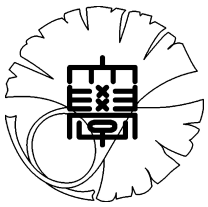
修士論文

道案内音声対話システムにおける応答音声生成手法



2006 年 1 月 30 日

指導教員　広瀬 啓吉 教授

東京大学大学院新領域創成科学研究科
基盤情報学専攻　46325

高田 靖也

内容梗概

音声は人間の最も基本的なコミュニケーション手段であり，これを計算機との情報授受に利用することの要求は高い．近年の，音声認識や自然言語処理，音声合成といった要素技術の発展に伴い，これら技術の統合により構成される音声対話システムの研究が盛んに行なわれるようになってきている．その多くのシステムでは，モジュール的な独立性が高く，音声対話システムに組み込みやすいという理由により，音声出力に既存のテキスト音声合成（TTS : Text-To-Speech）ソフトウェアが用いられている．しかしながら，TTS システムは，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．音声対話システムにおいては，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．TTS システムによる音声合成では，この利点を生かすことは困難であり，これらを応答音声に反映できる音声合成の枠組み，すなわち概念音声合成（CTS : Concept-To-Speech）[1] の実現が求められている．TTS がテキストを入力とするのに対し，CTS ではシステムの内部表現（概念）から直接音声を合成するため，文の生成過程で正確な言語情報が得られ，統語構造を韻律に反映させたり，談話情報で韻律の制御をしたりすることが容易に行なうことができる [2]．音声対話システムでは，朗読調のみならず対話調の応答音声が求められ，さらに，発話の意図や感情を反映させることも求められるが [3]，テキスト音声合成ソフトウェアでは，単調な朗読調音声しか出力することができない．

従来の音声対話システムは，音声認識や人工知能といった，入力・中間処理を対象としたものがほとんどであり，応答音声という出力を扱っているものは皆無に等しい．しかし，入力・中間処理の精度がどれほど進歩しても，出力の面での進歩が伴わなければ，ロボット産業はじめ様々なメディアコンテンツ産業において利用されるであろう音声対話システムのクオリティ全体の向上は果たせない．

本稿で述べた応答音声生成手法，韻律制御手法は，タスクによらず音声対話システムに適応できるという汎用性を持ち，本手法を用いることで，統語構造や文の焦点といった高次の言語情報を応答音声に反映することができる音声対話システムを構築することができると考えられる．

目次

図目次

表目次

第1章

序論

1.1　本論文の背景

音声は人間の最も基本的なコミュニケーション手段であり，これを計算機との情報授受に利用することの要求は高い．近年の，音声認識や自然言語処理，音声合成といった要素技術の発展に伴い，これら技術の統合により構成される音声対話システムの研究が盛んに行なわれるようになってきている．その多くのシステムでは，モジュール的な独立性が高く，音声対話システムに組み込みやすいという理由により，音声出力に既存のテキスト音声合成（TTS : Text-To-Speech）ソフトウェアが用いられている．しかしながら，TTS システムは，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．音声対話システムにおいては，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．TTS システムによる音声合成では，この利点を生かすことは困難であり，これらを応答音声に反映できる音声合成の枠組み，すなわち概念音声合成（CTS : Concept-To-Speech）[1] の実現が求められている．TTS がテキストを入力とするのに対し，CTS ではシステムの内部表現（概念）から直接音声を合成するため，文の生成過程で正確な言語情報が得られ，統語構造を韻律に反映させたり，談話情報で韻律の制御をしたりすることが容易に行なうことができる [2]．音声対話システムでは，朗読調のみならず対話調の応答音声が求められ，さらに，発話の意図や感情を反映させることも求められるが [3]，テキスト音声合成ソフトウェアでは，単調な朗読調音声しか出力することができない．

従来の音声対話システムは，音声認識や人工知能といった，入力・中間処理を対象としたものがほとんどであり，応答音声という出力を扱っているものは皆無に等しい．しかし，入力・中間処理の精度がどれほど進歩しても，出力の面での進歩が伴わなければ，ロボット産業はじめ様々なメディアコンテンツ産業において利用されるであろう音声対話システムのクオリティ全体の向上は果たせない．

1.2　本論文の目的

多胡らにより，音声対話システムにおける応答文生成手法と韻律制御手法についての研究が行われている [7]．その中で，CTS のための日本語の言語情報の取り扱い手法が示され，音声対話システムに実装することで，より自然な応答音声の生成を実現している．しかしながら，[7] のシステム（以下，従来システムという）では，より複雑な状況下における応答音声の生成や韻律の制御方法など，依然として多くの問題点が残されている．そこで，本論文では従来システムにおける問題点として挙げられる，応答文生成と韻律制御に基づく応答生成を中心に，音声対話システムを改良する，ということを目的とする．

従来システムにおける応答音声は，いかにも機械的であり，違和感の残るものであったが，これは人間同士の対話のような，言葉の言い回しや表現の豊富さの欠落，人間特有の感情表現といったパラ言語情報や非言語情報が，応答音声において反映されていないことが原因であると考えられる．より自然な，より違和感のない応答音声の生成は，ユーザにとって使いやすく，ストレスの少ないシステムを目指す上で必要不可欠なものである．本

研究において開発された道案内音声対話システム（以下，本対話システムという）は，応答文生成の新たな枠組みの導入と，応答音声に韻律規則やアクセント結合規則，文節間結合規則の導入を行なうことで，システム中で既に得られている高次の言語情報を韻律に反映させ，より自然な応答音声を生成することを目的としている．また，聴取実験により，従来のシステムとの比較，検討を行ない，本提案手法の妥当性を示す．

1.3　本論文の構成

本論文は，以下のように６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，音声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，道案内音声対話システムのタスクやユーザインタフェースを始めとするシステムの概要について述べる．第４章では，音声対話システムの対話処理の部分に関して，概念音声合成と親和性の高い言語情報の取り扱い手法について述べる．第５章では，音声対話システムにおける音声合成について，韻律制御手法を中心に述べる．第６章では，本論文をまとめ，今後の展望について述べる．

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，音声による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った．現在，いくつかの音声対話システムが実用化に至っている [8][9].

本章では，一般的な音声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音声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要素技術について述べる．

2.2　一般的な音声対話システム

2.2.1　システム概要

一般の音声対話システムは，おおむね図 2.1 のように 5 つのモジュールと，タスクのためのデータベースから構成される．入力された音声は，音声認識部（Speech recognizer）により文字列に変換される．そして，その文字列の意味解析を言語理解部（Language interpreter）が実行し，対話制御部（Dialog manager）に渡す．対話制御部は，データベースを参照して返答する内容を生成する．それを言語生成部（Sentence generator）が文字列に変換し，音声合成部（Speech synthesizer）において音声として出力される．この様な処理を繰り返すことで，人と計算機の対話を処理し，タスクを達成する．

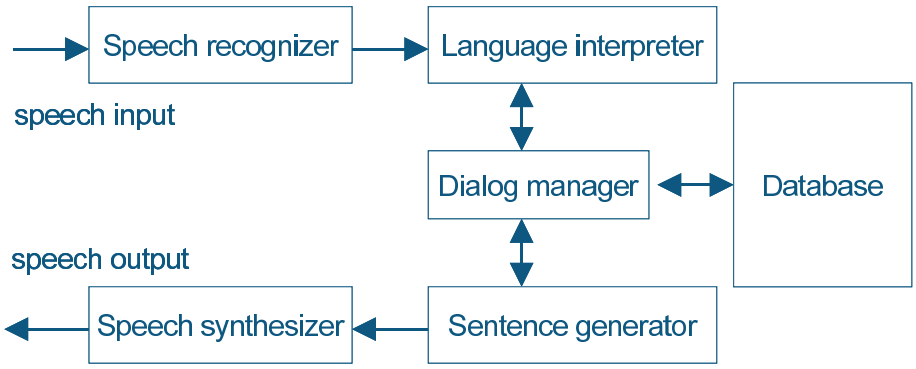


図 2.1: 一般的な音声対話システムの構成

2.2.2　音声認識部

ユーザの発話音声を文字列に変換する役割を担う．音声認識は，音響分析・特徴抽出・探索の 3 段階の処理によって実現される．探索の際に，音響モデル・言語モデルを用いて探索空間を制限する．音響モデルとしては，統計的なモデルである HMM（Hidden Markov Model）が多く用いられる．大語彙連続音声認識を行なうディクテーションソフトなどでは，言語モデルとして，新聞記事などから得られた確率的言語モデルである N-gram（bigram や trigram）を用いるのが一般的である．一方，音声対話システムの場合には，タスクに無関係の語彙はあまり発話されないため，タスクに関係ある単語を用いてのネットワーク文法（有限オートマトン）で言語モデルを記述する場合が多い．

2.2.3　言語理解部

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

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抽出し，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なえない場合は，ユーザに再入力を促すなど，対話の全体的な制御を行なう．

2.2.5　言語生成部

対話制御部から受け取った意味表現を文字列に変換する．

2.2.6　音声合成部

言語生成部より受け取った文字列，を音声信号の形にして出力する．ここでは TTS（Text-To-Speech）システムがよく用いられるが，音声対話システムにおいては，CTS（Concept-To-Speech）システムが望ましいと考えられる．

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することができる．

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．

2.2.8　音声対話システムの例：GALAXY[10]

i)　概要

MIT の Zue らは，音声対話システムアーキテクチャGALAXY を開発した．1989 年以来，開発が行なわれていた VOYAGER[11] などの電話での利用を目的とした音声対話システムを，1 つの統一したアーキテクチャ上に実現した．そのアーキテクチャが GALAXY である．図 2.2 は GALAXY のシステムの図である．

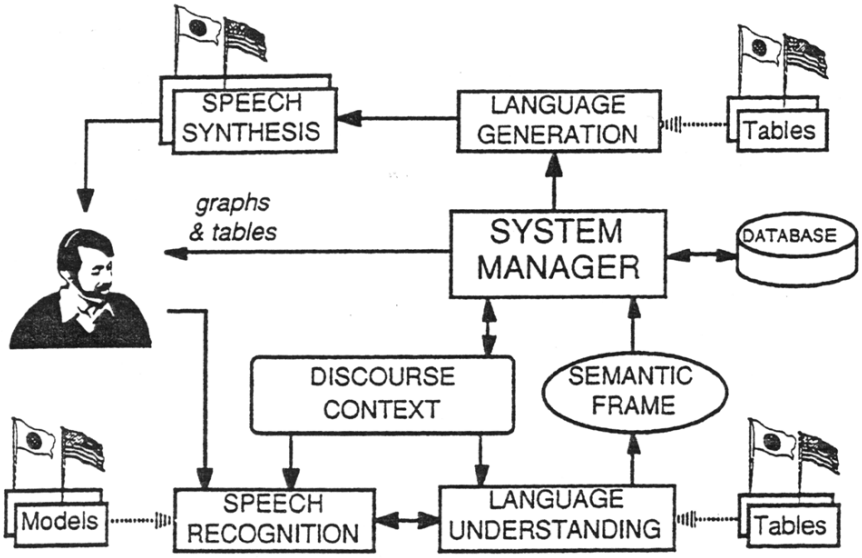


図 2.2: GALAXY のシステム構成

GALAXY は，human-to-computer conversation を可能とする音声対話システムを構築するためのアーキテクチャである．音声認識部に SUMMIT[12]，言語理解部に TINA[13]，言語生成部に GENESIS，音声合成部には市販の音声合成器を用いている．

GALAXY アーキテクチャを用いたシステムには，天気予報を行なう JUPITER，ケンブリッジ市内の案内を行なう VOYAGER，航空便の座席情報を示す PEGASUS[14]，航空便のプランニングを行なう MERCURY などがある．

ii)　音声認識部

音響モデルにより，音声信号を N-best 候補に変換する機能を持つ SUMMIT を用いている．ケンブリッジ市内の案内を行なう VOYAGER を例にとると，“Where is the library

near Central Square?” という音声が入力された場合，SUMMIT は次のような認識候補を出力する．

- Where is the library near Central Square?
- Uh, where is the library near Central Square?
- Where are the library near Central Square?

iii)　言語理解部

言語理解部には，TINA と呼ばれるモジュールが用いられている．入力された認識候補のうち，第 1 候補を品詞に分解し，さらに意味フレームに分解する．上の例では，意味フレームは次のようになる．

```
Clause: LOCATE
  Topic: PUBLIC-BUILDING
    Quantifier: DEF
    Name: library
    Predicate: NEAR
    Topic: SQUARE
      Name: Central
```

iv)　対話管理部

対話管理部では，受け取った意味フレームを SQL query に変換し，それをもとにデータベースを検索し，その結果を意味フレームに変換する．

v)　言語生成部

言語生成部には，GENESIS というモジュールを用いる．GENESIS は，対話管理部から受け取った意味フレームを発話文の形に変換する．上記の例では，発話文は “The library near Central Square is located on Massachusetts Avenue between Green Street and State Street.” といったものになる．

vi)　音声生成部

音声生成部には，市販の音声合成器を用いる．発話文を音声に変換して出力する．

2.3　ソフトウェアロボットによるエージェント音声対話システム

2.3.1　ソフトウェアロボット

仮想世界上に存在するロボットのことをソフトウェアロボットと呼ぶ．ソフトウェアロボットによるエージェント音声対話システムとは，自然言語をインタフェースとしてロボットを制御し，このシステム上における自然言語と空間的位置・エージェントの行動の関係を調べるためのシステムである．

