明日のトップランナー

NTT人間情報研究所

石井　亮 特別研究員

音声・言語・身体動作を複合的に扱い対話の仕組みを解明.「マルチモーダルインタラクション」の研究

IOWN構想の3つの主要技術分野の1つである「デジタルツインコンピューティング(DTC: Digital Twin Computing)」．外面だけではなく，意識や思考などヒトの内面を含む分身をデジタル世界に構築するには，ヒトが行うコミュニケーションの仕組みの理解・モデル化が不可欠です．今回は，音声・言語・身体動作などのマルチモーダル情報を扱い，心理状態の伝達メカニズムなど人間のコミュニケーションの機序の解明をめざす石井亮特別研究員にお話を伺いました．

◆PROFILE：2008年日本電信電話株式会社入社．NTTサイバースペース研究所（2008年～2012年），NTTコミュニケーション科学基礎研究所（2012年～2016年），NTTメディアインテリジェンス研究所(2016年～2021年, うち2019年～2020年はカーネギーメロン大学客員研究員），NTT人間情報研究所（2021年7月～），NTTデジタルツインコンピューティング研究センタ（2021年1月～）に所属.

コミュニケーションは音声や身振り手振りなど複数のモダリティの相互作用により成り立っている

◆「マルチモーダルインタラクション」とはどのような研究分野なのでしょうか.

人が他者と対話をするときには，声や言語だけでなく視線，表情，身振り手振りなど複数のモダリティ，すなわち「マルチモーダル情報」を発信します．それらは，複合的に利用され，人どうしで相互に影響を与えながら情報伝達を行います．このような，マルチモーダル情報を利用した相互作用を「マルチモーダルインタラクション」と呼んでいます．このインタラクションにおいて，特に重要なのは，発信されるマルチモーダル情報が統合的に扱われている点です．例えば，対話中に「ふざけるな」という発言が出てきたとします．文字だけをみると，発言者が怒っているのか冗談を言っているのか分かりません．しかし，声色や顔の表情を見てみると，声色は柔らかく，表情は満面の笑顔であれば，これは冗談を言っているのだろうということが分かります．このようなマルチモーダル情報を総合的に扱って，人の伝達する意図・メッセージの理解，これらの情報伝達の相互の伝達メカニズムの解明やモデル化を行うことが大きな研究トピックの1つです．さらにこのような認識・モデルを基盤として，人どうしのコミュニケーションの支援や，人と対話システムの円滑なコミュニケーションを実現することもマルチモーダルインタラクションの大きな研究トピックです．

本領域で私は，現在の心理状態を発信するための音声・言語・身体動作の表出生成メカニズムを解明しモデル化する「A）マルチモーダル表出生成」，また逆に表出された音声・言語・身体動作から送信者の心理状態を推定する「B）マルチモーダル心理状態推定」，そのような心理状態が受信者にどのように認知されたかを推定する「C）受信者の認知状態推定」といった，マルチモーダル情報の表出・認識メカニズムを多面的に扱う研究を行っています（**図1**）．さらに，応用研究として，送信者の心理状態と受信者の認知状態の差異を認識して，送信者の心理状態を受信者に正しく伝達することをサポートする「D）心理伝達支援手法」などに関する研究を行っています.

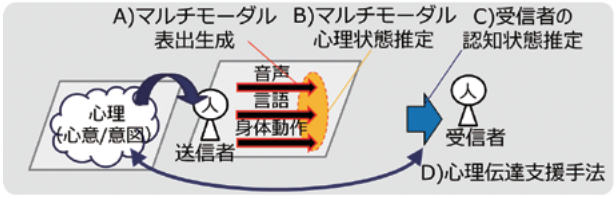

図1　マルチモーダルインタラクションの研究分野

◆具体的にはどのようなものが研究対象となるのでしょうか.

例えば，NTTコミュニケーション科学基礎研究所在籍時の私のテーマであった，「次の発話者と発話タイミングの予測」です．将来的に対話システムが，人と対話を行う中で適切なタイミングで話すためには，システムが次に誰が話すべきなのかを適時予測する必要があります．そこで，視線や頭の動き，発話前の呼吸などの人の微動作から次に誰がいつ発言するかを予測するモデルを構築しました．特に,対話中の人の呼吸動作を計測し，発話行動と関連付けて分析する研究は当時なかったため，この研究成果をまとめた論文はマルチモーダルインタラクション分野のトップ会議であるACM International Conference on Multimodal Interaction (ICMI)のベストペーパーを受賞しました.

