音声認識技術を使った人と機械との対話

技術解説

駒　谷　和　範*

Human-Machine Dialogue based on Speech Recognition Technology

Key Words：Spoken Dialogue Systems, Automatic Speech Recognition, Language Understanding

1．はじめに

人間の発話に対して，それを理解し応答を返すシステムが一般ユーザにも広く使われ始めている．音声認識技術に基づき，人と言葉を使ってやりとりするシステムを音声対話システムと呼ぶ．有名なものとして，Siriやしゃべってコンシェル，音声アシストなどのスマートフォン上の音声応答アプリや，ソフトバンクのPepperのようなロボットが挙げられる．

音声対話システムの基本構成を図1に示す．主に，音声認識，言語理解，対話管理，応答文生成，音声合成の5つのモジュールにより構成される[1]．まず音声認識部では，ユーザの話す音声信号をテキスト（単語列）へと変換する．言語理解部では，その単語列に対する理解結果を得る．ここまでが入力理解と呼ばれる処理である．この後，入力理解結果に基づいて，対話管理部でシステムの内部状態を更新し，例えば天気検索の場合は天気のデータベース（DB）を検索するなどして，応答内容を得る．応答文生成部でこの内容を表す文を生成し，それを音声合成部でシステム発話の音声信号に変換する．これが音声対話システムの一連の処理の流れである．

本稿では，このような音声対話システムの構成のうち音声認識や言語理解などから成る入力理解部分について主に述べる．2節では，音声認識技術のこれまでの発展について，3節では言語理解について簡単に述べる．これらは現在のスマートフォン上の応答アプリなどでも使用されている技術である．続いて4節では，入力理解に関して図1の枠組みに含まれない課題について述べる．5節では音声対話システムを対話ロボットに展開する際の課題について述べ，今後の展望を示す．

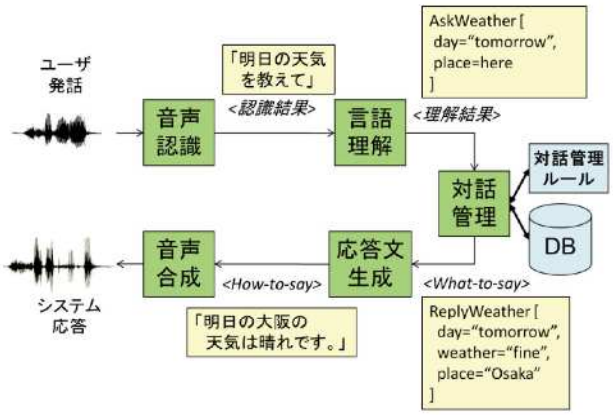

図1　音声対話システムの5つのモジュール

2．音声認識技術の進歩

音声認識とは，人の話す音声信号を入力として，それに対応する文字列や単語列を出力する処理である．

音声認識の研究は古くから行われており，日本でも1960年前後からパターン認識の一分野として研究が続けられてきた[2]．2000年前後にはPCの普及の流れを受け，IBM社が一般ユーザ向けの音声認識ソフトウェアViaVoiceを開発して販売し，一時テレビCMにも登場していたが，広く使われるには至らなかった．その理由として，使用の前に話者ごとに数百文程度の音声を録音するエンロールと呼ばれる作業が必要であったことなど，当時の技術が未成熟であったことが挙げられる．また，持ち運びしにくいデスクトップ型PCが当時は主流であり，これを使って音声認識を行うというシチュエーショ

* Kazunori KOMATANI
1975年12月生まれ
京都大学大学院情報学研究科知能情報学
専攻　博士後期課程修了（2002年）
現在、大阪大学　産業科学研究所
教授　博士（情報学）　知能情報学
TEL：06-6879-8415
FAX：06-6879-2123
E-mail：komatani@sanken.osaka-u.ac.jp

