メディア教育研究　第９巻　第１号
Journal of Multimedia Education Research 2012, Vol.9, No.1, S29－S41

特集（招待論文）

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―

中野　幹生[1)]

本論文では実用的な対話ロボットの構築における技術課題を述べ，対話ロボットの知能モデルを提示する。まず，対話ロボットの価値を議論し，対話サービスロボットが有望であることを示す。そして知能ロボットと対話システムのモデルを統合するための拡張性の高い対話行動制御フレームワークHRIMEについて述べる。また，個別の技術課題と，現在までの取り組みを述べ，今後対話ロボット研究が進展していくために解決すべき課題を議論する。

キーワード
対話ロボット，知能ロボット，音声対話システム，マルチモーダルインタラクション

1. はじめに

本論文では，音声認識技術の有用なアプリケーションとして考えられている対話ロボットの技術について述べる。対話ロボットとは，人間と音声言語で対話しながら情報を授受し，指示されたタスクをこなすロボットのことである。ロボットの定義は様々であるが，本稿では，物理的な実体を持ち，動作する機械を指す。すなわちソフトウエアエージェントは含めない。人間が操作しなくても動作するロボットを自律ロボットと呼び，その中でも，認識やプランニングなどの高次の知的処理が含まれているものを知能ロボットと呼ぶが，対話ロボットは知能ロボットに属する。

知能を持ったロボットが，人間と対話することで人間の役に立つことは，ロボット研究者のみならず多くの人の夢であった。SFに登場するロボットには，人間と対話ができるものが多い。研究者もロボットに対話機能を持たせることを試みてきた。白井ら（白井・小林・岩田・深沢，1985）は，早稲田大学で開発された鍵盤楽器演奏ロボットWABOT-2に音声認識・音声合成を組み込み，人間の声による命令をロボットに理解し応答させることを可能にした。以来，様々な対話ロボットが開発されてきている（石黒・宮下・神田，2005；Bohus, Horvitz, Kanda, Mutlu, & Raux, 2010）。

対話ロボットは音声認識技術の有力な応用先であるが，様々な技術の集合体である対話ロボットのソフトウエア・ハードウエア全体から見ると，音声認識はその一部の要素技術に過ぎない。しかしながら，必須の技術であることは間違いない。これは自動車全体とステアリングのような関係である。人がステアリングを操作することで，クルマが方向を変える。しかし，ステアリングが唯一の操作系ではない（アクセルやブレーキもある）し，駆動源でもない。同様に，対話ロボットはそれ自体が自律的に動作するものでありながら，音声やその他のユーザ入力に応じて動作を変更する。

本論文では，音声認識を含む対話ロボットの全体の技術を概観した上で，我々が提案している対話ロボット構築フレームワークを紹介する。さらに，対話ロボット研究の今後の課題や将来の方向性を述べる。

2. 知能ロボットと対話機能

2.1　知能ロボットの種類

知能ロボットは，一般的に，与えられたタスクを達成するための行動を計画する行動計画部と，外界の状況を認識する状況認識部，ひとつひとつの行動を実行する行動実行部からなり，これらが相互作用しながらタスクを達成する。これらのモジュールはソフトウエアで実装されており，センサの入力を受け取ったり，ハードウエアを制御したりする。

知能ロボットにはいくつかの種類がある。産業用ロボットは，工場などで組み立て作業などを行うロボットである。対話機能を求められることはほとんどない。サービスロボットは，運搬，医療，警備などのサービスを人間に代わって行って行う。パーソナルロボットは，家庭などで動作するロボットである。家事や掃除などのサービスを行ったり，話し相手になったり，人間を楽しませたりするエンタテイメントの機能も期待されている。コミュニケーションロボットは，サービスは行わず，人間

1) ㈱ホンダ・リサーチ・インスティチュート・ジャパン

とコミュニケーションをすることを主なタスクとするロボットである。これらの分類は，主なタスクや働く場所などの複数の視点によるもので，厳密なものではなく，重複もある。

2.2　ロボットにおける対話機能の価値

これらの知能ロボットが人間と対話することが必要な場面は，次のようなものが考えられる。

- ●人間のロボットに対するタスク指示や情報要求を理解するとき
- ●ロボットが人間にタスクの実行終了，失敗，途中状態などを報告するとき
- ●人間が要求した情報をロボットが人間に伝達するとき
- ●ロボットが人間にアドバイスなどを行うとき（予定の時間になるとリマインダを出すなど）
- ●時間つぶしや，単に楽しむために会話を行うとき

これらの場面でロボットが対話をすることは，一般に有用だと考えられている。しかしながら，対話機能をもつためには，そのためのハードウエア・ソフトウエアを備えている必要があるため，それらの実装コストと，得られる有用性のトレードオフを考える必要がある。

ここで，対話機能の価値とコストを議論するときに，いくつかの機能が混同されがちであることを述べておきたい。それは，音声認識機能，言語理解機能，対話機能である。

音声認識機能があれば，音声で指示を伝えられる。音声が使えなければ，タッチパネル・リモコン・キーボード等を用いて指示を行う必要がある。音声が使えれば，これらのデバイスの使い方を覚える必要がない。また，メニューからコマンドを選ぶようなインタフェースの場合，コマンドの数が多いときに，メニューが複雑になるという問題があるが，音声が使えればこの問題は解消される。リモコンは紛失しやすいが，音声が使えれば探さなくても良い。ただし，少数の音声コマンドが使えるだけの場合，音声コマンドをおぼえなくてはならない。また音声認識はリモコンやキーボードに比べて確実性が低い。また，マイク，音声認識用のソフトウエア，CPUパワー・メモリが必要になり，システムが複雑になる。

言語理解は，比較的自由な文型で発話しても，どのような指示や要求なのかを理解する技術である。具体的には，単語列から，機械が理解できる意味の表現（コマンドもこれに含まれる）へのマッピングを行う技術である。言語理解機能があれば，コマンドを覚えなくても理解してもらえる。しかしながら，コマンドを確実に覚えたときに比べると理解の精度は低くなる。特別なソフトウエアが必要になるのは音声認識と同じである。ちなみに，言語理解は必ずしも音声認識を前提としない。キーボード入力から言語理解を行うようなインタフェースも考えられる。

音声認識や言語理解が，１度の入力からコマンドや意図を推定するのに対し，ユーザと複数回やりとりしながら理解を進めて行く機能を対話機能と呼ぶ。対話機能によって，音声認識や言語理解に失敗したときや，言語理解の結果が曖昧なときに，言い直すことができるようになる。また，複雑なコマンドを，一度の発話で言わなくても，複数の発話に分割して伝えることができるようになる。ただし，対話機能の実装は複雑で，一般的には専門家による作業が必要となる。

2.3　対話相手としてのロボットの価値

対話ロボットの価値を論じるときに，ロボットを用いない対話システムに比べ，対話ロボットの方がユーザにとって話しやすいと主張されることがある。このとき，ロボットは何か仕事をするわけではなく，３次元の実体として存在し，ユーザの発話に反応してジェスチャーをしたり，特定の部位を光らせたりすることが仮定されている。

しかしながら，同様の効果はディスプレイ上のバーチャルエージェントでも期待できるかもしれない。バーチャルエージェントは，実ロボットに比べて低いコストで作成でき，また制御も容易である。単に話しやすいというだけで，コストに見合う価値がロボットにあるかどうかを見極める必要がある。

ロボットとバーチャルエージェントの違いの一つとして，ジェスチャーが３次元空間上で行われることがあげられる。たとえば，道順を教える際に，ロボットが実際の道を指し示す方が，バーチャルエージェントが仮想空間内の道を指し示すより，人間にとってわかりやすいと考えられる。また，ロボットは移動することができるため，人間と適切な位置関係をとることができる。人間との距離や，人間に対する体の向きなどが違えば，人間のロボットに対する話しやすさも変わってくることが期待できる。

