音声言語情報処理 30－3
（2000．2．4）

事情通ロボットの音声対話インターフェース

麻生 英樹, ジョン フライ†, 浅野 太, 速水 悟, 伊藤 克亘
本村 陽一, 原 功, 栗田 多喜夫, 松井 俊浩
電子技術総合研究所 実世界知能研究推進センター
† スタンフォード大学 CSLI

和文抄録: 事情通ロボット *Jijo-2* の音声対話インターフェースについて報告する．雑音の多い実環境下でのロバストかつ自然な対話を実現するために，マイクロフォンアレイによる音源位置推定とビームフォーミング，異なる辞書を持つ複数の音声認識モジュールの対話状況による切り替え，タスク指向なフレームによる意味表現，対話の文脈情報を利用した欠落情報の補完，等の技術を統合的に利用している．全システムは実ロボット上に実装され，地図の学習，訪問者の案内，所在データベース検索／登録等のタスク実行が可能になっている．

Spoken Dialog Interface of the Office Conversant Robot, *Jijo-2*

Hideki Asoh, John Fry†, Futoshi Asano, Satoru Hayamizu, Katsunobu Itou
Yoichi Motomura, Isao Hara, Takio Kurita, and Toshihiro Matsui
Real World Intelligence Center, Electrotechnical Laboratory
† Center for the Study of Language and Information, Stanford University

Abstract: A spoken dialog interface of the mobile office robot, *Jijo-2* is reported. To realize robust and flexible spoken dialog in noisy office environments three techniques are integratedly exploited. They are: a microphone array system for sound source detection and beam forming, switching multiple speech recognition processes with different dictionaries depending on dialog state, task dependent semantic frames, and keeping track of contextual information of dialog to fill omitted information. The system is implemented on a real mobile robot base and evaluated with some tasks such as dialog based map learning, guiding visitors, and accessing databases.

1. はじめに

情報処理システムやロボットの活躍の場が銀行や工場のような定型的業務の場から家庭や個人へと広がるにつれて，あらかじめ限定された「閉じた」世界ではなく，多様かつ変化に富んだ実世界で人間と柔軟にコミュニケートし，協調して働くことのできる知的システム「実世界知能 (Real World Intelligence)」の実現が幅広く期待されるようになってきている．

われわれは，実世界知能の一つのプロトタイプとして，通常のオフィス環境において自律的に行動し，さまざまな情報を収集することを通じてそのオフィスに関する事情通に育ってゆくような自律学習型移動ロボット「事情通ロボット」を構想し，Real World Computing Program の下でその実現に向けた研究を進めてきている [9, 18, 20, 23, 24].

事情通ロボットの主たるサービスは「オフィス環境についての情報を提供すること」である．より具体的には，訪問者の案内や配達といった通常の受付ロボットとしての機能の他に，遠隔地からのアクセスに対する情報提供サービス，メンバのスケジュール管理，ミーティングマネジメントといった，一種のグループウェアとしてのサービス，など多様なオフィス環境における知的作業支援が想定されている [26]．これらのサービスを提供するためには，オフィスに存在するさまざまな種類の情報について「事情通であること」が重要であるが，オフィス環境は変化に富んでいるため，ロボットは自律的にデータを収集し常に学習しつづけることが必要になる．このようなロボットにとって，音声対話はユーザとのインタラクションを円滑にするとともに，情報収集のための有力な手段の一つでもある [6, 7].

ロボットのための音声対話インターフェースというアイデアは古くからあるが，実際のロボット上に実装された例は多くはない．また，実際に実装されたシステムの多くは，単純なコマンドをロボットに伝えるものであった．しかし，近年，音声認識技術の水準の向上にも支えられて，音声対話をロボットとのインターフェースに利用する試みが盛んになってきている [13, 25, 29].

実際のオフィス環境で複数のタスクをカバーする

音声対話システムを，多様なセンサと限定された計算資源を持つロボット上に実装するためには，以下のような課題を解決する必要がある．

1. ノイズや反響のある実環境での頑健な音声認識と理解
2. 多様なセンサ情報を統合的に利用しながら多様なタスクをカバーする柔軟な対話制御
3. 限定された計算資源を用いた実時間の応答

現在のシステムでは，最初の課題に対処するために，マイクロフォンアレイを用いた話者位置推定とビームフォーミング，および異なる辞書を持つ複数の音声認識プロセスの対話状況依存な切り替えを，二番目の課題に対処するために，タスク指向なフレームによる意味記述を利用した対話管理を，三番目の課題に対処するために，DSP によるベクトル量子化を利用している．また，日本語発話に多い省略に対処するために，対話の文脈情報を用いた省略部分の補完機能を組み込んだ．