ンが限定的であったことも一因であろう．

2000年代後半から2010年代にかけて，クラウドコンピューティングに基づくクラウド型音声認識が登場し，音声認識システムの開発運用パラダイムは大きく変わった[3]．それまではユーザの手元にあるコンピュータ上で音声認識処理が行われていたが，クラウド型音声認識では音声データをサーバに送り，サーバ上で音声認識を行う．これによる運用上のメリットは非常に大きい．ユーザはプログラムや認識用モデルを端末にダウンロードしたりアップデートしたりすることを強いられずに，既存の音声端末（例えばスマートフォン）から音声認識を気軽に使うことができる．また辞書などのモデルもサーバ側にあるため，システム開発者が随時アップデート可能である．さらには音声データはサーバに送られるため，開発者から見るとデータ収集が極めて容易であり，収集したデータを機械学習による性能向上に使うことができる．上記のパラダイムシフトにより，スマートフォンなどで使える一般ユーザ向けの音声認識アプリが多数登場した．ここ数年（2010年代後半）では，スマートスピーカと呼ばれる家庭用の商品が登場している．日本では，2017年10月にGoogle HomeやLINEのClova WAVEが，11月にはAmazon Echoの販売が開始された．

技術的には，ここ数年は深層学習を用いた技術が席巻している[4]．音声認識の分野では，音声認識率（認識結果に占める正解単語の割合）という明確な評価尺度があり，大規模音声データベースが共有されていることから，深層学習の適用が可能である．端的に言えば，学習データ量が多いほど，またそのデータが実使用に近い環境で得られたデータであるほど，音声認識の性能は高くなる．このため，利用ユーザが増えるとデータが集まり，データが集まると性能が上がるのでさらにユーザが増える，という正のスパイラルが回り始めると，性能向上は加速する．

現状，これらの技術を実現したソフトウェアはほぼ無料で提供されている．スマートフォン上の音声アプリは無料であり，スマートスピーカもほぼハードウェアの価格であるように思われることから，ソフトウェアを販売することで投資を回収するというビジネスモデルであるとは思えない．上述したデータ収集や，自社の技術力の誇示によるブランド価値の向上，また自社の他サービスへの導線とするユーザインタフェースなどとして位置づけられていると思われる．

3．言語内容の理解

言語理解部の処理は，音声認識結果として得られる文字列（単語列）に対して行われる．言語理解結果は，対象とするタスクやドメインに依存して定められる．タスクとは検索や予約などシステムが遂行すべき課題を指し，ドメインとは天気予報やホテル情報などの話題である．つまり言語理解とは，そのドメインにおいて事前に定義されたキーワードや意図理解結果を，入力である単語列から取り出すという処理に相当する．一般的に想定される「人間が言語を理解する」ということとはやや乖離があることに注意が必要である．

具体例として，例えば「週末の大阪の天気を教えて」というユーザ発話を言語理解することを考える．ここでは，「週末」が日時を表すことや，「大阪」が地名を表すこと，さらにはこの発話全体で天気予報を検索しようとしていることを理解する必要がある．この結果，AskWeather（Place＝“Osaka”，Date＝“weekend”）といった内容を取り出すことに相当する．

上で述べた，入力文全体からその発話の意図を同定する問題は，多クラス識別問題として定式化でき，機械学習が用いてラベル（ここでは「検索」や「取消」など発話の意図）を予測するのが一般的である．また，単語列の中から，日時や場所に該当する単語の範囲を同定するのも，系列ラベリング問題として定式化できる．これらはいずれも教師あり学習であり，大量の学習データとそれに対する正解ラベルを用意することで同定可能である．またいずれの問題に対しても，近年では深層学習の利用が進んでいる．

なお，上記の「週末」という表現を例えばホテルの予約に使う場合には，これを具体的な日付に変換する必要がある．この際，まず週末が土曜と日曜を指すことや，さらにはこの状況では「前の週末」ではなく「次の週末」を意味することは，人間にとっては常識であるが，コンピュータには明示的に指示する必要がある．あまりにも常識である知識はデータの中には明示的に現れないことが多く，深層学習を含む機械学習では解決できないため，実用的にはルールで補う部分も依然必要である．

ここまでで述べた入力理解部における処理は，入力音声に含まれる言語情報に基づくものであり，昨今のスマートフォン上の音声応答アプリでも行われている．なお，これらのシステムは基本的に「一つの質問に対してシステムが応答する」という一問一答型システムであるため，図1に示した対話管理は行わず，入力理解結果に応じて用意した応答文を，音声合成部に直接渡している．このような構成が採られる理由は，音声認識誤りを含む入力理解の誤りの可能性を保持したうえで，誤解を解消するのが困難なためである [5]．一問一答ではなく，それまでの履歴や文脈を考慮して，長く続く対話を実現する音声対話システムは，未だ研究途上にある．

