Vol.2014-SLP-100 No.8
2014/2/1

テーマセッション：音声応用（音声対話）は今後こうなる！

中野 幹生[†1]　李 晃伸[†2]　駒谷 和範[†3]　東中 竜一郎[†4]

情報処理研究会 音声言語情報処理研究会（SIG-SLP）第 100 回記念シンポジウムにおいて，音声応用（音声対話）研究の流れを俯瞰し，今後の目標・応用や方法論を探ることを目的としたテーマセッションを実施する．本稿はこのテーマセッションのうち，音声対話に関する発表内容の概要を，登壇者がそれぞれ執筆したものである．

1. 「音声対話」の先にあるもの（中野）

音声対話システムの現状

音声対話システムの基本的な構成は DARPA Communicator Project が行われた 2000 年前後に確立され，音声理解とともに，文脈理解（意図理解），エラーハンドリング，話者交替，適応的対話制御などが基本的な課題として研究者間で共有された．現在は研究者がこれらの課題に引き続き取り組む一方，音声理解技術を用いた商用サービスが広く使われはじめている．上記の課題を解決していくことでよりよい音声対話システムが使われるようになるというのが現在の延長線上にあるストーリーであるが，果たしてそうだろうか．

対話システムの将来

音声対話にかぎらず，対話システムの将来を想像してみよう．どのようなタイプの対話システムが今後重要になるかは，どのような場面で対話システムが使われやすいかに依存する．

電車の中やオフィスなど社会生活において音声が使えない場面は多い．そのような場面ではキーボード入力が有効である．今や多くの人が非常に速い速度でキーボード入力ができるようになっており，キーボードによるインタフェースは今後も重要だと考えられる．したがってテキスト入力の対話システムが今後発展すると考える（出力はテキストだけでなく画像もあり得る）．

音声が使える場面では，音声だけではなく，画像入出力やその他のセンサ入力が併用可能な場合や複数のユーザが一つのシステムを使える場合が多い．すでにスマートフォンの音声 IF は画像出力を前提にしているし，GPS 情報など音声以外のセンサ入力も利用されている．一つのタブレットを二人以上で見ながら使う場合もある．このような状況を考えると，今後はマルチモーダル・マルチパーティ対話システムの研究が主流になっていくだろう．家庭やオフィスで役に立つロボットも研究が進むと考えられるが，ロボットのインタフェースとしてもマルチモーダル・マルチパーティ対話システムは重要である．

まとめるとテキスト対話システムとマルチモーダル・マルチパーティ対話システムの二つが今後の対話システムの主流になると考える．

意味処理の進化

その際に重要になる技術は何だろうか．一つは深い意味処理だと考える．すでに人々は対話システムが人工知能であることを期待しており，多様な言語表現を理解することが望まれている．特に，テキスト対話システムでは複雑な言語表現が入力され得る．現状では大量のテキストデータを利用した手法が多く用いられており，そこでは表現レベルのマッチングが使われている．しかし，それだけでは不十分である．例えば，「渋谷で 3 歳の子供を連れてっても大丈夫なイタリアンはあるかな」といったような依頼を理解するには，現状ではレビューに類似した表現があるようなレストランを探すのがせいぜいだが，本当に「3 歳の子供を連れて行っても大丈夫な」レストランをリストアップするには，その意味を理解しなくてはならない．「3 歳の子が食べられるものがある」という意味や「3 歳の子が騒いでも大丈夫な雰囲気である」という意味などの候補を出してユーザに聞き返し，その結果に基づいてデータベースを検索する必要があるだろう．

マルチモーダル対話システムの場合は，実世界にグラウンドされた意味理解が重要になる．例えば掃除ロボットに「あの辺の汚れているところを掃除して」と言って理解してもらえると良いだろう．この場合，「あの辺の汚れているところ」を画像センサ等の出力と合わせて理解し，地図上のエリアにマッピングしなくてはならない．このような実世界意味理解のための知識をあらかじめ用意するのは難しいので，インタラクションの中で知識を獲得する必要も出てくる [1]．

状況依存対話

マルチモーダル・マルチパーティ対話の研究は対話参加度推定や受話者推定などで成果が出ているが [2]，今後は対話の状況を利用した意図理解や行動生成の研究が進むだろう．例えば，画像認識を使うことで，対話の場面を理解し，ユーザが話す前に意図を推察することでより良いコミュニケーションができる．メニューのどこを指さしたり見ていたりするかを認識することで，何を注文しそうか推察し，音声理解がうまくいかなくてもコミュニケーションできる注文受付システムなどが作られるだろう．究極的には，ユーザがなるべく発話しなくても済むようなインタフェースが

