



# 気の利く対話 AI のための「空気を読む」技術 ——マルチモーダル情報を用いた対話の場・関係の 理解とインクリメンタル応答生成

対話システムが人間の日常生活に溶け込み、「気の利いた」会話ができるようになるには、会話の場面や対話者どうしの関係を認識し、適切な間合いで応答を行うといった「空気を読む」能力が求められます。私たちはこれまで、「空気を読む」対話システムの実現をめざし、さまざまな対話処理技術の研究に取り組んできました。本稿ではそのうち、日常会話のシーンを理解する対話状況認識、対話者間の心理的距離感を理解する親しみ認識、人間の対話のテンポに追従するためのインクリメンタル応答生成の3つの研究をピックアップして紹介します。

キーワード：#対話状況推定、#心理的距離感の推定、#インクリメンタル応答生成

ちば ゆうや  
千葉 祐弥

NTTコミュニケーション科学基礎研究所

## 「空気を読む」対話処理技術

大規模言語モデル (LLM: Large Language Model) の進展により、対話システムの応答の自然さは飛躍的に向上しました。ChatGPTをはじめとする多くの AI (人工知能) アプリケーションでは、会話による操作がすでに一般的になっています。しかし、対話システムが私たちの日常生活に参画し、「気の利いた」振る舞いをするためには、システムにも「空気を読む」ことが求められます。例えば、友人どうしの雑談で会話を盛り上げるように話題を振ったり、あるいは議論が硬直したときに適切なアドバイスをしたりするには、対話参加者間の関係や対話の目的などの対話状況を的確に把握する必要があります。話者間の対立を仲裁するなど、ときには人々の心理的距離感までも考慮した振る舞いが求められる場合もあるでしょう。加えて、

対話のテンポに追従して適切なタイミングで応答できることも重要です。私たちはこれまで、人間と共生するパートナーとしての対話システムの実現をめざし、さまざまな「空気を読む」対話処理技術の研究に取り組んできました。本稿では、このうち3つの研究について紹介します。

## 対話状況の推定

対話システムが私たちの日常会話に自然に加わるには、状況に応じた振る舞いが求められます。例えば、喫茶店での友人どうしの会話と会社での同僚との議論の場面では、会話に参加するシステムに求められる対話は異なります。前者のような場面ではカジュアルな話し方で会話を盛り上げてほしいですし、後者のような場面では議論を邪魔しないように話しかけるタイミングを計ったり、フォーマルな話し方をしたりする必

要があります。本研究では、こうした状況に応じた対話制御の実現をめざしました。

最初に、状況ごとに異なる会話の言語的特徴を、因子分析によって調査しました<sup>(1)</sup>。図1に因子分析の手続きの概要を示します。まず、それぞれの発話から、意図や話者の心的態度、文法的要素などを抽出します。それらに対話全体で集積し、特徴量ベクトルとして扱います。このようにして抽出した特徴量ベクトルに対して因子分析を行うことで、会話の特徴を説明する軸、すなわち因子を得ることができます。国立国語研究所が提供する日本語日常会話コーパス<sup>(2)</sup>を用いた分析では、7つの因子を抽出しました。言語学や社会学などで蓄積されてきた会話分析の知見と照らし合わせ、それぞれ「説明」「依頼」「語り」「丁寧」「感情」「関与」「提案」と名付けました。これらの因子は対話の仕方や目的に関する要素であると解釈できます。各状況の対話に対する因



図1 因子分析の手続き

子の重みを比較した結果、例えば会議における会話では説明的な成分、同僚どうしの会話では丁寧さの成分、友人・知人との会話では感情を表す発話の成分が大きかった。この知見は、日常生活に参画する対話システムの行動方針を設計するための指針となると期待できます。

続いて、対話中の周囲の情報から実際に対話状況を推定する技術に取り組みました。提案手法では、対話場面を収録した動画画像情報、音響情報、会話内容を用いて、4つの対話状況、すなわち対話の「形式」、「場所」、「会話に伴う活動」、および「対話参加者間の関係」を推定します(図2)。推定モデルの基盤には、当時の自然言語処理分野で分類タスクに広く用いられていたBERT (Bidirectional Encoder Representations from Transformers)を採用しました。図3に手法の概要を示し