2.3.2　傀儡（かいらい）[6]

i)　概要

新山らは「傀儡」[6] と呼ばれる自然言語処理による音声対話システムを構築した．このシステムは，時々刻々と変化する環境をターゲットとした自然言語処理を行なうことを目的としている．図 2.3 はその実行画面である．計算機上に仮想世界を構築し，そこにいくつかの物体とロボット（ソフトウェアロボット）を配置する．ソフトウェアロボットに対して，日本語で指示を行なうことができる．このシステムでは，仮想空間において，カメラ（ユーザの視点）とロボットとの位置関係に応じて変化する照応・省略の問題を取り扱っている．

図 2.3: 傀儡システム

ユーザやソフトウェアロボットの視界を常に計算しているので，視界に応じた自然言語処理を行なうことができる．図 2.4 にこのシステムの構造を示す．

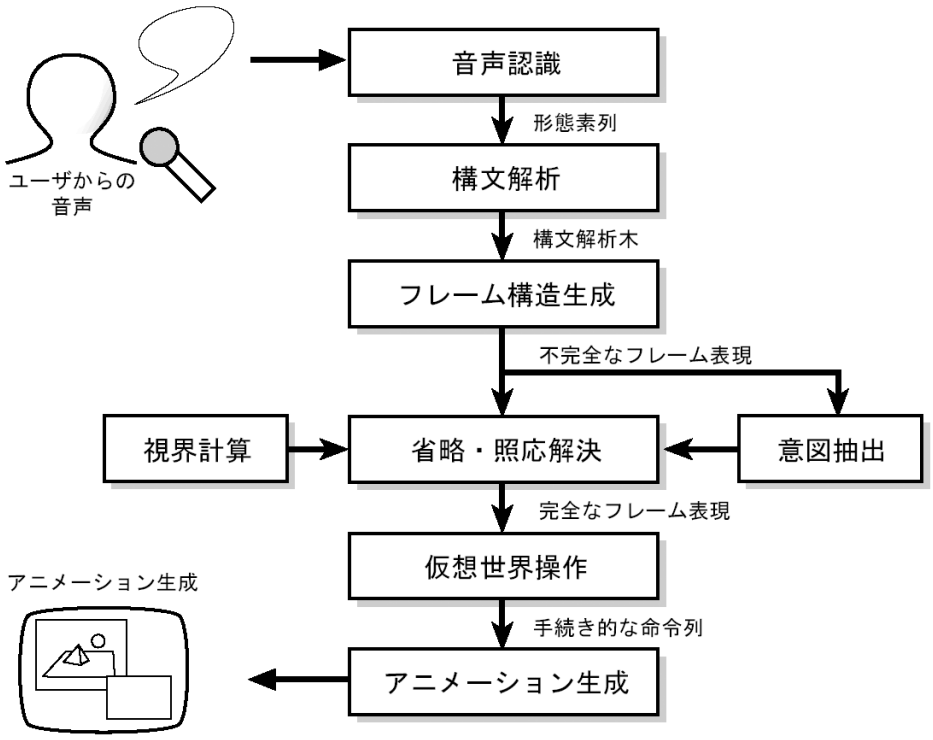


図 2.4: 傀儡のシステム構成

ii)　音声認識部

市販の IBM ViaVoice を用いている．

iii)　意味理解部

構文解析とフレーム構造生成を行なう．また，フレームが不完全だった場合は，意図抽出を行なう．さらに，ユーザやソフトウェアロボットの視界を考慮し，省略・照応の解決を行なう．

iv)　対話制御部

ユーザの意図に応じて仮想世界を操作し，アニメーションを作成する．

v)　マルチモーダルインタフェース

省略・照応を解決するために，発話情報だけではなくユーザの視界を用いている．ユーザの視界の中心に映っている物体は，コ系やソ系の直示的照応詞で指示するといったこと

を考慮している．ユーザが「ロボット A はそれを押して」と発話し，先行詞「それ」の候補となる名詞が現れていない場合などは，ユーザの視界を考慮して「それ」の指す先を同定する．

2.4　省略・照応の解決

2.4.1　入力における省略・照応の解決

人間同士が対話を行なう場合，相手がわかっていることは省略され，相手と共有しているものに関しては，指示語が用いられる．音声対話システムは当然，何が省略されているのか，指示語が何を指しているのかを解決しなければならない．例えば以下の文章を例に考えてみる．

A: 「机の上にあるものは何?」

B: 「花瓶だよ.」

A: 「それを椅子のそばに持ってきて.」

B: 「わかった.」

A: 「あと，お皿も.」

A の 2 回目の発話文で「それ」が出てくるが，これはもちろん「花瓶」のことを指す．A の 3 回目の発話では,「あと，お皿も持ってきて」の「持ってきて」が省略されている．

これらの省略・照応を解決するには，対話の履歴を参照すればよい．[6] では，スタックを用いてこれらの問題を解決している．この方法では，まず自然言語表現を一度フレーム構造 [15] に変換する．発話行為の分析や照応の解決といった作業は，全てこのフレーム構造を用いて行なわれている．対話履歴を参照し，これまでの発話のフレーム表現から現在の発話のフレームのうち，省略されているもの，照応されているものを補って解決することができる．

例えば,「ロボット A は球を押して」という文のフレーム構造は,

agent:　ロボット A

object:　球

verb:　押す

となる．そして「それを押して」という文が入力されたとき，そのフレーム構造は

agent:　（不明）

object:　それ

verb:　押す

となる．省略されている agent フレームには，スタックに格納された過去の agent フレームを参照して「ロボット A」を補完し,「それ」という指示語で照応されている object フレームは「球」に置き換えることで「ロボット A は球を押して」という文が生成され，省略・照応を解決し，システムはその動作を決定することができる．

また，[5] では，情報の要素ごとにスタックを用意している．要素とは，文献一覧の番号と質問の内容である．適宜スタックを参照することで，省略されている語を補うことが可能となる．

2.4.2　出力における省略

システムの応答にユーザにとって分かりきった情報が入っていると，ユーザにとってはくどいものになり，ユーザに不快感を与える．さらに，必要としている情報が，この分かりきった情報の中に埋没してしまう可能性がある．また，音声の対話では，文字による対話とは異なり，過去の対話が記録として残らないので，省略表現を多用しすぎると対話内容が曖昧になり，ユーザに正しく伝わらない可能性がある．従って，システムの応答は適切な省略を伴ったものでなくてはならない．

[5] では，ユーザの質問文に含まれない要素を補完し，ユーザが発言した要素を省略するのがもっとも好ましい，という実験結果を出している．文献検索を例に取ると,「3 番」の項目について話をしている時に,「著者名は何ですか」とユーザが質問した場合，システムは「著者名は○○です」と答えるのではなく，ユーザの省略した項目である 3 番を補い，ユーザの発話した「著者名」という情報は省略して,「3 番は○○です」と答えることで，これがもっとも好ましい応答である，としている．

2.5　応答文生成

多くの音声対話システムで用いられている，テキスト形式の表層文から音声を合成する TTS による方式では，多様な応答文に対して適切な韻律情報を付与した応答音声を出力することは困難である．一般的に音声対話システムにおいては，応答内容の決定にあたり，その時点までの対話履歴を参照することが可能であり，対話の焦点などの情報をこの対話履歴から抽出して応答に反映させる方式が有効である，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音声対話システムについて，その概要を述べる．

2.5.1　談話情報を用いた音声合成における韻律の制御 [16]

i)　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 [16]．システムの概要を図 2.5 に示す．

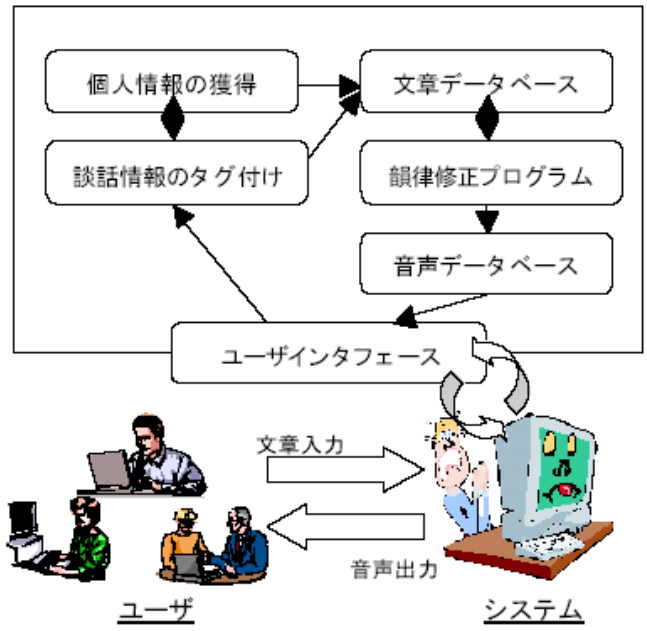


図 2.5: システム概要

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．本研究における「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．保持するデータには要約を施したものを用いる．テキストに出現する語の統計的性質を利用した N-gram モデルの考え方を基に話題の展開を考え，その配置を行なう [17].

iii)　談話情報のタグ付け

談話情報として用いる対人態度，心理のタグセットの一部を表 2.1 に示す.

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.「相手への態度」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

表 2.1: タグセットの一部

自分の態度	相手への態度	心理
確信度	支配・指図	怒り
意欲/興味	批判・不同意	悲しみ
素直	譲歩・撤回	喜び
真剣	同意・協力	驚き
	親切・親しみ	
	冷淡	
	失礼・ぶしつけ	
	皮肉	

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 [18].

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 [19].

v)　出力フェーズ

システムからの出力は，話題の流れに沿って行なわれる．韻律修正された発言を，その話題，その談話情報タグセットとともに出力する．話題の遍歴はもちろん，タグセットの履歴を示すことで，全体における発言の位置付けを確認できる．

2.5.2　学術情報検索音声対話システム [5]

i)　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 [5] を開発した．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．このシステムの構成図を図 2.6 に示す．また，このシステムの画面表示の様子を図 2.7 に示す.

ii)　音声認識部

音声認識部には，連続音声認識パーザ Julian v.2.2[20] を用いている．Julian とは,「日本語ディクテーション基本ソフトウェア」である Julius の言語モデルの代わりに，ネットワー

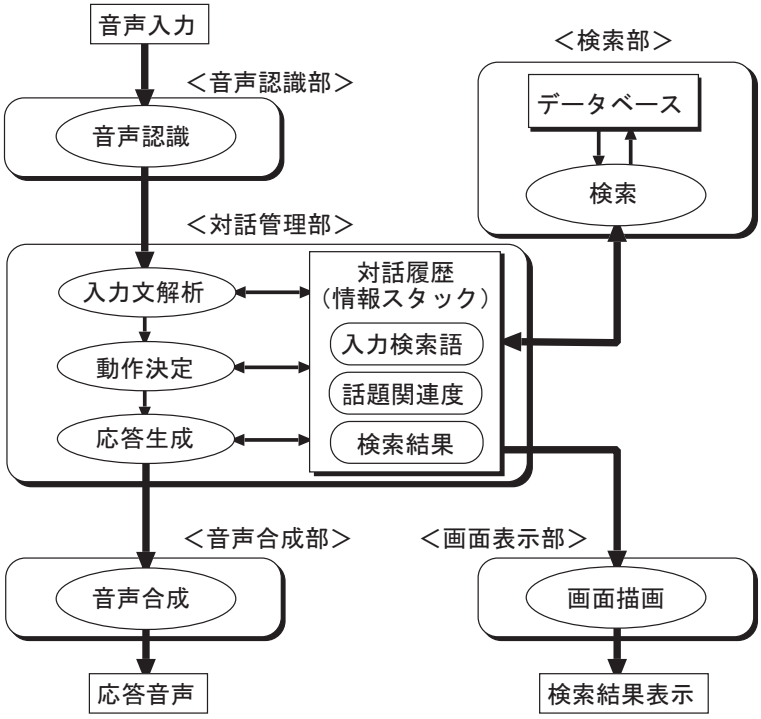


図 2.6: 学術情報検索音声対話システムの構成

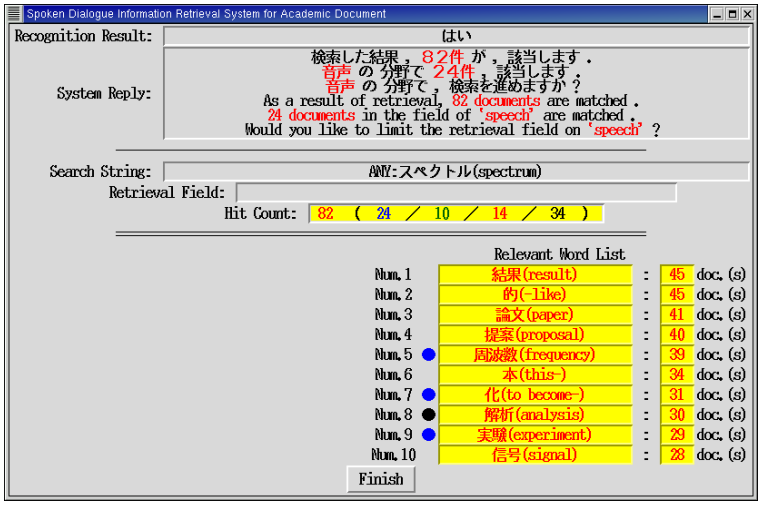


図 2.7: システムの画面表示

ク文法による言語モデルを用いる音声認識ソフトウェアである．音響モデルは,「日本語ディクテーションソフトウェア 98 年度版」[21] に含まれる，状態数 1000，混合数 4 の triphone モデルである．言語モデルとしては，タスクに依存した文脈自由文法（CFG : Context Free Grammar）を作成したものが用いられている．この CFG は，語彙数：425，構文規則数：88，終端シンボル数：84 である．

iii)　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．

iv)　韻律規則

[5] では，[22] の韻律規則を用いて音声合成を行なっている．[22] は，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，朗読調の音声に対して構築された規則 [23] 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・起伏型のアクセント立ち上げと，両者の立ち下げを表す 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．詳しくは 5.3 節で述べる．

