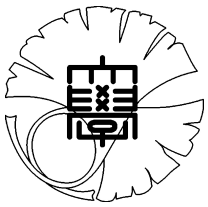
修士論文

番組検索音声対話システムにおける ユーザの知識を考慮した応答生成



2008 年 2 月 4 日

東京大学大学院
情報理工学系研究科
電子情報学専攻

篠田　知宏

内容梗概

音声認識技術や自然言語処理技術の進歩とともに機械との対話を行う音声対話システムへの期待は高まっている．一昔前のコンピュータは, 扱いやすさ, 処理能力, 価格など, 一般的な人達が使用できるような代物ではなかった．しかし現在ではコンピュータ技術の急速な進歩により, 驚異的な処理能力を持つコンピュータを一般家庭の人々が当たり前のように所有するようになってきた．このようなコンピュータの急速な進歩により, 以前は困難だった音声認識のリアルタイム処理が可能になり, IBM の ViaVoice[1] などの商用音声認識プログラムも登場し音声対話システムを構築する際の音声認識の敷居を低くした．

一方, 社会の情報化が進展し, 家庭内でのインターネット等を介した情報収集が一般的になっている．既に, テレビジョンなど, 家電製品が情報機器化し, それらのネットワーク化が進んでいる．未来住宅では, これがさらに発展し, ユーザは機器を意識せずに, 操作し得ることが求められる．このためには, 人間の最も根源的な情報伝達媒体である音声による情報検索や情報授受の確立が必要である．音声を情報授受の媒体として用いることの利点としては,「特別な練習を必要としない」,「対話中に手を自由に使える（マルチモーダルインターフェース）」,「情報伝達速度が速い」,「音声しか用いることの出来ない状況でも操作が行なえる」などが挙げられる．

このような背景もあり, 音声対話システムはめざましい進歩を遂げているが, そうしたシステムを実際に利用しているユーザは多くはいない．その理由として, 現在の音声対話システムでは, 人間同士のような柔軟な対話が困難なことが挙げられる．これまで構築された一般的な音声対話システムでは, 話題とするトピックが変わらない限り, ユーザの同じ発話には, （省略, 照応のレベルの違いはあるものの）基本的に同じ応答が生成される．しかしながら, 同じ質問であっても, 状況によって, ユーザの求めている情報は異なり, またユーザの情報の理解力は, ユーザの知識により異なってくる．このため, ユーザが真に求めている情報を, ユーザが理解しやすいように提示するという観点から, 音声対話システムには, 画一的でない応答を生成する能力が求められる．ユーザの情報を利用し, それに即した応答を生成する研究は存在するが, ユーザの知識レベルを判定して応答を動的に変化させるものではない．ここでは, トピックに対する, ユーザの知識の静的なレベル, 動的なレベルを判定し, 判定された知識レベルに応じて適切な応答を生成することを提案し, システムの実装を進めた結果について報告する．

目次

図目次

表目次

第1章

序論

1.1　本論文の背景

音声認識技術や自然言語処理技術の進歩とともに機械との対話を行う音声対話システムへの期待は高まっている. 一昔前のコンピュータは, 扱いやすさ, 処理能力, 価格など, 一般的な人達が使用できるような代物ではなかった. しかし現在ではコンピュータ技術の急速な進歩により, 驚異的な処理能力を持つコンピュータを一般家庭の人々が当たり前のように所有するようになってきた. このようなコンピュータの急速な進歩により, 以前は困難だった音声認識のリアルタイム処理が可能になり, IBM の ViaVoice[1] などの商用音声認識プログラムも登場し音声対話システムを構築する際の音声認識の敷居を低くした.

一方, 社会の情報化が進展し, 家庭内でのインターネット等を介した情報収集が一般的になっている. 既に, テレビジョンなど, 家電製品が情報機器化し, それらのネットワーク化が進んでいる. 未来住宅では, これがさらに発展し, ユーザは機器を意識せずに, 操作し得ることが求められる. このためには, 人間の最も根源的な情報伝達媒体である音声による情報検索や情報授受の確立が必要である. 音声を情報授受の媒体として用いることの利点としては, 「特別な練習を必要としない」, 「対話中に手を自由に使える（マルチモーダルインターフェース)」, 「情報伝達速度が速い」, 「音声しか用いることの出来ない状況でも操作が行なえる」などが挙げられる.

1.2　本論文の目的

音声対話システムはめざましい進歩を遂げているが, そうしたシステムを実際に利用しているユーザは多くはいない. その理由として, 現在の音声対話システムでは, 人間同士のような柔軟な対話が困難なことが挙げられる. これまで構築された一般的な音声対話システムでは, 話題とするトピックが変わらない限り, ユーザの同じ発話には, （省略, 照応のレベルの違いはあるものの）基本的に同じ応答が生成される. しかしながら, 同じ質問であっても, 状況によって, ユーザの求めている情報は異なり, またユーザの情報の理解力は, ユーザの知識により異なってくる. このため, ユーザが真に求めている情報を, ユーザが理解しやすいように提示するという観点から, 音声対話システムには, 画一的でない応答を生成する能力が求められる. ユーザの情報を利用し, それに即した応答を生成する研究には, 大人か子供かによって応答を変化させる「たけまるくん」[2] などがあるが, ユーザの知識レベルを判定して応答を動的に変化させるものではない. ここでは, トピックに対する, ユーザの知識の静的なレベル, 動的なレベルを判定し, 判定された知識レベルに応じて適切な応答を生成することを提案し, システムの実装を進めた結果について報告する.

1.3　本論文の構成

本論文は，以下のように 5 つの章より構成される.

第 1 章（本章）では，本論文の背景・目的などを述べている．第 2 章では，一般的な音声対話システムについて述べる．また，対話システムにおいて重要となる各要素技術につ

いても解説する．第 3 章では，音声対話システムを構築する際に重要となる, 対話の主導権, 対話モデルについて述べた後，対話記述言語に関しても説明を行う. 第 4 章では，本研究における対話処理, 及び応答生成をテレビ番組情報の検索を行う音声対話システムに実装する. 第 5 章では, 本論文をまとめ, 今後の課題について述べる.

以降，次章より本論を進めていくこととする.

第2章

音声対話システム

2.1　はじめに

音声対話システムとは，音声対話を行ないながらユーザと共同でタスクを実行するシステムである．1990 年代に入って，音声を登録せずに任意の単語を認識できるようになり，また自由発話の中から単語を識別できるようになった．さらに，ソフトウェアだけで音声認識が実現可能となった．このような音声認識技術，言語処理技術，さらには音声合成技術の向上に伴い，これらの技術の実用化が検討され始めた．実用化の検討に際して，これらの要素技術を統合して実現される音声対話システムは格好の研究材料である．そして，いくつかの音声対話システムはすでに実用化されている [3][4]．本章では，一般的な対話システムについて，例を交えながら説明を行なう．また，ソフトウェアロボットによるエージェント対話システムについて説明する．さらに，音声対話システムにおける必要技術について述べる．

2.2　一般的な対話システム

2.2.1　システム概要

音声対話システムとは音声コミュニケーションを行いながらタスクを実行するシステムである．

音声コミュニケーションは多くの要素が複雑に絡み合って成立するものだが, 主な処理の流れは以下のようになる．

1. まず耳で聞く．計算機で処理する場合は音を文章（テキスト）に変換する音声認識部 (Speech recognizer) がこれに相当する．このとき, 不必要な雑音などは除去される必要がある．
2. 次に脳で考える．つまり, テキストの中身を解釈し, 応答を作成する．言語理解 (Language interpreter), 対話制御 (Dialog manager), 言語生成 (Sentence generator) がこれを担う．
3. 最後に口で話す．作成された応答（テキスト）を音声に変換する音声合成技術 (Speech synthesizer) がこれを担う．

言葉と同様に, 目を使った状況把握や会話相手の感情認識, 表情や発話音声の調子を変化させることによる感情の表現などの非音声情報も音声コミュニケーションの重要な要素である．主な計算機での流れは図 2.1 のようになる．

2.2.2　音声認識部

音声認識技術は, その認識対象とする単語辞書の大きさによって小語彙（数百語程度）と大語彙（数千語以上, 数万語程度）に区分することができる．

一般的に, タスクを限定することによって音声認識での認識の対象とする語彙を抑えることはある程度可能である．その場合人手で文法を記述し, タスクに応じた語彙を登録する

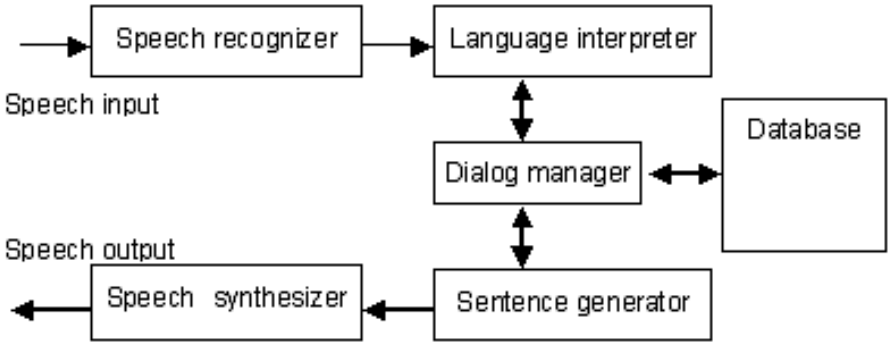


図 2.1: 一般的な対話システムの構成

ことによって認識を実現する.

しかし, 人間の発話表現は多種多様であり, ユーザーに小語彙を要求するのは厳しいものがある. また, 現在の音声認識技術では辞書に登録されていない発話があると, 認識不可能な未知語が認識誤りを起こすのみではなく, その前後にまで影響が派生することが知られている. 結果として, 辞書に含まれる単語で構成される誤った文章が出力され, 認識後の対話処理に悪影響を与えかねない.

よって近年では大語彙連続音声認識システムが多く使われるようになってきている. 例えば, 2 万語の単語辞書を用いることで 75 か月分の新聞記事に出現する全単語の約 97% を被覆することができる.

大語彙連続音声認識システムでは, 音声認識の際に音響モデル, 単語辞書, 言語モデルを用いる. 音響モデルとは, 音素の並びを統計的に学習したものであり, HMM(Hidden Markov Model) がよく用いられる. HMM は特徴ベクトル時系列の確率モデルであり, 自己遷移を持つ複数の状態間を遷移することで, 音声のような長さの一定しない時系列信号を効率良くモデル化することが可能である. HMM の学習には EM（Expectation Maximization）アルゴリズムと呼ばれる最尤パラメータ推定手法が用いられる.

音響モデルが話者性や音声入力環境などの音声認識における音響的特徴を担うものであるに対し, 言語モデルと単語辞書は, 言い回しなどの文章表現や認識対象単語などの言語的特徴を定めるものである. 統計的言語モデルを用いた大語彙連続音声認識では, 認識結果を開発者があらかじめ決定的に定義することは難しく, 一見するとアプリケーションに組み込むのには向かない. しかし, 柔軟に様々な発話を受理することが可能であるため利用する価値は高い.

2.2.3　言語理解部

言語理解部では音声認識部から受け取った文の意味構造を解析し, 対話制御部が理解できる形にして対話制御部に送る. 単語列の文法的な構造を解析して品詞を同定し, 意味表現（文の意味を論理式などで曖昧性なく表現したもの）に変換する.

2.2.4　対話制御部

対話制御部ではユーザの意図を理解し, データベースを参照してユーザへの返答を生成する. そして, それを意味表現の形で言語生成部に送る. 言語理解部から受け取った意味構造からユーザが何を尋ねているのかを判断し, データベースから適切なデータを引き出し, それを再び意味構造に落とし込む. このとき, ユーザの発話が不完全でデータの検索を行えない場合はユーザに再入力を促すなど, 対話の全体的な制御を行う.

2.2.5　言語生成部

言語生成部では対話制御部から受け取った意味表現を文の形に変換する.

2.2.6　音声合成部

言語生成部より受け取った文を音声信号の形にして出力する. 初期は TTS（Text-To-Speech）システムがよく用いられていたが, 最近では CTS(Concept-To-Speech)[5] なども用いられる. 対話システムにおいてはより高度な言語情報を韻律に反映させることのできる CTS 方式は注目されている.

2.2.7　音声対話システムの特徴

音声対話システムは，音声認識技術・音声合成技術・自然言語処理技術の集大成であり，さまざまなアプリケーションに適用できるものと考えられる.

音声対話は即興的に行なわれるので，文字言語に比べて誤り，曖昧さ，省略，語順変更などの不確実さや重複が多くなる. 従って，音声対話においては，音声認識と言語理解を密接に結びつける必要がある.

2.3　音声対話システムのアーキテクチャ

音声対話システムにおいて複数のモジュールを協調して動作させるためには, 全体を制御する何らかの方式が必要となる. この方式をアーキテクチャと呼ぶ. ここでは, これまで研究レベルの音声対話システムで利用されてきた二つのアーキテクチャと現在標準化が進められている実用的なマルチモーダル対話システムのアーキテクチャを説明する.

