特集論文 「知的対話システム」

潜在キャラクタモデルによる聞き手のふるまいに基づく対話エンゲージメントの推定

Engagement Recognition from Listener's Behaviors in Spoken Dialogue Using a Latent Character Model

井上 昂治 Koji Inoue
京都大学 大学院情報学研究科 Graduate School of Informatics, Kyoto University
inoue@sap.ist.i.kyoto-u.ac.jp

Divesh Lala （同 上）
lala@sap.ist.i.kyoto-u.ac.jp

吉井 和佳 Kazuyoshi Yoshii （同 上）
yoshii@sap.ist.i.kyoto-u.ac.jp

高梨 克也 Katsuya Takanashi （同 上）
takanasi@sap.ist.i.kyoto-u.ac.jp

河原 達也 Tatsuya Kawahara （同 上）
kawahara@sap.ist.i.kyoto-u.ac.jp

keywords: dialogue, engagement, behavior, character, latent model

Summary

This article addresses the estimation of engagement level based on the listener's behaviors such as backchannel, laughing, head nodding, and eye-gaze. Engagement is defined as the level of how much a user is being interested in and willing to continue the current interaction. When the engagement level is evaluated by multiple annotators, the criteria for annotating the engagement level would depend on each annotator. We assume that each annotator has its own character which affects the way of perceiving the engagement level. We propose a latent character model which estimates the engagement level and also the character of each annotator as a latent variable. The experimental results show that the latent character model can predict the engagement label of each annotator in higher accuracy than other models which do not take the character into account.

1. は じ め に

近年，会話ロボットなどの知的対話システムが様々な場面で実用化されている．これらのシステムは，質問応答などの特定のタスクにおいては，適切な応答を生成することができる [DeVault 14, Higashinaka 14, Skantze 15, Wilcock 15]．しかしながら，そこでなされるインタラクションは，人間どうしの自然なそれとは異なる．例えば，コマンドなどの小語彙をマイクに向かって丁寧に発話される傾向にある [河原 13]．将来，会話ロボットなどが，社会で人間と共生するためには，自然でかつ円滑なインタラクションを指向する必要がある．

対話システムとのインタラクションに求められる要素の一つとして，本論文では対話におけるエンゲージメントを扱う．エンゲージメントとは，対話参与者間において，コミュニケーションが成立，維持，終了する過程を表す [Cerrato 16, Sidner 02, Sidner 05]．つまり，エンゲージメントが成立している場合には，参与者間でのコミュニケーションの質が保証されることになる．ヒューマン・コンピュータ・インタラクションの分野では，ユーザの状態に焦点があてられており，どの程度対話に対して興味や意欲があり，対話の継続を望んでいるかを表す指標としてエンゲージメントが用いられている．ユーザのエンゲージメントを推定することで，適応的な対話システムの行動やふるまいを生成することが考えられる．この適応的な行動やふるまいは，社会的スキル [Breazeal 04] と呼ばれ，人間との共生実現を目指す対話システムにとって重要な能力であるといえる．

本論文では，ユーザが聞き手であるときの，そのふるまいに基づいてエンゲージメントの推定を試みる．聞き手のふるまいとは，相槌，笑い，うなずき，視線などを表す．これらのふるまいは，話し手の発話に対する聞き手の反応であるだけでなく，聞き手の心的状態の表出であるともいえる [高梨 09]．したがって，これらのふるまいはエンゲージメントと相関すると期待される．エンゲージメントの推定モデルを構築するためには，その正解デー

タが必要である．ユーザの真のエンゲージメントを対話中に記録することは困難であるため，通常は第三者のアノテータが対話を後から観察して知覚したエンゲージメントで近似する．しかしながら，エンゲージメントの判断は主観的であり，その結果はアノテータにより異なることがしばしばである．この問題に対処するために，ユーザのエンゲージメントを知覚する側であるアノテータ，または対話システムは，各自のキャラクタを潜在的に保持しており，このキャラクタはエンゲージメントの知覚に影響を及ぼす，という仮定をおく．そして，エンゲージメントの推定のための潜在キャラクタモデルを提案する．このモデルは，各アノテータのキャラクタを潜在変数として，キャラクタの分布とエンゲージメントの分布を同時にデータから学習する．これにより，エンゲージメントを知覚する要因に関して，アノテータ間での共通部分または差異を理解することができる．また，各アノテータに適したエンゲージメントの推定結果を出力することができる．

本論文の構成を以下に示す．2 章では，エンゲージメントの定義とその推定手法についての関連研究を概観する．3 章では，本研究で用いる対話データとエンゲージメントのアノテーションについて述べる．4 章では，提案する潜在キャラクタモデルについて説明する．5 章では，提案モデルの推定精度を評価する．また，実際にデータから学習した提案モデルの分布について分析する．最後に，本論文のまとめと今後の展望について，6 章で述べる．

2. 関 連 研 究

エンゲージメントに関する研究の源流は Goffman の文献 [Goffman 66] まで遡ることができる．その後，多数の研究がなされ，その都度エンゲージメントの定義も様々ではあるが，ある程度の共通点がみられる [Glas 15]．コミュニケーションが成立する過程に焦点をあてると，エンゲージメントの定義として以下が挙げられる．

> “the process by which two (or more) participants establish, maintain, and end their perceived connection” [Sidner 02]

ヒューマン・コンピュータ・インタラクションの分野では，ユーザの状態に焦点があてられる傾向にあり，エンゲージメントの推定を想定して，以下のような定義がなされている．

> “the value that a participant in an interaction attributes to the goal of being together with the other participant(s) and of continuing the interaction” [Poggi 07]

> “how much a participant is interested in and attentive to a conversation” [Yu 04]

エンゲージメントの推定はこれまでに様々な研究で取り組まれてきた．多くの場合で，エンゲージメントが高いか否かの二値分類問題として定式化されている．特徴量には，言語情報は音声認識誤りや背景雑音などの影響を受けるため，非言語情報が主に用いられてきた．本研究で扱う聞き手のふるまいもこの背景に基づく．先行研究で最も用いられてきた特徴量は，視線のふるまいである [Bohus 10, Morency 06, Nakano 10, Peters 05]. 視線のふるまいは視覚的な注意の表出であり，エンゲージメントと相関すると考えられる [Langton 00]．他の特徴量として，音響特徴量 [Yu 04]，空間位置情報 [Michalowski 06]，頭部や体の方向 [Kuno 07] などが検討されてきた．

エンゲージメントの推定モデルとして，近年ではルールベースではなく機械学習によるアプローチがとられている．Yu ら [Yu 04] は，電話会話における音響特徴量について，サポートベクトルマシン（SVM）や隠れマルコフモデル（HMM）を用いてエンゲージメントを推定した．Nakano ら [Nakano 10] や石井ら [石井 11] は，エージェントがユーザに対して商品説明を行う場面において，ユーザの視線のふるまいを N-gram 遷移により表現し，エンゲージメントに基づくクラスタリングや，SVM によるエンゲージメントの推定を行った．Xu ら [Xu 13] は，移動型ロボットとの対話において，音響特徴量，頭部方向，ユーザとロボットとの距離などのマルチモーダル特徴量を SVM によりモデル化することでエンゲージメントを推定した．ここでのエンゲージメントは，ユーザがロボットとの会話を開始しようとする，または発話権を取得しようとする程度を表していた．千葉ら [千葉 16] は，Wizard-of-Oz（WOZ）形式のインタビュー対話において，音響，言語，画像のマルチモーダル特徴量を，SVM，ランダムフォレスト，ニューラルネットによってモデル化することで，インタビューを受ける側の対話意欲度を推定した．また，Huang ら [Huang 16] の研究のように，人間どうしの対話におけるエンゲージメントの推定も取り組まれている．ここでは，対話参与者の画像データを特徴量とする畳み込みニューラルネットが推定モデルとして用いられた．