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．

v)　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するものである．入力側の抽象的な文概念を抽象文概念コードと呼んでいる．これに対して，応答文型を定める文概念を決定文概念コードと呼び，便宜上両者を区別している．文概念は，システムに依存したものとなっている．例えば定型文，検索語入力，実行命令通知，といった具合である．

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，記述したものである．

単語コード　単語を辞書引きするためのコードである．辞書には，単語を画面表示するための単語辞書と，その単語の韻律を表現するための韻律句辞書がある．

ユーザの「著者名は何ですか」という質問に対し，文献番号を補って「3 番は，広瀬啓吉，峯松信明です」という応答文を生成する過程を図 2.8 に示し，以下に流れを述べる．

1. “質問応答” という抽象文概念の入力を,「抽象文概念コード：07」という形式で受け取る．
2. 情報スタックを参照して,「文献リスト 3 番の著者名を，文献番号を付加して回答する」という応答文内容が決定され，これを表す「決定文概念コード：0721」が生成される．
3. 文概念辞書により，文概念コードから韻律句コード列を得る．
4. 韻律句辞書により，韻律句コードから単語コード列を得る．情報スタックを参照して，単語分類コードを単語コードに変換する．
5. 単語辞書により，単語コードを単語の見出しに変換して表層文を得る．
6. 単語辞書から各単語の韻律情報を読み込む．
7. 抽象文概念が “質問応答” であるため，焦点位置を質問の回答情報にあたる韻律句に決定する．
8. 韻律句コード単位で文全体を走査し，韻律規則にのっとってフレーズ指令をしかるべき位置に立てる．
9. 直前のフレーズ指令からの位置と単語の品詞情報，ならびに焦点情報に依存するパラメータ値を算出し，アクセント指令を決定する．
10. 韻律句ごとの記号列をつなぎ合わせ，文全体の音韻・韻律記号列を生成する．

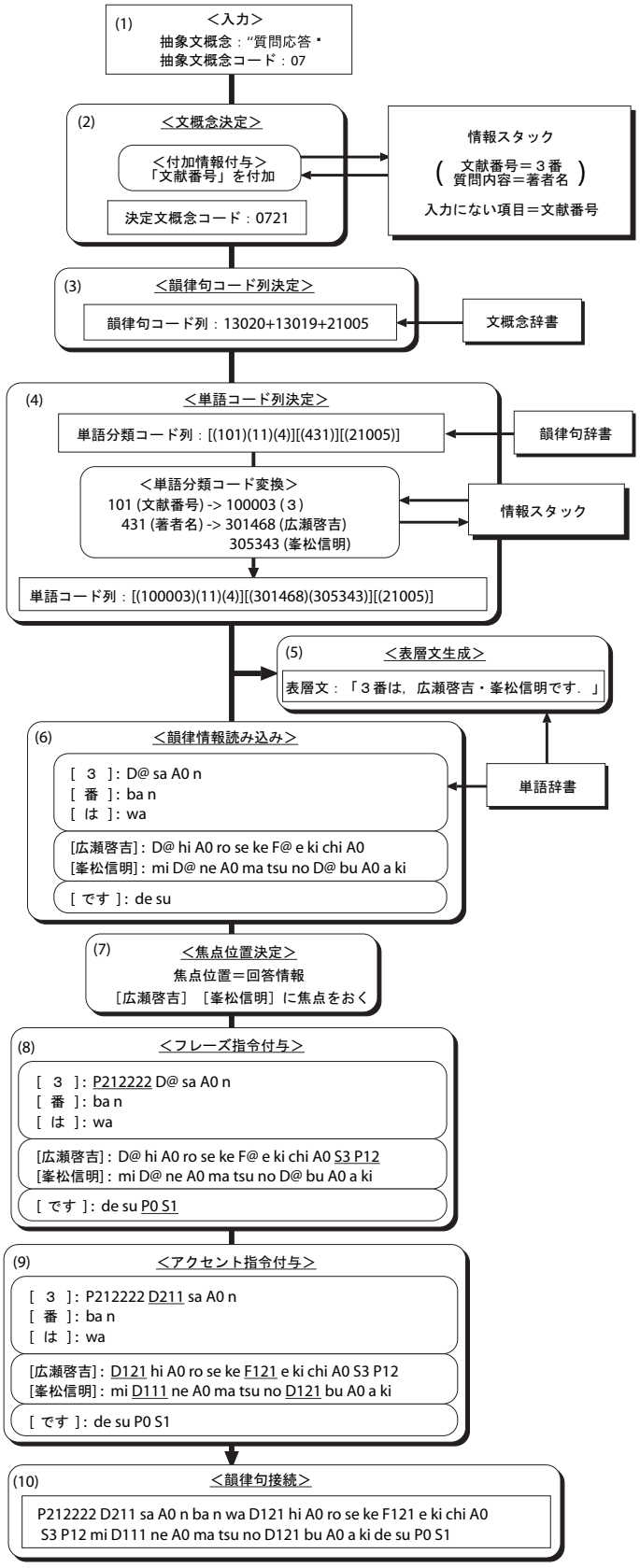


図 2.8: 文献検索システムにおける応答文の生成過程

2.5.3　エージェント音声対話システム [7]

i)　概要

多胡らはエージェントによって仮想空間中の物体を操作することのできるエージェント音声対話システム [7] を開発した．操作は，仮想空間にいるエージェントに自然言語で指示することで行なう．仮想空間の状態はリアルタイムで画像として出力され，ユーザはそれを見ながらエージェントに質問や指示などを行なうことにより物体を操作する．このシステムの概念的な構成は図 2.9 のようになる．また，モニタ出力の様子を図 2.10 に示す．

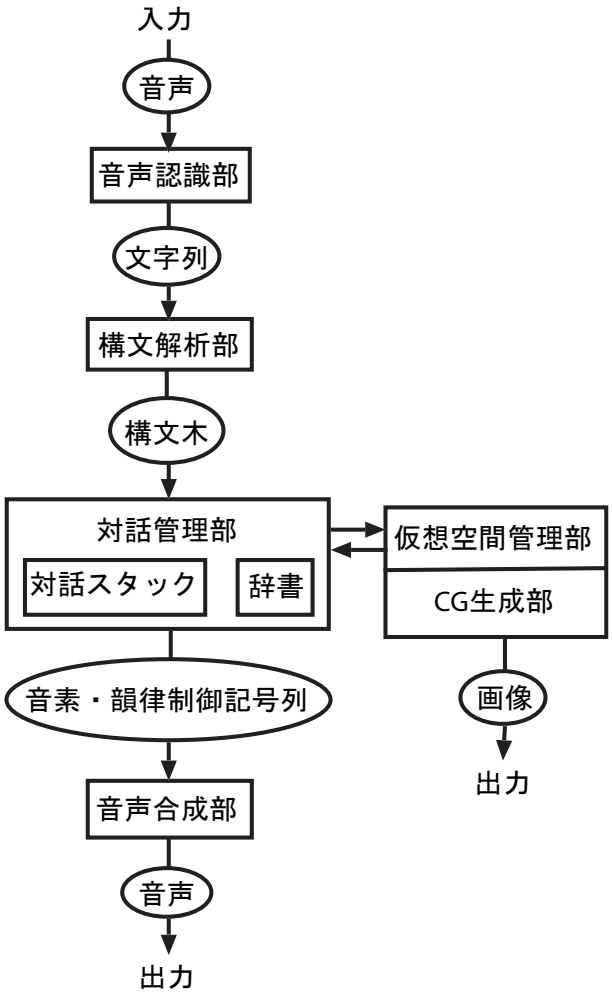


図 2.9: エージェント音声対話システムの構成

図 2.10: エージント音声対話システム

ii)　音声認識部

音声認識部には，Julian v.3.2[20] を用いている．音響モデルとしては，状態数 1000，混合数 4 の triphone モデルを用いる．言語モデルとしては，タスクに依存した文脈自由文法を作成したものが用いられている．音声認識部は，ユーザの発話文を文字列に変換して，その結果を構文解析部に渡す．

iii)　構文解析部

構文解析部は，音声認識部より受け取った文の形態素解析・構文解析を行ない，構文木構造を持った文として対話管理部に渡す．

iv)　対話管理部

対話管理部では，構文解析された入力文により，現在の空間の状態に応じて応答文の生成，エージェントの動作指示などを行なう．また，対話の必要がある場合は，応答文を生成して対話を開始する．その際には，韻律制御記号を含む音素記号列を音声合成器に出力する．また，対話のログを取るのもこのモジュールである．

v)　対話スタック

現在までの対話を記憶するためのスタック．対話管理部は，アイテム・場所などの主要な項目についてもスタックを持つ．

vi)　辞書

対話管理部は，ユーザ発話文の理解・応答文生成のための辞書を持つ．この辞書は，構文解析部でも用いられる．

vii)　音声合成部

音声合成部では，対話管理部より受け取った韻律制御記号を含む音素記号列から，音声を合成する．

viii)　仮想空間管理部

仮想空間管理部は，空間の状態を管理し，対話管理部からの指令により，エージェントの動作を行なう．また，対話管理部に空間状態の情報を提供する．エージェントのパスプランニングもこのモジュールが行なうので，対話管理部はエージェントのパスに気を配る必要はない．

ix)　CG 生成部

CG 生成部は，空間の状態をリアルタイムに描画する．描画には，OpenGL API[43] を用いている．

2.6　マルチモーダル

2.6.1　擬人化音声対話エージェントツールキット Galatea[24]

i)　概要

嵯峨山らは，擬人化音声対話エージェントのソフトウェアツールキット “Galatea[24]” を開発した．このツールキットは，オープンソース，ライセンスフリーであるのが特徴である．

このツールキットは，以下のような特徴を備えている．

- 高いカスタマイズ性（顔，合成音声，認識文法，対話制御等）
- 標準化動向に対応（Voice XML[25]，W3C[26]，JEIDA-62-2000[27] 等）
- 簡明なモジュール通信，部品交換が容易，モジュール別に別々の PC に分散して実行可能
- 最新の高度な技術内容を実現．特に，初の無償の日本語テキスト音声合成システムが含まれている

- ソース公開，無償使用許諾

全体構成を図 2.11 に示す．基本的な構成では，対話音声認識モジュール（SRM），対話音声合成モジュール（SSM），顔画像合成モジュール（FSM）の 3 機能モジュールをモジュール統合処理部（Agent Manager : AM）が統合し，タスク制御モジュール（TM）あるいは対話制御モジュール（DM）の下で動作する．

各モジュールは独立したプロセスとして，単一の PC，もしくは複数の PC 上で並行に動作することを想定している．モジュール統合処理部は，各モジュールが連動して 1 つの音声対話システムとして円滑に動作するためのシステム制御，情報管理等を司る．

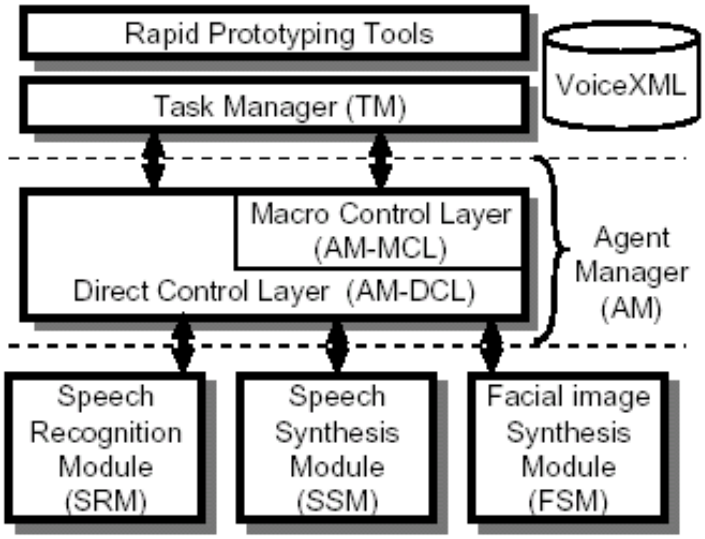


図 2.11: GALATEA の全体構成

ii)　音声認識モジュール

音声認識モジュールは，音声認識エンジン，通信・制御モジュール，文法変換モジュールの 3 つのサブモジュールからなる．音声認識実行部と通信・制御部を独立させることで，外部プログラムと音声入力から非同期に発生する通信イベントと音声入力イベントに対してそれぞれ専属のプロセスを割り当て，イベントの取りこぼしや遅延を防ぐ設計となっている．

モジュールの中心となる音声認識部において，音声認識エンジンは Julian[20] を想定しているが，文法や出力形式のインタフェースを汎用的なものとし，他の認識エンジンに置き換えても外部モジュールからは等価に扱える仕組みになっている．

認識対象とする発話の語彙や構文規則は，外部モジュールから与えられる．Julian はオートマトン文法のみを扱うので，文法は専用コンパイラによって有限状態オートマトン（FA）

