概念音声合成の枠組を用いた音声対話システムにおける応答生成手法の構築

東京大学大学院 工学系研究科
電子工学専攻

八木 裕司

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ものがある．また，近年における音声認識や音声合成，自然言語処理といった音声言語情報処理技術の顕著発展を背景とし，これらの技術を統合して実現される音声対話システムの研究・構築が盛んに行なわれるようになってきた．

音声対話システムの研究を行なう上で重要となってくる点として,「対話音声を取り扱う」ことが挙げられる．音声言語情報処理の分野において非常に数多くの研究がなされているが，それらの研究は必ずしも対話音声を取り扱ったものではなく，音声対話システムへの応用を考えた場合には，テキストに現れる文字情報のみならず，意図や感情といったパラ言語情報，性別や年齢といった非言語情報にも対応できる必要がある．音声対話システム研究の大半は，このような観点から研究が進められている．

音声対話システムは，一般的に音声認識→言語理解→対話処理→言語生成→音声合成の流れで処理が行なわれる．音声対話システム研究のほとんどは，このうちの音声認識・言語理解（これらを合わせて音声理解と呼ぶ）に焦点が置かれたものとなっている．このような研究は，言い間違いや言い澱みに対する頑健な音声理解やユーザ発話の韻律的特徴からユーザの意図を判断する，といったように非常に多角的な観点から研究が行なわれている．その一方，言語生成・音声合成といった，音声出力に関する研究は国内ではほとんどなされておらず，海外に目を向けても数えるほどしか存在しない．既に研究・構築されている音声対話システムの多く（特に，音声出力を研究対象としていないシステム）では，音声出力にテキスト音声合成（TTS：Text-to-Speech）に基づいた既存のソフトウェアを用いている．しかしながらこのテキスト音声合成手法では，表層テキストのみから音声を合成するため，テキストに現れない情報を合成音声の韻律に反映させることができず,「対話音声の合成」を考えた場合に適したものとは言えない．

音声対話システムでは，応答内容を全てシステム自身で作り出すため，テキスト情報のみならず統語構造や談話情報といった高次の言語情報も容易に得ることができる．そこで，高次の言語情報を応答音声に反映できる概念音声合成（CTS：Concept-to-Speech）の枠組の実現が求められる．概念音声合成という考え方は古くから提唱されてきたものであるが，この概念音声合成の枠組を実際にシステムとして実装し，対話システムに組み込んだり応答音声を出力させたりした研究例はこれまでのところ存在しない．

本論文では，このような観点から，概念音声合成の枠組の実現を目指した応答生成手法の構築を行ない，特に統語構造と談話情報を適切に応答音声の韻律，その中でも主に基本周波数（F_0）に反映させることを目指す．また，実際にシステムに組み込むことによって，ユーザにもわかりやすい音声を合成することを目標とする．

応答文生成手法の構築にあたり，本論文では2つの音声対話システムを構築している．その2つはエージェント音声対話システム（本論文第3章）と道案内音声対話システム（同4章）である．エージェント音声対話システムは，仮想空間内にいるエージェントに対してユーザが指示を行なうことで，仮想空間内の物体を移動させる，というタスクを取り扱うシステムである．道案内音声対話システムは，出発地点から目標まで，システムがユーザに音声を用いて指示を出すことで案内をする，というタスクを扱っている．

統語構造は，最終的な応答音声の韻律においては，主に文のイントネーションに深く関わってくる．音声対話システムでは，自らが1から応答文生成を行なうため，始めから100%正しい構文情報を得ることができる．そのため，この構文情報を常に保持しておくことが重要となる．本論文で提案する手法でも，応答文の言語情報は常に構文木構造を保持したまま扱う方針をとる．

エージェント音声対話システムでは，構文木構造を持たせた定型文をテンプレートとして用意しておき，対話状況に応じて適切な定型文を選択する手法をとる．この定型文にはタグを埋め込んでおき，同じ属性の単語は同様に扱えるようにする等の統一的な処理を可能とする．応答生成の際には，このタグに適切な単語を挿入する．しかしながら，この手法では，システムがより多様な応答生成を求められるような状況では，必要な定型文の数が膨大になる，という問題点があった．この問題を解決する際に，エージェント音声対話システムではタスク拡張によって多様な応答生成を実現することが困難であったことから，新たに道案内音声対話システムを構築し，その中でテンプレートの単位を文ではなく文節（定型フレーズ）とする手法を新たに提案する．これを実現するために，エージェント対話システムでは単語しか挿入できなかったタグに，文節も挿入できるようにした．その結果，ある対話状況を実現するために必要な応答文の種類が多様になればなるほど，定型フレーズを用いる手法の方が定型文を用いる手法に比べてテンプレートの増加数を大幅に減らすことができ，より少ないテンプレート数でより多様な種類の応答生成を実現できることが示された．また同時に，この手法は，タスクに依らず汎用的な手法であることも示唆された．

一方，談話情報は，最終的な応答音声の韻律においては，主に個々の単語のアクセント成分の大きさに深く関わっている．本論文では，単語の「重要度」と「新規性」に着目し，これらの情報を適切に応答音声の韻律に反映させることで焦点制御を行なう，という手法を提案する．これらの情報もまた，音声対話システムが自ら作り出す情報であるため，これらの情報を応答音声に反映させないとこれらの情報が無駄になってしまう.「重要度」と「新規性」の情報は，応答文生成過程においてテンプレート中のタグに単語を挿入する際に同時に付与する．そして，それらの情報が音声合成器（本システムでは波形接続方式の合成器を用いている）に渡され，応答音声の韻律に反映されて合成音声として出力される．エージェント音声対話システムでは，付属語アクセント結合規則を韻律制御規則に採り入れ，道案内音声対話システムでは，さらに文節間結合規則も韻律制御規則に採り入れた．聴取実験の結果から，これらの手法を採り入れることの妥当性が示された．

また，道案内音声対話システム内において，統語構造と談話情報の観点からの聴取実験も行なった．統語構造においては，テキスト音声合成で用いられるような一般的な構文解

析ツールを用いた場合に得られる構文情報を基に合成した音声との比較によって，談話情報においては,「重要度」「新規性」という情報を用いない場合の合成音声との比較によって評価を行なった．その結果，本論文における提案手法の妥当性が示された．さらに，この聴取実験で得られた知見の検証を行なうために，道案内音声対話システムを拡張することで応答生成のバリエーションを増やし，再度聴取実験を行なった．その結果，提案手法の有効性が再確認されるとともに，新たな知見も得ることができた．

本研究で取り上げた統語構造，談話情報以外にも，意図や感情といったパラ言語情報，性別や年齢といった非言語情報等多数挙げられる．そして，より多様な対話音声を合成するためには，これらの観点からの研究も必要になってくると考えられる．これまでの音声対話システム研究では，このような観点からの音声出力に関する研究はほとんど行なわれてきていないが，本研究が今後の展開における 1 つの方向性を示すことができたと考えている．

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への要求は高いものがある．実用上の観点から言えば，キーボードやマウスを用いるシステムに抵抗のある人にとっても，音声を用いることでシステムとの意思伝達を図ることができれば，感じられる抵抗感は軽減されると考えられる．また，公共の場等においてキーボードやマウスといったデバイスを設置するのが様々な面で問題となるような場合にも，音声を用いるのであればマイクとスピーカー（これらを内蔵型にすれば盗難の恐れもない）を用意するだけでユーザが容易に利用することが可能となる．このような観点から，音声対話システムの研究開発が盛んに行なわれるようになり，[1, 2] のように実用化されたシステムも出現している．

音声対話システム研究における技術的側面からの背景として，近年における音声認識や音声合成，自然言語処理といった音声言語情報処理技術の顕著な発展が挙げられる．音声対話システムは，様々な音声言語情報処理技術を統合して実現されるものであるため，これらの技術を統合したシステムを構築することは，各研究の実用化ということを考えた場合にも重要なものと言える．

音声対話システムということを考えた時，考慮すべき重要な点の 1 つとして「対話音声を取り扱う」ことが挙げられる．音声言語情報処理の分野において非常に数多くの研究がなされているが，それらの研究は必ずしも「対話」音声を取り扱う，目標としているものではなく，理想的な音声[1]を対象としているものも多数見受けられる．しかしながら，実際に人が話す音声には様々な「対話」音声の特徴が含まれる．例えば，言い間違いや言い澱みのようなものやフィラーなどテキストにも現れるものから，意図や感情，性別や年齢といったテキストに現れないものまで非常に様々な要因が「対話」音声に含まれている．そのため，音声対話システム研究の多くは,「システム内の某の箇所について，いかに対話音声を取り扱うか」について研究しているものだと言い替えることができる．

音声対話システム研究の多くは，音声認識・理解に焦点を当てたものとなっている．一方，音声出力（音声合成）に関する研究は非常に少ない．特に，国内の研究[2]ではほぼ皆無と言っても過言ではなく，海外の研究例においても文献 [3] が知られている程度である．実際に音声出力に焦点を当てていない研究では，音声出力にテキスト音声合成（TTS : Text-to-Speech）と呼ばれる手法を用いた既存のソフトウェアを用いている．しかしながら，この TTS システムとは，一般のテキストから所謂「朗読」音声を生成することを目的としたものが多く，そのような音声合成システムでは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，その用途にもよるが，朗読調のみならず対話調の応答音声が求められ，またそれに発話の意図や感情を反映させることも求められる [4]．これらの要求を満たすことを考えると，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といった高次の言語情報を容易に得ることができるため，これらを応答音声に反映できる音声合成の枠組，すなわち

[1] 多くの場合，朗読音声．例えば無音室でアナウンサーが原稿を読み上げたもの．
[2] 音声は言語情報とは切り離せないものであり，必然的に言語依存となる部分が生じる．

概念音声合成（CTS : Concept-to-Speech）[5] 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行なうといったことが容易に行なえる [6]．統語構造や談話情報等の高次の言語情報，あるいは意図や感情等のパラ・非言語情報は，音声の韻律と関連する点が多く，この観点からの研究が重要であるが，実際にこのような観点から研究を行ない，音声合成システムとして構築した研究は，少なくとも国内では見受けられない．

このような背景を踏まえて，本論文では，概念音声合成の枠組を実現し，統語構造や談話情報等の高次の言語情報を応答音声の韻律に反映させる手法を構築する．また，その手法を音声対話システムに組み込み，実際のユーザにもわかりやすい音声を合成することを目指す．

1.2　本論文の目的

音声対話システムでは，応答音声がユーザにとって「わかりやすい」ものであることが求められる．この「わかりやすい」には，応答音声自体の明瞭性等の音質に関わるものもあれば，適切な韻律制御による音声の自然性や，適切な焦点制御による意図の伝達といったものも要因として挙げられる．

本論文では，概念音声合成の枠組の実現に向けて，特に統語構造と談話情報を応答音声の韻律に反映させ，適切な韻律制御を行なう手法の構築を目的とする．

統語構造に関しては，生成する応答文の言語情報を常に構文木構造を保持したまま扱う，という手法を提案する．統語構造は，最終的な応答音声の韻律においては，主に文のイントネーションに深く関わってくる．そのため，正確な統語構造を保持することは非常に重要である．音声対話システムでは，自らが 1 から応答文生成を行なうため，一般的な構文解析ツールとは異なり，始めから 100%正しい構文情報を得ることができる．そのため，システム内部情報として始めから構文木構造を保持したまま扱う手法を構築する．また，構文木構造内にタグを用いることにより，同じ属性の単語は同様に扱えるようにする等の統一的な処理を可能とする．これらの内容は，以降「言語情報の取り扱い手法」と述べることがある．

談話情報に関しては，上記構文木構造中のタグに「重要度」と「新規性」という 2 つのパラメータを同時に保持させ，これらを適切に応答音声の韻律に反映させることで焦点制御を行なう，という手法を提案する．談話情報は，最終的な応答音声の韻律においては，主に個々の単語のアクセントに深く関わっている．本論文で構築している道案内音声対話システムのような情報提示型のシステムにおいては，伝えるべき単語が強調されることによって，システムの意図がユーザに伝わりやすくなることが期待できる.「重要度」や「新規性」といった情報もまたシステムが 1 から作り出す情報であるため，これらの情報を応答音声に反映させずに破棄するのは，無駄にしてしまうことになる．そのため，これら談話情報を適切に設定し，また応答音声の韻律に適切に反映させる手法を構築する．これらの内容は，以降「韻律制御手法」と述べることがある．

また，エージェント音声対話システムや道案内音声対話システムといったシステムを構築し，システムの中にこれらの提案手法を組み込む．実際のシステム発話をユーザに評価してもらうことで，これらの提案手法の妥当性を検証する．

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・目的について述べた．第2章では，一般的な音声対話システムについての概観を述べる．また，音声対話システムにとって重要な「対話調の」音声を先行研究がどのように取り扱っているかについて，それぞれの研究がどこに着目しているかによって分類しながら紹介する．第3章では，エージェント音声対話システムについて述べる．これは修士課程において構築していたシステムであるが，本論文の目的としている「概念音声合成の枠組の実現」に必要な「言語情報の取り扱い手法」及び「韻律制御手法」について，現在構築している手法の基となっている手法をこのシステム内で構築しているため，本論文でも第3章という形で触れる．第4章では，道案内音声対話システムについて述べる．このシステムは，第3章で述べたエージェント音声対話システムにおける様々な問題点を解決するために構築したシステムである．ここでは，特に「言語情報の取り扱い手法」及び「韻律制御手法」についての従来手法（第3章の手法）の問題点を明確にするとともに，それらを改良した手法を提案する．また，応答音声の聴取実験により，改良提案手法の妥当性を示す．最後に，第5章で本論文をまとめ，今後の課題と展望について述べる．

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 [7][8].

音声対話ということを考えた際，対話音声を考慮することは重要な課題の 1 つである．音声認識や音声合成といった個々の要素技術研究が必ずしも対話音声を対象とせず，アナウンサーの読み上げ音声等の朗読音声のような対話音声特有の現象を含まない音声を対象としている．しかし，音声対話システムに組み込むに当たっては対話音声ならではの特徴を熟慮する必要がある．そのため，音声対話システムの研究は，2.4 節以下で紹介するように，ほとんどが特定のモジュールに特化した研究となっている．しかしながら，後述する通り音声出力（特に音声合成）に関する研究はほとんどなされておらず，一般的なテキスト音声合成ソフトウェアが用いられているのが現状である．

本章の構成は以下のようになっている．まず，一般的な音声対話システムの構成について，例を交えながら説明する．続いて，音声対話システム研究において重要な要因である「対話音声」と「朗読音声」の違いについて触れ，実際の音声対話システム研究がどのように行なわれているか，どのような箇所に研究の焦点を置いているかについて，研究例を紹介しながら述べる．

2.2　一般的な対話システム

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（Speech input）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（Database）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）で音声として出力される（Speech output）．このような処理を繰り返すことで，人と計算機との対話を処理し，タスクを達成する．

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であると考えられている．

以下，各モジュールについての説明を行なう．

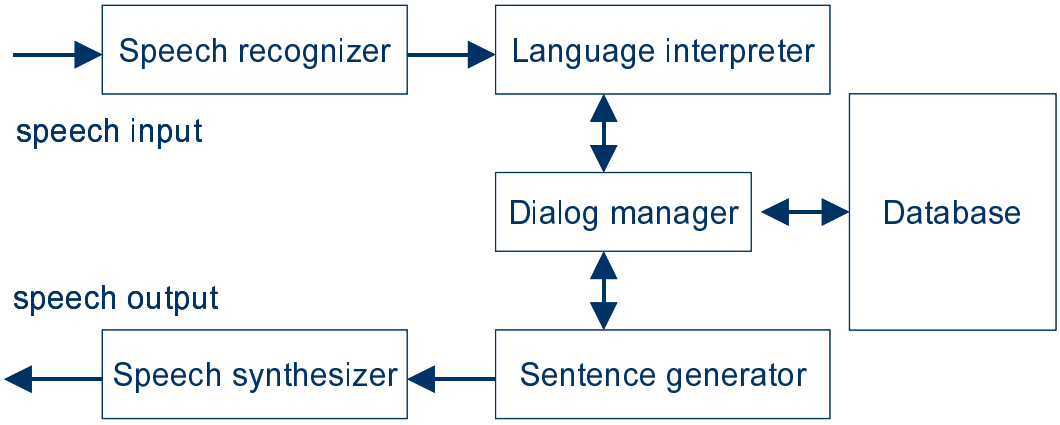


図 2.1: 一般的な音声対話システム構成図

2.2.2　音声認識部

音声認識部は，ユーザの発話音声を文に変換する．音声認識の際に，音響モデル，言語モデルを用いる．最近の音声認識では，音響モデルとしては統計的なモデルである HMM（Hidden Markov Model）[9] が，言語モデルとしては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）がよく用いられる．しかし，言語モデルとして N-gram を用いるのは大語彙連続音声認識を行なうような場合であり，対話システムにおいては，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のネットワークあるいは CFG（文脈自由文法）で言語モデルを記述する場合が多い．