最近では，エンゲージメントを推定した後に，対話システムがどのような行動およびふるまいを生成するべきかについての研究も着手されている．前述の Xu らの研究 [Xu 13] では，推定したエンゲージメントに応じてターンテイキングにおけるふるまいが制御されていた．たとえば，対話システムの発話中に，ユーザのエンゲージメントが高いと判断した場合には，ユーザへ発話権を譲る，などの機能が実装された．被験者実験の結果より，対話システムの人間らしさや知的さなどの主観評価が有意に向上することが確認された．Yu ら [Yu 16] は，WOZ 形式の対話において，人手により判断されたエンゲージメントに応じて対話戦略を切り替える対話システムを実装した．たとえば，ユーザのエンゲージメントが低下した

場合には，現在の対話モジュールを別のものへとランダムに切り替える．これにより，被験者実験では，システム発話の適確さなどの主観評価が向上した．また，遠隔操作されたアンドロイドとの対話において，ユーザのエンゲージメントと，直後のターンテイキングに関するふるまいとの関係について分析が行われた [Inoue 16a]．その結果，ユーザのエンゲージメントが高い場合には，直後のユーザのターンにおける継続時間長とアンドロイド側がうつ相槌の頻度が増加することがわかった．本研究では，エンゲージメントを推定した後の対話システムの行動およびふるまいについては扱わないが，エンゲージメントレベルを推定する目的は上記の研究と同様である．

本研究では，多様なふるまいからエンゲージメントを推定する．先行研究の多くは，視線などの単独のふるまいを用いるか，センサーから得られた音響や画像の信号情報をそのまま用いる場合がほとんどである．ここでのふるまいとは，信号情報とエンゲージメントとの中間状態に相当し，客観的に検出可能なものである．ただし，ここでは聞き手のふるまい，つまり相槌，笑い，うなずき，視線などの非言語的ふるまいに焦点をあてる．多様なふるまいに基づくエンゲージメントの推定モデルを構築することで，これらの関係性について調べる．この関係性が明らかになれば，将来，対話システムが自身のエンゲージメントを表現するときに，適切なふるまいを用いることができると期待される．

さらに本研究では，アノテータ間でのエンゲージメントの判断の違いについて検討する．先行研究では，複数のアノテータによるエンゲージメントのアノテーションが行われていた．しかしながら，エンゲージメントの判断は主観的であり，アノテータ間での不一致も散見される．最終的には多数決，または一致した箇所のみを用いるなどの方法がとられ，アノテータ間での判断の違いは検討されていなかった．これに対して，アノテータ間での判断の違いを考慮した推定モデルを提案する．これにより，推定精度の向上だけでなく，アノテータ間の判断傾向の相違点や共通点をデータから学習および理解することが可能になる．エンゲージメントの推定以外のタスクにおいて，アノテータ間での判断の違いを考慮した従来研究について述べる．Ozkan ら [Ozkan 10, Ozkan 11] は，相槌の予測において，アノテータ毎に予測モデルを作成し，各モデルの予測結果を潜在変数により統合する二段階の手法を提案した．本研究で提案するモデルは，上記のようにアノテータ毎にモデルを作成するのではなく，潜在変数とそれに応じた予測モデルを同時にデータから学習する．したがって，各アノテータの学習データが少ない場合でも，モデルの学習が頑健であることが期待される．熊野ら [Kumano 13, 熊野 17] は，対話行動情報に基づく共感状態の推定において，アノテータの特性（性格特性）による判断傾向の違いをモデル化した．正解ラベルが与えられたもとで，アノテータの特性と対話行動情報は独立であると仮定して，アノテータの特性に基づく推定と，対話行動情報に基づく推定とをそれぞれ独立にモデル化した．本研究では，アノテータの特性と入力である聞き手のふるまいには依存性がある（アノテータによってふるまいのとらえ方が異なる）とみなし，これらを同時に扱うモデルを提案する．ただし，熊野らの手法は未知のアノテータに対しても，アノテータの性格特性を指定すれば，これに応じた推定が可能である．

図 1 対話収録の様子

3. エンゲージメントのアノテーション

エンゲージメントをアノテーションするために用いた対話データについて述べる．我々は，自律型アンドロイド ERICA [Glas 16, Inoue 16b] を用いたヒューマン・ロボット・インタラクションコーパスの構築に取り組んでいる．今回用いた対話データは，被験者と遠隔操作された ERICA との一対一対話である．ここでは，ERICA を遠隔操作した別の被験者を ERICA 役と呼ぶ．対話のシナリオは以下の通りである．ERICA は，ある研究室の秘書であり，被験者はその研究室の教授に用事があり来訪した．しかし，教授は不在のため，教授が戻るまで ERICA と対話をする．ERICA のふるまいを統制するための ERICA 役への教示として，対話の進め方，社会的役割，研究室に関する背景知識を伝えた．対話の進め方として，来訪者の用件が主題ではあるが，その際に雑談をしてもよいと伝え，またそのための話題一覧を与えた．話題は，天気や研究室へどのようにして来たのか，などの初対面対話でよく見られるものを選んだ．研究室秘書の社会的役割として，来訪者の目的が達成されるようにすること，好意的に接すること，終了後に来訪者の用件やどのように対応したかを報告すること，などを示した．研究室に関する背景知識として，当該研究室の教員や学生に関する情報，また大学秘書に関する一般常識を与えた．対話の長さは約 10 分である．図 1 に対話の様子，および収録に用いたセンサなどを示す．両者は椅子に座り，テーブルを挟んで対面し，ERICA 役はその様子を映像と音声でモニタリングした．ERICA 役は口元のマイクに向かって発話し，その音声は ERICA 本体の近くに設置したス

ピーカからそのままリアルタイムに再生される．ERICA 役の発話の音韻情報を基に，ERICA の発話中の口唇および頭部動作を自動的に生成している [Ishi 12, 境 16]．また，ERICA 役は ERICA のうなずきや視線のふるまいを手元のコントローラによって操作した．収録に用いたセンサは，ショットガンマイクロフォン，16 チャネルマイクロフォンアレイ，RGB カメラ，Kinect v2 である．収録した対話データに対して，発話，相槌，笑い，フィラー，うなずき，視線，談話行為，隣接ペア，ターンなどの情報を人手により付与した．

本研究では，上記の対話コーパスから 20 セッションを用いてエンゲージメントのアノテーションを行った．この 20 セッションでは，被験者はすべて異なる人物であったが，ERICA 役は複数人が交代で務めた．被験者の年代は 10 代から 70 代までの男性 8 名，女性 12 名であった．ERICA 役は 20 代から 30 代の女性 6 名であった．これらすべての対話参加者は日本語母語話者であった．

エンゲージメントのアノテーションは，上記の対話には参加しなかった別の女性 12 名により行ってもらった．彼女らはすべて大学学部または大学院に所属している．以下，彼女らをアノテータと呼ぶ．各対話セッションに 12 名のうちの 5 名のアノテータをランダムに割り当てた．つまり，各アノテータは 8 または 9 セッションのアノテーションを行った．アノテータへの指示は以下の通りである．エンゲージメントの定義は先行研究にならい，「どの程度対話に興味や意欲があり，対話の継続を望んでいるか」とした．アノテータには，ERICA 役の視点に立ち，被験者正面を映した映像と音声を確認しながら，エンゲージメントを判定してもらった．具体的には，以下の条件をすべて満たした瞬間に手元のボタンを押してもらった．

(1) 被験者が ERICA の話を聞いているとき，つまり聞き手であるとき
(2) 被験者が聞き手のふるまいを表出しているとき
(3) エンゲージメントが高くなったと判断したとき