4．言語内容以外の理解

音声対話システムが対話を円滑に進めるには，音声認識結果に含まれる言語内容に加えて，音声対話に特有な，言語内容以外の要素も理解する必要がある．具体的には，発話の時間構造，つまりタイミングの管理が挙げられる．また実空間内での対話では，ユーザが誰に向けて話したのかを理解することも必要である．

対話では，基本的に参加者は交互に話すことが想定される．すなわちある参加者が話し，それが終了した後に他の参加者が話し始める．これは話者交替，つまりターンテイキング（turn taking）と呼ばれる．単信式のトランシーバ（同時に送信か受信のどちらかしかできない）での対話では，発話の終わりに「どうぞ」と言うなどして発話の終了を明示することで，ターンテイキングが実現される．

しかしながら，自然な音声対話では，参加者は必ずしも交互に話してはいない．つまり，同時に話し始めたり，話している最中に他の参加者が話し始めたりといった現象が頻繁に生じる．対話システムにおいても，ユーザにトランシーバのような不自然な対話を強いるのではない場合には，このような現象がしばしば生じる．

以降ではまずこのようなターンテイキングの基礎となる事項を述べた後，それを円滑にするための処理について述べる．さらに，受話者推定についても述べる．

4．1　発話区間の認定

発話は，音声対話を理解するうえで基本的な単位であるが，常に明確に規定できるわけではない．単純には，「400 ミリ秒以上の無音区間で区切られた音声区間」という単位が考えられるが，例えば促音（「っ」で表記）は物理的には無音であり，この長さが 400 ミリ秒を超えることもよくある．したがって，単純に一定長の無音区間で区切っても必ずしも適切な単位とはならない．

ここでまず発話が過剰に細かく区切られてしまう例を示す．例えば，新幹線の予約窓口で，「帰り・・・は自由席でお願いします」という発話があったとしよう．話し手は考えながら話していて言い淀んでおり，この「・・・」は無音区間を表す．この時，単純に無音区間を用いて発話を認定すると，「帰り」と「は自由席でお願いします」という2つの発話の断片が得られるが，これらは「帰りは自由席でお願いします」という発話として理解するのが妥当であろう．

一方で，無音区間がなくても内容から2発話に区切るべき場合もある．例えば，「新横浜からどこまで行きますか？」という質問に対して，「違います新大阪です」と無音区間なしに話したとしよう．これは「違います」という否定と，「新大阪（から）です」という出発地を示す2つの発話がなされたとして，「新横浜ではなく新大阪からどこかに行く」と言っていると理解するのが妥当である．しかしこれを一発話と捉えてキーワードだけを拾うと，「新横浜から新大阪まで」のように誤って理解してしまうことになる．

音声認識では，まず発話区間検出（Voice Activity Detection; VAD）と呼ばれる処理で，発話に相当する区間が切り出される [6]．一般的な発話区間検出では，音声信号から得られる特徴である振幅や零交差数に基づき，ボトムアップに音声区間と無音区間が判別される．零交差数とは，単位時間内に音声波形が振幅0の軸を交差する回数で定義される．これらを用いて無音とされた区間が，事前に設定したしきい値より長く続いたときに，ユーザの発話が終了したとみなされる．なおシステムに素早く応答させるには，このしきい値を小さく設定するとよい．一方で，これを小さく設定した場合，発話内に含まれる短い無音区間で，誤って発話が終了したとみなさ

れ，システムが誤って話し始めてしまうリスクが高まる．つまりシステムの素早い応答とシステムが誤って話し始めてしまうリスクは，トレードオフの関係にある．

通常，VADにより検出された区間に対して音声認識が行われるため，VADが誤っていた場合は，正しい音声認識結果が得られない場合も多い．また雑音環境下では，前述の振幅と零交差数を用いたVADの性能は劣化することが知られている．