2.2.3　言語理解部

言語理解部は，音声認識部において得られた単語列に対して，形態素解析・構文解析・意味理解・文脈理解等が行なわれる．これらは，音声認識においても認識率の向上に寄与する部分も多く，音声認識部において既に処理されている場合も多い．これらの情報を，意味ネットワークやスロットフィリング，ワードスポッティング等の方法により，対話制御部において処理が可能な意味表現（内部表現や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

対話制御部は，言語理解部から渡された意味表現からユーザの意図を理解する．また，スムーズな対話を実現するための対話の全体的な制御を行なう．

対話制御のために，タスクに応じて適切な対話モデルを構築する必要がある．対話モデルは，対話の履歴を参照し，適切な応答内容の生成やユーザの次発話の予測，対話の主導権の管理や誤認の確認・回復といった役割を果たす．

2.2.5　言語生成部

言語生成部は，対話制御部から受け取った意味表現を文の形に変換し，それを音声合成部に送る．文生成には，スロット法や CFG によるランダム生成を用いる方法がある．前者は，文のフレームをあらかじめ複数個用意しておき，適切な単語で埋めていく方法で，後者は，初期記号 S から始めて，書き換え規則をランダムに適用しながら生成を行なうものである．

2.2.6　音声合成部

音声合成部は，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．音声合成の手法には，波形編集方式や分析合成方式などがある．文字列を音声に変換するテキスト音声合成（TTS：Text-To-Speech）がよく用いられるが，言語生成部との一貫した処理によって，意味表現から直接音声を生成しようとする概念音声合成（CTS：Concept-To-Speech）[5] 方式が今後多く利用されることが予測される．

2.2.7　音声対話システムの例：GALAXY[10, 7]

2.2.7.1　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．これは，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ものである．そのシステム構成を図 2.2 に示す．

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．GALAXY アーキテクチャを用いたシステムには，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[12] などがある．

2.2.7.2　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[13] 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 は次のような音声認識結果の候補を出力する．