2.3.1　GALAXY アーキテクチャ

i)　概要

Galaxy[6] は, ハブ・スポーク構造を持つ, クライアント／サーバ方式のアーキテクチャである. MIT で開発され, 現在はオープンソースプロジェクトとして MITRE が配布を行っている. [7] 音声対話システムの各モジュールがサーバとして機能し, ハブがメッセージの

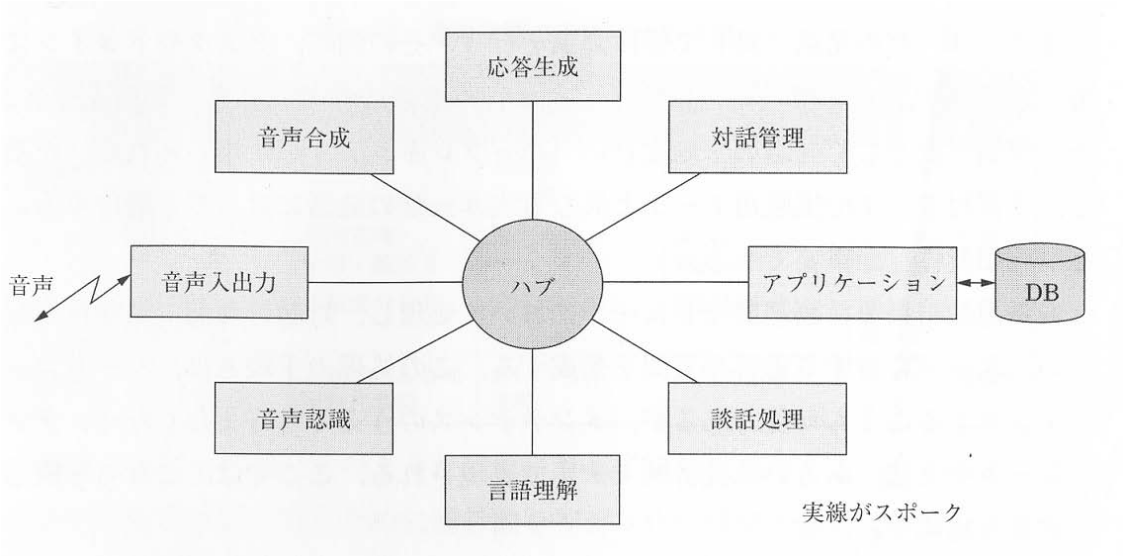


図 2.2: Galaxy アーキテクチャ

受け渡しおよび全体の制御の機能を有する. このアーキテクチャによる典型的なシステムの構成例を図 2.2 に示す.

ハブとモジュールの間の情報伝達はフレームを用いて行われる. ハブにおいて, どのようなフレームを受け取ったら, 次にどのようなモジュールを呼び出すかは規則の形式で記述されている.

このアーキテクチャでは, 各サーバはハブとのインタフェース仕様に従って実装すれば差し替えが可能である. これによって, 多言語化やタスク拡張に柔軟に対応できるアーキテクチャとなっている.

2.3.2　オープンエージェントアーキテクチャ(OAA)

OAA(Open Agent Architecture)[8] は自律分散環境で動作するマルチエージェントシステムを構築するフレームワークである. ここでのエージェントの定義は, 共通のインタフェースを備えた自立的なソフトウェアプロセスである. Galaxy アーキテクチャにおけるサーバが, OAA におけるエージェントに相当する. 相違点は, OAA におけるエージェントの方が自律性が高く, 複数のエージェントの並列動作や実行時の動的な結合などを可能としている点である. その代償として, エージェントを制御する言語が複雑になっている.

Galaxy アーキテクチャのハブ (クライアント) に相当する役割を果たすエージェントをファシリテータと呼ぶ. また, スポーク (サーバ) に相当するエージェントをクライアントと呼ぶ. ファシリテータはメッセージの仲介だけではなく, クライアントから受け取った要求を分割し, 解決可能なクライアントを探すという, よりエージェント指向に近い問題解決機能を持つ. クライアントエージェントは, 個別のサービス (音声認識・言語理解・DB アクセスなど) を提供するアプリケーションエージェント, ドメイン依存の制御情報を用いてファ

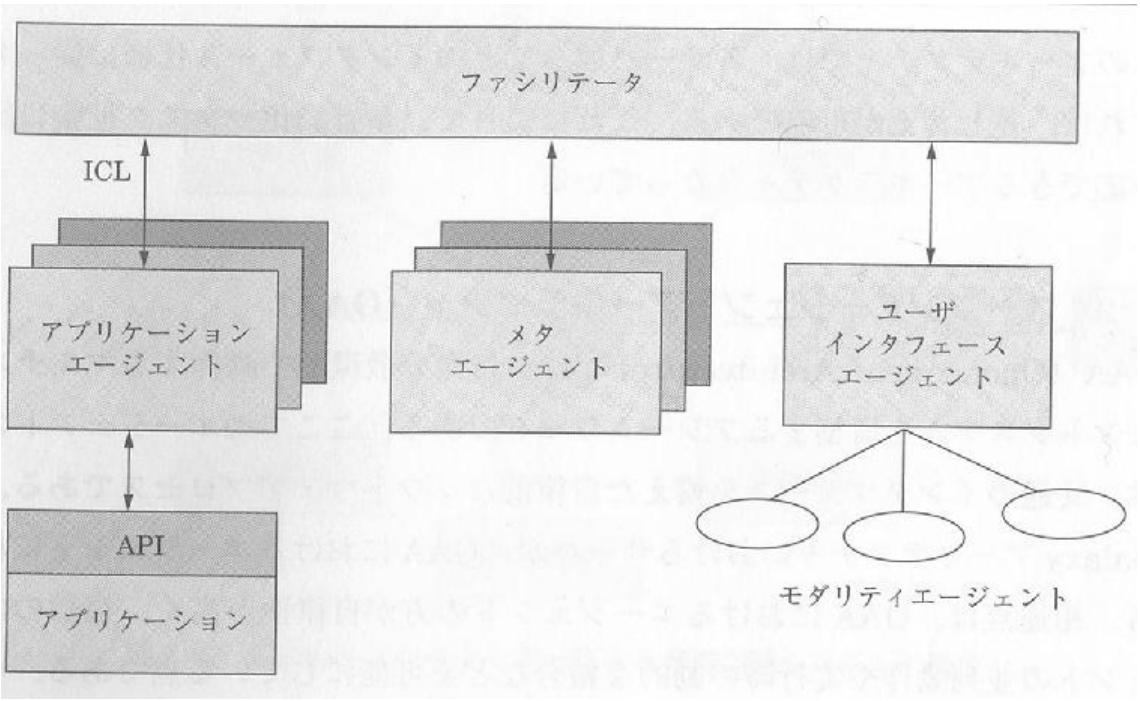


図 2.3: OAA のシステム構成

シリテータを補助するメタエージェント, ユーザとのインタラクションを担当するユーザインタフェースエージェントから成る. (図 2.3). さらにユーザインターフェースエージェントは下位にマイクロエージェントを持ち, 各マイクロエージェントは音声・ペン入力・キー入力などのイベントを監視し, 相互に協力して意図の解釈を行う.

2.3.3　W3C マルチモーダルインタラクションアーキテクチャ

Web 技術の標準化団体である W3C[9] では, アクセシビリティ向上やモバイル環境での利用を想定して, マルチモーダルインタラクション (MMI) に関連する技術の標準化も行っている. このような標準化活動は通常, 記述言語の標準化が主要な目標となることが多いが, MMI 標準化活動の場合は, 記述言語に先立ってアーキテクチャが提案されている.

このアーキテクチャ設計の目的は, モダリティに依存したさまざまなコンポーネント (音声認識・文字認識など) 間の相互結合性を保証し, 柔軟なフレームワークを提供することである.

フレームワークは以下の指針を基本設計目標としている.

- カプセル化... コンポーネントはブラックボックスとして扱う
- 分散環境... ネットワークで結合された複数の機器で構成可能
- 拡張性... 新たなモダリティ要素を容易に統合できる
- 再帰性... コンポーネントの集合を新たなコンポーネントとして定義できる
- モジュール化... アーキテクチャはデータ・制御・表示を分離する.

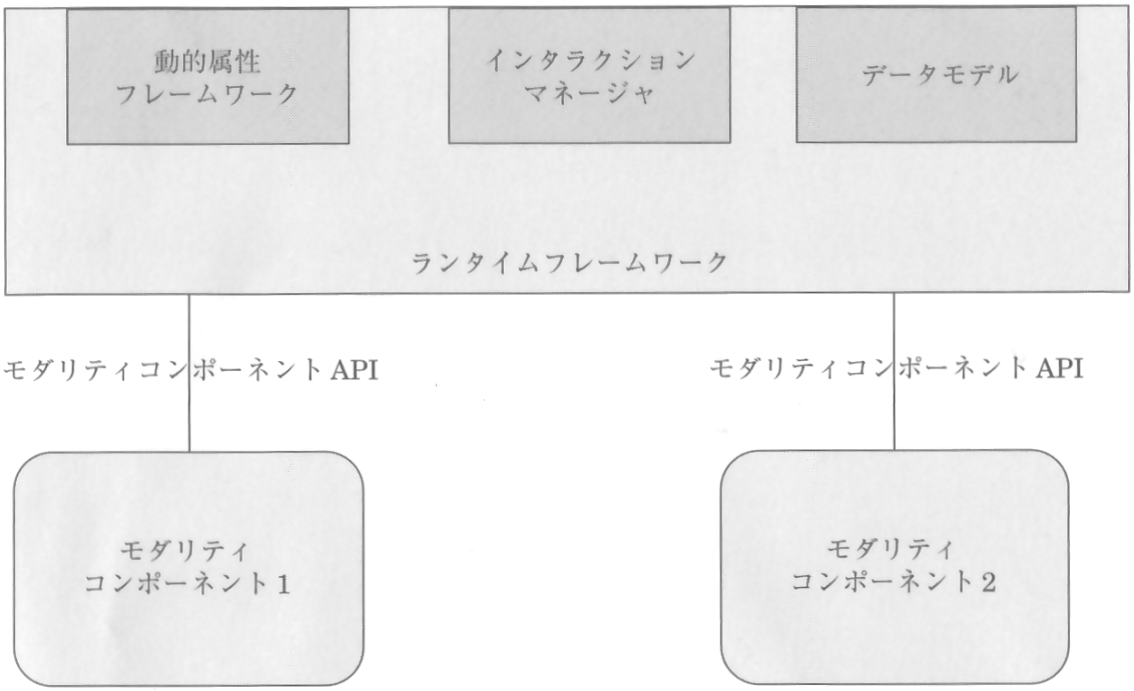


図 2.4: W3C の MMI アーキテクチャ

W3C の MMI アーキテクチャを図 2.4 に示す.

各モダリティコンポーネントはそれぞれ独立の記述言語でその動作が規定されていてもよい. 例えば, GUI モダリティであれば XHTML. 音声モダリティであれば VoiceXML や SALT などが用いられる. データモデルとしては HTML のフォームを一般化した XForms が想定されている. インタラクションマネージャは, 他のコンポーネントからのイベントを処理する役割を果たし, その記述言語としては状態遷移が記述可能な CCXML[10] や SCXML[11] などがある.

コンポーネント間の通信はイベントのみで定義される. したがって, 各コンポーネントは仕様で定義されたイベント生成とイベントハンドラを実装すればよく, 他のコンポーネントの動作に影響されない.

このように W3C のアーキテクチャは既存の技術を最小限の変更で利用できるように考慮されたもので, Galaxy アーキテクチャや OAA に見られるような柔軟性はあまりない. これは仕様をユーザとのインタラクション部分に限定しているためで, 高いレベルの処理はサーバサイドの動的な処理で実装されることを想定している.

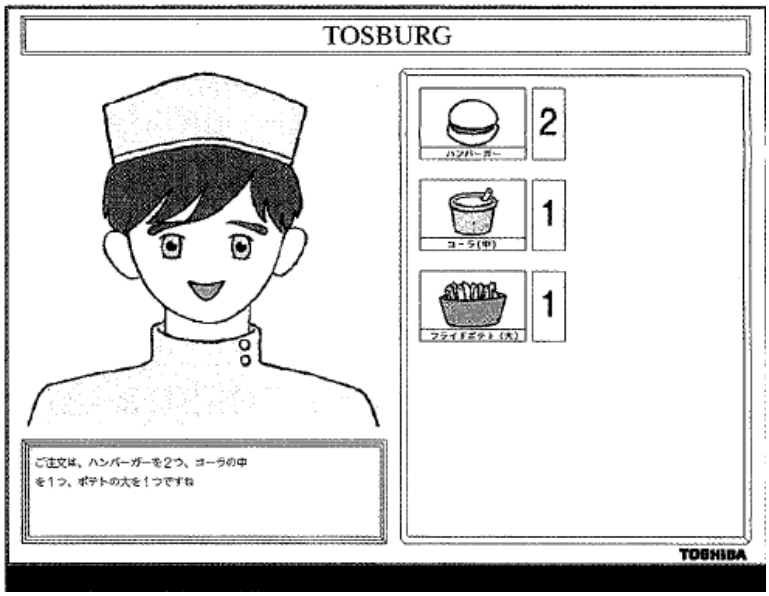


図 2.5: TOSBURG　system

2.4　音声認識に関する研究

2.4.1　小語彙音声認識システム:TOSBURG

i)　概要

東芝研究開発センターの竹林らは音声自由対話システム TOSBURG　(Task-Oriented dialogue System Based on speech Understanding and Response Generation)[12] を開発した. 図 2.5 にその応答画面を示す. これはハンバーガーショップでの注文をタスクとしたマルチモーダルインターフェースの音声自由対話システムである.