ます。動画画像情報と音響情報は、それぞれ事前学習済みのモデルを用いて埋め込みベクトルに変換し、言語情報と組み合わせてモデルに入力します。ここで、4つの状況の間には一定の相関関係が存在することが分かっています。例えば、家族との会話は自宅で起こることが多く、社会的な関係の相手との会話は会社や学校で行われやすいといった傾向がみられます。そのため、各状況を個別の推定タスクとして扱い、タスク間の関係性を考慮するマルチタスク学習を導入しました。実験の結果、動画像・音響・言語のすべての情報を統合して用いることで性能が向上し、さらにマルチタスク学習を併用することで、より高い性能が得られることが示されました<sup>(3)</sup>。

### 話者の心理的距離感の推定

人間の振る舞いは、相手に対して抱く親

しみによっても変化します。例えば、話者の言語的振る舞いが相手との心理的距離や上下関係に影響されるとするポライトネス理論<sup>(4)</sup>や、親密さに応じて自己開示の深さが変化するとする社会浸透理論<sup>(5)</sup>などがよく知られています。また、非言語的な振る舞いに関しても、親しみを感ずる相手に対しては、しぐさや態度が相手に近似する「同調」などの現象がみられるといわれています<sup>(6)</sup>。このような背景から、システムが話者の親しみの度合いを理解できることは、人間と社会的関係を築くうえで重要な要素といえます。そこで本研究では、話者の言語的・非言語的振る舞いから対話相手に対する心理的距離感、すなわち親しみを推定する手法を研究しました。

まず、大学生・大学院生どうしの雑談を収録したSMOC (Spontaneous Multimodal One-on-one Chat-talk) コーパスを利用し、話者間の親しみを表現する言語的・非言語的振る舞いを分析しました。言語的な振る舞いとしては発話の意図と発話中の心理的・社会的要素を抽出し、非言語的振る舞いとしては発話の韻律的情報(話速、基本周波数、応答タイミング)や表情、視線、姿勢を抽出しました。言語的振る舞いの分析では、親しみが低い話者間の会話ではネガティブな感情を伝達する発話が多いこと、親しみが低い話者間の会話では相手の欲求や習慣を聞く質問の頻度が高いことが示されました。このことから、対話者は仲の良い間柄ではより率直に自身の意見を表明する傾向がある一方で、関係の初期段階では相手のことをよく知る会話が多く行われる傾向があることが示唆されます。非言語的