2.4　対話サービスロボット

以上のような状況から，われわれはサービスロボットに対話機能がついた対話サービスロボットが有用だと考えている。様々なサービス，特に，運搬のように移動を伴うサービスは，ロボット以外では代替できない。したがって，ロボットとしての価値が高く，今後も積極的に研究が続けられると考えている。そのようなサービスロボットが家庭やオフィスに導入されたとき，特別なデバイスを必要としない対話インタフェースは，ロボットに指示を与えたり情報を受け取ったりするためには有効な手段だと考えている。

3. 対話ロボットの構成と要素技術

本節では，対話ロボットのモジュール構成と，各モジュールで用いられる要素技術を説明する。

3.1 知能ロボット

ここでは，知能ロボットの概要を説明する。詳細な解説は教科書（浅田・國吉，2005）に譲り，対話ロボットに関連する事項を簡単に説明する。

ロボットは一般的にセンサ群とアクチュエータ群を持つ。その間に知能処理部（呼ばれ方は様々である）がある。知能処理部は，センサの出力を解釈し，それをもとに内部状態を変更する。内部状態は，ロボット自身の状態の推定結果，外界の認識結果，実行すべきタスクやその実行状況などを保持しているデータである。

知能処理部では，内部状態に基づき，実行すべき行動を決め，それに基づいてアクチュエータを制御する。実際に計画通り行動が行えるかどうかは，ダイナミックに変化しうる外界やロボット自身の状況に依存するので，行動を起こそうとした結果どのような状態になったかをモニタしなくてはならない。

知能処理部はソフトウエアであり，そのアーキテクチャが重要である。アーキテクチャは，各モジュールがどのような処理をし，どのようなモジュール間通信を行うかを規定する。初期の知能ロボットでは，次のようなモジュールが順次駆動されるアーキテクチャであった（図１）。

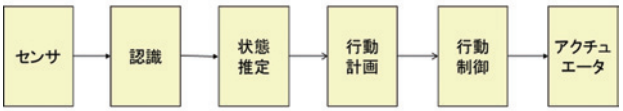

図１　シーケンシャルアーキテクチャ

行動計画部は，どのような行動系列を実行すれば，現在の状態から目標状態に変化するかを計算する。このとき，環境が複雑になればなるほど行動計画時に考慮すべき要因が増えてしまい，現実的には計算が不可能になってしまう（これは人工知能でフレーム問題と呼ばれている）。障害物を発見して衝突回避をしようとするときなど，処理が遅れてしまう。

この問題に対処すべく様々なアーキテクチャが提案されてきたが（Brooks, 1986; Kanda, Ishiguro, Imai, Ono, & Mase, 2002; Hoshino, Takagi, Diprofio, & Fujita, 2004），ポイントはモジュール化と階層化である。モジュール化は，単純な処理を行う複数のモジュールを組み合わせることで，全体として複雑な処理も行えるようにすることである。階層化は，衝突回避のように，センサ出力とアクチュエータ制御がダイレクトに結びついている反射行動モジュールを下位に配置し，タスクの遂行の計画を行うような熟考モジュールを上位に配置する。上位のモジュールが下位のモジュールを制御することで，反射行動と熟考を組み合わせた処理ができる（図２）。

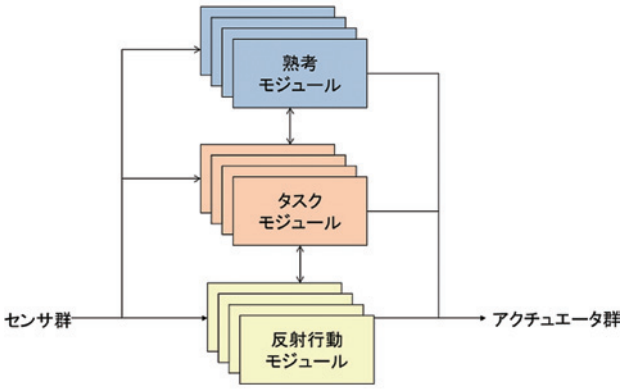

図２　階層化とモジュール化に基づく知能ロボットのアーキテクチャ

3.2 対話システムの構成

ここでは，ロボットから離れ，一般的な対話システムの技術を説明する。しかしながら，対話システムも知能ロボットと同じ人工知能システムであり，その構成は似ている（図３）。

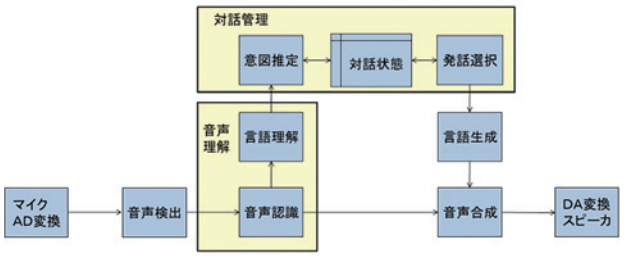

図３　対話システムの構成

センサに相当するものが，マイクとマイクから入ってきた信号をAD変換するAD変換器である。また，アクチュエータに相当するものは，DA変換器とスピーカーである。その間に知能処理部に相当するモジュールがある。

音声検出部は，マイク入力から入ってきた信号をAD変換し，雑音除去などの処理を行った上で，人間の声が含まれている時間区間を音声区間として検出する。音声理解部は，送られてきた音声区間の音声を意味表現に変換する。一般に，音声理解は音声認識と言語理解の組み合わせである。意味表現とは，対話管理部の対話状態を更新するためのコマンドであり，記号表現を用いたデータ構造を持つ。このデータ構造は，一般に対話行為（dialog act）と呼ばれる。意図推定部は，ユーザの意図を含む，対話状態と呼ばれる内部データを，対話行為に基づいて更新する。行動生成部は，対話状態に基づきどのよ

うな行動を行うかを決定する。行動は対話行為で表現される。意図推定部と行動生成部および対話状態を合わせて対話管理部と呼ぶ。対話行為は言語生成部で文字列に変換され，音声合成部で音声波形に変換される。

これはロボットのシーケンシャルなアーキテクチャに対応する。ロボットアーキテクチャでモジュール化や階層化が進んだように，対話システムアーキテクチャでもモジュール化と階層化が試みられている。

モジュール化に相当するのは，分散型マルチドメイン対話システムアーキテクチャである。特急券の予約を受け付ける電話対話システムなどのように，多くの音声対話システムが単一のドメインの対話を行うシステムであるが，音声対話システムの適用範囲が広がるにしたがって，複数のドメインの対話を行うシステムが増えてきている。たとえば，ホテル，バス，電車，天気の情報を提供するシステムが開発されている（Lin, Wang, & Lee, 1999）。このようなマルチドメインシステムでは，分散型アーキテクチャが用いられている（図４）。このアーキテクチャでは，対話管理を行う対話管理部がドメイン毎に用意され，入力発話の理解結果に応じてどのドメインの対話管理部を駆動するかが決まる。ロボットの対話インタフェースを考えた場合，基本的にロボットは多くのドメインのタスクを行うことが期待されているので，その対話インタフェースも必然的にマルチドメインシステムとなる。

