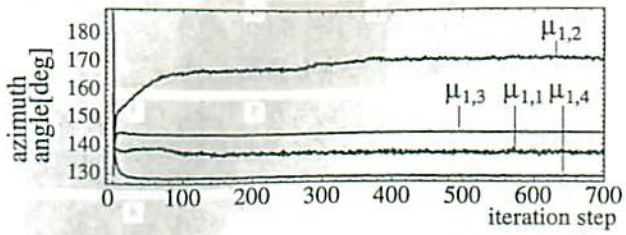


図 5 ギブスサンプリングにおける頭部方向尤度分布の平均値 $\mu_{1,1}$, $\mu_{1,2}$, $\mu_{1,3}$, $\mu_{1,4}$ の遷移（G1-C1 の場合）

Fig. 5 Transition of $\mu_{1,1}, \mu_{1,2}, \mu_{1,3}, \mu_{1,4}$ through iteration of Gibbs sampler: G1-C1 case.

4.3 ギブスサンプリング

ギブスサンプリングの反復回数は，反復回に従ってパラメータの値がどのように遷移するかを示したグラフを作成し，それを目視することで，収束に至る反復回数を把握し，$\mathcal{N} = 700$, $\mathcal{N}' = 500$ のように設定した．**図 5** にはそのようなグラフの一例として，会話データ G1-C1 について得られた，頭部方向の尤度分布の平均値 $\{\mu_{1,j}\}_{j=1}^{4}$ の遷移を示す．

4.4 評価 1：視線方向の推定精度

提案した会話モデルによって，会話レジームと人物行動の関係が適切にモデル化されていることを確認するために，視線方向の正解ラベルを用いて，視線方向の推定精度を検証した．なお，視線方向の正解ラベルは，撮影された映像を詳細に観察することにより人手で付与された．

表 1 には，各会話データ，各参加者について，推定された視線方向と正解ラベルが一致した時間ステップ数の割合，および，それらの全参加者についての平均値を示す．また，**図 6** (a) には，推定結果の一例として，会話 G1-C1 の一部（最初の 2 分間）における各人の視線方向の時系列を正解ラベルとあわせて図示する．推定結果と正解ラベルが一致しない状況をすべての会話データについて調べた結果，視線を逸らした状態（凝視回避と呼ぶ）を凝視として誤判定したケースが，すべての誤りのうち，約 60%を占め，また，凝視の状態を凝視回避と誤ったケースが同約 13%を占め，推定誤りの多くは視線を逸らした状態に関連することが分かった．このような誤りの要因を探るため，**図 7** に，推定された頭部方向の尤度分布，および，正解ラベルを用いて作成された各視線方向ごとの頭部方向のヒストグラムを示す．図 7 からは，凝視回避時の分布と凝視時の分布が，尤度分布，ヒストグラムともに大幅にオーバラップしていることが分かる．このことは，頭部方向を固定したまま，視線のみを動かすことができる，つまり，流し目を使う，という人間の性質

表 1 視線方向の推定精度 [%]

Table 1 Accuracy of gaze direction estimates [%].

	Average	P1	P2	P3	P4
G1-C1	71.1	80.9	65.7	71.5	66.4
G1-C2	59.3	70.4	48.2	67.4	51.4
G2-C1	72.4	57.4	83.8	78.2	70.1
G2-C2	75.9	49.0	90.5	84.1	80.1

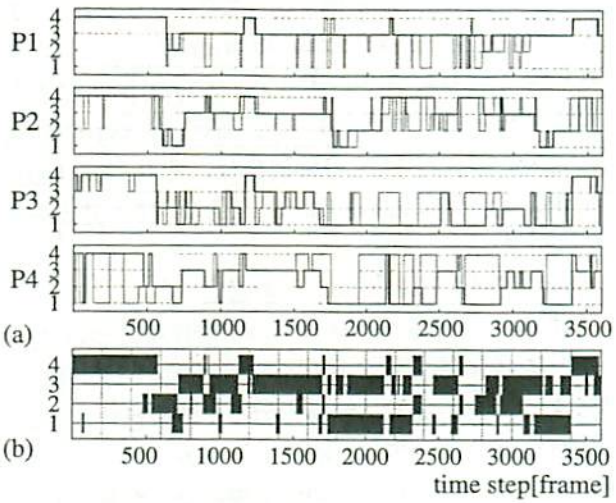


図 6 会話データ G1-C1 に対する推定結果：(a) 各人物 $P1$, $P2$, $P3$, $P4$ の視線方向 $\{X_{1,t}, X_{2,t}, X_{3,t}, X_{4,t}\}$（実線：推定値，破線：正解ラベルの方向），(b) 推定されたレジーム状態；各時刻 t において，1 者集中レジーム $S_t = R_i^C$ の場合，i の位置のみにバンド，2 者結合レジーム $R_{(i,j)}^{DL}$ の場合，i, j の位置にバンド，分散レジーム R^0 の場合は空白により示す

Fig. 6 Estimated sequences of (a) gaze pattern $\{X_{1,t}, X_{2,t}, X_{3,t}, X_{4,t}\}$ and (b) regime states: C1-G1 case. In (a), solid lines: estimates, dashed lines: ground truth. In (b), single band at a time slice indicates regime R_i^C (convergence), dual band at time slice indicates regime $R_{(i,j)}^{DL}$ (dyad link), and no band indicates R^0 (divergence).