このVAD結果を含め，音声対話における発話という単位は，以下の3つの側面から認定できる．単純なシステムではこの3つの単位が一致していることが暗に想定され，それが一発話とされるが，実際には常に一致するわけではない．

(1) 音響信号に基づく認定：
VADにより得られる発話区間がこれに相当する．

(2) 対話行為に基づく認定：
発話の言語内容が一つの対話行為をなすかどうかである．前述の「違います新大阪です」の例では，「違います」という否定と，「新大阪です」という情報提供という2つの対話行為からなると考えられる．一方で，「帰り・・・は自由席でお願いします」も，「・・・」で表した言い淀みの前後を繋ぐことで，依頼という一つの対話行為と考えることができる．つまり言語内容により単位を区切るという考え方である．

(3) ターンに基づく認定：
ターン，つまり発話の順番が交互に取られることを仮定した場合である．例えば話者Aが話し終わってから，次に再度話者Aが話し始めるまでの区間は，話者Bの発話であったと考える．つまり，やりとりをもとに単位を規定するという考え方である．ただしこれが適切でない例を以下に示す．

A1: 席の予約はどうされますか？
B1: 指定席の窓側
A2: はい，窓側
B2: を予約します

この例では，対話行為の観点から見ると，B1とB2を繋げて一発話とするのが適当だと考えられる．しかし，相槌のような役割を持つ部分A2が挿入されていることで，ターンと対話行為の単位が一致していない．

上述した3つの単位が一致するのは，相手の発話が完全に終わるまで話し始めず，一つの対話行為のみを含めて，自分の発話を必ず一息で（無音区間なしに）話す場合である．一方で，このような発話しか許容しないのは自然ではなく，また，システムが誤ってターンを取ってしまう場合も現実にしばしば起こる．したがって，この前提が満たされない場合にでも，システムが適切に動作できることが望ましい．

4．2　話者交替を円滑にするための処理

システム応答の大きな遅延や話者間の発話の重なりがない状態を円滑なターンテイキングとすると，システムはユーザが話し終わればすぐにそれを検出し，遅滞なく発話を開始する必要がある．一方で，ユーザが発話を終了しないうちにシステムが話し始めると発話が重なってしまう．したがって，ユーザが話し終わったかどうかを正確に検出する必要がある．この問題は，発話の終端検出（endpoint detection）と呼ばれる[1]．既述のように，素早く終端を検出しようとした場合には誤検出が増えるというトレードオフの関係にある．

我々のグループでは，終端検出自体の高精度化ではなく，終端検出の誤りを事後的に検出して，その誤りに起因する問題を修復するという新たなアプローチで研究を行った[7]．終端検出が正しく行われなかった場合，不適切なタイミングでシステムが応答するだけでなく，本来の発話の一部区間のみに対する音声認識結果が得られるため，それに基づく応答内容も誤りとなる．これらの問題に対して，まずシステム発話の冒頭でユーザ発話の開始を検出した場合には，ただちに終端検出が誤っていたとみなしてシステム発話を停止する．これによりターンテイキングを修復し，さらに誤分割された発話区間を連結して再度音声認識を行うことで，応答内容も修復する．図2に例を示す．

話者交替については，現状の一問一答のやりとり

[1] この検出は，話者交替が起こり得る時間的な場所を推定する移行適格場所（Transition Relevance Place; TRP）の推定とも関連し，相づちを打ってもよいタイミングを推定する問題ではTRPという用語がよく用いられる．

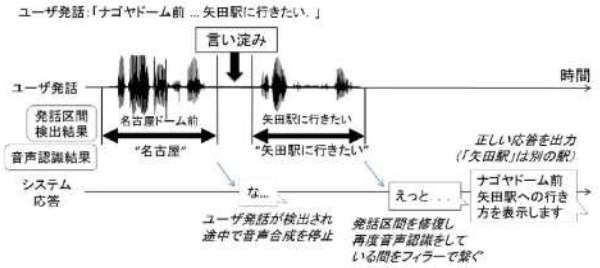

図2　発話の誤分割の事後的修復の例 [7]