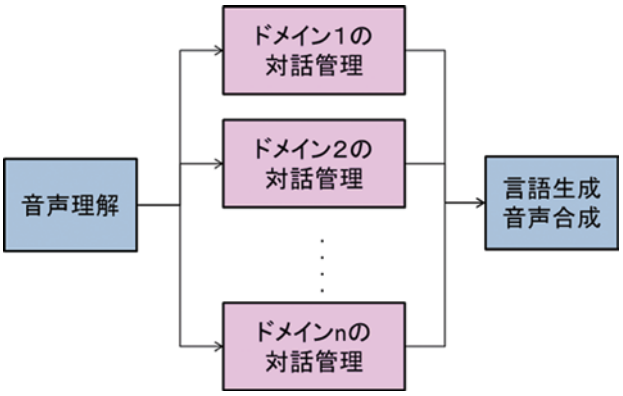

図４　分散型マルチドメインシステムアーキテクチャ

階層化に相当するものとして，対話管理において反射行動と熟考に基づく行動の組み合わせが試みられている（Traum & Allen, 1994; Hasegawa, Nakano, & Kato, 1997）。また，システムの発話中にユーザの発話が開始された場合（バージインと呼ぶ）に，音声検出部がそれを発見して，音声合成部に直接信号を送り，システム発話を止める仕組みなども階層化である。

3.3　対話技術

こでは対話管理部の主要技術を説明する。対話管理部は対話状態をどのような形式で表現するかによってその構成が大きく変わってくる。一問一答式の対話の場合，対話状態はユーザ発話の分類結果だけを保持している（西村・西原・鶴身・李・猿渡・鹿野，2004）。その分類結果に応じて応答を選択するルールがあり，応答が出力される。このような方式では，文脈情報を利用できないため，ユーザが複数の発話で要求を相手に伝えたり，システムから確認したりすることができない。そこで，より複雑な対話状態を保持する。90年代初頭までの対話システムの研究では，プラン認識結果や発話計画の結果など非常にリッチな情報を保持することで，柔軟で協調的な対話が行えることが示された。しかし，2000年前後から音声認識と結び付けた音声対話システムの研究が盛んになり，音声認識誤りがある程度の頻度で起こることを考慮するとあまり複雑な構造は用いても意味がないことが判明し，フレームなどのより単純な表現方法が用いられるようになった。図５に特急券予約システムのフレームの例を示す。フレーム中の一つ一つの情報をスロットと呼ぶ。対話が進むにしたがってスロットの情報が埋まっていく。図５には対話の途中の状態を示している。

乗車駅	新宿
下車駅	—
日付	10月2日
発車時間帯	—

図５　フレームによる対話状態表現

対話管理で重要な技術がエラーハンドリングである。エラーハンドリングとは，音声理解誤りなどで生じた問題を対話で修正するプロセスである。

エラーハンドリングの方法の一つに，理解結果を確認する方法がある。音声理解の結果，出発駅スロットの値が「新宿」になれば，「新宿からですか？」のように確認を求め，ユーザが肯定的な応答をすれば，確認済みとする。確認済みかどうかの情報を対話状態中に保持する。このようなプロセスを，Clarkが提唱したコモングラウンディングの理論（Clark, 1996）に基づき，グラウンディングと呼ぶ。

毎回確認していると非効率なため，なるべく確認の回数を減らすことが試みられている。正しい確率が高い理解結果は確認しないという方法が用いられている（駒谷・河原，2002）。理解結果の正しさ（確信度と呼ぶ）の推定には，音声認識・理解・文脈の情報が有用であることが示されている（Higashinaka, Sudoh, & Nakano, 2006）。確信度は，確認をするかしないかだけではなく，理解結果の棄却（理解しなかったことにする）にも用いられている。このような場合を，理解結果が間違ってい

るmisunderstandingと区別してnon-understandingと呼ぶ。misunderstandingやnon-understandingに対してどのような戦略で確認や問い返しを行えばよいかが研究されている（Bohus & Rudnicky, 2005）。

音声理解結果は曖昧であることがある。音声認識や言語理解のモデルが不十分である場合もあれば，人間が聞いても発話の意味が曖昧な場合もある。そこで，音声認識や音声理解が複数の候補を出力し，その中から文脈に会うものを選ぶ方法が用いられる。対話の途中では，その時点までのユーザの意図推定結果は曖昧にならざるを得ないので，複数の推定結果を確率的に保持しておき，対話が進むにしたがって，確率を変化させていく方法が提案されている（Paek & Horvitz, 2000; Higashinaka, Nakano, & Aikawa, 2003; Williams & Young, 2007; Ma, Raux, Ramachandran, & Gupta, 2012）これをBelief Trackingと呼ぶ。

3.4　ロボットの制御技術と対話技術との統合

対話ロボットの構築に向けて，ロボットの知能処理部と対話システムを結びつけることが試みられてきた（Asoh, Motomura, Asano, Hara, Hayamizu, Itou, Kurita, Matsui, Vlassis, Bunschoten, & Kröse, 2001; Makihara, Takizawa, Ninokata, Shirai, Miura, & Shimada, 2002; Yoshimi, Matsuhira, Suzuki, Yamamoto, Ozaki, Hirokawa, & Ogawa, 2004; Zobel, Denzler, Heigl, Noth, Paulus, Schmidt, & Stemmer, 2001）。それらのロボットでは，ロボット知能処理部と独立に対話処理部が動作しており，対話と行動を統合して制御することはできない。しかし，対話ロボットがタスクを達成するには，物理行動と対話の組み合わせが必要になる。たとえば，人をある場所に案内する場合，依頼者をその場所まで連れて行き，その場所についたことを対話で知らせる。また，移動などの物理行動中に人が声で停止要求をした場合にはそれに従う必要がある。そのためには，行動と対話を統合した形で，行動計画・発話選択を行う必要がある。

4. 対話行動制御部構築フレームワークHRIME

4.1　HRIMEの概要

前節までにみたように，ロボット知能処理部は対話技術と類似のアーキテクチャを持つ。したがって，音声入力をセンサ入力と考え，発話と物理行動を同一視することでロボットの知能処理部に音声理解・音声生成を組み込める。しかしながら，フレームによる対話管理やドメイン選択などをどのように組み込むかは自明ではない。

我々は，HRIME（HRI Intelligence Platform with Multiple Experts）[2] と呼ぶフレームワークを提案し，対話と物理行動の統合的な制御を行うモジュールの開発に用いている（Nakano, Hasegawa, Funakoshi, Takeuchi, Torii, Nakadai, Kanda, Komatani, Okuno, & Tsujino, 2011）。HRIMEは，ロボット／エージェントの知能システムの中で，記号レベルでの状況理解と行動選択を司るモジュール（対話行動制御部と呼ぶ）を構築するためのフレームワークである。対話行動制御部は，音声・画像認識や他のセンサ解釈モジュールによってシンボル化された入力を受け取り，それを基に人の意図や状況を推測し，適切な行動を選択し，ロボットハードウエア制御や音声合成などの行動実行モジュールに送る。

HRIMEでは，特定の種類のサブタスクに特化した知識と内部状態を持つ，エキスパートと呼ぶモジュールを用いる。これは，マルチドメイン音声対話システムで用いられているドメイン毎の対話管理部を拡張したものである。たとえば，天気予報に関する質問に答えられるロボットであれば，「天気予報に関する質問を理解する」というサブタスクのためのエキスパートや「天気予報を人に伝える」というサブタスクのためのエキスパートを持つ。また，「特定の場所に移動する」という物理行動を行うサブタスクのためのエキスパートなども用いることができる。これらのエキスパートを順次利用することにより，複雑なタスクを遂行することができる。たとえば，ある物を説明するタスクは，その物のところに人を案内して，言葉で説明するという２つのサブタスクを順次遂行することによって行うことができる。