に変換される．音響モデルはサブワード単位の HMM を用いる．ファイルのフォーマットは標準的な HTK[28] のフォーマットに対応する．

iii)　対話音声合成モジュール

Galatea における対話音声合成モジュール（Galatea Talk）は，漢字仮名混じり文で表記された日本語テキストを合成音声に変換する，いわゆる日本語テキスト音声合成を行なう基本的な機能

1. 形態素解析

2. 読み，アクセント型の付与

3. 韻律生成

4. 合成波形生成

5. 合成音声出力

に加えて，顔画像生成を伴う音声対話システムを構成するための音声合成モジュールとして，以下の機能

6. 出力発話（合成音声）における各音素の継続時間長の出力

7. 埋め込みタグによる韻律の制御

8. 音声出力の途中停止，及び中断における既出力音素列の出力

を持つ．6. は顔画像出力における口唇の動きと合成音声を同期させるために用いられる．

1.，2. では，アクセント情報を付加した辞書を用いて “茶筅 [29]” で形態素解析したのち，[30] で示されるアクセント処理を行なう．3.，4. では，HMM に基づいた音声合成 [31, 32, 33] により，合成波形を生成する．音声合成部で必要となる話者の音響モデルとしては，男女各 1 名の基本話者のモデルが提供される．

Galatea Talk は，独立した 4 つのモジュール，コマンド解析部，テキスト解析部，音声合成部，音声出力部からなり，図 2.12 の構成をとる．

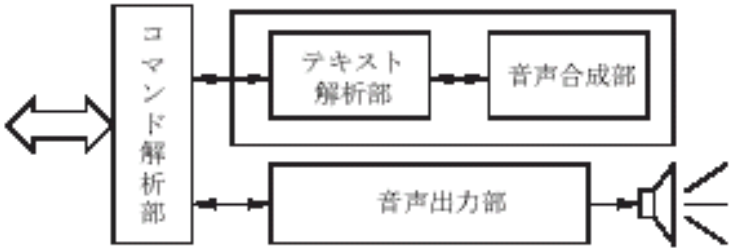


図 2.12: Galatea Talk の構成

Galatea Talk では，音声出力する発話文の内容は，set コマンドで Text スロットの値を設定することによって行なう（例：「set Text = こんにちは.」）．発話文の表現形式としては，プレインテキストによる漢字仮名混じり文に加えて，7. の機能を実現するために，[27, 34] におけるテキスト埋め込み制御タグ及び仮名レベルの韻律記号に準拠したタグ付きテキストを受け付ける [35]．この記述例を図 2.13 に示す．

```
<SPEECH> <VOICE OPTIONAL="male1">
これは<PRON SYM="アイピーエー">IPA</PRON>のプ
ロジェクトで開発された<EMPH>対話</EMPH>音声合成
システムです。
</VOICE> </SPEECH>
```

図 2.13: Galatea Talk による発話文の記述例

iv)　顔画像合成モジュール

Galatea の顔画像合成の基盤として用いたのは，IPA プロジェクト「感性擬人化エージェントのための顔情報処理システムの開発」（1995.6∼1998.3）で開発したソフトウェア [36] である．このソフトウェアは無償公開されており，正面方向から撮影した 1 枚の顔画像と標準顔モデルを整合させ，各個人のモデルを生成できるため，顔画像を準備するだけでエージェントの顔をカスタマイズできる．今回新たに，より人間らいし対話を実現するために，精密な Lip Sync のための他のモジュールとの連係，喜びや怒りを表現するための任意の表情付加機能，自然な瞬きの制御機能を付加している．

2.6.2　富士山観光案内音声対話システム [4, 37]

i)　システム概要

伊藤らは，音声対話システムにおいてユーザの自由な発話を許す，より頑健な言語理解手法を実現している．図 2.14 にシステムのモニタ出力の様子を示す．

ii)　人間の理解手法を用いた頑健な言語理解手法

ユーザに自由な発話を許す音声対話システムの構築には，誤りを含んだ認識結果を解析する必要がある．そのため，誤認識文から発話文の意味を復元する機構を音声対話システムに組み入れる必要がある．

この研究では，まず人間に音声認識システムによって得られた誤認識を含んだ認識結果を見せ，元の文を復元する実験を行なっている．この実験では，音声認識システムだけでの平均文認識率は 57.4%であったが，認識結果からのエキスパートの復元訂正によって意味理解率は 87%まで向上している．

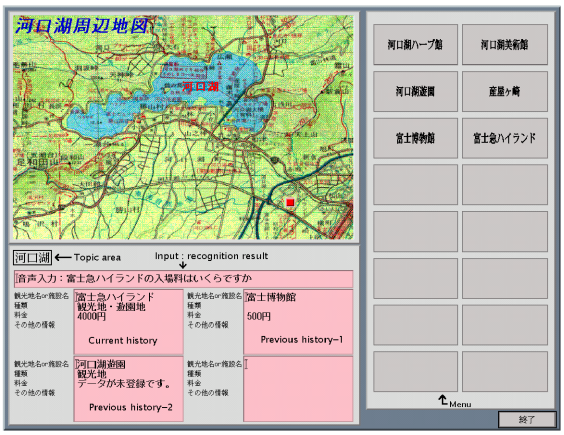


図 2.14: 富士山観光案内システムのモニタ出力

この実験によって獲得されたストラテジーを元に，言語理解手法を構築し，音声対話システムに組み入れた．その結果，助詞落ちや言い直し，間投詞を含む初心者が発声した自然言語に対して，文認識率 52%，意味理解率 72%を得ることができ，自然な発話や誤認識を含んだ認識文に対してある程度のロバスト性を持たせることができた，としている．

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，ソフトウェアロボットによるエージェント音声対話システムを構築するにあたり，必要となる要素技術について述べた．具体的には，省略・照応の解決，応答文生成について述べた．

音声対話システムにはまだ不完全な部分が多い．省略・照応の解決については，ここで述べた手法が必ずしも適用できるとは限らず，むしろ適用できない場合の方が多い．応答文生成に関しては，韻律に着目した応答生成を行なうシステムはまだまだ少ないのが現状である．本節で述べたシステムのほとんどが，実際の応答文生成では単純に単語を並べるだけで，文の統語構造までは扱っておらず，アクセント結合なども考慮されていない，という問題がある．[7] や [24] では，アクセント結合は考慮されているが，それがきちんと応答に反映されているとは言い難いのが現状である．応答文生成について，次章で統語構造や談話情報など高次の言語情報の取り扱い手法について説明し，またアクセント結合なども考慮した，より一般的な応答音声生成手法を第 4 章にて提案する．

第3章

道案内音声対話システムの概要

3.1　はじめに

本章では，道案内対話システムの概要について述べる．まず，本対話システムのタスクとインターフェースについて述べる．さらに，本対話システムの構成について述べ，最後に，誤解の解決手段について述べる．

3.2　道案内音声対話システムのタスク

本章で構築する道案内音声対話システムのタスクは，仮想地図内をシステムの案内対話によって，ユーザの目標地点への到着を目指すことである．システム・ユーザ間の情報のやりとりは全て自然言語音声によって行なう．ユーザは仮想地図内を移動し，得た情報を基にシステムに質問や指示を行い目標地点を目指す．

3.2.1　地図

本対話システムのタスクに持ちられる地図を図 3.1 に示す．地図の主な特徴としては，

- 交差点や三叉路などに目標物が設定されている．
- 目標物間にリンクが張られている．
- 目標物間に距離情報が設定されている．

などがあげられる．

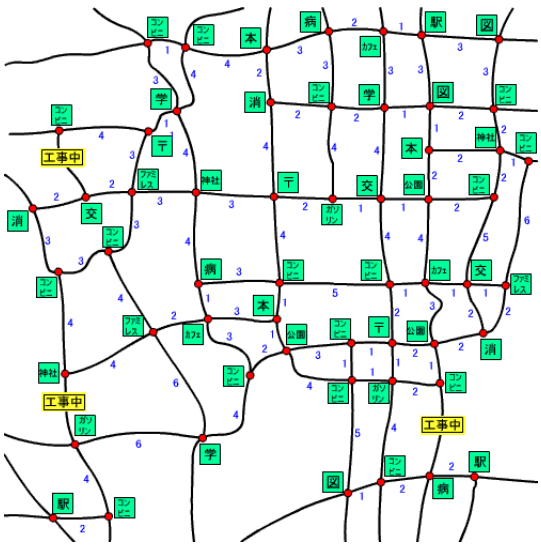


図 3.1: 地図

3.2.2　扱う主な項目

本章で構築する道案内対話システムは，主に目標物・場所および方位・向きという 4 項目を扱う．

i)　目標物

目標物は，店舗など仮想空間中の施設の種類を表すものである．同一の名称の目標物（図書館など）は複数存在するため，目標物名だけでユーザの位置を特定することはできない．

ii)　場所

場所は，仮想地図内の位置を一意に表すものである．システムは，仮想地図内の全て目標物について一意のノード番号を保持している．また，ノード間のリンク情報の行列情報により，システムは仮想地図の全体像を把握している．

iii)　方位

仮想地図上の東西南北に相当する情報であり，システムはこの方位に基づいてユーザへの指示を行なう．

iv)　向き

向きは，仮想地図内でのユーザからみた相対的な方向（左右前後）を表すものである．例えば,「右に曲がってください」という場合，直前の場所から現在の場所到着した瞬間に向いている方向が「前」となり，これを基準に前後左右が決定される．システムは，次にユーザが進むべき「方位」を把握しているが，この方位がユーザにとってはどの「向き」になるのかを変換して指示を行なう．

3.2.3　ユーザインタフェース

ユーザには地図の全体像は分からず，視界に相当する部分（図 3.2 赤丸の内側）の情報のみ得ることができる．赤丸は，マウス操作により移動することができ，ユーザの地図上の移動による視界の変化はこの操作により代替される．

3.2.4　システム・ユーザ間の道路情報の差異

道路工事による通行止めなどの一時的な道路状況の変化について，その情報をシステムは保持しないが，ユーザが実際に道路上を移動する上では遭遇する状況である．こうした突発的な障害により，システム・ユーザ間の情報に差異を生じさせることで，システム・ユーザ間に誤解を誘発し，より複雑な対話を実現する．

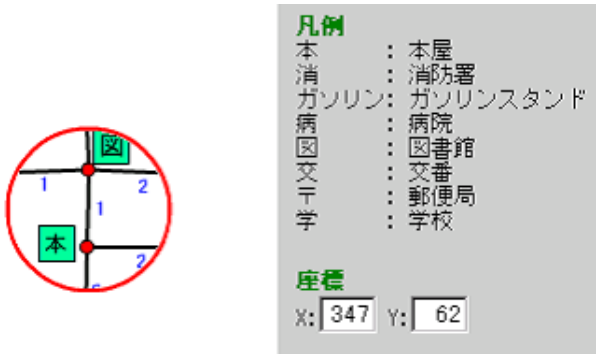


図 3.2: ユーザインタフェース

3.2.5　最短経路探索

システムは，あらかじめユーザの現在地点と目標地点を把握し，この間の最短距離を計算することにより，ユーザに道案内を行なう．最短距離の計算方法は，目標物をノード，施設間の道路をリンクとし，ダイクストラの最短経路探索アルゴリズムを使っている．

3.3　道案内対話システムの構成

本対話システムの概念的な構成はおおむね図 3.3 のようになる．
以下，各モジュールについて説明をする．

3.3.1　音声認識部

音声認識部には，Julian v.3.2[20] を用いる．Julian とは,「日本語ディクテーション基本ソフトウェア」である Julius の言語モデルの代わりに，ネットワーク文法による言語モデルを用いる音声認識ソフトウェアである．音響モデルとしては，状態数 1000，混合数 4 の triphone モデルを用いる．言語モデルとしては，タスクに依存した文脈自由文法（CFG : Context Free Grammar）を作成したものが用いられている．音声認識部は，ユーザの発話文を文字列に変換して，その結果を構文解析部に渡す．

3.3.2　構文解析部

構文解析部は，音声認識部より受け取った文の形態素解析・構文解析を行ない，構文木構造を持った文として対話管理部に渡す．

3.3.3　対話管理部

対話管理部では，構文解析された入力文により，現在のユーザの状態に応じて応答文を生成する．その際，韻律制御記号を含む音素記号列を音声合成器に出力する．

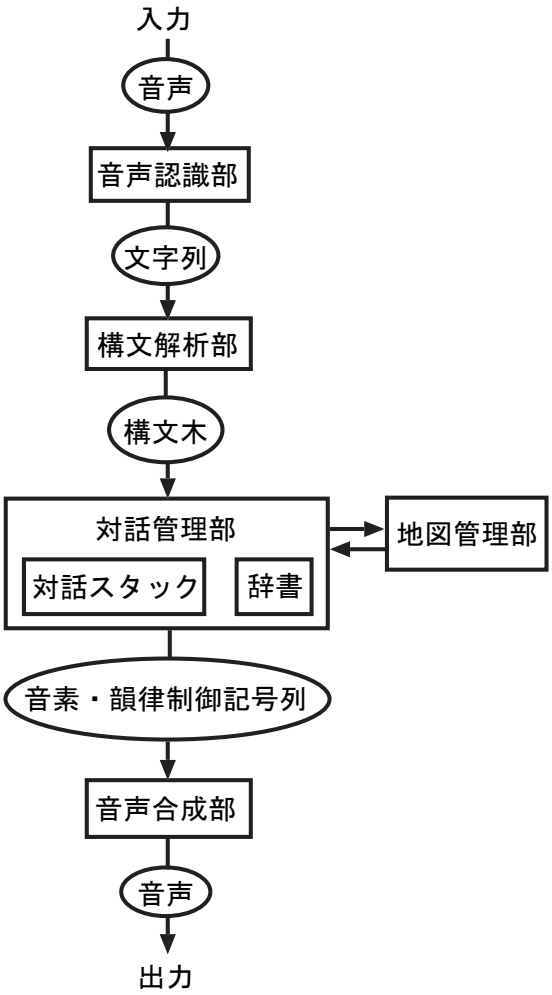


図 3.3: 道案内対話システムの構成

i)　対話スタック

現在までの対話を記憶するためのスタック．対話管理部は，主要な項目（直前にユーザがいた場所など）についてもスタックを持つ．

ii)　辞書

対話管理部は，ユーザ発話文の理解・応答文生成のための辞書を持つ．この辞書は，構文解析部でも用いられる．

3.3.4　音声合成部

音声合成部では，対話管理部より受け取った韻律制御記号を含む音素記号列から，音声を合成する．（本対話システムで用いる音声合成器については，5.4.2 節で後述）

3.3.5　地図管理部

地図管理部は，地図を管理し，対話管理部からの指令により，現在地情報の提供と，と次に進むべき方位の決定を行なう．道案内の経路プランニングもこのモジュールが行なうので，対話管理部は道案内の経路に気を配る必要がなく，受け取った情報のみからユーザへの指示を決定することができる．ここで管理しているのはあくまで方位であり，ユーザの向きという相対的な情報については対話管理部で処理する．