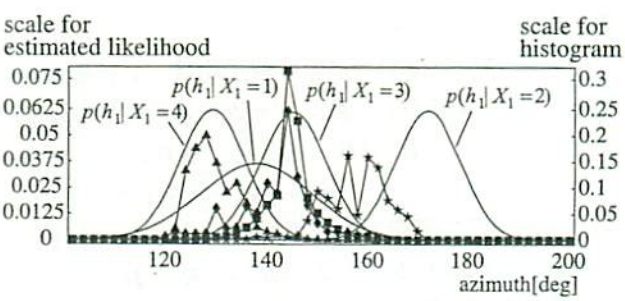


図 7 会話データ G1-C1 より推定された頭部方向の尤度分布 $p(h_1|X_1 = i)$（曲線），および，頭部方向のヒストグラム（折れ線）．$p(h_1|X_1 = i)$, $i \neq 1$：人物 1 が人物 i を見ている．$p(h_1|X_1 = 1)$：人物 1 が視線を逸らしている．折れ線の記号で視線方向を表す；◆視線を逸らした状態，★人物 2，■人物 3，▲人物 4

Fig. 7 Estimated likelihood function $p(h_1|X_1 = i)$ of head direction, from G1-C1 case. (person 1 looks at person if $i \neq 1$, or avert gaze if $i = 1$); line with symbol shows corresponding histogram, (symbol = diamond: avert, star: gaze at P2, square: gaze at P3, triangle: gaze at P4).

が反映された結果であり，視線方向の推定誤りの要因になっていると考えられる．

また，比較のために，視線方向の正解ラベルを用いて，頭部方向尤度分布（ガウス分布）を求め，視線方向を最大事後確率推定により求めた結果，各会話データについての平均推定精度は，それぞれ，G1-C1：68.6%，G1-C2：65.4%，G2-C1：74.7%，G2-C2：71.1%となった．これらの精度は，床面に水平な頭部方向成分をデータとして用い，その尤度分布にガウス分布を採用する際の一種の限界を示唆するものと考えられる．表1記載の提案法の推定精度は，これらの推定精度より若干，劣る程度，あるいは，場合によっては上回ることが分かった．

本節の結果からは次のような考察が可能である．提案方法では，各人の相対的な位置関係に関する知識は使用しているものの，頭部方向の尤度分布のパラメータは未知という条件のもと，そのパラメータや視線方向の推定を行っている．この問題設定に対処するため，提案方法では会話モデルを導入し，それにより会話構造に応じて出現しやすい視線パターンが想定できることから，曖昧な頭部方向のデータからも精度良く視線方向が推定できることを期待している．ここで会話モデルが適切ではない場合，各人物の頭部方向の尤度分布が精度良く求まらず，結果的に視線方向の精度も低いものになると考えられ，また逆に，視線方向の推定精度が高いことは，会話モデルが適切なものであった証拠であると考えられる．本実験において，提案方法の推定精度が，上記，比較対象の手法の精度に匹敵しうるものであったことから，会話レジームと参加者行動の関係のモデル化という観点から会話モデルの妥当性が示唆されたといえる．

4.5 評価2：会話レジームについての評価

4.5.1 定性的な側面

図6(b)に，会話データG1-C1より推定されたレジームの時系列の一部を示す．また，図8には，この会話から特徴的な個所を取り出し，その会話の流れに沿った3時刻（$t = 310$，$t = 485$，$t = 578$）における各参加者の様子（図8(a)），および，推定されたレジームと視線パターン（図8(b)）を示す．この個所において，最初（$t = 310$），人物4は，他のすべての参加者に対して意見を述べており（P4：「結婚は考えなくても，誰かと付き合うってことを考えるわけじゃない．っていう意味で」），他の参加者は人物4の発話を聞いている．このときの会話レジームは，人物4への1者集中 R_4^C と推定されており，実際の会話の構造に合致していると考えられる．次に（$t = 485$），人物

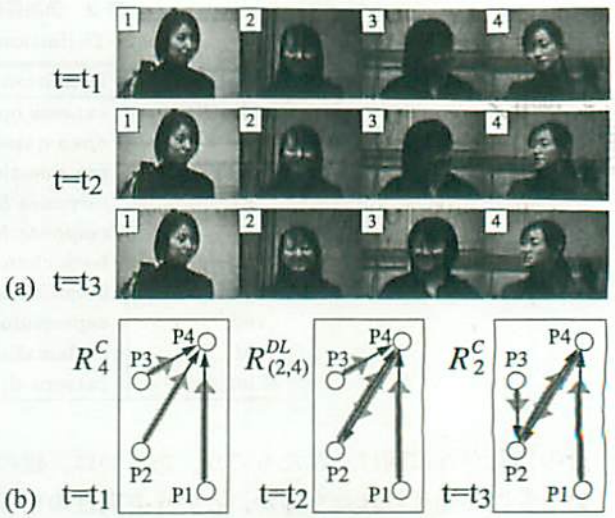


図8 推定されたレジームの時間遷移の様子（会話データC1-G1中の $t_1 = 310$，$t_2 = 485$，$t_3 = 578$）．(a) 各参加者のバストショット，(b) 推定されたレジーム，および，視線パターン（細い矢印：推定された視線方向，幅の広い矢印：正解ラベル）

Fig. 8 An example of regime transition ($t_1 = 310$, $t_2 = 485$, $t_3 = 578$) from C1-G1. (a) snapshot of each participant, (b) regime estimates and gaze patterns (solid arrows: estimates, wide arrows: ground truth).