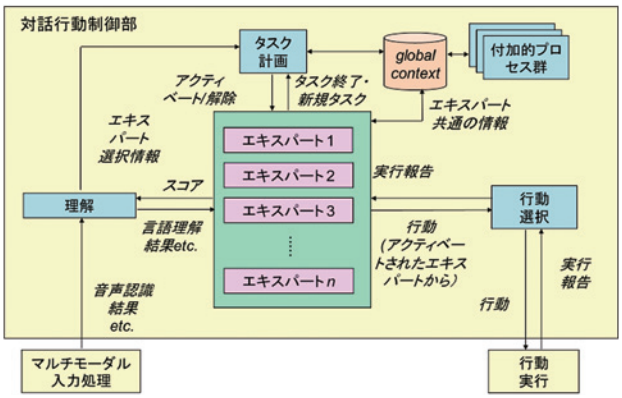

図６　HRIMEのモジュール構成

HRIMEでは，このようなエキスパートを利用して全体のシステムを動作させるためのプロセス群（調整プロセスと呼ぶ）が走っている。HRIMEのモジュール構成を図６に示す。調整プロセスは３つあり，並行動作する。理解プロセスは音声認識結果を言語理解してエキスパートに送信し，最適なエキスパートを選択し，タスク計画プロセスにその情報を送る。行動選択プロセスは，選択されたエキスパートに対し，次の動作の決定を逐次要求する。タスク計画プロセスは，タスクを遂行したり，音声認識結果に反応したりするために，どのエキスパート

[2] 以前の文献ではRIMEまたはRIME-TKと呼んでいた。

をアクティベートし，どのエキスパートのアクティベートを解除するかを決定する。これら３つのプロセスは並列に動作する。

それぞれのエキスパートは内部状態にアクセスするためのメソッドを持っていなければならない。initializeメソッドはエキスパートが作られたときに呼ばれ，内部状態を初期化する。understandメソッドは音声理解結果を受け取った際に理解プロセスから呼び出され，音声認識結果に基づいて情報を更新する。select-actionメソッドは，行動選択プロセスから継続的に呼び出され，発話待ちの状態でなければ，内部状態に基づき行動を１つ出力する。その他に割り込み発話を扱うためのメソッドなどを持っていなくてはならない。understandメソッドの返り値は，その音声理解結果がどのくらいそのエキスパートで処理されるべきかを表すスコアである。理解プロセスは，音声理解結果を，現在アクティベートされているエキスパートおよび新規にアクティベートされる可能性のあるエキスパートに，このunderstandメソッドを用いてる。そして最も高いスコアを返したエキスパートを選択して，その情報をタスク計画部に送る。これは，マルチドメイン音声対話システムにおけるドメイン選択の機能にあたる。

これらのインタフェースを実装しさえすれば，内部で知識や状態をどのような形で保持しているか，また，どのようなアルゴリズムで理解や行動選択を行うかに関わらず，どのようなエキスパートでも導入することができる。対話を行うエキスパートには，先に述べたようなエラーハンドリングやBelief Trackingの機能を持たせることができる。したがって，対話技術とロボット行動制御技術を統合して制御させることが可能になる。

各エキスパートは，global contextと呼ばれるデータ格納部を介して，共通に使える情報（例えば，話題になった事物，ユーザの興味，周囲の状況など）を授受できる。global contextが環境情報をセンサから取り込めるように，上記の３つの調整プロセスに加え，付加的なプロセスを走らせることができる。

4.3　実装

HRIMEは，オブジェクト指向言語Javaで実装されている。HRIMEには，理解，行動選択の各プロセスおよびその他のモジュールを作成するのに必要な抽象クラスがあらかじめ実装してある。また，それらの抽象クラスの実装例をライブラリとして用意している（中野・船越・長谷川・辻野，2008）。

4.4　タスクの例

HRIMEを用いてさまざまなタスクを行うロボット・エージェントシステムを実現している。以下にタスクの例を示す。

天気予報・内線番号などの情報提供：情報要求理解を行うエキスパートと情報提供を行うエキスパートを用いる。

呼んだ人のところへの接近：呼ぶ要求の理解を行うエキスパートと移動を行うエキスパートを用いる。

場所の案内：案内の要求を理解するエキスパート，案内する場所へ移動するエキスパート，情報提供を行うエキスパートを用いる。

非タスク指向対話：ユーザ発話のキーワードを検出して応答する（Nakano, Hoshino, Takeuchi, Hasegawa, Torii, Nakadai, Kato, & Tsujino, 2006）。

説明や案内：説明や案内の開始要求の理解は要求理解を行うエキスパートとプレゼンテーションを行うエキスパートを用いる。プレゼンテーションコンテンツの記述を容易にするための言語MPML-HR（Multi-modal Presentation Markup Language for Humanoid Robots）（Nishimura, Kushida, Dohi, Ishizuka, Takeuchi, Nakano, & Tsujino, 2007）を用いることができる。

質問応答：あらかじめ用意された質問応答データベースに沿って応答を行うエキスパートを用いる。

図７，図８にインタラクションの例を示す。図８は行動中に割り込み発話があったときの例である。

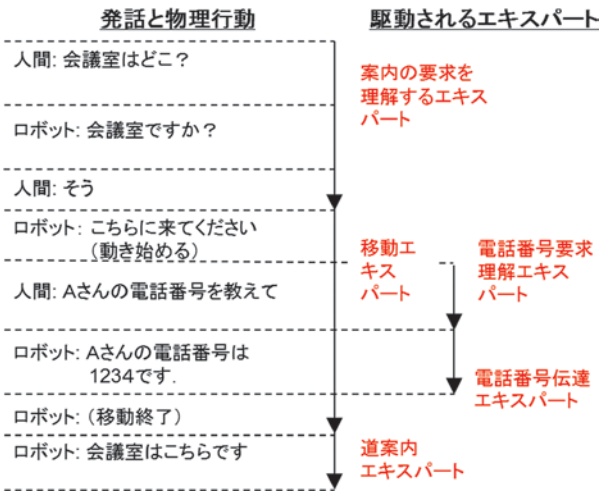

図７　インタラクションの例

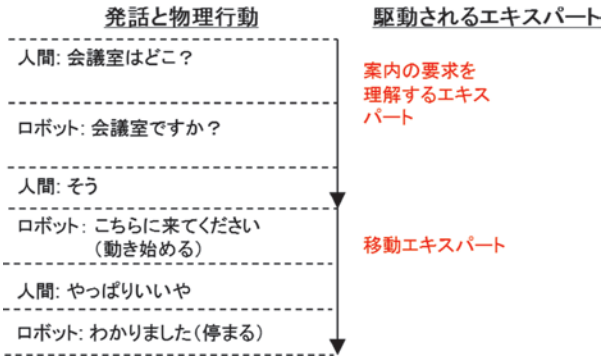

図８　行動への割り込み発話への対処の例

4.5　HRIMEの利点

HRIMEは，共通インタフェースさえ実装すればどのような制御を行うエキスパートも導入できるという点で拡張性が高い。また，対話も物理行動も同様に扱うことができる。各エキスパートの対話戦略は異なっても良いので，音声認識結果中のキーワードに基づいて応答を返す雑談のエキスパートと，フレームに基づく対話管理を行う要求理解エキスパートのように，まったく異なるエキスパートも統合可能である（Nakano, Hoshino, Takeuchi, Hasegawa, Torii, Nakadai, Kato, & Tsujino, 2006）。また，タスク計画部が複数のエキスパートを統合的に制御するため，対話と物理行動を組み合わせてタスクを遂行することが可能である。我々はHRIMEを様々なタスクに適用することで，対話サービスロボットの可能性を探りたいと考えている。

4.6　HRIMEにおける音声認識・音声理解

HRIMEでは，音声認識はマルチモーダル入力処理部で，言語理解は理解部で行う。HRIMEとして，どのような音声認識や言語理解を用いるかは規定していない。ただ，我々の今までの研究で，どのようなエキスパートを用いる場合にどのような音声認識・理解を用いればよいかという指針を得ている。

