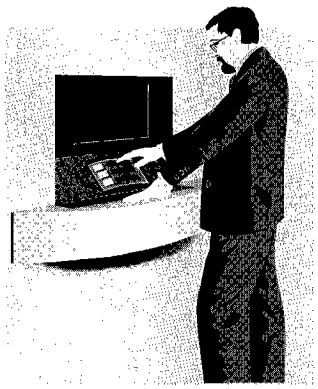
3 話がはずむ音声対話システム

中野幹生
NTTコミュニケーション科学基礎研究所

なぜコンピュータが会話しなくてはならないのか

音声認識技術が身近になってきた．パソコン用の安価な音声ワープロ（ディクテーションソフト）や，音声認識を使ったゲームが売れている．音声認識の性能は着実に上がっており，さまざまな分野での応用が期待されている．そして次に求められるのは，単に音声入力を受け付けるだけではなく，人間と音声で会話する「音声対話システム」である．完全な音声対話システムができれば，パソコンと音声で会話するだけで，すべての用が足せる．キーボードが嫌いな人にとっては，夢のシステムである．しかし，現実には，音声認識の性能は完全ではない．ただでさえ複雑で使いにくいコンピュータが，誤認識でわけの分からない動作をすると，ストレスなしには使えないだろう．CGIやタッチパネルなど，誤認識の起こり得ないメディアを使えば，音声対話なんかいらないと考える人も多い．

それでも音声対話システムは有望であると，我々音声対話の研究者は考えている．音声対話システムが実現すれば，多少の欠点はあっても，便利なことがある．キーボードがいらないので，小さい携帯電話に話しかけるだけで，会社のコンピュータの情報が得られる．銀行のATMで振込み先を指定するのに，タッチパネルでいちいち入力しなくてもよい．自動車が便利になってもオートバイや自転車がなくならないのは，オートバイや自転車にはそれなりの利点があるからであるが，これと同じで，タッチパネルが有効な場面もあれば，たとえ誤認識があっても音声対話が有効な場面があるのである．しかも，音声対話は，単なる音声コマンドと違って，誤解があってもインタラクションによって解消できるという利点があり，その意味からも有望であると考えられる．



世界的にみると，音声対話システムの研究に力が注がれるようになってきている．将来音声対話システムの性能が上がって，コンピュータが気軽に音声で使えるようになったとき，英語の音声対話インタフェースのみが存在して，日本語は受け付けられないとしたら悲惨な状況になる．それは，コンピュータが日本語を扱えなかったころのことを想像するとよい．コンピュータを使ったビジネスや技術開発で，日本語話者は大きなハンディを負う可能性がある．しかし，日本語の音声対話システムの研究や開発を行うには，どうしても日本語のネイティブスピーカの努力が必要である．日本語の音声や言語に関する知識が必要だからである．したがって，日本でも音声対話システムの研究をますます発展させていく必要があるだろう．

対話の言語処理技術

一般的に，音声対話システムは，図-1のような構成で作られている．まず，ポーズなどの情報を使ってユーザの発話の音声区間を切り出し，その音声データを音声認識システムによって単語列に変換する．ここではどんな単語列が発話されやす