2. システムの構成

現在の事情通ロボットシステム *Jijo-2* のハードウェア構成を図 1 に示す．ベースとなっているのは，米国 Nomadic 社製の Nomad 200 という移動ロボットベースで，環境中の物体との距離計測のための超音波センサ，赤外線近接距離センサ，衝突検出のための接触センサと，移動距離・回転角度計測のための車輪回転のエンコーダを備えている．これらのセンサおよびモーターは，Linux で動作する通常の PC によって制御されている．ロボットは無線イーサネットを通じて LAN に接続しており，リモートのワークステーションと 1M bit/sec 程度の速度で情報をやりとりすることができる．

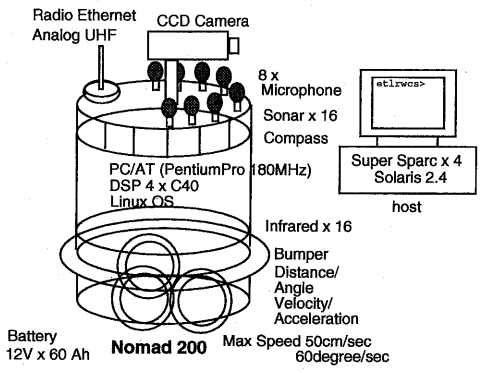

Figure 1. 事情通ロボット *Jijo-2* のハードウェア

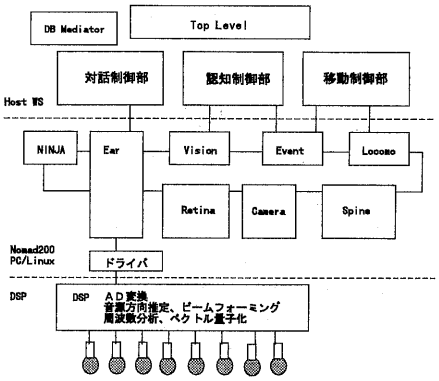

Figure 2. *Jijo-2* のソフトウェアモジュール

このロボットベースに，音声対話のための装置として 8 チャンネルのマイクロフォンアレイ，音響・音声処理用の DSP モジュール，日本語音声合成装置（NTT データ通信（株）の「しゃべりん坊」）を追加した．また，顔やジェスチャの認識を行うためのパン・ティルト可能な CCD カメラと，場所認識，障害物検出，移動物体追跡等のための球面鏡のついた全方位視覚用カメラとを追加している．

障害物回避や音声認識などの即応的な処理と，経路計画や対話制御のような熟考的な処理とを統合するために，複数のモジュールから成るソフトウェア群をイベント駆動方式によって制御している [8, 22]．現在のソフトウェアの構成を対話部分を中心にして描くと図 2 のようになる．ソフトウェアは機能別にモジュール化され，それぞれ UNIX のプロセスとして実行されている．相互の通信は TCP/IP のソケットおよび共有メモリを介して行なわれる．中央上部の点線よりも下のモジュール群は障害物回避や音声認識など即応的な機能を受け持つもので，C で実装され，ロボットの CPU 上で実行されている．点線より上のモジュール群は，経路計画や対話管理など，より熟考的な機能を受け持つもので，ロボットプログラミング用オブジェクト指向 LISP である *EusLisp*[21] で記述されている．また，データベースメディエータモジュールの一部は Java のデータベースアクセス機能を用いて実装されている．

センサが特定のイベント（たとえば，「こんにちは」というユーザからの呼び掛け）を検出することが口火となり，それぞれのモジュールが他のモジュールに request メッセージを送って仕事を要請することで処理が進められる．request を受け取ったモジュールは要請された仕事を処理し，終わった時点で reply メッセージを返す．メッセージの到着がそれぞれのモジュールにとっての「イベント」であり，それ

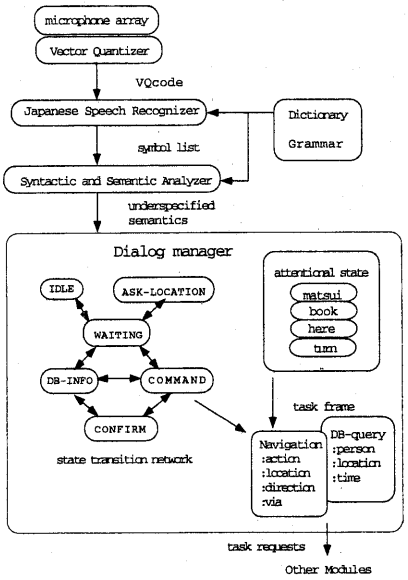

Figure 3. 発話処理の流れ

によって、個々のモジュールの処理が駆動される.