まず，音声認識は音響モデルと言語モデルを用いるが，音響モデルはタスクによらないので，ここでは言語モデルのことだけを述べ，音響モデルは5.1節で触れる。言語モデルは，一般にタスクドメイン毎に用意する必要がある。実際の発話例やそれに近い文が多く集められれば，それらの例文から統計言語モデルを作り，そうでなければ，有限状態文法による言語モデルを手作業で作る必要がある。対話インタフェースにおける音声認識は，正しい確率が高い認識結果を一つ出すよりも，言語理解部や意図理解部で正解が得られるよう，定型発話からくだけた発話まで様々な発話に対して多くの認識結果の可能性を出力した方が良い。統計言語モデルと有限状態文法を併用することで，様々なタイプの発話に対処できる（勝丸・中野・駒谷・船越・辻野・尾形・奥乃，2010）。言語理解部や意図理解部で利用するための，音声認識結果の確信度を計算するためには，ドメイン非依存の大語彙統計言語モデルを併用し，発話検証スコアを求めると良い。各ドメインの統計言語モデルや有限状態文法言語モデルは並列に用いても良いが，それぞれ一つづつのモデルに統合することもできる。

言語理解は，一般的に，キーフレーズ抽出に基づく手法，構文解析に基づく手法，有限状態変換器（FST）に基づく手法，条件付き確率場（CRF）に基づく手法などがよく使われているが，どれも，手作業または統計処理によって構築したドメイン毎のモデルが必要となる。また，単一のドメインでも複数の手法の組み合わせが有効である（勝丸・中野・駒谷・船越・辻野・尾形・奥乃，2010）。

どのエキスパートがどの言語モデルを用いた音声認識結果とどの言語理解モデルを用いた言語理解結果を用いるかを指定しておくことで，各エキスパート内でドメイン依存の音声理解結果を用いることができる。

エキスパート選択のために，各エキスパートが出力するスコアは音声理解結果と文脈情報を用いて算出する。我々は音声理解誤りにロバストなスコア算出法を提案しているが，詳細は割愛する（Nakano, Sato, Komatani, Matsuyama, Funakoshi, & Okuno, 2011）。

5. 対話ロボットの技術課題

前節までに見たように，対話ロボットのフレームワークができ，様々な対話ロボットを構築することができるようになってきたが，実際に役に立つ対話ロボットを構築するためには，多くの課題が残っている。本節ではそれらの課題を説明し，現在までの研究を紹介する。

5.1　ロボット音声認識

一般に音声認識は，話者の口元に近いマイクを用いることが仮定されている。しかし，ロボットと人間との距離は遠いので，ロボットから離れたデバイスを使ってリモートで音声入力を行うのでなければ，ロボットに取り付けたマイクで音声入力を行う必要がある。

発話者とマイクが離れている時の音声認識を遠隔音声認識（Distant Speech Recognition）という。遠隔音声認識の問題は，周囲の騒音に比べて人間の声のゲインが相対的に小さくなることや，壁等に反射した音（残響音）がマイクに入ることなどである。これらの問題に対処する方法として，指向性の非常に強いマイク（超指向性マイクロホン）を用いる方法と，二つ以上のマイク（マイクロホンアレイ）を用いる方法がある。超指向性マイクロホンを用いるときは，人間がマイクロホンに正対するか，ロボットがマイクロホンを人間に向けなくてはならないという制限が生じる。マイクロホンアレイを用いればその制限はなくなるが，音声がくる方向の推定（音源定位）やノイズ・残響音の除去が必要である（Mochiki, Ogawa, & Kobayashi, 2008; Nakajima, Nakadai, Hasegawa, & Tsujino, 2010）。また，ただの遠隔音声認識と異なり，ロボットの場合，ロボット自身が発する雑音も問題になる（Ince, Nakadai, Rodemann, Tsujino, & Imura, 2011）。また，ノイズがある中で，どの時点からどの時点までが音声か（音声認識の対象とすべきか）という音声区間検出の問題もある。さらにノイズは完全に除去できなかったり，ノイズ除去の結果歪みが生じたりするため，接話マイク用での音声認識のための音響モデルではうまく認識できない場合が多い。そこで，ロボット音声

認識用の音響モデルを構築しなくてはならない。以上のような問題は，個別に解決しても全体のパフォーマンスが上がらないため，まとめて扱うことが重要であり，ロボット聴覚という分野で統合的に扱われている。ロボット聴覚システムを効率的に構築するためのオープンソースソフトウェアHARKが開発されている（中臺・奥乃，2011）。

5.2　マルチパーティ会話

一般に音声対話システムは対話相手が一人であることを仮定しているが，ロボットの場合は必ずしもそうではない。３人以上の会話をマルチパーティ会話と呼ぶが，マルチパーティ会話に従事するロボットの機能も研究テーマの一つである。

複数の人間と対話するときには，誰が誰に話しているのかを認識し，適切なタイミングで適切な相手に向かって話しかける必要がある。そのような機能を持つロボットが研究されている（松坂・東條・小林，2001；Bohus & Horvitz, 2009b）。また，人間同士のコミュニケーションを促進することを目的としたロボットの研究も行われている（藤江・松山・谷山・小林，2012）。

5.3　エンゲージメント・発話交代

人間の音声を検出したとしても，それがロボットに対して話しかけたものとは限らない。独り言のこともあれば，隣にいる人に話しかけた場合もある。そこで，ロボットに対して話しかけたものかどうかを判定する必要がある。音声認識誤りもあり得るので，音声理解結果だけから判定することはできない。我々は音声理解結果と発話の行われた状況との整合性（Zuo, Iwahashi, Funakoshi, Nakano, Taguchi, Matsuda, Sugiura, & Oka, 2010）や発話の行われたタイミングをもとに対ロボット発話かどうかの判定を行う方法の研究を進めている。

また，ロボットが人間に対して発話をするときには，人間がロボットと対話をしても良いと思っている状態になっていなければならない。これをエンゲージメント（engagement）と呼ぶ（Rich, Ponsler, Holroyd, & Sidner, 2010）。エンゲージメントを推定して，いつ話しかけるかを決めることが重要である。エンゲージメントの推定には，視線，双方の発話の連続性などが有効である（Rich et al. 2010）。また，人間がロボットに対してどのように近づくかという情報からエンゲージメントを予測する研究もある（Bohus & Horvitz, 2009a）。音声対話システムでは，ユーザが発話開始時にボタンを押したり，発話中ボタンを押し続けたりすることが仮定されている場合がある（Push-to-Talk）が，ロボット対話では，リモコンを使わない限りボタンは使えない。そこで，いつ人間がロボットに向かって話し始めたのかを検出し，ロボットの発話を中断したり，人間の方に向いたりする必要がある。また，人間が発話を終えてロボットの返答を待っている状態（ターン委譲）を検出し，適切なタイミングで応答する必要がある。このような円滑な発話交代は音声対話システムで研究されてきたが（Raux & Eskenazi, 2009），ロボット対話でも研究されており，ジェスチャーや視線の情報が用いられている（Thomaz & Chao, 2011）。

これらの研究は人間同士の対話のような発話交代の実現を目指しているが，我々は現状の技術ではそれは難しいと考え，即座に応答するよりも，応答が遅れても対話が破綻しないことを目的とした。具体的には，LEDの明滅を用い，ロボットが人間の発話を聞き取ったことを直感的に示すことで，ユーザが不要な繰り返しを行ってユーザとロボットの発話が衝突することを避けることができることを実験的に示した（船越・小林・中野・小松・山田，2011）。我々はこのLEDの点滅やビープ音のような単純でかつ補完的な（音声対話のようなメインのコミュニケーションプロトコルを阻害しない）人工的なモダリティによる内部状態の表出法をASE（Artificial Subtle Expression）と呼んでおり，種々の実験を通してその有効性を確認している（小松・山田・小林・船越・中野，2010）。