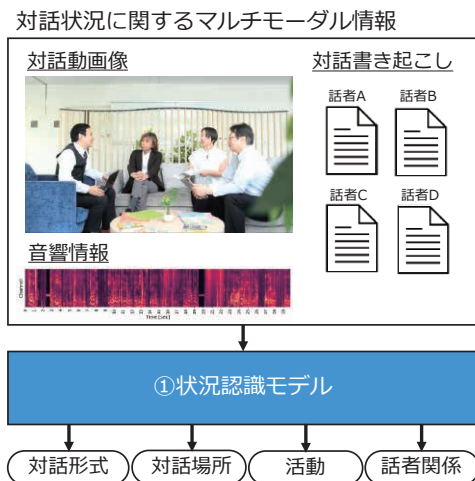


図2 対話状況の推定

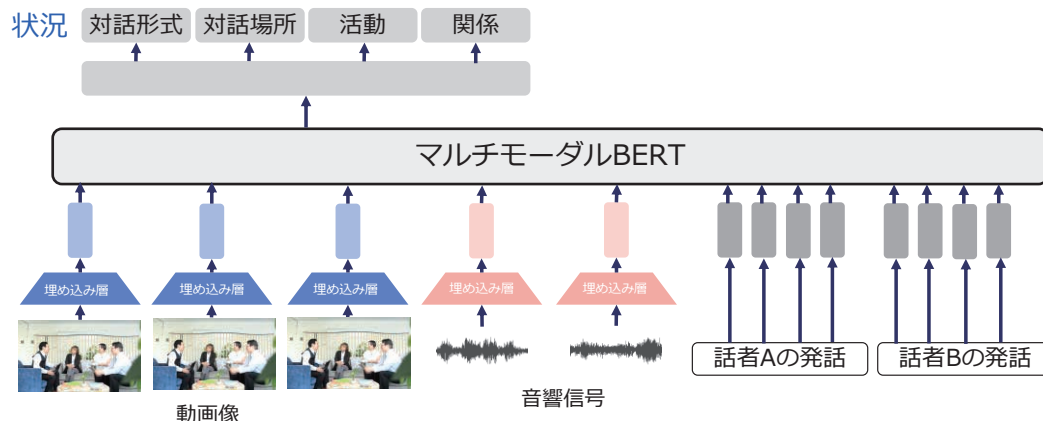


図3 対話状況推定モデル

振る舞いの分析では、表情や姿勢、一部のジェスチャが親しみの高い話者間でより同調していることが示されました。これらの結果から、抽出された言語・非言語的振る舞いは対話相手に対する親しみの推定に有効であることが示唆されます。

そこで本研究では、これらの特徴量を用いて話者の親しみの度合いを推定するモデルを構築しました<sup>7)</sup>(図4)。図5に提案モデルの概要を示します。まず、発話から言語情報、韻律情報、表情・姿勢情報を抽出し、MS-BLSTM (Multi-stream Bi-directional Long Short-Term Memory) を用いて統合します。これを、文脈を扱うBLSTMに入力し親しみを3段階で推定します。実験では、複数のベースラインモデルと提案モデルを比較し、提案

手法がこれらのベースライン手法よりも性能が高いことを示しました。特に、テキスト情報のみを用いて親しみを推定する先行研究のモデルよりも高い性能を得ることができ、マルチモーダル情報の利用が話者間の心理的距離感の推定に有効であることを示しました。

### 人間の会話の間合いに追従するインクリメンタル応答生成

人間の会話では、会話の内容だけでなくターン交替\*も重要です。人間は相手の発話を聞きながら自身が次に話す内容や、話し出すタイミングを予測しながら会話をしているといわれており、これによって非常に短い時間で話者の交替が行われます。こ

れに対して、対話システムの応答生成の仕方は大きく異なっています。一般的なパイプライン型の音声対話システムでは音声認識、応答生成、音声合成の各モジュールが同期的に駆動し、前のモジュールの処理が終わったら次のモジュールの処理というかたちで順次実行されます。そのため、人間と比べると話者交替にかかる時間が長くなってしまいます。本研究では、人間の会話の仕方を参考にした応答生成を実施することで、システムのターン交替の時間を短くすることをめざしました。

本研究で構築したインクリメンタル応答生成の概要を図6に示します。ユーザの発話はストリーミング音声認識により逐次認識され、システムは部分的な音声認識結果が得られるたびに応答生成を実施します。これにより、ユーザ発話が終わる前に複数の応答候補を保持することができます。その後、ターン交替のタイミングで、応答候補の中からもっとも適切な応答を選択し、システム発話としてユーザに提示します。この応答生成の手法は、名古屋大学と共同して開発した対話システム構築ツールキット<sup>9)</sup>に導入され、多くの人々が利用できるかたちで公開されています。ターン交替の予測方法にはいくつかの方法がありますが、ここでは最先端の予測手法であるVAP

\* ターン交替：それぞれの話者は発話する番(ターン)を持ち、別の話者とターンを授受することで会話が成立します。このプロセスをターン交替や順番交替、ターンテイクングなどと呼びます。