上記を判断するために，アノテータにいくつかの補助情報を提示した．まず，(1) のために，被験者が聞き手であるときの時間区間を，映像の下に再生時間と連動する形で表示した．これにはターンの情報を用いた．また，(2) のために，予めエンゲージメントと相関すると考えられる聞き手のふるまいのリストを提示した．このリストには，表情，笑い，視線，相槌，うなずき，姿勢，肩の動き，腕・手の動きを含めた．ただし，最終的にアノテータから付与されるラベルは，エンゲージメントが高くなった瞬間の時間情報のみである．ふるまいの表出の判断については，アノテーション作業の効率性から記録しなかった．したがって，このアノテーション作業は対話データと同じ時間で行うことができる．機械学習を行うためには，大量のデータを効率的に収集する必要がある．そのため，必要かつ重要な情報をできるだけ効率的に収集する方法として，このアノテーション方法を着想した．各セッションのアノテーション作業後に，アノテータに対してアンケート調査を行った．アンケートの内容は，エンゲージメントの判断方法，アノテーション作業の簡易さ，アノテーション結果に対する自信度などである．アノテーション作業の簡易さは，7 点の間隔尺度（1 ＝とても難しい，7 ＝とても簡単）で，平均 3.61，標準偏差 2.02 であった．また，アノテーション結果に対する自信度は，同様の尺度（1 ＝全く自信がない，7 ＝とても自信がある）で，平均 3.67，標準偏差 1.86 であった．したがって，ある程度の簡易さと結果への自信があるといえる一方で，エンゲージメントの判断は容易ではないこともうかがえる．

表 1 エンゲージメントの判断に有用だったと回答された回数

聞き手のふるまい	回数
表情	77
相槌	67
うなずき	65
視線	40
笑い	39
姿勢	32
肩の動き	3
腕・手の動き	2
その他	4

アノテーション結果の概要を以下に示す．1 セッションあたりの 1 人のアノテータがボタンを押した回数の平均は 18.13，標準偏差は 12.88 である．このばらつきは，セッションまたはアノテータによるものかを比較した．まず，各セッションにおけるアノテータ間での平均回数を求めたところ，この値のセッション間での標準偏差は 8.08 であった．また，各アノテータにおけるセッション間での平均回数を求めたところ，この値のアノテータ間での標準偏差は 6.99 であった．したがって，セッション間だけでなく，アノテータ間でもボタンを押す回数にある程度のばらつきがあるといえる．アノテータ間での一致率も算出した．一致率を算出するために，ERICA のターンを一致率を判断する単位とした．つまり，ある ERICA のターン内で，あるアノテータがボタンを 1 回以上押せば，そのターンはエンゲージメントが高いとみなした．ERICA のターン数は 1 セッションあたりの平均 33.35 回である．各セッションですべてのアノテータのペアについて Cohen のカッパ係数を求めた．これを全セッションでのすべてのペアについて平均をとると 0.338 であり，標準偏差は 0.215 であった．したがって，アノテータ間での一致率は高いとは言えず，アノテータ毎にエンゲージメントの判断傾向が異なることが推察される．アノテーション作業後のアンケート調査では，どのふるまいがエンゲージメントが高いと判断するときに有用であったかを複数選択してもらった．回答結果を表 1 に示す．ただ

し，このアンケート調査は計 100 回（5 人のアノテータに対して 20 セッション）行ったものである．この結果から，表情，相槌，うなずき，視線，笑い，姿勢がエンゲージメントに相関すると考えられる．ふるまいとエンゲージメントとの関係についてデータのみから分析することも考えられるが，可能性があるすべてのふるまいについてのアノテーションデータが必要である．本研究では，関係するふるまいを発見するためにアンケート調査を採用した．後の実験では，上記のふるまいのうち，相槌，うなずき，視線，笑いの 4 種類について，その生起を人手によりアノテーションしたものを特徴量として用いる．表情と姿勢はその曖昧性からアノテーション作業が比較的困難であるため今回は用いないが，将来的には特徴量に加える予定である．ここでの相槌の定義は，Den らの分類 [Den 11] における，応答系感動詞と感情表出系感動詞に対応する．また，視線のふるまいは，ある ERICA のターン内で，ERICA の頭部への連続注視時間が 10 秒以上の場合に生起とした．この 10 秒は連続注視時間のヒストグラムの分布から，ある程度の数が生起することを確認したうえで判断した．ただし，0.5 秒以下の視線の逸脱は無視した．全セッションの ERICA の各ターンにおいて，エンゲージメントが高いと判定したアノテータの人数と，これら 4 種類のふるまいの生起とのスピアマンの順位相関係数は，それぞれ相槌が 0.381，笑いが 0.245，視線が 0.362，うなずきが 0.549 であった（いずれも $p < 0.001$）．

4. 潜在キャラクタモデルによるエンゲージメントの推定

前章のアノテーションの結果からわかるように，エンゲージメントは，それを知覚する側（アノテータ）の主観に左右される．つまり，アノテータ毎に非言語行動の解釈の方法が異なると推察される．この要因として文化，性別，性格などの違いが挙げられる．ここでは，アノテータの性格（キャラクタ）の違いが主であると考える．キャラクタを分類する軸として，Big Five と呼ばれる 5 つの因子，外向性，情緒不安定性，開放性，誠実性，調和性がある [Barrick 91]．また，このキャラクタを将来的に対話システムに実装することで，設定されたキャラクタに応じたエンゲージメントの推定，および適応的な行動やふるまいの生成が期待される．以上より，アノテータ間でのキャラクタの違いを考慮したエンゲージメントの推定モデルを検討する．

問題設定について述べる．エンゲージメントの推定は対話システム側（ERICA）のターン毎に行う．入力は，前章で定義した聞き手の各ふるまいの生起を表す二値とする．入力特徴量の表現方法に関しては，将来的には，複数回の生起を表現するための非負の整数や，ふるまいの自動検出を想定した連続値へと拡張する予定である．出力はエンゲージメントが高いか否かの二値である．これは前章のアノテーション結果に対応する．ただし，あるターンでアノテータが手元のボタンを 1 回以上押した場合に，そのターンはエンゲージメントが高いとする．

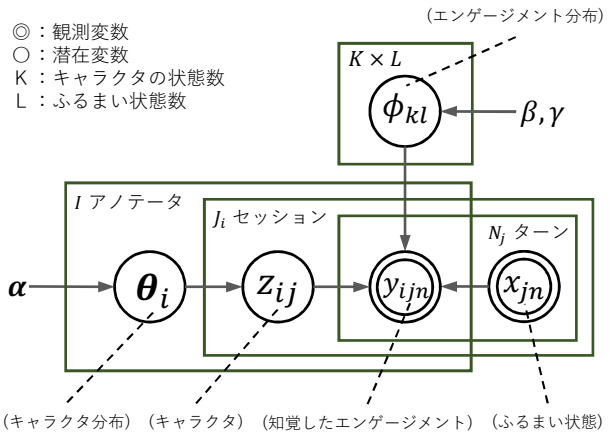

図 2 潜在キャラクタモデルのグラフィカルモデル

エンゲージメントを判断する側である対話システム（アノテータ）のキャラクタを潜在変数とする階層ベイズモデルを提案する．本論文では，これを潜在キャラクタモデルと呼ぶ．このモデルは潜在ディリクレ配分法（LDA: latent dirichlet allocation）[Blei 03] などの潜在変数モデルを参考にしている．エンゲージメントとキャラクタの分布を同時にデータから学習することで，アノテータ毎に異なる適切なエンゲージメントの判断結果を出力する．例えば，ユーザの笑いが生起した場合に，あるキャラクタを持つアノテータはその笑いからエンゲージメントが高いと判断するが，別のキャラクタを持つアノテータは異なった判断をする，といったことが可能になる．つまり，複数のキャラクタによるエンゲージメントの判断を同時に推定することができる．以下では，複数のふるまいの統合方法について，特徴量レベルの統合（4·1 節）と確率レベルの統合（4·3 節）の 2 種類について述べる．まず，前者を基にして潜在キャラクタモデルを説明する．

4·1 階層ベイズモデル

潜在キャラクタモデルのグラフィカルモデルを図 2 に示す．生成過程を以下に述べる．まず，各アノテータにおいて，キャラクタの分布が生成される．

$$\boldsymbol{\theta}_i \sim \mathrm{Dir}(\boldsymbol{\alpha}) \,,\, 1 \leq i \leq I \tag{1}$$