3.4　誤解の解決

道に迷うなどしてシステム・ユーザ間の現在位置情報が相違した場合，システムは対話を誤解解決のためフェーズに移行する．ここで行われるとは，1) システムがユーザの現在位置情報を把握し，2) 新たに得られた現在位置情報から，目標地点までの最短経路を再探索する，という 2 点である．1 のために，システムは以下のような方法をとる．

1. ユーザから得られた現在位置の目的物情報が，システムの想定していた現在位置の隣のノードの目的物名と位置する場合，ユーザによる「向き」の聞き取りエラーと判断し，ユーザへの指示なくシステムの現在位置を当該近隣ノードに移動する．
2. ユーザから得られた現在位置の目的物情報が，システムの想定していた現在位置と直前の位置の同一方位上にある場合，ユーザの目的物の見落としと判断し，ユーザへの指示なくシステムの現在位置を当該同一方位上ノードに移動する．
3. 上記のどちらでもない場合，ユーザに対して道に迷っていることを告げ，さらに直前に通過した目的物名を問い合わせる質問を行なうことにより，ユーザの現在位置の目的物名と質問によって得られた目的物が直線上に隣り合って存在する地点を検索し，当該ノードの現在位置を移動する．

3.5　対話例

対話例を以下に示す．U はユーザ，S はシステムである．システムの応答のうち，本システムにおいて重要であるとみなされる（焦点の当たる）語は**太字**で記してある．

S:　**左**に曲がって**二つめのコンビニ**まで行って下さい

U:　**コンビニ**に着きました

S:　**右**に曲がって**病院**まで行って下さい

U:　**右**に曲がることができません

S:　**まっすぐ**行くことはできますか

U:　はい

S:　では，**まっすぐ学校**まで行って下さい

⋮

S: **右**に曲がって**駅**まで行って下さい

U: **駅**に着きました

S: そこが目標地点です

3.6　まとめ

本章では，本研究のために開発された道案内対話システムの概要について述べた．まず，本対話システムのタスクとインターフェースについて述べた．さらに，本対話システムの構成について述べ，最後に，誤解の解決手段について述べた．また，本対話システムにおいて想定される対話例も示した．

第4章

道案内音声対話システムおける言語情報の取り扱い

4.1　はじめに

音声対話システムで柔軟な応答音声の生成を実現するためには，言語情報は構文木構造で扱うのが自然であるが，従来の音声対話システムでは，言語情報を単なる単語の一次元列として扱ってきた．本研究で構築した音声対話システムでは，言語情報を一貫して構文木構造を保持したまま扱うことで，柔軟で容易な応答文生成を実現する．また，言語情報の統一的な処理のために，タグを付与し，言語情報の取り扱いを容易にする．

本章では，対話システムにおける言語情報の取り扱い手法として，[7] の手法を拡張した手法について述べる．まず，本手法で想定する音声対話システムの構成を示し，続いて本手法で用いる辞書の構成について述べ，内部表現への言語情報の付加について述べる．さらに，文の接続方法や音声合成についても説明する．

4.2　辞書

言語情報を取り扱う上で，辞書は必要不可欠である．辞書は，ユーザの発話の理解と応答文生成に用いる．辞書には以下の 3 種類があり，それぞれの辞書は拡張性を考慮して XML[38] で記述されている．

品詞辞書　品詞情報を格納

活用辞書　活用型・活用形を格納

単語辞書　個々の単語を格納

4.2.1　品詞辞書

品詞情報を格納する．例えば,「名詞」は以下のように記述する．

```
<node>
  <name>名詞</name>
  <independence>YES</independence>
  <accent_3rd_param>1</accent_3rd_param>
  <connection>基本形</connection>
  <node><name>一般</name></node>
  <node>
    <name>固有名詞</name>
    <node><name>一般</name></node>
    <node>
      <name>人名</name>
      <node><name>一般</name></node>
      <node><name>姓</name></node>
      <node><name>名</name></node>
```

```
    </node>
  </node>
  <node>
    <name>代名詞</name>
    <accent_3rd_param>4</accent_3rd_param>
    <node><name>一般</name></node>
    <node><name>縮約</name></node>
  </node>
  <node>
    <name>非自立</name>
    <node><name>一般</name></node>
  </node>
</node>
```

品詞（node）は，以下のようなパラメータを持つ．

name	品詞名
independence	自立語（YES）or（NO）
accent_3rd_param	アクセント指令の 3 番目のパラメータ（品詞を表す）
connection	接続を表す

品詞は入れ子状にすることができ，省略したパラメータは親の品詞（自分を内包するノード）を参照する．

4.2.2　活用辞書

活用型・活用形を表す．例えば，一段活用の場合は以下のように記述する．

```
<node>
  <name>一段</name>
  <form> <name>基本形</name> <display>る</display>
         <phoneme>ru</phoneme> </form>
  <form> <name>未然形</name> <display>*</display>
         <phoneme>*</phoneme> </form>
  <form> <name>未然ウ接続</name> <display>よ</display>
         <phoneme>jo</phoneme> </form>
  <form> <name>連用形</name> <display>*</display>
         <phoneme>*</phoneme> </form>
  <form> <name>仮定形</name> <display>れ</display>
         <phoneme>re</phoneme> </form>
```

```
  <form> <name>命令 yo</name> <display>よ</display>
         <phoneme>jo</phoneme> </form>
  <form> <name>命令 ro</name> <display>ろ</display>
         <phoneme>ro</phoneme> </form>
  <form> <name>仮定縮約 1</name> <display>りゃ</display>
         <phoneme>rja</phoneme> </form>
  <form> <name>体言接続特殊</name> <display>ん</display>
         <phoneme>n</phoneme> </form>
</node>
```

活用型（node）は以下のようなパラメータを持つ．

name　活用型名

form　活用形（複数持つことができる）

さらに，活用形（form）は以下のようなパラメータを持つ．

name　　活用形名

display　活用形の表示する場合の文字列（*の場合は何も表示しない）

phoneme　発音する場合の音素記号列（*の場合は発音しない）

本手法において，辞書は音声合成にも用いられるので，音素記号を記述する必要がある．

4.2.3　単語辞書

単語辞書には，それぞれの単語の持つ情報を格納している．例えば,「進む」の場合は以下のように記述する．

```
<node>
  <identifier>進む</identifier>
  <part>動詞<part>自立</part></part>
  <stem>進</stem><stem_read>ス ス</stem_read>
  <phoneme>su su<accent_core>0</accent_core></phoneme>
  <inflection>五段・マ行</inflection>
  <dialog_data>
    <identity>go</identity>
    <attribute>action</attribute>
  </dialog_data>
</node>
```

単語の持つパラメータには，以下のようなものがある．

identifier	単語を特定する.
display	単語を表示する時に用いる．省略した場合は identifier が代わりに用いられる.
part	単語の品詞を表す．入れ子にすることが可能である.
stem	単語の語幹を表す（活用語のみ）.
inflection	活用型を表す（活用語のみ）.
connection	単語の接続を表す．省略した場合は品詞に登録された接続を用いる.
phoneme	単語の発音を音素記号列により示す．アクセント核（accent_core）の情報も含む.
dialog_data	対話用データ（対話システム特有の情報を含む）

対話用データは，対話システムに応じて値を設定し，ユーザ発話の理解や応答文生成に用いる．対話用データとして単語の意味や種類を記述することで，それを実現する．対話用データは逆引き，すなわち対話用データベースから単語を検索することができるようになっている.

また，付属語である格助詞の「が」の例を以下に示す.

```
<node>
  <identifier>が・格</identifier>
  <display>が</display>
  <part>助詞<part>格助詞</part></part>
  <phoneme>ga</phoneme>
  <accent_verb_connection_method>F5</accent_verb_connection_method>
  <accent_verb_connection_value>0</accent_verb_connection_value>
  <accent_adj_connection_method>F1</accent_adj_connection_method>
  <accent_adj_connection_value>0</accent_adj_connection_value>
  <accent_noun_connection_method>F1</accent_noun_connection_method>
  <accent_noun_connection_value>0</accent_noun_connection_value>
</node>
```

付属語には，文献 [39] によって得られたアクセント結合規則を付与した．詳しくは 5.3.3 節で述べる．辞書におけるパラメータの意味は以下に示すとおりである.

accent_verb_connection_method	動詞に接続する場合の付属語アクセント結合様式
accent_verb_connection_value	動詞に接続する場合の付属語結合アクセント価
accent_adj_connection_method	形容詞に接続する場合の付属語アクセント結合様式
accent_adj_connection_value	形容詞に接続する場合の付属語結合アクセント価
accent_noun_connection_method	名詞に接続する場合の付属語アクセント結合様式

accent_noun_connection_value　　名詞に接続する場合の付属語結合アクセント価

4.3　言語情報

4.3.1　言語情報の表現

音声対話システムでは，システムの内部状態を言語に変換してユーザに伝える必要がある．そのためには，内部状態に，言語情報を付加する必要がある．本手法では，言語情報は構文木構造を保持しなくてはならないので，言語情報を LISP 形式で表現する．例えば,「イスを机の前に置いて」という文の構文木構造は図 4.1 に示すとおりである．このとき LISP 形式では,「(て（置く（を（イス)）(に（前（の（机)))))」と表すことができる．対話管理部への入力・ファイルへの入出力は，この LISP 形式を用いて行なう．

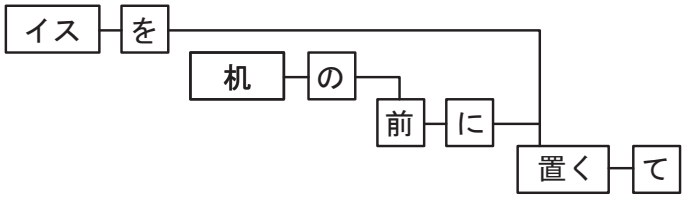


図 4.1: 「イスを机の前に置いて」の構文木構造

4.3.2　タグの付与

単語にタグを与えたり，単語の代わりにタグを用いたりしておくことで，文の単語にアクセスしたり，文を接続したりすることができる．例えば「**アイテム**を**場所**に置く」という文のタグを含む構造は，図 4.2 のとおりである．これを「(置く$PRED（を（$ITEM)）(に（$POS)))」のように$PRED，$ITEM，$POS というタグを埋め込んで表しておく．これにより，$ITEM，$POS タグの部分に単語や句を接続したり，$PRED タグを参照することで述語にアクセスしたりすることができる．また，同じ種類の内部表現には同じタグ名を使用することで，同じ種類の内部表現には個々の内部表現によらず共通の処理を定義することができる．

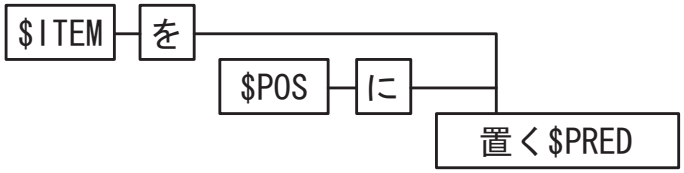


図 4.2: 「**アイテム**を**場所**に置く」のタグと構文木構造

4.4　タグの拡張

前節までが [7] の手法であるが，この手法においては，$ITEM タグや$POS タグを用いてアイテムや場所の情報を表現していたが，この手法のみでは定型文の$ITEM タグや$POS タグに単語を埋め込むといういわゆる「穴埋め問題」になっており，必ずしも柔軟な応答生成とは言えない．例えば 4.2 は，従来構築していたシステムにおける「置く」という動作の定義そのものである．つまり，一つの定型文からは一通りの応答文しか生成することができず，少しでも異なる応答文を生成しようと思えば，異なる定型文を容易する必要があった．この問題を解決し，より汎用的な応答生成を行なうために，従来のタグつき LISP を拡張する，ということを行なう．今回提案する手法は，上記のような単語単位のタグのみではなく，連文節を複数接続して応答文を生成する，というものである．本手法は従来の「穴埋め問題」とは異なり，定型文として与えるのは必要最小限（つまり連文節単位）に止める．この連文節の接続によって，より柔軟な，より汎用的な応答生成を行なう．