その他，テレビ会議システムの映像を三次元化しユーザの視線や身体動作の自然なインタラクションを促進する「窓越しインタフェースMoPaCo」，話者の音声・言語から身振り手振りを自動生成する「身体動作生成技術」，対話中の音声・言語・画像情報から参加者の性格やコミュニケーションスキルなどを高精度に推定する「性格特性・スキル推定技術」，など，社内外の研究者と密に連携しながらマルチモーダルインタラクションに関連するさまざまな研究を行っています.

◆現在の課題は何でしょうか.

昨今，盛んに研究・利用されている機械学習技術を，マルチモーダルインタラクションの分野に適用して，人の伝達する意図の理解，情報伝達の相互の伝達メカニズムの解明やモデル化を行うことは有用なアプローチの1つです．機械学習技術適用においては，大量のデータ利用が前提となりますが，人のコミュニケーションを扱ううえで，シチュエーションの違い，人数，文化，人間関係, 場所など, 多くの要因によって大きく変わるコミュニケーション様式の多様性等によりデータ収集に多大な労力を要するといった，データ取得の困難性が大きな課題になります.

例えば，対話シーンのデータを収集する場合，まずは発言内容のデータを収集するために（前処理として音声認識技術を利用することもありますが），音声を聞きながら○○秒～○○秒まで××と発言したといった発話区間と発話言語を人手で書き起こすことを行います．表情，視線，身体姿勢に関しては，人画像を自動処理してある程度はデータ化できますが，モダリティによっては認識精度が十分ではないために，画像1枚1枚を人手で確認しながら，例えば，「誰が誰を見ているか」を確認して視線行動をラベリングしていきます．1秒間の動画は多くの場合，30枚程度の静止画像で構成されていますから，これらを人力でやるとなると相当な時間と手間がかかります．このように，人の対話のマルチモーダル情報を含むコーパスデータ（ここでは対話研究のためのデータ群）を構築するためには多大な労力がかかるため，すべてのデータを収集できずに，ある特定のシーンの少量のデータを用いて研究を行うことしかできないことが多くあります.

現状，研究者の間では，少し大げさな表現ですが，「データが集まれば研究は7，8割がた終わり」といわれるほど，データの収集はとても重要で膨大なコストがかかる作業です．人の対話の大量データを効率的に集めることは大きな課題ですが，これがなされれば，この分野の研究は一気に進むものと期待しています.

コミュニケーションを取り巻く森羅万象のモデル化をめざす

◆今後の研究の方向性について教えてください.

新たな試みとして，人間のコミュニケーションの機序を究極的に理解するために，コミュニケーションの森羅万象をモデリングする研究を推進しています．現状のコミュニケーションのモデリング技術は，人の行動を入力Xとして，人の感情，個人特性，場の状況といった1つの状態を出力Yとして定義し，その関係性のみを解明することに取り組んでいます．このような個々の研究は非常に興味深く面白い研究ですが，人のコミュニケーション全体のメカニズムの解明，モデリングという視点からみると，ごく一部の事象を切り取って扱っているにすぎません．私は，人のコミュニケーションで起きているさまざまな事象の関連性を統合的に解明，モデル化することへの取り組みを始めています.

一例ですが，人は音声・言語・身体動作といったマルチモーダル情報を実際の行動として表出（発信）するわけですが，そのような表出に至るまでには人の内部に心意・意図があるはずです．また，個人には性格，価値観といった個人特性があるはずです．さらに，対話の場において，人は人間関係や役割，雰囲気に応じて影響を受けながら，対話を施行しているはずです．**図2**はこのような関係性を簡易に図示した例です．ここでは，人間関係，役割，場の雰囲気などの状態を保持する「高次状態レイヤ」，音声・言語・身体動作といったマルチモーダル情報を実際の行動として発信・受信する「実行動レイヤ」，人の内部にある心意・意図「心理レイヤ」，性格，価値観といった情報を保持する「個人特性レ

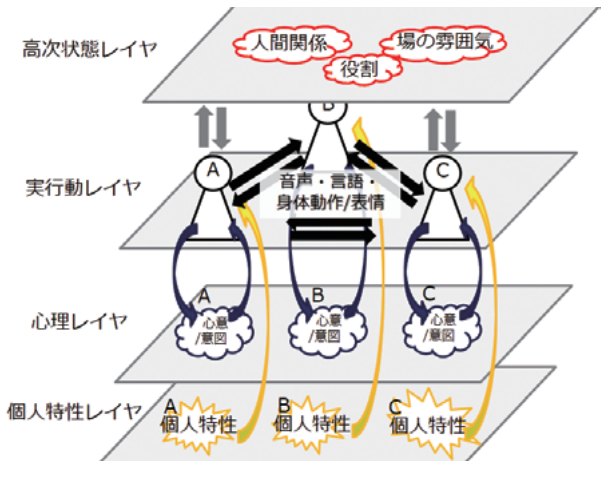

図2　石井特別研究員が考案するコミュニケーションモデル

イヤ」といった4つのレイヤに分けてコミュニケーションの万象をモデル化しています．

前述のとおり，実行動から次の行動を予測，実行動から個人特性や人間関係を個別に推定するといった研究は従来から多くありましたが，本来，コミュニケーションは４つのレイヤが相互に関係性を持ち統合的に動作しているモデルとして考えることができます．さらに図2はある一時刻でのコミュニケーション状態を示したものですが，このモデルは時間の経過とともに時系列に変化していくことでしょう．

