複数人対話を対象とした顔表情の共起パターンの確率モデルに基づく共感状態の推定

熊野　史朗† 　大塚　和弘† 　三上　弾† 　大和　淳司†

† NTT コミュニケーション科学基礎研究所　〒 243-0198 神奈川県厚木市森の里若宮 3-1
E-mail: †{kumano.shiro,otsuka.kazuhiro,mikami.dan,yamato.junji}@lab.ntt.co.jp

あらまし　本研究では，複数人対話における対話者間の共感状態を推定するという新しい研究の枠組みを提案する．ここでの共感状態とは，共感，無関心，反感の三種類の状態をとり，外部観察可能な行動を伴い対話者間で感情を伝達しあっている場面を対象として，二者の間に一つ存在する状態であると定義する．キーとなる対話者行動として，二者間での瞬時の感情伝達を可能とする表情と視線の組み合わせに注目し，各共感状態と視線の状態に対して二者の表情の間に特有の共起パターンが存在するという仮説を立てる．この仮説に基づき，共感状態の層，表情，視線及び発話からなる行動の層，及び，映像音声信号の層からなる階層的な対話の確率モデルを提案する．そして，このモデルを用いて，ベイズ推定の枠組みに基づき，観測が与えられたもとでの共感状態及びモデルパラメータの同時事後確率分布をマルコフ連鎖モンテカルロ法を用いて近似的に推定する．4 人対話のデータを用いた評価実験により，仮説の妥当性，及び，共感状態についての人間の間での判断のばらつきを提案手法により事後確率分布として精度よく推定できる可能性が示唆された．

キーワード　対話分析，感情推定，共感，表情認識，視線，マルコフ連鎖モンテカルロ法

1. はじめに

対面対話は，人が社会的生活を営む上で，他者との情報共有，相手の感情の理解，及び，意思決定などを行うための最も基本的なコミュニケーション形態である．そのため，近年，複数人対話の自動分析に関する研究が盛んに行われている [1]．もし，複数人対話においてシステムが，話者交代がどのような流れで行われているか，現在の主役は誰か，各対話者がどのような感情を抱いているのか，といった様々な状況を踏まえて対話を上手く盛り上げてくれるならば，対話はもっと楽しく，有意義で生産的なものになることだろう．しかし，これまで複数人対話の自動分析研究では，対話者の感情はほとんど対象とされてこなかった．

複数人対話を深く理解する上で重要なのは，対話者同士がどのような感情をどのように伝え合い，その結果としてどのような対人関係が構築されていったのか，あるいは，会議などの場面ではどのように合意が形成されていったのかを把握することである．感情を伝え合う際の基本要素は共感，無関心，及び，反感であると考える．例えば，対話者同士が繰り返し共感できれば，親密な対人関係が構築され，グループがひとつにまとまっていく．

このような共感や反感に関わる対話者行動としては，表情と視線が特に重要である．なぜなら，表情と視線が組み合わさることで，お互いに自分の感情を相手に対して瞬時に伝達することができるためである．まず，表情は感情的なメッセージが伝達される主要な経路である [2], [3]．そして，対話者間の表情の一致／不一致の傾向が共感や反感と深く関わる [4], [5]．対話者間の表情の一致傾向は共感を引き起こすとともに，共感がまた表情の一致傾向を引き起こす [4]．この共感時の表情の一致傾向としては，微笑に対する微笑 [4] といった肯定的な感情と深く関わる表情同士ばかりでなく，誰かが辛い思いをしているときに哀れみの表情を浮かべる [5] といった否定的な感情と深く関わる表情同士の場合もある．他方，反感や無関心は表情の不一致傾向と関係深い．

一方，視線については，様々な機能 [6], [7] のうち，特にモニタ機能及びトリガ機能が重要である．モニタ機能は他者の表情を観察してそこから相手の感情を読み取るために必須である．トリガ機能は，ある人物に視線を向けた際にその人物の視線を引きつけ，さらに反応をも引き起こす．例えば，人物 A が人物 B に視線を向けた際には，人物 B も人物 A に視線を向けることで相互凝視，つまり，アイコンタクトの状態となり，さらに，相互凝視の状態では対話者間の行動の一致傾向が高まる [5] ため，結果として共感が促進される．

このように対話中では，対話者同士がお互いに相手の視線や表情に対する反応として視線や表情を返すという短時間のインタラクションを繰り返しており，それにより二者の間で感情を伝達し合い共感や反感が生じている．その共感や反感がさらに次の行動を引き起こしもする．本研究では，このような外部から観察可能な何らかの行動を伴って対話者間で感情を伝え合って生じる共感，無関心，反感を対象とし，これら 3 つをまとめて共感状態と呼ぶ．

ここで，表情と視線の組み合わせによって二者間でど

のように共感状態が生成されるのかの例を挙げる．今，人物 A と人物 B が相互凝視中であるものとする．まず最初に A が B に対して微笑を投げかけ，B の肯定的な反応を期待しつつ B の感情を推察しようとする．この A からの微笑に対して，B は A に微笑み返すことで A に対して共感していることを表現する．A はその B からの微笑により B が自分に対して共感してくれていることを確認すると同時に，B も A が自分の感情を受け取ってくれたことを把握する．これにより，A と B がお互いに共感しあっていることを共有することとなる．もし，B が A に対して無関心であったり反感を抱いているとすれば，B は A の微笑に対して特に無表情のまま反応を示さない，苦笑を浮かべる，あるいは，目をそらせることで，その共感状態が二者の間で共有される．以上のように，共感状態と視線の状態は，二者の表情がどう組み合わさるのかに深く関係している．

本研究では，複数人対話の場面における対話者間の共感状態を推定するという新しい研究の枠組みを提案する．ベースとなっているのは，共感状態と視線のそれぞれの状態に対して二者の表情間に特有の共起パターンが存在するという独自の仮説であり，この仮説に基づく階層構造をもつ対話の確率モデルを提案する．このモデルは，マルコフ過程に従う共感状態の層と，表情，視線及び発話からなる行動の層，さらに，映像音声信号の層という 3 つの層からなる．提案モデルでは，インタラクションを行っている各対話者ペアが，それぞれの時刻において，共感，無関心，反感のいずれかの状態にあり，その共感状態と視線の状態により決まる表情の共起パターンに従って，両者がそれぞれ表情を表出することを仮定する．この対話者間の表情の共起のパターンを確率分布としてモデル化する．