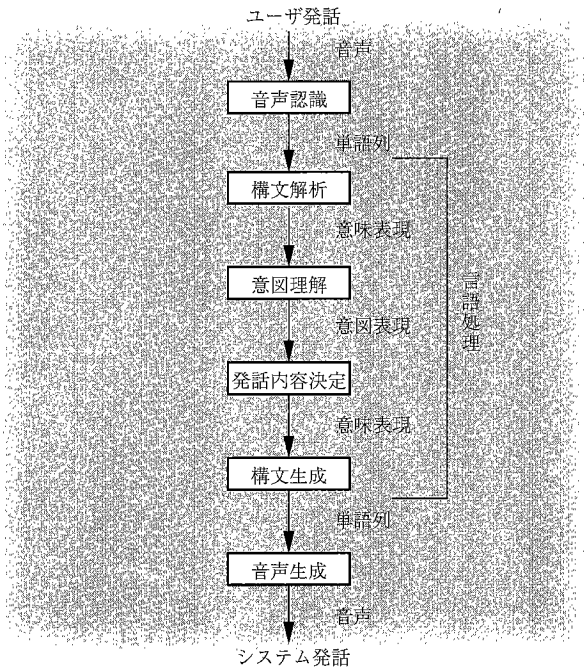


図-1　音声対話システムの構成

いかという情報を使う．これを言語モデルと呼んでいる．ディクテーションソフトなどでは新聞記事などから得られた確率的言語モデルを使うのが普通だが，対話システムの場合はタスクに無関係な語彙はあまり発話されないため，タスクに関係ある単語のネットワーク（有限オートマトン）で言語モデルを記述する場合も多い．構文解析部では，単語列の文法的な構造を解析して，意味表現（文の意味の論理式などで曖昧性なく表現したもの）を計算する．意図理解部では，発話意味表現からユーザの意図を推測し，発話内容決定部で適切な応答や説明の内容を決定する．さらにこの内容を表す単語列を，構文生成部で得る．構文解析から構文生成までのプロセスをまとめて言語処理と呼ぶことができる．音声生成部では単語列を音声に変える．音声合成システムが使われることが多い．対話システム全体の性能の向上のために，これらの要素技術の研究が行われている．これに関しては，最近の文献2)，7)，10）を参照していただきたい．

音声対話システムの研究動向を把握するのに重要な点は，音声認識と音声生成は音声処理研究者によって，そして言語処理は言語処理研究者（AI研究者と呼んでもよいだろう）によって研究されているということである．各々が，自分たちの研究している要素技術の良さを見せるために，総合的な音声対話システムを作っているが，それはあくまでアプリケーションに過ぎない．したがって，一口に音声対話システムといっても，そのフォーカスが当たっているところは違うのである．

音声処理研究者が主に音声認識・理解に焦点を当てているのに対し，言語処理研究者は，音声を直接には扱わないで，発話を文字列に変換した後のことだけに専念している．これは，音声を扱うためには，信号処理の知識が必要だが，言語処理研究者は記号処理が専門で，必ずしも信号処理の知識を持っているとは限らないことが1つの理由である．さらに，言語処理研究と音声処理研究は異なる分野とされてきたので，両方の知識がある人は少なかった．

言語処理での対話研究はもともとキーボード入力を対象として始まり，自然言語によるデータベース操作の研究（いわゆるキーボード入力による質問応答システム）を経て，プラン（ある目標を達成するにはどのような行動をとればよいかを記述したもの）を用いたユーザの意図理解や説明の生成といった，AI技術を駆使した対話システムの研究に発展していった．これらの研究で扱われてきた問題は，たとえば，「ユーザの意図を正確に理解をするには，どのようなプラン認識手法が必要か?」といったものである．ユーザの意図を正確に理解することによって，「第3会議室は空いていますか?」という質問に対して，「空いていません」だけではなくて，「第4会議室なら空いています」と答えることが可能になる．すべてがそうだとはいえないのだが，これらの研究の興味の対象は知的エージェントとしての人間の対話行動のモデル化であり，言語哲学や心理学などの周辺領域ともからめて，深い議論が行われている[1)]．

音声対話システム実現の問題点

1980年代末ごろから，音声認識の性能が飛躍的に高くなったことと，音声認識の知識がなくても使えるパッケージ化された音声認識・合成システムが使えるようになったことから，AI的な対話システムと音声認識システムおよび音声合成システムをつなげて音声対話システムが作られるようになった．このキーボード対話システムから音声対話システムへの発展は，今まで言語処理研究者があまり考えてこなかった問題を浮かび上がらせた．特に発話理解に関する問題をいくつかリストアップしてみよう．

• 書き言葉の文法では扱えない発話

ユーザの発話の仕方にまったく制限を加えない

と，ユーザの発話は書き言葉の文とは大きく異なる．言い直し，言い淀み，省略など，書き言葉の自然言語処理技術が対応していなかった現象を扱う必要がある．たとえば，従来は，「新大阪までの新幹線を予約したい」というような発話を対象としていたが，実際のユーザの話し方は，「えーとですね，新大阪まで行きたいんですけどー，あのー，新幹線予約したいんですがー」といった感じである．

- **音声認識の誤認識**

音声認識は完全ではない．特に，ユーザに自由に話させると，認識率は大きく下がってしまう．細かい機能語（助詞や助動詞）までは正しく認識できない．内容語（名詞や動詞など）の認識率だけでもある程度確保しようとすると，認識可能な語彙数を減らさざるを得ず，結果としてタスクは単純なものになってしまう．将来音声認識技術が向上すれば，100%の認識率が得られると考える人もいるかもしれないが，人間でも聞き間違えることは頻繁にあるし，完全に聞き取れる場合でも，言語以外のいろんな知識を総合して聞き取っていると考えられる．したがって音声認識部だけで100%の認識率を達成できる可能性は非常に少ないといわざるを得ない．