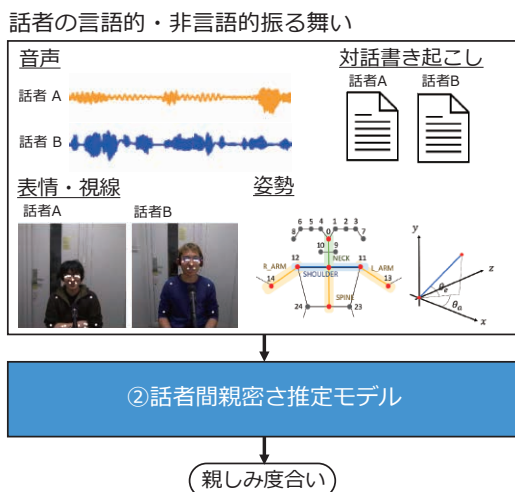


図4 心的距離感（親しみ）の推定

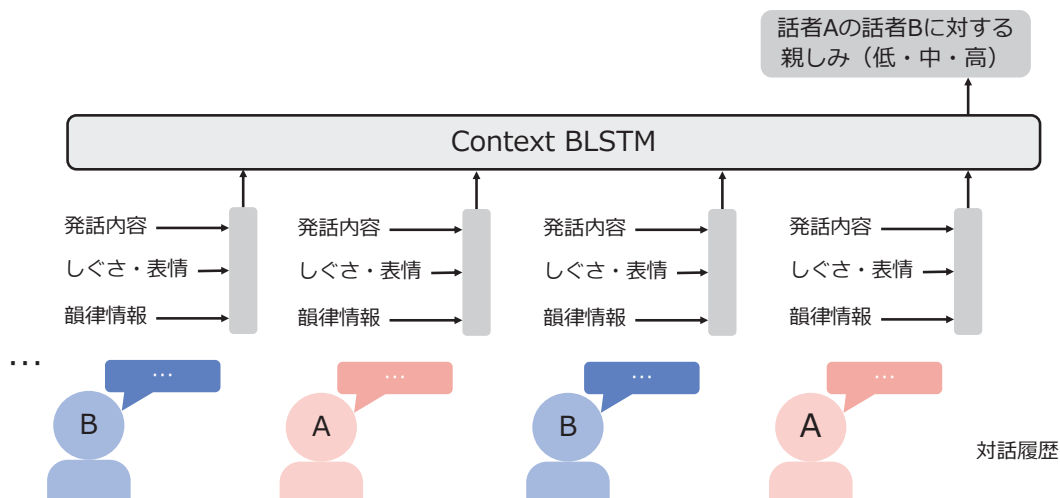


図5 心的距離感推定モデル

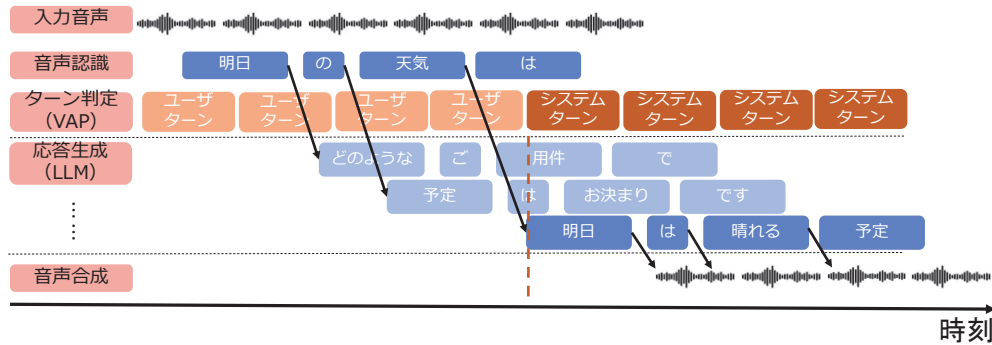


図6 インクリメンタル応答生成

(Voice Activity Projection) を採用しました。

このシステムを用いて、インクリメンタル応答生成がユーザ評価やタスク達成率など対話に与える影響を調査しました<sup>(9)</sup>。評価指標として、満足度、一貫性、自然性といった対話システムの評価に一般的に用いられる評価尺度と、音声インタフェースの評価尺度<sup>(10)</sup>を採用しました。タスク指向の対話においては、これに加えてタスク達成率も比較しました。結果から、インクリメンタル応答生成は、ターン交替までの時間自体は短縮できるものの、ユーザ評価が従来手法よりも低下してしまうことが示されました。1つの大きな理由は、ターン交替の時間が短い場合は必然的に応答生成に使えるユーザ発話が短くなるため、破綻した応答が出やすくなってしまっていることにありました。これは、満足度の高い会話を行うためには単に高度なターン交替手法を用いるだけでは不十分で、それに合わせた正確な応答生成を行うことが必要であることを示唆しています。人間は対話相手の特性や過去の対話のやり取りを考慮したり、まだ話されていない相手の発話を予測したりすることで破綻のない会話ができていると考えられ、インクリメンタル応答生成においてもこのような情報を取り入れる必要があると考えています。

## まとめと今後の展望

本稿では、私たちのグループで取り組んでいる対話処理技術の中から、「空気を読む」能力に焦点を当てた研究を紹介しました。対話システムが人間の会話に参画し、人々と共生するパートナーとなるためには、

個々の技術についてさらなる精度向上が求められます。一方で、より多様な対話処理技術にも目を向ける必要があります。例えば、現在の対話システムは基本的に1対1の会話を前提としており、複数人による対話の管理には依然として多くの課題があります。また、高齢者や子どもといった多様なユーザに対しても頑健に会話が行える技術の開発が重要です。さらに、対話中に変化する状況に応じてタスクを柔軟に遂行したり、コミュニケーション上の齟齬を適切にフォローしたりするためには、システムの対話戦略の研究も欠かせません。今後は、これらの対話処理技術に関する研究をさらに発展させていきます。

本稿で紹介した成果の一部は名古屋大学、東北大学との共同研究によるものです。

## 参考文献

- (1) Y. Chiba and R. Higashinaka: "Analyzing variations of everyday Japanese conversations based on semantic labels of functional expressions," ACM Transactions on Asian and Low-Resource Language Information Processing, Vol. 22, No. 2, pp. 1-26, 2023.