このシステムの特徴は, 小語彙の音声認識技術を用いているにも関わらず, ユーザーの自由な発話に対応している点である. 要素技術として雑音や自由発話に対するロバスト性を高めるために, 雑音免疫ワードスポッティングを用いている. また, 音声応答のキャンセルを行うことによりテンポのよい対話を可能としている.

ii)　雑音免疫ワードスポッティング

ワードスポッティングとは登録された特定の語だけに反応して, 登録されていない音声はスルーする方法である. この方法は, 計算量は多いが音声区間検出が不要でありロバスト性も向上させることができる反面細かい対応はできない. 実際に TOSBURG では 49 語のキーワードを用いてワードスポッティングを行っている.

また, 環境騒音や特に意味を持たない発声に対する耐性を強化するため雑音免疫学習を行っている. 雑音免疫学習では, 雑音を除去するのではなく雑音を徐々に高めて学習を行い, ワードスポッティングの際の耐雑音性の向上を図っている.

iii)　音声応答キャンセル

音声応答のキャンセルとして騒音能動制御技術を用いている．これはマイクロホンに入力されたユーザの入力音声とシステムの応答音声が混合した信号中から音声応答の成分を常に引き去ることによって音声応答を行いながらユーザーの音声認識を行うことができる．音声応答キャンセルを行う際にはスピーカーからマイクロホンに直接届く成分以外に, 周囲の物体に反射して届く成分も考慮する必要があるため, リアルタイムにスピーカからマイクロホンへの伝達特性の推定を行う必要がある．

2.4.2　大語彙音声認識システム

奈良先端科学技術大学の西村らは受付案内ロボット「ASKA」[13], 音声情報案内システム「たけまるくん」[2] などの大語彙連続音声認識を用いた音声対話システムを開発した．これらのシステムは半自動に言語モデルを構築でき, 開発のコスト削減に貢献するものであり, その手順は,

1. コーパスの自動作成とトピック依存 N-gram モデルの構築
2. モデル融合でのタスク操作
3. 文法の適用によるモデル高精度化

の三段階から構成される．

i)　ASKA

ASKA は奈良先端科学技術大学院大学情報科学研究科の複数の研究室が合同で開発した人との対話機能を持つ人型の受付案内ロボットである．ここではその多くの要素技術のうち音声インターフェースを中心に紹介する．想定されるタスクは来客の案内であり, 次のような質問に応対できるように設計されている．

- 教官および研究室の場所と内線番号
- 学内及び周辺の施設
- ASKA 自身に関する事柄
- その他, いくつかの挨拶

実際の対話処理は以下のような流れで行われる．

1. 目の中の CCD カメラを用いたステレオ画像処理によって, ASKA の前に立つ発話者を検知する
2. 検知の後, 音声認識理解部が音声の入力を開始する．同時に顔をユーザに向け, 質問の受け付けが可能な状態であることを示す．
3. ユーザが ASKA に質問をする（音声入力）．入力には ASKA の前の据え置き型のマイクロホンを用いる．
4. 音声認識理解部が入力音声に対する応答文を作成, 結果をサーバに送信する．

（人）こんにちは.
（ASKA）こんにちは.
（人）公衆電話はどこですか？

（ASKA）公衆電話は私の後ろにあります.
（人）内線電話をかけたいんですが.
（ASKA）内線電話はこちらです.
（人）写真を撮ってもいいですか？

（ASKA）ヤメテクダサイ！

図 2.6: ASKA の対話例

5. 音声合成部は, 応答文から TTS（Text To Speech）プログラムにより合成音声を作成, 発話待ちの状態で待機する.
6. 胴体と頭のジェスチャ部は, 応答文等の必要なパラメータがサーバに蓄えられたことを検出, ジェスチャの定義パターンに基づいて動作を開始する.
7. 音声合成部は, ジェスチャと同時に発話を開始する.
8. すべてが終了すると, ユーザからの発話待ち状態に戻る.

実際の対話例を図 2.6 に示す.

音声認識エンジンは言語モデルを用いた Julius[14, 15] と記述文法を用いた Julian[16] が併用されている. 併用することによってそれぞれ単独で用いた場合よりも精度を高めている. 言語モデルとしては学生に対して行ったアンケート結果である ASKA への質問文テキスト（769 文, 総単語 10k 個, 異なり単語 1.1k 個）, Web 検索エンジンを用いて収集した奈良先端科学技術大学院大学の関連 Web ページテキスト（22,078 文, 総単語 600k 個, 異なり単語 26.8k 個）, 過去約 2 年間の学内学生連絡用メーリングリストに流れたメールテキスト（8,183 文, 総単語 253k 個, 異なり単語 9.5k 個）をコーパス結合した状態から学習した単語 N-gram モデルを使用している. これらの融合言語モデルにネットワーク記述文法の単語間制約を適用して N-gram 確率を強化した文法適用言語モデルを作成して用いている. 音響モデルとしては日本音響学会新聞記事読み上げ音声コーパス（JNAS）[17] から学習した性別依存モデルを使用している.

言語理解部にはキーワードスポッティングを用いている. 音声認識結果の形態素と用意されている用例テキストの形態素とのキーワードの一致回数を数え上げ, その結果を対話制御部に送る. この時応答生成プログラムの入力である音声認識結果には, N-best 出力結果を用い, 正解形態素が 1-best 結果に出現しない場合も 2-best 以下を参照できる.

言語理解部から渡されたデータからあらかじめ用意された来客の質問に対する応答文の候補を選択する. この応答候補は, ASKA に答えてほしい事項に関する学内アンケートを

```
100 おはようございます.
103 ご用件はなんでしょうか？
200 私の名前は，アスカです.
204 施設の案内や，先生方のお部屋の案内ができます.
302 <is-staff:3>先生の部屋は，<is-staff:5>です.
303 <is-staff:3>先生の内線番号は，<is-staff:8>です.
404 内線電話は，こちらにあります.
405 公衆電話は，私の後ろにあります.
415 バス停は，そこの玄関を出てまっすぐ道路へ出て左側にあります.



073 李晃伸リアキノブ B613 B 6 5282 音情報処理学鹿野オトジョウホウショリガクコウザシカノケン
```

図 2.7: 応答文とスロットに挿入するデータ

実施し, その結果から必要性が高いと思われる質問を選びだし, その質問に応じた応答を用意する.

応答文の候補には, 定型なものと他のデータベースからデータを検索してスロットに挿入できるもの（スロット型）の 2 種類がある. 図 2.7 に応答文候補とスロットに挿入するデータの例を示す.

ii)　たけまるくん

たけまるくんは ASKA と同様西村らによって作られた, 奈良県生駒市のコミュニティセンターにて音声情報案内を行うシステムであり, 北コミュニティセンターの館内施設や生駒市の観光情報, 周辺情報などの各種案内を行うものである. システムは同センター内に常設され, 一般の来訪者がいつでも気軽に利用できる音声インタフェースによる館内案内サービスを提供することを目指している. （図 2.8）また, 長期間フィールドテストを通じてシステムの利用実態を調査し, 同時にインタラクションの観察に必要な発話の大規模収集も同時に行っている. 音声認識部は基本的には ASKA と同じであり, Julius を用いるが, 言語モデルとして以下のテキストから 2-gram および逆向き 3-gram を作成している.

1. Web 検索を用いて収集した生駒市関連及び生駒市ホームページ内の Web ページテキスト 1,080,272 文, 総単語 31,265k 個, 異なり単語 218.7k 個
2. 人手で収集した本システムを想定した質問文テキスト 6,488 文, 総単語 56k 個, 異なり単語 3.2k 個

音響モデルには, 日本音響学会新聞記事読み上げ音声コーパス（JNAS） のクリーン音声に 25dB SNR で電子協騒音データベースの展示会場の雑音を重畳した音声から学習した性別非依存 PTM モデルを使用している.

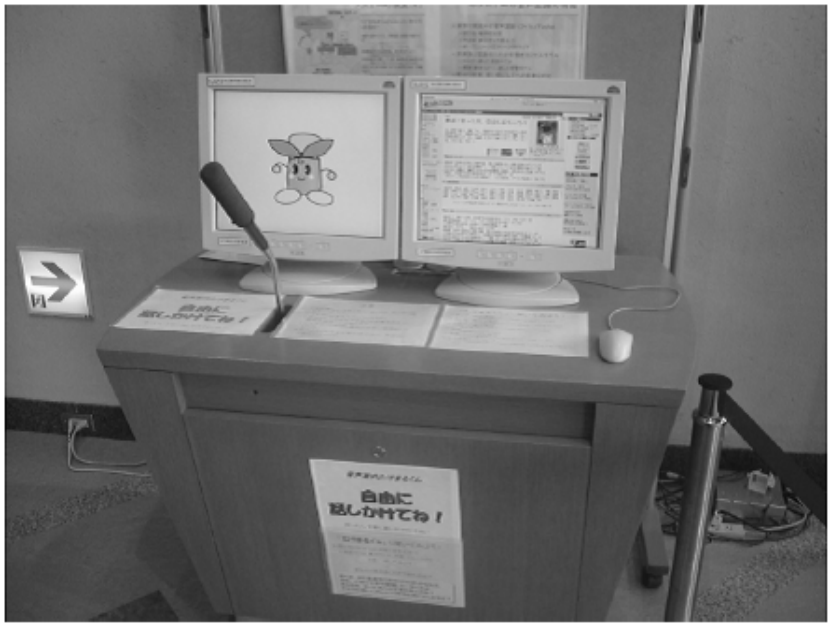


図 2.8: たけまるくんシステム

実際に運用した結果, 子供の発話への対応などが必要だということなどが課題として挙げられている.

2.5　対話管理に関する研究

音声対話システムの構築にあたって最も核となる要素技術が対話管理であり，様々な側面から研究が進められている.

対話管理の根幹となる部分は，言語処理である．しかし，音声対話システムというものが，言語処理だけでなく，音声認識や音声合成といったものが統合して実現されるものであるため，言語処理単体での研究，というよりは，音声認識や音声合成と一体となって進められている研究が多い.

対話管理に着目した対話システムとして，本節では，相槌を打ったり，ユーザの割り込みに対しても適切な対応を行なう，といった，人と計算機の円滑な対話の実現を目的とした研究例を紹介する.

2.5.1　飛遊夢（ひゅうむ）

i)　システム概要

NTT コミュニケーション基礎科学研究所は，マルチモーダルインタフェースを備えたエージェントである「飛遊夢」[18] を開発した．システム構成を図 2.9 に示す．このシステムは，ユーザとの対話を通して気象情報案内を行なうシステムである.