ではまだ重要視されていないが，より自然な音声対話を実現する際には重要となる．スマートフォン上のアプリでは起動ボタンにより発話の始端の候補区間を限定したり，スマートスピーカでは発話の冒頭にマジックワード（例：「アレクサ,」「オーケーグーグル,」など）を言わせたりすることで，発話区間検出の一部（発話の始端検出）の問題を緩和している．しかし今後，人型ロボットなどボタンを持たないインタフェースでの音声対話や，マジックワードなしでの音声対話を実現するには，本節で述べたような技術が重要となる．また，特に一般のユーザがシステムを使用する場合，発話に言い淀みが生じるのは不可避であるため，これを許容したシステム設計が望ましい．さらに音声対話システムをユーザインタフェースとして見た場合，ユーザの入力に対して素早く反応するのは必要条件と考えられるため，終端検出の高度化やその誤りの修復機構も，今後重要性が増すと考えられる．

4．3　受話者推定

音声信号が入力されたとしても，それが環境雑音である場合や，自分以外に話しかけられた音声（例えば独り言）である場合には，システムは応答すべきではない．つまり，ユーザがシステムに向けて話した場合にのみシステムは応答すべきであり，それ以外の場合にシステムが勝手に話し始めると煩わしい．したがって発話の受け手となる受話者を推定し

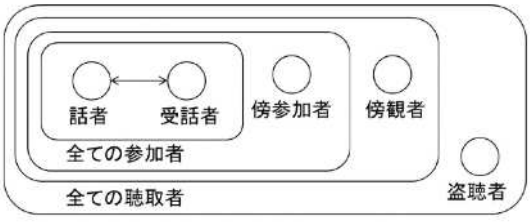

図3　多人数対話における参加構造 [8]

理解する必要がある．

実空間中での，複数のユーザが参加する対話（複数人対話；multi-party dialogue）では，受話者推定の必要性は顕著となる．つまり，あるユーザが話した発話が誰に向けられたものであるかの判定や，システムがそれに応答する必要があるかどうかの判定が常に必要となる．複数人対話において，誰が誰に向けて話しているかという情報は，対話の参加構造を理解するうえで重要な手がかりである．複数人対話には図3で表される参加構造 [8] があり，どのユーザがどの程度対話に参加しているのかを表すエンゲージメント（engagement）の推定は，対話システムが応答をするか否かの決定やその内容の決定において重要な課題である．このエンゲージメントを適切に推定できれば，対話に参加していない人に話しかけてしまうのを防いだり，対話ロボットが自身の視線や体の向きを適切に制御したりすることが可能になる．

つまりこれらの技術は，今後，対話ロボットが実社会で人との共生を目指すうえでは不可欠である．一方で現状のスマートフォン上のアプリやスマートスピーカでは，タスクが一問一答であることからこれらの技術はあまり必要とされていない．複数の参加者が存在する状況において対話を実現するには，受話者推定技術の重要性が増すと考えられる．

5．ロボット対話への展開

スマートフォン上のアプリとの限られた一問一答ではなく，実空間中に存在する物理的なロボットとの間で比較的自由な音声対話を実現させるには，さらにいくつかの課題がある．

5．1　ロボットでの音声認識や音響信号処理

1点目は実空間中に存在するロボットを使うことに伴う課題である．

まず，スマートフォン上のアプリのようにユーザの口元にあるマイク（接話型マイク）を使う場合と，ロボット自身に備え付けられたマイクを使う場合では，音声認識の難しさが大きく異なる．音源である口の位置と，それを受けるマイクの位置との距離が大きくなることで，SN比が下がり，さらには周辺雑音や残響の影響を受けることから音声認識性能が劣化する．

次に，音源定位も必要となる．ユーザは必ずしもロボットの正面から話しかけるとは限らないため，話しかけたユーザの方向を向いて応答することは，顔のあるコミュニケーションロボットにおいて必須の機能である．しかし，特に小型のロボット上でこれを実現するには，従来の自由空間中でのマイクロホンアレーを使った音源定位と比べて，1. マイク間の距離が小さい，2. マイクの近くにモータ用のファンなど雑音音源がある，3. 頭部の外装や肩など音響伝達特性に影響を与える障害物が近くにある，4. 各マイクの周波数特性も必ずしも同じではない，といった状況から問題は難化する．この問題に対して，我々はdeep neural network（DNN）を使った音源定位にも取り組んでいる [9]．なお認識すべき音声に他の音が混入している場合には，音声認識の前に音源分離を行う必要もある．