2が人物4に対して頷きをともなう発話（P2：「うん，うん，うん，うん，うん」）により強い反応を表出し，人物4も人物2に視線を向け，その反応を確認している．ここでは人物2と4の間で相互凝視が生じている．この個所での会話レジームは，この2人物間の2者結合 $R_{(2,4)}^{DL}$ と推定され，この状況を適切に反映したものになっていると考えられる．さらに（$t = 578$），人物2は発話を継続し（P2：「そう，ずっとっていう意味で」），それに対して人物4は相槌を打った後（P4：「うん，うん」），発話を中断し，人物2に発話権を譲渡した．また，同時に，人物3は，新しい話し手である人物2の発言に注意を向けるため，その視線を人物4から人物2へと移した．このような現象を反映して，この個所では会話レジームは人物2への1者集中 R_2^C と推定された．このような観測により，推定されたレジームの系列は，実際の会話の流れを反映したものであることが確認された．

4.5.2 推定された会話レジームの評価法

次に，推定されたレジームの精度を定量的に評価するための方法について述べる．現在のところ，会話構造を定量的に評価する一般的な方法が存在しないため，本論文では独自の評価方法を考案した．この方法は，会話中の各発話区間について，人手によって付与されたラベルと推定されたレジーム状態との照合を行うことで実行される．発話区間のラベルは，発話の種別と

表 2 発話区間に付与されるラベルの定義
Table 2 Definition of labels for utterance interval.

Label	Definition
a$d_1d_2,\cdots$	express opinion toward persons $d_1d_2,\cdots$
q$d_1d_2,\cdots$	open question toward persons $d_1d_2,\cdots$
Q$d_1d_2,\cdots$	tag question toward $d_1d_2,\cdots$
rd	response for question from person d
Rd	response for others' utterances (other than 'r')
nd	back-channel utterance for person d
sd	laugh caused by what person d said or did
ed	expressions with sound such as hum and sigh
!d	exclamation caused by what Pd said or acted
h$d_1d_2,\cdots$	persons $d_1d_2,\cdots$ are listening to the utterance

その発話が誰に向けられたものか，あるいは，誰の発話を受けて発せられたものか，という方向性の情報を含む．このラベルは，具体的には表 2 に示すような「発話種別＋方向性」の形式をとり，各発話区間について，その区間に該当する項目のラベルをすべて付与した．発話種別には，意見の表明 {a}，他者への問いかけ（質問 {r}，同意の要求 {R}），問いかけに対する応答（質問に対する応答 {r}，要求された同意に対する応答 {R}），そのほか，他者の発話に対する反応（相槌 {n}，笑い {s}，驚き {!}），その他の発声 {e}（溜息，唸り声など）が含まれる．加えてラベルには，聞き手が記号 h に続いて記述される．意見の表明は，1 者集中レジームにおける話し手から受け手への情報伝達に関連し，また，問いかけやそれに対する応答は，2 者結合レジームにおける 2 者間の双方向的な情報交換に関連する．また，分散レジームに関連したラベルとしては，方向性のラベルがない，つまり，他者に向けられていない独り言や，聞き手のいない発話などがあげられる．

たとえば，ラベル {a234,h234,q2} が人物 1 の発話区間について与えられた場合，人物 1 は，人物 2，3，4 に向かって自分の意見を表明しており，また，人物 2，3，4 は人物 1 の発話に注意を傾けて聞いていることが表される．さらに，人物 1 は，その発話区間の終端において，人物 2 に対して問いかけ（質問）を発している．また，この発話区間に後続する人物 2 の発話区間にラベル {r1,a1,h134} が与えられた場合，人物 2 は，人物 1 の問いかけに対して応答し，自らの意見を人物 1 のみに向かって表明しており，その発話を人物 1，3，4 が聞いているという状況が表される．このようなラベル付けによって，各人の発話区間の間の因果関係を記述することができる．

次に，レジームの推定精度として，各時刻 t において，推定されたレジーム $\hat{R}_t$ に対応する会話構造が，発話ラベルによって示唆された現象中に含まれるかどうか判定し，含まれていると判定された時間ステップの割合を，レジームの正答率として計算する☆．そのため，まず，各時刻 t について，その時刻を範囲に含む発話区間のラベル集合 $L(t)$ を抽出する．ここでは，レジームの時間変化のスケールが単位時間ステップ（=1/30 [sec]）よりも大きいことを考慮し，推定されたジレーム状態 $\hat{R}_t$ と同じ推定値を持つ連続時間区間 $[t_0,t_1]$（$t_0 \le t \le t_1$，$\hat{R}_t = \hat{R}_{t'}$，$\forall t' \in [t_0,t_1]$）との交差を持つ発話区間の集合に含まれるすべてのラベルを抽出し，照合に用いるラベル集合 $L(t)$ とした．