提案手法では，この対話モデルに基づき，視線及び表情から共感状態を推定する．ベイズ推定の枠組みに基づき，観測情報が与えられたもとでの共感状態及びモデルパラメータの同時事後確率分布を推定する．共感状態については，事後確率分布のばらつきが人間の間での判断のばらつきに対応する．その事後確率分布を，複雑なモデルに対しても適用可能な，マルコフ連鎖モンテカルロ法の一種であるギブスサンプラー [8] を用いて近似的に推定する．

以下では，まず 2. にて関連研究の整理を行い，本研究の位置付けを明確にする．次いで，3. では提案モデルの主要部分である，表情共起パターンについて説明する．続いて 4. では，提案モデルの全体を説明するとともに，5. でその事後分布をギブスサンプラーを用いて近似する方法について述べる．次に 6. にて提案手法についての評価実験の結果を示す．最後に 7. にて，本研究のまとめを行うとともに，将来展望について併せて説明する．

2. 関 連 研 究

複数人対話における共感状態の自動推定の試みは，その重要性にもかかわらず，筆者の知る限りにおいて見当たらない．そこで，本章では，関連の深い従来技術として，(a) 表情認識，(b) 対人感情の推定，(c) グループ全体としての複数人対話状態の推定，及び，(d) ソーシャルネットワークの推定の 4 種類を順に取り上げ，本研究との違いを明確にする．

(a) 近年，表情認識研究分野では，従来の意図的で大げさな幸福，怒り，悲しみといった 6 基本表情から，自発的でより微細な表情認識対象のシフトが見られる [9], [10]. しかしながら，いずれも表情を内なる感情の表れと捉えて表出している人物のみに注目しており，本研究のように対話中に表出された微細な表情の対話上での意味や機能に着目した研究は現時点では皆無に近い．

(b) 文献 [11] では，表情が誰に対して向けられたのかに注目し，映像から認識される笑顔の表出量を，表出した人物がその相手に対して抱いている好意的感情の度合いと捉えて対人感情の推測の手掛かりとするアプローチが提案されている．つまり，表出された表情がその向けられた人物に対する感情を直接的に表すことが前提とされている．一方，本研究では，お互いに好んでいないモノや第三者に対する否定的な感情を，否定的な表情を表出しあって共感する場面も対象の範囲内である．

(c) 複数人対話に関する上位レベルの状態として，モノノーグやダイアローグといった対話のレジームの状態 [12] や，主導権を持った人物 [13] を自動的に推定する技術が提案されている．特に，文献 [12] では，レジームの状態，誰と誰がインタラクションしているのか，及び，対話者の行動からなる 3 層の階層構造の対話モデルを提案している．しかし，本研究のような対話者の感情に関わる状態は対象とされていない．

(d) ソーシャルネットワーク，あるいは，ソシオメトリと呼ばれる直敵的な人間関係のある人物同士をリンクでつないだ大規模なネットワーク構造を，物理的な近接性 [14] や画像中から計算される動き特徴 [15] といった低次の情報から推定する研究が存在する．しかし，共感という短い時間スケールで変化する対人感情を推定の対象とした研究は見当たらない．

3. 表情共起に関する仮説とその検証

本章では，共感状態を推定するための独自の仮説について妥当性を検証する．その仮説とは，共感状態の種類と視線状態に対して二者の表情間に特有の共起パターンが存在するという仮説である．仮説の検証には，6.1 にて述べる実際の 4 人対話のデータを用いる．

3.1 共感状態，表情及び視線の定義

まず，本研究における共感状態，表情，及び，視線を

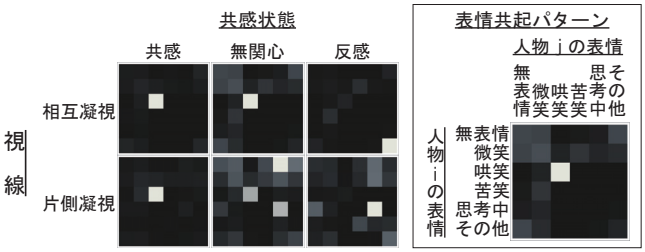

図 1: 表情共起パターンの一例：各ブロックは，共感状態及び視線の各状態に対する表情の共起パターンを表す．それらの 6×6 の行列はそれぞれ，人物 i と人物 j についての表情の共起パターンを表す．要素の色が明るいほど，その表情の組み合わせが起きやすいことを表す．片側凝視については，人物 i が見ている人物，人物 j が見られている人物を表す．

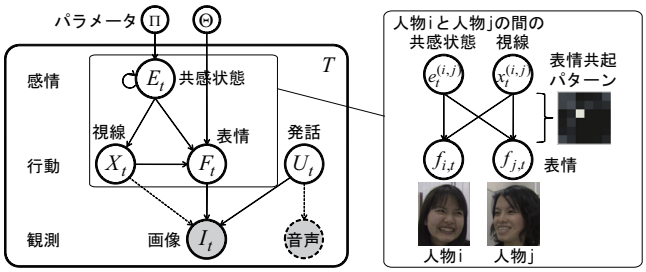

図 2: （左）本研究における対話の階層的モデルのグラフ表示：観測となるノード（映像信号 I）はグレーで塗られている．(右) 表情共起行列に関わる，共感状態，視線，及び，表情の間の関係を表すグラフ表示．

定義する．共感状態とは，それぞれの人物ペア，すなわち，二者の間に 1 つ存在する状態であり,「共感」,「無関心」，及び,「反感」の 3 状態のうちのいずれかの状態をとるものと定義する．

表情は人物毎に定義され，複数のカテゴリに分類できるものとする．本研究では，今回の対話データ中に比較的多く観察された,「無表情」,「微笑」,「哄笑」,「苦笑」,「思考中」及び「その他」の 6 つのカテゴリを対象とする.「その他」の表情は,「不満」,「攻撃」及び「驚き」といった表情を含んでいる．