ただし，i はアノテータのインデクス，I はアノテータの数，Dir はディリクレ分布，$\boldsymbol{\alpha}$ はハイパーパラメータをそれぞれ表す．また，$\boldsymbol{\theta}_i$ の各要素は下記で定義される．

$$\boldsymbol{\theta}_i = (\theta_{i1}, \cdots, \theta_{ik}, \cdots, \theta_{iK}) \tag{2}$$

ここで，K はキャラクタの状態数に対応する．続いて，各アノテータが，各対話セッションにおいて持つキャラク

タが生成される．

$$z_{ij} \sim \text{Cat}(\boldsymbol{\theta}_i)\ ,\ 1 \leq j \leq J_i \tag{3}$$

ただし，j はセッションのインデクス，J_i は i 番目のアノテータがアノテーション作業を行ったセッション数，Cat はカテゴリカル分布をそれぞれ表す．つまり，アノテータ毎に 1 つのキャラクタの分布 $\boldsymbol{\theta}_i$ を持ち，これに基づいて，アノテータとセッションの組合せ毎にキャラクタ z_{ij} が 1 つ決まる．各アノテータが持つキャラクタは各対話セッション内では一貫して同じであるとする．ここで，j 番目のセッションでの，n 番目の対話システムのターンにおいて，ふるまい b の生起を $x_{bjn} \in \{0,1\}$ で表す．また，各ふるまいの生起の組合せ x_{jn} をふるまい状態と呼び，以下で表す．

$$x_{jn} := \text{comb}\left(x_{1jn}, \cdots, x_{bjn}, \cdots, x_{Bjn}\right)\ ,\ 1 \leq n \leq N_j \tag{4}$$

ただし，B はふるまいの種類数，N_j は j 番目のセッションにおける対話システムのターン数，$\text{comb}\left(\cdot\right)$ は $\cdot$ の組合せの種類をそれぞれ表す．また，ふるまい状態 x_{jn} がとり得る状態数 L は，今回の問題設定では 2^B である．キャラクタ k とふるまい状態 l の組合せについて，エンゲージメントの分布 ϕ_{kl} が生成される．

$$\phi_{kl} \sim \text{Beta}(\beta, \gamma)\ ,\ 1 \leq k \leq K\ ,\ 1 \leq l \leq L \tag{5}$$

ただし，Beta はベータ分布，β と γ はハイパーパラメータをそれぞれ表している．この分布は $K \times L$ 個用意される．キャラクタ z_{ij} を持つアノテータがふるまい状態 x_{jn} を観察したとき，そのアノテータはエンゲージメント y_{ijn} を知覚する．

$$y_{ijn} \sim \text{Ber}(\phi_{z_{ij}x_{jn}}) \tag{6}$$

ただし，Ber はベルヌーイ分布を表す．

モデルを学習するときには，ふるまい状態 x_{jn} とエンゲージメント y_{ijn} は観測変数であり，キャラクタ z_{ij} は潜在変数として推定する．上記の変数のデータセットが与えられたときの条件付き分布は以下で表される．

$$p(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\Theta}, \boldsymbol{\Phi} | \boldsymbol{X}) = p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Phi}) p(\boldsymbol{Z}|\boldsymbol{\Theta}) p(\boldsymbol{\Theta}) p(\boldsymbol{\Phi}) \tag{7}$$

ただし，これらの変数はそれぞれの小文字で表される変数のデータセットに対応し，$\boldsymbol{\Theta}$ と $\boldsymbol{\Phi}$ は分布のパラメータ集合である．また，ふるまい状態のデータセット $\boldsymbol{X}$ はテスト時にも与えられる．(7) 式の右辺において，$p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Phi})$ はふるまい状態とキャラクタの組合せの各々について，エンゲージメントが高いとどの程度知覚されるか，$p(\boldsymbol{Z}|\boldsymbol{\Theta})$ は各アノテータがどのキャラクタを持つのかをそれぞれ表している．このモデルはマルコフ連鎖モンテカルロ法（MCMC）によって学習データから各分布を推定することができる．キャラクタとエンゲージメントの分布のパラメータ集合 $\boldsymbol{\Theta}$ と $\boldsymbol{\Phi}$ はサンプリング結果の期待値により推定する．

推定したモデルをテストするときには，アノテータのインデクス i，ふるまい状態 x_t，学習データから推定したパラメータ集合 $\tilde{\boldsymbol{\Theta}}$ と $\tilde{\boldsymbol{\Phi}}$ が与えられる．ただし，テスト時の対話システムのターンのインデクスを t で表す．これらから未知のエンゲージメント y_{it} の予測分布を以下の式で求める．

$$p(y_{it}|x_t, i, \tilde{\boldsymbol{\Theta}}, \tilde{\boldsymbol{\Phi}}) = \sum_{k=1}^{K} \tilde{\theta}_{ik}\ \tilde{\phi}_{kx_t} \tag{8}$$

ただし，$\tilde{\theta}_{ik}$ は i 番目のアノテータがキャラクタ k を持つ確率，$\tilde{\phi}_{kx_t}$ はふるまい x_t がキャラクタ k を持つアノテータにエンゲージメントが高いと知覚される確率について，学習データから推定した分布である．エンゲージメントが高いと知覚される確率 $p(y_{it} = 1|x_t, i, \tilde{\boldsymbol{\Theta}}, \tilde{\boldsymbol{\Phi}})$ が閾値以上の場合に，t 番目のターンをエンゲージメントが高いと出力する．また，このモデルは未知のアノテータについては，エンゲージメントの推定を行うことができない．しかし，熊野らの手法 [Kumano 13, 熊野 17] のように，アノテータの性格特性などをアノテータのインデクスの代わりとすることで対応可能になると考えられる．

4·2 文脈情報の利用

直前のターンでのエンゲージメントの推定結果を利用することで，文脈情報を考慮したものへと提案モデルを拡張する．ここでの仮定は，エンゲージメントはいくつかのターンにわたって継続する場合もあることである．例えば，一度エンゲージメントが高くなると，以降のターンも継続してエンゲージメントが高くなると考えられる．前章でのモデルの定義を以下のように拡張する．(5) 式のエンゲージメントの分布と (6) 式の知覚されるエンゲージメントはそれぞれ以下のように表される．

$$\phi_{kly} \sim \text{Beta}(\beta, \gamma)\ ,\ y \in \{0,1\} \tag{9}$$

$$y_{ijn} \sim \text{Ber}\left(\phi_{z_{ij}x_{jn}y_{ij(n-1)}}\right) \tag{10}$$

ただし，(9) 式での y は，直前のターンでのエンゲージメントの推定結果に対応する．また，(7) 式の条件付き分布は以下となる．

$$\begin{aligned} &p(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\Theta}, \boldsymbol{\Phi} | \boldsymbol{Y}_{1:n-1}, \boldsymbol{X}) \\ &= p(\boldsymbol{Y}|\boldsymbol{Y}_{1:n-1}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Phi}) p(\boldsymbol{Z}|\boldsymbol{\Theta}) p(\boldsymbol{\Theta}) p(\boldsymbol{\Phi}) \end{aligned} \tag{11}$$

ただし，$\boldsymbol{Y}_{1:n-1}$ は直前のターンまでのエンゲージメントの判断結果のデータセットを表す．(8) 式のエンゲージメントの予測分布は，再帰的に以下の前向きアルゴリズム

により求める．

$$p(y_{it}|x_t,i,\tilde{\mathbf{\Theta}},\tilde{\mathbf{\Phi}}) = \sum_{k=1}^{K}\tilde{\theta}_{ik}\sum_{y_{i(t-1)}\in\{0,1\}}\tilde{\phi}_{kx_ty_{i(t-1)}}p(y_{i(t-1)}|x_{t-1},i,\tilde{\mathbf{\Theta}},\tilde{\mathbf{\Phi}}) \quad (12)$$