次にレジームの種別ごとに異なる条件を用い，時刻ごとに評価を行った．1 者集中 $\hat{R}_t = R_i^C$ の場合，ある時刻のレジーム状態の推定値は，ラベル集合 $L(t)$ が次の 2 つのラベル要素のいずれかを含む場合に正解とした；条件 i-i) 人物 i が他のすべての参加者に向けて意見を表明していた，条件 i-ii) 人物 i は 1 人の人物に向けて意見を表明しているが，すべての人物がその話に注意を傾けて聞いていた☆☆．また，2 者結合 $\hat{R}_t = R_{(i,j)}^{DL}$ の場合，正解の条件はラベル集合 $L(t)$ が次の条件のいずれかを満足する場合とした；条件 ii-i) 人物 i は人物 j のみに向けて問いかけ，もしくは，応答・反応を行った，条件 ii-ii) 人物 i は人物 j のみに向けて意見を表明し，人物 j のみがその話を注意を傾けて聞いていた．条件 ii-iii) 先の条件 ii-i)，ii-ii) において，人物 i と j を入れ替えた条件．分散 $\hat{R}_t = R^0$

☆ 発話ラベルは，物理的な発話区間を単位としているため，1 つの連続した発話区間内に，複数の異なる意図を持った発話が含まれる可能性がある．また，複数の人物がそれぞれ話者として同時に発話を試みるような状況も存在しうる．よって，1 時刻に対応する発話ラベルには，複数の現象が含まれる場合がある．そのため，レジームとラベルとの一致率という観点からではなく，「含まれるかどうか」という観点からの評価を行うこととした．

☆☆ この場合，人物 i としては，特定の 1 者への話しかけという意図を持っているのであるが，結果的にこの場で生じているコミュニケーションは，1 対多の 1 方向性の情報伝達であると考えられるため，2 章で定義した 1 者集中レジームに該当するものと考えた．

表 3 レジームの正答率 [%], (a) 提案手法 $\mathcal{M}$, (b) 視線既知 $\mathcal{M}_{GG}$, (c) 視線のみ $\mathcal{M}_{GO}$, (d) 発話のみ $\mathcal{M}_{UO}$. (Conv.：1 者集中レジーム, DL.：2 者結合レジーム, Div.：分散レジーム, Total：全レジーム)

Table 3 Accuracy in regime estimates [%], (a) Our method $\mathcal{M}$, (b) Gaze given $\mathcal{M}_{GG}$, (c) Gaze only $\mathcal{M}_{GO}$, (d) Utterance only $\mathcal{M}_{UO}$. (Conv.: Convergence, DL.: Dyad-Link, Div.: Divergence, Total: All regimes).

	(a) Our method $\mathcal{M}$				(b) Gaze Given $\mathcal{M}_{GG}$			
	Total	Conv.	DL.	Div.	Total	Conv.	DL.	Div.
G1-C1	81.8	85.0	77.7	53.2	78.1	85.8	70.1	68.6
G1-C2	92.1	94.9	85.9	69.0	81.0	91.3	76.7	42.9
G2-C1	91.4	95.7	72.6	100	92.6	93.5	91.3	88.5
G2-C2	96.3	98.8	83.5	100	96.0	97.6	91.3	88.5

(c) Gaze Only $\mathcal{M}_{GO}$				(d) Utterance Only $\mathcal{M}_{UO}$			
Total	Conv.	DL.	Div.	Total	Conv.	DL.	Div.
69.7	86.3	52.7	54.1	78.5	80.2	64.6	100
79.0	92.1	67.5	56.9	83.4	85.9	73.4	100
84.0	95.4	64.6	52.5	89.6	88.4	78.5	100
92.0	98.6	41.7	49.2	92.4	93.2	83.4	98.7

の場合，正解の条件は，ラベル集合が空集合 $L(t) = \emptyset$ である，もしくは，次の 2 つの条件を同時に満たす場合とした；条件 iii-i) 発話は独り言であった，条件 iii-ii) 他者の発話に対して誰も応答をしなかった．

4.5.3 レジームの推定結果の評価結果

表 3 に，各会話データについて得られたレジーム推定値の正答率を，各レジームのクラスごと，および，全体について示す．この表には提案方法（$\mathcal{M}$ と記す）の結果に加えて，比較対象として 3 つの異なる方法による結果も併記した．この比較対象の方法は，提案した会話モデルにおいて，観測可能な変数の条件を変えたものであり，i) 視線既知 $\mathcal{M}_{GG}$，ii) 視線のみ $\mathcal{M}_{GO}$，iii) 発話のみ $\mathcal{M}_{UO}$ とした．視線のみ $\mathcal{M}_{GO}$ と発話のみ $\mathcal{M}_{UO}$ の手法については，それぞれのみを観測値として与えた条件のもと，ギブスサンプリングによりレジームの系列を推定した．また，視線既知 $\mathcal{M}_{GG}$ と視線のみ $\mathcal{M}_{GO}$ においては，視線パターン系列 $\boldsymbol{X}_{1:T}$ について正解ラベルを観測値として与えた．なお，これら方法において，新たに既知とした変数以外の条件やパラメータは，提案方法 $\mathcal{M}$ と同一とした．これらの方法との比較は，正答率の大小という観点を補完し，より多角的に提案方法を検証することを狙ったものである．具体的には，視線のみ $\mathcal{M}_{GO}$，発話のみ $\mathcal{M}_{UO}$ の各方法との比較により，着目した人物行動の観点から会話モデルの妥当性を検証する．また，提案方法 $\mathcal{M}$ と視線既知 $\mathcal{M}_{GG}$ の方法を比較することで，頭部方向を視線方向の代用として計測することの妥当性を議論する．