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

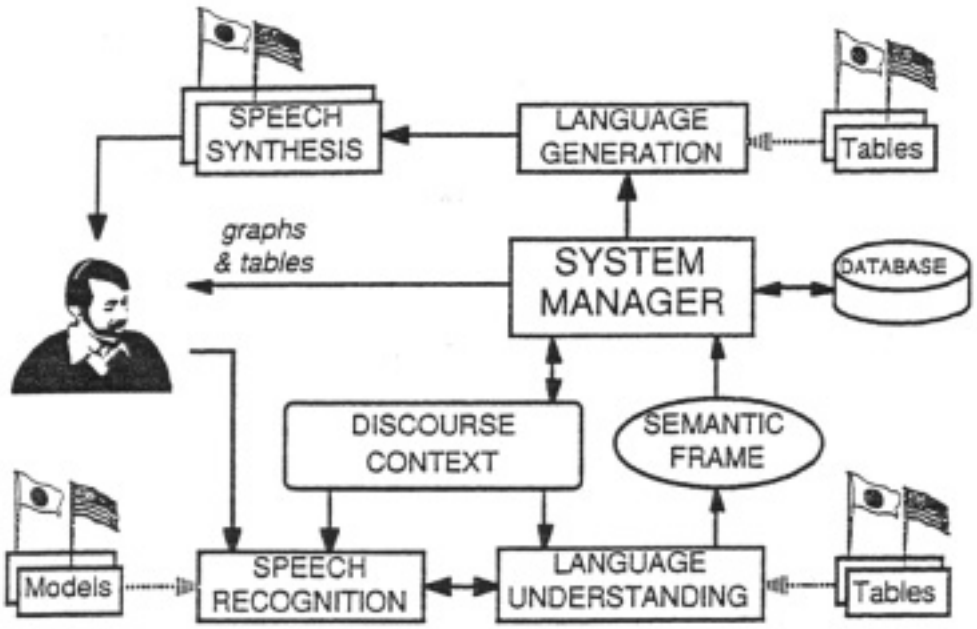


図 2.2: 音声対話システムアーキテクチャGALAXY

2.2.7.3　言語処理部

言語処理部には，TINA[14] という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
```

2.2.7.4　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，それを元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．

2.2.7.5　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．

2.2.7.6　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　対話音声特有の現象への対処

これまで，数多くの音声言語情報処理に関する研究が進められており，これらの技術を統合することで音声対話システムが実現される．しかしながら，実際にそれぞれの要素技術を組み込むにあたり，各要素技術を単純に組み合わせるだけでは不十分である．それは，個々の要素技術研究が朗読音声のような音声を対象としたものであり，対話音声を想定して構築されたものではない，という理由による．具体的には，音声認識の際に言い淀みや間投詞のような対話特有のものを扱う必要があったり，音声合成の場合には前後の文脈によってどこを強調すべきかを考慮する必要があったり，ということが挙げられる．そのため，それぞれの要素技術を対話システム用に再構築する必要がある．このような観点から，音声対話システム研究のほとんどは,「特定のモジュールについて対話音声を適切に取り扱う」という方針で研究が進められている．

現在，数多くの音声対話システムが研究・開発されているが，全てのモジュールを扱うのは多大な時間を要するため，そのほとんどは特定のモジュールに特化した研究となっている．次節以降では，各研究が対話音声をどのように取り扱っているかを中心に様々な研究例について述べる．

2.4　音声認識・理解に関する研究

音声対話システムの入力対象である対話音声（話し言葉）に特有の特徴としては，間投詞・助詞落ち・倒置の多様・言い間違い・言い直し・言い淀みといったものが挙げられる．こうした特徴を持つ対話音声の認識においては，文全体を一字一句聞き取るディクテーションではなく，キーワードなど発話内容の理解に必要な部分だけを正しく抽出するスポッティング方式というアプローチが有効であると考えられる．

2.4.1　富士観光案内音声対話システム [15]

2.4.1.1　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．また，このシステムは，マルチモーダルインタフェースとしての性質も備えている．図 2.3 にシステムのモニタ出力の様子を示す．

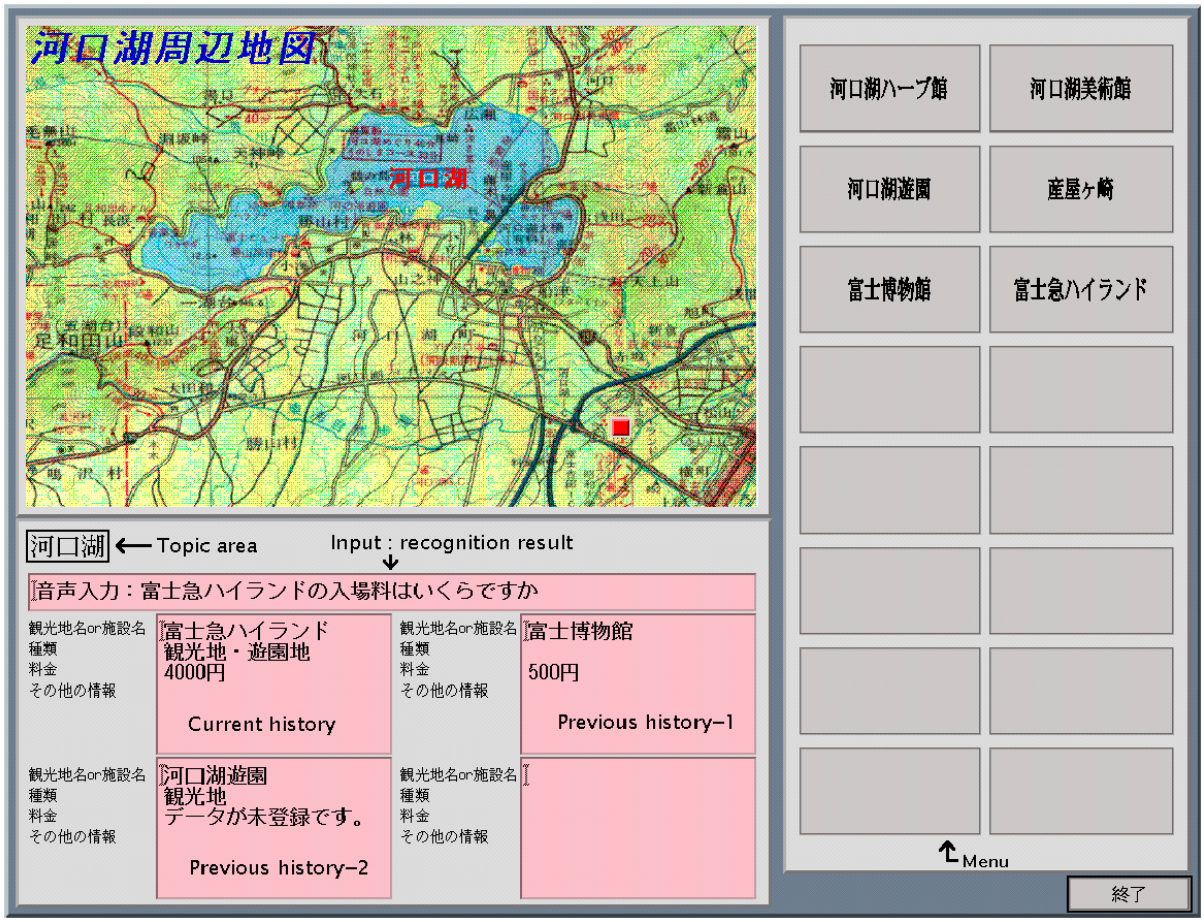


図 2.3: 富士観光案内音声対話システム

2.4.1.2　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．

この実験によって獲得されたストラテジーを基に言語理解手法を構築し，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．

2.4.2　韻律情報を用いた否定/肯定的態度の認識 [16]

2.4.2.1　システム概要

音声言語により表現され，伝達される情報は，必ずしも明確に境界を引くものではないが，主に言語情報・パラ言語情報・非言語情報の 3 つに大別することができる [17]．言語情報とは，文字言語による表記およびその前後の文脈から容易に，一義的に導出し得るものを指す．それに対し，態度・感情・話者の状態など，記号情報以外で伝わる情報のうち，話し手が意図的に制御できるものをパラ言語情報，話し手が意識的に制御できないものを非言語情報と呼ぶ．

藤江らは [16] において，聴覚的なパラ言語情報として，発話に含まれる韻律情報から，forward selection 法により主要な特徴量を抽出し，それを用いて発話が肯定的であるか否定的であるかを識別する手法を検討し，その有効性を確認している．

このシステムは，ある事柄に関して決断を迷っているユーザに対して，ユーザの要求を聞き出し，提案を行なう対話を想定している．具体的には，昼食に何を食べるか，またどの店へ行くかを迷っているユーザの相談に答えるタスクを対象としている．

2.4.2.2　パラ言語情報を用いた音声対話システム

韻律情報による態度の認識と，頭部ジェスチャの認識 [18] の結果をパラ言語情報として利用する音声対話システムを，対話ロボット ROBISUKE（図 2.4）上に構築した．

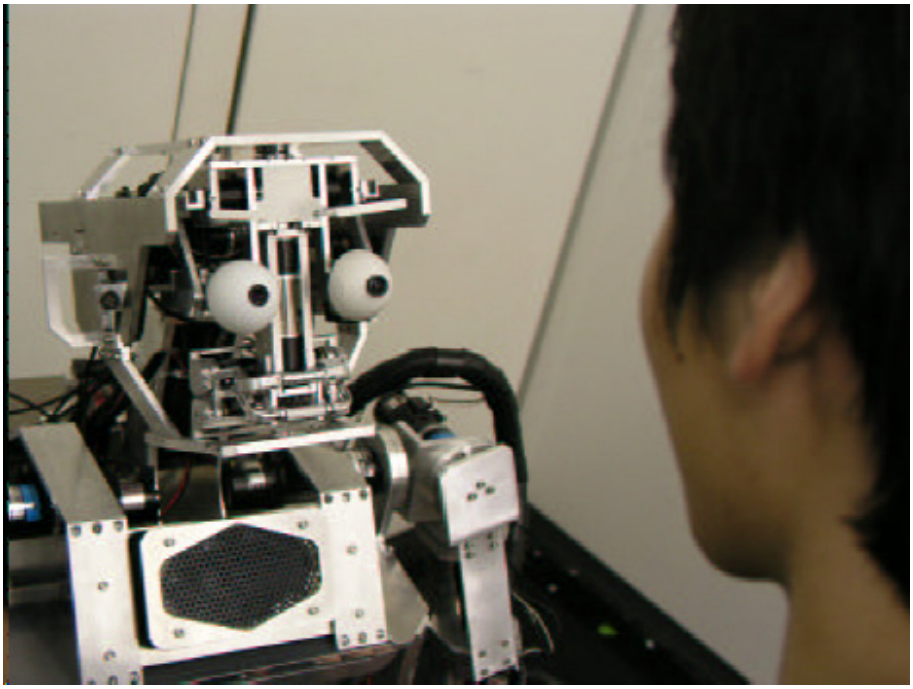


図 2.4: 対話ロボット ROBISUKE

対話例を以下に示す．U はユーザ，R は ROBISUKE の発話である．

U: お昼ごはんなんだけど，どこかいいところ無いかな

R: カレーなんかどう

U: カレーか**（否定的）**

R: それじゃあ，弁当なんかどうかな

U: ああ，弁当ね**（肯定的）**

R: 弁当なら近くにホカ弁があるよ

システムは，提案に対してユーザの応答が否定的な場合は代案を，肯定的な場合にはより具体的な提案を行なう．

2.4.3　雑音下での音声認識

音声対話システムを実環境で用いる場合，音声入力の際にどうしても周辺の雑音が混ざってしまう．特に，カーナビのようなシステムでは非常に重要な問題である．そこで，耐雑音音声認識に関する研究が非常に多くなされており，雑音データベースとしても JIS 生活環境データベース [19] や環境騒音データベース [20] 等，多数公開・利用されている．

雑音下音声認識にもいくつかアプローチがあり，音声認識時の音響モデルに対する雑音処理を行なうもの [21] や音声認識の前処理（特徴抽出）の際に雑音を抑えてしまうもの [22] 等がある．[21] では，推定された雑音と HMM 合成法を用いて音響モデルを逐次更新し，非定常雑音が重畳した音声を認識している．一方，[22] では，ベイズ推定法の一種であるパーティクルフィルタ [23] によって非定常雑音の逐次推定を行なうことを提案しており，この手法では雑音抑圧後のデータを用いた音響モデル適応等複合処理が可能である利点がある，としている．

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）[24]

2.5.1.1　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」[24] を開発した．システム構成を図 2.5 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．

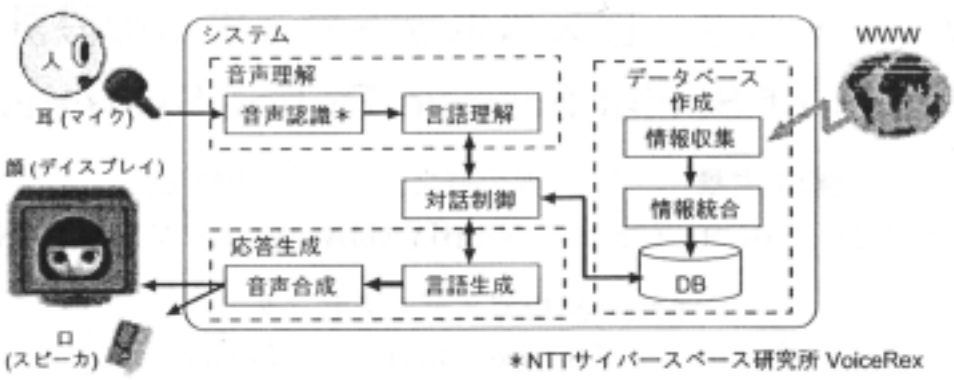


図 2.5: 「飛遊夢」システム構成図

2.5.1.2　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[25] 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 [26] 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．

2.5.1.3　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．

2.5.1.4　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．

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 [27] 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 [28]

2.5.2.1　システム概要

竹内らは [28] 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 2.6 に示す．

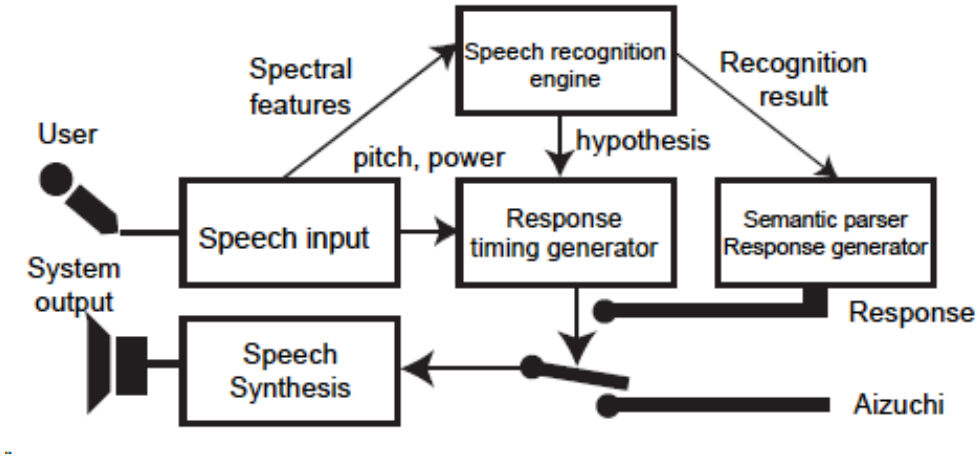


図 2.6: 雑談対話システム

2.5.2.2　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）[29] 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[30] 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 [31]，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 [32, 33]，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 [34]

2.6.1.1　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．システムの概要を図 2.7 に示す．

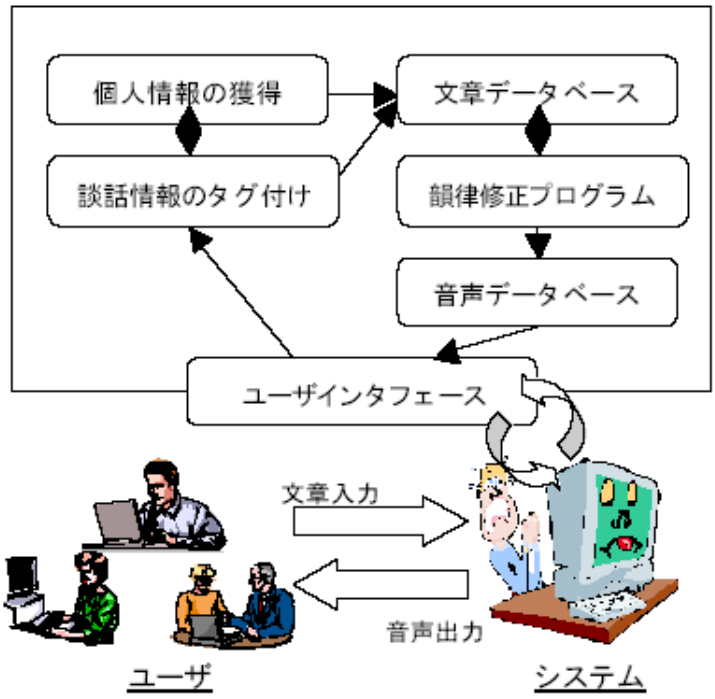


図 2.7: 談話情報を用いた韻律制御システム

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．

2.6.1.2　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．

2.6.1.3　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 [35].

2.6.1.4　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 [36].

2.6.2　学術情報検索音声対話システム [37]

2.6.2.1　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．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 2.8 に示す．また，システムの画面表示の様子を図 2.9 に示す.

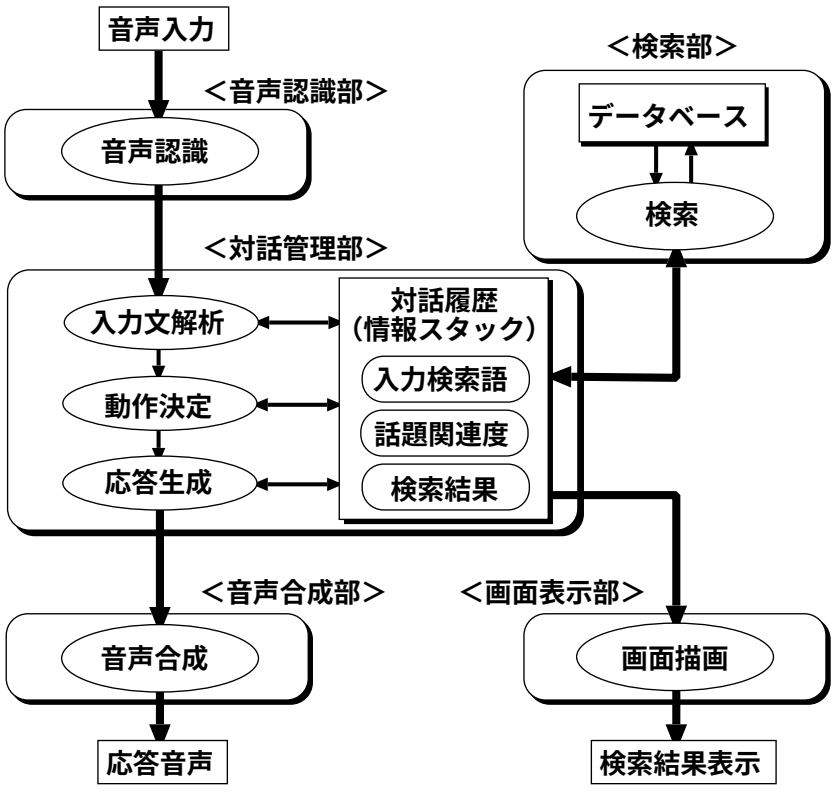


図 2.8: 学術文献検索音声対話システム

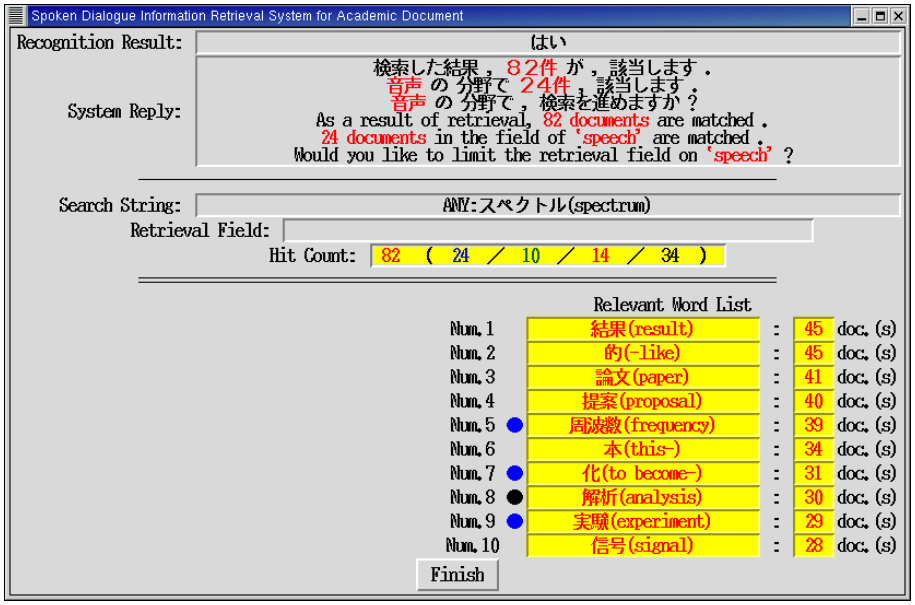


図 2.9: システムの画面表示

2.6.2.2　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．

2.6.2.3　韻律規則

[38] の韻律規則を用いて音声合成を行なっている．[38] 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 [39] 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．

2.6.2.4　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[3]

2.6.3.1　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．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 2.10 に示す．

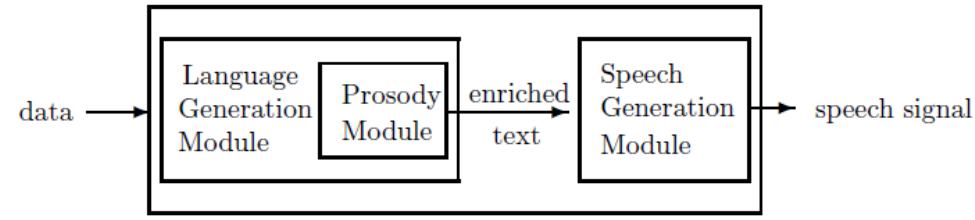


図 2.10: Data-to-Speech

「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 2.11 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．

2.6.3.2　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

図 2.11: GoalGetter

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．

2.6.3.3　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　その他の音声対話システム研究

2.7.1　傀儡（かいらい）[40]

2.7.1.1　概要

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．

新山らは「傀儡」[40] と呼ばれる自然言語処理による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 2.12 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 2.12: 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 2.13 にこのシステムの構造を示す．

2.7.1.2　音声認識部

IBM[1] の ViaVoice[41] を用いている．

2.7.1.3　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．

2.7.1.4　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．

[1]現在は NUANCE が販売・サポートを行なっている．

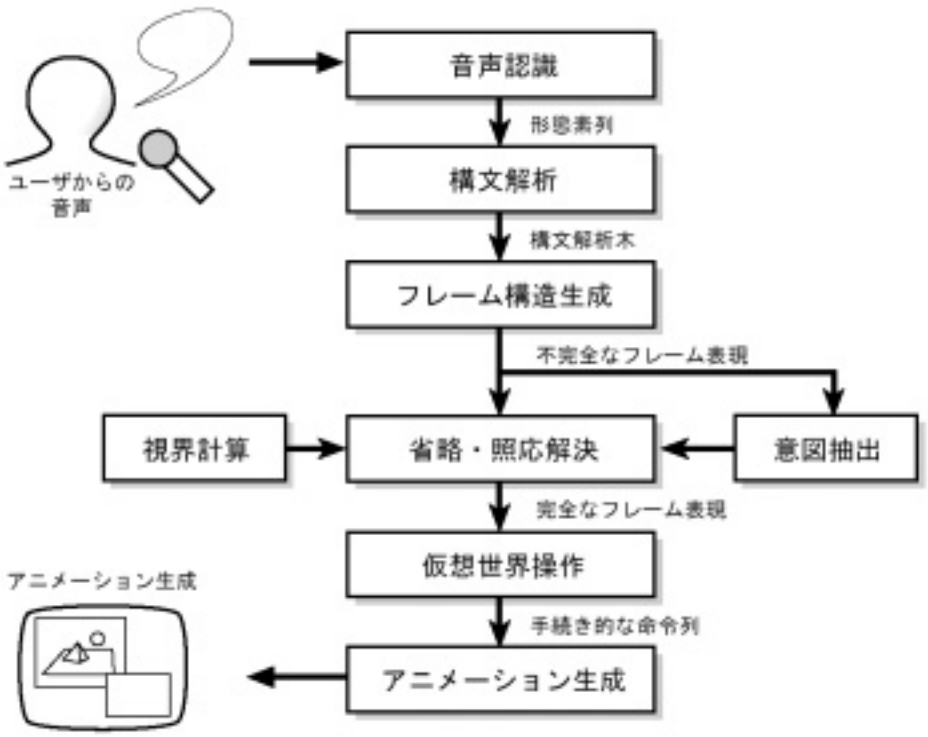


図 2.13: 傀儡のシステム構成

2.7.1.5　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．

2.7.2　擬人化音声対話エージェントツールキット Galatea[42]

2.7.2.1　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[42]” 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[43]，W3C[44]，JEIDA-62-2000[45] 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 2.14 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．

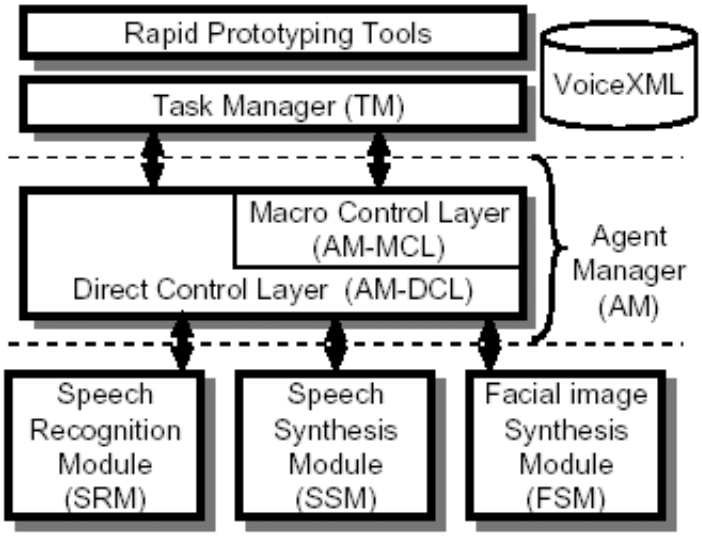


図 2.14: GALATEA の全体構成

2.7.2.2　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[46] 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[47] のフォーマットに対応する．

2.7.2.3　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 [48]” で形態素解析したのち，[49] 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 [50, 51, 52] 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 2.15 の構成をとる．

Galate 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，[45, 53] 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 [54]．この記述例を図 2.16 に示す．

2.7.2.4　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 [55]

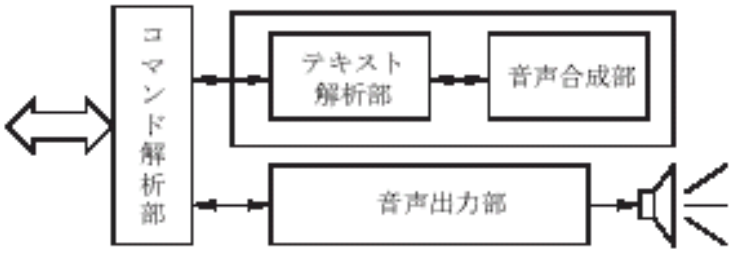


図 2.15: Galatea Talk の構成

```
<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 2.16: Galatea Talk による発話文の記述例

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．

2.8　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．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においても，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いない．また，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．[42] では，アクセント結合は考慮はされているが，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．

本論文では，応答生成（＝応答文生成+韻律制御）について次章以降でより詳細に取り上げ，概念音声合成の枠組を用いたより汎用的な応答生成手法について提案する．

第3章

エージェント音声対話システム

3.1　はじめに

音声対話システムにおいては，応答音声に関しても対話音声のような，より自然な韻律制御がなされた音声であることが望ましい．しかしながら，多くの音声対話システム研究開発において，システム応答音声に関してはテキスト音声合成に依っているものが非常に多くを占めており，対話音声合成に関する研究はほとんどなされていない．

テキスト音声合成は，テキスト表層文のみから音声合成を行なう手法である．しかしながら，実際の対話音声を合成することを考えると，統語構造や談話情報といった高次の言語情報についても考慮する必要がある．音声対話システムにおいては，応答内容を 1 からシステムが生成するため，高次の言語情報も容易に得ることができる．そのため，これらの情報を応答音声に反映させることのできる概念音声合成 [5] の枠組を用いるのが望ましいと考えられる．

日本語における対話音声合成の研究は，[56] によって概念音声合成の枠組を用いた応答生成を行なうことが提案されている（実装にはいたっていない）が，現在にいたるまでこのような研究は非常に少なく，[37] において研究がなされている程度である．

そこで本章では，概念音声合成の枠組を用いた応答生成手法についての提案手法を述べる．具体的には，応答生成における言語生成，音声合成の 2 ステップに関して概念音声合成の実現のために提案する手法について詳述する．言語生成においては，言語情報を常に構文木構造を保持したまま扱い，またタグを用いることで言語情報の統一的な処理を実現している．音声合成については，重要度や新規性といった談話情報を応答音声の韻律に反映させ，さらにアクセント結合規則の導入によってより自然な応答音声の生成を実現している．さらに，これらの手法をエージェント音声対話システムに実装し，応答音声の評価を行なう．

本章の構成は以下のようになっている．まず始めに，本章で述べる概念音声合成の枠組を実装したエージェント音声対話システムについての概要を述べる．続いて，応答文の言語生成に関して，言語情報の取り扱い手法について述べる．そして，音声合成に関して，適切な韻律情報の応答音声への反映について述べる．最後に，これらの手法をシステムに実装し，聴取実験を行なった結果について述べる．

3.2　システム概要

本システムのタスクは，仮想空間中の物体を操作することであり，仮想空間内にいるエージェントに自然言語で指示することで操作を行なう．仮想空間の状態はリアルタイムで画像（図 3.1）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したり指示したりして物体を操作する．

図 3.2 に，エージェント音声対話システムの概念的な構成を示す．以下，各モジュールについて詳述する．

図 3.1: エージェント音声対話システムの画像出力

3.2.1　音声認識部

音声認識部には，Julian v.3.2[46] 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．

3.2.2　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．本システムにおいては，音声入力は研究の対象とはしておらず，ユーザの入力は非常に限られたものとなっている．そのため，構文解析部での実際の作業は，あらかじめ用意された入力文用テンプレートのどれに入力文がマッチするかを探索し，それを対話管理部に渡している．