こうした各レイヤどうしの相互関係性，時間変化を扱って，いわば「究極のコミュニケーションの森羅万象」をモデル化することが現在の目標です．

◆本技術によりどのようなことが可能になるのでしょうか．

究極のコミュニケーションの森羅万象モデルが実現されれば，対話の状況をシステムが高度に理解することはもちろん，将来の状況を予測・シミュレートすることができると考えています．これにより，主に3つのことが可能になると期待しています．

1番目は人のコミュニケーションのリアルタイム支援です．対話中に，例えば誰かが少し不機嫌になっていたらフォローするとか，1人がずっと話していたら「○○さんはどう思いますか？」と話者の切り替えを試みるといったシステムによるファシリテーションを行い，場を和やかにしてコミュニケーションの促進に寄与することが実現されます．

2番目は，究極の対話システムの実現です．対話システムが，人と同様に対話相手やコミュニケーションの状況を理解して，適切にコミュニケーションを取れるような存在になります．

3番目は，人のコミュニケーション能力のトレーニングです．現在，「褒め方の上手さの推定，トレーニングを自動化する」研究に取り組んでいます．「人を褒める」ことは大切なスキルで，どう褒めてよいのか悩んでいる方も多いと思います．コミュニケーションの森羅万象モデルでは，人が対話相手を上手く褒められていたかを，対話相手の心意・意図，特性，場の状況，対話相手の特性といった多様な観点から評価することを可能にします．評価結果により，「○○な対話状況で，△△な性格の対話相手が××な意図を持っているのだから，もっと□□なところを褒めてあげたほうが良い関係を構築できますよ」といったような，アドバイスをシステムが行うことで，人のコミュニケーションスキルを高めることができるアプリケーションを創出することをねらっています．

（今回はリモートにてインタビューを実施しました）

◆研究者をめざす方々へのメッセージをお願いします．

私は，めざしたいもの・実現したいものは学生時代からほとんど変わっていません．現在の研究テーマは自分のライフワークだと思っていて，かなりの思い入れがあります．やはり自分のめざしたいものを，人生をかけて取り組む，あきらめないということが大切なのではないかと思います．大学，企業，さまざまな環境で研究開発を行う機会があると思いますが，必ずしも自身が100%満足できる環境は存在しないと思います．時には，自身のめざす方向とマッチしない研究テーマに取り組まなければいけない状況も起こり得ると思います．そのような環境の中で，不平不満による研究意欲や自身の能力開発行動が消極的になってしまったとしたら，それは非常にもったいないことだと思います．今与えられている状況の中で，自身の最終目標に到達するために学べることは何か，どうしたら効率的に成果を出せるか，どうしたら自分自身の能力を活かして成功できるか，ということを一生懸命考えるべきです．そして，1つひとつ結果を積み上げていくことがとても大切だと思っています．そうする中で，必ず自分自身のやりたい研究テーマを行うチャンスはめぐってくると思いますし，また自分自身でそういう研究テーマを立ち上げるチャンスもあるはずです．

また，「人を巻き込む」ことも大切だと思います．私は現在，社外の研究者と4件の共同研究を進めていて，社外にも自身の研究を共に推進してくれる強力な仲間がたくさんいます．1人だけでやれることは非常に限られています．自身のめざすもの，実現したいものを一緒になって実現できる仲間をつくっていくことは非常に大事だと思います．

「マルチモーダルインタラクション」という研究分野は，人が好き，コミュニケーションに興味がある人に向いていると思います．マルチモーダルインタラクションは，人文学，工学の幅広い学術分野が関連する学際的な分野です．よって，1つの専門領域さえあれば，それだけで良い技術を生み出すことはできませんし，幅広い専門性が求められます．このような幅広い領域での専門性を習得するのは容易ではありません．そして，本分野はまだまだ発展途上の新しい領域です．このような分野を開拓し研究者として成功するためには，人のコミュニケーションのメカニズムを理解したい，新たなインタラクション技術で世の中を変えていきたい，という強い思いが重要だと思います．今は必ずしも専門的な知識がなかったとしても，そんな強い思いを持って研究に取り組めば，成功できる研究分野だと思っています．