図3に, ear モジュールおよびその子プロセスである音声認識モジュール, そして上位の対話管理モジュールによる一つの発話の処理の流れを示した. 以下では, この流れに沿って個々の部分システムついてより詳しく述べる.

2.1 マイクロフォンアレイ

音声入力部には, 8チャンネルのマイクロフォン・アレイを構築し, 1個の DSP(TI TMS320C44) を用い, 遅延和法によってダイナミック (大きな入力音がある間約1秒毎) な音源方向の推定とその方向へのビームフォーミングを行なっている. これによって周囲雑音を 1.5 kHz から 3 kHz の周波数帯域で最高で 10dB 程度軽減し, ノイズの多い環境下での単語音声認識の正答率を 10% 程度改善することができた [5].

2.2 音声認識

音声認識には, 電総研で開発された HMM ベースの不特定話者連続音声認識システム niNja を用いている [17]. niNja は, 音声の周波数分析結果から VQ コードを生成する部分と, VQ コードから辞書と文法を用いて認識結果 (語記号の並び) を生成する部分に分かれている. 周波数分析とベクトル量子化をもう一つの DSP(C44), 言語情報を用いた VQ コードのデコーディングをロボット内の PC (Pentium Pro 180 MHz) で処理している.

マイク	音韻モデル	ビーム幅	1位	5位以内
口元	mono	250	74 %	91 %
口元	mono+tri	200	86 %	94 %
遠隔	mono	200	47 %	64 %

Table 1. 発話レベルの認識率

VQ 用の音韻モデルには離散型 HMM によるモノフォンモデルを用いている. デコーダ用の辞書と文法には, ロボットのタスクにチューンしたものを用意した. 登録語数は固有名詞も含めて約 200 語, 文法ルールの数は約 90 である. すべての語彙を含む辞書, 文法 (full-grammer) の他に, 「はい」と「いいえ」という応答に関する語だけを含む reply-grammer, 固有名詞や場所の名称だけを含む name-grammer の二つを用意し, それぞれの辞書, 文法を用いた niNja のプロセス3つを並行して走らせて, 対話の状況による辞書の切り替えを実現した. ベクトル量子化に DSP を用いたことで, 3つのプロセスを並行に走らせた場合でも, 話し終わりから認識結果を得るまで1秒以内程度の良い応答性能を得ている.

音声認識部分の性能を, 認識可能な 95 文を読み上げたデータを用いて予備的に評価した. 用意した文章は, 「こんにちは」のような 1 語文から「松井さんの予定を教えてください」「原さんのスケジュールを知っていますか」「これを山崎さんに届けてください」程度の語数の単文で, 6 人 (すべて男性) の話者が, 通常のマイクが口元にある状態で読み上げたデータ (口元) および反響の多い部屋で口元から約 50cm 離れたところにマイクをセットした状態で読み上げたデータ (遠隔) の二つのデータセットを用意した. 発話レベルの認識率を表1にまとめた.

音韻モデルの mono + tri は, モノフォンのモデルで得たスコアをもとに, トライフォンモデルでリスコアリングした結果を用いたことを示している. この結果から, 通常の状態では十分な認識性能が得られるシステムでも, 口元からマイクが離れた状態では高い認識率を得ることがむずかしいことがわかる.

2.3 構文・意味解析

多様な言い回しの可能性を持つユーザ発話の意図を理解し, 適切な応答をするために, 表層的な認識結果から, 言い回しによらない意味内容を抽出することが望ましい. まず, 音声認識用の出力記号のレベルで, 語彙レベルの簡単な同義化を行っている. たとえば, 「こんにちは」「こんばんは」はどちらも (hello) という認識結果ラベルを出力する. 「ここは松井さん

の部屋です」「ここは松井のところです」はどちらも (**here Matsui s office is**) というラベルリストを出力する.

この認識結果を用いて, さらに構文および意味の解析を行っている. 構文・意味解析には, 音声認識部と同じ辞書と文法が用いられる. 品詞情報や係り受け情報を出力する通常の日本語構文解析とは異なり, 下記のように文法中に埋め込まれた意味記述用のトークンを用いて構文解析と意味解析を一体として行い, 文の種別 (挨拶, 命令, 平叙, 疑問) の識別や, それに基づくタスク種別の識別および, 各タスクの実行に必要な情報の抽出を行っている.

```
sentence: greeting
sentence: imperative
sentence: declarative
imperative: action
imperative:  action please
action: motion
action: direction to motion
direction: RIGHT
direction: LEFT
```

終端記号は, その語彙クラスに属する単語定義辞書ファイルに対応する. 構文・意味解析の結果は, 以下のような形式の意味表現である.

```
「こんにちは」  (greeting hello)
「右にまわって」(imperative :action turn
                            :direction right)
```

2.4 対話制御

構文・意味解析結果は対話管理モジュールに渡される. 対話管理モジュールでは, 双方向性の対話を実現するために, 状態遷移ネットワークとタスク毎のフレーム的な意味記述 (form-based[28]) とを組み合わせた対話制御を行っている. 現在のシステムでは以下の 6 つの状態からなる状態遷移ネットワークが用いられている.