表 3 より，提案方法 $\mathcal{M}$ により推定されたレジームの正答率は，81.8～96.3%と平均的に高い値を示していることが分かる．また，総合的にみて，提案方法 $\mathcal{M}$ の結果は，視線のみ $\mathcal{M}_{GO}$，発話のみ $\mathcal{M}_{UO}$ の方法の正答率を上回り，また，視線既知 $\mathcal{M}_{GG}$ に匹敵する正答率を実現していることが分かる．

また，表 3 からは，視線，発話を単独に用いた方法（$\mathcal{M}_{GO}$，$\mathcal{M}_{UO}$）よりも，それら双方を用いた方法（$\mathcal{M}$，$\mathcal{M}_{GG}$）の方が全体的に高い正答率を達成していることも読み取れる．なお，数値上は，発話のみの方法 $\mathcal{M}_{UO}$ でも十分な精度が得られているようにみえるが，この方法で得られたレジームの系列は，2 者以上の同時発話や瞬時的な沈黙など発話の状態に敏感なものとなっており，会話の流れを適切に反映しているとはいい難いことが観察により判明した．それに対して，視線パターンと発話の双方を利用した方法では，発話の状態に敏感に左右されることもなく，発話情報だけからは誰が話し手であるか特定困難な同時発話などの状況においても，発話権を保持している話し手を安定に同定できることが確認された．これらの結果により，視線パターンと発話状態の双方をモデルに組み込むことの有効性が示唆された．

さらに表 3 より，提案方法 $\mathcal{M}$ は，視線既知の方法 $\mathcal{M}_{GG}$ と比較して，同程度かそれを上回るレジームの正答率を達成していることが分かり，視線方向の代わりに頭部方向を計測することの妥当性が示唆された．本方法 $\mathcal{M}$ で推定された視線方向には，4.4 節で述べたように誤差が含まれている．それにもかかわらず，このような正答率が得られた原因として，以下のような考察が成り立つ．まず，話し手が受け手に対して話しかける場合や，聞き手が話し手の発話を傾聴する場合，対象へのアテンションは継続するものの，その視

線は対象への凝視と凝視回避を繰り返すという特性があり[16),19),21)]，図6(a)からもそのような特性を反映した同一対象への凝視・凝視回避が読み取れる．一方，このような場面では，視線方向の変化にともなう頭部方向の変化が比較的小さく，この性質により4.4節で述べたように凝視回避を凝視として誤判定するケースが多く生じると考えられる．このように提案方法によって推定された視線方向は，実際の視線方向よりも頭部方向の影響を強く受けたアテンションの方向を示唆するものと考えられ，これがかえって，話しかけ・傾聴時のアテンションの方向と符合した結果，推定された会話レジームは実際の会話現象を十分に反映したものとなり，高いレジームの正答率が得られたものと考察できる．

4.6 評価のまとめ

4.4節で述べた視線方向の推定精度の評価から，会話レジームと参加者行動の関係が適切にモデル化されていることが確認され，また，4.5節において，推定された会話レジームが実際の会話構造を反映していることが確認された．これらの結果より，本論文で提案した会話モデル，ならびに，会話モデルに基づく会話構造の推定方法の有効性が示唆された．

5. 議　　論

本章では以下の各項目について議論を行う．

会話レジームとその評価について

本論文では，会話を構成する代表的な会話構造を定義するために，視線パターンの構造的な特徴に着目し，3クラスの会話レジームを仮説的に設定した．しかしながら，実際の会話においては，本論文の会話レジームでは想定していない状況，たとえば，話し手が1人の受け手のみに向かって1方的に話しかける場面や，参加者が複数のサブグループに分かれて会話が進行するような場面もありうる．今後は，これらの場面へ対処するために，データに基づいて適応的に会話レジームを設定することが検討課題としてあげられる．また，参加者数の異なる会話や，会話中の参加者数の増減，様々な人物配置，ノートや資料，黒板などの使用など，より多様な会話の状況に対処できるよう提案法を拡張し，評価を行うことが望まれる．

また，本論文では，推定されたレジームの評価のために，発話区間を単位として，その人物間の問いかけ・応答に関するラベルを用いて評価を実施した．しかしながら，会話中には，発話として現れない非言語的なメッセージ交換，たとえば，話し手の問いかけに対して，表情のみで無言で反応を返す場面など，も含まれており，今後は，それらも含めた評価法を検討する必要がある．

関連研究に対する本研究の位置付け

本研究は，Stiefelhagenらの研究[26),29)]の発展形であると位置付けることができる．彼らは，グループ会話において，参加者の頭部方向，および，発話の状態から各人物の視線方向（注意の焦点と呼んでいる）をベイズ推定により求める手法を提案している．本論文の方法との類似点として，頭部方向の尤度関数にガウス分布を仮定している点があげられ，また，相違点として以下の項目があげられる．まず，彼らは，個々の参加者の視線方向を独立に求めており，参加者間のインタラクションや会話の構造について陽にモデル化を行うというアプローチをとっていない．また，視線方向と発話の共起関係を用いて推定を行っている点において，部分的に会話の性質を利用していると考えられるが，彼らの提案手法は，会話の構造の推定を目的とするものではない．さらに，視線を逸らした状態を推定の対象から除外している点でも本研究とは立場を異にする．