5.4　マルチモーダルインタラクション

ロボットは，マイクだけではなく，カメラやその他のセンサを用いてユーザや環境の状態を推定することができる。また，スピーカーから音声を出すだけではなく，ジェスチャー，ディスプレイ，発光デバイスを用いることができる。このような多様なモダリティを用いてコミュニケーションを効率よく円滑に行うことができる。例えば，頭部ジェスチャーと韻律を用いて人間の肯定的または否定的な態度を検出する方法（藤江・江尻・菊池・小林，2005）が提案されている。また，道案内をするロボットが発話だけではなくジェスチャーを行うことで，聞き手にとってわかりやすくなることがわかっている（志和・奥野・神田・今井・石黒・萩田，2012）。

5.5　状況依存言語インタラクションとマルチモーダル学習

特急券予約などの音声対話システムでは，言語の意味は，データベースの要素や属性と結びついている。したがって，すべてが情報の世界で完結する。しかしながら，ロボットが人間の指示を解釈して実際に行動するには，言語による指示が物理世界のどのような行動によって実現されるのかがわからなくてはならない。また，物理世界の状況を人間に言語で説明することも必要である。このような研究として，例えば，言語による道案内を実際のルートにマッピングする研究（Kollar, Tellex, Roy, & Roy, 2010）や，物理世界での言語による命令を理解す

る研究（Tellex, Kollar, Dickerson, Walter, Banerjee, Teller, & Roy, 2011）が行われている。

実世界での言語理解生成において重要な要素の一つに参照表現（referring expression）がある。参照表現とは事物を指す名詞句表現である。参照表現を解釈し，実世界上の事物のどれを指すかを決定すること（参照解決）と，実世界上の事物が与えられたとき，曖昧なくそのものと解釈できる表現を生成すること（参照表現生成）が重要である。参照表現には，直示（deixis，外界への指差しを伴っての「これ，あのコップ」など），記述（description,「テーブルの上の白い本」など），照応（coreference，先行文脈を伴っての「それ」など）の３つがあり，それぞれ個別に研究されてきたが，実際の対話では区別なく用いられるため，統一的に扱う必要がある。我々は，このような多様な参照表現を統一的に扱うために，ベイジアンネットに基づく方法を提案し，マルチモーダル対話コーパスを用いた参照表現解決課題に適用し有効性を示した（Funakoshi, Nakano, Tokunaga, & Iida, 2012）。

上記に「コップ」や「テーブル」などの物体を表す表現が出てくるが，ロボットがコップを見てそれがコップであることを知るためには，事前の学習が必要である。また，物の名前はオフィスや家庭ごとに異なるので，名前と物体を結びつけて覚えさせる必要がある。画像と言語を同時に学習する手法は多く提案されているが，その多くは，複数の発話のセットから統計的に獲得するものである（Roy, 2000; Yu & Ballard, 2004；田口・岩橋・船越・中野・能勢・新田，2010）。対話の中で物体を覚える研究として，Holzapfel ら（Holzapfel, Neubig, & Waibel, 2008）の研究がある。ロボットは対話の中で未知語（out-of-vocabulary word）を発見すると語彙を学習する対話を行う。Holzapfelらは定型的なパターンの中で未知語が現れる場合のみを扱っているが，定型的でない発話も扱っていく必要がある。

我々は，マルチドメイン対話の中で，名前を教える発話を発見し，インタラクションによって名前の正しい音素列を獲得する方法を提案し，ロボットに実装した（Nakano, Iwahashi, Nagai, Sumii, Zuo, Taguchi, Nose, Mizutani, Nakamura, Attamimi, Narimatsu, Funakoshi, & Hasegawa, 2010）（図９）。このロボットは，電話番号などの問い合わせに答えられるとともに，場所や物体の名前を憶え，あとで聞かれたときにこたえることができる。物体画像の学習・認識はCCDカメラと距離センサの両方を使って行っている（Attamimi, Mizutani, Nakamura, Nagai, Funakoshi, & Nakano, 2010）。名前を教える発話かどうかは，言語理解結果と文脈の情報を用いて決定する。名前を教える発話が検出され，名前を覚えるドメインの対話が開始されると，検出した未知語の音素列をユーザに確認し，間違っている場合は言い直してもらう。すなわち，

人間：この場所の名前はロボノマ
ロボット：ロボロマですか？
人間：ロボノマ
ロボット：ロボノマですか？

のようなやり取りが行われる。このとき複数の発話の認識結果から，より良い音素列を取り出すことでなるべく早く正解音素列にたどりつくと考えられるが，音素列毎に計算した確信度を用いることで，早く正解にたどり着く手法（Zuo, Sumii, Iwahashi, Nakano, Funakoshi, & Oka to appear）を用いている。

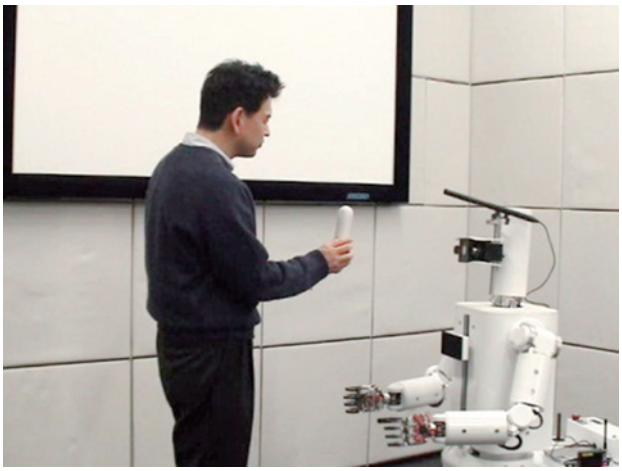

図９　物の名前と画像を覚える対話ロボット

6. 対話ロボット研究の発展に向けて

以上見てきたように対話ロボットに関する様々な課題が研究されているが，異なる分野の研究者の間のコミュニケーションが少ないことが問題である。対話ロボットに関連している研究者は，少なくとも次の５つの分野から来ている。

- 対話システム研究：音声・マルチモーダル対話システムの研究をバックグラウンドに持ち，対話技術をロボット対話に適用しようとしている。対話ロボットを構築し，一般のユーザに使ってもらって評価することを目指す。
- 音声認識・理解研究：音声認識・理解技術のアプリケーションとしての対話ロボットに興味があり，主に音声認識率・音声理解率の向上を目指す。
- サービスロボット研究：サービスロボットを実際に開発することに興味があり，対話は一つのインタフェースととらえている。
- ヒューマンロボットインタラクション研究：人間とロボットがより良いインタラクションを行うためには，ロボットがどのような振る舞いをすべきかに興味がある。ロボットの振る舞いに対して人がどのよ

うな印象を持つかを心理実験で調べるが，必ずしも自律ロボットを構築せず，Wizard-of-Oz方式で実験を行うことも多い。

- ●発達ロボティクス研究：人間がどのように言語コミュニケーションや行動を学習するのかに興味があり，それをロボットでシミュレーションすることで理解しようとしている。

これらの分野の各々で，前提となる知識や目標が異なっており，相互理解が容易ではない。例えば，対話技術で重要なエラーハンドリング・意図推定といった技術は他の分野の研究者にはあまり知られていない。そのため，サービスロボット研究者が対話機能を導入しようとしたときに，とりあえず音声認識を使って音声コマンド理解ができるようにしたものの，実際に使用する場面でエラーハンドリングが必要になり，アーキテクチャを見直すというようなケースが見受けられる。また，対話システム研究者がロボットを用いるときには，ロボットの制御との統合は考慮にいれないため，移動を含むようなタスクが扱えない場合がある。対話ロボットの研究を効率的に行うためには，対話ロボット関連の研究者間での情報共有を進める必要がある。