IDLE	ユーザの呼び掛け待ち状態 (初期状態)
WAITING	ユーザのタスク依頼待ち状態
DB-INFO	データベース対話の状態
COMMAND	移動や人の呼出し対話の状態
ASK-LOCATION	自己位置確認の対話の状態
CONFIRMATION	ユーザ入力の確認対話の状態

状態に依存した応答および状態遷移のためのルールは, 状態, 入力文, 文脈 (期待する文) の 3 つの情報から取るべきアクションおよび状態遷移への if-then ルールで, Prolog 風に記述され, Lisp 上に実現された Prolog インタプリタによって解釈される.

フォームは、各タスクの実行に必要な情報をスロットとする意味記述用のフレームである. 現在, DB-query (データベース問い合わせ), DB-update (データベース更新), identify-person (人物同定), navigation (移動), call (人物呼び出し) の 5 種類のタスク用フォームが用意されている. それぞれのフォーム (プログラム中では transaction-context と呼んでいる) は *EusLisp* のオブジェクトとして定義され, タスクの実行に必要な情報のスロットを持つ. たとえば人物呼び出しタスクに対応する call-transaction-context の場合, タスクの実行に必要な情報は呼ぶ相手を示す person スロットおよび, タスク実行の依頼者による確認が済んでいるか否かを記録する confirmation スロットの二つである.

一つの発話の内容だけでタスクの実行に必要なスロットが埋まらない場合には, まずデフォルト値を用いた補完を試み, さらに, 先行する対話文脈から得られる情報によって省略を補うことを試みる. それでも欠落がある場合にはそのスロットの値をユーザに能動的に尋ねる. 先行する対話の文脈情報はキューによって管理されており, 一つのタスク実行が終了した時点で, そのタスク実行のために作成されたフォームの内容が文脈管理用のキューに加えられ, 古い情報から順に失われてゆく [11]. 新しいタスクフォームのスロットを埋める情報は, スロットの種別をキーとして検索され, 一致するものが存在した場合に新しいものを優先に取り出されて利用される.

このようにして, ユーザとのやりとりを進め, タスクの実行に必要な情報が得られたところで, 他のソフトウェアモジュールに対するタスク実行の要求が送出される.

誤認識の影響が継続的に残ることを避けるために, 「取り消して」などの対話文脈をクリアするためのメタコマンドが利用可能になっている. また, 発話生成は, 用意された少数の発話テンプレートに基づく単純なものである.

3. 音声対話例

図 4 に複数の行動／タスク実行を含む典型的な対話例を示した. この対話の中で, ロボットは

1. 話者の方向を向き,
2. 話者の顔を認識し,
3. オフィスメンバーの所在に関する情報を提供し,
4. 電子メールでオフィスメンバーを呼び出し,
5. 話者をメンバの居室スペースまで案内し,
6. 所在情報の変更をデータベースに登録する,

松井 こんにちは
話者のほうを向いて顔を検出し、認識する。
Jijo-2 こんにちは松井さん
松井: 麻生さんはどこですか？
Jijo-2: 麻生さんは本館です。
松井: 呼んでください。
Jijo-2: 麻生さんにメールを送りましょうか？
松井: はい。
Jijo-2: メールを送りました。
松井: 本村さんは？
Jijo-2: 本村さんはオフィスにいます。
松井: 連れていってください。
Jijo-2: 本村さんの部屋に行きます。
本村の部屋までの経路を計画し、実行する。
Jijo-2: 本村さんの部屋に着きました。
松井: ありがとう。
Jijo-2: どうしたしまして。

本村： 食事に行きます。
Jijo-2: 本村さんは食事、でよろしいですか？
本： はい。
Jijo-2: スケジュール登録します。
松井： 松井も
Jijo-2: 松井さんは食事、でよろしいですか？
松井： はい。
Jijo-2: スケジュール登録します。