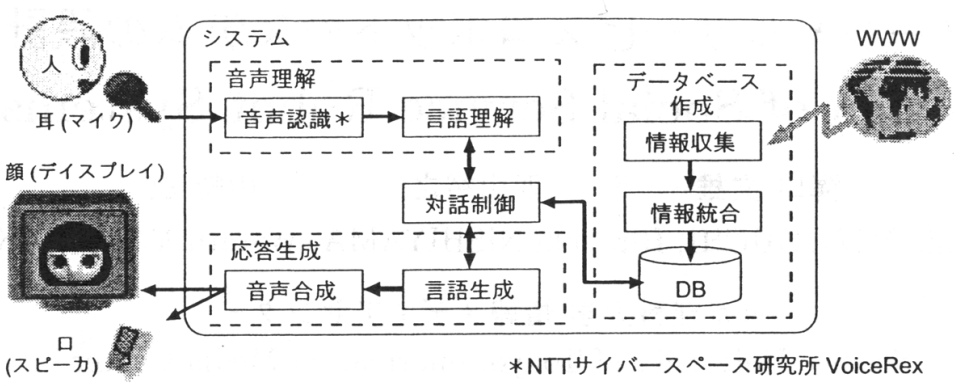


図 2.9: 「飛遊夢」システム構成図

ii)　音声理解部

音声理解部は，音声認識部と言語理解部から成る．

従来の音声認識は，認識単位が文であったため，音声入力文が終わってから認識を行なう必要があった．これでは理解が遅れ，円滑な対話に必要とされる適切な相槌を打つことができないなどの問題があった．

音声認識部においては，NTT サイバースペース研究所が開発した，不特定話者の連続音声認識器 VoiceRex[19] を用いている．これは，音声認識結果を逐次出力することができるのが特徴である．言語理解部では，逐次理解法 [20] によりユーザ発話を理解する．これは，ユーザが発話を終了する前に理解を開始する方法である．また，複数文脈を用いたビームサーチによる音声理解を行なうことで，文節などの短い単位で理解することを可能としている．このようにすることで，上記の問題を解決している．

iii)　対話制御部

対話制御部は，2 つのフェーズを持つ．

ユーザ要求確定フェーズにおいては，システムが相槌発話を行なうか，ユーザ発話理解結果を確認するための確認発話を行なうか，ユーザに情報を要求する要求発話を行なうかのいずれかを決定する．

システム情報提供フェーズにおいては，確定したユーザ要求内容に応じてシステム発話内容を設定する．

iv)　応答生成部

応答生成部は，音声合成部と言語生成部から成る．音声生成部は，前もって文節単位で録音した人の音声を再生する録音編集方式（波形編集方式）を採っている．

ユーザ要求確定フェーズでは，言語生成部は対話制御部の決定に従ってシステム発話を生成する．

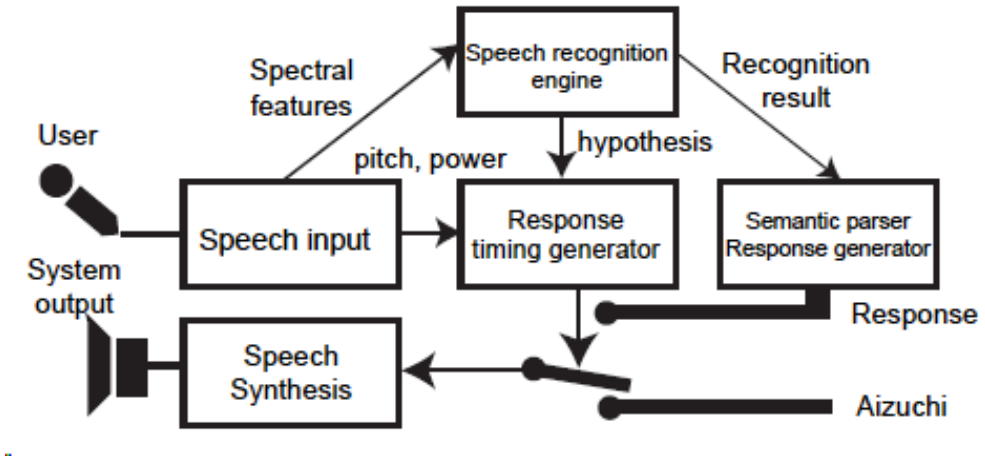


図 2.10: 雑談対話システム

システム情報提供フェーズでは，言語生成部は発話内容を伝達するための応答文を逐次生成法 [21] により生成する．これは，ユーザに伝達済みの情報を逐次管理しながら発話を生成する方法であり，システムが発話している途中にユーザが割り込むと，その時点で伝達済みの情報と照合することによりユーザ意図を理解する．ユーザが話の進め方を変更する意図を持っているなら，対話制御部は言語生成部に発話を中断することを命じ，ユーザ意図に合致するようにシステム応答文を変更した上で発話を再開する．

2.5.2　雑談対話システム

i)　システム概要

竹内らは雑談対話システム [22] において，自然な雑談対話をする上で最も重要であるタイミング生成，すなわち，相槌，割り込みのタイミングの判定を，人間同士の対話の特徴から学習した決定木を用いて行なっている．この決定木は，韻律情報と言語情報を素性として用いているが，このうち韻律情報は発話句音声末のおよそ 100ms の変動，言語情報は発話終端単語の品詞や発話の最後に現れた自立語の品詞情報などを考えている．システム構成を図 2.10 に示す．

ii)　決定木の構築

決定木の構築に先立ち，人間の対話の分析を行なう必要がある．この研究では，人工知能学会コーパス利用研究グループの談話タグつき対話コーパス（全 29 対話）[23] のうち，雑談タスク，旅行案内タスク，テレフォンオペレータとの対話の 3 タスク 11 対話（全 1842 発話）をトレーニングとテストに用いている．対話コーパスの分析を，韻律情報と言語情報に分けて行なう．

この分析を元に，決定木を用いて話者交替，相槌のタイミングを検出することを行なう．

決定木の生成には，与えられた学習データで初期決定木を構築し，その後枝刈りを行なう帰納学習システム C4.5[24] を用いている．

このようにして構築した決定木に現れた主要な素性は，発話長，ピッチやパワーの変動といった韻律情報がほとんどであった．言語情報の主要な素性は，発話中の最後の自立語のみであった．また，これはタスクの違いによらないものであったことから，相槌や話者交替の検出には，韻律情報に加え自立語であるという判定などの表層的な言語情報が大きく関係している，といえる．

この結果を，天気案内をタスクとした雑談対話システムに応用し，被験者（4 名）との対話による評価実験を行なったところ，システムの返答内容に関しては改良が必要であるものの，相槌タイミング自体はよい，という評価を得ている．

2.6　音声合成に関する研究

従来，音声合成研究においては，音声によるテキストの流暢な読み上げを目標とした研究が精力的に進められ [25]，現在多数の高品質のテキスト音声合成（Text-To-Speech）システムが市販されるまでになっており [26, 27]，最近までは，ほとんどの音声対話システムの音声出力がこの TTS システムによって合成されている．

しかしながら，この TTS システムとは，一般のテキストから音声を生成することを目的としたものであり，高次の言語情報を反映した音声合成を想定していない，という問題点がある．音声対話システムでは，応答文がシステムにより生成されるため，統語構造や談話情報などの高次の言語情報を得ることができる．そのため，これらを応答音声に反映させることのできる音声合成の枠組み，すなわち概念音声合成（Concept-To-Speech）[5] の実現が望ましいと考えられる．

これまでに述べた音声認識・対話管理に関する研究と比べると，音声合成を主眼とした研究例は少ないのが現状である．この理由としては，合成音声の評価が主観的なものにならざるを得ない，という理由が最も大きいと考えられる．本節では，特に韻律の制御に着目した応答文生成を考慮している対話システムについての概要を述べる．

2.6.1　談話情報を用いた音声合成における韻律の制御

i)　システム概要

遠山らは，対話データベースから特に対人態度に関わる談話情報を抽出し，発言ごとに談話情報のタグセットを用意することで，特徴的な音声出力を行なうシステムを提案した．[28] システムの概要を図 2.11 に示す．

システムは，複数のユーザがある話題について述べたデータベースを保有している．保有されたデータからまず個人情報を獲得する．これにより，ユーザの話題ごとの知識や興味関心を獲得できる．また同時に談話情報を獲得する．この研究における「談話情報」とは，文章同士の構造や関連性といった言語概念的なものではなく，対話に関連する対人態度や心理，感情といったパラ言語的概念によるものである．文章に現れない意図，ユーザ

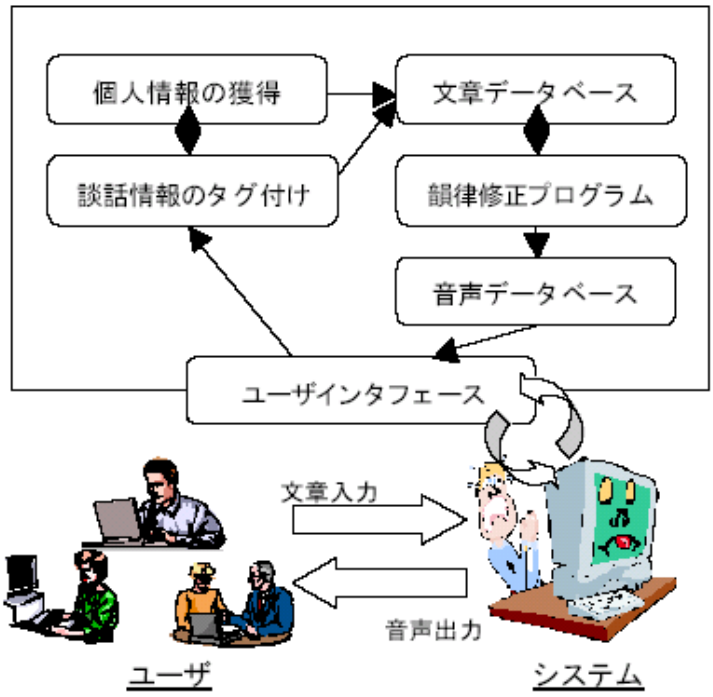


図 2.11: 談話情報を用いた韻律制御システム

同士の立場等を文，意見単位でタグ化する．タグ化されたパラ言語情報に応じて，出力する音声を変換する韻律修正プログラムを設計する．

ii)　個人情報の獲得

内容語の語彙連鎖と語の統計的性質に着目し，同一の概念に属する語が集まって形成される語彙的連鎖の情報，語の重み付けによる値を用いて話題構造を生成し，話題の境界の特定を行なう．

iii)　談話情報のタグ付け

タグセットとしてはおおまかに，自分の態度，相手への態度，心理の 3 種類を想定している．「自分の態度」とは，自分の意見に対する考え方である．「相手の態度」とは，相手の発言に対する働きかけである．「心理」については，音声合成における重要な情報として古来から研究が行なわれており，システムにも適用されている [29]．

iv)　韻律修正プログラムの作成

タグ付けされた談話情報に基づいてユーザの各発言に音声合成を施す．音声合成においては，韻律の修正を行なう．音声合成における韻律の制御には，大きくパワー，ピッチ，タイミングの 3 つのパラメータを用いる必要がある．本システムでは，タグとの関係をプロ

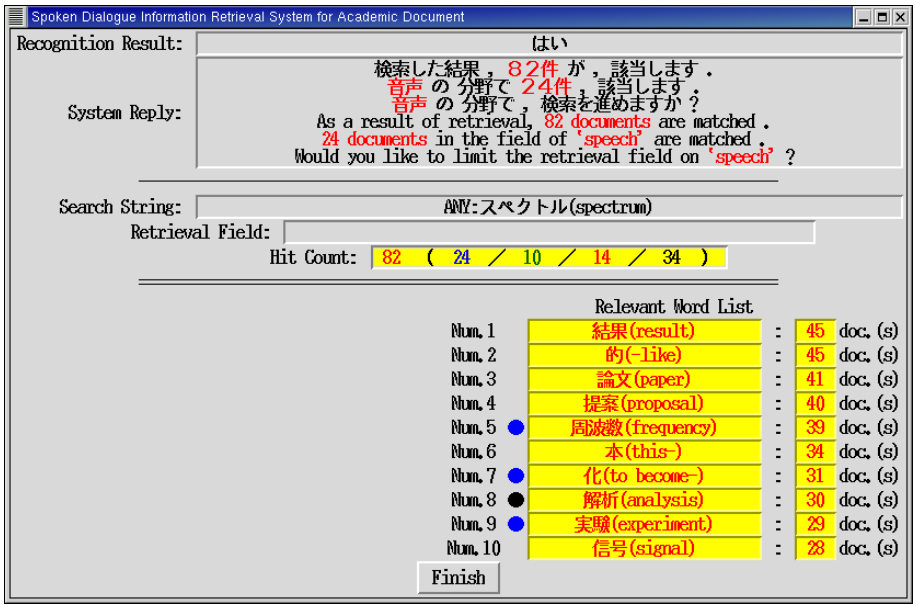


図 2.12: システムの画面表示

グラム化するという観点から，藤崎モデルを元にした基本周波数 F_0，及び継続時間長の修正（フレーズ成分，アクセント成分）を検討している [30].