実用的な音声対話システムの実現には，このような問題を解決することが必要である．人間の対話行動の計算モデルを作るという観点ではなく，実際に動く音声対話システムを作るという観点から考えると，キーボード対話システムや対話の計算モデルの研究で検討されてきた多くの問題を解決することの緊急度は下がる．音声認識率の制限によりタスクを単純なものにするならば，ユーザの発話のパターンは比較的限られるので，ユーザの意図の理解はあまり困難ではなくなり，複雑なプラン認識手法などを用いる必要がなくなってしまう．また，現在のプラン認識は音声認識の誤りを想定したものではないので，そのまま使ってもうまくいかない．Ward[9]は，音声理解において，従来のAI的なアプローチを用いることの限界について考察している．

音声処理研究者は，比較的早くから音声認識を使って音声対話システムの研究をしてきており，上で述べたような問題に取り組んでいる．たとえばTOSBURG II[8]というシステムでは，ユーザの自由な発話を理解するため，キーワードスポッティングを用いている．キーワードとはそのシステムのタスクで重要な働きをする単語のことである．

S1: はいどうぞ
U1: えーとですね 松本幸四郎の娘が出てるドラマ 予約したいん<はい>ですけど
S2: 松たか子が出てるドラマですか
U2: はい
S3: はい
U3: はい そうです
S4: えーっとラブジェネレーションという番組が<はい>あります 夜のドラマで<はい>松たか子が<はい>出演します
U4: ラブジェネレーションですか
S5: はい
U5: はい
S6: ラブジェネレーション
U6: はい
S7: 予約しますか
U7: 予約します
S8: かしこまりました ほかにございますか
U8: それだけでいいです
S9: ありがとうございました

図-2 対話例
Uはユーザ発話，Sはシステム発話．＜＞は聞き手のあいづち

TOSBURG IIは，音声認識をしてから言語処理によって意味を求めるのではなく，音声からダイレクトにキーワードの候補を抽出し，それを組み合わせることによって意味を理解する．TOSBURG IIのタスクはハンバーガー店での注文タスクで，「ハンバーガー」「コーヒー」「3つ」「ください」などがキーワードである．このようなシステムでは，ユーザの自由な発話を理解することに主眼がおかれているため，言語処理，特に応答内容決定などは比較的簡単なものになっている．

一方言語処理の方からも音声理解の難しさにトライする研究が出てきている．たとえば，ロチェスター大学では，発話の音声認識結果に含まれる誤認識を検出するために，タスクのドメインに依存した確率モデルを用いて認識結果のフィルタリングを行う方法を用いている．これと，ロバスト性を高めたプラン認識やプランニングの技法を結合して，物流の計画立案のサポートを行うTRIPSというシステムを構築している[3]．このように，AI的な対話技術を音声対話システムに用いるためには，新たな工夫が必要である．

音声処理と言語処理を結合してできた他のシステムについては，たとえばヨーロッパの研究動向を解説した文献4）が参考になるだろう．

話がはずむ対話システムへ

言語処理と音声処理の結合により，単純なタスクならば，音声対話を用いて用を足すことができ

図-3　DUG-1の外観

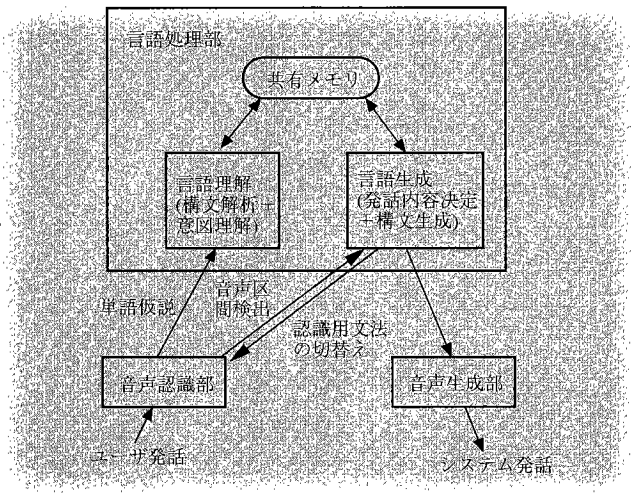


図-4　DUG-1の構成
各モジュールは並行動作している．詳しくは文献6）を参照

るようになってきている．しかしながら，実際にみんながそのようなシステムを使ってチケット予約などをしようと考えるかというと，残念ながらまだ課題は多い．

問題点の1つに，会話が堅苦しくなってしまうという点をあげることができる．音声対話システムとユーザの会話は，まるでトランシーバを通した会話のようになっていて，ユーザが話している間はシステムは無反応で，また，システムが話している間はユーザが何を言ってもシステムは関係なく話し続ける．これは，ユーザにとってはあまり気分のよいものではない．また，ユーザが話し終わってから（つまり長いポーズをおいてから）理解して応答するので，応答が遅れてしまう．