Figure 4. 対話例

という行動を実行している．ここでは、この対話例に沿って，いくつかの発話について実際に行われる処理を，対話以外のものも含めて説明する．

1. 既に述べたようにシステムは始め IDLE の状態にある． この状態では，システムは「こんにちは」等の greeting カテゴリの入力のみを受け付ける．「こんにちは」という発話は (**greeting hello**) という意味解析結果に変換されて，対話管理モジュールへの入力イベントとなる．

   入力を受け取った対話管理モジュールは

   ```
   +(s *x *con *class *foc (greeting hello))
     -(clear-conversation-context)
     -(submit-task (:hello))
     -(kaiwa-state :waiting).
   ```

   という if-then ルールに従って，1) 推定音源方向に向き直る，2) 話者の顔部分を検出し顔認識を行う，3) 「こんにちは xxx さん」と発話する，という一連のアクションの実行要求を他のモジュールに出すとともに，4) WAITING 状態に移行する．上のルールの +( ) の部分が if 部，-( ) の部分が then 部である．

   顔の検出と認識には，ロボットとユーザとの距離が変動することを考慮し，Log-Polar 変換と高次自己相関特徴を組み合わせた大きさの変化に強い特徴量を用いている [14, 15, 19]．探索時間を短くするために，顔発見には肌色部分の抽出による粗い探索と，モノクロの顔画像とそれ以外の画像のデータを用いた判別分析による詳細探索とを組み合わせて用いた [12].

2. 「麻生さんはどこですか？」という発話はオフィスメンバの所在に関する問い合わせ (location-query) と解釈され，DB-query のタスクフォーム (query-transaction-context というクラスのインスタンス) が作成されて，データベース名や検索キーなど問い合わせに必要な情報が埋められる． メディエータモジュールを介して LAN 上の他のマシンにあるデータベースへの問い合わせ要求が送られ，それに対する回答が得られると，「麻生さんは本館に居ます」という話者への答えが発話される．

   所在データベースは汎用のデータベースサーバ PostgresSQL を用いて実装されており，Java applet による Web ブラウザを通じてのデータの表示・更新と，データベースメディエータモジュールによる SQL 文での問い合わせの両方を可能にしている．たとえば，ロボットによって更新された結果は，Java applet 側の表示にも反映される [26].

3. 「本村さんは？」という発話は述語の省略を含むものであるが，システムは先行文脈情報を用いて一つ前の発話と同様に，所在データベース問い合わせのタスクを実行している．

4. 「連れていってください」という発話は，navigation カテゴリの命令文と解釈され，navigation-transaction-context クラスのインスタンスが作成される． このタスクフォームは person, action, direction, adverb というスロットを持つ．現在はすべての移動目的地はオフィスメンバの居室であり，その人の名前によって表現されている．従って，person スロットが移動目的地を表す．action には行動の種類（この場合は go-to), direction および adverb には action の実行のための補助情報（たとえば，回転の方向や，移動速度）が入る．

   「連れていって」という発話の意味解析結果から，要求されている action の種類が go-to であることが得られる．この action を実行するために最低限必要な情報は移動目的地＝ person スロットの値であるため，スロット種別 person をキーとして先行文脈情報が検索される． 結果として「本村さんは（どこ）？」という発話の解釈結果中の Motomura という人名が得られ，これを目的地スロットの値として，移動の要求メッセージが移動管理モジュールに伝えられて実行される．

移動管理モジュールによる経路計画には，オフィスメンバの居室や交差点などのランドマーク間をその間を遷移するための要素移動行動でむすんだ位相的な地図を用いている．地図はあらかじめロボットに与られるのではなく，対話を通じて教示されたものである [6]．要素移動行動としては，障害物回避をしつつ通路に沿って進む (go-straight)，右に 90 度回転する (turn-to-right)，後ろを向く (turn-around) などが用意されている．

この例には含まれていないが，音声入力以外のイベントが対話のきっかけになることもある．たとえば，センサ情報のノイズや環境の変化などのためにが移動中にロボットの自己位置が不確実になることがある．その場合には，移動管理モジュールが対話管理モジュールに対して自己位置確認対話を request する．これを受けた対話管理モジュールは ASK-LOCATION 状態になり，近くにいる人間に「ここはどこですか」と尋ね，「原さんのところです.」などの答えを受け取って移動管理モジュールに返す．移動管理モジュールはその結果を用いて経路を再計画し，目的地への移動を継続する．

4. 考察と今後の課題

事情通ロボット *Jijo-2* 上に実装された音声対話インターフェースの現状および典型的な対話例について述べた．現在のシステムは原初的なものであり，まだまだ不十分な点が多い．以下では，その構築を通じての考察および今後の課題について述べる．