人物ペア間の視線の状態としては,「相互凝視」,「片側凝視」，及び,「相互そらし」の 3 状態がある．相互凝視とは二者がお互いに相手の方を見合っている状態，すなわち，アイコンタクトをしている状態のことである．片側凝視とはペアの一方の人物のみが他方の人物を見ている状態のことである．相互そらしとは両者がお互いに相手の方を見ていない状態のことである．本研究では，相互そらし状態の場合にはいずれの種類の共感状態も行われていないものとみなす．

3.2 表情の共起パターン

本稿では，対話者ペアの間でどのような表情の共起がどのくらいの確率で生じるのかのパターンを行列の形で表したものを表情共起行列と呼ぶ．この表情共起行列は，図 1 に示すように，それぞれの共感状態及び視線の状態によって異なる．これらの表情共起行列の作成には，6. にて述べる，共感状態，表情，及び，視線についての状態が人手で付与された 4 人対話のデータ（約 30 分間）を用いた．

図 1 ではそれぞれの表情共起行列の間に明確なパターンの違いが見られる．これは，共感状態と視線の状態に対して二者の表情間に特有の共起パターンが存在するという独自の仮説の妥当性を示唆している．さらに，これらの表情共起のパターンの特徴の多くは，これまでに得られている心理学的知見と概ね一致する．まず，a) 共感時には肯定的表情（ここでは微笑及び哄笑）の共起が多く見られる [4]．b) 相互凝視は表情の一致傾向をより強める [5]．例えば，相互凝視，片側凝視，相互そらしの順で，肯定的表情の共起の頻度が高くなっていく．その他にも，反感時には，少なくとも一方の人物が苦笑，思考中といった表情を表出しやすい，無関心時には無表情と他の表情の組み合わせが顕著であるなど，我々の直感と概ね一致する傾向も見られる．以上から，この表情共起のパターンがある程度の汎用性を持ち，未知の対話者ペアに対しても適用可能なことが示唆される．

なお，図 1 に示した表情共起行列には，各表情の発生頻度に基づく正規化を施してある．まず，各表情の組み合わせの頻度を数え，次いで，各頻度をそれぞれの対話者がその表情を表出した頻度で割り，さらに，行列の要素が 1 になるようにスケーリングを行う．これをペア毎に行い平均化する．こうするのは，人はソーシャルな場面では微笑を多く表出しやすく [16]，単に各表情の組み合わせの頻度を求めただけでは，微笑同士の共起の頻度が突出して高くなり，共感状態や視線状態の間のパターンの明確な違いが現れないためである．

4. 対話モデルの提案

4.1 モデル構造の概要

共感状態と対話者の行動との間の関係性をモデル化する 1 つの方法として，本研究では階層的動的ベイジアンネットワークを用いる．ここで階層的とは，最上層の状態がマルコフ過程に従って遷移する離散的状態をとり，下位の層の状態を支配することを意味する．提案モデルにおいては，共感状態が上層であり，表情と視線を含む対話者行動が中層である．さらに，対話者行動はその下層である映像信号に影響を及ぼす．

図 2 に示す映像信号から共感状態を推定するための階層的動的ベイジアンネットワークを提案する．ここで，t は離散的な時間ステップを表し，$t = 1, \cdots, T$ である．また，ノードは確率変数を表し，矢印は変数間の確率的な因果関係を表す．図 2 のモデルでは，感情及び行動についての確率変数は，共感状態 $\{E_t\}_{t=1}^T$，表情 $\{F_t\}_{t=1}^T$，視線 $\{X_t\}_{t=1}^T$，及び，発話 $\{U_t\}_{t=1}^T$ の 4 つである．3. で検証した通り，各ペアの共感状態と視線の状態は，両者

の表情の共起パターンに影響を及ぼすと仮定する．また，共感状態は視線の状態にも影響を及ぼすとする．これは，今回使用した対話データに，共感や反感と判断された場面では相互凝視や片側凝視が，無関心においては相互そらしが多く生じていたことによる．

時刻 t における全対話者ペアの共感状態の集合を，$E_t = \{e_t^{(i,j)}\}_{(i,j)\in \boldsymbol{r}}$ にて表す．ここで，$e^{(i,j)} \in \boldsymbol{e} = \{1, \cdots, N_e\}$ は人物 i と人物 j のペア (i,j) の間の共感状態を表し，$\boldsymbol{r}$ は $N \times (N-1)/2$ 組ある対話者ペアの集合を表す．N_e は共感状態の数である．本研究では，共感状態 e は「共感」，「無関心」，「反感」のいずれかの状態をとる $(N_e = 3)$．時刻 t における全ての対話者ペアの視線状態の集合を $X_t = \{x_t^{(i,j)}\}_{(i,j)\in \boldsymbol{r}}$ にて表す．ここで，$x_t^{(i,j)}$ はペア (i,j) の視線状態を表し，$N_x = 3$ 種類，すなわち，相互凝視，片側凝視，相互そらしのうちのいずれか 1 状態をとる．時刻 t における全ての対話者の表情及び発話の状態の集合を，それぞれ，$F_t = \{f_{i,t}\}_{i=1}^N$ 及び $U_t = \{u_{i,t}\}_{i=1}^T$ にて表す．ここで，$f_i \in \boldsymbol{f} = \{1, \cdots, N_f\}$ は人物 i の表情状態を表し，N_f は表情カテゴリの数を表す．本研究で対象とする表情 f の種類は「無表情」,「微笑」,「哄笑」,「苦笑」,「思考中」，及び,「その他」$(N_f = 6)$ である．$u_{i,t}$ は人物 i が時刻 t において発話していない $(u = 0)$，あるいは，発話をしている $(u = 1)$ ことを表す．時刻 t における画像（映像中の 1 フレーム）を I_t にて表し，全ての対話者がこの画像中に含まれているものとする．