4·3 確率レベルのふるまい統合

4·1 節では，複数のふるまいは特徴量の時点で，組合せの状態として表現された．ここでは，各ふるまいを独立に確率モデルとして表現し，確率レベルで統合することを考える．具体的には，Product of Experts (PoE) [Hinton 02] の枠組みを用いる．各ふるまいにおけるエンゲージメントの分布を以下で表す．

$$\phi_{kbb'} \sim \text{Beta}(\beta,\gamma)\ ,\ 1\le b\le B\ ,\ b'\in\{0,1\} \quad (13)$$

ただし，b' はふるまい b の生起に対応する．そして，これらを以下の式で統合する．

$$\phi_{kl} = \frac{\prod_b^B \phi_{kbb'}}{\prod_b^B(1-\phi_{kbb'})+\prod_b^B\phi_{kbb'}} \quad (14)$$

ここで，4·1 節でのふるまい状態 l は，ふるまいの種類 b とその生起 b' に分解されていることがわかる．このモデルでは，エンゲージメントの分布の数は，ふるまいの種類数に対して線形であり，特徴量レベルの統合に比べて，ふるまいの種類を容易に増やすことができる．また，このモデルでも，4·2 節と同様にして，直前のターンでの推定結果 $y_{ij(n-1)}$ を文脈情報として利用することができる．

5. 評 価 実 験

提案モデルの有効性を確認するために，3 章で述べた 20 セッションの対話データを用いて交差検定を行った．ここでは，19 セッションを学習用，1 セッションをテスト用とした．各セッションでは 5 人のアノテータによる正解データがある．これらの正解データについて，すべて同時に評価した．ただし，提案モデルでは，与えられたアノテータのインデクス i に応じて，それぞれ異なる推定結果を出力する．評価手順は以下の通りである．まず，エンゲージメントが高いと知覚される確率 $p(y_{it}=1|x_t,i,\tilde{\mathbf{\Theta}},\tilde{\mathbf{\Phi}})$ の閾値を変化させて，以下の式により適合率と再現率の曲線をテスト用のセッション毎に求める．

$$\text{適合率} = \frac{\#\text{TP}}{\#\text{TP}+\#\text{FP}} \quad (15)$$

$$\text{再現率} = \frac{\#\text{TP}}{\#\text{TP}+\#\text{FN}} \quad (16)$$

ただし，#TP は正解ラベルが正のうちで，これを正しく推定したターン数，#FP は正解ラベルが負のうちで，正と誤って推定したターン数，#FN は正解ラベルが正のうちで，負と誤って推定したターン数をそれぞれ表す．ある 1 つのターンにおける評価の一例を表 2 に示す．このターンでは，適合率は 1/2，再現率は 1/3 である．実際には，セッション毎に，すべてのターンのデータを用いてこの評価を行う．そして，全セッションでの平均曲線を求め，その Area Under the Curve（AUC）を評価指標とした．したがって，AUC が高いほど良い推定精度であるといえる．また，実際のシステム運用では，各アノテータに対してどの程度近い推定がなされているかが重要である．このための評価指標として，アノテータ毎に F 値を求めた．F 値を算出するための閾値として，上記で求めたセッション毎の適合率と再現率の曲線において適合率と再現率が等しくなる点（Equal Error Rate）の値を用いて，アノテータ毎に適合率と再現率およびその調和平均である F 値を求めた．そして，全セッションにおける全アノテータの平均 F 値を評価した．データセット全体で，エンゲージメントが高いとアノテータにより判断されたターン数は 1,140，そうではないターン数は 2,195 であった．したがって，すべてのターンでエンゲージメントが高いと推定した場合のチャンスレートは 0.342 である．また，入力特徴量は，3 章で述べたように，4 種類のふるまい，相槌，笑い，うなずき，視線の生起について，人手によりアノテーションしたものである．

表 2 評価の例（このターンでは，適合率=1/2，再現率=1/3）

アノテータ (i)	1	3	4	7	10
正解ラベル	1	1	0	0	1
モデル出力	0.45	0.61	0.39	0.55	0.41
結果 (閾値=0.50)	FN	TP	TN	FP	FN

提案モデルのパラメータは以下の通りである．キャラクタの状態数（K）は 2 から 4 を試した．MCMC の実装には，PyMC 2.36[*1]を用いた．MCMC のサンプリングアルゴリズムにはメトロポリス・ヘイスティングス法を用いた．サンプリング回数は 30,000 回で，うち 20,000 回はバーンインとした．事前分布はすべて一様分布とした．

5·1 キャラクタの効果

キャラクタの違いを考慮することによる効果を比較した．ただし，4·2 節で述べた文脈情報はここでは用いない．これについては次節で比較を行う．比較手法は，ロジスティック回帰モデルと，潜在キャラクタモデルと同じ枠組みでキャラクタ数が 1 の場合（$K=1$）である．ロジスティック回帰モデルは scikit-learn 0.18.1[*2]を用いて実装した．また，ロジスティック回帰モデルを学習する際の教師ラベルの使用方法は，アノテータ非依存とアノテータ依存の 2 種類を検討した．前者のアノテータ非依存では，すべてのアノテータのデータをそのまま用いて，1 つの

∗1 https://github.com/pymc-devs/pymc
∗2 http://scikit-learn.org/stable/

表 3 キャラクタの効果 (セッション毎の Area Under the Curve およびアノテータ毎の F 値の平均)

	手法	K	AUC	F 値
ロジスティック回帰	アノテータ非依存		0.569	0.496
	アノテータ非依存（多数決）		0.567	0.491
	アノテータ依存		0.597	0.440
	アノテータ依存（投票）		0.550	0.484
潜在キャラクタモデル	特徴量レベル	1	0.568	0.497
		2	0.640	0.535
		3	**0.645**	**0.547**
		4	0.643	0.545
	確率レベル	1	0.569	0.494
		2	0.642	0.529
		3	0.642	0.520
		4	0.643	0.518

モデルを学習する．テスト時には，複数のアノテータに対して同一のモデルを用いるため，アノテータのインデクスによらず同じ推定結果を出力する．さらに，多数決によるラベリング手法と比較するために，アノテータ非依存では，アノテータ間での多数決ラベルによるモデルの学習も試みる．具体的には，５人のアノテータのうち３人以上がエンゲージメントが高いと判断したものには正，２人以下の場合には負のラベルを用意して，これを教師ラベルとしてモデルを学習する．後者のアノテータ依存では，各アノテータによる学習データのみを用いて，アノテータ毎にモデルを学習する．テスト時には，各アノテータに対応するモデルを用いて評価する．この方法では，アノテータ毎の傾向を学習できるが，学習データ数は少なくなる．また，集合知による手法と比較するために，アノテータ依存では，他のアノテータのモデルによる推定結果の統合も試みる．具体的には，推定対象であるアノテータ以外の４人のアノテータのアノテータ依存モデルを用いて推定を行い，エンゲージメントが高いと判断する確率の平均をとったものを出力とした．ロジスティック回帰モデルでは，学習データを９対１に相当するようにセッション単位で分割して，前者をモデルの学習，後者を L2 ノルム正則化項の重み（ハイパーパラメータ）のチューニングに用いた．この重みの値の範囲は，$\{10^n | n = -3, -2, -1, 0, 1, 2, 3\}$ であり，AUC が最も高くなる値を選択した．

結果を表 3 に示す．まず，ロジスティック回帰モデルにおいて，セッション毎の AUC ではアノテータ依存が，アノテータ毎の F 値ではアノテータ非依存が，それぞれ高い精度を示した．ただし，アノテータ依存ではデータ数が少ないことから，モデルが十分に学習されていないと推察される．実際に，出力された事後確率はアノテータ毎にその範囲が大きく異なっていたため，閾値を変化させると，特定のアノテータにはすべて正，別のアノテータにはすべて負と推定するセッションがしばしばみられた．このような推定であっても，今回のタスクでは AUC