†1 ホンダ・リサーチ・インスティチュート・ジャパン
†2 名古屋工業大学
†3 名古屋大学
†4 NTT

できるかもしれない．

研究の分業体制

マルチモーダル・マルチパーティ対話システムは非常に複雑なシステムであり，音声処理，画像処理，言語処理，ロボット制御などを統合しなくてはならない．対話システム研究者は統合に注力せざるをえず，個別の要素技術には手が回らないだろう．したがって音声処理研究と対話システム研究は分業が進むと考えられる． そのような状況では，対話システム研究者と音声処理研究者が独立に研究開発ができるよう, 音声処理の API がリッチになって行くだろう．韻律情報，話者 ID，話者位置を含め，どのような情報が必要かを対話システム研究者側から提案していく必要がある．

DARPA の ATIS (Air Travel Information Service) プロジェクトで研究された自由発話の音声理解技術は 20 年経って広く使われるようになった．上に挙げたような技術の研究は今端緒についたばかりであるが, それが 20 年後に花開くと期待している．

参考文献

1) 中野, 実用的な対話ロボットの構築に向けて―物理世界での言語インタラクションのモデルと技術課題―, メディア教育研究, 9(1), pp. S29--S41 (2012).
2) D. Bohus, E. Horvitz, Facilitating Multiparty Dialog with Gaze, Gesture and Speech, Proc. ICMI (2010).

2. 音声 IF・音声対話の本質的応用に向けて（李）

近年の音声認識精度の向上に伴い、テキスト入力手段としての音声認識は実用化したといってよい段階にある。これより数年は、Web 検索、ホームオートメーション、車室内機器操作、FAQ やマニュアルの検索、一問一答の情報エージェントとの対話といった、単純なタスクを遂行するシステムが多く開発されるであろう。また、家庭用ゲーム機やデジタルサイネージのように、画像やジェスチャーといった他のモーダルと組み合わせたマルチモーダルインタラクションの実用化も進み、人々が音声入力に触れる機会もますます増えることは確実である。

一方で、ユーザから見たときに「準備→話しかけ→確認」という一連の過程を要するという音声入力モードの性質は昔から大きく変わっていないように思われる。例えば、指先で機器を操作するタッチ操作は、スマートフォンやタブレット型端末の基本モーダルとして、近年細かい改善の積み重ねにより大きく使い勝手が進化し、近年デスクトップ OS までがタッチ操作を前提とした設計となるに至った。一方、音声は、VAD やハンズフリー音声入力を軸として頑健な入力に関する要素技術の積み重ねが着実に行われているものの、潜在的な入力エラーの可能性から、多くのシステムでは事前の準備（push-to-talk 等）および事後の確認（応答画面での確認や確認用対話）という長年のパラダイムから大きく離れていない。これらは現実的な機器への入力手段として、他のモードに比べて大きな弱点である。この弱点は本質的に不可避であるが、テキスト入力フロントエンドとしての応用ではその弱点が強調されてしまう。単純なタスクのみで用いられる限定された応用だけが広がる状況が続けば、音声研究者以外は「結局音声の利用は限定的なものでしかない」と感じるようになるのではないか。

「音声は人間の最も根源的なコミュニケーションメディアである」とは昔からの文言である。人間は有史以来ほぼ人とのみ会話しており、それゆえ、対面して話しかける相手に対してある程度の「人間的性質」を暗に要求する性質を持つ。機械への音声入力においても、少なくとも訓練されていない人間は、発話するという行為においてこの人間的能力、すなわち知的処理能力を相手（システム）が具備することを潜在的に求める。よって、音声対話のみならずあらゆる音声 IF において、システムが知的能力を持つ（とユーザが感じる）ことは本質的に重要である。これができない音声対話システムは、どれだけデザインが素晴らしく動きが流麗であっても、ただの玩具でしかなく、前述で述べた音声モードの弱点が勝って飽きられてしまう。

今後の音声対話の実用化のためには、人間らしい音声対話のための多様な知的処理の実現とともに、それを自然かつ直感的な形でユーザに表出する方法についても、知的処理と密接に関連付けて工学的に捉え研究していくことが必要であろう。意味解釈や発話解析といった知的処理の研究のほかに、エラーハンドリング、対話の継続性、状況や周辺環境、プロンプトな機械からの話しかけ、イニシアティブ制御といった様々な側面を、それらの適切なユーザへの表出と関連付けて取り扱う必要があろう。