また，坊農らは，ポスター発表会場における説明者と来訪者との間の会話を対象とした研究を進めている[13),14)]．この研究では，来訪者の参与役割が，非参与者から傍観者を経て傍参与者，受け手，話し手へと動的に遷移する過程に着目し，参加者の音声，および，赤外線IDシステムから得られる位置・頭部方向などの情報を用いて，来訪者の参与役割の同定を試みている[13)]．また，レクチャモードとインタラクションモードと呼ばれる2つの会話のモードを定義し，説明者の無音区間持続長の長短によりこれら2つのモードが判別できることを示唆している[14)]．この研究が，会話参加者の増減や移動を含むいわば「開放環境」を対象とし，ポスター発表という限定された会話の状況を対象としているのに対し，本論文では，参加者の増減・移動が含まれない「閉じた環境」を対象とし，対等な立場の参加者による自由な会話を対象とした研究であると位置付けられる．さらに，本研究は，会話の構造のモデル化を志向している点においても，坊農らの研究とは方向性が異なる．

確率的なアプローチの利点・発展性

本論文で提案した確率的なアプローチの利点としては，まず，複数の人物から表出されるマルチモーダルな情報を統一的な枠組みのもと統合することができるという点があげられる．そのため，頭部方向や発話状態に加えて，他の種類の人物行動を組み込んだ会話モデルにも容易に拡張できると考えられる．たとえば，

頷きや相槌，胴体姿勢，表情，ジェスチャなどの非言語的な行動も会話においては重要な役割を果たしていることが知られているが[16)~18)]，それらを会話モデルに組み込むことで，より精緻に会話の状態を推測することが可能になると思われる．

また，提案方法では，この確率的な枠組みのもとで，視線パターンと発話状態から会話レジームを推定するという問題と，頭部方向から視線方向を推定するという2つの問題を同時に解いていると考えることができる．これにより，それぞれの問題を単独に解く場合よりも，より精度の高い推定が実現されていると考えられる．その理由としては，聞き手は話し手の方をよく見るというような会話の性質に関する情報を利用することで，視線方向の推定精度が高まり，また，より高精度の視線方向を入力とすることで，会話レジームの推定の精度も向上するというような，互いの推定情報が相補的に作用する点があげられる．このような利点をさらに活用することにより，人物が密集して着席する場合など，頭部方向による視線方向の判別が困難な状況にも対処できる可能性がある．たとえば，1者集中の場合において，一部の聞き手の視線方向が判別できない場合でも，他の聞き手を含めた全体での視線方向の多数決により視線の焦点を求め，さらにその結果を発話状態の情報と統合することで，ロバストに会話構造を同定できるものと期待される．

アプリケーションと課題

提案方法の応用先としては，会議映像の自動編集・アーカイブや多地点間の複数人対複数人の遠隔会議，会話エージェント・ロボットなどがあげられ，これらの応用において，会話構造の推定結果は有力な入力情報になると期待される．たとえば，会議映像の自動編集システムにおいては，話し手の映像と受け手の映像を時間的に切り替えて表示するなど，参加者の役割に応じた様々な映像表現が可能となり，誰が誰に向かって話をしているかなど，視聴者にとって会話の内容がより分かりやすい映像を自動的に編集することが可能になると期待される．

さらに，これらの応用を実現するためには以下のような課題の解決が求められる．まず，本論文では，参加者の頭部方向の計測のために装着式のセンサを用いたが，実際の応用場面では，画像を用いた頭部追跡の手法[40)]などによる非接触方式の計測が望まれる．また，3.2節で示した方法は，対象となる会話の時間区間についてバッチ処理を行うものであったが，実時間性が要求されるアプリケーションに提案方法を供する場合，オンライン推定法の利用が必須となる．

また，提案方法は，ハイパーパラメータの設定が煩雑であり，ヒューリスティクスが介在する点が欠点としてあげられる．しかし，これらのパラメータやモデル構造の設定は，会話現象の性質と密接に関連するため，それらの吟味・検討のプロセスは，会話の様々な側面を分析・定量化するツールの開発へと発展する可能性がある．たとえば，推定されたパラメータに基づいて，会話に対する積極性や他者への影響力などといった参加者個人の特性や，会話の円滑さなどのグループとしての特性などを定量化することも興味深い検討課題の1つである．このように，従来，定量的な分析が困難であった複数人物の会話という現象に対して，本論文は，1つの新しいアプローチを提供するものと位置付けられる．

6. むすび

本論文では，複数人物による対面会話を対象として，その会話の構造を推定するための確率的枠組みを提案した．会話の構造として，会話中の各参加者の参与役割に着目した．また，会話の構造に依存して参加者の行動が規定されると仮説を立て，マルコフ切替えモデルに基づく会話モデルを提案した．さらに，参加者の頭部方向，および，発話状態を観測データとし，会話構造に対応すると想定される会話レジームの状態，視線パターン，および，モデルパラメータを推定するため，ギブスサンプリングを用いた方法を提案した．4人会話を対象とした実験を行い，視線方向の推定精度，および，会話レジームの正答率を評価し，提案した会話モデル，および，会話構造推定方法の有効性を確認した．