7. おわりに

本論文では，対話ロボット，特に，対話サービスロボットのモデルとして，我々の提案したHRIMEを紹介した。HRIMEは様々なタスクに応用できる汎用的なモデルであり，様々なロボットの開発に役立てることが期待できる。また，本論文では対話ロボット研究の課題について述べた。特に，対話ロボットに関係する様々な分野の研究者の情報交換が大事であることに触れた。

対話ロボット研究の発展のためには，何よりも，実際に役に立つ対話ロボットを早く作って，その将来性を示すことが必要だと考えている。そのためには，具体的な目標を立てた上で，ハードウエアの研究者も含めた協力体制を作る必要があると思われる。サービスロボットの分野では，ロボカップ@ホームリーグというコンテストが行われており，日本のチームも好成績をおさめている（岡田・大森，2010）。ロボカップ@ホームリーグでは，キッチンやリビングルームなどの家庭環境において人の名前と顔を覚えたり，移動してゴミを集めたりするようなタスクを扱う。対話システム研究者が対象にするような制限のない対話は扱われていないが，今後そのような対話も課題に入れば，対話サービスロボット研究が加速すると考えられる。

謝辞

本論文で述べた研究は㈱ホンダ・リサーチ・インスティチュート・ジャパンの皆様や共同研究者の皆様とともに進めたものです。心より感謝いたします。

引用文献

浅田稔，國吉康夫（2005）．『ロボットインテリジェンス』岩波書店.

Asoh, H., Motomura, Y., Asano, F., Hara, I., Hayamizu, S., Itou, K., Kurita, T., Matsui, T., Vlassis, N., Bunschoten, R., & Kröse, B.（2001）. Jijo-2: An Office Robot that Communicates and Learns. IEEE Intelligent Systems, 16(5), 46–55.

Muhammad Attamimi, Akira Mizutani, Tomoaki Nakamura, Takayuki Nagai, Kotaro Funakoshi, Mikio Nakano (2010) : Real-time 3D visual sensor for robust object recognition. Proc. IROS 2010, 4560–4565.

Bohus, D. & Horvitz, E.（2009a）. Learning in Open-World Settings. Proc. SIGDIAL 2009, 244–252.

Bohus, D. & Horvitz, E.（2009b）. Models for Multiparty Engagement in Open-World Dialog. Proc. SIGDIAL 2009, 225–234.

Bohus, D., Horvitz, E., Kanda, T., Mutlu, B., & Raux, A.（Eds.）(2010). Dialog with Robots: Papers from the AAAI Fall Symposium. AAAI.

Dan Bohus, Alexander I. Rudnicky（2005）: Error Handling in the RavenClaw Dialog Management Architecture. Proc. HLT/EMNLP.

Brooks, R. A.（1986）. A Robust Layered Control System for a Mobile Robot. IEEE Trans. on Robotics and Automation, 2, 14–23.

Clark, H. H.（1996）. Using Language. Cambridge University Press.

藤江真也，江尻康，菊池英明，小林哲則（2005）．肯定的／否定的発話態度の認識とその音声対話システムへの応用　『電子情報通信学会論文誌（D–II）』, J88–D2(3), 489–498.

藤江真也，松山洋一，谷山輝，小林哲則（2012）. 人同士のコミュニケーションに参加し活性化する会話ロボット『電子情報通信学会論文誌A』, J95–A(1), 37–45.

船越孝太郎，小林一樹，中野幹生，小松孝徳，山田誠二（2011）．対話の低速化とArtificial Subtle Expressionによる発話衝突の抑制　『人工知能学会論文誌』, 26(2), 353–365.

Funakoshi, K., Nakano, M., Tokunaga, T., & Iida, R.（2012）. A unified probabilistic approach to referring expressions. Proc. SIGDIAL 2012, 237–246.

Hasegawa, T., Nakano, Y. I., & Kato, T.（1997）. A Collaborative Dialogue Model Based on Interaction between Reactivity and Deliberation. Proc. 1st Int' l Conf. on Autonomous Agents, 75–82.

Higashinaka, R., Nakano, M., & Aikawa, K.（2003）. Corpus-based Discourse Understanding in Spoken Dialogue Systems. Proc. 41st ACL.

Higashinaka, R., Sudoh, K., & Nakano, M.（2006）. Incorporating Discourse Features into Confidence Scoring of Intention Recognition Results in Spoken

Dialogue Systems. Speech Communication, 48（3-4）, 417-436.

Holzapfel, H., Neubig, D., & Waibel, A.（2008）. A dialogue approach to learning object descriptions and semantic categories. Robotics and Autonomous Systems, 56(11), 1004-1013.

Hoshino, Y., Takagi, T., Diprofio, U., & Fujita, M.（2004）. Behavior description and control using behavior module for personal robot. Proc. ICRA-2004, 4165-4171.

Ince, G., Nakadai, K., Rodemann, T., Tsujino, H., & Imura, J.（2011）. Whole Body Motion Noise Cancellation of a Robot for Improved Automatic Speech Recognition. Advanced Robotics, 25（11-12）, 1405-1426.

石黒浩，宮下敬宏，神田崇行（2005）.『コミュニケーションロボット』 オーム社.

Kanda, T., Ishiguro, H., Imai, M., Ono, T., & Mase, K.（2002）. A constructive approach for developing interactive humanoid robots. Proc. IROS-2002, 1265-1270.

勝丸真樹，中野幹生，駒谷和範，船越孝太郎，辻野広司，尾形哲也，奥乃博（2010）：複数の言語モデルと言語理解モデルによる音声理解の高精度化 『電子情報通信学会論文誌』, Vol. J93-D, No. 6, 879-888.

Kollar, T., Tellex, S., Roy, D., & Roy, N.（2010）. Toward understanding natural language directions. Proc. 5th HRI, 259-266.

駒谷和範，河原達也（2002）. 音声認識結果の信頼度を用いた効率的な確認・誘導を行う対話管理 『情報処理学会論文誌』, 43(10), 3078-3086.

小松孝徳，山田誠二，小林一樹，船越孝太郎，中野幹生（2010）. Artificial Subtle Expressions：エージェントの内部状態を直感的に伝達する手法の提案 『人工知能学会論文誌』, 25(6), 733-741.

Lin, B., Wang, H., & Lee, L.（1999）. A Distributed Architecture for Cooperative Spoken Dialogue Agents with Coherent Dialogue State and History. Proc. ASRU-99.

Ma, Y., Raux, A., Ramachandran, D., & Gupta, R.(2012). Landmark-Based Location Belief Tracking in a Spoken Dialog System. Proc. SIGDIAL 2012, 169-178.

Makihara, Y., Takizawa, M., Ninokata, K., Shirai, Y., Miura, J., & Shimada, N.（2002）. A Service Robot Acting by Occasional Dialog —Object Recognition Using Dialog with User and Sensor-Based Manipulation—. Journal of Robotics and Mechatronics, 14(2), 124-132.

松坂要佐，東條剛史，小林哲則（2001）. グループ会話に参与する対話ロボットの構築 『電子情報通信学会論文誌D-II』, 84(6), 898-908.

Mochiki, N., Ogawa, T., & Kobayashi, T.（2008）. Ears of the robot: Direction of arrival estimation based on pattern recognition using robot-mounted microphones. IEICE Transactions on Information and Systems, E91D(5), 1522-1530.