4.2 同時事後確率分布

本研究では，ベイズ推論の枠組みに従い，観測 Z から共感状態，未知の対話者行動，及び，モデルパラメータ φ を同時に推定する．本節では，その際に計算すべき全ての未知の確率変数の同時事後確率分布を提案モデルに基づき定義する．ここで，モデルパラメータとは，確率変数の間にどのような確率的因果関係があるのかを表すパラメータのことであり，例えば，共感状態，視線，及び，表情間の関係については，表情共起行列により表される．

本研究の目標は，画像信号の時系列 $I_{1:T}$ 及び音声信号の時系列のみから共感状態の時系列 $E_{1:T}$ を正しく推定することである．視線及び発話については，自動検出方法として，例えば文献 [17] 及び文献 [18] などが存在する．そこで，将来的にそのような技術を組み合わせることを前提として，ここでは，映像信号 $I_{1:T}$，視線 $X_{1:T}$，及び，発話 $U_{1:T}$ が観測 Z として与えられており，未知の確率変数が共感状態の時系列 $E_{1:T}$，表情の時系列 $F_{1:T}$ であるときの同時事後確率分布 $p(E_{1:T}, F_{1:T}, \varphi|Z)$ を推定する場合について説明する．

この推定すべき同時事後確率分布を図 2 のベイジアンネットワークに従って次のように展開する．

$$
\begin{aligned}
&p(E_{1:T}, F_{1:T}, X_{1:T}, I_{1:T}, \varphi) \\
&:= p(\varphi)P(E_{1:T}|\varphi)P(X_{1:T}|E_{1:T}, \varphi) \\
&\quad P(F_{1:T}|E_{1:T}, X_{1:T}, \varphi)P(I_{1:T}|F_{1:T}, \varphi). \qquad (1)
\end{aligned}
$$

以降，右辺について，説明の都合上，第 2 項から第 5 項，最後に第 1 項の順で説明する．なお，特に必要がなければ，簡略化のためモデルパラメータ φ は省略する．

まず，共感状態の事前確率 $P(E_{1:T})$ については，ここでは，共感状態がペア間で独立であり，それぞれ 1 次マルコフ過程に従うと仮定して，

$$
P(E_{1:T}) := \prod_{t=1}^{T} \prod_{(i,j)\in \boldsymbol{r}} P(e_0^{(i,j)})P(e_t^{(i,j)}|e_{t-1}^{(i,j)}) \qquad (2)
$$

と表す．ここで，$P(e_0^{(i,j)} = e)$ はペア (i,j) の共感状態が e である確率（初期確率）を，$P(e_t^{(i,j)} = e'|P(e_{t-1}^{(i,j)} = e)$ は共感状態が e から e' へと遷移する確率を表す．これらの確率はいずれも時間不変であり，ペア毎に用意される．

視線状態に対する共感状態の尤度 $P(X_{1:T}|E_{1:T})$ については，ペア及び時間についての独立性を仮定し，

$$
P(X_{1:T}|E_{1:T}) := \prod_{t=1}^{T} \prod_{(i,j)\in \boldsymbol{r}} P(x_t^{(i,j)}|e_t^{i,j}) \qquad (3)
$$

と定義する．ここで，$P(x_t^{(i,j)} = x|e_t^{(i,j)} = e)$ は，ペア (i,j) の共感状態が e であるときにそのペアの視線状態が x である確率を表す．

共感状態及び視線の状態が与えられたもとでの表情の条件付き確率 $P(F_{1:T}|E_{1:T}, X_{1:T})$ については，表情の事前確率，及び，表情共起行列の積，すなわち，

$$
\begin{aligned}
P(F_{1:T}|E_{1:T}, X_{1:T}) &:= \prod_{t=1}^{T}\prod_{i=1}^{N} P(f_{i,t}) \cdot \\
&\prod_{t=1}^{T} \prod_{(i,j)\in \boldsymbol{r}} \mathcal{M}^{(i,j)}_{e_t^{(i,j)}, x^{(i,j)}}(f_{i,t}, f_{j,t}) \qquad (4)
\end{aligned}
$$

と表現する．ここで，$P(f_{i,\cdot} = f)$ は人物 i の表情が f である事前確率であり，時不変性と対話者間での独立性を仮定している．$\mathcal{M}^{(i,j)}_{e,x}(f, f')$ は，ペア (i,j) について，共感状態が e でありかつ視線が x であるときに，人物 i の表情が f でありかつ人物 j の表情が f' である確率を表す．これは，共感状態が e かつ視線が x の状態についての表情共起行列の (f, f') 成分として表される．表情共起行列 $\mathcal{M}$ はそれぞれの共感状態及び視線の状態について存在し，合計で $N_e \times N_x$ 個存在する．それぞれの表情共起行列 $\mathcal{M}$ の大きさは $N_f \times N_f$ であり，それらの各行列の要素の和は 1 である．ただし，相互そらしの視線状態においてはどの表情の組み合わせも常に等しい確率で生じるものと仮定して，全ての要素が $1/(N_f \times N_f)$ である行列とする．各ペア及び各人物が表情の事前確率及び表情共起行列についてのパラメータを持つ．

観測画像に対する表情及び発話の尤度 $P(I_{1:T}|F_{1:T}, U_{1:T})$ については，対話者間の独立性を仮定し，

$$P(I_{1:T}|F_{1:T}, U_{1:T}) = \prod_{t=1}^{T}\prod_{i=1}^{N} P(I_t|f_{i,t}, u_{i,t}) \qquad (5)$$

と定義する．本研究では，人物 i の発話状態 u 用の表情認識器 $\mathrm{FER}_{i,u}(I)$ を用いて，$P(I_t|f_{i,t}, u_{i,t}) := P(\mathrm{FER}_{i,u_{i,t}}(I_t)|f_{i,t})$ と定義する．この表情認識器は画像 I を入力として表情の識別結果 $\tilde{f}$ を返す．今回用いた表情認識器の詳細については 6.2 にて述べる．