謝辞　本研究を進めるにあたり，熱心にご指導・ご議論いただきましたNTTコミュニケーション科学基礎研究所メディア情報研究部の皆様，ならびに，NTTサイバーソリューション研究所主任研究員大野健彦氏，東京電機大学教授武川直樹氏，名古屋大学大学院情報科学研究科教授末永康仁氏，同助教授井手一郎氏の諸氏に深く感謝いたします．

参考文献

1) Cutler, R., et al.: Distributed Meetings: A Meeting Capture and Broadcasting System, *Proc. ACM Multimedia '02*, pp.503–512 (2002).
2) Bett, M., Gross, R., Yu, H., Zhu, X., Pan, Y., Yang, J. and Waibel, A.: Multimodal Meeting Tracker, *Proc. RIAO 2000: Content-Based Multimodal Inform. Access* (2000).
3) 竹前嘉修，大塚和弘，武川直樹：対面の複数人

対話を撮影対象とした対話参加者の視線に基づく映像切り替え方法とその効果，情報処理学会論文誌，Vol.46, No.7, pp.1752–1767 (2005).
4) Heylen, D., Es, I.V., Nijholt, A. and Dijk, B.V.: Experimenting with the Gaze of a Conversational Agent, *Proc. Int. CLASS Workshop on Natural Intelligent and Effective Interaction in Multimodal Dialogue Systems*, pp.93–100 (2002).
5) 松坂要佐，東條剛史，小林哲則：グループ会話に参与する対話ロボットの構築，信学論 D-II, Vol.J84-D-II, No.6, pp.898–908 (2001).
6) Katzenmaier, M., Stiefelhagen, R. and Schultz, T.: Identifying the Addressee in Human-Human-Robot Interactions based on Head Pose and Speech, *Proc. ACM Int. Conf. Multimodal Interface* (*ICMI'04*), pp.13–15 (2004).
7) McCowan, I., Perez, D., Bengio, S., Lathoud, G., Barnard, M. and Zhang, D.: Automatic Analysis of Multimodal Group Actions in Meetings,, *IEEE Trans. PAMI*, Vol.27, No.3, pp.305–317 (2005).
8) Zhang, D., Perez, D.G., Bengio, S., McCowan, I. and Lathoud, G.: Modeling Individual and Group Actions in Meetings: A Two-Layer HMM Framework, *Proc. 2nd IEEE Workshop on Event Mining* (2004).
9) Basu, S., Choudhury, T. and Clarkson, B.: Learning Human Interactions with the Influence Model, *MIT Media Lab. TR#539* (2001).
10) Dielmann, A. and Renals, S.: Dynamics Bayesian Networks for Meeting Structuring, *Proc. IEEE ICASSP'04* (2004).
11) Goffman, E.: *Forms of Talks*, University of Pennsylvania Press, Philadelphia (1981).
12) Clark, H.H. and Carlson, T.B.: Hearers and Speech Acts, *Language*, Vol.58, pp.332–373 (1982).
13) 坊農真弓，鈴木紀子，片桐恭弘：ユビキタスセンサを用いた会話参与手続きの認識，第 41 回人工知能学会音声・言語理解と対話処理研究会（SIG-SLUD），pp.27–32 (2004).
14) 坊農真弓，鈴木紀子，片桐恭弘：多人数会話を対象としたデータ収集と分析—参与構造分析を例として，国際文化学，Vol.11, pp.81–94 (2004).
15) 榎本美香，伝 康晴：3 人会話における参与役割の交代に関わる非言語行動の分析，第 38 回人工知能学会音声・言語理解と対話処理研究会（SIG-SLUD-A301），pp.25–30 (2003).
16) Argyle, M.: *Bodily Communication — 2nd ed.*, Routledge, London and New York (1988).
17) M.F. ヴァーガス：非言語コミュニケーション，新潮社 (1987).
18) M.L. パターソン：非言語コミュニケーションの基礎理論，誠信書房 (1995).
19) Kendon, A.: Some Functions of Gaze-Direction in Social Interaction, *Acta Psychologica*, Vol.26, pp.22–63 (1967).
20) Goodwin, C.: *Conversational Organization: Interaction between Speakers and Hearers*, Academic Press, New York (1981).
21) 福井康之：まなざしの心理学，創元社 (1984).
22) Vertegaal, R., Slagter, R., Veer, G. and Nijholt, A.: Eye Gaze Patterns in Conversations: There is More to Conversational Agents than Meets the Eyes, *Proc. ACM CHI'01*, pp.301–308 (2001).
23) Jovanovic, N. and Akker, R.: Towards Automatic Addressee Identification in Multi-party Dialogues, *Proc. SIGdial'04*, pp.89–92 (2004).
24) Ohno, T. and Mukawa, N.: A Free-Head, Simple Calibration, Gaze Tracking System That Enables Gaze-Based Interaction, *Proc. Eye Tracking Research & Application Symposium* (*ETRA*'04), pp.115–122 (2004).
25) Matsumoto, Y. and Zelinsky, A.: An Algorithm for Real-Time Stereo Vision Implementation of Head Pose and Gaze Direction Measurement, *Proc. Int. Conf. Automatic Face and Gesture Recognition '04*, pp. 499–504 (2000).
26) Stiefelhagen, R., Yang, J. and Waibel, A.: Modeling Focus of Attention for Meeting Index Based on Multiple Cues, *IEEE Trans. Neural Networks*, Vol.13, No.4, pp.928–938 (2002).
27) Reidsma, D., et al.: Virtual Meeting Rooms: From Observation to Simulation, *Proc. Social Intelli. Design '05* (2005).
28) 伊藤禎宣，岩澤昭一郎，土川 仁，角 康之，間瀬健二，片桐恭弘，小暮 潔，萩田紀博：装着型体験記録装置による対話インタラクションの判別機能実装と評価，ヒューマンインタフェース学会論文誌，Vol.7, No.1, pp.167–178 (2005).
29) Stiefelhagen, R. and Zhu, J.: Head Orientation and Gaze Direction in Meetings, *Proc. ACM CHI'02* (2002).
30) Kim, C.-J. and Nelson, C.R.: *State-Space Models with Regime Switching*, MIT Press (1999).
31) Frühwirth-Schnatter, S.: Markov Chain Monte Carlo Estimation of Classical and Dynamic Switching and Mixture Models, *Journal of American Statistical Association*, Vol.96, No.453, pp.194–209 (2001).
32) Gilks, W.R., Richardson, S. and Spiegelhalter, D.J.: *Markov Chain Monte Carlo in Practice*, Chapman & Hall/CRC (1996).
33) 中妻照雄：ファイナンスのための MCMC 法に