4.1 実環境での頑健な音声認識

まず，実環境での音声認識は依然として大きな課題である．音声認識の評価でも述べたとおり，雑音や反射の多い実環境でマイクから離れて話しかけるような状態では，高い認識率を得ることは容易ではなく，対話者のフラストレーションをある程度に抑えるためには，登録語数および文法の複雑さをかなり制限する必要があると思われる．定量的な評価は行っていないが，対話の状況に依存した辞書の切替えは有効に機能していると思われる．また，省略語の補完機能は，自然な対話を可能にするとともに，音声認識の誤りを減らすことにも寄与していると思われる．

マイクロフォンアレイは，一般に横幅が広いほど低い周波数まで雑音を抑制できるが，今回用いたマイクロフォンアレイは，ロボットの半径長による制約から，約 1.5kHz 以下での雑音抑制効果が少なく，実環境における低域優勢の雑音に対して，十分な効果が得られているとはいえない．現在，Wiener Filter などを併用することにより，低域での雑音抑制効果の改善を検討している [2]．

認識率が十分高くならないことの理由の一つとして，これらの前処理による音声特性の歪の影響も考えられる．歪の原因の一つは，ビームフォーミングによる音響フォーカスの位置と実際の音源位置とのずれであり，現在，サブスペース法による高精度な音源の 2 次元位置推定法を導入することにより，ビームフォーミングの精度の向上を検討している [3]．また，環境に存在する雑音は，競合話者などの方向性が強いものと，壁からの反射音による残響などのように方向性が低下しているものがあり，実際の環境では，これらが混在しているため，信号処理により，十分な雑音抑制ができない原因となっている．現在，逆フィルタとビームフォーミングとを組み合わせたハイブリッドなシステムを用いて，方向性と非方向性の雑音の両方に効果のあるシステムの搭載についても検討している [3, 4, 5]．しかし，特に，母音の残響など，低域優勢でなおかつ方向性が低下しているものについては，アレイ処理で抑制するのに限界があり，実際の環境における，反射 / 残響を含む音声データから HMM のモデルを作成する [30] など，よりきめ細かなチューニングを行う必要があるものと考えている．

4.2 柔軟な対話制御

対話の流れをどのように制御するかも大きな課題である．今回は，ユーザが話しそうなこととそれに対する応答を考えて，文法や応答ルールを作成しているが，この方式では，応答ルールやキーワードを熟知している設計者周辺の人以外が破綻せずに対話を継続することは不可能に近い．

Bernsen らは,SLDS (Spoken Language Dialogue System) における実際的な問題を整理している [10]．その中で，オフィスでのロボットの対話にとって重要と思われるのは以下の点である．

- システムの能力をユーザに知らしめる，
- ユーザのスキル／熟練度に応じた対話を行なう，
- ユーザのコマンドとシステムが実行しようとしている処理の確認を取る，
- 対話理解の結果を即座にフィードバックする，
- 会話の間違い（言い間違えや誤認識）の修正を可能にする．

さらに，我々の経験からは，対話の円滑さ，信頼性を増すために，次の点も重要であると思われる．

- 話題（タスク）の変化を許す，

- 会話を開始したり, 話題（タスク）を変更するためのキーワードを限定する.

前者は, たとえば所在表データベース登録の会話の中に, 「今日は何日」, 「ちょっと後ろに下がって」といった発話が挿入され得るからである. 後者は, 電話での対話や条件のよい環境での対話と異なり, オフィス対話で混入しやすい他者の声や雑音に対する耐性を増すためである.

しかし, 話題切替えのキーワード／入力文特性をどのように選ぶか, はむずかしい問題である. 現在の実装では, WAITING 状態から, 命令文なら移動タスクか呼び出しタスク, 平叙文ならデータベース登録タスク, 疑問文ならデータベース検索タスク, というような形で, ごく粗い制御をしているが, 今後, タスクの種類が増えた場合には問題が生じることが予想される. この問題に対して, マルチエージェントプログラミングにおける「競り」を用いた対話制御を検討している. それぞれのタスクごとに担当するエージェントを割り当て, 入力発話に対してそれぞれのエージェントが自分のタスクとの関連度を判断しつつ競りを行うことで, リアルタイムに柔軟なタスクの切替えを可能にしようというものである [32].

より自然な対話を実現するには, Wizard of OZ のような方式で対話データを収集し, それに基づいて対話のデザインをすることや, ユーザの不適切な応答を, システムが理解可能なものに自然に誘導するような仕掛けを組み込むこと, などが必要だろう. また, システム側が対話のイニシアチブを取り, ユーザの応答を絞り込むことができれば, ユーザの応答から予期されているキーワードだけをスポッティングすることが, ユーザの応答の揺らぎや, 音声認識の小さな誤りを許容するのにかなり有効である.