さらに4.3節で述べたロボットの応答義務の推定も重要な課題である．我々は，発話者の顔の向きや体の動き，直前のロボットの発話行為などを考慮したうえで，機械学習により応答義務の有無を推定する研究も行なっている [10]．

5．2　ユーザを含んだ系で考える対話システム

2点目は実ユーザの問題である．研究室環境で統制されたユーザとは異なり，実際のユーザはシステム開発者の想定を上回る多様な挙動をする．一方で，対話は共同行為，つまり，相手と互いに調整し合いながら行う行為である [8]．すなわち，対話の参加者は，相手とは独立に行動するのではなく，相手に応じたふるまいをする．

対話システムという観点から考えると，システムが受け身にあらゆるユーザから入力を処理できるようにするという枠組みではなく，ユーザを含んだ系として対話システムを考える必要があると考える．つまり，システムがうまく人間の助けを得られるような対話システムの設計を考える必要がある．

この一環として，対話を通じて知識を獲得する対話システムの研究も行なっている [11]．対話システムにおいて未知語の問題は避けられない．つまり，どんなにシステム知識を大規模化したとしても，そこにはない語彙が必ず現れる．特に，家庭用の対話システムが普及した場合を考えると，ローカルな地名や知り合いの人名など，一般的ではない語彙が対話において重要な役割を占めると考えられる．そのような語彙を，各家庭向けに人手でカスタマイズするのはコスト的に現実的ではない．このことからも，対話を通じて対象ドメインの知識を獲得する技術，ひいてはユーザの助けをうまく得ながら対話を通じてシステムを賢くする技術も，今後必要性が増すと考えられる．

参考文献

[1] 中野 幹生，駒谷 和範，船越 孝太郎，中野 有紀子，“対話システム”コロナ社，2015.

[2] Toshiyuki Sakai and Shuji Doshita, “An Automatic Recognition System of Speech Sounds,” Studia phonologica, Vol.2, 1962, pp.83-95. http://hdl.handle.net/2433/52632

[3]「特集　人に近づく音声インタフェース　第1部＜市場動向編＞ クラウドの利用で飛躍 用途はスマホ以外にも拡大」日経エレクトロニクス 2012年12月24日号，2014. http://techon.nikkeibp.co.jp/article/HONSHI/20121221/257654/

[4] 河原　達也，“音声認識技術の展開”信学技報 PRMU2015-111, pp.111-116, 2015.

[5] 河原　達也，“音声対話システムの進化と淘汰ー歴史と最近の技術動向ー”人工知能学会誌，Vol.28, No.1, pp.45-51, 2013.

[6] 石塚 健太郎，藤本 雅清，中谷 智広，“音声区間検出技術の最近の研究動向”日本音響学会誌，65巻，10号，pp.537-543, Oct. 2009.

[7] Kazunori Komatani, Naoki Hotta, Satoshi Sato, Mikio Nakano, “User-Adaptive A Posteriori Restoration for Incorrectly Segmented Utterances in Spoken Dialogue Systems,” Dialogue and Discourse, Vol.8, No.2, pp.206-224, 2017.

[8] Herbert Clark, “Using Language,” Cambridge University Press, 1996.

[9] Ryu Takeda, Kazunori Komatani, “Unsupervised Adaptation of Deep Neural Networks for Sound Source Localization using Entropy Minimization,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),

pp.2217-2221, 2017.

[10] Takaaki Sugiyama, Kotaro Funakoshi, Mikio Nakano, Kazunori Komatani, “Estimating Response Obligation in Multi-Party Human-Robot Dialogues,” IEEE-RAS International Conference on Humanoid Robots (Humanoids 2015), pp.166–172, 2015.

[11] Kohei Ono, Ryu Takeda, Eric Nichols, Mikio Nakano, Kazunori Komatani, “Lexical Acquisition through Implicit Confirmations over Multiple Dialogues,” 18th Annual SIGDIAL Meeting on Discourse and Dialogue, pp.50-59, 2017.