- (2) H. Koiso, Y. Den, Y. Iseki, W. Kashino, Y. Kawabata, K. Nishikawa, Y. Tanaka, and Y. Usuda: "Construction of the corpus of everyday Japanese conversation: An interim report," Proc. of LREC, pp. 4259-4264, 2018.
- (3) Y. Chiba and R. Higashinaka: "Dialogue situation recognition in everyday conversation from audio, visual, and linguistic information," IEEE Access, Vol. 11, pp. 70819-70832, 2023.
- (4) P. Brown and S. C. Levinson: "Politeness: Some universals in language use," Cambridge University Press, 1987.
- (5) I. Altman and D. Taylor, Social penetration: "The development of interpersonal relationships," Holt, Rinehart & Winston, Vol. 212, 1973.
- (6) R. Levitan and J. Hirschberg: "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," Proc. of INTERSPEECH, pp. 3081-3084, 2011.
- (7) Y. Chiba and A. Ito: "Speaker intimacy estimation in chat-talks based on verbal and non-verbal information," IEEE Access, Vol. 12, pp. 184592-184606, 2024.
- (8) 東中・光田・千葉・李: "Pythonと大規模言語モデルで作るリアルタイムマルチモーダル対話システム," 科学情報出版, 2025.
- (9) Y. Chiba and R. Higashinaka: "Investigating the impact of incremental processing and voice activity projection on spoken dialogue systems," Proc. of COLING, 2024.
- (10) K. S. Hone and R. Graham: "Towards a tool for the subjective assessment of speech system interfaces (SASSI)," Natural Language Engineering, Vol. 6, No. 3-4, pp. 287-303, 2000.



千葉 祐弥

今後も人間と共生するパートナーとしての対話システムの実現をめざし、人間どうし、人間・機械間のコミュニケーション技術に関する研究を進めます。

## ◆問い合わせ先

NTTコミュニケーション科学基礎研究所  
協創情報研究部