オントロジーに基づいた領域知識, 問題解決知識の組み込みも必要である [31]. 現在は, タスクフレームのスロット設計とタスク依頼をする関数の中にそうした知識が混ぜ込まれているが, 将来的には, オフィスイベント／タスクに関する領域／問題解決知識を別に持ち, 学習的に獲得される知識を参照しながらタスクフレームを動的に生成するようなことを考える必要があるかもしれない.

4.3 マルチモーダル情報の統合

多数のセンサ情報やアクチュエータを統合的に利用するような複雑な対話／タスク実行をどのように制御するかも, 大きな課題である. 対話が複雑になるにつれ, 詳細な応答ルールを記述することはかなり面倒な仕事になる. また, 音声入力以外のイベント, たとえばカメラやソナーなどのセンサから得られるイベントを対話の中に組み込み, 有効に利用することも必要である.

こうした問題への対応として, 秋葉らは, マルチモーダルな対話記述のためのスクリプト言語とインタプリタを提案している [1]. より少数の応答合成のための一般的なルール（テンプレート）と, 対話の進行度や効用の評価関数を用いて, 期待効用を最大化する経路を探索することで応答生成を行なうことも考えられる [16]. タスクフォームという generic なストラクチャを用い, タスク主導のスロット埋めという一般的な形で, タスク実行のために必要な情報を収集する対話パターンを記述しているのことは, こうした方向への第一歩と言える. さらに, ロボット全体のソフトウェアアーキテクチャを確率的な状況依存エージェントによってタスク主導, 状況依存的に実現することの検討も始めている [27].

これと関連して, よりきめ細かな文脈情報管理も今後の課題である. 現在の実装では, 文脈情報として, 深さ一定のスタックに使い終ったタスクフレームの情報をほぼそのまま保存しているが, センタリング理論のような談話の理論に基づいたよりきめ細かな文脈情報管理を用いる必要があるだろう [11, 33]. また, 音声入力以外のイベント, たとえば, ある物体が視覚的に認知された, というイベントの文脈情報への反映も重要である. 現在は, 顔認識結果を文脈情報に入れ, 「こんにちは」に対する応答生成に利用しているが, より一般的な仕組みが必要であろう.

謝 辞

本研究は通商産業省次世代情報処理基盤技術開発研究 (Real World Computing Project) の一環として電子技術総合研究所 RWI センターで行なわれたものです. 研究の機会を与えていただいた 大津展之 電総研 RWI センター長に感謝いたします.

参考文献

[1] 秋葉, 伊藤, 神嶌. マルチモーダル対話記述言語 MILES. 人工知能学会全国大会, 1998.

[2] 浅野, 速水, 松井. 話者方向同定と雑音抑制による音声認識の改善. 日本音響学会誌, Vol.53, No.11, 889–894, 1997.

[3] 浅野, 麻生, 松井. サブスペース法と空間逆フィルタを用いた音源分離. 信学技報, EA99-22, 1999.

[4] 浅野, 本村, 麻生, 松井. ブラインド信号分離における PCA フィルタの効果. 信学技報, DSP99-118, 1999.

[5] Asano, Hayamizu, Yamada and Nakamura. Speech enhancement based on the subspace method. IEEE Trans. Speech Audio Processing (in printing).

[6] H. Asoh, Y. Motomura, T. Matsui, S. Hayamizu, and I. Hara. Combining probabilistic map and dialogue for robust life-long office navigation. *Proceedings of IROS'96*, 807–812, 1996.

[7] H. Asoh, S. Hayamizu, I. Hara, Y. Motomura, S. Akaho and T. Matsui. Socially embedded learning of the office-conversant mobile robot, *Jijo-2*, *Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI'97)*, 880–885, 1997.

[8] H. Asoh, I. Hara, and T. Matsui. Dynamic structured multi-agent architecture for controlling office-conversant mobile robot. *Proceedings of 1998 IEEE International Conference on Robotics and Automation (ICRA'98)*, 1552–1557, 1998.

[9] H. Asoh, T. Matsui, J. Fry, F. Asano, and S. Hayamizu. A spoken dialog system for a mobile office robot. *Proceedings of 6th European Conference on Speech Communication and Technology (Eurospeech'99)*, 1139–1142, 1999.

[10] N. Bernsen, H. Dybkjoer, L. Dybkjoer. What should your speech system say? *IEEE COMPUTER*, **30**, 25–31, 1997.

[11] J. Fry, H. Asoh and T. Matsui. Natural Dialogue with the *Jijo-2* Office Robot. *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'98)*, 1278–1283, 1998.