4.4.1　連文節

連文節単位で応答文の構成要素を生成する．例としては,「(に（$NOUN))」や「(まで（$NOUN （$ADJ)))」のようなものが挙げられる．この段階で，既に統語構造や単語の重要度等を保持する．

4.4.2　文生成

前節で得られた連文節を，統語構造を考慮しながら接続することで応答文全体を生成する．接続方法によって，単文のみならず重文や複文といった種々多用な応答文が生成できると期待できる．2.4 提案手法における応答文生成の例例として,「右に曲がって駅まで行ってください」という応答文を生成する手順を以下に示す．1. 「(に（$DIRECTION))」というフレームから,「右に」という名詞句を生成する．2. 「(まで（$LANDMARK))」というフレームから,「駅まで」という名詞句を生成する．3. 「(て（$VERB（$NOUN PHRASE)))」というフレームから,「右に曲がって」・「駅まで行って」という動詞句を生成する．4. 2 つの動詞句を連結することで「右に曲がって駅まで行って」という連文節を生成する．5. 「(ください（$VERB PHRASE))」というフレームから,「右に曲がって駅まで行ってください」という応答を生成する．上記の例は重文であるが，同様の手法によって「右に曲がってください．そして駅まで行ってください」という 2 つの単文を生成することもできる．その際には,「そして」等の接続詞を適宜挿入する．

4.5　音声合成

本対話システムで用いる音声合成器は，対話音声の分析に基づいて韻律規則を規定したものである [22]．音声合成器の韻律規則は，基本周波数パターン生成過程モデル [40, 41] に基づいて構築されたものであり，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

規則からなる．これらの規則は，朗読調の音声に対して構築された規則 [23] を，対話音声の分析結果に基づいて対話調音声向けに変換したものである．詳しくは第 5 章で述べる．

4.6　まとめ

本章では，音声対話システムの対話処理における言語情報の取り扱い手法について述べた．言語情報を一貫して構文木構造のまま扱うことで，概念音声合成と親和性の高い応答文生成手法を示した．従来の概念音声合成を扱う音声対話システムは，穴の空いた文に単語を入れることで文を生成しており，文全体を直接生成することは行なっていなかった．本手法では，文全体を直接生成するため，正しい統語情報を音声の韻律に反映することができる．

また，言語情報にタグを付与することで，統一的な処理を定義することができ，言語情報を容易に扱うことのできる方法であることを示した．節や単語を接続し，タグを参照して重要度を設定するだけで，高次の言語情報を韻律に反映した概念音声合成を実現することが可能となった．

次章では，本章で詳しく取り上げなかった韻律制御手法について，より詳しく述べる．

第5章

道案内音声対話システムにおける韻律制御

5.1　はじめに

現在，研究・構築が行なわれている音声対話システムの多くは，音声出力にテキスト音声合成（TTS:Text-To-Speech）ソフトウェアを用いている．しかしながら，TTS システムは一般のテキストから音声合成を行なうことを目的としたものであるため，より自然な対話音声の合成に必要とされる，統語構造や談話情報などの高次の言語情報を応答音声に反映することが困難である．

本稿では，概念音声合成（CTS:Concept-To-Speech）を用いた対話システムにおける音声合成について，韻律制御の観点から述べる．さらに，本韻律制御手法によって得られた応答音声の聴取実験により，従来システムに比べ，応答音声が改善されていることを示す．

5.2　テキスト音声合成（Text-To-Speech）

5.2.1　概要

任意のテキストを音声化するテキスト音声合成の試みは 1960 年代後半に始まり，1970 年代の後半には実用的な英語音声合成システムが開発された．その後，日本語や各言語についてテキスト音声合成システムが開発され，現在では PC のソフトウェアとして一般的になっている．

日本語におけるテキスト音声合成について考えると，漢字かな混じり文章で書かれたテキストが入力となる．この場合，音声波形を生成する処理の前に，テキストがどのような単語から構成されているか，主部と述部の境界はどこにあるか，といった情報を抽出する言語レベルでの処理，さらにこれらの言語情報と音声の音響的特徴とを関連づける操作が不可欠である．

一般に，テキストから音声を合成するためには，言語処理，音韻処理，音響処理の各段階の処理が必要となる．具体的な手順を以下に述べる．

1. 書かれたテキストの構文及び意味解析
2. 各単語の読み仮名，構文・意味情報，及び音韻規則による音韻記号への変換
3. 各単語のアクセント位置，構文情報，韻律情報及び韻律規則による継続時間長，アクセント，イントネーション，ポーズなどの決定
4. 音声合成器への制御パラメータへの変換
5. 音声合成器による音声の生成

5.2.2　対話システムにおける音声合成

テキスト音声合成では，高次の言語情報を利用した音声出力を行なうことができない．なぜなら，テキスト音声合成とは，テキスト形式の表層文から音声を合成する，いわば朗読調の音声合成を目的としたものであり，多様な応答文に対して適切な韻律情報を付与した音声応答を出力することが困難だからである．簡単な対話システムでは，定型文に応答内容の

語句を挿入する録音編集によって応答音声を生成するが，対話システムが高度になり，多様な内容の文で応答するためには，まずユーザに伝えたい内容の意味表現を生成し，それを音声化して出力する必要がある．そのため，統語構造や談話情報などの高次の言語情報を応答音声に反映できる音声合成の枠組み，すなわち概念音声合成（CTS : Concept-To-Speech）[1] の実現が求められている．

概念からの音声合成では，前節のテキスト音声合成の処理において文解析の代わりに文生成のプロセスが必要となるが，その過程で高次の言語情報が正確に得られるため，統語構造を韻律に反映させたり，談話情報で韻律の制御をしたりすることが容易に行なえる [2]．このような観点から，対話システムにおける応答生成には概念音声合成を用いるのが適切であると考えられる．概念音声合成のための言語情報の取り扱い手法については第 4 章で述べた．本章では，その言語情報から実際に音声として出力する手順を述べる．

5.3　音声合成の韻律規則

5.3.1　基本周波数パターン生成過程モデル

ピッチアクセント型言語である日本語においては，韻律的特徴量のうちでも，基本周波数が特に重要となる．基本周波数は一般的に対数軸で取り扱われるが，それは声帯の特性として，ある線形の微少伸縮に対し，周波数が対数的に変化することが知られているからである．基本周波数は，F_0 とも表現され，声の高さに当たる情報である．

図 5.1 に示すように，このモデルでは，2 つの成分に分けて考えている．

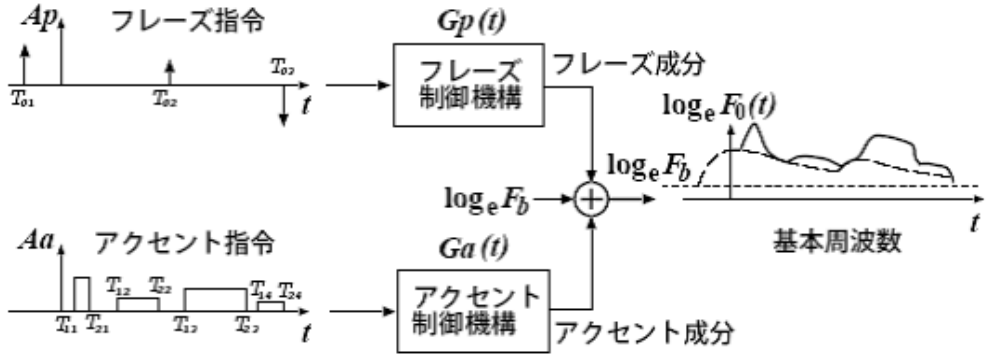


図 5.1: 基本周波数パターン生成過程モデル

その 1 つは，句頭から句末に向かう比較的緩やかな下降に対応する土台部分で，これをフレーズ成分と呼ぶ．もう 1 つは，個々の単語又は単語の連鎖に付属する局所的な起伏に対応するもので，これをアクセント成分と呼ぶ，実測される F_0 パターンを話者毎のほぼ一定な基準値にこれらの 2 種類の成分が加えられたものと考えれば，単語および文音声の F_0 パターンを統一的に把握することが出来る．

こうしたフレーズ成分およびアクセント成分はいくつかのパラメータによって表すこと

ができることが知られている．このうち，フレーズ成分の制御をするパラメータをフレーズ指令，アクセント成分の制御をするパラメータをアクセント指令と呼ぶ．

5.3.2　フレーズ指令

フレーズ指令には，文頭・文中・文末の 3 種類の指令があり，文頭フレーズ指令（Ph）は PI1I2I3I4I5I6，文中フレーズ指令（Pm）は PI2I5，文末フレーズ指令は P0 と表される．フレーズ指令におけるパラメータの意味とその値を表 5.1 に示す．ただし，FRD とは Fundamental Routine of Dialogue のことで，質問や要求の発話とその応答の組からなる対話の基本単位のことを指す

パラメータ	値	意味
1st digit	1	FRD を開く
	2	FRD を閉じる
2nd digit	1	フレーズ中に重要な情報を持つ単語を含む
	2	フレーズ中に重要な情報を持つ単語を含まない
3rd digit	1	話題を変更している
	2	話題を変更していない
4th digit	1	接続詞に従属するフレーズである
	2	接続詞に従属するフレーズでない
5th digit	1	対応するフレーズ成分のモーラ数が 7 以下
	2	対応するフレーズ成分のモーラ数が 8 以上
6th digit	1	疑問の終助詞「か」でわるフレーズである
	2	疑問の終助詞「か」で終わるフレーズでない

表 5.1: フレーズ指令

本節では，生成された応答文構文木から韻律制御を行なう上で必要となる，フレーズ指令およびアクセント指令の挿入位置，およびそれらのパラメータの決定方法について述べる．

i)　フレーズ指令挿入位置

本対話システムにおける，フレーズ指令挿入位置決定規則を以下に示す．

1. 文の先頭には Ph を挿入し，文末には $P0$ を挿入する
2. 最も深い ICRLB 境界に Pm を挿入する
3. 区切られた区間のモーラ数 $> L$ ならば，同様の操作を繰り返し Pm を挿入する
4. ICRLB が存在せず，かつプレーズに含まれるモーラ数が L より大きい場合，単語間結合強度に従って Pm を挿入する
5. 区切られた区間のモーラ数 $> L$ ならば，同様の操作を繰り返し Pm を挿入する

ここで，Ph は文頭フレーズ指令，Pm は文中フレーズ指令を表している．文末フレーズ指令はパラメータを持たずに $P0$ という記号で表される．
本システムでは変数 $L = 12$ を用いた．また，ICRLB 境界とは Immediate Constituent with Recursively Left-Branching structure の略であり，文の構文木において，右枝分かれ境界で前後を区切られ，かつ左枝分かれ境界のみを含む単語連鎖のことである．ICRLB 境界の例を図 5.2 に示す．

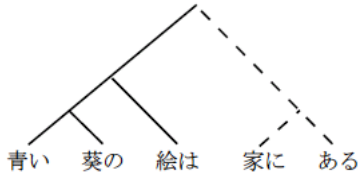


図 5.2: An example of ICRLB boundary

ii)　単語間の結合強度

前節の構文木によるフレーズ分割手法では，必ずしも全てのフレーズが適切に分割できるわけではない．不自然に長いフレーズが残る可能性がある．それら長いフレーズに対し，フレーズ内の単語間の言語的なつながり易さ（単語間結合強度）によって，さらに適当な数のフレーズに区切る．そこでまず，毎日新聞 97 年度版 1 年分から図 5.3 のようなデータベースを PostgreSQL を用いて構築し，bigram を作成した．

単語 ID	単語

頻度	単語 ID（文節前）	単語 ID（文節後）

図 5.3: The design of our database

作成した bigram から，文節境界前後の単語についての結合強度 P を以下の式から求める．ここで，$f(w)$ は，単語 w の出現頻度とする．

$$P = \frac{f(w_2, w_1)}{f(w_1)}$$

iii)　フレーズ指令挿入位置決定の例

フレーズ指令挿入位置の決定方法について,「委員会の方から何度もお誘いを受けていました」という文を例に解説する．
まず,「委員会の方から」と「何度も」の間に ICRLB 境界が存在するので，ここにフレーズ指令を挿入する．さらに，ICRLB 境界の後半部分「何度もお誘いを受けていました」は 16 モーラのため，どこかにフレーズ指令を挿入する必要が生じる．しかし，この部分には ICRLB 境界が存在しない．そこで，単語間結合強度に基づきフレーズ指令の挿入を行なう．

「何度もお誘いを受けていました」の文節前後の単語結合強度を求めると以下のようになる.（括弧内が単語間結合強度）

何度も (0.001) お誘いを (0.017) 受けて (0.474) いました.

文節境界前後の単語間の結合確率の最も小さい箇所を，言語的な結合が弱いとみなし，フレーズ指令を挿入する．よって,「委員会の方から」,「何度も」,「お誘いを受けていました」の 3 つのフレーズに区切られる.

5.3.3　アクセント指令

アクセント指令におけるパラメータの意味を表 5.2 に示す.