DUG-1[6]は，このような問題点にチャレンジしてできた音声対話システムである．DUG-1はユーザの発話を話すそばから理解し，ユーザの発話中でもあいづちを打つなどの反応をすることができる．また，ユーザの発話終了後すぐに応答することができる．さらに，ユーザの反応を考慮しながら少しずつ説明を進めていくので，ユーザの割り込みに適切に対処できる．これらの特長に加えて，もしユーザの話したいことをうまく引き出せるような応答や説明ができれば，「話がはずむ」対話をすることができるだろう．

図-2はTV番組の録画予約をタスクとした対話例である．ユーザが気軽に話せるように，ディスプレイ上に擬人エージェントを表示し，生成内容に合わせてうなずいたり首をかしげたりすることができる（図-3）．

DUG-1に上記のような特長があるのは，アーキテクチャが従来の音声対話システムとは違うからである（図-4）．このアーキテクチャでは，理解を行うプロセスと，応答を行うプロセスが並行して動作しているので，聞いている途中でも応答ができ，また，話している途中でも割り込みに対処できる．さらに，話すそばから理解を進めていけるのは，音声認識部と発話理解部が密結合しているからである．従来の結合方式では，音声認識部は音声区間の終了（ポーズなど）まで認識結果を言語処理部に渡すことができなかったが，DUG-1では，ISTARプロトコル[5]を使って，音声認識と同時に単語候補を言語処理部に逐次的に送る（図-5）．これにより，ユーザの発話中でも理解を進めることができ，あいづち応答を行うことができる．

実用性という観点から見ると，DUG-1の問題点がいくつかあげられる．1つは認識できる語彙が少ないことである．語彙数を増やせば，認識速度も遅くなり誤認識も増える．また，誤解が起こったとき，それを解消する機構が十分でないため，タスクの達成率が100%ではない．このような点を改良していくことにより，実用的な音声対話インタフェースを作ることができるだろう．

実用的な音声対話システムに向けて

おわりに，音声対話システム研究の今後の課題をいくつかあげる．

• 要素技術の密結合

音声対話システムの実現には，さまざまな要素技術が必要で，さらに，それらの有機的な結合が必要である．上に述べたISTARも1つの例だが，ほかにもあげると，たとえば，音声生成と対話処理

の密統合が考えられる．生成される音声の速度，パワー，イントネーションなどは，対話のどの場面でその音声が発せられるかによってダイナミックに変える必要があるだろう．音声生成部に単純な文字列を送っただけではこの制御はできない．また，ユーザからの割り込み発話があったとき，生成しようとしていた発話のうちどこまで生成していたのかということが分からないと，割り込み発話の意味は正確には理解できないので，音声生成部はその情報を言語処理部にフィードバックする必要がある．

• 音声対話システムを「作る」こと

人間同士の対話を分析して，人間の対話行動のモデルを作ることは，言語学的，認知科学的に重要であるだけでなく，音声対話システム研究に重要な示唆を与えることは疑うべくもないが，それだけでは，インタフェースとしての音声対話システムの研究を本質的に進展させることはできない．理由はいくつかあるが，まず，人間同士の対話では，対話者の社会的な関係などの要因がからむのに対して，音声対話システムはあくまで機械であり，ユーザの行動は人間相手のときとは異なる．また，人間同士で対話させても，人間の対話者すべてが，機械のモデルになれるほど，「良い」対話ができるわけではない．説明がうまくない場合，流暢に喋れない場合，相手に対して失礼なことを言う場合もある．さらに，人間同士の対話を分析するには人手による書き起こしテキストを使うのが一般的だが，書き起こすときには，その対話の内容に関する知識を用いている．しかしながら，これと同じことを現状の音声対話システムに行わせるのは無理なので，書き起こしを分析して得られた知見が現状のシステムに使えるとは限らない．したがって，音声対話システムの構築技術の進展には，人間同士の対話の分析だけでは不十分で，やはり，音声認識を使って実験システムを組み上げ，実際にユーザに使ってもらいながら，何が本質的な問題なのかを探っていくことを避けて通るわけにはいかない．