3.2.3　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，現在の空間の状態に応じて応答文を生成したり，エージェントを動かしたりする．また，対話の必要がある場合は，応答文を生

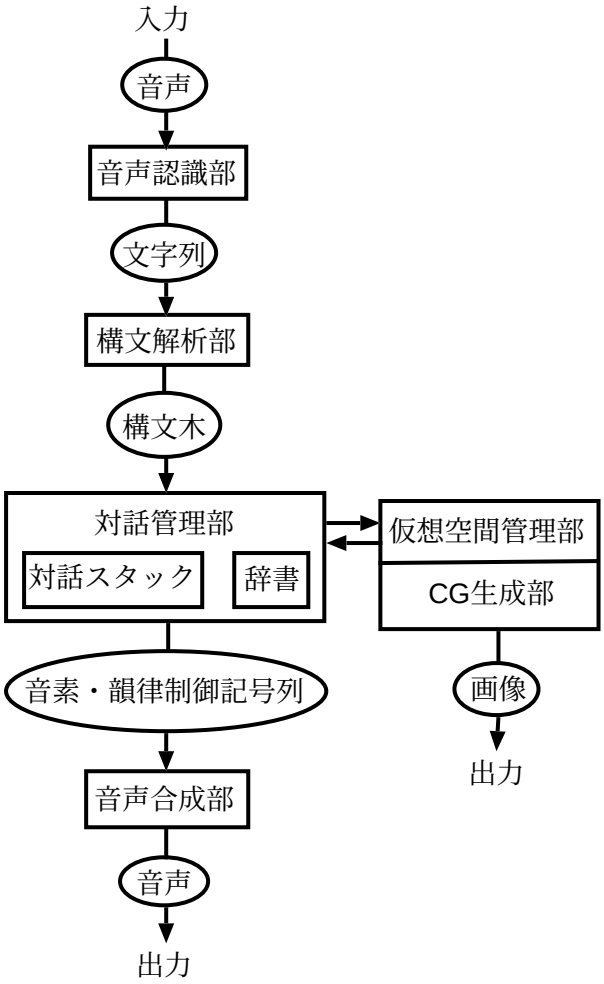


図 3.2: エージェント音声対話システム構成

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．

詳細は 3.5 節，特に 3.5.1 節〜3.5.9 節で述べる．

3.2.3.1　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（アイテム・場所など，詳細は 3.5.2 節）についてもスタックを持つ．

3.2.3.2　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．詳細は 3.5.1 節で述べる．

3.2.4　音声合成部

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．詳細は 3.4 節で詳述する．

3.2.5　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．詳細は，3.5 節のうちの 3.5.10〜3.5.11 節で述べる．

3.2.6　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[57] を用いる．詳細は，3.5 節のうちの 3.5.12〜3.5.13 節で述べる．

3.3　言語情報の取り扱い

音声対話システムでは，システムの内部情報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，言語に変換する必要のある内部状態に，言語情報を付加する必要がある．

概念音声合成の枠組による音声合成の実現のためには，応答音声の韻律に反映されるべき情報である統語構造や談話情報といった高次の言語情報をいかに保持するかが焦点となる．

そこで本論文では，言語情報を常に構文木構造（統語構造）を保持したまま扱い，さらにタグを用いることで談話情報の保持も行なう手法を提案する．

詳細を次節以降にて述べる．

3.3.1　言語情報の表現

本提案手法は，言語情報を一貫して構文木構造を保持したまま扱う，という方針を取る．その実現のために，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 3.3 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．

3.3.2　タグの付与

前節で述べた例では,「イスを机の前に置いて」という文しか表現できない．そこで，単語の代わりにタグを用いることで，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ようにする．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 3.4 に示すとおりである．これを LISP 形式で表現する際に,「(置く$PRED（を（$ITEM））

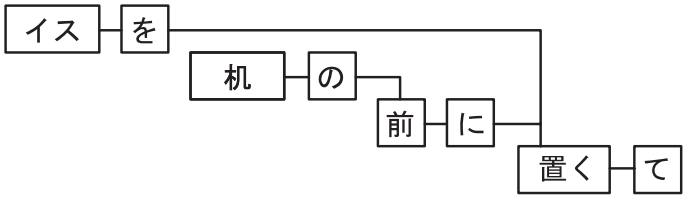


図 3.3: 「イスを机の前に置いて」の構文木構造

(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．

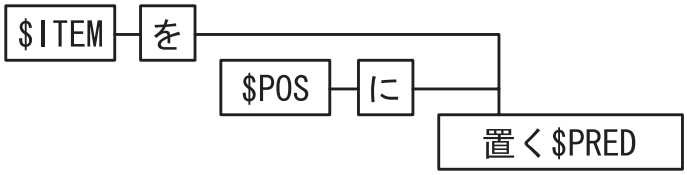


図 3.4: 「**アイテム**を**場所**に置く」のタグと構文木構造

また，タグは，表層文のみならず，文章における重要度や新規性といった情報も保持する．重要度や新規性は，それぞれ 0（重要でない/新しくない）または 1（重要/新しい）の値を取る．今回のシステムでは，重要度は方向・場所を表す語を 1，それ以外を 0 とし，タグがこの情報を保持する．また，新規性は対話の履歴を参照し，初出（かつ重要度 1）の語であれば 1，それ以外を 0 としている．それが韻律制御（3.4.2 節）の際に用いられる．

3.4　音声合成

現在，音声対話システムの研究・構築が盛んに行なわれるようになってきているが，それらの多くは，音声出力には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．しかしこの TTS システムは，テキストのみから音声を合成することを目的として作られたものであり，高次の言語情報を韻律へ反映させることまでは考慮されていない．そのため，音声対話システムに求められるような対話音声を合成する用途には不向きである．

本論文では，概念音声合成の枠組を用いた音声合成の実現を目標としている．3.3 節で提案する手法により，統語構造や談話情報といった高次の言語情報を取り扱うことができるようになる．そこで以下では，これら高次の言語情報をいかに応答音声に反映させるかについて，テキスト音声合成との比較を行ないながら説明する．

3.4.1　テキスト音声合成（Text-to-Speech）

3.4.1.1　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，ポーズ等の決定
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

3.4.1.2　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ったからである．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等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，すなわち概念音声合成（CTS : Concept-to-Speech）[5] 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 [6]．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，3.3 節で既に述べた通りである．以降では，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．

3.4.2　音声合成の韻律規則

本対話システムで用いる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 [58, 36] 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（詳細は付録 A 章）．これらの規則は，朗読音声に対して構築された規則 [39] を，対話音声の分析結果に基づいて対話音声向けに変換したものである [38]．各指令の大きさは数量化分析により決められており，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表す．

以下，フレーズ指令とアクセント指令のそれぞれについて，各々の指令を配置する位置を決定する指令の生成規則と，指令を構成するパラメータの意味について説明する．

3.4.2.1　フレーズ指令

フレーズ指令には，文頭・文中・文末の 3 種類の指令がある．これらは，それぞれ P111222・P12・P0 といった記号で表される．フレーズ指令のパラメータは表 3.1 の 6 種類がある．ここで 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．

表 3.1: フレーズ指令のパラメータ

パラメータ	値	意味
item1	1	FRD を開く
	2	FRD を閉じる
item2	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
item3	1	話題を変更している
	2	話題を変更していない
item4	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
item5	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
item6	1	疑問の終助詞「か」で終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

文頭フレーズ指令（Ph）は，これらのパラメータのうち $\mathrm{P}I_1I_2I_3I_4I_5I_6$ の 6 つのパラメータを持つ記号で表される．文中フレーズ指令（Pm）は，これらのパラメータのうち $\mathrm{P}I_2I_5$ の 2 つのパラメータを持つ記号で表される．文末フレーズ指令（P0）は，パラメータを持たずに P0 という記号で表される．

各パラメータの設定のうち，重要語（重要度が 1 となる語）を含むかどうか（I_2），また生成文の構文構造（I_4，I_6）については第 3.3.1 節で述べた手法により得られる．他のパラ

メータについては，対話履歴や単語辞書を参照することで得られる．そして，各パラメータの値によって，フレーズ指令の大きさが決定される（詳細は付録 A.3.2.1〜A.3.2.2 節）．

フレーズ指令の挿入規則を以下に示す．

- 文の先頭には Ph を挿入し，文末には P0 を挿入する．文の境界には S1 を挿入する．
- ICRLB 境界には Pm を挿入する．ただし，直前の Ph/Pm からの距離が L1 モーラ以下であるときは，Pm を省略する．
- 列挙表現の境界には S3Pm を置く．ただし，直前の Ph/Pm からの距離が L1 モーラ以下であるときは，Pm を省略する．
- ICRLB 境界が L2 モーラより長ければ，全ての韻律句が L2 モーラ以下になるように Pm を挿入する．この際，分割してできる韻律句の長さがなるべく均等になるようにする．
- 節境界に S3Pm を挿入する．直前の韻律文境界までの距離が L2 モーラ以下なら，Pm のみを置く．

ここで，S1･S2･S3 は休止記号であり，数字が小さいほど休止の長さが長い．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．変数は L1=5，L2=15 を用いている．

3.4.2.2　アクセント指令

アクセント指令には，D$I_1I_2I_3$・F$I_1I_2I_3$・A0 の 3 種類があり，D$I_1I_2I_3$ は起伏型または頭高型のアクセントの立ち上げ，F$I_1I_2I_3$ は平版型のアクセントの立ち上げ，A0 は両者の立ち下げを示す記号である．

これらの指令を配置する位置について，頭高型のアクセントの立ち上げ指令は，対応する韻律語素の第 1 モーラの直前に配置し，起伏型および平版型のアクセントの立ち上げ指令は第 1 モーラの直後に配置する．アクセントの立ち下げ指令については，頭高型・起伏型はアクセント核の直後に配置する．平版型のアクセント立ち下げ指令は，対応する韻律句の最後尾，または後続の韻律語素が頭高型の場合は最終モーラの直後に配置し，それ以外の場合は後続韻律語素の第 1 モーラの直後に配置する．アクセント立ち上げ指令におけるパラメータを表 3.2 に示す．

各パラメータの設定については，I_1・I_2 に関してはそれぞれ，第 3.3.1 節の手法によって重要度・新規性，構文構造を参照することで設定を行なう．また，品詞（I_3）については，辞書を参照することによって設定する．そして，各パラメータの値によって，アクセント指令の大きさが決定される（詳細は A.3.2.4〜A.3.2.5 節）．

3.4.2.3　アクセント結合規則 [59]

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

表 3.2: アクセント指令のパラメータ

パラメータ	値	意味
item1	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
item2	1	フレーズの先頭
	2	フレーズの途中
item3	1	名詞
	2	動詞
	3	形容詞・副詞
	4	指示語・疑問詞
	5	接続詞

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だけでなく，このアクセント結合についても考慮する必要がある．

[59] は，匂坂らにより一連の結合規則 [49] を聴取実験により見直したものであり，付属語アクセント規則の他に，複合名詞や接頭辞に関する規則，文節間規則がある．本対話システムでは，現在の対話システムで生成される応答文の範囲を考え，付属語アクセント規則のみを考慮している．将来的には，複合名詞や接頭辞に関する規則，文節間規則も記述する必要がある．

アクセント結合様式には表 3.3 に示すものがある．N はアクセント核で，M はアクセント価である．

3.4.2.4　韻律制御記号を含む音素記号列の生成

構文木構造と単語ごとに重要度と新規性を持つ文から，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．

1. 接続に従って活用形を決定
2. アクセント結合規則に従ってアクセント指令の位置を決定
3. 単語の重要度に従ってアクセント指令のパラメータを決定
4. 構文木構造に従ってフレーズ指令の位置を決定
5. フレーズ指令のパラメータを決定

3.4.2.5　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「電話の所に移動できないのですがどうすればいいでしょうか」という音声の音声合成

表 3.3: 付属語アクセント結合様式
(N_1 モーラ M_1 型 + N_2 モーラ $\widetilde{M_2}$ 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	
	$M_1 = 0$ のとき	$M_1 \neq 0$ のとき
(F1) 従属型	M_1	
(F2) 不完全支配型	$N_1 + \widetilde{M_2}$	M_1
(F3) 融合型	M_1	$N_1 + \widetilde{M_2}$
(F4) 支配型 1	$N_1 + \widetilde{M_2}$	
(F5) 支配型 2	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)	$N_1 + \widetilde{M_2}$	M_1 and $N_1 + \widetilde{M_2}'$

器への入力は,「P121211 de F311 n wa no A0 to D411 ko ro A0 ni P22 i D412 do o de ki A0 na i no de su ga P22 D413 do A0 o D412 su A0 re ba D413 i A0 i de sho o ka P0 S1」となる.

3.4.2.6　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等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/白色雑音音源や残差音源といった音源で駆動する.

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等，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．本研究では，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 [60] を基にしたものを用いる．[60] 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．

3.5　実装

3.5.1　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[61] 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

3.5.1.1　品詞辞書

品詞情報を格納する．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 付属語（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）
（アクセント指令については 3.4.2.2 節にて詳述）

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．

3.5.1.2　活用辞書

活用型・活用形を表す．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

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．

3.5.1.3　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．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	単語を表示する場合の文字列．省略した場合は identifier が代わりに用いられる
part	単語の品詞．入れ子にすることが可能
stem	単語の語幹（活用語のみ）
inflection	活用型（活用語のみ）
connection	単語の接続．省略した場合は品詞に登録された接続を用いる
phoneme_symbol	単語の発音の音素記号列．アクセント核（accent_nucleus）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（3.5.2 節）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．

付属語には，[59] によって得られたアクセント結合規則を付与した．詳細は 3.4.2.3 節で詳述する．辞書におけるパラメータの意味は以下に示す通りである．

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

3.5.2　対話用データ

本章で述べるエージェント音声対話システムでは，対話用データとして主にアイテム・場所という 2 項目を扱う．

本提案手法の単語辞書（3.5.1 節）に記述する対話用データは，音声対話システムのタスクごとに異なる．本章で構築するエージェント対話システムにおいて，単語に関連付ける対話用データには表 3.4 のようなものがある．対話用データは，ユーザ発話の理解と応答文生成に用いられる．

3.5.2.1　アイテム

アイテムとは，仮想空間中に置かれたオブジェクトのことであり，種類・色といった属性を持つ．例えば図 3.5 の場合，左の図は種類が「電話」で色が「赤」であり，右の図は種類が「イス」で色が「灰色」である．種類・色はそれぞれ複数持つことができる．

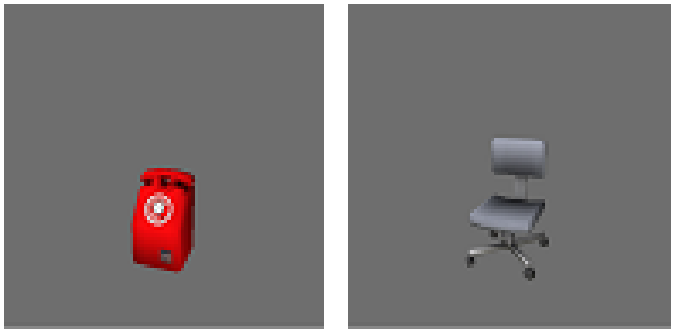


図 3.5: アイテムの例

3.5.2.2　場所

場所は，仮想空間中の位置を表すものである．空間はグリッドに区切られており，アイテムはそのグリッドに置くことができる．場所はシステム内部ではグリッド座標で扱われる．

表 3.4: 対話データ

attribute	単語の対話システムでの分類を表す
identity	単語の対話システムでの値を表す

対話用データの例を表 3.5 に示す．

表 3.5: 対話データの例

attribute	identity	内容
item_type		アイテムの種類を表す
	desk	机
	shelf	棚
item_color		アイテムの色を表す
	red	赤色
	blue	青色
position		場所を表す
	up	上
	down	下
item		アイテムを表す
	mono	もの
agent_action		エージェントの動作を表す
	take	持つ
	put	置く

3.5.3　対話管理部の処理

対話管理部での処理を図 3.6 に示す．まず，初期状態から構文木構造を含んだ発話文を受け取ると，その内容に応じて処理を分岐する．現状では「エージェントへの命令」と「アイテムについての質問」の 2 つの機能があるが，このような仕様にしておくことで，さらに新しい機能を容易に付け加えることができる．

発話内容による処理の分岐では，得られた構文木構造を持ったユーザの発話文に含まれる対話用データの構造にルールを定めることで実現する．例えば「エージェントへの命令」のルールは,「ルートとなる文節が命令を表す単語で命令を表す助詞で終わっている」である.「アイテムについての質問」のルールは,「ルートとなる文節に「何」という語が含まれている」である．