これらは単純なルールでカバーできるものではなく、表出はデザインなど曖昧な側面を多く含むため、少なくとも統計に基づくデータドリブンなアプローチは肝要であろう。しかし、現状では機械と人間の実際のインタラクションに関するデータは言語資源や音声資源に比べ圧倒的に少なく、個別のシステムごとのケーススタディに陥りがちで、大規模なデータ収集も難しく、研究の手がかりが得にくい。

李らは徳田・大浦とともに、実用に関連した多様な音声対話システムの研究を行うための共通基盤としてオープンソースの音声インタラクション構築ツールキット MMDAgent を構築した。これは（1）システムモジュール間、およびシステムとコンテンツ間をできる限りシンプルに設計することでタスクごとの表層的構成要素（アバター、声、辞書、モーション等）および対話シナリオをシステムから独立して切り離し（2）それらを音声対話コンテンツとして独立させ自由な構築・流通を許すことでオープンプラットフォームによるデータの Web 的な蓄積を目指し、さらに（3）クリエイターを巻き込んだ作り手側からの音声対話システムの普及および多種多様な音声対話システムの出現を狙ったものである。2013 年度末には PC 版に続き

Android OS でコンパイルできるバージョン 1.4 を年末に公開した。対話記述は単純な FST であるが、独立モジュール化されており、任意のテキストベース知的処理システムと容易に入れ替えることができる。世の中に多くのデータが流通し、そこからより広い範囲を工学的に扱う音声対話研究が多く生まれることを期待している。

3. 言ったことの理解から意図の理解へ（駒谷）

音声言語を使った対話に期待されるのは，気軽に話しかけて，意図が通じることである．例えば，有能なアシスタントに,「昨日渡した書類を印刷しておいてください」と頼んだとしよう．この依頼だけからタスクを実行するには，昨日渡した書類がどの書類を指すのか（言語表現のシンボルグラウンディング）や，両面印刷かどうか，白黒印刷でよいのか（ユーザの選好の理解）などといった確認が必要となる．しかし，もし有能なアシスタント（かつ指示側と日頃からコミュニケーションが十分に取れている）なら，いちいち細かい条件を確認せず，その意図を短く確かめる程度で,指示側の意図を汲んでタスクを遂行できるだろう．音声言語を使う時点で，意図を全て言語化するということはほぼなく，音声による発話は，その対話の場における何らかの前提に基づいている．この前提なしに，相手の意図を理解することはできない．我々がたまに目にする難解で長大な書類として，契約や特許に関する文書があるが，このようになる原因のひとつは，場の前提が共有されていない状態で，誤解が生じないように全てを説明する必要があるからだと考える．音声言語による対話，特に身近で繰り返し話をする相手との対話は，この対極にあることが望まれている．

これが意味することは,「相手が言ったことだけを解析しても，相手の意図は必ずしも理解できない」ということである．つまり,「音声ファイルに対して音声認識を行い，その結果に対して，システムの知識の中から最も適切なものを一発話で応答する」という問題設定では，上記は達成できない．この問題設定は，音声認識や自然言語処理という既存の研究分野に沿って切り取られた，音声言語による発話の側面のひとつである．これがコミュニケーションにおいて最も重要な情報を含んでいることは間違いないが，それだけでは足りないことも事実である.「対話は共同行為である」という原点を見直す必要がある[1].

今後，以下の各点が音声対話システムにおいて（再度）重要な課題となると考える．

1. 相手のモデルを持ったシステム

現存のスマートフォン上のアプリでは，様々なデフォルトルールを設けることにより，一問一答を越える対話を回避しており，これにより「対話管理は退化[2]」している．素早く情報提供するためにこれは合理的な設定であるが，一方で，システム開発者が想定したデフォルトルールを，全てのユーザが一律に押し付けられているとも言える．使用するユーザやその状況に関するモデル化がさらに必要である．このモデル化には，音声以外のモダリティの利用も考えられるが，音声コミュニケーションに関するユーザのモデルも依然必要である．この一環として，後述するように，話す内容（言語レベル）とともに，話し方（信号レベル）や社会的規範に関するモデルも必要である．

2. 膨大な背景知識の構築とその自動獲得

これは 20 世紀の人工知能研究で取り組まれてきた課題で，そこでは人手で知識を論理式などで記述するというアプローチが採られた．しかしこれら全てを人手で記述するのに限界があるのは周知のとおりである．特に，矛盾なく一貫した知識を決定的に記述するのは難しいため，何らかの自動化が必要である．また，音声対話システムでは，一般的な辞書的知識もさることながら，対象ドメインにおける知識が重要である．このようなドメイン知識のラピッドプロトタイピング技術も,工学的には必要となる.さらに,現存の音声対話システムを「賢く」しているのはシステム開発者であり,システム自身は対話から何も学ばない.様々な知識を対話から獲得・学習する枠組みも必要である．