以上の共感状態に関するモデルパラメータ Π と表情に関するモデルパラメータ Θ を併せて $\varphi = \{\Pi, \Theta\}$ と表す．このパラメータの事前確率分布 $p(\varphi)$ は，それぞれのパラメータの事前分布の積として表される．各パラメータの事前分布 $p(\varphi)$ としては，数学的に扱いやすいという理由から事後分布に対する共役事前分布とし，中でも離散変数に対して一般によく使用されるディリクレ分布を使用する．これらのパラメータの事前分布のパラメータのことをここではハイパーパラメータと呼ぶ．

5. ギブスサンプラーを用いたベイズ推定

(1) 式の同時事後確率分布は複雑で解析的に解を得ることが困難である．そこで，本研究では，このような複雑な分布に対して近似解を得るのに優れた 1 つの方法として，マルコフ連鎖モンテカルロ法の一種であるギブスサンプラーを用いることとする．ギブスサンプラーは，推定する全ての確率変数を 1 つずつ順番にサンプリングすることを 1 セットとして，このセットを多数回繰り返す．各サンプリングにおいて，各変数を順番に，その変数以外の全ての変数を固定とした全条件付き事後分布からランダムに抽出する．マルコフ連鎖が収束した後の不偏分布は推定すべき事後分布に一致するため，対象の事後分布の統計量を抽出したサンプルから計算できる．

5.1 全条件付き確率分布からのサンプリング

各確率変数に対する全条件付き確率分布は，それ以外の変数が全て固定されているため，(1) 式で表される同時事後確率分布から対象の変数を含まない項を正規化定数とした分布となる．時刻 t における共感状態 $e_t^{(i,j)}$ についての全条件付き確率分布は，

$$\begin{aligned} &P(e_t^{(i,j)}|E_{1:T}\backslash e_t^{(i,j)}, F_{1:T}, \varphi, I_{1:T}) \\ &\propto P(e_t^{(i,j)}|e_{t-1}^{(i,j)})P(e_{t+1}^{(i,j)}|e_t^{(i,j)}) \\ &\qquad \mathcal{M}^{(i,j)}_{e_t^{(i,j)}, x_t^{(i,j)}}(f_i, f_j)P(x_t^{(i,j)}|e_t^{(i,j)}) \end{aligned} \qquad (6)$$

となる．同様に，時刻 t における表情の状態 $f_{i,t}$ についての全条件付き確率分布は，

$$\begin{aligned} &P(f_{i,t}|E_{1:T}, F_{1:T}\backslash e_t^{(i,j)}, \varphi, I_{1:T}) \propto P(f_{i,t}) \cdot \\ &\qquad \prod_j \mathcal{M}^{(i,j)}_{e_t^{(i,j)}, x_t^{(i,j)}}(f_i, f_j) \cdot P(I_t|f_{i,t}, u_{i,t}) \end{aligned} \qquad (7)$$

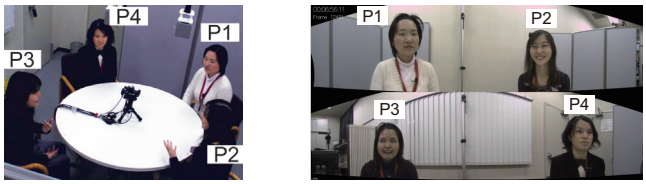

図 3: 対話シーンの一例.（左）デジタルカメラで撮影した俯瞰画像.（右）提案手法の入力とする全方位映像.

となる．

モデルパラメータについては，その事前分布として共役事前分布を使用することから，全条件付き確率分布が事前分布と同じ分布形状となる．ここでは全てのモデルパラメータがディリクレ分布に従うことを仮定している．よって，ハイパーパラメータが表す事前分布からサンプリングされた値に対して，対象事象がサンプリング時点における共感状態及び表情のサンプルにより表される発生回数を加えた値として求めることができる．この反復計算の詳細については例えば文献 [19] を参考されたい．

5.2 共感状態の事後分布及び表情の認識結果

推定結果となる各時刻の共感状態の事後分布（周辺事後確率分布）$\hat{e}_t^{(i,j)}$，及び，表情カテゴリの推定値 $\hat{f}_{i,t}$ については，ギブスサンプラーでの反復 $M'+1$ 回から $M(>M'+1)$ 回目までに得られたサンプル集合 $\{E_{1:T}^{(q)}, F_{1:T}^{(q)}, \varphi^{(q)}\}_{q=M'+1}^{M}$ から算出する．共感状態の事後分布については，$\hat{e}_t^{(i,j)} = e$ である確率を $\sum_{q=M'+1}^{M}\delta_e(e_t^{(i,j)(q)})$ として算出する．ここで，$\delta_\xi(\xi')$ は，$\xi = \xi'$ であれば 1 を，そうでなければ 0 を返す関数である．表情カテゴリの推定値は表情の周辺事後確率分布を最大化するカテゴリとして算出する．

6. 実　　験

本章では，提案手法が共感状態について判断の曖昧性を含めてどれだけ人間の判断に近い推定ができるのかについての定性的及び定量的な評価を行う．

6.1 対話データ

本研究では 4 人対話を対象とする．ここで使用した対話データは，図 3 に示すように，複数の被験者が円卓テーブルに座り，ある与えられた議題について 8 分間以内でグループで意見をまとめて報告する，というタスクのもとで行われたものである．初対面かつ同年代の女性 4 人からなるグループが 4 つの議題についてそれぞれ討論している．その 4 つとは，(対話 A) 男女どちらが得か，(対話 B) 結婚は必要かどうか，(対話 C) 公共空間での全面喫煙を法的に義務付けるかどうか，(対話 D) 恋愛と結婚は別か，である．対話は，卓上型のコンパクトな全方位カメラ・マイク統合システム [20] （図 3 左のテーブル上）を用いて撮影されたものである．撮影された画像（図 3 右）のサイズは $2448 \times (512 \times 2)$pixels，フレームレートは 30fps である．