中臺一博，奥乃博（2011）. ロボット聴覚用オープンソースソフトウェアHARKの展開 『情報処理学会デジタルプラクティス』, 2(2), 133-140.

Nakajima, H., Nakadai, K., Hasegawa, Y., & Tsujino,H.（2010）. Blind Source Separation With Parameter-Free Adaptive Step-Size Method for Robot Audition. IEEE Trans. on Audio, Speech and Language Processing, 18(6), 1476-1485.

Nakano,M., Hoshino,A., Takeuchi,J., Hasegawa,Y., Torii, T., Nakadai, K., Kato, K., & Tsujino, H.（2006）. A Robot That Can Engage in Both Task-Oriented and Non-Task-Oriented Dialogues. Proc. Humanoids-2006, 404-411.

中野幹生，船越孝太郎，長谷川雄二，辻野広司（2008），オブジェクト指向に基づくロボット・エージェントのマルチドメイン対話行動制御モジュール構築ツールRIME-TK 人工知能学会研究会資料, SIG-SLUD-54.

Nakano, M., Iwahashi, N., Nagai, T., Sumii, T., Zuo, X., Taguchi, R., Nose, T., Mizutani, A., Nakamura, T., Attamimi, M., Narimatsu, H., Funakoshi, K., & Hasegawa, Y.（2010）. Grounding New Words on the Physical World in Multi-Domain Human-Robot Dialogue. Proc. AAAI 2010 Fall Symposium on Dialog with Robots, 74-79.

Nakano, M., Hasegawa, Y., Funakoshi, K., Takeuchi,J., Torii,T., Nakadai,K., Kanda,N., Komatani, K., Okuno, H. G., & Tsujino, H.（2011）. A multi-expert model for dialogue and behavior control of conversational robots and agents. Knowledge-Based Systems, 24(2), 248-256.

Mikio Nakano, Shun Sato, Kazunori Komatani, Kyoko Matsuyama, Kotaro Funakoshi, and Hiroshi G. Okuno（2011）: A Two-Stage Domain Selection Framework for Extensible Multi-Domain Spoken Dialogue Systems. Proc. SIGDIAL 2011 Conference, 18-29.

西村竜一，西原洋平，鶴身玲典，李晃伸，猿渡洋，鹿野清宏（2004）. 実環境研究プラットホームとしての音声情報案内システムの運用 『電子情報通信学会論文誌』, J87-D-II(3), 789-798.

Nishimura,Y., Kushida,K., Dohi,H., Ishizuka,M., Takeuchi, J., Nakano, M., & Tsujino, H.（2007）. Development of Multimodal Presentation Markup Language MPML-HR for Humanoid Robots and Psychological Evaluation. Int. J. of Humanoid Robotics, 4(1), 1-20.

O'Neill, I. M. & McTear, M. F.（2001）. Objectoriented modelling of spoken language dialogue systems. Natural Language Engineering, 6（3&4）, 341-362.

岡田浩之，大森隆司（2010）. ロボカップ@ホーム―人とロボットの共存を目指して― 『人工知能学会

誌』，25(2), 229-236.
Paek, T. & Horvitz, E.（2000）. Conversation as Action Under Uncertainty. Proc. UAI-2000.
Raux, A. & Eskenazi, M.（2009）. A finite-state turn-taking model for spoken dialog systems. Proc. NAACL-HLT 09, 629-637.
Rich, C., Ponsler, B., Holroyd, A., & Sidner, C. L.（2010）. Recognizing Engagement in HumanRobot Interaction. Proc. 5th HRI, 375-382.
Roy, D.（2000）. Integration of Speech and Vision Using Mutual Information. Proc. ICASSP-2000, 2369-2372.
白井克彦，小林哲則，岩田和彦，深沢克夫（1985）. ロボットとの柔軟な対話を目的とした音声入出力システム　『日本ロボット学会誌』，3(4), 362-372.
志和敏之，奥野佑将，神田崇行，今井倫太，石黒浩，萩田紀博（2012）. コミュニケーションロボットによる道案内―ジェスチャの有用性と発話タイミングのモデル化―　『電子情報通信学会論文誌D』，J95-D(10), 1818-1828.
田口亮，岩橋直人，船越孝太郎，中野幹生，能勢隆，新田恒雄（2010）. 統計的モデル選択に基づいた連続音声からの語彙学習　『人工知能学会論文誌』，25(4), 549-559.
Tellex, S., Kollar, T., Dickerson, S., Walter, M. R., Banerjee, A. G., Teller, S., & Roy, N.（2011）. Understanding Natural Language Commands for Robotic Navigation and Mobile Manipulation. Proc. 25th AAAI.
Thomaz, A. L. & Chao, C.（2011）. Turn-taking based on information flow for fluent human-robot interaction. AI Magazine, 32(4).
Traum, D. R. & Allen, J. F.（1994）. Discourse Obligations in Dialogue Processing. Proc. 32nd ACL, 1-8.
Williams, J. D. & Young, S.（2007）. Partially Observable Markov Decision Processes for Spoken Dialog Systems. Comp. Speech and Lang., 21(2), 393-422.
Yoshimi, T., Matsuhira, N., Suzuki, K., Yamamoto, D., Ozaki, F., Hirokawa, J., & Ogawa, H.（2004）. Development of a Concept Model of a Robotic Information Home Appliance, ApriAlpha. Proc. IROS-2004, 205-211.
Yu, C. & Ballard, D.（2004）. On the Integration of Grounding Language and Learning Objects. Proc. 19th AAAI, 488-494.
Zobel, M., Denzler, J., Heigl, B., Noth, E., Paulus,D., Schmidt, J., & Stemmer, G.（2001）. MOBSY: Integration of Vision and Dialogue in Service Robots. Proc. Second International Workshop on Computer Vision Systems（ICVS）, 50-62.
Zuo, X., Iwahashi, N., Funakoshi, K., Nakano, M., Taguchi, R., Matsuda, S., Sugiura, K., & Oka, N.（2010）. Detecting Robot-Directed Speech by Situated Understanding in Physical Interaction. Transactions of the Japanese Society for Artificial Intelligence, 25(6), 670-682.
Zuo, X., Sumii, T., Iwahashi, N., Nakano, M., Funakoshi, K., & Oka, N.（to appear）. Correcting phoneme recognition errors in learning out-of-vocabulary words through speech interaction. Speech Communication.

中野　幹生（なかの　みきお）
1988年東京大学教養学部基礎科学科第一卒業。1990年東京大学大学院理学系研究科相関理化学専攻修士課程修了。1990年～2004年日本電信電話㈱にて，自然言語処理，音声対話システムの研究に従事。1998年博士（理学）取得。2004年㈱ホンダ・リサーチ・インスティチュート・ジャパン入社。現在同社プリンパル・リサーチャ。2011年より早稲田大学客員教授。対話ロボット，対話システム，音声言語理解の研究に従事。人工知能学会，情報処理学会，言語処理学会，電子情報通信学会，AAAI，ACM，IEEE，ISCA各会員。

メディア教育研究　第９巻　第１号
Journal of Multimedia Education Research 2012, Vol.9, No.1, S29－S41

Toward the Development of Usable Dialog Robots —Models and Issues of Linguistic Interactions in the Physical World—

Mikio Nakano[1)]

This paper discusses technological challenges in the development of usable dialog robots, and presents intelligence models for dialog robots. We first discuss the values of dialog robots and shows conversational service robots are promising. Then we present HRIME, a framework for dialog and behavior controllers that integrate models for intelligent robots and spoken dialog systems. In addition we describe issues in building dialog robots and existing work on those issues. We finally discuss the problems to solve to promote dialog robot research.

Keywords
dialog robot, intelligent robot, spoken dialog system, multimodal interaction

[1)] Honda Research Institute Japan Co., Ltd.