パラメータ	値	意味
1st digit	1	主要で新しい
	2	主要でないが新しい
	3	すでに現れているが主要
	4	すでに現れていて主要でない
2nd digit	1	フレーズの先頭
	2	フレーズの途中
3rd digit	1	名詞
	2	動詞
	3	形容詞・副詞
	4	指示語・疑問詞
	5	接続詞

表 5.2: アクセント指令

i)　アクセント規則

対的な音の高低 (High/Low) のパターンでアクセントが決定する．例えば,「桜」は LHH というアクセントを持ち,「野原」は HLL というアクセントを持っている．このように，日本語のアクセントは 2 値の時系列パターンであるため，N モーラの単語を考えた場合，理論上は 2^N 個のアクセントのパターンが存在するはずであるが，本研究が扱う東京方言の場合は，実際に存在するアクセント型は $N+1$ 個に限定されている．これは次に示すような規則が働いているためである.

第一規則

第 1 モーラから第 2 モーラにかけて必ず明確なピッチの上昇あるいは，下降がある.

第二規則
1 単語中のピッチの下降は高々1 箇所である．

この結果，単語のアクセント型は単語中のモーラの音が High から Low に移る位置 (以下，アクセント核) によって一意に定まることになる．このアクセント核が M モーラ目にあるとき，M 型のアクセント型を持つという（図 5.4）．

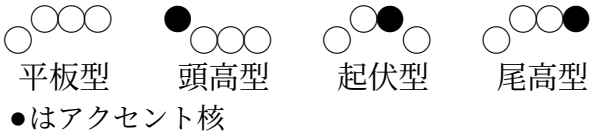


図 5.4: 日本語アクセント型（4 モーラの場合）

N モーラの単語が，0 型，1 型，$\cdots$，N 型の $N+1$ 個のアクセント型を持つ様子を，1 モーラから 4 モーラの単語を例にとって表 5.3 に示した．後ろに助詞「が」を接続した時のアクセントパターンを示している．0 型 (平板型) 語と N モーラ語における N 型 (尾高型) 語について，その単語を単独で発声した時には違いはないが，後ろに助詞などが接続した時に，アクセントが H のままか，L になるかという違いが現れる．その他，1 型に対応したものを頭高型，上述のもの以外を中高型という．

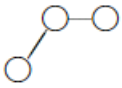
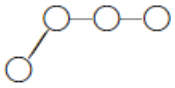
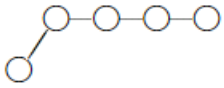
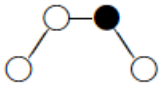
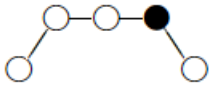
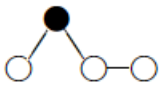
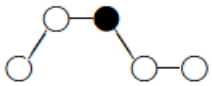
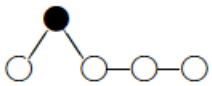
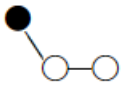
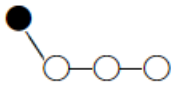
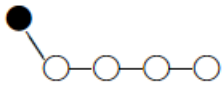
日本語において，単語と単語が統合して文節や複合単語ができるとき，そのアクセントは構成要素それぞれを単独に発声したときのものとは異なるものになり，アクセント核の移動 · 生起 · 消失が起こる．この現象をアクセント結合と呼ぶ．音声対話システムにおいては，より自然な応答音声の生成が求められるため，個々の単語のアクセント型に加えて，このアクセント結合についても考慮する必要がある．[47] は，匂坂らによる一連の結合規則 [45] を聴取実験により見直したものである．本対話システムでは，より自然な応答音声生成のため，上述の，付属語アクセント規則 [47] によるアクセント結合に加え，文節間結合規則の一部 [46] を適用している．表 i) に付属語アクセント結合様式を，図 i) に文節間結合規則の一部を示す．ここで，D，F はそれぞれ起伏型，平板型アクセントとし，添え字は文節の出現位置を表すものとする．

ii)　アクセント指令挿入位置

本対話システムにおける，アクセント指令の挿入規則を以下に示す．

1. 付属語アクセント規則に従い，各文節内の仮アクセント核を決定する．
2. 文節間結合規則に従い，文節間のアクセント結合を文頭から巡回評価する．ただし，フレーズ指令を挟んだ結合は行わない．また，焦点の当てられている文節につても結合は行わない．

アクセント指令には，$DI_1I_2I_3$，$FI_1I_2I_3$，$A0$ の 3 種類があり，$DI_1I_2I_3$ は起伏型または頭高型のアクセントの立ち上げ，$FI_1I_2I_3$ は平板型のアクセントの立ち上げ，$A0$ は両者の立ち下げを示す指令である．パラメータ $I_1I_2I_3$ は表 5.2 による．

型の種類	モーラ数			
	1	2	3	4
平板型	 ハ (ガ) (葉)	 ミ　ズ (ガ) (水)	 サ　ク　ラ (ガ) (桜)	 オ　ハ　ナ　ミ (ガ) (お花見)
尾高型		 ヤ　マ (ガ) (山)	 ヤ　ス　ミ (ガ) (休み)	 イ　モ　ー　ト (ガ) (妹)
中高型			 オ　カ　シ (ガ) (お菓子)	 ナ　イ　ヤ シュ (ガ) (内野手)  ヒ　コ　ー　キ (ガ) (飛行機)
頭高型	 キ (ガ) (木)	 ハ　ル (ガ) (春)	 ミ　ド　リ (ガ) (緑)	 サ　ン　ガ　ツ (ガ) (三月)

(● アクセント核)

表 5.3: N モーラ単語におけるアクセント型

5.4　韻律制御記号を含む音素記号列の生成

構文木構造などから，韻律制御記号を含む音素記号列を生成する手法は，以下に述べるような流れで行なう．

(N_1 モーラ M_1 型+N_2 モーラ M_2 価 →N_c モーラ M_c 型)

アクセント結合様式	文節のアクセント型 M_c	
	$M_1 = 0$	$M_1 \neq 0$
(F1)	M_1	
(F2)	$N_1 + \widetilde{M_2}$	M_1
(F3)	M_1	$N_1 + \widetilde{M_2}$
(F4)	$N_1 + \widetilde{M_2}$	
(F5)	0	
(F6)	$N_1 + \widetilde{M_2}$	$N_1 + \widetilde{M_2}'$
(F7)	0	M_1 and $N_1 + \widetilde{M_2}$
(F8)	$N_1 + \widetilde{M_2}$	M_1 and $N_1 + \widetilde{M_2}$
(F9)	$N_1 + \widetilde{M_2}$	$N_1 + \widetilde{M_2}$

表 5.4: 日本語付属語アクセント結合様式

$$F_1 \quad D_2 \quad \longrightarrow \quad D_{1\ 2}$$
$$F_1 \quad F_2 \quad \longrightarrow \quad F_{1\ 2}$$

図 5.5: A concatenation rule of two *bunsetsu*'s

1. 接続に従って活用形を決定
2. 構文木構造，フレーズ内モーラ数，単語間結合強度に従って，フレーズ指令挿入位置を決定
3. フレーズ内モーラ数，単語間結合強度に従って，フレーズ指令第 5 パラメータを決定
4. 単語の重要度，新規性に従って，フレーズ指令第 2 パラメータを決定
5. 品詞に従って，アクセント指令第 3 パラメータを決定
6. 単語の重要度，新規性に従って，アクセント指令第 1 パラメータを決定
7. フレーズ指令挿入位置に従って，アクセント指令第 2 パラメータを決定
8. アクセント結合規則に従って，アクセント指令挿入位置を決定

5.4.1　音声合成器への入力

音声合成器には，前節で得られた音素記号列と韻律制御記号列の両方を入力する．例えば「左へ曲がって神社まで行ってください」という音声の音声合成器への入力は,「P111212 hi F311 da ri e ma ga SX te A0 P11 D311 zi A0 n zja ma de P21 i F412 SX te ku da sa A0 i P0 S1」というような音素記号列が生成され，音声合成器へ送られる．

5.4.2　音声合成器

音声合成の方式には，大きく分けて以下の 4 つの方式がある．

波形編集方式:　自然音声波形から，合成単位の音声波形を前後の音素環境・韻律的情報と共に切り出し，波形辞書として蓄積しておく．合成時には，音韻環境がテキストの音韻処理結果と最も合致する波形を選択して接続する．波形そのものを用いるために，個々の合成音声の品質は高い．

分析合成方式:　線形予測法やケプストラム法などによって音声を分析し，スペクトル包絡特性と音源特性に分離する．これを音節程度の単位で蓄積し，必要に応じて取り出して接続することにより，連続音声の制御パラメータ時系列を得る．そして，得られたパラメータ時系列をパルス列/白色雑音の音源で駆動する．

ターミナルアナログ方式:　声道の伝達特性を，極に対応する共振回路・零点に対応する反共振回路を組み合わせることで模擬し，音源波で励振する方式である．母音の場合の共振周波数はフォルマント周波数として与えられるなど，音声の物理的特徴との対応が直接的であり，規則による合成に適している．蓄積パターンの接続による場合でも，百数十個の音節パターンを用意することで，比較的品質の高い音声合成が可能である．

声道アナログ方式:　声道内の音波の伝達特性まで遡って模擬する方式である．調音との対応を，ターミナルアナログ方式よりも直接的にとることが可能であり，言語情報との対応もつけやすいと考えられるが，その反面，規則の導出に必要な，正確な生理的データが得にくいという欠点がある．

これらの合成方式は，上ほど蓄積量が多くなり，柔軟性が低くなるが，音質は良くなる．本研究は，対話音声合成の品質に関しての研究であるため，波形接続方式の音声合成器を用いることとした．

本対話システムで用いる音声合成器には，単音・韻律記号列を入力とする波形接続方式の音声合成器 [42] を元にしたものを用いる．[42] は様々な言語に対応したテキスト音声合成器であり，本研究ではこの日本語版を，前述した韻律規則（つまり韻律制御記号列を含む音素記号列）を適用できるよう改良したものを用いている．

5.5　聴取実験

今回，韻律制御手法として新たに制御手法（単語間結合強度，文節間結合規則）を取り入れた有効性を調べるために，聴取実験を行なった．対話例の中から 8 文を用意し，従来手法と提案手法の両方で合成音声を生成し，被験者 19 人に対してどちらの音声が良いかを評価してもらった．従来手法と提案手法をランダムに提示することで，被験者にはどちらの手法で合成した音声かはわからないようにした．評価としては,「提案手法が良い」を 1 点,「従来手法が良い」を-1 点,「どちらとも言えない」を 0 点とし，全被験者に対する平均点を取った．結果を表 5.5 に示す．

表 5.5: 実験結果

文 No.	1	2	3	4
平均点	1.00	1.00	0.89	0.53
文 No.	5	6	7	8
平均点	0.58	0.21	0.79	0.79

5.5.1　考察

「改良手法が良い」の有意性を調べるために,「すべてランダムに評価した」という帰無仮説を用意する．この時，平均点が 0.53 以上となる確率は 0.31%となるため，文 No.6 を除いては有意水準 1%で帰無仮説を棄却でき,「改良手法が良い」ということができる．唯一棄却できなかった文 No.6 に関しては，文の構文木構造の関係で改良手法の効果が表れにくかったため，従来手法とほとんど応答音声の差異がなく，平均点が小さくなったと考えられる．

5.6　まとめ

本章では，対話システムにおける音声合成について，テキスト音声合成との比較を行ないながら本システムにおける応答音声合成手法について述べた．

既存のテキスト音声合成ソフトウェアは，一般のテキストから朗読調音声を合成することを目的としたものであり，対話システムにおいては，より自然な応答音声の生成のために，高次の言語情報を応答音声に反映させる必要がある．そのため，対話システムにおける音声合成には，概念音声合成を用いるのが望ましいと考えられる．

従来のシステムでは，構文木構造を用いることで，高次の言語情報の取り扱いを行なっていた．しかし，その高次の言語情報が応答音声に必ずしも適切に反映されていなかった,という課題があった．そこで，音声合成における韻律規則やアクセント結合規則を適用する本提案手法により，従来のテキスト音声合成による合成音声に比べてより自然な対話音声の合成が可能となった．

第6章

結論

6.1　まとめ

本論文では，音声対話システムにおける応答文の生成手法と応答音声の韻律制御手法について提案を行ない，その提案手法を道案内対話システムに実装した．

第 2 章では，これまでに研究されてきたさまざまな対話システムについて説明し，また対話システムの要素技術について述べた．

第 3 章では，道案内対話システムの概要，タスク，ユーザインタフェース，構成について述べた．また，本システムにおける誤解の解決のための手法を示した．

第 4 章では，今回実装した道案内対話システムで用いられている，音声対話システムの対話管理部における言語情報の取り扱いに関する手法について述べた．文などの言語情報は，常に構文木構造で扱うことで，音声合成との親和性の高い応答生成が実現できることが示されている．内部状態（概念）に自身を表す言語情報を与えておき，応答文生成の際にそれらを接続することで，柔軟な応答文生成を実現している．また，タグを用いることで異なる言語情報に関して統一的な処理を加えることができ，言語処理の記述を容易にした．さらに，タグを拡張して単語単位から連文節単位での挿入を可能とし，より柔軟な応答文生成を実現した．

第 5 章では，本体対話システムの韻律制御手法の改良について述べた．フレーズ指令について，第 4 章で求められた構文木によって得られる ICRLB による挿入位置の決定手法に加え，bigram といった統計的手法を併用することでより自然な韻律制御が行えるようになった．アクセント指令については，従来のアクセント結合規則に加え，文節間結合規則を導入することで，より自然な応答音声の生成が可能となった．これらにより得られた応答音声についての聴取実験を行ない，本提案手法により従来システムより自然な応答音声が実現できたことを示した．