3.5.4　エージェントへの命令の処理

エージェントへの命令の処理を図 3.7 に示す．まず，発話文から命令の種類を判別する．この命令を基に，発話文中からアイテム・場所を表す句を探し，それの指すものを特定する．続いて，エージェントの動作を決定する.「アイテム・場所の特定」,「エージェントの動作の決定」のそれぞれの段階で，システム自身で問題を解決できない場合は，ユーザとの対話を通して問題を解決する．

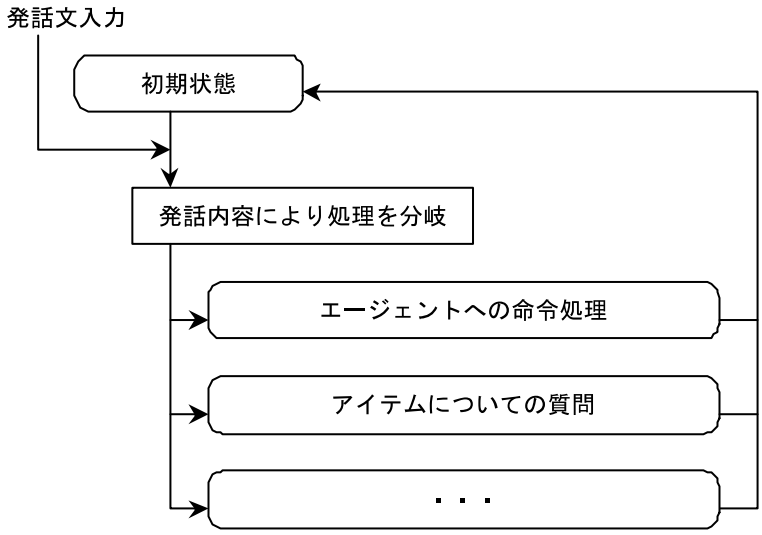


図 3.6: 対話管理部での処理の流れ

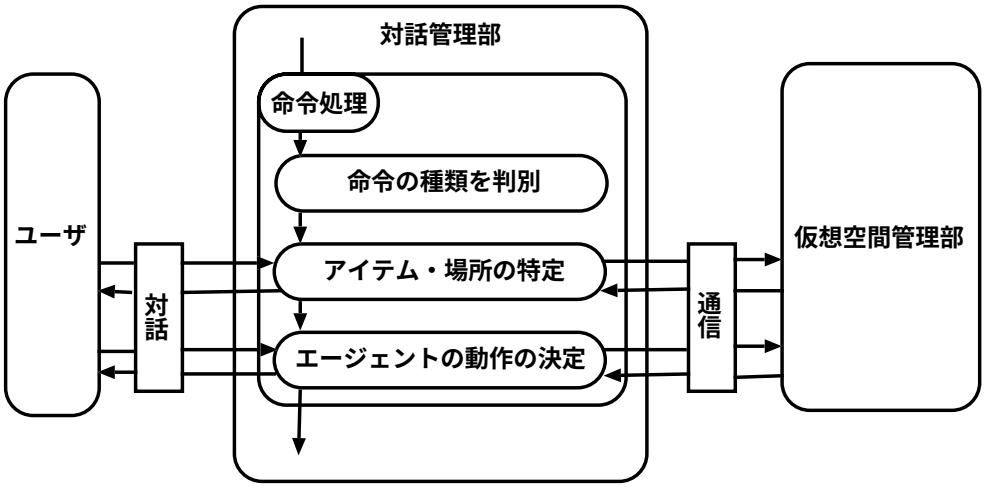


図 3.7: エージェントへの命令の処理の流れ

3.5.5　省略・照応の解決

対話管理部は，アイテム・場所に関するスタックを持っている．命令の引数となったアイテム・場所をスタックに保存する．アイテム・場所を表す語が省略されたり,「それ」などの直示的照応詞が用いられたりした場合は，スタックを探索してエージェントの意図を汲み取る．省略・照応の解決が行なわれた場合は，ユーザとの確認対話を行なう．

3.5.6　アイテム・場所の特定

ユーザが「机」と言っても，仮想空間中に机が複数ある場合は，ユーザがどの机を指しているのかを特定しなくてはならない．アイテム・場所の特定は，ユーザの発話文の構文木の枝側のアイテム・場所から決定していく．

アイテムの特定の処理は，命令を表す動詞に助詞「を」を伴ってかかる節において，そこに属する単語の対話用データを参照し，アイテムを表す単語（すなわち **attribute** が **item_type** であるもの）を探す．その単語の対話用データの **identity** を用いて仮想空間管理部と通信を行ない,「机」を探すことでアイテムを特定する．

3.5.7　エージェントの動作の決定

エージェントの動作を命令できるようにするために,「状態」・「動作」・「命令」を定義する．これらは，辞書と同様に XML で記述される．

3.5.7.1　状態

仮想空間の状態を記述する．状態には表 3.6 のものがある．

表 3.6: 状態

名前	状態の表示（内容）
movablef_o	**アイテム**の前に移動できる
movablef_p	**場所**の前に移動できる状態
have_nothing	手が空いている状態
frontof_o	**アイテム**が目の前にある状態
have_o	（エージェントが）**アイテム**を持っている状態
nothing_on_p	**場所**が空いている状態
frontof_p	**場所**が目の前にある状態
o_on_p	**アイテム**が**場所**にある状態

状態の持つパラメータは以下のようになる．

name　名前

item0　アイテムを引数に持つか否か

position0　場所を引数に持つか否か

sentence_key　状態を表す文の名前

3.5.7.2　動作

個々の動作は前提状態と終了状態を持ち，動作とは空間の状態を前提状態から終了状態にするものである．動作には表 3.7 のようなものがある．

表 3.7: 動作

名前	前提状態	終了状態
場所の前にある	場所の前に移動できる	場所が目の前にある
アイテムの前に移動する	アイテムの前に移動できる	アイテムが目の前にある
アイテムを持つ	アイテムが目の前にある 手が空いている	アイテムを持っている
アイテムを場所に置く	場所が空いている アイテムを持っている 場所が目の前にある	アイテムが場所にある

動作の持つパラメータは以下のようになる．

name　名前

item0　アイテムを引数に持つか否か

position0　場所を引数に持つか否か

sentence_key　動作を表す文の名前

pre_state　前提状態（複数指定することができる）

end_state　終了状態

3.5.7.3　命令

命令は目標状態を持ち，命令とはエージェントに空間の状態を目標状態にするように促すものである．命令には表 3.8 のようなものがある．

表 3.8: 命令

名前	目標状態
持つ（take）	**アイテム**を持っている
移動する（move）	**場所**が目の前にある
置く（put）	**アイテム**が**場所**にある

命令のパラメータは以下のようになる．

name　名前

item0　アイテムを引数に持つか否か

position0　場所を引数に持つか否か

sentence_key　命令を表す文の名前

target_state　目標状態

3.5.7.4　エージェントの動作の決定の手順

システムは「動作」を格納するスタックを持つ．エージェントの動作を決定するには，まずユーザ発話から「命令」を抽出し，目標状態を設定する．続いて，設定した目標状態を終了状態とする「動作」を検索し，その「動作」をスタックに格納し，その「動作」の前提状態が満たされているかどうかを判断する．満たされていれば「動作」を実行し，スタックから「動作」を取り出す．満たされていなければ，その前提状態を新しい目標状態とする．以上を再帰的に実行し,「動作」スタックが空になるまで繰り返す．そして，全ての目標状態を実現することでユーザの命令を実行する．このとき，エージェント自身では問題を解決できない場合は，ユーザと対話を行なうことで問題を解決する．

3.5.8　アイテムについての質問

アイテムについての質問での処理は，図 3.8 に示す流れで行なわれる．アイテムについての質問では，前述した通り「何」という単語を含むが，これにかかるアイテムを表す構文木のノードを探し，それを 3.5.6 節で述べた手法で決定する．エージェント自身で決定できなければ，ユーザと対話を行なう．アイテムが特定できれば，その種類を発話する．

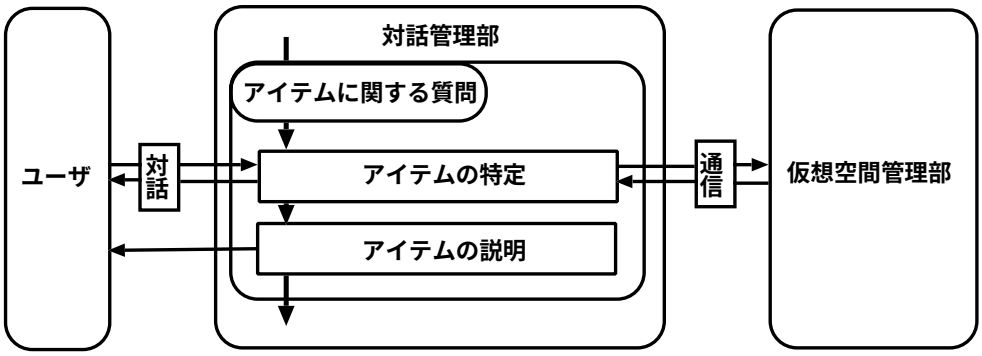


図 3.8: 「アイテムについての質問」での処理過程

3.5.9　応答文生成

3.5.9.1　「アイテム・場所の特定」での応答文生成

アイテムの特定では，図 3.9 に示すように，アイテムの検索結果に言語情報を付加し，それを適宜組み合わせることで文を生成する．具体的には，該当するアイテムが複数ある状態には「**アイテム**はいくつかある」という文，該当するアイテムがひとつも見つからないという状態には「**アイテム**はひとつもない」という文を与えておき，ユーザの発話からアイテムを表す言葉を取り出して**アイテム**に接続し，できた文を回答を促すための文に接続する．アイテムが省略されたり,「それ」などの直示的照応詞で表されている場合は，スタックから直前のアイテムを取り出して補完することで省略・照応の問題を解決している．

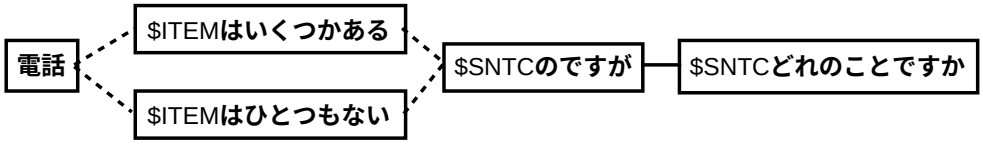


図 3.9: アイテムの特定での応答文生成法

3.5.9.2　「エージェントの動作の決定」での応答文生成

エージェントの動作の決定の段階では，システムが何をしようとしていて，何が問題なのかを伝える必要がある．全てを伝えると図 3.10 のようになり，非常に冗長な文が生成され，重要な情報がわかりづらくなっている．そのため，実際のシステムでは，直前の解決できない状態と要求だけを出力し，この場合は「黒い電話の前に移動できないのですがどうすればいいでしょうか」と出力する．

命令	黒い電話をテレビの前に置く	てと命令されましたが
状態	黒い電話を持っている	ないので
動作	黒い電話を持つ	うと思ったのですが
状態	黒い電話の前にいる	ないので
動作	黒い電話の前に移動する	うと思ったのですが
状態	黒い電話の前に移動できる	ないのですが
要求	どうすればいいでしょうか	

図 3.10: システムの思考過程の表示

3.5.10　仮想空間管理部

仮想空間は XML で記述される．仮想空間はグリッドに区切られており，アイテムはそのグリッド上に置くことができる．1 つのグリッドにはひとつのアイテムを置くことができる．仮想空間には，エージェントとアイテムが配置されている．エージェントは仮想空間中のアイテムのないところを動きまわることができる．また，エージェントはアイテムを持ち上げ，アイテムを持ったまま移動し，アイテムを空間の空いている場所に置くことができる．また，アイテムを右回り・左回りに回転することもできる．

3.5.11　仮想空間中での出来事の記録

3.5.11.1　イベントとそのデータ構造

音声対話システムにおいて，位置の情報と時間の情報をきちんと扱った研究は意外と少ない．多くのシステムでは空間のみを扱うか，対話履歴という時間のみを扱うに留まり，これら 2 つを同時に扱ってはいない．

過去に起こった出来事を対話で参照するためには，それらの出来事を記憶している必要がある．本システムでは，過去の出来事を Event として記憶するようにする．Event は，図 3.11 のように起こった時間，場所，そこに関係した人，物等を要素として持っていて，それぞれをすぐに参照できるようにする．要素は，Event の種類によって変化してくる．Event は，図 3.12 のようり新しいものから参照できるように，連結データとして管理しておく．Event は，様々なものから参照できるような仕組になっていなくてはならない．

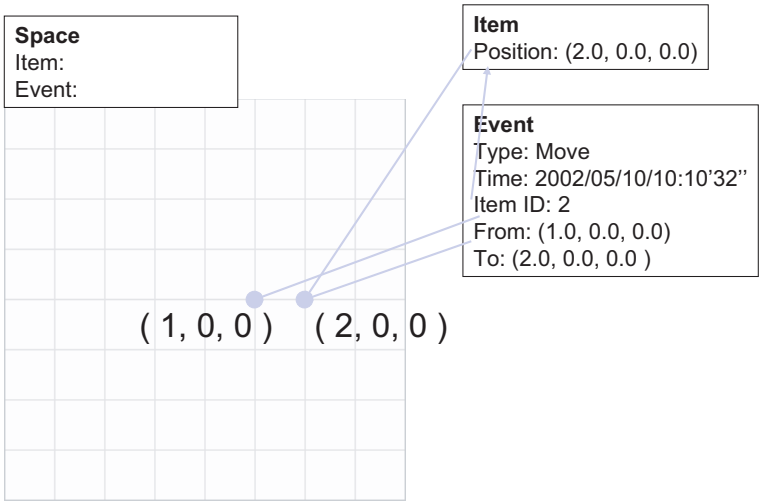


図 3.11: イベントデータの関係

例えば，場所についての対話をしているところを想定してみる.「ユーザが先週あそこに何かを置いていたような気がするのだけど，何だっけ?」と尋ねた場合，システムはその場所に関連するイベントを走査して，先週その場所にあったものを探し出す必要がある．こ

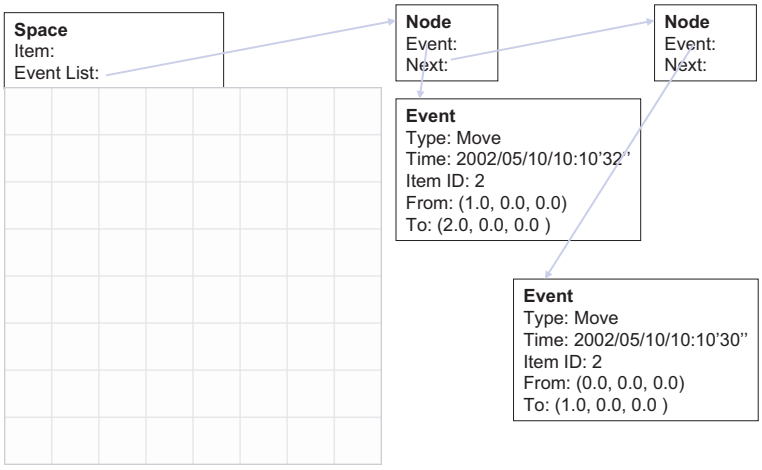


図 3.12: イベントデータリスト

の時，場所に関するイベントの走査は，図 3.13 のような連結リストをたどっていくことで可能となるようにする．

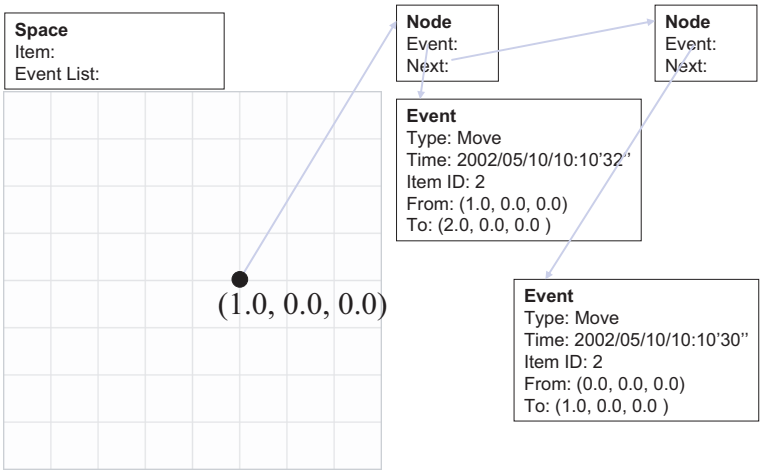


図 3.13: 場所によるイベントの検索

また，ユーザと「もの」について対話を行なっている場合，その「もの」についての Event を参照する必要がある．例えば,「この机昨日はどこにあったっけ?」と尋ねられた場合,「机」に関する Event を走査する必要がある．この時，物体に関するイベントの走査は，図 3.14 のリストをたどっていくことで実現できるようになっている必要がある．