6.2 推定に関する設定

提案手法の性能評価において 2 つの条件を設定した．1 つめ（条件 A と呼ぶ）は，視線及び表情が与えられたもとで，すなわち，観測 $Z = \{X_{1:T}, F_{1:T}\}$ のときに共感状態を推定するものである．これは，対話者の視線と表情から共感状態をどの程度正しく推定できるかという提案手法の基本性能を評価するための条件である．もう 1 つの条件（条件 B と呼ぶ）は，共感状態と表情の同時推定 ($Z = \{X_{1:T}, U_{1:T}, I_{1:T}\}$) である．なお，条件 A については，4. や 5. 中の各式において観測のみを含む項を正規化定数として扱えばよい．

表情認識器として今回はサポートベクターマシン (SVM) を用いた．特徴量として，幾何学的特徴，アピアランス特徴，及び，視線方向を用いた．幾何学的特徴として，文献 [21], [22] を参考にして，眉中心と目中心，目の左右外側の端点と口の左右の端点，鼻下端と唇下端のそれぞれの距離を使用した．アピアランス特徴には，歯がどのくらい見えているかの情報として，口領域の輝度値が閾値以上の画素数とした．視線方向としては，6.3 で述べる方法で与えた他者方向，上方向，下方向，その他の 4 方向を使用した．また，画像中の顔面上の顔特徴点の追跡には既存の追跡器 (*faceAPI*, Seeing Machines: http://www.seeingmachines.com/product/faceapi/) を使用した．以上の SVM を，各個人に対して発話時用と沈黙時用を用意した．クロス検定法に従い認識対象の対話以外の全ての対話データから学習を行った．

各モデルパラメータに対するハイパーパラメータについては今回使用する 4 つの対話データから学習した(注1)．これらのハイパーパラメータは，全ての対話データ，及び，全ての対話者ペアあるいは全ての対話者に対して共通であるとした．ただし，パラメータは確率変数であり，ギブスサンプラーによりその事後分布が推定される．ギブスサンプラーの反復回数については，経験的に，収束させるために破棄する回数を $M' = 600$，それ以降の推定結果算出に使用する回数を $M - M' = 200$ 回とした．

6.3 人手によるラベル付け

対話データ中の全てのフレームに対して，共感状態についてのラベルを 5 名のラベラがそれぞれ音声を聴かずに付与した．さらにそのうちの 1 名のラベラは表情及び視線についてのラベルも付けた．また別の 1 名のラベラが各対話者に対して発話しているのかしていないかのラベルを付けた．いずれのラベラも非対話参加者であった．これらのラベルはハイパーパラメータの学習，及び，手法の評価にのみ使用しており，推定時には視線及び発話情報のみ（条件 A についてはさらに表情）を観測として使用している．なお，視線ラベルから相互そらしであるフレームについては定量的な評価の対象外とした．ラベル付けの詳細については，文献 [23] を参照されたい．

(注1)：今回のデータにおいては，学習でなく人手で経験的に定めた値を使用した場合でも推定性能に大きな低下は見られていない．

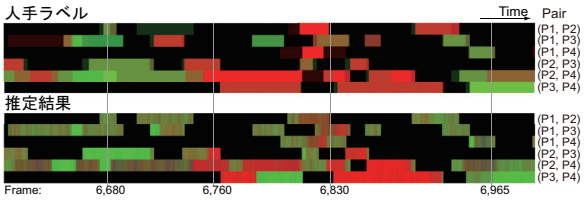

図 4: 共感状態の人手ラベル（上），及び，条件 A による推定結果（下）の時系列．横軸は時間を表し，対話 C のフレーム番号 6,600-7,000(≈ 13 sec) の範囲を表す．縦軸は対話者ペアを表す．色は共感状態についての人手ラベルの投票結果あるいは推定結果を表す．R，G，及び，B 成分は，それぞれ，共感，無関心，及び，反感の人手ラベルにおける得票率あるいは推定した事後確率を表す．すなわち，例えば，真赤はラベラ全員が一致して共感のラベルを付与したこと，あるいは，推定された共感の事後確率が他に比べて突出して高かったことを表している．一方，混色は，人手ラベルについてはラベラ間で判断が割れたことを，推定結果については事後分布が広がっていることをそれぞれ表す．

各共感状態の頻度は，15.2[%]（強い共感），29.2[%]（弱い共感），52.8[%]（無関心），2.5[%]（弱い反感），及び，0.3[%]（強い反感）であった．表情のラベルの頻度については，36.5[%]（無表情），52.4[%]（微笑），3.0[%]（哄笑），0.3[%]（苦笑），3.7[%]（思考中），及び，4.1[%]（その他）であった．

6.4 推定結果の定性的評価

図 4 及び図 5 に，共感状態についての人手ラベルと条件 A で推定した結果の一例を示す．それぞれの図の見方についてはキャプションにて説明している．図 4 については両者の色のパターンが近いほど，図 5 については両者の棒グラフ及び人物（ノード）間のリンクの色が近いほど，提案手法がラベラ間での共感状態の判断のばらつきを含めて人手ラベルに近い結果を出力していることを表す．両図より提案手法の有効性が示唆される．

6.5 推定結果の定量的評価

続いて定量評価を行う．本研究では人手ラベルの分布と推定した事後分布の分布間の類似度合いを評価尺度とする．というのも，そもそも人の感情というのは他者が直接観測して一つの正解値を得ることができないものであり，特にコミュニケーションにおいては，対話者の行動からどのような感情がどの程度推測できるのかが重要である．そして，この他者による感情の推測は，図 4 にて混色が多く存在することが意味する通り，本質的に個人間でばらつく性質がある．このため，本研究では，このような人間の間での判断の違いを分布として捉え，それを提案した対話モデルで事後分布という形で推測することを狙っている．よって，分布間の類似度でもって評価する