6.2　課題と今後の展望

応答音声の韻律制御に利用する語の重要度・新規性をどのように抽出するかについて，現在までは対話の履歴を利用することを行なってきた．しかしながら，実際に人間がどの単語を強調して発声しているかはより複雑であり，語が一般的であるか否か，相手の対話に対する理解をどのように想定しているかなどにも関係する．すなわち，ユーザの発話内容に関する知識をシステムは常に観測を行ない，それに従って韻律の制御を行なう必要がある．また，強調の方法については，韻律を制御することはもちろんであるが，長い語句などで強調すると不自然になる場合，あるいは統語などの情報との兼ね合いで強調しにくい場合などがある．このような場合，語順の制御も重要になる．具体的には，焦点の置かれた単語は優先的に前に持ってくるなどの処理を行なうことで，より自然な応答音声を生成することができる，と考えられる．

今回提案した応答音声生成手法を用いることで，聴取実験を行なうことにより従来と比較してより自然な応答音声を生成することができた．しかし，重要度・新規性の扱い方について，まだ不完全な部分が残されている．特に，語の新規性については，まだ正確に反

映されているとは言えない状況である．そのため，言語情報をより詳細に扱うことで，韻律のさらなる細かな制御を行なう必要がある．特に新規性については，対話履歴を対話システムに組み込むのではなく，言語処理手法に組み込むことで単語の新規性を扱い，それを応答音声に反映する必要がある．また，より細かな韻律の制御には辞書の充実も重要な課題となる．

本対話システムのタスクについて，従来研究よりも複雑な応答音声を生成する枠組みを作成することができた．しかし，道案内タスクそのものがそれほど複雑なタスクではなく，システムの発話文のバリエーションが少ないため，システム・ユーザ間で行なわれる対話も簡単なものに限られている．より複雑な状況設定をするために，現在の単純な道案内ではなく，よリ多様な周辺タスクを組み入れていく必要がある．今後の具体的な方針としては，施設についての属性情報を持たせることを考えている．例えば，タバコを購入するためにコンビニを探しているユーザに対して，単純にコンビニ間での道案内をするのではなく，他のユーザの希望が満足できる施設（タバコの売っている駅などへの案内）のような提案もできるようなシステムにすることで，人間との「相談」という高度，複雑かつ様々なレパートリーの対話が考えられる状況を設定したい．

音声合成について，現在の音声合成器では，音素単位の継続長の変化や，ピッチの微妙な制御については行なうことができない．そこで今後の構想として，1) コーパスを利用した立ち上がり幅，継続長の制御，2) 使用する音声合成器の見直しを考えている．より自然な音声を出力のため，コーパスを利用した音声合成を考えている．コーパスによる韻律制御は音声合成器においてではなく，より応答文生成部に近い部分で実現することが望ましい．これによって，話者つまり対話システムの意図をより音声に反映することが可能になると期待できる．現在，コーパス作成に必要となる音声（道案内タスク 500 文程度）を修習し，近々収録を開始する予定である．また，収録した音声から HTK により HMM を作成し，HMM 合成を行なう予定である．これに伴い，既存の音声合成器の代替が必要となる．現在，汎用的であり，オープンソースの音声合成器である Gtalk[24] が候補の一つと考えている．Gtalk は TTS システムであり，そのままでは今回の対話システムの音声合成器としての利用はできない．そこで，ソースを変更して本対話システムのインターフェースに整合させ，コーパスベースの合成という目的が達せられるようなシステムに変更する．

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助教授には，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．

また，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，秘書の武田祥子さん，笠島恵美さんに，深く感謝いたします．

本研究を共同研究者である八木裕司氏には，研究を進めていく上で様々な助言を頂きました．また，日頃研究室生活を行なう上で，様々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みなさまに深く感謝いたします．

また，本研究を行なうにあたり，聴取実験に協力して下さった被験者の方にも，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．

2006 年 1 月 30 日
高田 靖也

参考文献

[1] 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.

[2] 広瀬啓吉: “音声合成技術,” 情報処理学会誌, vol.38, no.11, pp.984-991, 1997.

[3] K. Hirose: “Speech reply generation for a spoken dialogue system on academic document retrieval,” Proc. International Symposium on Spoken Dialogue, Japan Society for the Promotion of Science “Research for the Future” Program, Beijing, pp.8.1-5, 2000.

[4] 中川聖一, 傳田明弘, 伊藤敏彦: “マルチモーダル観光案内対話システム,” 人工知能学会誌, vol.13, no.2, pp.241-251, 1998.

[5] 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.

[6] Y. Shinyama, T. Tokunaga and H. Tanaka: “Kairai - Software Robots Understanding Natural Language,” Third International Workshop on Human-Computer Conversation, 2000.

[7] K. Hirose, J. Tago and N. Minematsu: “Speech Generation from Concept for Realizing Conversation with an Agent in a Virtual Room,” Proc. EUROSPEECH 2003, vol.3, pp.1693-1696, 2003.

[8] S. Bennacef, L. Devillers, S. Rosset and L.Lamel: “Dialog in the RAILTEL Telephone-Based System,” Fourth International Conference of Spoken Language Proceedings, vol.1, pp.550-553, 1996.

[9] C. Popovici and P. Baggia: “Language Modelling For Task-Oriented Domains,” Proc. Eurospeech ’97, vol.3, pp.1459-1462, 1997.

[10] 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.

[11] J. Glass, D. Goodine, M. Phillips, S. Sakai, S. Seneff and V. Zue: “A Bilingual VOYAGER System,” Proc. Eurocpeech ’93, pp.2063-2066, 1993.

[12] M. Phillips, J. Glass and V. Zue: “Automatic Learning of Lexical Representations for Sub-Word Unit Based Speech Recongnition System,” Proc. of Eurospeech ’91, pp.577-580, 1991.

[13] S. Seneff: “TINA : A Natural Language System for Spoken Language Applications,” Computational Linguistics, vol.18, no.1, pp.61-86, 1992.

[14] V. Zue, J. Glass, D. GOddeau, D. Goodine, L, Hirschman, M. Phillips, J. Polfroni and S. Seneff: “A Spoken Dialogue Interface for On-Line Air Travel Planning,” Proc. International Symposium of Spoken Dialogue, pp.157-160, 1993.

[15] G. Ringland: “Structured Object Representation - Schemata and Frames,” in Ringland, G. A. and D. A. Duce eds., Approaches to Knowledge Representation, Researach Studies Press Ltd., 1988.

[16] 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, 3B3-03, 2001.

[17] 遠山義洋, 西田豊明: “話題構造の抽出と変形による対話録の自動要約,” 2000 年度人工知能学会全国大会（第 14 回）論文集, 07-06, pp.157-160, 2000.

[18] 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.

[19] 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.

[20] `http://winnie.kuis.kyoto-u.ac.jp/pub/julius/`

[21] 河原達也, 李晃伸, 小林哲則, 武田一哉, 峯松信明, 伊藤克亘, 山本幹男, 山田篤, 宇津呂武仁, 鹿野清宏: “日本語ディクテーション基本ソフトウェア（98 年度版）,” 音響学会誌, vol.56, no.4, pp.255-259, 2000.

[22] K. Hirose, M. Sakata and M. Kawanami: “Synthesizing dialogue speech of Japanese based on the quantitative analysis of prosodic features,” Proc. ICSLP96, vol.1, pp.378-381, 1996.

[23] 河合剛, 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.50, no.6, pp.433-442, 1994.

[24] 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.

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

[26] 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/.

[27] （社）日本電子工業振興協会: “日本語テキスト音声合成用記号の規格,” JEIDA-62-2000, 2000.

[28] S. Young, J. Jansen, J. Ordell, D. Ollason and P. Woodland: “The HTK Book,” 1995.

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

[30] 匂坂芳典, 佐藤大和: “日本語単語連鎖のアクセント規則,” 電子情報通信学会論文誌, vol.J66-D, no.7, pp.849-856, 1983.

[31] 益子貴史, 徳田恵一, 小林隆夫, 今井聖: “動的特徴を用いた HMM に基づく音声合成,” 電子情報通信学会論文誌, vol.J79-D-II, no.12, pp.2184-2190, 1996.

[32] 益子貴史, 徳田恵一, 宮崎昇, 小林隆夫: “多空間確率分布 HMM によるピッチパターン生成,” 電子情報通信学会論文誌, vol.J83-D-II, no.7, pp.1600-1609, 2000.

[33] http://hts.ics.nitech.ac.jp/

[34] 赤羽誠, 蓑輪利光, 板橋秀一: “音声合成用記号の標準化について,” 日本音響学会誌, vol.57, no.12, pp.776-782, 2001.

[35] 山下洋一, 喜多竜二, 峯松信明, 吉村貴克, 徳田恵一, 田村正統, 益子貴史, 小林隆夫, 広瀬啓吉: “マルチモーダルコミュニケーションのための音声合成プラットフォーム,” 情報処理学会研究報告（音声言語情報処理研究会）, 2002-SLP-40-12, pp.67-72, 2002.

[36] 森島繁生, 八木康史, 金子正秀, 原島博, 谷内田正彦, 原文雄, 橋本周司: “顔の認識・合成のための標準ソフトウェアの開発,” 電子情報通信学会技術報告（パターン認識・メディア理解研究会）, PRMU97-282, 1998-3.

[37] Satoru Kogure, Toshihiko Itoh, Akihiro Denda and Seiichi Nakagawa: “A Semantic Interpreter for a Robust Spoken Dialogue System,” The Second International Conference on Multimodal Interface, Hong Kong, Vol.II, pp.61–66, 1999.

[38] http://www.w3c.org/XML/

[39] N. Minematsu, R. Kita and K. Hirose: “Automatic estimation of accentual attribute values of words to realize accent sandhi in Japanese text-to-speech conversion,” Proc. IEEE 2002 Workshop on Speech Synthesis, Santa Monica, 2002.

[40] H. Fujisaki and S. Nagashima: “A model for synthesis of pitch contours of connected speech,” Annual Report of Engineering Research Institute, University of Tokyo, vol.28, pp.53-60, 1969.

[41] 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.

[42] http://tcts.fpms.ac.be/synthesis/mbrola.html

[43] http://www.opengl.org/

[44] Lander, Jeff: “Slashing Through Real-Time Character Animation,” Game Developer, vol.5, no.4, pp.13-15, 1998.

[45] 匂坂芳典，佐藤大和,“ 日本語単語連鎖のアクセント規則， ”電子情報通信学会論文誌，J66-D，7，pp.849-856, 1983.

[46] 広瀬啓吉，藤崎博也,“ 音声合成とアクセント・イントネーション， ”電子情報通信学会誌，第 70 巻 4 号，pp.378-385, 1987.

[47] N.Minematsu, R.Kita and K.Hirose: “Automatic estimation of accentual attribute values of words to realize accent sandhi in Japanese text-to-speech conversion,” Proc. IEEE 2002 Workshop on Speech Synthesis, Santa Monica, 2002.

発表文献

[1] Keikichi Hirose, Yuji Yagi, Seiya Takada, Yasufumi Asano and Nobuaki Minematsu: “Speech Synthesis from Concept and its Prosodic Control for Reply Speech Generation in a Dialogue System,” Annual Project Report Grant-in-Aid for Creative Basic Research “Language Understanding and Action COntrol,” 2006-3.(出版予定)

[2] 八木裕司, 高田靖也, 広瀬啓吉, 峯松信明: “道案内音声対話システムにおける応答生成手法の評価” 日本音響学会 2006 年春期研究発表会講演論文集, 2-11-13, 2006-3.（出版予定）

[3] 高田靖也, 八木裕司, 広瀬啓吉, 峯松信明: “道案内音声対話システムにおける韻律制御手法の改良” 日本音響学会 2006 年春期研究発表会講演論文集, 2-11-14, 2006-3.（出版予定）

[4] Y.Yagi, S.Takada, K.Hirose and N.Minematsu: “Improved concept-to-speech generation in a dialogue system on road guidance,” Proc. LUAR2005 (2nd International Workshop on Language Understanding and Agents for Real World Interaction), pp.429-436 (2005-11).

[5] Y.Yagi, S.Takada, K.Hirose and N.Minematsu: “An improved method of generating speech from concept and its application to a dialogue system of road guidance,” Proc. SPECOM2005 (10th International Conference on Speech and Computer), pp.703-706 (2005-10).

[6] K.Hirose, N.Minematsu, Y.Yagi and S.Takada: “Generating Reply Speech of a Dialogue System,” Proc. International Symposium on Advanced Electronics for Future Generations - ”Secure-Life Electronics” for Quality Life and Society -, pp.229-234. 2005-10.

[7] 八木裕司, 高田靖也, 広瀬啓吉, 峯松信明: “道案内音声対話システムにおける応答生成の改良” 日本音響学会 2005 年秋期研究発表会講演論文集, 1-7-6, pp.3-4, 2005-9.(CD-ROM)

[8] 八木裕司, 高田靖也, 広瀬啓吉, 峯松信明:“ 対話システムにおける応答生成手法の改良とその実装 ”情報処理学会研究会報告 2005-SLP-57, pp.93-98, 2005.

[9] K.Hirose, Y.Yagi, S.Takada, Y.Asano and N.Minematsu: “Generation of Speech Reply in a Dialogue System,” Annaul Project Report Grant-in-Aid for Creative Basic Research ”Language Understanding and Action Control,” pp.209-218, 2005-3.

[10] 八木裕司, 高田靖也, 広瀬啓吉, 峯松信明: “音声対話システムにおける応答生成手法の検討” 日本音響学会 2005 年春期研究発表会講演論文集, vol.1, 3-5-14, pp.653-654, 2005-3.