3.5.11.2　イベントの保存

仮想空間中で起きた出来事は，Event として記憶されている．本システムでは，対話が終了しても（プログラムが終了しても）これらの Event を保存しておいて，次回の起動時

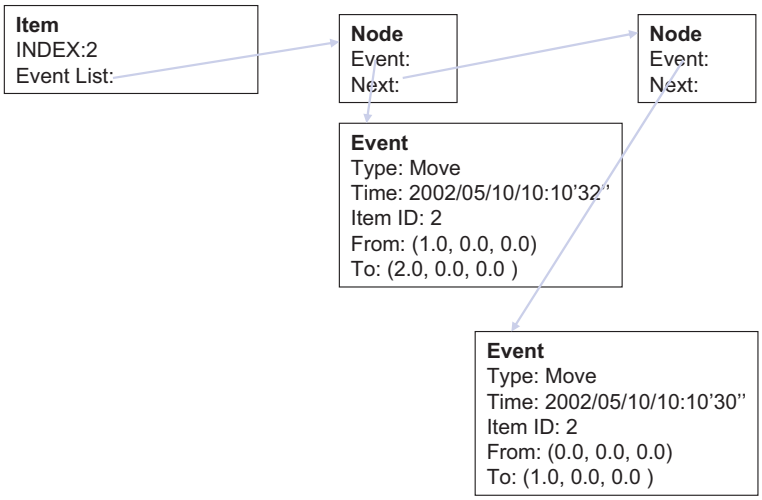


図 3.14: アイテムによるイベントの検索

もこれらの Event を参照しなくてはならない．そこで，空間の保存形式として XML 形式を用いる．

3.5.12　エージェント

エージェント（図 3.15）は，仮想空間中で様々な動作を行なう．エージェントの 3 次元モデルとモーションは，LightWave Scene ファイルにより記述される．LightWave Scene ファイルは，Netwek 社 [62] の 3DCG ソフトウェアである LightWave3D のファイルである．

図 3.15: エージェント

3.5.12.1　エージェントのモーションの決定

エージェントの動作は複数ある（図 3.16）ので，同じ構造を持つモデルに対して複数のモーションを定義しなくてはならないが，LightWave Scene ファイルは 1 つのモーションしか定義することができない．そこで，同じモデルに対して異なるモーションを定義した LightWave Scene ファイルを用意し，それを同時に読み込むことで解決する．

図 3.16: アイテムを持つエージェント

エージェントのモーションが切り替わるところでは，モーション間で補間を行なう．LightWave Scene ファイルでは，モーションはオブジェクトの位置 (x, y, z)・大きさ (S_x, S_y, S_z)・回転 (h, p, b) により定義される．LightWave Scene ファイルは，オブジェクト毎にモーションに対して時間軸に沿ったキーフレームが設定され，ある時間のモーションは，その前後のキーフレームから補間することで決定される．このようにして得られたモーションの値に対して，さらにモーション間で線形補間を行なう．ここで回転については，ジンバルロックを防ぐために，3 軸回転からクオータニオンに変換し，そこで線形補間を行なう．

3.5.12.2　ボーンデフォーメーション

エージェントの 3 次元モデルは，LightWave Object ファイルにより頂点とポリゴンとテクスチャの集合として定義されている．エージェントは頂点を移動することでその形を変える．エージェントにはボーンがアサインされており，ボーンを変形することでボーンにアサインされた頂点が移動し，エージェントは歩いたり，アイテムを持つ格好をすることができる．1 つの頂点は複数のボーンにアサインされ，ボーンによる変形をウェイトに応じてブレンディングすることで頂点の位置を決定する [63].

3.5.12.3　エージェントのパスプランニング

エージェントは，アイテムを持つ等する場合には移動しなくてはならない．アイテムのところまで行くことができるか否か，行くことができる場合はさらに，そのパスを調べな

くてはならない．ここでは A*探索を用いてエージェントのパスを求める．

3.5.12.4　アイテム

アイテムは仮想空間中に配置され，エージェントにより移動される．アイテムの 3 次元モデルは Wavefront OBJ 形式で定義する．テクスチャは 1 毎の Targa 形式の画像で定義される．アイテムの位置は仮想空間のグリッド上に限られる．回転は 90° 単位である．アイテムの移動・回転は，イベントとして仮想空間に常に記録されている．

アイテムは Wavefront OBJ 形式で定義されており，頂点の座標とポリゴンの頂点構成とテクスチャ座標により定義される．アイテムは OpenGL API によりテクスチャ付きのポリゴンとして描画する．

アイテムには固有の名前が与えられており，内部ではその名前で扱われる．また，色・種類といった属性を持つ．更に，そのアイテムが対象固有軸を持つかどうかも与えられる．

3.5.13　描画効果

3.5.13.1　鏡面処理

アイテムやエージェントの位置を視覚的にわかりやすくするために，仮想空間の床に鏡面処理を加える．鏡面処理をしない画像出力を図 3.17 に，鏡面処理をした画像出力を図 3.18 に示す．

図 3.17: 鏡面処理なし

3.5.13.2　影の描画

鏡面処理と同様に，アイテムやエージェントの位置を知覚しやすくするために，エージェントやアイテムの影を床に描画する．影を描画しない画像出力を図 3.19 に，影を描画した画像出力を図 3.20 に示す．

図 3.18: 鏡面処理あり

図 3.19: 影なし

3.6　聴取実験

3.6.1　概要

上記の提案手法を用いて実装したエージェント音声対話システムにおける対話例を以下に示す．U はユーザ発話，S はシステム応答を表す．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．

U1　電話をパソコンの前に置いて

S1　**電話**はいくつかあるのですが，どれのことでしょうか

U2　黒い電話です

S2　電話の所に移動できないのですが，どうすればいいでしょうか

図 3.20: 影あり

U3　花瓶を持って

S3　手が空いていないのですが，どうすればいいでしょうか

U4　テレビの前に置いて

S4　**花瓶**を置いていいですか

U5　はい

上記対話例のうち，システムの応答音声について，被験者 15 人に対して聴取実験を行なった．実験は,「聞きやすさ」（何を言っているのかはっきり聞き取れるかどうか）,「自然さ」（イントネーションやアクセントが自然かどうか）の 2 点について，実際の対話の流れを追いながら，システム発話の箇所で従来システムによる応答音声と本提案手法による応答音声との比較を行なってもらった．具体的には，従来のシステムにおける応答音声を 3 とした時に 5 段階で評価してもらい，最後にコメントをしてもらう，という形で行なった．

聴取実験結果を表 3.9 に示す．表 3.9 のうち，Sn（n：1〜4）となっているものは，それぞれ前節の対話例と対応している．

表 3.9: 聴取実験結果

評価	聞きやすさ				自然さ			
	S1	S2	S3	S4	S1	S2	S3	S4
1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0
3	7	9	9	14	3	3	7	15
4	7	6	6	1	11	11	5	0
5	1	0	0	0	1	1	3	0

3.6.2　考察

S4 について，評価がほとんど 3 となっているのは，本研究における提案手法によって得られる音素記号列+韻律制御記号列（つまり応答音声）が全く変化しなかったことによる．

以降は S1〜S3 についての考察を行なう．

「聞きやすさ」について，評価 3 をつけた被験者が多かった．これは，従来システムにおいても「何を言っているか」はわかったため，本研究における応答音声との差が感じられなかった，という原因が考えられる．そのため，今回の実験において,「聞きやすさ」の項目は適切でなかった，と言えるかもしれない．

「自然さ」については，多くの被験者が 4 以上（つまり従来よりも自然さが増した）という評価を行なった．そのため，本研究における「より自然な対話音声の合成」という目的のために，提案手法が有効である，と言える結果が得られた．

また，被験者にはコメントも求めたが，一番多かったのは「語尾」についてのものだった.「語尾」について言及した被験者のほとんどが関西出身であった．今回の実験では被験者は特に制限を設けなかったが，今後このような実験を行なう場合，東京方言話者等の制限を設ける，あるいは出身を明記してもらい，出身ごとに傾向を分析する，という方法が適切であると思われる．

3.7　まとめ

本章では，修士課程時代に構築したエージェント音声対話システムについて，その概要を述べると共に，本研究で特に主眼を置いている応答生成手法について，言語情報の取り扱い手法及び韻律制御手法についてそれぞれ提案手法の説明を述べた．

言語情報の取り扱い手法については，言語情報を一貫して構文木構造のまま扱い，概念音声合成と親和性の高い応答文生成手法を提案した．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．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．

韻律制御手法については，音声合成における韻律規則（基本周波数パターン生成過程モデルを基にしたもの）やアクセント結合規則を適用する手法を提案した．これにより，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．

また，実際のエージェント対話システム構築にあたっての各処理（応答音声の生成，仮想空間管理，CG 生成）について述べた後，本提案手法の有効性を示すために聴取実験を行なった．その聴取実験の結果，本提案手法の妥当性を示すことができた．

聴取実験結果について示された「提案手法の妥当性」はあくまで「従来手法との比較」によるものであり,「韻律の面から見た音声の品質」という点ではまだ様々な問題点を残している．また，言語情報の取り扱い手法についてもまだまだ改良の余地を残している．

次章では，道案内音声対話システムを構築し，その中でこれらの手法をさらに改良することを試みる．

第4章

道案内音声対話システム

4.1　はじめに

第 3 章において，概念音声合成の実現にとって重要な課題である，言語情報の取扱い手法及びそれと親和性の高い韻律制御手法についての提案を行なった．

これらの提案手法について考察を行なった結果，言語情報の取扱い手法については，柔軟な応答文生成という観点から考えた場合に，必ずしもその要求に応えるものではなかった．また，韻律制御手法についても，アクセントやイントネーションといった部分において，人間の話す声とはまだ大きな差がある．

これらの課題に対するより詳細な検討を行なうためには，より豊富な種類の応答生成が要求され，またタスクの拡張も容易であると考えられるようなタスクの音声対話システムを構築するのが良いとの判断のもと，新たに道案内をタスクとするシステムを構築するに至った．この道案内音声対話システムの下で，第 3 章で提案した手法の検討，改良を試みた．

本章の構成は以下のようになっている．まず，道案内音声対話システムの概要について述べる．続いて，言語情報の取扱い手法，韻律制御手法の双方について，問題点を述べるとともに，それを基に改良した手法について述べる．これらの改良を行なった後，提案手法における優位性を検証するために，言語情報の取扱い手法，韻律制御手法の双方について行なった聴取実験の結果について述べ，最後に本章をまとめる．

4.2　システム概要

本システムのタスクは，仮想地図を用意し，システムがユーザに対して指示を行なうことで，ある地点から目標地点まで移動する，というものである．タスクとして道案内を選んだ理由は，システム応答に多様な応答生成が求められるため，本論文における提案手法の有効性の検証に適切であると考えたからである．

図 4.1 は本システムにおける仮想地図である．仮想地図において提示される情報としては，目標物の場所・名前，経路の距離があり，システム側はこれらの情報を全て保持している．図 4.2 は，実際にユーザに提示されるインタフェースであり，円はユーザの視界に相当する．つまり，ユーザは現在地にある目標物と道が出ている方向しかわからない．システムはあらかじめ目標地点までの最短距離を算出しておき，その経路をユーザに指示することでタスクの完了を目指す（4.3.3.1 節参照）．

図 4.2 における「工事中」等の一時的な情報はシステムは保持していない．このようなシステムとユーザの間で保持する情報の異なりにより，両者間で誤解が生じうる．そのような状況においてユーザが正解経路から外れたことが判明した場合，システムはユーザと協調的な対話を行なって現在地を推定し，その地点からの最短経路を再計算し，道案内を再開する（4.3.3.3 節参照）．

4.2.1　音声認識部

音声認識部には，Julian v.3.2[46] を用いる．音声対話システムでは，ユーザの発話もタスクに限定されたものとなるため，ネットワーク文法によってモデルを構築できる音声認

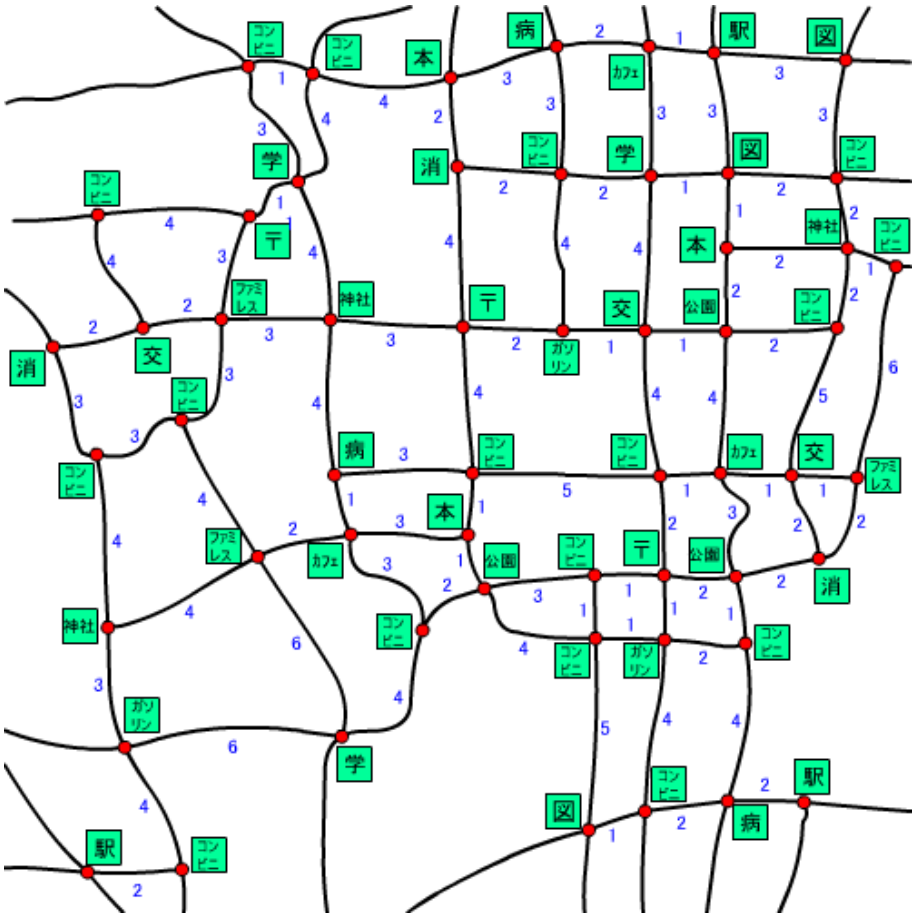
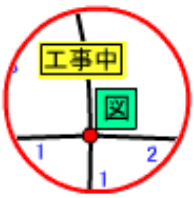


図 4.1: 道案内音声対話システムにおける仮想地図



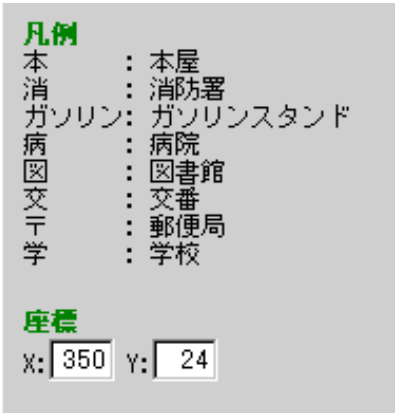


図 4.2: ユーザの視界と凡例

識器の方が，より認識誤りを起こしにくいと考えられる．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．

4.2.2　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．本システムにおいては，音声入力は研究の対象とはしておらず，ユーザの入力は非常に限られたものとなっている．そのため，構文解析部での実際の作業は，あらかじめ用意された入力文用テンプレートのどれに入力文がマッチするかを探索し，それを対話管理部に渡している．

4.2.3　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，現在のユーザの位置を確認したり，道を外れる等の問題を認識したりして，それらの内容に対して応答文を生成することで対話を進める．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．

4.2.3.1　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（向き，場所等）についてもスタックを持つ．

4.2.3.2　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．詳細については，3.5.1 節にて既に述べている内容と同様で，実際に用意される語彙が異なるのみである．

4.2.4　音声合成部

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．概要は 3.4 節で述べたものと同様であるが，本章にて行なった改良の詳細については 4.5 節で述べる．

4.3　実装

4.3.1　GUI インタフェース

道案内音声対話システムの開発当初，インタフェースとしては，図 4.1 上に図 4.2 のように赤枠の窓をかぶせることを，研究室のサーバ上で CGI を動かし，web ブラウザ上で動作するアプリケーションとして実装していた．このアプリケーションは，道案内音声対話システムとは完全に独立しており，単独でも動かせるものとなっていた．これは，ユーザが赤枠の窓を動かすことを対話制御部が感知していた場合，カーナビにおける GPS によ