2.6.2　学術情報検索音声対話システム

i)　システム概要

桐山らは，論文検索をタスクとした学術情報検索音声対話システムを開発した．[31] このシステムでは，対話管理手法および音声応答生成手法の高度化によって，ユーザにとって有益な学術論文の検索を分り易い音声応答によって提示することを目的としている．また，システムの画面表示の様子を図 2.12 に示す.

ii)　対話の焦点

対話中のある時点での発話の中における，相手に伝達される情報の中心となるもの，すなわち，発話者が相手にもっとも把握してもらいたいと考える情報を，対話の焦点と位置付ける．応答生成にあたって，この焦点の置かれている部位を強調することで，ユーザにとって理解しやすい音声応答を生成できるようになる，と期待される.

iii)　韻律規則

[32] の韻律規則を用いて音声合成を行なっている．[32] において，対話音声の韻律規則はフレーズ指令とアクセント指令の 2 種類の指令に対する規則からなっている．これは，朗

読音声に対して構築された規則 [33] を，対話音声の分析結果に基づき対話音声向けに変換したものである．

フレーズ指令は，文頭・文中・文末の 3 種類の指令があり，アクセント指令には，平板型・頭高型のアクセント立ち上げ指令，起伏型のアクセント立ち上げ指令，両者の立ち下げ指令の 3 種類の指令がある．各指令の大きさは，数量化分析によって決められており，数字は指令決定の際に考慮される各項目（パラメータ）のどのカテゴリに分類されるかの値を示す．

このパラメータの 1 つに，フレーズ指令についてはそのフレーズが重要度を持つか否か，アクセント指令についてはその韻律語が重要か否か，という単語の文脈における重要度を表すものがあり，対話の焦点をこれらのパラメータ値に反映させて音声応答を生成する．

iv)　システム内部表現

応答文のシステムの内部表現を 3 種類用意した上で，抽象度の高い概念表現を入力としてこれを段階的に変化していくことで，音声合成器への入力となる音韻記号と韻律記号の列を生成している．

文概念コード　抽象的な文概念に付加情報を与えて決定される応答文の文型を記述したもの

韻律句コード　応答文を意味的なまとまりのある韻律句に準ずる単位で分割し，記述したもの

単語コード　単語を辞書引きするためのコード

2.6.3　GoalGetter

i)　概要

Thenue らは，Data-to-Speech（D2S）と呼ばれる手法による応答生成手法を提案し，GoalGetter システムに実装している．[34] D2S とは，CTS と考え方は似ているが，「システム内では概念ではなくデータとして扱っており，また概念だけではなくあらゆる情報をデータとして用いることから，Concept-to-Speech よりも Data-to-Speech と呼ぶ方がより一般的な呼称としてふさわしい」と主張して命名されたものである．D2S の概念図を図 2.13 に示す．

「GoalGetter」システムは，「サッカーの試合結果の案内」をタスクとし，「どの試合で誰がいつ点を取った」のような情報をユーザに提示する．「GoalGetter」のシステム画面を図 2.14 に示す．このシステムはオランダ語で開発されたものであるが，システム内に実装された D2S の枠組を用いれば，オランダ語のみならずドイツ語や英語といった Germanic language にも適用できる，としている．

ii)　言語生成手法

構文テンプレートをいくつか用意しておき，テンプレートに含まれるスロットに単語を挿入することで文を完成させる．このテンプレートは完全な文単位で保持しており，ユー

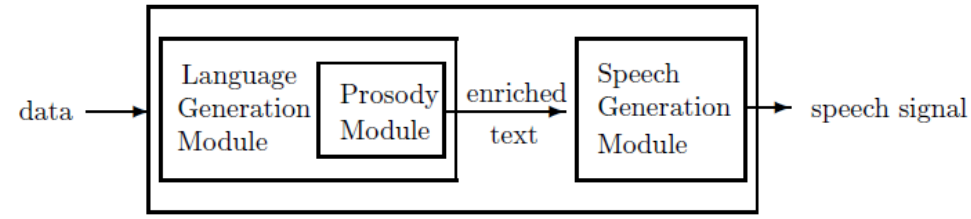


図 2.13: Data-to-Speech

図 2.14: GoalGetter

ザの知識量（ユーザが知りたいことがどこまではっきりしているか）や伝える内容（得点状況やカード状況）によってテンプレートを選択する．この際，状況によっては，ある状況に対して数種類のテンプレートを用意しており，その中から任意のテンプレートを選択することによって応答文のバリエーションを持たせている．また，スロットに挿入する単語にもバリエーションを持たせている．

iii)　韻律制御手法

文の構文情報から，焦点位置と強勢/弱勢の判別を行なう．談話情報（重要度，新規性）から焦点を当てるかどうかの決定を行ない，文の構文情報から強勢/弱勢を決定する．ピッチの上げ下げに関しては，定量的な制御を行なっているわけではなく，2（焦点が当たっている/いない）×3（強勢/弱勢境界，文末，それ以外）の 6 通りに場合分けを行なっており，該当するピッチパターン音声を収録したコーパス音声の中から選択している．

2.7　まとめ

本章では，一般的な音声対話システムの概要とシステム構成について述べ，例をあげて各モジュールの構成を示した．また，音声対話システム研究において留意すべき事項を，研究例を紹介しながら説明した．音声対話システムにはまだ不完全な部分が多い．ユーザ発話の意図を汲み取るシステムは存在するが, 精度が高いとは言えない. また, ここで述べられている手法では画一的な応答しか得られない. 実際の対話においては, ユーザの知識などに応じて同じ質問に対して違う返答を行うこともあるのが自然である. 本論文では, 次章にて対話システムを構築するための対話の主導権や, 対話モデルについてより詳細に説明し, 第 4 章にて, 実際にユーザに応じて対応を変化させる柔軟な音声対話システムを構築する.

第3章

対話の主導権と対話モデル

3.1　はじめに

音声対話システムにおいて, 音声理解においては文構造に関する理論や意味に関する理論が背景となって個々の処理が実現されている. これらの構文理論や意味論は文の妥当性や言語表現から記号表現への写像を扱っており, それぞれの理論で閉じた系で議論ができる. しかし, 対話理解においては, 連続する発話の関連を導くためにタスクドメインに関する知識や, 話し相手の心の中を推察するような処理が必要となり, その表現すべき知識や規則の対象は言語理論の中に閉じることはできない. そこで, タスクの構造や対話参加者の意図をできるだけ一般的な形式で表現し, そのもとで対話の構造を記述するという方法で対話の理論化が試みられてきた.

本章では対話の理論化における対話の主導権と対話モデルについて説明し, それらをふまえた上で次章にて実際に音声対話システムを構築する.

3.2　対話の主導権

音声対話システムにおいては, 対象とするタスクの違いや対象とするユーザグループの違いによって実現すべき対話のスタイルが異なる. チケット予約のように対話の流れが比較的定型な場合や, 想定されるユーザの範囲が広く, ユーザの行動が予想しにくい場合は主としてシステムが対話を制御し, より確実に目的達成に導くスタイルが望ましい. 一方, ヘルプデスクのように対象とするタスクが広い場合や, 想定されるユーザが熟練者の場合にはユーザが対話を制御した方が効率が良いことが多い.

このように質問や要求を発する立場にある側を対話の主導権を持つと表現する. 対話の主導権をシステムが持つものをシステム主導, ユーザが持つものをユーザ主導と呼ぶ. また, 対話の途中で主導権が入れ替わり得るものを混合主導と呼ぶ.

3.2.1　システム主導

システムがあらかじめ決められた手順でユーザから情報を順次引き出すことによって対話の目的を達成するスタイルをシステム主導スタイルと呼ぶ. 図 3.1 にシステム主導の対話例を示す. このシステム主導スタイルは, 基本的にはシステムからの情報要求に対してユーザが応答するというターンの繰り返しで構成される. システムからの情報要求項目数を 1 ターンにつき 1 項目に限定すれば, ユーザの応答発話を単純なものに限定できる. たとえば, システムの「どの地域のレストランをお探しですか？」質問に対しては, ユーザの回答を地域名に限定することができる. したがって比較的高い音声認識率が期待でき, このことが高いタスク達成率につながる.

一方このシステム主導スタイルでは一般に対話のターン数が多くなるのでタスク達成に要する時間が長くなる傾向がある. また, そもそもユーザの意図に沿った情報検索が可能なのかどうかは対話が終了するまでわからず, 対話そのものは破綻なく終了してもユーザが意図した結果が得られたかどうかは分からない場合もある.

S(システム): 番組案内システムです．いつ放送されている番組について知りたいですか？
U(ユーザ): 明日の 10 時
S: どんなカテゴリの番組について知りたいですか？
U: ドラマ
S: 明日 10 時に放送されているスポーツ番組は「渡る世間は鬼ばかり」です．
S: 出演者についてお知らせしましょうか？

図 3.1: システム主導の対話例

U(ユーザ): 明日の 17 時にやっているニュース番組を教えてください．
S(システム): イブニング・ニュースです．
U: その番組には誰が出演していますか？
S: 長峰由紀，伊藤隆太が出演しています．

図 3.2: ユーザ主導の対話例

3.2.2　ユーザ主導

ユーザ主導スタイルとはユーザがシステムに対して質問し, システムから情報を引き出すスタイルである．図 3.2 にユーザ主導の対話例を示す．ユーザ主導スタイルでは, ユーザの入力発話が適切であれば効率よく対話が進行する．しかし, 一般にこのスタイルではユーザの発話が複雑になる傾向があるため, 音声認識・理解が難しくなる．さらにシステムを使い慣れないユーザには, どのように話せばよいのかが分かりにくいという問題がある．

また,「では予算はどれくらいですか？」のように前のターンの内容と関連した省略発話が頻繁に現れる．システムの対話管理部では予算をたずねている対象の飲食店が前述の店であることを補わなければ正しく回答することができない．省略された情報は直前のターンに現れることが多いが, 数ターン前方に現れた情報が省略されることも珍しくはない．このことが省略発話の理解を難しくしている．

3.2.3　混合主導

混合主導スタイルは, 主導権の移動が可能なスタイルである．図 3.3 に混合主導の対話例を示す．人間同士の会話も, 多くは主導権が一貫して固定しているわけではなく, 一般には話の流れに応じて主導権は移動する．しかし, 人間同士の会話ではどちらに主導権があるのか分からない状況もよく見られ, そのような問題を解決して対話の最終目的に着実に至るような音声対話システムを作るのは難しい．したがって, 音声対話システムにおける混合主導対話はあらかじめ決められた一定のスタイルを前提とするものが多い．

よく見られるスタイルは基本的にはシステム主導で対話を進め, 必要に応じてユーザに主導権を取らせることによって対話を円滑に進めるものである．一時的にしかユーザに主導権を渡さない方式であっても, 熟練したユーザには対話の効率が大きく向上する傾向が

S(システム): 番組案内システムです．何か質問はございますか？
U(ユーザ): 明日の 17 時にやっているニュース番組を教えてください．
S: どの放送局の番組ですか？
U: TBS です．
S: 該当する番組は「ニュース・イブニング」です．

図 3.3: 混合主導の対話例

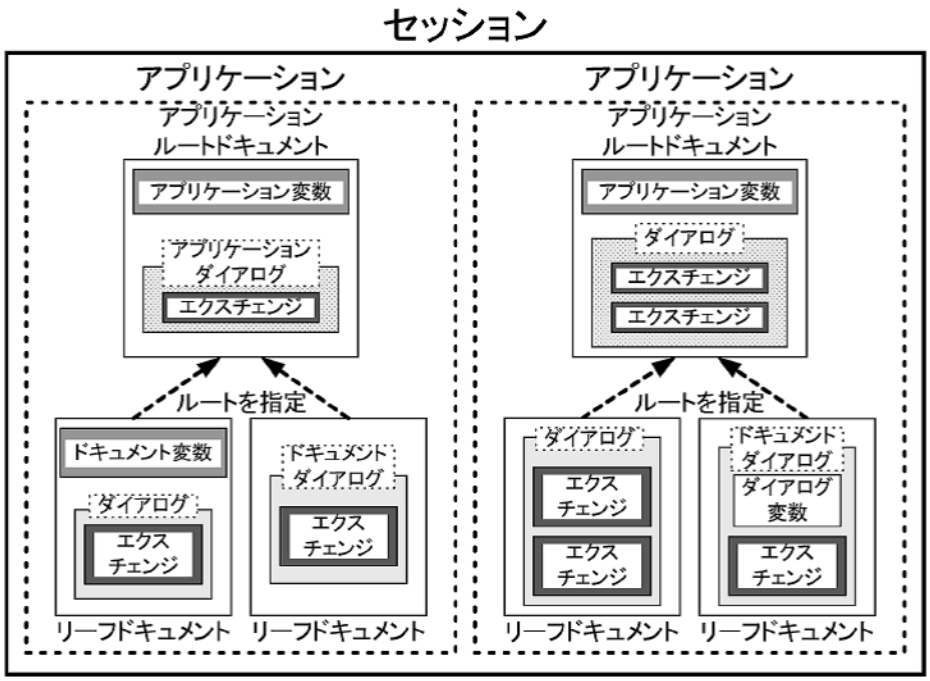


図 3.4: XISL の対話構造

見られる．

一方, システムとユーザの役割を入れ替えた逆のスタイルもありうる．基本的にはユーザ主導で対話を進め, 必要に応じてシステムが主導権を取るものである．しかし, このスタイルはユーザ主導スタイルに見られた発話理解および対話管理が複雑になるという問題に加え, システムが主導権をとるタイミングの制御という新たな問題も生じ, 一般には実現は難しい．

3.3　対話モデル

3.3.1　状態遷移モデル

状態遷移モデルはシステム主導対話システムにおける対話管理に適している．状態遷移モデルの実現方法として現在の音声対話システムで最も広く使われている方法は, 有限状態オートマトンによる対話管理手法である．

オートマトンは入力あるいは出力に伴って状態変化を起こす機械である．有限状態オートマトンは状態があらかじめ規定できるオートマトンであり, 以下の要素で定義される．

- 状態集合：$\{s|s \in S\}$
- 入力記号集合：$\{a|a \in A\}$
- 状態遷移規則集合：$\delta : S \times A \rightarrow S$
- 初期状態：$s_0 \in S$
- 終了状態集合：$F \in S$

このオートマトンモデルで対話システムの動作を定義することができる．システムの発話を状態 s に, ユーザ発話またはシステムの処理結果を入力記号 a に, 可能な対話の遷移を δ に対応させることによってシステム主導の対話管理部の動作が規定できる．このようなオートマトンによる対話制御の長所としては, システム開発者にとって対話の流れが容易に理解できるということがあげられる．ＧＵＩによる開発ツールキットを利用することで, 分岐や状態の遷移などが直感的に把握でき, 開発効率や保守性が高まる．

一方, 有限状態オートマトンによる対話制御の短所としては, ユーザ主導対話や混合主導対話を記述する際に, 状態数や遷移条件記述が多くなりすぎるために, 作成, 保守が困難になるということがあげられる．

3.3.2　知識駆動モデル

知識駆動モデルは混合主導スタイルの音声対話システムを実現するのに適している．混合主導スタイルでは, 変数間の制約や対話の目標に応じた変数間の動的な優先順序制御などによって対話の遷移を決めるべきである．このようなことを実現することによって, 対話が効率的になり, 混合主導スタイルを採用したメリットが出てくる．

変数間の制約や対話の目標と変数との関係などを対話制御に反映させるためには, タスク全体の知識を明示的にまとめて表現する方法が必要になる．そのような知識表現としては, フレームや AND-OR 木があげられる．

3.3.3　相互作用モデル

ユーザの目的が多様であるような状況では, ユーザが対話システムを使う目的やユーザの好みを把握しなければならない．このような対話システムを実現するためには, システムがあらかじめ用意した状態・知識表現によって対話を進めるモデルではなく, ユーザとシ

ステムの相互作用によって対話を進める必要がある. 相互作用による対話制御は, 混合主導スタイルもしくはユーザ主導スタイルで実現するのが適している.