3. 発話状況の理解

「気軽に話しかける」ということは，話す内容を全て整理して，流暢に話すことを前提にはできない．逐次的に，言い淀みを挟みながら話すような場合でも，システムはその状況を理解し，適切に総合して解釈する必要がある．つまり，言い淀みにより誤って分割された音声ファイルを個別に扱っていても，相手の意図は理解できない．音声認識のロバストネスという観点でも，この状況の理解は必須である．

発話という行為には階層性がある[1,3]．上記の例は発話の信号レベルでの理解に失敗している例である．音声ファイルから音声認識結果として得られるのは言語レベルの情報であり，発話の一側面である．これを扱うだけでは不十分であることは前述のとおりである．

さらに,Clark が提唱するように,対話ではそれぞれのレベルで 2 者の間に共同行為が成り立つはずである[1]．我々は社会的レベルでも対話が成り立っていることを入力音の解釈に利用しようとしている．具体的には，人間がロボットに対して話しかけられると感じるか否かを予測し，これを雑音棄却の事前確率として利用する[4]．ここでは，ユーザがロボットに対して社会性を感じており，話しかけるタイミングを決める際に，任意のタイミングではなく，相手（ロボット）の状況を慮ることを仮定している．これは，人間が人工物を無意識に擬人化して扱うという心理学的知見[5]に加え，音声言語を使っており，かつ相手がヒューマノイドロボットである場合には，人間は，人間同士の対話のプロトコルに準じてシステムと対話するに違いない，と

いう考えに基づいている．現状の音声対話システムでは，システムは任意のタイミングでの任意の入力を処理している．これに対して，社会的レベルで対話が成り立っており，人間同士の社会的規範が適用されているなら，任意の入力を同じ重みで解釈する必要はなく，社会的規範に沿った発話の方が，相手に対する発話としてもっともらしい，と解釈することができる．このように，発話を信号レベルから社会レベルまで階層的に捉え，その内容だけでなく，それが行われた状況を理解することは，ロバストネスを高めるうえでも有用である．

以上のように，「気軽に話しかけて意図が通じる」音声対話システムを実現するには，解決すべき課題は山積している．「既に切り取られた音声ファイルに対する最適な応答を探す」という問題設定を超え，より広い視野で音声対話システム研究を考えていきたい．

参考文献

1) Herbert H. Clark, Using Language, Cambridge University Press (1996).
2) 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).
3) John L. Austin, How to Do Things with Words, Oxford University Press (1962); 坂本百大 訳，言語と行為，大修館書店 (1978).
4) 杉山貴昭, 駒谷和範, 佐藤理史, ロボットへの話しかけやすさモデルの評価と個人差や教示による変動への対応, 人工知能学会論文誌, Vol.29, No.1, pp.32-40 (2014).
5) Byron Reeves and Clifford Nass, The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places, Cambridge University Press (1996).

4. 雑談対話システムに向けて（東中）

タスク指向型対話システムについては日常生活に入り込む程度に実現するだろう．ただ，コンピュータによる雑多な話題での日常会話（いわゆるおしゃべり，我々は雑談対話と呼ぶ）については限定的にしか実現しないと思う．前者は今後の着実な進歩の先にあると思うが，後者については，要素技術がまだまだ不十分である．

タスク指向型対話システムについて述べると， 2000 年代初頭から，インタラクションの質は大きく変わっていない．しかし，音声理解，対話制御，発話生成などの各モジュールは着実に進歩している．音声入力の便利さも，一般ユーザに少しずつ理解されてきており，これからタスク指向型対話システムについては必要とされる場所に導入が進んでいくのは確実だ．タスク指向型対話システムは，近年の Dialogue State Tracking Challenge [1]などの評価型ワークショップなどの動向を見ても，よりロバストになり，信頼のおけるモダリティとして使用される方向にある．タスク指向型対話システムについて言えば，タスク達成の一つのモダリティとしての利用が進められるだろう．なお，音声対話単体でのタスク達成率はどうしても 100%にはならないため，音声対話のみのアプリケーションはほとんど普及しないだろう．タスク達成率の観点から言えば，将来的なデバイスは，より多くのモダリティを扱えるように進化すると思う．

雑談対話について述べると，これはタスク指向型対話とは本質的に違う．何を達成すればよいのかも不明確で，扱うべき知識の質・量も比較にならないくらいに多い．私はここ 1〜2 年雑談対話システムの研究に着手し[2]，また，そのサービス化（ドコモのドライブネットインフォにおける雑談対話機能[3,4]）にも関わっているが，ユーザの話す内容は多様かつ複雑で，現在の言語処理技術でうまく扱えないものばかりである．すぐに思いつくものだけでも，照応解析，述語項構造解析，談話関係認識，含意認識，多義語解消，因果関係認識といった要素技術の進展が必須である．そして，現時点では，どの課題も解決が難しい．