る位置把握と同じような状況を想定したものとなる．しかしながら，本対話システムでは，より多様な応答生成が必要となるような対話状況を作り出したいという要求から，後述の「システムとユーザの間の誤解」のような状況を作り出すことも念頭に置いたシステム作りを行なっている．そのため，アプリケーションと対話管理部は完全に独立したものとしている．

しかしながら，音声対話システムとしての使い勝手や，音声入力のみでは雑音環境下での音声認識精度の悪化により対話が破綻することが頻繁に起こるという実体験や，スタンドアローンで動かしたいという観点から，図 4.3 に示すような GUI インタフェースを，Perl/TK を用いて構築した．

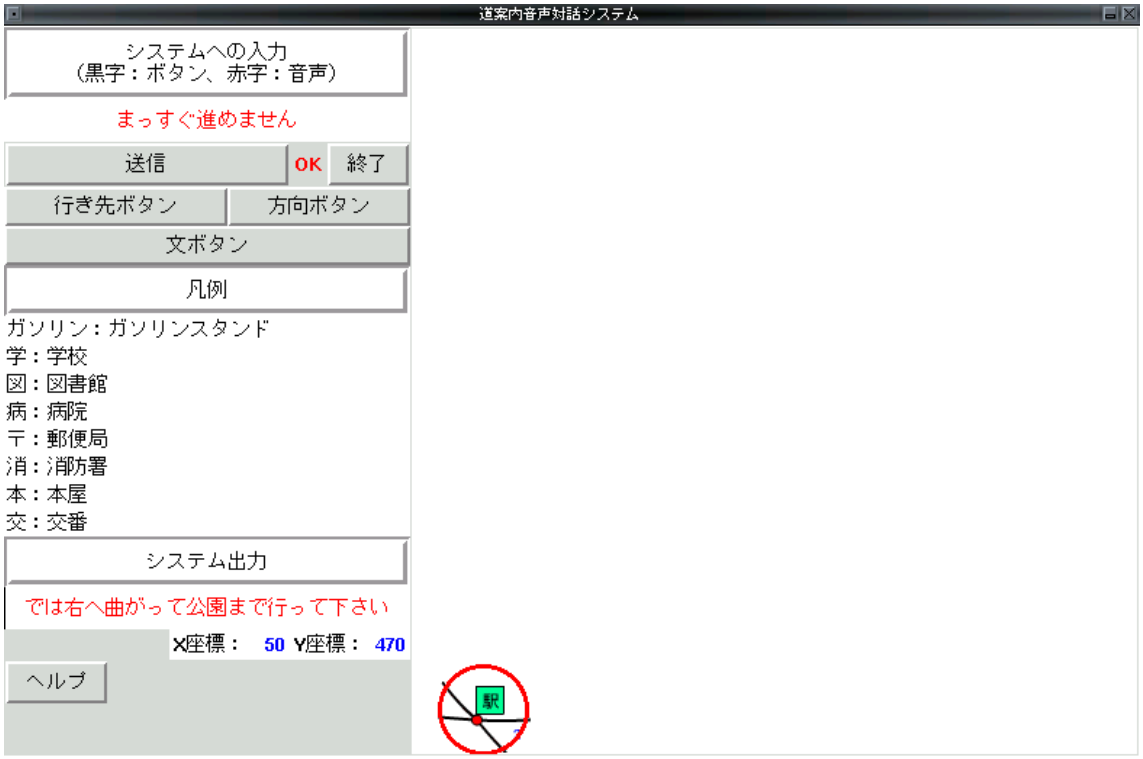


図 4.3: GUI インタフェース

この GUI インタフェースは，音声入力・音声出力に関しては道案内音声対話システムと連動して動作する．音声入力に関しては，雑音環境下でも動かしたいという観点から，マイク入力のみならずボタン選択（図 4.3 中の左上の各種ボタン）を用いて入力できるようにした．4.2.2 節でも述べた通り，システムへの入力に関しては，入力文の種類が限られたものであることから，このような入力方法も可能となるようにした．また，音声出力に関しては，システムの音声出力をテキストとしても表示できるようにした．ただし，地図および赤枠の窓（図 4.3 の右側）に関しては，以前と同様に対話管理部とは完全に独立したものとなっている．

本 GUI インタフェースは，大きく左右 2 つの画面に分けられる．右側については，上述の通り，地図および赤枠の窓に関する表示部である．ユーザは，この赤枠の窓を動かすこ

とによって，仮想地図の中を移動する，という動作を行なう．以前とは異なり，スタンドアローンでも動かせる仕様になっている．

続いて左側について，上から順に述べる．一番上の背景が白い部分は,「システムへの入力」に関するフィールドである．これは，ユーザからの入力をテキスト表示する場所であり，ユーザからの入力がボタン選択による時は，テキストは黒字で表される．マイク入力によって入力された場合は，テキストは赤字で表される．

続く背景が灰色の部分は,「ボタン選択による入力」に関するフィールドである．まず「行き先ボタン」「方向ボタン」を用いて，システムに入力したい情報を選択する．続いて「文ボタン」を選択すると,「(行き先）に着きました」等の定型文が提示され，その中から 1 つを選択すると,「行き先ボタン」or「方向ボタン」を用いて選択した「行き先」もしくは「方向」が挿入された状態で「システムへの入力」フィールドに表示される．図 4.3 中の「OK」と表示されている場所は,「文ボタン」によって決定された文章がシステムの受けつけられる文章であった場合に表示され，それ以外の場合には常に「NG」と表示されている．表示が「OK」の場合のみ,「送信」ボタンを押すことによってシステムに入力を与えることができる．また,「終了」ボタンは，GUI インタフェース・道案内音声対話システム双方を終了させるボタンである．

その下の背景が白い部分は,「凡例」が図 4.3 右側の地図の凡例を表しており,「システム出力」がシステムの音声出力をテキストで表示させる場所となっている．また,「x 座標・y 座標」は，地図中におけるユーザ位置（赤い窓枠の中心）の座標を表している.「ヘルプ」ボタンは，GUI インタフェースの使用方法を表示する．

4.3.2　対話用データ

本章で述べる道案内音声対話システムでは，対話用データとして主に場所・方向という 2 項目を扱う．

4.3.2.1　場所

場所とは，仮想地図内に存在する「コンビニ」「図書館」などのランドマークの総称である．仮想地図内には全部で 56 のランドマークがあり，その全てが以下の情報を持つ．

- 場所 ID．0〜55 までの値を取る．
- その場所にある「コンビニ」等のランドマークの名称．全 13 種類[1]．
- 東西南北の各方向に繋がっている道の長さ．実際の値は距離によって 1〜6，どこにも繋がっていない場合は，4.3.3.1 節における道順決定を想定して 9999 としている．
- 東西南北の各方向に繋がっている場所の ID．

[1]後述の 4.7 節において全 27 種類に増加.

4.3.2.2　方向

方向とは，現在のユーザの向きから見てどの方向かを表すもので,「右」「左」「まっすぐ」「後ろに」等が挙げられる．道案内の際には，4.3.3.1 節の手順で道順を決定するが，この道順決定の際には「東西南北」の情報で処理を行なっている．しかしながら，ユーザが「どちらが北か」といった情報は，その土地に不慣れであった場合にはわからないと見込まれることから，現在のユーザの向きからどちらの方向に進むべきかに変換してユーザに指示を行なう．

4.3.3　対話処理

4.3.3.1　道順の決定

目標地点が定まれば，次にどのような道順で案内するかを考える必要がある．方針としては

- 最短経路を通る
- 曲がる回数を少なくする
- 直線距離をなるべく長くする

といった辺りが考えられるが，本システムでは「最短経路を通る」という方針で道順を決定する．

最短経路決定問題としては，非常に良く知られているダイクストラアルゴリズムを用いる．

ダイクストラアルゴリズム

ダイクストラアルゴリズムは，以下のような状況において，あるノードからの最短経路を求める方法である．

ノード $1, 2, \ldots, n$ があり，ノード i からノード j に直行する経路の長さ w_{ij} が各 i, j について与えられているとする．ただし，直行する道がなければ，$w_{ij} = \infty$ とする．

この時，ノード 1 から各ノードまでの最短経路の初期値を $d_1 = 0, d_2 = d_3 = \cdots = d_n = \infty$ とする．$i = 1$ から始め，また訪れていない各ノード j について，距離を $d_j \leftarrow \min(d_i + w_{ij}, d_j)$ と更新する．この距離が最小のノード i を訪れ，上記のことを繰り返す．距離最小のノードを訪れる際，直前のノードを記憶しておけば，それを逆にたどることで最短経路を求めることができる．

以上の手順により，ノード 1 から他の全てのノードへの最短距離・経路が求められる．道案内音声対話システムでは，後述の「誤解への対処」のような何らかの事情により本来の道順を外れた場合でも破綻なく道案内を続行できるように，目標地点が定まった後に目標地点から全てのノードへの最短経路を求める．これは，ダイクストラアルゴリズムを逆向きに適用することで実現できる．また，直行する道がない場合に用いる ∞ は，システム内では「距離 9999 の道がある」と設定することにより，事実上その道を通る経路が得られないようにしている．

なお，この時点（詳細は 4.7 節参照）では，出発地点は図 4.1 中の左下の「駅」，目標地点は図 4.1 中の右上の「図書館」で固定されている．

4.3.3.2　道案内

出発地点，目標地点，案内経路が決定したら，システムは，4.3.3.1 節で決定した案内経路を基にユーザに対して道案内を行なう．この時，指示としては,「右へ曲がって公園まで行ってください」のように，基本的に「次に曲がる場所」までユーザに進んでもらうような指示を行なう．ただし，オプションとして「右へ曲がってください．そして公園まで行ってください」のように，より細かい単位で指示を出すこともできるようになっている．

このような指示を行ない，ユーザからの返答を受け取る．ユーザからの入力が「公園に着きました」のようにシステムの指示通りであった場合は，続けて次の指示を行なう．この時,「着きました」のように「公園」等のランドマークが省略されていたり,「そこに着きました」のように指示代名詞を用いた入力があった場合には，ユーザがシステムの指示通り動いているとみなして対話を続行する．

以上の対話を繰り返すことで目標地点までの案内を行なう．

4.3.3.3　誤解への対処

道案内の対話を行なう上で，ユーザがシステムの意図通りに動かない可能性は無視できない．このような，ユーザが道を「間違える」要因としては，

- システムの指示を聞き間違える
- 「工事中」等の要因で，システムの指示通り進むことができない
- システムの指示通りに進んだが，止まるべき場所が異なる

等が挙げられる．このような場合，システムはユーザとの対話によって，ユーザがシステムの意図とは異なる道を進んでいることを感知する．

まず,「何らかの要因でシステムの指示通り進めない」場合は，ユーザに対して「道はありますか」という質問を行なう．その質問に対して「工事中」という返事が返ってきた場合は，その時点まではユーザはシステムの意図通り進んでいるとみなす（「工事中」という一時的な情報は，システムは事前には知らない）．この時，システムにはユーザの位置が把握できているため,「工事中」となっている道を「道がない」ものとして，再度ダイクストラアルゴリズムによって最短経路を計算し，その結果に従って道案内を再開する．

上記の場合において「道がない」という返答が返ってきた場合，およびユーザがシステムの意図とは異なる場所にいた場合は，ユーザが実際はどこにいるのかを対話によって推測する必要がある．その手順を以下に示す．

1. 直前のユーザ位置の周辺を探索
2. システムが想定していた現在地の周辺を探索
3. どこを通過したかをユーザに尋ね、直前位置・通過地点・ユーザ現在地が一直線に並ぶ箇所を網羅的に探索

4.3.4　対話例

以下に道案内音声対話システムにおける対話例の一部を示す．S がシステム，U がユーザの発話を表し，数字は対話例中においてそれぞれの何番目の発話かを表す．システム発話のうち 下線 を引いてある単語（方向・場所を表す単語）は重要であるとみなし，**太字**で表されている単語（対話中において初出の場所を表す語[2]）は新しいとみなす語である．

S1: 東 へ行ってください

U1: コンビニに着きました

S2: 左 へ曲がって **神社** まで行ってください

U2: ガソリンスタンドに着きました

S3: 神社 はまだ先ですね

S4: まっすぐ 神社 まで行ってください

U3: まっすぐ進めません

S5: 道はありますか

U4: 工事中です

S6: では 右 に曲がって **公園** まで行ってください

U5: 公園に着きました

S7: 左 へ曲がって **郵便局** まで行ってください

U6: 消防署に着きました

S8: 郵便局 を通り過ぎたようですね

S9: 右 へ曲がって二つ目の **コンビニ** まで行ってください

U7: 学校に着きました

S10: どこを通りましたか

U8: コンビニです

S11: 二つ目の コンビニ はまだ先ですね

S12: まっすぐ二つ目の コンビニ まで行ってください

U9: 着きました

S13: 左 へ曲がって **図書館** まで行ってください

U10: 図書館に着きました

S14: そこが目標地点です

[2]表層テキストでは同じでも，内部的に別のものを表していれば初出とみなす．

4.4　応答文生成手法の改良

4.4.1　従来手法の問題点

3.3.1 節における言語情報の取扱い手法では，$ITEM タグや$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，という問題があった．例えば図 3.4 は,「$ITEM を$POS に置く」という比較的短い文章であるが，これに「〜にある$ITEM を$POS に置く」のように修飾語がついたり，また語順の入れ換えや文末表現が変わるといった（同じ意味の）多様な応答生成を行なう際，そのつど異なる定型文を用意する必要がある．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．

これは，今後のシステム拡張，もしくは新たな対話システムの構築を考えた場合，新たに必要となるすべての応答文に対し，再度行なわれなければならないことを意味する．

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．

4.4.2　フレーズ単位での応答文生成

まず，タグに単語しか挿入できない，そのため修飾語がつくといった微小な変化に対しても定型文を用意する必要があるという問題を解決するため，タグに単語だけでなく連文節の挿入も可能となるようにした．これにより，例えば$DIR に修飾語がつく（「〜を$DIR へ」等）ような場合にも，新たな定型文を用意する必要がなくなる．連文節をタグに挿入する際，挿入する連文節にタグが含まれていた場合，そのタグが持つ重要度や新規性はそのまま保持される．

また，語順の入れ換えや文末表現が変わるといった応答文の変化に対しても，それぞれ個別に定型文を用意する必要があった．この問題点の解決のため，テンプレートの単位を「文」から「フレーズ」とした．これにより，応答文生成は定型フレーズの組み合わせによって実現されることとなり，微小な変化に対しても，それに相当する定型フレーズさえ用意すれば，他の定型フレーズがどのように変化しようとも用意した定型フレーズ 1 つで済む．

以下，例を挙げながら実際の応答文生成の流れを述べる．

4.4.2.1　連文節

まず，連文節単位での定型フレーズを用意する．この定型フレーズには以下の 5 種類がある．

- 名詞句生成フレーズ
  例：(に（$DIR))

- 動詞句生成フレーズ
  例：(て（$VERB（$NOUN_PHR））)
- 重文フレーズ
  例：($NOUN_PHR1 $NOUN_PHR2)
- 複文フレーズ
  例：($NOUN_PHR1（$NOUN_PHR2))
- 文頭・文末フレーズ
  例：(ください（$VERB_PHR))

それぞれのタグには，単語もしくは連文節が挿入される．

応答文生成の際には，これらの定型フレーズを用いて，連文節単位で応答文の構成要素を生成する．単文節を生成する際には，タグに単語が挿入されることになるが，この時点で単語の重要度や新規性といった談話情報を付与し，以降の応答文生成過程においても常に保持する．

4.4.2.2　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．これは，文の構造を決定する重文フレーズ・複文フレーズ等の定型フレーズの該当タグに複数の連文節を挿入することで実現される．この際，挿入する連文節中の単語に談話情報が設定されていた場合，それらの情報も引き継いで保持する．

4.4.2.3　種々多様な応答文生成

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（$DIRECTION))」(名詞句生成フレーズ）から名詞句「右に」を生成
2. 「(まで（$LANDMARK))」(名詞句生成フレーズ）から名詞句「駅まで」を生成
3. 「(て（$VERB（$N_PHRASE)))」(動詞句生成フレーズ）から動詞句「右に曲がって」・「駅まで行って」を生成
4. 2 つの動詞句を連結して「右に曲がって駅まで行って」を生成
5. 「(ください（$VERB_PHRASE))」(文末フレーズ）から「右に曲がって駅まで行ってください」を生成

こ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．この例では，この際に「そして」等の接続詞を適宜挿入するが，これは「そして」に対応する定型フレーズを導入するだけで実現できる．