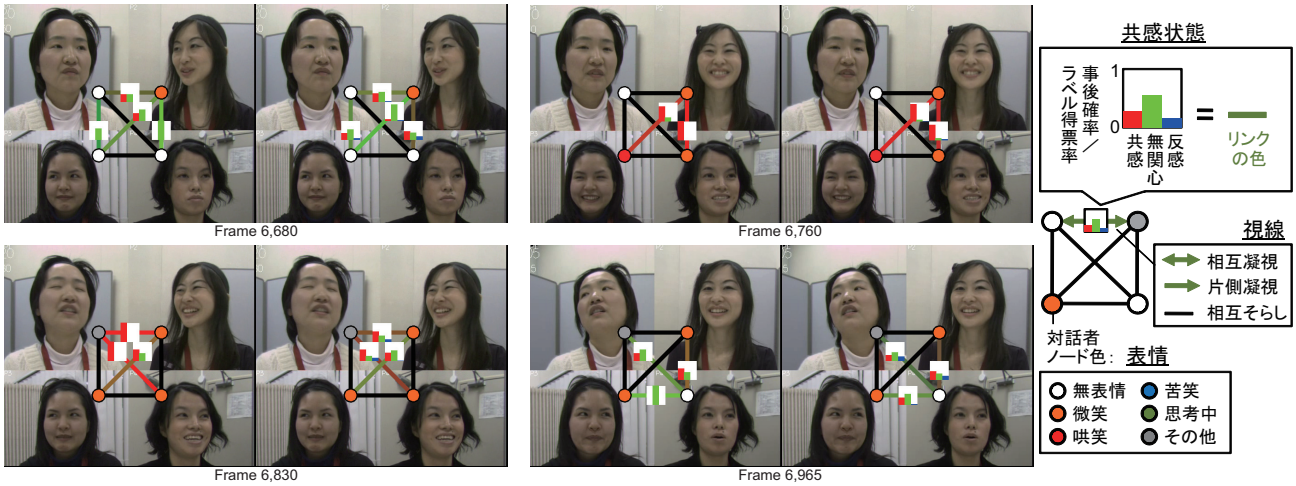

図 5: （左）人手ラベルと（右）条件 A での推定結果についてのスナップショット（図 4 中の 4 シーン）．それぞれの映像の上に重ねて描画されたグラフ構造については，ノードは対話者を表し，その色は人手により与えられた表情のカテゴリを表す．人物間をつなぐリンク，及び，その上に置かれた棒グラフは，そのペアについての共感状態を表す．リンクの色は図 4 と同じカラースキームで表現されている．棒グラフ中の左の赤，中央の緑，及び，右の青のバーは，それぞれ共感，無関心，及び，反感の人手ラベルにおける得票率あるいは推定した事後確率を表している．

表 1: 提案手法の推定性能

(a) 共感状態推定性能

条件	BC	最多得票クラスに基づく一致率			
	Total	Total	共感	無関心	反感
条件 A	0.874	0.673	0.729	0.628	0.671
条件 B	0.860	0.662	0.697	0.636	0.163

BC: Bhattacharyya coefficient

(b) 表情認識性能（再現率）

条件	Total	Ntr	Sml	Lgh	Wry	Thn	Oth
条件 B	0.597	0.642	0.610	0.682	0.122	0.489	0.117
FER	0.570	0.708	0.507	0.692	0.253	0.504	0.142

FER：表情認識器を単体で用いた際の性能
Ntr から Oth：無表情，微笑，哄笑，苦笑，思考中，その他

のが適切であると考える．ここでは，分布間の類似度の尺度として Bhattacharyya 係数 (BC) を用いる．BC は，確率分布 p 及び q に対して，$\mathrm{BC}(p,q) = \sum_e \sqrt{p(e)q(e)}$ $(0 \leqq \mathrm{BC} \leqq 1)$ にて計算される．

さらに，共感状態の推定に対する補足的な評価尺度として，パターン認識の問題でよく用いられる，推定値が正解値にどれだけ一致するかという値同士を比較する尺度についても掲載する．ここでは，人手ラベルの最多得票クラス（複数可）が推定結果の事後確率最大クラス $\hat{e}$ を含んでいれば一致とみなした．ここではこの尺度を最多得票クラスに基づく一致率と呼ぶ．なお，表情についても，推定値 $\hat{f}$ が人手ラベル（ラベラは 1 名）と一致していれば成功，そうでなければ失敗とみなす．

表 1 に提案手法の推定性能を示す．表 1(a) に示すとおり，視線と表情が与えられた場合（条件 A）では，共感状態の推定性能は高い BC(= 0.874) を示している．これは，6.4 でも述べたとおり，もし正確な視線及び表情の情報が与えられれば，提案手法の枠組みで共感状態を人間の判断のばらつきを含めて精度よく推定可能であることを示唆している．また，最多得票クラスに基づく一致率については，全ての共感状態の種類について 0.6 以上，平均で 0.673 と高い値が得られている．以上から提案手法の基本部分についての妥当性は示唆されたと考える．

表 1(a) に，さらに，共感状態と表情を同時に推定した場合（条件 B) の推定性能を示す．反感を除いて条件 A の結果と大差ない．また，表 1(b) に，条件 B での表情の認識率と表情認識器単体での表情の認識率を示す．どちらも，無表情，微笑，及び哄笑の認識率は高いが，苦笑，及び，その他の表情の認識率は比較的低い．この理由としては，まず，これらの表情では顔部品の移動量が顔部品追跡器の推定誤差と同程度あるいはより小さく，識別のために必要な情報が特徴量として含まれていない可能性があることが挙げられる．また，今回使用した対話データ中でこれらの表情の頻度が少なく，識別器の十分な学習ができていない可能性もある．

今回使用した識別器には改良の余地がまだあるものの，これらの表情は表出の強度が低いため，現時点での最先端の技術を適用しても高い認識率の達成は困難であると予想される．本研究によりこのような微細な表情を認識する重要性と難しさが明らかとなったことから，微細表情の認識は今後も引き続き取り組むべき問題と言える．提案手法の枠組みは，共感状態を各対話者の表情の表出に対する文脈情報として扱っていると捉えることもできるため，将来的に微細表情認識の性能が向上すれば，共感状態をより正確に推定できるようになるとともに，表情認識器単体の推定性能を上回る表情の推定性能を達成できる可能性もある．