これらの要素技術の改善には共通のコーパスや評価セットから作らなくてはならない．また，個々のモジュールの精度が向上したとしても，統合の方法論も必要である．さらに，やりとりをどのようにして改善していくかという方法論も必要である．これまでの自然言語処理研究の歴史を見ても，要素技術の研究が始まって，世の中に使われるようになるまで 10 年以上はかかる．たとえば，NTT で質問応答システムに取り組み始めたのが 2000 年ごろであるが，しゃべってコンシェルにその技術が用いられるようになったのは 2012 年である[5]．また，タスク指向型対話システムの歴史から考えても，要素技術の統合にさらに 10 年はかかる．対話処理のような研究コミュニティの小さい分野で，かつ，コーパスの作成に時間がかかるような分野ではなおさら時間がかかるだろう．10 年から 20 年後では各要素技術が進展しつつも，人間と同様の会話が可能な雑談対話システムの実現には至らないと思う．

技術的に成熟していないにもかかわらず，雑談対話システムのニーズは高い．独居世帯や介護の現場でも必要とされている．商用の音声エージェントサービスにおいても，ユーザが雑談をしようとする現象が見て取れる．このような背景から，雑談対話の機能の中で重要と思われるもの，もしくは，実現可能なものから順次切り出されて実装されていくことになるだろう．手続きとしては，まず雑談対話とはどういう機能からなるかを洗い出し，それらを重要性と実現性（用いられる要素技術の成熟度）の観点から優先度付けをして取り組んでいくことになる．

重要性という意味では，雑談が持つ，人間社会で実際に役立っている機能が重要と言えるだろう．たとえば，雑談はビジネスの場では「相手の情報を得る」ためのツールとして重要である．たとえば，営業担当者は雑談の中で顧客の好みを把握し，的確な提案につなげている．カウンセリングや心理療法などにおける「相手の話を聞く」機能も重要である．傾聴ボランティアが存在するように，人に話を聞いてもらうこと自体に価値を見出す人は多い．「相手と人

間関係を築く」ための雑談も重要である．雑談によって相手との共通項などを知り，今後の親密な人間関係に発展させることができる．ここで挙げたような機能をその時々の要素技術で実現していくことになる．これらについてヒューリスティックを用いた取り組みはすでにいくつかあるが，一般的な解決法はまだ提案されていない．

雑談のどの機能を実現するにしてもユーザの発話を理解することは第一ステップである．研究はまずここから進めていくことになるだろう．すなわち，「オープンドメイン発話理解」が当面の研究課題となる．関係する要素技術としては，照応解析，述語項構造解析，談話関係認識などがある．幸い，これらについては比較的研究はされている方ではあると思うので，要素技術の進歩のペースから言えば10年後には，これらの要素技術を組み合わせたようなオープンドメインな発話理解系が実現できているだろう．発話理解の直接的なアプリケーションは「相手の話を聞く」システムであるが，これは有望なサービスになると思っている．自分のことを伝えることに多くの人は熱心であるし，システムによる聞き返しがあったとしても，伝えようとすると思われるからである．このシステム実現の後，理解結果に基づいて，ユーザに働きかけていくアプリケーションも作られていくだろう．ただ，人間同士のように，議論をしたり，意見の交換をしたりといった会話の実現はまだ先になるだろう．システムが自分で考えて人間と会話をする必要性から議論していく必要があると思う．

参考文献

1) Jason Williams, Antoine Raux, Deepak Ramachandran and Alan Black, The Dialog State Tracking Challenge, In Proc. SIGDIAL, pp.404-413, 2013.
2) Hiroaki Sugiyama, Toyomi Meguro, Ryuichiro Higashinaka and Yasuhiro Minami, Open-domain Utterance Generation for Conversational Dialogue Systems using Web-scale Dependency Structures, In Proc. SIGDIAL, pp.334-338, 2013.
3) https://www.nttdocomo.co.jp/service/information/drive_net/drive_net_info/
4) 大西可奈子，吉村健，コンピュータとの自然な会話を実現する雑談対話技術，NTT DOCOMO テクニカル・ジャーナル，Vol.21，No.4，pp.17-21，2014.
5) 東中竜一郎，貞光九月，内田渉，吉村健，しゃべってコンシェルにおける質問応答技術，NTT 技術ジャーナル，Vol.25，No.2，pp.56-59，2013.