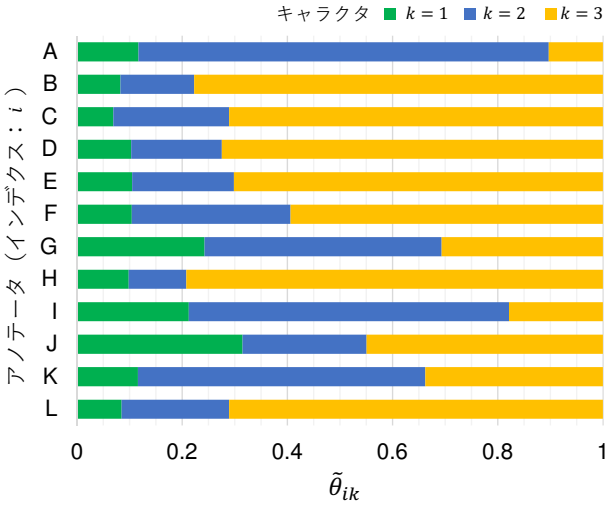

図 3 各アノテータのキャラクタの分布

が多少の改善を示すため，AUC および F 値の両方を改善することが望ましいといえる．また，アノテータ非依存（多数決）とアノテータ依存（投票）では，大きな精度改善は特にみられなかった．次に，潜在キャラクタモデル（$K \geq 2$）に注目すると，いずれの評価指標においても，比較手法に比べて推定精度が向上していることがわかる．したがって，アノテータ間でのキャラクタの違いを考慮することは有効であるといえる．潜在キャラクタモデルにおけるふるまいの統合手法を比較すると，若干ではあるが特徴量レベルの方が確率レベルよりも高い精度を示した．このことは，特徴量の段階で各ふるまいの共起を考慮する必要があることを示唆している．キャラクタの状態数 K を変化させたときの推定精度を比較すると，有意な差はみられなかった．したがって，ここでのキャラクタは潜在的に２または３程度の次元で表現できると推察される．

潜在キャラクタモデルの学習例について分析する．ここでの例は，キャラクタ数が３で，特徴量レベルの統合

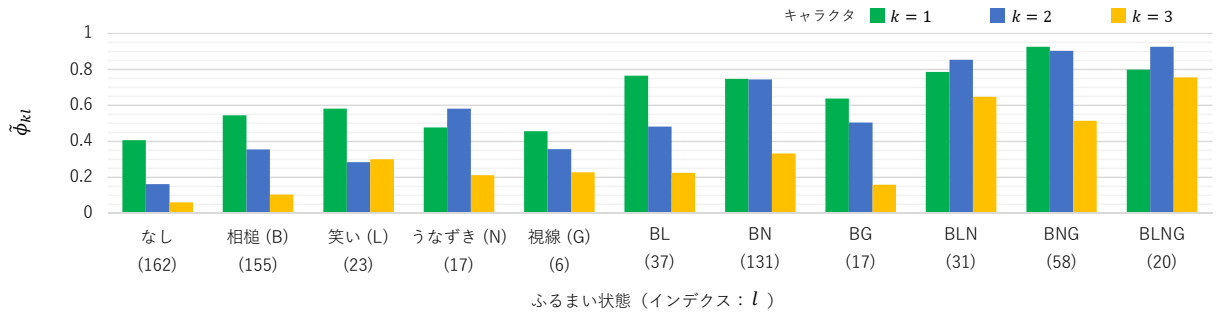

図 4 各ふるまい状態に対してエンゲージメントが高いと各キャラクタが知覚する確率（ラベル下の括弧内はコーパス内での頻度）

表 4 文脈情報の効果 (セッション毎の Area Under the Curve およびアノテータ毎の F 値の平均)

	手法	K	文脈情報	AUC	F 値
ロジスティック回帰	アノテータ非依存			0.569	0.496
	アノテータ非依存（多数決）			0.567	0.491
	アノテータ依存			0.597	0.440
	アノテータ依存（投票）			0.550	0.484
条件付確率場（CRF）	アノテータ非依存		✓	0.599	0.554
	アノテータ非依存（多数決）		✓	0.586	0.533
	アノテータ依存		✓	0.609	0.447
	アノテータ依存（投票）		✓	0.566	0.505
潜在キャラクタモデル	特徴量レベル	1		0.568	0.497
			✓	0.619	0.552
		3		0.645	0.547
			✓	**0.667**	**0.565**
	確率レベル	1		0.569	0.494
			✓	0.615	0.544
		3		0.642	0.520
			✓	0.657	0.542

の場合である．図 3 にアノテータ毎のキャラクタの分布（$\tilde{\theta}_{ik}$）を示す．いくつかアノテータ間で共通した傾向がみられる．例えば，アノテータ A と I は似た分布である．また，アノテータ B，C，D，E，H，L の間でも似た分布を示している．図 4 に，各キャラクタにおいて，各ふるまい状態が入力されたときにエンゲージメントが高いと知覚する確率（$\tilde{\phi}_{kl}$）を示す．ただし，コーパス全体での頻度が 5 回以下のふるまい状態は除外した．この例より，キャラクタ毎に異なる傾向が学習できていることがわかる．例えば，1 つ目のキャラクタ（$k=1$）は，相槌（B）と笑い（L）が，エンゲージメントが高いと知覚する傾向にあることがわかる．2 つ目のキャラクタ（$k=2$）ではうなずき（N）が，3 つ目のキャラクタ（$k=3$）では笑い（L）が比較的優位であることがわかる．また，3 つ目のキャラクタ（$k=3$）は，エンゲージメントが高いと知覚する確率が他の 2 つのキャラクタに比べて全体的に低い．すべてのキャラクタに共通して，複数のふるまいが共起した場合（図の右側）には，エンゲージメントが高いと知覚する傾向にある．

5·2 文脈情報の効果

直前のターンでの推定結果を文脈情報として用いる効果を比較した．比較手法は，条件付き確率場（CRF: conditional random field）と，文脈情報ありの潜在キャラクタモデルと同じ枠組みでキャラクタ数が 1 の場合（$K=1$）である．CRF は，CRF suite 0.12[*3] を用いて実装した．CRF を学習する際の教師ラベルの使用方法は，前節と同様に，アノテータ非依存とアノテータ依存の 2 種類を検討した．CRF の L2 ノルム正則化項の重み（ハイパーパラメータ）は，前節と同様にして，学習データセットの 1 割に相当するセッションを用いてチューニングを行った．前節の結果に基づいて，潜在キャラクタモデルにおけるキャラクタの状態数は 3 に固定した（$K=3$）．結果を表 4 に示す．文脈情報を利用することにより，セッション毎の AUC およびアノテータ毎の F 値は全体的に向上した．キャラクタを考慮することによる効果は特徴量レベルの統合でみられた．また，ふるまいの統合方法の比較についても前節と同様の傾向がみられ，特徴量レベルの方が若干の高い精度を示した．

*3 http://www.chokkan.org/software/crfsuite/

表 5 各アノテータが各キャラクタを持つ確率と Big Five 尺度の各因子とのスピアマンの順位相関係数（$^-$ は $p < 0.10$，*は $p < 0.05$, **は $p < 0.01$ をそれぞれ表す.）

Big Five 因子	キャラクタ		
	$k=1$	$k=2$	$k=3$
外向性	0.723**	0.295	−0.368
情緒不安定性	0.196	0.298	−0.263
開放性	0.172	−0.305	0.256
誠実性	−0.487	−0.557$^-$	0.655*
調和性	−0.095	−0.102	0.193