3.4　対話記述言語

データベース検索を行う音声対話システムは, 前述した複数のモデルによって構築することができる. 音声対話の記述言語には XISL[35],VoiceXML[36] などがあり, これらは XML 構文に基づくマルチモーダル対話記述言語である. VoiceXML は音声認識, DTMF(Dual Tone Multiple Frequency:電話のトーン信号) キー入力, 録音, 音声合成, オーディオファイルの再生, 電話転送昨日などを君合わせて, 音声アプリケーションを開発するための言語である. 現在主流である VoiceXML が限定されたモダリティを対象としているのに対して XISL はモダリティの拡張性を高めている.

3.5　まとめ

本章では音声対話システムの対話理解における対話の主導権について概要について説明し, 状態遷移モデル, 知識駆動モデル, 相互作用モデルの 3 つの対話モデルについてその特徴を述べた. また, 対話システムを構築するための対話記述言語についても言及した.

これらのモデルをふまえた上で, 次章では実際に, ユーザの特徴や知識を利用することによって画一的でない応答を生成し, ユーザとの自然な対話を実現する音声対話システムを構築する.

第4章

ユーザの特徴・知識を考慮した情報検索音声対話システム

4.1　はじめに

音声認識技術や自然言語処理技術の進歩とともに機械との対話を行なう音声対話システムへの期待が高まっている. キーボードでの操作が困難なカーナビなど, 実用化が進んでいる応用分野もあるが, 一般には, 簡単な電話応答, あるいは, イベント会場でのデモに一部利用されているに過ぎず, 日常社会において音声対話システムはまだまだ身近なものとはなっていない.

一方, 社会の情報化が進展し, 家庭内でのインターネット等を介した情報収集が一般的になっている. 既に, テレビジョンなど, 家電製品が情報機器化し, それらのネットワーク化が進んでいる. 未来住宅では, これがさらに発展し, ユーザは機器を意識せずに, 操作し得ることが求められる. このためには, 人間の最も根源的な情報伝達媒体である音声による情報検索や情報授受の確立が必要である.

一般的な音声対話システムでは, 話題とするトピックが変わらない限り, ユーザの同じ発話には, （省略, 照応のレベルの違いはあるものの）基本的に同じ応答が生成される. しかしながら, 同じ質問であっても, 状況によって, ユーザの求めている情報は異なり, またユーザの情報の理解力は, ユーザの知識により異なってくる. このため, ユーザが真に求めている情報を, ユーザが理解しやすいように提示するという観点から, 音声対話システムには, 画一的でない応答を生成する能力が求められる. ユーザの情報を利用し, それに即した応答を生成する研究には, 大人か子供かによって応答を変化させる「たけまるくん」[2] などがあるが, ユーザの知識レベルを判定して応答を動的に変化させるものではない. ここでは, トピックに対する, ユーザの知識の静的なレベル, 動的なレベルを判定し, 判定された知識レベルに応じて適切な応答を生成することを提案し, システムの実装を進めた結果について報告する.

4.2　システム概要

4.2.1　構成

本研究における対話システムは, テレビ番組情報の検索をタスクとしている. システムはユーザとの自然な対話を通じてテレビ番組の情報を与え, ユーザの番組選択を助ける. 対話を行う際, 自然な対話を行うためにユーザの特徴や知識を利用して対応を行う. システム処理の概要を図 4.1 に示す.

大枠のシステムとしては擬人化音声対話エージェントのツールキットである Galatea Toolkit を用いる [37]. 音声認識部としてネットワーク文法を用いた Julian[16] を使い, 音声合成部として Galatea Talk を使用する.

4.2.2　音声認識部

音声認識部ではユーザの発声を文字列に変換し, 出力する. 音声認識技術は, その認識対象とする単語辞書の大きさによって小語彙（数百語程度）と大語彙（数千語以上, 数万語

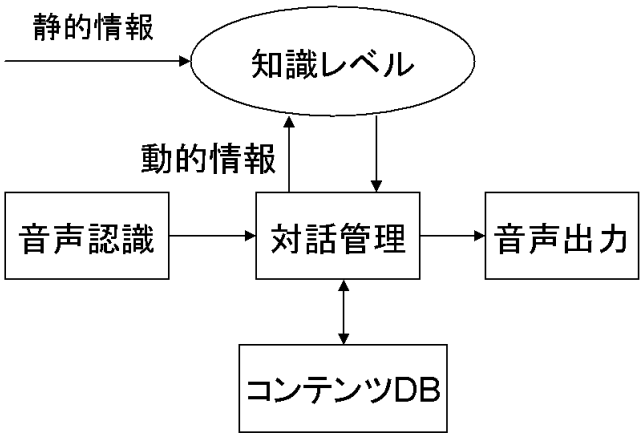


図 4.1: システム構成

```
S              :NS_B SENTENCE NS_E

SENTENCE            :program NOISE
SENTENCE            :program NOISE WO PLEASE
SENTENCE            :program NOISE GA PLEASE
```

図 4.2: grammar ファイル

程度）に区分することができる.

大語彙連続音声認識システムでは, 音声認識の際に音響モデル, 単語辞書, 言語モデルを用いる. 音響モデルとは, 音素の並びを統計的に学習したものであり, HMM(Hidden Markov Model) がよく用いられる. HMM は特徴ベクトル時系列の確率モデルであり, 自己遷移を持つ複数の状態間を遷移することで, 音声のような長さの一定しない時系列信号を効率良くモデル化することが可能である. HMM の学習には EM（Expectation Maximization）アルゴリズムと呼ばれる最尤パラメータ推定手法が用いられる.

音響モデルが話者性や音声入力環境などの音声認識における音響的特徴を担うものであるのに対し, 言語モデルと単語辞書は, 言い回しなどの文章表現や認識対象単語などの言語的特徴を定めるものである. 統計的言語モデルを用いた大語彙連続音声認識では, 認識結果を開発者があらかじめ決定的に定義することは難しく, アプリケーションに組み込むのには向かない. しかし, 柔軟に様々な発話を受理することが可能であるため利用する価値は高い.

それに対して, タスクを限定することによって音声認識での認識の対象とする語彙を抑えることはある程度可能である. その場合人手で文法を記述し, タスクに応じた語彙を登録することによって認識を実現する.

本システムでは音声認識部として Julian を使用する. Julian は, 有限状態文法 (DFA) に基づいて，与えられた文法規則の元で入力音声に対して最尤の単語系列を探しだす音声認

```
%WO
を                    o
について              n i ts u i t e

%GA
が                    g a

%PLEASE
お願いします          o n e g a i sh i m a s u
よろしく              y o r o sh i k u
教えてください        o sh i e t e k u d a s a i
教えて                o sh i e t e
```

図 4.3: voca ファイル

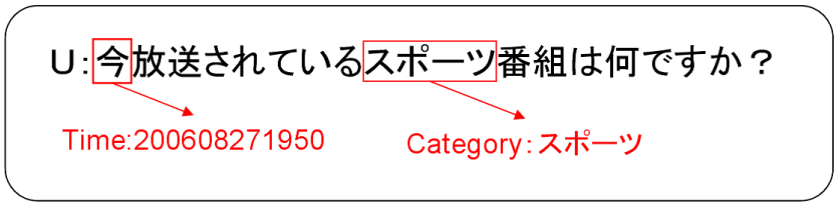


図 4.4: キーワードスポッティング

識エンジンであり, その動作には，そのタスクにおいて認識の対象とする文のパターンを記述した「認識用文法」（あるいはタスク文法）が必要である. タスク文法は構文制約を単語のカテゴリを終端規則として記述する grammar ファイルとカテゴリごとに単語の表記と読み（音素列）を登録する voca ファイルから成る. これらのタスク文法の例を図 4.2, 図 4.3 に示す.

4.2.3　対話管理部

対話管理部ではデータベースを用いて対話を構築する. 以下に対話処理の基本的な流れを示す.

1. ユーザの自由な発話より必要な検索情報を取得.

↓

2. 検索情報をデータベースと照合.

表 4.1: データベース例

start_datetime	finish_datetime	title	kana_title	category	...
200608271800	200608271955	東京シティ競馬	トウキョウシティケイバ	スポーツ	...
200608272000	200608272054	どうぶつ奇想天外！	ドウブツキソウテンガイ	...	...
...	...	...	...	...	

S:<Time>放送されている<Category>番組は<Title>です。

S:2006年8月27日19時50分に放送されているスポーツ番組は東京シティ競馬です。

図 4.5: 返答例

↓

3. 十分な情報があればデータベースから必要な番組情報を得て返答．検索情報が足りなければユーザに情報を要求．

↓

4. 次の対話に遷移．この時知識レベルの判定を行うことにより柔軟な対応を行う．

1. では図 4.4 のようにキーワードスポッティングを用いてユーザの発話から必要な文字情報を取得する．図 4.4 の例では Time 変数に「200608271950」, Category 変数に「スポーツ」が格納される．2. では 1. で取得した情報をデータベースと照会する．データベースの例を表 4.1 に示す．この例では Time 変数が「200608271950」, Category 変数が「スポーツ」である場合, title「東京シティ競馬」が選択される．3. では 1. と 2. で取得した番組情報を用いて返答を返す．図 4.5 に返答例を示す．2. の段階において返答に十分な情報を取得できなかった場合にはユーザに再び発話を促す．一通りの対話が終了した後, 4. では次の対話に遷移を行うが, このときに知識レベルの判定を行うことによって, より自然な対話を実現する．

4.2.4　音声合成部

音声合成部には GalateaTalk を用いる．GlateaTalk は HMM に基づいた Text to Speech の音声合成システムである．形態素解析は chasen[38], 複合語処理, 音韻交代処理には chaone を用いている．本システムでは基本的にテキスト音声合成 (TTS Text-To-Speech) によって行うが, 重要語句にアクセントをつけたり, 知識レベルに応じて発話速度の最適化を行って

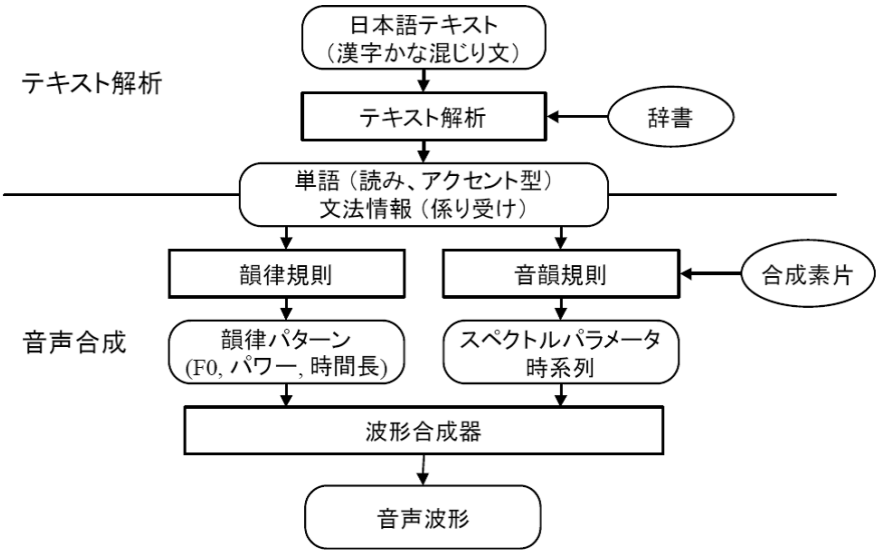


図 4.6: Speech synthesis system

いる.

4.2.5　データベース

本システムではテレビ番組情報のデータソースとして EPG(Electronic Program Guide) を用いる. 今回用いたデータには 2006 年 8 月 19 日〜2006 年 12 月 8 日までのテレビ放送内容が収録されており, このソースに含まれるデータには, 「番組名」, 「時刻類」「カテゴリ」などがある. これらのデータを RDB として扱い, SQL で操作する. 図 4.7 にデータベースの構造を示す.