よるベイズ分析，(財) 三菱経済研究所 (2003).
34) R. ディーステル：グラフ理論，シュプリンガー・フェアラーク東京 (2000).
35) Novic, D.G., Hansen, B. and Ward, K.: Coordinating Turn-Taking with Gaze, *Proc. Int. Conf. Spoken Language '96*, pp.1888–1891 (1996).
36) Bernardo, J.M. and Smith, A.F.M.: *Bayesian Theory*, John Wiley & Sons, Ltd. (1994).
37) 繁桝算男：ベイズ統計入門，東京大学出版会 (1985).
38) 渡部　洋：ベイズ統計学入門，福村出版（株）(1999).
39) Chen, R. and Li, T.-H.: Blind Restoration of Linearly Degraded Discrete Signals by Gibbs Sampling, *IEEE Trans. PAMI*, Vol.43, No.10, pp.2410–2413 (1995).
40) Morency, L.-P., Rahimi, A. and Darrell, T.: Adaptive View-based Appearance Model, *Proc. CVPR'03*, pp.803–810 (2003).

(平成 17 年 8 月 11 日受付)
(平成 18 年 4 月 4 日採録)

大塚　和弘（正会員）

平成 5 年横浜国立大学工学部電子情報工学科卒業．平成 7 年同大学院工学研究科博士課程前期修了．同年日本電信電話（株）入社．現在，NTT コミュニケーション科学基礎研究所研究主任．名古屋大学大学院情報科学研究科博士課程（後期）在学中．コンピュータビジョン，時系列画像解析，コミュニケーションシーンの分析に興味を持つ．第 55 回全国大会優秀賞，平成 9 年度電子情報通信学会学術奨励賞，IAPR 10th Int. Conf. Image Analysis and Processing Best Paper Award 各賞受賞．電子情報通信学会，IEEE 各会員．

竹前　嘉修

平成 11 年慶應義塾大学理工学部電気工学科卒業．平成 13 年同大学院理工学研究科修士課程修了．同年日本電信電話（株）入社．平成 13〜16 年 NTT コミュニケーション科学基礎研究所．平成 17 年より NTT サイバーソリューション研究所．平成 17 年より慶應義塾大学大学院理工学研究科後期博士課程在学中．映像コミュニケーション，人物の行動理解・解析，次世代ネットワークサービス等の研究開発に従事．平成 14 年度電子情報通信学会学術奨励賞受賞．電子情報通信学会会員．

大和　淳司

NTT コミュニケーション科学基礎研究所メディア情報研究部メディア認識研究グループ主幹研究員/グループリーダ．東京大学工学部精密機械工学科卒業，同大学院修士課程修了．MIT Electrical Engineering and Computer Science 修士課程修了．1990 年 NTT 入社．NTT ヒューマンインターフェース研究所等を経て，現在，NTT コミュニケーション科学基礎研究所．画像認識，ロボットビジョン，人間・ロボット間コミュニケーション等の研究に従事．博士（工学）．

村瀬　洋（正会員）

昭和 30 年生．昭和 55 年名古屋大学大学院工学研究科電気電子工学専攻修士課程修了．同年日本電信電話公社（現在の NTT）に入社．平成 4 年から 5 年にかけて米国コロンビア大学客員研究員．平成 15 年より名古屋大学大学院情報科学研究科教授．文字認識，画像認識，マルチメディア認識の研究に従事．工学博士．平成 6 年 IEEE-CVPR 最優秀論文賞，平成 7 年山下記念研究賞，平成 13 年高柳記念奨励賞，平成 14 年電子情報通信学会業績賞，平成 15 年文部科学大臣賞，平成 16 年 IEEE Trans.MM 論文賞ほか受賞．電子情報通信学会，IEEE-CS 各会員．