5·3 主観的なキャラクタとの関係

潜在キャラクタモデルにより推定したキャラクタの分布と主観的なキャラクタとの関係について調べた．主観的なキャラクタを得るために，和田 [和田 96] によって作成された Big Five 尺度を用いた．これは，60 項目の 7 段階評定から，性格に関する 5 因子，外向性，情緒不安定性，開放性，誠実性，調和性についての尺度得点を算出するものである．3 章のアノテーション作業を実施するにあたり，各アノテータに対してこの尺度を測定した．図 3 で示した各アノテータが各キャラクタを持つ確率と，Big Five 尺度の各因子との組合せについて，スピアマンの順位相関係数を求めた．結果を表 5 に示す．1 つ目のキャラクタ（$k=1$）では，外向性と有意な正の相関がみられた（$p=0.008$）．藤本らの研究 [藤本 07] では，言語および非言語を通して相手の考えや気持ちを正しく読み取る「解読力」について，外向性が高い人は低い人よりも解読力が高いことが示されている．1 つ目のキャラクタ（$k=1$）は，図 4 において，11 個中 7 個のふるまい状態について，他の 2 つのキャラクタよりもエンゲージメントが高いと知覚する確率が高かった．したがって，このキャラクタは，ふるまいの解読力が高いことが示唆される．また，誠実性について比較的高い負の相関係数が示されたが，有意であるとはいえなかった（$p=0.108$）．2 つ目のキャラクタ（$k=2$）でも，誠実性について比較的高い負の相関係数が示されたが，こちらも有意であるとはいえなかった（$p=0.060$）．3 つ目のキャラクタ（$k=3$）では，誠実性と有意な正の相関がみられた（$p=0.021$）．以上より，潜在キャラクタモデルによって推定されるキャラクタの分布が，Big Five 尺度による主観的なキャラクタと一部相関することがわかった．ただし，より多くのアノテータのデータを用いて有意性を検証する必要があるといえる．

6. お わ り に

本論文では，聞き手の多様なふるまいに基づく対話エンゲージメントの推定について述べた．ここでは，エンゲージメントを知覚する側の内部にキャラクタがあり，これが知覚の方法に影響を及ぼすと仮定した．そして，キャラクタを潜在変数とする潜在キャラクタモデルを提案した．提案モデルは，キャラクタとエンゲージメントの分布を同時に推定することができる．実験の結果，キャラクタの違いを考慮することで，各アノテータに適したエンゲージメントの推定ができることを確認した．さらに，文脈情報を利用することによる精度向上も示した．また，実際に学習された分布の分析を行い，キャラクタ毎にエンゲージメントの推定傾向が異なることも確認した．この分布は，自律型アンドロイドのような身体性を伴う対話ロボットが自身のエンゲージメントを表現する際に有用であると期待できる．

今後は，ふるまいの自動検出 [Chen 15, Gosztolya 15, Inoue 15, Kaushik 15, Morency 07] を実装して，エンゲージメントの推定モデルとの統合を図る．これにより，リアルタイムでエンゲージメントを推定する対話ロボットの実現を目指す．さらに，関連研究で述べたように，エンゲージメントを推定した後の対話システムの行動およびふるまいについても研究を展開していく予定である．

謝 辞

対話収録，およびアノテーション作業にご協力いただいた皆様に感謝いたします．本研究は，JSPS 科研費（課題番号 : 15J07337），ならびに JST ERATO 石黒共生ヒューマンロボットインタラクションプロジェクト（課題番号 : JPMJER1401）の支援を受けて実施されたものである．

◇ 参 考 文 献 ◇

[Barrick 91] Barrick, M. R. and Mount, M. K.: The big five personality dimensions and job performance: A meta-analysis, *Personnel Psychology*, Vol. 44, No. 1, pp. 1–26 (1991)

[Blei 03] Blei, D. M., Ng, A. Y., and Jordan, M. I.: Latent dirichlet allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022 (2003)

[Bohus 10] Bohus, D. and Horvitz, E.: Facilitating multiparty dialog with gaze, gesture, and speech, in *Proceedings of the International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI)*, No. 5 (2010)

[Breazeal 04] Breazeal, C.: Social interactions in HRI: The robot view, *IEEE Transactions on Man, Cybernetics, and Systems*, Vol. 34, No. 2, pp. 181–186 (2004)

[Cerrato 16] Cerrato, L. and Campbell, N.: Engagement in dialogue with social robots, in *Proceedings of the International Workshop on Spoken Dialogue Systems (IWSDS)* (2016)

[Chen 15] Chen, Y., Yu, Y., and Odobez, J.-M.: Head nod detection from a full 3D model, in *Proceedings of the International Conference on Computer Vision Workshops (ICCVW)*, pp. 136–144 (2015)

[千葉 16] 千葉 祐弥, 伊藤 彰則 : WOZ システムとの対話におけるユーザの対話意欲の段階識別と特徴量の分析, 人工知能学会研究会資料 言語・音声理解と対話処理研究会 (SLUD), SIG-SLUD-B505-02, pp. 7–12 (2016)

[Den 11] Den, Y., Yoshida, N., Takanashi, K., and Koiso, H.: Annotation of Japanese response tokens and preliminary analysis on their distribution in three-party conversations, in *Proceedings of Oriental COCOSDA*, pp. 168–173 (2011)

[DeVault 14] DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., Lucas, G., Marsella, S., Morbini, F., Nazarian, A., Scherer, S., Stratou, G., Suri, A., Traum, D., Wood, R., Xu, Y., Rizzo, A., and

Morency, L. P.: SimSensei Kiosk: A virtual human interviewer for healthcare decision support, in *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 1061–1068 (2014)

[藤本 07] 藤本 学, 大坊 郁夫 : コミュニケーション・スキルに関する諸因子の階層構造への統合の試み, パーソナリティ研究, Vol. 15, No. 3, pp. 347–361 (2007)

[Glas 15] Glas, N. and Pelachaud, C.: Definitions of engagement in human-agent interaction, in *Proceedings of the International Workshop on Engagement in Human Computer Interaction (ENHANCE)*, pp. 944–949 (2015)

[Glas 16] Glas, D. F., Minaot, T., Ishi, C. T., Kawahara, T., and Ishiguro, H.: ERICA: The ERATO Intelligent Conversational Android, in *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (2016)

[Goffman 66] Goffman, E.: *Behavior in Public Places: Notes on the Social Organization of Gatherings*, Simon & Schuster (1966)

[Gosztolya 15] Gosztolya, G.: On evaluation metrics for social signal detection, in *Proceedings of Interspeech*, pp. 2504–2508 (2015)

[Higashinaka 14] Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., Hirano, T., Makino, T., and Matsuo, Y.: Towards an open-domain conversational system fully based on natural language processing, in *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp. 928–939 (2014)

[Hinton 02] Hinton, G. E.: Training products of experts by minimizing contrastive divergence, *Neural Computation*, Vol. 14, No. 8, pp. 1771–1800 (2002)

[Huang 16] Huang, Y., Gilmartin, E., and Campbell, N.: Engagement recognition using auditory and visual cues, in *Proceedings of Interspeech* (2016)

[Inoue 15] Inoue, K., Wakabayashi, Y., Yoshimoto, H., Takanashi, K., and Kawahara, T.: Enhanced speaker diarization with detection of backchannels using eye-gaze information in poster conversations, in *Proceedings of Interspeech*, pp. 3086–3090 (2015)

[Inoue 16a] Inoue, K., Lala, D., Nakamura, S., Takanashi, K., and Kawahara, T.: Annotation and analysis of listener's engagement based on multi-modal behaviors, in *Proceedings of the International Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction (MA3HMI)* (2016)

[Inoue 16b] Inoue, K., Milhorat, P., Lala, D., Zhao, T., and Kawahara, T.: Talking with ERICA, an autonomous android, in *Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, pp. 212–215 (2016)

[Ishi 12] Ishi, C. T., Ishiguro, H., and Hagita, N.: Evaluation of formant-based lip motion generation in tele-operated humanoid robots, in *Proceeding of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2377–2382 (2012)

[石井 11] 石井 亮, 大古 亮太, 中野 有紀子, 西田 豊明 : 視線と頭部動作に基づくユーザの会話参加態度の推定, 情報処理学会論文誌, Vol. 52, No. 12, pp. 3625–3636 (2011)