4.3　ユーザー情報の利用

4.3.1　概要

一般的な音声対話システムでは, ある同じトピックを話題にしている場合, ユーザが同じ発話をすると同じ答えを返すのが普通である. しかし, 同じ質問であってもユーザによって求めている情報や, 情報の理解力は異なるのが実情であり, 対話システムにおいてもユーザの真に求めている情報をユーザが理解しやすいように対応することが望ましい. そこで本システムではユーザの静的情報, 動的情報を利用することによって, ユーザのトピックに対する知識レベルを判定し, 知識レベルに応じて適切な対応を行うことを提案する.

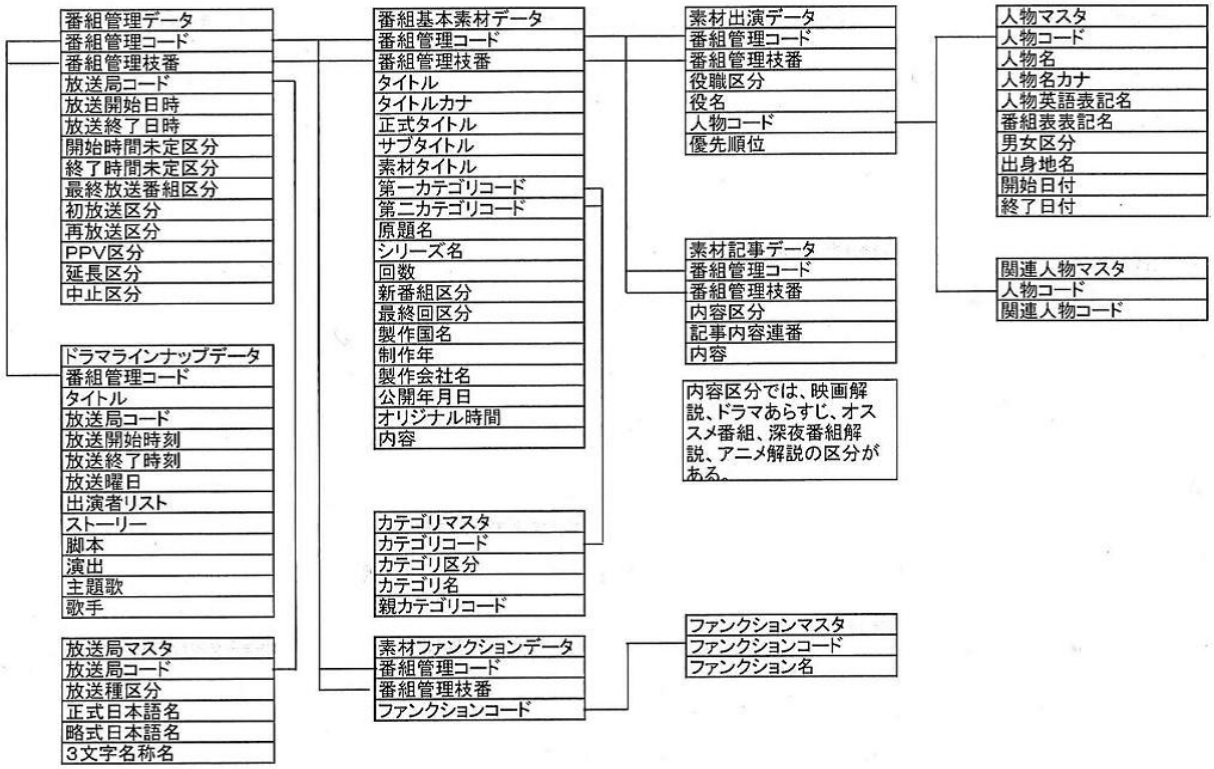


図 4.7: データベース構造

4.3.2　知識レベルの導入

ユーザの特徴, 知識の度合いを表すパラメータとして知識レベルというものを定義する. 知識レベルはユーザの特徴データから決定され, 対話システムの応答生成に影響を与える. この知識レベルは各カテゴリと各番組ごとに設定する.

4.3.3　知識レベルの決定方法

知識レベルの具体的な仕様については以下のようになる.

1. 知識レベルは各カテゴリ, サブカテゴリ, 番組ごとに-4〜+4 の 9 段階で設定する
2. カテゴリ, サブカテゴリのレベルはユーザの静的情報から初期値を決定する.
3. 各番組のレベルの初期値は 0 とする.
4. 各番組の知識レベルが変化した場合, それに応じてカテゴリ, サブカテゴリのレベルも変化させる.

ユーザの特徴のうち, 年齢, 体格などのような実際に人間が対話したとして前提条件として持つような知識を静的情報とする. 静的情報は人のコミュニケーションにおいて, 対話以外からあらかじめ持っている知識であるため, システム上では最初に入力するという形をとる. 実際の処理としては, カテゴリ, サブカテゴリに対してアンケートを取り, 興味がある

S: 「刺客請負人」の出演者を知っていますか？

U: はい→増加
U:いいえ→減少
//直接的な知識の差

図 4.8: ユーザの受動的な発言における知識レベルの変化

U: 野球について教えてください
//知識はある？ない？興味はある．

U: エムエルビーイズムゼロロクについて教えて
//タイトル名を自分で言い出しているため，先ほどよりは知識が深い．

図 4.9: ユーザの能動的な発言における知識レベルの変化

ものの初期値を+2, 興味のないものの初期値を-2 としてあらかじめ入力しておく．それに対して, 対話を実際に行いながら得るユーザ情報を動的情報とする．以下に対話中における知識レベルの動的変化について述べる．

i)　ユーザの受動的な発言

システム主導の対話におけるユーザの受動的な発言からはユーザのトピックに対する知識レベルを得る事ができる．図 4.8 にその対話例を示す．

ii)　ユーザの能動的な発言

それに対して, ユーザ主導の対話におけるユーザの能動的な発言からは新しいトピックに関する知識レベルを得ることができる．ユーザによって発せられた新たな単語はそれぞれ意味を持っているが, その種別によって情報量は異なってくる．そのため以下のように重み付けを行う．

1. 内容（出演者, ラテ欄など）
2. タイトル
3. サブカテゴリ
4. カテゴリ

図 4.9 に対話例を示す．

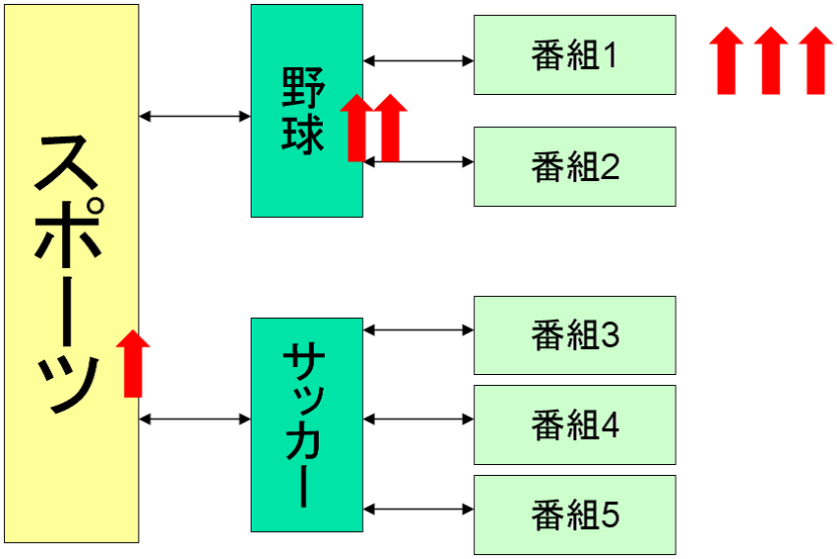


図 4.10: 知識レベルフィードバック

iii)　フィードバック

知識レベルは番組, カテゴリ, サブカテゴリごとに設定するが, それぞれの知識レベルの変化をフィードバックすることによって番組単体の知識にとどまらないユーザモデルの構築を目指す. 図 4.10 に処理の概念図を示す.

番組 1 の知識レベルが増加した際, サブカテゴリの「野球」とカテゴリの「スポーツ」のレベルが増加する. この時他の番組のレベルは変化しないが, 知識レベル判定の際には, カテゴリ, サブカテゴリ, 番組単体の 3 つの知識レベルを用いて判定するため, 番組 2〜5 にも影響は波及する. このような知識レベルフィードバックにより, 少しの対話から多くの情報を得ることができる.

4.3.4　知識レベルに伴う対話の変化

4.3.3 で定めた知識レベルに応じて対話を変化させる. 具体的には単純にカテゴリ, サブカテゴリ, 番組のレベルの和が 0 より多ければ知識レベルが高いと判断. 0 以下ならば知識レベルが低いと判断する. 知識レベルが高い場合には, 省略を増やしユーザ主導で対話を進める. 知識レベルが低い場合には, 省略を減らし, システム主導で対話を行い, また, 応答音声の発話速度を遅めに設定する. 図 4.11 に対話例を示す.

4.3.5　対話例

4.3.4,4.3.3 で説明した手法を含んだ対話例を図 4.12 に示す. ユーザはバラエティにはあまり興味がないという前情報があり, バラエティの知識レベルの初期値は-2, 他のレベルは 0 であったと仮定している. ここで「浅草お茶の間寄席」はカテゴリがバラエティ, サブカ

S: 番組案内システムです．何か質問はございますか？
U: 今 TBS で放送している番組はなんですか？
S: 現在放送している番組は「エムエルビーイズムゼロロク」です．

//スポーツカテゴリの知識レベルが低い場合
S: 「エムエルビーイズムゼロロク」は「スポーツ」カテゴリーの番組です．
S: 今日の「エムエルビーイズムゼロロク」の内容についてお伝えしましょうか？
U: お願いします．
S: 今日の「エムエルビーイズムゼロロク」の内容は「千葉ロッテ・西岡剛が登場．イチローから学んだものほか」です．
S: 出演者についてお伝えしましょうか？

//スポーツカテゴリの知識レベルが高い場合
S: 「エムエルビーイズムゼロロク」について何か知りたいことはありますか？
U:誰が出演していますか？
S:「イシバシタカアキ」，「ニシオカツヨシ」が出演しています．
U:石橋貴明が出ている番組は他に何がありますか？

図 4.11: 知識レベルに伴う対話の変化例

テゴリが寄席／演芸であった．

まずはじめにユーザが自分からバラエティの話題を振ったため, 興味があると判断され, バラエティレベルは 1 増加した．その後, ユーザは「浅草お茶の間寄席」について質問したが, これはシステムが提示した情報なので知識レベルに変化はない．「浅草お茶の間寄席」についての知識レベルはバラエティレベル, 寄席／演芸レベル, 番組のレベルから判断され, この場合閾値である 0 よりも低いため, システム主導の対話が始まる．システムが情報を提示した後, ユーザに知識を問うたところ, ユーザに知識があることが判明したために, 番組のレベルは増加した．この時同時にサブカテゴリとカテゴリのレベルも増加している．この変化により, 知識レベルは閾値を越えたため, 以後ユーザ主導の対話に切り替わっている．

4.4　実装

これらの提案手法を用いた番組情報検索音声対話システムを Galatea ツールキットをメインシステムとして実装した． システムのデモ画像を図 4.13 に, 構成を図 4.14 に示す． GalateaFE が音声認識部 (Julian), 顔画像合成部 (FSM), 音声合成部 (gtalk) を統合し, GalateaDM が対話管理を行う． この時複雑な処理は外部サーバにて PHP で処理し, データベースは MySQL を用いて管理する．

S:何か質問はございますか？
U:バラエティの番組は放送してない？
//バラエティ-2 →-1

S:いつの番組ですか？
U:今日の午後.
S:今日の午後は「あいのり」，「浅草お茶の間寄席」，「@なまてれ」，他が放送しています.
S:どれについて知りたいですか？
U:「浅草お茶の間寄席」について教えて.
//新規情報ではないので知識レベル変化なし
//知識レベル判定-1+0+0=-1 =< 0
//システム主導に切り替え

S:「浅草お茶の間寄席」は 8 月 28 日 15 時 00 分から放送されます.
S:「浅草お茶の間寄席」はどんな番組か知っていますか？

U:知ってるよ.
//バラエティ-1 → 0
//トラベル 0 → 2
//浅草お茶の間寄席 0 → 3
//知識レベル判定 0+2+3=5 > 0
//ユーザ主導に切替