7. まとめと今後の課題

本研究では，複数人対話中に対話者間で取り交わされる共感状態を推定するという新しい研究の枠組みを提案した．そのキーとなる対話者行動として，お互いに自分の感情を相手に対して瞬時に伝達することを可能とする表情と視線の組み合わせに注目した．その上で，それぞれの共感状態の種類と視線状態に対して二者の表情間に特有の共起パターンが存在するという性質に基づき，共感状態，対話者行動，映像音声信号の各層からなる階層的な確率モデルを提案した．このモデルを用いて，ベイズ推定の枠組みに基づき，観測情報が与えられたもとでの共感状態の事後確率分布をマルコフ連鎖モンテカルロ法を用いて近似的に推定する方法を提案した．4 人対話のデータを用いた評価実験により，共感状態についての人間の間での判断のばらつきを提案手法により事後確率分布として推定できる可能性が示唆された．

今後の課題として，まず，人数や性別といった異なる対話者構成や異なるトピックの対話に対する評価を行い，提案した枠組みの汎用性をより検証する予定である．また，手法の改良・拡張としては，共感状態について，同調圧力と呼ばれる現象をモデルに含めるため，ペア間の共感状態の相互作用を追加するとともに，より正確な現象の記述のために頭部ジェスチャや音声特徴などの他の行動要素をモデルに加える予定である．最後に，複数人対話中の対話者の間の感情の状態を推定するという問題は，非常に重要かつ先駆的な研究であると考える．

文　献

[1] D. Gatica-Perez: “Analyzing group interactions in conversations: a review”, Proc. IEEE Int’l Conf. Multisensor Fusion and Integration for Intelligent Systems, pp. 41–46 (2006).

[2] N. Chovil: “Discourse-oriented facial displays in conversation”, Research on Language and Social Interaction, **25**, pp. 163–194 (1991).

[3] N. Ambady and R. Rosenthal: “Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis”, Psychological Bulletin, **111**, 2, pp. 256–274 (1992).

[4] T. Chartrand and J. Bargh: “The chameleon effect: the perception-behavior link and social interaction”, Journal of Personality and Social Psychology, **76**, 6, pp. 893–910 (1999).

[5] J. B. Bavelas, A. Black, C. R. Lemery and J. Mullett: “”I show how you feel”: Motor mimicry as a communicative act”, J. Personality and Social Psychology, **50**, pp. 322–329 (1986).

[6] M. Argyle and J. Dean: “Eye contact, distance and affiliation”, Socimetry, **28**, pp. 289–304 (1965).

[7] A. Kendon: “Some functions of gaze-direction in social interaction”, Acta Psychologica, **26**, pp. 22–63 (1967).

[8] S. Geman and D. Geman: “Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images”, IEEE Trans. Pattern Analysis and Machine Intelligence, **6**, 1, pp. 721–741 (1984).

[9] M. Pantic and M. Bartlett: “Machine Analysis of Facial Expressions”, I-Tech Education and Publishing (2007).

[10] Z. Zeng, M. Pantic, G. I. Roisman and T. S. Huang: “A survey of affect recognition methods: Audio, visual, and spontaneous expressions”, IEEE Trans. Pattern Analysis and Machine Intelligence, **31**, 1, pp. 39–58 (2009).

[11] S. Kumano, K. Otsuka, D. Mikami and J. Yamato: “Recognizing communicative facial expressions for discovering interpersonal emotions in group meetings”, Proc. Int’l Conf. Multimodal Interfaces (ICMI-MLMI) (2009).

[12] K. Otsuka, H. Sawada and J. Yamato: “Automatic inference of cross-modal nonverbal interactions in multiparty conversations”, In Proc. Int’l Conf. Multimodal Interfaces (ICMI), pp. 255–262 (2007).

[13] D. Jayagopi, H. Hung, C. Yeo and D. Gatica-Perez: “Modeling dominance in group conversations from nonverbal activity cues”, IEEE Trans. Audio, Speech, and Language Processing, **17**, 3, pp. 501–513 (2009).

[14] N. Eagle, A. Pentland and D. Lazer: “Inferring social network structure using mobile phone data”, Proc. the National Academy of Sciences (PNAS), Vol. 106, pp. 15274–15278 (2009).

[15] L. Ding and A. Yilmaz: “Learning relations among movie characters: A social network perspective”, In Proc. European Conference on Computer Vision (ECCV2010), Vol. IV, pp. 410–423 (2010).

[16] R. E. Kraut and R. E. Johnston: “Social and emotional messages of smiling: An ethological approach”, J. Personality and Social Psychology, **37**, 9, pp. 1539–1553 (1979).

[17] S. Gorga and K. Otsuka: “Conversation scene analysis based on dynamic Bayesian network and image-based gaze detection”, Int’l Conf. Multimodal Interfaces and Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI) (2010).

[18] M. Fujimoto, K. Ishizuka and T. Nakatani: “A voice activity detection based on the adaptive integration of multiple speech features and a signal decision scheme”, In Proc. Int’l Conf. Acoustics, Speech and Signal Processing (ICASSP), pp. 4441–4444 (2008).

[19] R. Chen and T. H. Li: “Blind restoration of linearly degraded discrete signals by gibbs sampling”, IEEE Trans. Signal Processing, **43**, 10, pp. 2410–2413 (1995).

[20] K. Otsuka, S. Araki, K. Ishizuka, M. Fujimoto, M. Heinrich and J. Yamato: “A realtime multimodal system for analyzing group meetings by combining face pose tracking and speaker diarization”, In Proc. Int’l Conf. Multimodal Interfaces (ICMI), pp. 257–264 (2008).

[21] M. Pantic and L. Rothkrantz: “Expert system for automatic analysis of facial expression”, J. Image and Vision Computing, **18**, 11, pp. 881–905 (2000).

[22] R. Kaliouby and P. Robinson: “Real-time inference of complex mental states from facial expressions and head gestures”, In Proc. IEEE Int’l W. Real Time Computer Vision for Human Computer Interaction at CVPR, pp. 181–200 (2004).

[23] S. Kumano, K. Otsuka, D. Mikami and J. Yamato: “Analyzing empathetic interactions based on the probabilistic modeling of the co-occurrence patterns of facial expressions in group meetings”, In Proc. the IEEE Int’l Conf. on Automatic Face and Gesture Recognition (FG) (2011).