[12] I. Hara, A. Zelinsky, T. Matsui, H. Asoh, T. Kurita, M. Tanaka, and K. Hotta. Communicative functions to support human robot cooperation. *Proceedings of International Conference on Intelligent Robots and Systems (IROS'99)*, 683–688, 1999.

[13] S. Hasimoto et.al. Humanoid robot – Development of an information assistant robot *Hadaly*, *Proceedings of the Sixth IEEE International Workshop on Robot and Human Communication*, 1997.

[14] K. Hotta, M. Kurita, and T. Mishima. Scale invariant face detection method using Higher-order Local Auto-Correlation features extracted from Log-Polar image. *Proceedings of International Conference on Automatic Face and Gesture Recognition*, 70–75, 1998.

[15] K. Hotta, M. Tanaka, T. Kurita, and T. Mishima. Multilevel ising search for human face detection. *Proceedings of SPIE Conference on Applications of Digital Image Processing*, 202–213, 1998.

[16] 乾 健太郎. 自然言語生成における相互依存制約の扱いに関する研究. 博士論文, 東京工業大学, 1995.

[17] K. Itou, S. Hayamizu, K. Tanaka, and H. Tanaka. System design, data collection and evaluation of a speech dialogue system. *IEICE Transactions on Information and Systems*, E76-D, 121–127, 1993.

[18] 事情通ロボット Home Page
http://www.etl.go.jp/~7440/

[19] T. Kurita, K. Hotta, and T. Mishima. Scale and rotation invariant recognition method using higher-order local autocorrelation features of log-polar image. *Proceedings of Third Asian Conference on Computer Vision (ACCV'98)*, Vol.II, pp.89–96, 1998.

[20] 松井, 速水, 麻生, 原, 本村. 所内事情通ロボットの計画. 1995 年度日本ロボット学会全国大会, 1995.

[21] T. Matsui and I. Hara. *EusLisp Reference Manual Ver.8.00*, ETL-TR-95-2, 1995.

[22] T. Matsui, H. Asoh, and I. Hara. An event-driven architecture for controlling behaviors of the office conversant mobile robot *Jijo-2*. *Proceedings of 1997 IEEE International Conference on Robotics and Automation (ICRA'97)*, 3367–3371, 1997.

[23] T. Matsui, H. Asoh, J. Fry, Y. Motomura, F. Asano, T. Kurita, I. Hara, and N. Otsu. Integrated natural spoken dialog system of *Jijo-2* mobile robot for office services, *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, 621–627, 1999.

[24] 松井, 麻生, Fry, 浅野, 本村, 原, 栗田, 速水, 山崎. オフィス移動ロボット *Jijo-2* の音声対話システム. 日本ロボット学会誌 (in printing).

[25] Y. Matsusaka, T. Tojo, S. Kubota, K. Furukawa, D. Tamiya, K. Hayata, Y. Nakano, T. Kobayashi. Multi-person conversation via multi-modal interface – A robot who communicate with multi-user –. *Proceedings of 6th European Conference on Speech Communication and Technology (Eurospeech'99)*, 1723–1726, 1999.

[26] 本村, 松井, 麻生, 浅野, 原, Fry. 事情通ロボットによるオフィス環境における知的作業支援. 人工知能学会 AI シンポジウム'98 SIG-J-9801-21, 1998.

[27] 本村, 原, 田中. 学習知能ロボットにおける状況依存エージェントの協調. 1999 マルチエージェントと協調計算ワークショップ (MACC'99), 1999.

[28] K.A. Papineni, S. Roukos, and R.T. Ward. Free-flow dialog management using forms. *Proceedings of 6th European Conference on Speech Communication and Technology (Eurospeech'99)*, 1411–1414, 1999.

[29] R100 Home Page
http://www.incx.nec.co.jp/robot/

[30] 清水, 梶田, 武田, 板倉. 空間音響特性依存 HMM によるスペースダイバーシチ型音声認識. 日本音響学会講演論文集, 37–38, 1999 年 9 月.

[31] H. Takeda, N. Kobayashi, Y. Matsumoto, and T. Nishida. Toward ubiquitus human-robot interaction. *Workshop Notes for IJCAI-97 Workshop on Intelligent Multimodal Systems*, 1–8, 1997.

[32] 田中, 本村, 橋田. 多重文脈に即応的な対話インターフェース: 半可通. 1999 マルチエージェントと協調計算ワークショップ (MACC'99), 1999.

[33] M. Walker, M. Iida, and S. Cote. Japanese discourse and the process of centering. *Computational Linguistics*, 20(2), 193–233, 1994.