[Kaushik 15] Kaushik, L., Sangwan, A., and Hansen, J. H. L.: Laughter and filler detection in naturalistic audio, in *Proceedings of Interspeech*, pp. 2509–2513 (2015)

[河原 13] 河原 達也 : 音声対話システムの進化と淘汰 －歴史と最近の技術動向－, 人工知能学会誌, Vol. 28, No. 1, pp. 45–51 (2013)

[Kumano 13] Kumano, S., Otsuka, K., Matsuda, M., Ishii, R., and Yamato, J.: Using a probabilistic topic model to link observers' perception tendency to personality, in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 588–593 (2013)

[熊野 17] 熊野 史朗, 石井 亮, 大塚 和弘 : 評定者個人に特化した他者感情理解モデル, 2017 年度人工知能学会全国大会 (第 31 回), 2H4-OS-35b-3in2 (2017)

[Kuno 07] Kuno, Y., Sadazuka, K., Kawashima, M., Yamazaki, K., Yamazaki, A., and Kuzuoka, H.: Museum guide robot based on sociological interaction analysis, in *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pp. 1191–1194 (2007)

[Langton 00] Langton, S. R. H., Watt, R., and Bruce, V.: Do the eyes have it? Cues to the direction of social attention, *Trends in Cognitive Sciences*, Vol. 4, No. 2, pp. 50–59 (2000)

[Michalowski 06] Michalowski, M. P., Sabanovic, S., and Simmons, R.: A spatial model of engagement for a social robot, in *Proceedings of the International Workshop on Advanced Motion Control (AMC)*, pp. 762–767 (2006)

[Morency 06] Morency, L. P., Christoudias, C. M., and Darrell, T.: Recognizing gaze aversion gestures in embodied conversational discourse, in *Proceedings of the ACM International Conference on Multimodal Interfaces (ICMI)*, pp. 287–294 (2006)

[Morency 07] Morency, L. P., Quattoni, A., and Darrell, T.: Latent-dynamic discriminative models for continuous gesture recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2007)

[Nakano 10] Nakano, Y. I. and Ishii, R.: Estimating user's engagement from eye-gaze behaviors in human-agent conversations, in *Proceedings of the ACM Conference on Intelligent User Interfaces (IUI)*, pp. 139–148 (2010)

[Ozkan 10] Ozkan, D., Sagae, K., and Morency, L. P.: Latent mixture of discriminative experts for multimodal prediction modeling, in *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp. 860–868 (2010)

[Ozkan 11] Ozkan, D. and Morency, L. P.: Modeling wisdom of crowds using latent mixture of discriminative experts, in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 335–340 (2011)

[Peters 05] Peters, C.: Direction of attention perception for conversation initiation in virtual environments, in *Proceedings of the International Workshop on Intelligent Virtual Agents (IVA)*, pp. 215–228 (2005)

[Poggi 07] Poggi, I.: *Mind, Hands, Face and Body: A Goal and Belief View of Multimodal Communication*, Weidler (2007)

[境 16] 境 くりま, 石井 カルロス寿憲, 港 隆史, 石黒 浩 : 音声に対応する頭部動作のオンライン生成システムと遠隔操作における効果, 電子情報通信学会論文誌, Vol. J99-A, No. 1, pp. 14–24 (2016)

[Sidner 02] Sidner, C. L. and Dzikovska, M.: Human-robot interaction: Engagement between humans and robots for hosting activities, in *Proceedings of the ACM International Conference on Multimodal Interfaces (ICMI)*, p. 123 (2002)

[Sidner 05] Sidner, C. L., Lee, C., Kidd, C. D., Lesh, N., and Rich, C.: Explorations in engagement for humans and robots, *Artificial Intelligence*, Vol. 166, No. 1-2, pp. 140–164 (2005)

[Skantze 15] Skantze, G. and Johansson, M.: Modelling situated human-robot interaction using IrisTK, in *Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, pp. 165–167 (2015)

[高梨 09] 高梨 克也, 榎本 美香 :「特集–聞き手行動から見たコミュニケーション」編集にあたって, 認知科学, Vol. 16, No. 1, pp. 5–11 (2009)

[和田 96] 和田 さゆり : 性格特性用語を用いた Big Five 尺度の作成, 心理学研究, Vol. 67, No. 1, pp. 61–67 (1996)

[Wilcock 15] Wilcock, G. and Jokinen, K.: Multilingual WikiTalk: Wikipedia-based talking robots that switch languages, in *Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, pp. 162–164 (2015)

[Xu 13] Xu, Q., Li, L., and Wang, G.: Designing engagement-aware agents for multiparty conversations, in *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pp. 2233–2242 (2013)

[Yu 04] Yu, C., Aoki, P. M., and Woodruff, A.: Detecting user engagement in everyday conversations, in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pp. 1329–1332 (2004)

[Yu 16] Yu, Z., Nicolich-Henkin, L., Black, A. W., and Rudnicky, A. I.: A Wizard-of-Oz study on a non-task-oriented dialog systems that reacts to user engagement, in *Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, pp. 55–63 (2016)

〔担当委員：石井 亮〕

2017 年 9 月 1 日　受理

著　者　紹　介

井上　昂治（学生会員）

2015 年 京都大学大学院情報学研究科修士課程修了．現在，同大学院博士後期課程在学中，および日本学術振興会特別研究員 (DC1)．音声言語処理，マルチモーダルインタラクションに関する研究に従事．情報処理学会，日本音響学会，電子情報通信学会，IEEE，ISCA 各会員．

Divesh Lala

2015 年 京都大学大学院情報学研究科博士後期課程修了．博士 (情報学)．同年，日本学術振興会外国人特別研究員．現在，京都大学大学院情報学研究科研究員．マルチモーダルインタラクションに関する研究に従事．

吉井　和佳

2008 年 京都大学大学院情報学研究科博士後期課程修了．博士 (情報学)．同年，産業技術総合研究所情報技術研究部門に入所．2014 年 京都大学大学院情報学研究科講師に着任．音楽情報処理，統計的音響信号処理の研究に従事．

高梨　克也（正会員）

2000 年 京都大学大学院人間・環境学研究科博士課程単位取得退学．博士 (情報学)．独立行政法人情報通信研究機構専攻研究員，京都大学学術情報メディアセンター特定助教，科学技術振興機構さきがけ専従研究者などを経て，現在，京都大学大学院情報学研究科研究員．コミュニケーションの組織化を支える認知的・社会的プロセスの解明に従事．言語処理学会，日本認知科学会，社会言語科学会，組織学会 各会員．一般社団法人社会対話技術研究所理事．

河原　達也（正会員）

1987 年 京都大学工学部情報工学科卒業．1989 年 同大学院修士課程修了．1990 年 京都大学工学部助手．1995 年 同助教授．2003 年 同大学学術情報メディアセンター／情報学研究科教授．現在に至る．この間，1995〜96 年 米国・ベル研究所客員研究員．1998〜2006 年 ＡＴＲ客員研究員．2006 年〜 情報通信研究機構短時間研究員・招へい専門員．音声情報処理，特に音声認識および対話システムに関する研究に従事．博士 (工学)．科学技術分野の文部科学大臣表彰 (2012 年度)，日本音響学会から粟屋潔学術奨励賞 (1997 年度)，情報処理学会から坂井記念特別賞 (2000 年度)，喜安記念業績賞 (2011 年度)，論文賞 (2012 年度) を受賞．IEEE ASRU 2007 General Chair，INTERSPEECH 2010 Tutorial Chair，IEEE ICASSP 2012 Local Arrangement Chair，言語処理学会理事，情報処理学会音声言語情報処理研究会主査，情報処理学会理事，APSIPA 理事，ISCA 理事を歴任．IEEE Fellow. 情報処理学会，日本音響学会，電子情報通信学会，言語処理学会，ISCA，APSIPA 各会員．日本学術会議連携会員．