• 良い音声対話システムの基準

要素技術の性能を向上させるのはもちろん重要だが，全体のシステムのパフォーマンスという観点から，個々の要素技術を洗い直していく必要がある．たとえば，音声認識部の性能は単語認識率で測定するのが一般的だが，「はい」と「いいえ」を間違えるのと，「は」と「が」を間違えるのでは，全体のパフォーマンスに与える影響は大きく異なる．しかしながら，システム全体のパフォーマンスをどう評価するかも難しく，今後検討していく必要がある．

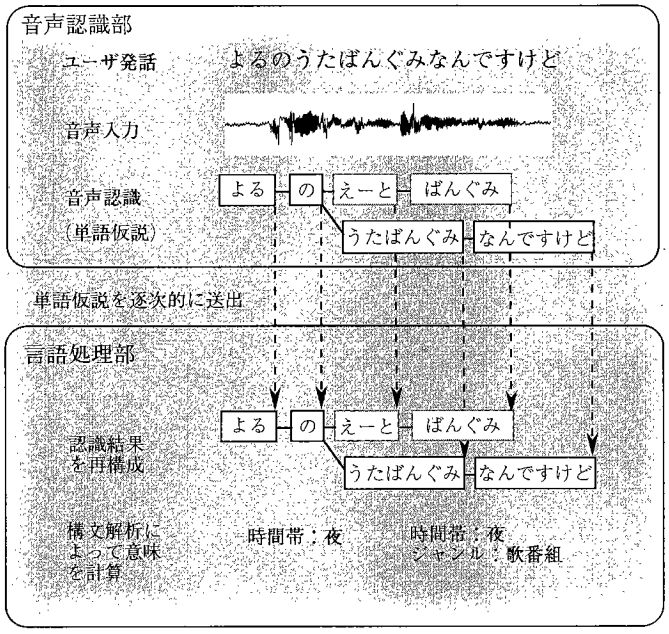


図-5　ISTARプロトコル

一般的には，今後の音声対話システム研究の進展の鍵となるのは音声処理技術と言語処理技術の統合ではないかと思われる．そしてそれは，音声処理研究者と言語処理研究者の密な連携によって可能になっていくだろう．

謝辞　DUG-1の製作にかかわった皆様，および貴重なアドバイスをいただいたNTTコミュニケーション科学基礎研究所の皆様に感謝いたします．

参考文献

1) Cohen, P.R., Morgan, J. and Pollack, M.E. (ed.): Intentions in Communication, MIT Press (1990).
2) 堂下, 新美, 白井, 田中, 溝口: 音声による人間と機械の対話, オーム社出版局 (1998).
3) Ferguson, G. and Allen, J.F.: TRIPS: An Intelligent Integrated Problem-Solving Assistant, pp.26-30 (1998).
4) Fraser, N.M. and Dalsgaard, P.: Spoken Dialogue Systems: A European Perspective, In Proceedings of 1996 International Symposium on Spoken Dialog, pp.25-36 (1996).
5) Hirasawa, J., Miyazaki, N., Nakano, M. and Kawabata, T.: Implementation of Coordinative Nodding Behavior on Spoken Dialogue Systems, ICSLP-98, pp.2347-2350 (1998).
6) 中野, 堂坂, 宮崎, 平沢, 田本, 川森, 杉山, 川端.: 柔軟な話者交代を行う音声対話システムDUG-1, 言語処理学会第5回年次大会論文集, pp.161-164 (1999).
7) 島津: コンピュータと人間の会話: 現状と課題, 情報処理, Vol.39, No.3 (Mar. 1998).
8) 竹林: 音声自由対話システムTOSBURG II－ユーザ中心のマルチモーダルインタフェースの実現に向けて－, 電子情報通信学会論文誌, J77-D-II, pp.1417-1428 (1994).
9) Ward, N.: Second Thoughts on an Artificial Intelligence Approach to Speech Understanding, 人工知能学会研究会資料SIG-SLUD-9601-3, pp.16-23 (1996).
10) 小特集「音声対話システムの実力と課題」, 日本音響学会誌, Vol.36, No.11 (1999).

（平成11年3月3日受付）