S:では何か他に知りたいことはありませんか？
U:誰が出ているか教えてよ.
…

図 4.12: 提案手法を用いた対話例

図 4.13: 番組情報検索音声対話システム

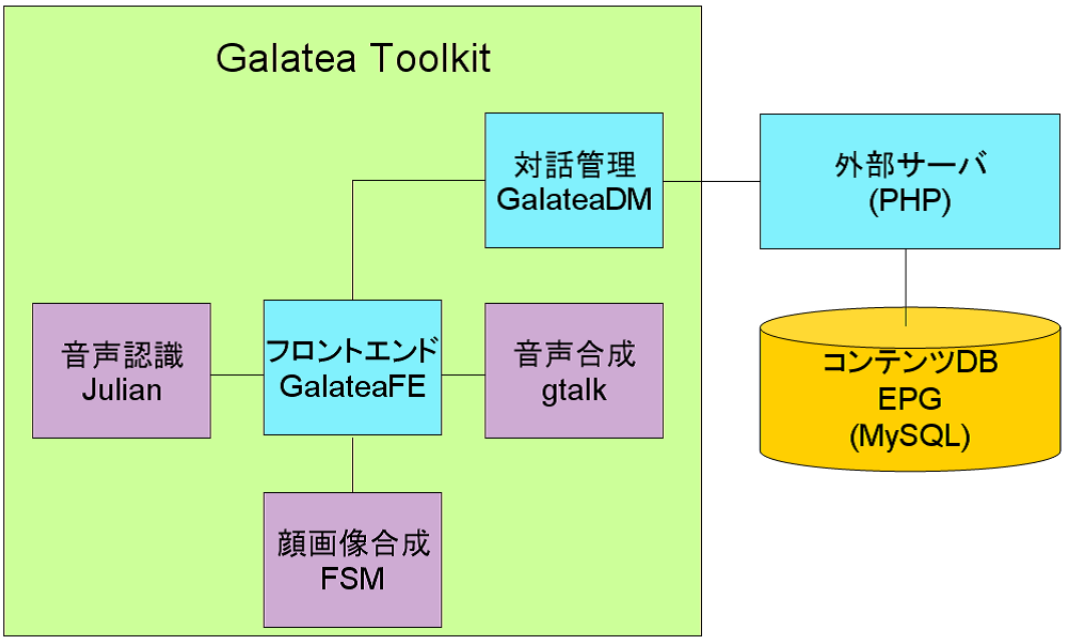


図 4.14: 実装構成図

4.5　まとめ

本章では, ユーザの特徴・知識を考慮することによる適切な応答生成手法を提案し, テレビ番組情報検索システムの実装を行った. 知識レベルという概念を導入し, そのレベルに応じた応答を生成することによってユーザとのより自然で柔軟な対話を実現することに成功した. また, 知識レベルの変化を番組ごと, サブカテゴリごと, カテゴリごとにそれぞれフィードバックすることによって少ない対話でより正確な知識レベルの判定を行えるようになった.

第5章

結論

5.1　まとめ

本論文では, 音声対話システムにおける画一的でない応答生成手法を提案し, その提案手法を実装した番組情報検索音声対話システムを構築した.

第 2 章では, これまでに研究されてきたさまざまな音声対話システムについて説明し, 対話システムにおいて重要となる各要素技術について述べた.

第 3 章では, 音声対話システムを構築する際に重要となる対話の主導権, 対話モデルについて説明した後, 対話記述言語に関して説明を行った.

第 4 章では，ユーザの知識を判定し, 適切な応答を生成する手法をについて述べ, 実際にテレビ番組情報の検索を行う音声対話システムへの実装における詳細について記述した.

5.2　今後の課題

知識レベルを利用した音声合成について, 本論文では単純に詳しい場合, 詳しくない場合の話速の変化, また重要語句へのアクセントだけについて行ったが, それが有効に働いているかの評価は行っていない. 正しい評価を行うとともに, ピッチの振れ幅の変更なども考えていく必要がある. また, 知識レベルの推定においてはまだ推定を行うことのできる文の量が不足している. 大規模な文例の分析を行ったり, ユーザの音声情報を利用することによってより正確に知識レベルの推定が可能になることが期待される.

また, 知識レベルの変化に伴う対話の分岐については今回は 2 パターンの分岐しか行っていない. より柔軟な対話を行うためには, 省略などの文体の変化を含めたより複雑な分岐が重要になってくる.

今回提案した手法によって従来よりも自然な対話を実現することができたと考えているが, きちんとした評価実験が行えていないため, 評価実験の実施も課題の一つである.

謝辞

本論文を執筆するにあたり，指導教官である広瀬啓吉教授，また研究室の共同運営者である峯松信明准教授には，日頃から研究や論文執筆等において，様々なご指導，ご鞭撻を賜わりました．深く感謝の意を表します．

研究室環境の整備など，本研究を様々な面で支援してくださった高橋登技官，秘書の楠本 由香里さん，元秘書の武田祥子さんに，深く感謝いたします．対話システムに関する研究をを私の前に行っておられた八木裕司氏, 高田靖也氏には研究を進める上でさまざまな助言をいただきました. 東京電力の小杉康宏氏には, データベースの提供をしていただいたり, さまざまな助言をいただきました. また, 日頃研究室生活を行う上でさまざまな面で相談に乗って頂いたり助言を下さったりした, 今まで関わってきた研究室の皆様に深く感謝いたします．

最後に，日頃から多岐にわたり私を支えて下さった友人，家族に深く感謝いたします．

2008 年 2 月 4 日
篠田 知宏

参考文献

[1] http://www.scansoft.co.jp/viavoice/.

[2] R. Nishimura, Y. Nishihara, R. Tsurumi, A. Lee, H. Saruwatari and K. Shikano: "Takemaru-kun: Speech-Oriented Information System for Real World Research Platform," International Workshop on Language Understanding and Agents for Real World Interaction, pp.70-78, 2003.

[3] S. Bennacef, L. Devillers, S. Rosset and L. Lamel: "Dialog in the RAILTEL Telephone-Based System," Fourth International Conference of Spoken Language Proceedings, Proc. ICSLP'96, Vol.1, pp.550–553, 1996.

[4] C. Popovici and P. Baggia: "Language Modelling For Task-Oriented Domains," Proc. Eurospeech '97, vol.3, pp.1459-1462, 1997.

[5] S. J. Young and F.Fallside: "Speech Synthesis from concept : A method for speech output from information systems," J. Acoust. Soc. Am., vol.66, no.3, pp.685-695, 1979.

[6] D. Goddeau, E. Brill, J. Glass, C. Pao, M. Phillips, J. Polifroni, S. Seneff and V. Zue: "Galaxy : A Human-Language Interface to On-Line Travel Information," Proc. ICSLP'94, Vol.2, pp.707–710, 1994.

[7] MITRE. Galaxy. http://communicator.sourceforge.net/index.shtml.

[8] SRI. Oaa. http://www.ai.sri.com/ oaa/.

[9] Speech Recognition Grammar Specification for the W3C Speech Interface Framework - W3C Working Draft 20 August 2001, http://www.w3.org/TR/2001/WD-speech-grammar-20010820/.

[10] W3C. ccxml. http://www.w3.org/TR/ccxml.

[11] W3C. scxml. http://www.w3.org/TR/scxml.

[12] 竹林洋一: "音声自由対話システム TOSBURG2 －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, Vol.J77-D-2, No.8, pp.1417-1428, 1994.

[13] 西村竜一, 内田賢志, 李晃伸, 猿渡洋, 鹿野清宏: “Julius を用いた学内案内ロボット用音声対話システムの作成,” 電子情報通信学会技術研究報告,pp.93-98, 2001.

[14] 李晃伸, 河原達也, 堂下修司: “単語トレリスインデックスを用いた段階的探索による大語彙連続音声認識,” 電子情報通信学会論文誌,pp.1-9, 1999.

[15] A. Lee, T. Kawahara and K. Shikano: “Julius - An Open Source Real-TimeLarge Vocabulary Recognition Engine,” EUROSPEECH2001,pp.1691-1694, 2001.

[16] 河原達也, 住吉貴志, 李晃伸, 坂野秀樹, 武田一哉, 三村正人, 山田武志, 西浦敬信, 伊藤克亘, 伊藤彰則, 鹿野清宏: “連続音声認識コンソーシアム 2001 年度版ソフトウエアの概要,” 情報処理学会研究報告, 2002-SLP-43-3, pp.13-18, 2002.

[17] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka and T. Kobayashi, K. Shikano and S. Itahashi: “JNAS: Japanese Speech Corpus for Large Vocabulary Continuous Speech Recognition Research,” The Journal of the Acoustical Society of Japan, Vol.20, No.3, pp.199-206, 1999.

[18] 堂坂浩二, 安田宣仁, 宮崎昇, 中野幹生, 相川清明: “音声対話システム「飛遊夢（ひゅうむ）」,” 電子情報通信学会総合大会, Vol.1, pp.506–507, 2001.

[19] 野田喜昭, 山口義和, 大附克年, 今村明弘: “音声認識エンジン VoiceRex を開発 - 幅広い応用に対応できる音声認識ソフトウェア,” NTT 技術ジャーナル, Vol.11, No.12, 1999.

[20] M. Nakano, N. Miyazaki, J. Hirasawa, K. Hohsaka and T. Kawabata: “Understanding unsegmented user utterances in real-time spoken dialogue systems,” Proc. 37th Annuam Meeting of the Association for Computational Linguistics, pp.200–207, 1999.

[21] 堂坂浩二, 島津明: “タスク指向型対話における漸次的発話生成モデル,” 情報処理学会論文誌, Vol.37, No.12, pp.2190–2200, 1996.

[22] 竹内真士, 北岡教英, 中川聖一: “韻律・表層的言語情報を発話タイミング制御に用いた雑談対話システム,” 情報処理学会研究報告（音声言語情報処理研究会）, 2004-SLP-50-14, pp.87–92, 2004. M. Takeuchi, N. kitaoka and S. Nakagawa: “Timing detection for real-time dialogue systems using prosodic and linguistic information,” Proc. Speech Prosody 2004, pp.529–532, 2004.

[23] 人工知能学会 談話・対話研究におけるコーパス利用研究グループ: “様々な応用研究に向けた談話タグ付き音声対話コーパス,” 人工知能学会研究会資料, SIG-SLUD-9903-4, 1999.

[24] J. Quinlan, R.: “C4.5 : Programs for machine learning,” Morgan Kaufmann, 1992.

[25] 広瀬啓吉: “音声の出力に関する研究の現状と将来,” 日本音響学会誌, Vol.52, No.11, pp.857–861, 1996.

[26] 山崎信英: “最近のテキスト音声合成とその技術,” bit, Vol.27, No.3, pp.11–20, 1995.

[27] （社）日本電子工業振興協会: “音声合成の製品動向,” 音声入出力方式に関する調査報告書, 00-標-2, pp.29–48, 2000.

[28] 遠山義洋, 西田豊明: “談話情報を用いた音声合成における韻律の制御,” 2001 年度人工知能学会全国大会（第 15 回）論文集, 07-06, pp.157–160, 2000.

[29] 飯田朱美, ニックキャンベル, 安村通晃: “感情表現が可能な合成音声の作成と評価,” 情報処理学会論文誌, Vol.40, No.2, pp.479–486, 1999.

[30] H. Fujisaki and K. Hirose: “Analysis of voice fundamental frequency contours for declarative sentences of Japanese,” J. Acoust. Soc. Jpn(E), Vol.5, No.4, pp.233–242, 1984.

[31] 桐山伸也, 広瀬啓吉: “応答生成に着目した学術文献検索音声対話システムの構築とその評価,” 電子情報通信学会論文誌, Vol.J83-D-II, No.11, pp.2318–2329, 2000.

[32] K. Hirose, M. Sakata and H. Kawanami: “Synthesizing dialogue speech of Japanese based on the quantitative analysis of prosodic features,” Proc. ICSLP’96, Vol.1, pp.378–381, 1996.

[33] 河合恒, 広瀬啓吉, 藤崎博也: “日本語文章音声合成のための韻律規則,” 音響学会誌, Vol.56, No.6, pp.433–442, 1994.

[34] M. Thenue, E. Klabbers, J. Odijk, J.R. de Pijper and E. Krahmer: “From Data to Speech: A General Approach,” Natural Language Engineering, 7(1), pp.47-86, 2001.

[35] 小林聡, 中村有作, 佳田浩一, 山田博文, 新田恒雄: “マルチモーダル対話記述言語 XISL の提案,” 情報処理学会研究報告, Vol.2001, No68, pp.43-48, 2001.

[36] http://www.voicexml.org/.

[37] 嵯峨山茂樹, 川本真一, 下平博, 新田恒雄, 西本卓也, 中村哲, 伊藤克亘, 森島繁生, 四倉達夫, 甲斐充彦, 李晃伸, 山下洋一, 小林隆夫, 徳田恵一, 広瀬啓吉, 峯松信明, 山田篤, 伝康晴, 宇津呂武仁: “擬人化音声対話エージェントツールキット Galatea,” 情報処理学会研究報告（音声言語情報処理研究会）, 2003-SLP-45-10, pp.57-64, 2003-2.

[38] http://chasen.aist-nara.ac.jp/.

発表文献

[1] 篠田知宏, 広瀬啓吉, 峯松信明, 小杉康宏: “ユーザの特徴・知識を考慮した番組情報検索音声対話システムの構築,” 日本音響学会 2007 年秋期研究発表会講演論文集, 2-3-7, pp.71-72, 2007-9.