また逆に，4.3.4 節における S11 と S12 を繋いで「二つ目のコンビニはまだ先ですので，まっすぐそこまで行ってください」というように，2 つの文を繋げて 1 つの応答文とする場合でも，同様に接続詞に対応する定型フレーズの導入のみで実現できる．このような場

合，この例では「コンビニ」の代わりに「そこまで」という照応表現を挿入しているように，対話の場面によって適宜省略・照応表現を挿入することも可能である．

また,「まっすぐそこまで行ってください」と「そこまでまっすぐ行ってください」のように，語順を入れ替えることも，新たなテンプレートを用意することなく実現できる．

4.4.3　評価

本システムにおける典型的な対話例（4.3.4 節）において，従来手法（定型文を用いる手法）と提案手法（定型フレーズを用いる手法）によるテンプレート数はそれぞれ 16，16 となる。しかしながら，これらの数字の意味するところは大きく異なる。前者におけるテンプレートとは定型文を表し，16 は全て完全な文の数であるのに対し，後者におけるテンプレートとは定型フレーズであり，16 は名詞句生成フレーズ，動詞句生成フレーズ等文の各構成要素のテンプレート数の合計である．典型的な対話例における各定型フレーズ数を以下に示す．

- 名詞句生成フレーズ：7
- 動詞句生成フレーズ：3
- 重文フレーズ：1
- 複文フレーズ：1
- 文頭・文末フレーズ：4

ここで，名詞句生成フレーズ等が複数存在するが，これらの数は助詞や接続詞等付属語の種類数にほぼ対応する．提案手法では付属語もタグを用いて表現することが可能である．しかし，タグを用いる最も大きな要因である「重要度・新規性の付与」を考えたとき，付属語は重要度・新規性を付与する対象とはなりえないため，今回では付属語はタグを用いず，それらに対応する定型フレーズを用意した．

次に，4.4.2.3 節のような種々多様な応答生成を考えた場合，従来手法だと定型文数が大きく増えるが，提案手法では，文頭・文末フレーズとして接続詞に対応するフレーズを追加するのみでよい．具体的に，4.4.2.3 節のような多様な応答生成を行なうことを考えれば，従来手法だと 22 定型文（6 増加）になるのに対し，提案手法だと 18 定型フレーズ（2 増加）で済む．このように，提案手法では応答文が多様になればなるほど必要な定型フレーム数の増加の割合を減らすことができ，従来手法よりも効率的な応答文の生成が可能となる。

4.4.3.1　様々な応答生成を考慮した評価

さらに様々な応答生成を行なうことを想定し，提案手法の有効性についてより詳細に検証するため，被験者 12 人を対象として対話例の収集を行なった．収集方法は，被験者がシステム，我々システム開発側がユーザ役を担当して，以下の条件以外は構築中のシステムと同様の条件で実際に道案内を行なってもらった．

- 被験者（システム役）に経路距離情報を与えない（明らかな遠回り等でない限りどのような経路を選択してもよい）

- 指示のスタイルについて制限を加えない（ユーザにとってわかりやすいと思うスタイルで指示してもらう

また，各被験者に対して，以下の 2 通りの状況で対話を行なってもらった．

実験 1　出発地点から目標地点までを通して対話を行なう

実験 2　実験 1 の途中でユーザが道に迷った状況から開始し，ユーザがどこにいるかを特定し，目的地まで導く

この結果得られた被験者（システム役）の総発話数 232 に対し，典型的な対話例の実現に必要であったテンプレートを可能な限り利用して応答生成を行なうように，文意を変えずに修正を行なった．具体的には，複数の同義語を一つに統一したり，語順を統一したりという修正である．

修正した応答発話の生成に必要なテンプレート数の増加を調べた結果，従来手法では計 47 定型文（25 増加）であるのに対し，提案手法では計 28 定型フレーズ（10 増加）となった．提案手法における増加した 10 定型フレーズの内訳は，名詞句生成フレーズ 1，動詞句生成フレーズ 1，文頭・文末フレーズ 8 である．今回の検証では語順を統一してあるが，これも考慮に入れた場合，従来手法で必要な定型文はさらに増加する（提案手法では増加しない）．

このように，提案手法では応答文が多様になればなるほど必要な定型フレーズ数の増加の割合を減らすことができ，従来手法よりも効率的な応答文の生成が可能となる．提案手法において追加すべき定型フレーズは，全てが新たに出現した付属語や接続詞等に対応するものであり，第 4.4.3 節で既に述べた通り，付属語や接続詞等にタグを付与しない方針を取っているため，不可避的なものかつ必要最低限のものであると言える．

また，定型フレーズを用いる今回の提案手法は，定型文を用意する従来手法とは異なり，どのような応答文の構文構造にも対応できるため，システムのタスクに依存しない汎用的な手法である．また，タグの識別子（$DIRECTION のような$で始まる文字列）に関しても，実装の際に分かりやすく名前を付けているだけのものであり，実際には任意の文字列を設定できるため，この点においてもシステムのタスクに依存しない手法である．これらのことから，同一システムを改良してより複雑な応答生成を行なうのみならず，新たなシステムを開発する際にも提案手法が有効であると言える．

韻律制御に関しては，定型文による従来手法と定型フレーズによる提案手法で全く同一の結果となる．

4.5　韻律制御手法の改良

3.4.2 節で述べた韻律制御手法によって，応答音声において，ある程度の自然性が実現できた．しかしながら，アクセントやイントネーション等，人間の話す声とはまだ大きな差がある．

これらの差を埋めることを目指し，新たな韻律制御規則を導入した．

4.5.1　文節間結合規則

アクセント結合は単語と単語の結合によって起こるものであるが，文節と文節の結合によっても，アクセント位置が変わる現象が確認されている [64]．そのため，この文節間結合規則も導入する．

本手法では，文節間結合規則のうちの 2 韻律句連鎖規則（表 4.1）を用いる．

表 4.1: 2 韻律句連鎖規則

D_1	X_2	$\longrightarrow$	D_1	x_2
F_1	D_2	$\longrightarrow$	$D_{1\ 2}$	
F_1	F_2	$\longrightarrow$	$F_{1\ 2}$	

ここで，D，F，X は，それぞれ起伏型, 平板型, 起伏 · 平板型のアクセント型を表す．左辺は，単独で発声した時のアクセント型を表し，右辺は，左辺の 2 韻律句を連続して発声した時に変化した結果のアクセント型を表す．なお，添字は何番目の韻律句であるかを示し，例えば $D_{1\ 2}$ は 1 番目と 2 番目の韻律句がアクセント結合して一つのアクセント句になることを示している．

この文節間結合規則は，具体的にはアクセント指令挿入位置の制御に用いられる．この結果，アクセント指令挿入位置の流れは以下のようになる．

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なわない．また，焦点の当てられている文節についても結合は行なわない．

4.5.2　音素・韻律記号列生成手法

上記の規則を導入した結果，韻律制御記号を含む音素記号列を生成する手法は，3.4.2.4 節でのものとは順序が異なっている．結果として，手順は以下のようになる．

1. 構文木の接続に従って単語の活用形を決定
2. 構文木構造やモーラ数に従って，フレーズ指令挿入位置を決定
3. モーラ数や単語の重要度・新規性に従って，フレーズ指令のパラメータを決定
4. 品詞や単語の重要度・新規性，フレーズ指令挿入位置に従って，アクセント指令のパラメータを決定
5. 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

4.5.3　評価

今回，韻律制御手法として新たに制御手法（文節間結合規則）を取り入れたことの有効性を調べるために，聴取実験を行なった．具体的なテキストの内容は，それぞれ以下に示

す通りである.

1. 左へ曲がって図書館まで行ってください
2. 東へ行ってください
3. 左へ曲がって神社まで行ってください
4. 道はありますか
5. まっすぐ神社まで行ってください
6. 公園を通りすぎたようですね
7. 左へ曲がって郵便局まで行ってください
8. 右へ曲がって公園まで行ってください

これらの 8 文に対し，従来手法（文節間結合規則を用いない手法）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．その結果を表 4.2 に示す.

表 4.2: 聴取実験結果

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

「改良手法が良い」ことの有意性を調べるために,「すべてランダムに評価された」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（「道はありますか?」）に関しては，文の構文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.

4.6　応答音声に関する聴取実験

4.6.1　概要

LISP 形式による統語構造の保持，タグを用いた談話情報の韻律への反映について，その有効性を確かめるために，聴取実験を行なった.

実験方法として，3 種類の音声を合成する．3 種類の合成方法を以下に示す.

提案手法： 提案手法の応答文生成手法，韻律制御手法により合成した音声

JUMAN+KNP 解析： JUMAN[65]+KNP[66] による統語構造解析結果を基に合成した音声

談話情報なし： 重要度・新規性を全てないものとして合成した音声

「JUMAN+KNP 解析」はテキスト音声合成の枠組を想定したものである．テキスト音声合成では，最初にテキストを用意してから，それを基に JUMAN+KNP 等の解析器によって構文解析を行なって韻律情報を付与する．「JUMAN+KNP 解析」は，このようなテキスト音声合成を想定したものであるが，この場合に得られる構文構造に誤りが含まれ，その結果合成音声の韻律に影響を与えることが考えられる．また，「談話情報なし」の合成音声については，タグによる談話情報の取扱いが適切に行なわれているかどうかを検証するために作成した．

実験方法としては，典型的な対話例（4.3.4 節）において現れるシステム応答のうち 8 文を上記の 3 種類の方法で合成し，被験者（24 名）に 8×3 個の音声に対して 5 段階評価（1：悪い〜5：良い）で評定してもらった．具体的なテキストの内容は，それぞれ以下に示す通りである（4.5.3 節による聴取実験と同じ）．

1. 左へ曲がって図書館まで行ってください
2. 東へ行ってください
3. 左へ曲がって神社まで行ってください
4. 道はありますか
5. まっすぐ神社まで行ってください
6. 公園を通りすぎたようですね
7. 左へ曲がって郵便局まで行ってください
8. 右へ曲がって公園まで行ってください

音声の提示方法については，各文ごとに 3 種類の合成音声をランダムな順序とした．被験者はどの合成方法によるものかの事前知識なしに聴取した．評価基準としては，8 文それぞれに強調すべき箇所を示し，適切に強調されているかどうか，また不適切な箇所で抑揚がついていないかどうか，に主に着目してもらうこととした．また，評価の際には，8 文それぞれにおいて 3 種類の音声を聞き比べてもらい（何度聞いてもよいこととした），3 種類の音声間に明らかに差異が認められる場合には，その差異を評価に反映させるよう指示した．

4.6.2　実験結果・考察

表 4.3 に，各 8 文における 3 種類の合成音声それぞれの評価点の平均点を示す．統計的有意性を検証するために，t 検定を行なった．t 検定にあたり，帰無仮説を「提案手法と（JUMAN+KNP 解析/談話情報なし）の平均が同じである」と設定し，有意水準 5%の片側検定を行なった．

表 4.3: 8 文 ×3 種類の合成音声の平均点

文 No.	1	2	3	4
提案手法	3.58	3.42	3.46	2.42
JUMAN+KNP 解析	3.04	3.25	2.92	2.38
談話情報なし	2.50	3.13	2.38	2.33
文 No.	5	6	7	8
提案手法	2.79	3.83	3.88	4.04
JUMAN+KNP 解析	2.17	3.17	3.25	3.29
談話情報なし	2.21	3.54	2.54	3.38

8 文のうち，文 No.2 に関しては，提案手法と「JUMAN+KNP 解析」が同じもの，文 No.4 に関しては，3 種類の合成音声が全て同じものとなっている．これらは，統語構造や談話情報の関係で，結果的に生成された合成音声に差異がなかったことを表す．表 4.3 のうち，これらに該当する項目に関しては有意な差は現れなかった．

その他の項目については，文 No.2 における「提案手法」と「談話情報なし」の間，文 No.6 における「提案手法」と「談話情報なし」の間に有意な差は認められなかった（上側確率はそれぞれ 14.3%，11.1%）.「提案手法」と「談話情報なし」で生じる差は，ピッチ（特にアクセント指令）の差となって現れてくる．上記の 2 文では，音素+韻律制御記号列では差が現れても，実際に合成音声として比較すると，ほとんど差がわかりづらかったことが原因と考えられる．

上記以外の項目については，統計的に有意な差が現れており，本提案手法の有効性が確認できる結果となっている．以下，詳細な分析結果について述べる．

「提案手法」と「JUMAN+KNP 解析」で異なる合成音声が生成された文章（文 No.2，4 以外）では，いずれも JUMAN+KNP 解析による構文解析結果が誤ったものとなっている．例えば文 No.1 では「左へ」と「曲がって」の間に ICRLB 境界が現れるという構文解析結果が得られている（正解は「曲がって」と「図書館」の間）．これは，形態素解析・構文解析の時点では，各形態素の意味まで考慮されることがないため，このような構文解析結果を出力していると考えられる．この結果が韻律に反映され，有意な差となって現れている．

24 名の被験者のうち，3 名は 8 文の手法ごとの平均点において「提案手法」が最高値とならなかった．これは，本提案手法で用いている韻律制御手法 [59] が東京方言アクセントを基準としたものであるため，それ以外の方言の話者にとっては違和感のある合成音声となっていることが一因として考えられる．

文章間でのスコアのつけ方については，実験を行なう際に特に指示を設けなかったためか，被験者ごとのスコアのつけ方（8 文 ×3 種類の合成音声のうちどの音声に最も高いスコアをつけたか等）は個人差が大きく，被験者間に共通する特別な傾向は見受けられなかった．

全体を通してみると，長い文章ほど手法間での得点差が大きいという傾向が見られた．これは，長い文章ほど手法間での差が現れやすく，その結果被験者がはっきりと差をつけ

やすかったためだと考えられる．

4.7　タスク拡張

4.7.1　概要

4.6.2 節における聴取実験では，テキストとして 4.3.4 節の対話例にあるようなものを用意した．しかしながら，この 8 文を見てみると，似通ったテキストが多くなってしまっていた．

そこで，より多様な応答生成をさせることを目的とし，タスクの拡張を行なった．

まず，これまでは固定されていた「目標地点」を，対話によって設定できるようにした．それに伴い，道案内に先立って目標地点を決定させるための対話をできるようにした．目標地点の決定にあたって，ユーザはコンビニ等の「場所」を直接指示するだけでなく,「遊びに行きたい」「お腹が空いた」等の若干曖昧性を持たせた入力もできるようにした．このような場合，システムが条件に合う選択肢をいくつか提示し，その中からユーザが選ぶことによって目標地点を決定する．出発地点は，図 4.1 の左下の「駅」で固定したままである．

また，曖昧な入力によってシステムが選択肢を提示する場合，従来用意していたランドマークの種類（13 種類）では少ないことから，27 種類に増やした[3]．

この結果できるようになった対話の例を以下に示す．**S** で始まるものがシステム発話，**U** で始まるものがユーザ発話，数字は対話例中の通し番号を表す．この対話例中では，**S1**〜**U2** までが，目標地点を決定するための対話となる．

S1:　どこへ行きたいですか

U1:　遊びに行きたいです

S2:　ボーリング場，ビリヤード，カラオケ，雀荘，漫画喫茶があります

U2:　漫画喫茶に行きたいです

S3:　では，東へ行ってください

U3:　ホテルに着きました

S4:　左へ曲がってボーリング場まで行ってください

U4:　ボーリング場に着きました

S5:　右へ曲がって郵便局まで行ってください

U5:　ラーメン屋に着きました

S6:　郵便局はまだ先ですね

[3]従来の 13 種類においても，本研究では考慮していない音声認識等の観点から，名称が変更されたり削除されたりしたものもある．そのため，純粋に新しく増やされた「場所」の名称は，$27 - 13 = 14$ 種類より多い．

S7: まっすぐ郵便局まで行ってください

U6: スーパーに着きました

S8: 郵便局を通りすぎたようですね

S9: では，左へ曲がってカフェまで行ってください

U7: カフェに着きました

S10: 左へ曲がって漫画喫茶まで行ってください

U8: 漫画喫茶に着きました

4.7.2　聴取実験

4.7.1 節のタスク拡張を行なうことで，従来に比べてより多様な応答生成を行なうことができるようになった．そこで，4.6 節で行なった聴取実験結果についてより詳細な検討を行なうことを目的として，さらなる聴取実験を行なった．

4.7.2.1　実験概要

今回の聴取実験では，テキストとして 4.7.1 節における対話例のうちのシステム発話（**S** で始まるもの）を用いる．

4.6 節での聴取実験と同様，今回の聴取実験でも「提案手法」・「JUMAN+KNP 解析」・「談話情報なし」の 3 種類の合成音声を用意した．

実験方法として，今回の聴取実験では 4.7.1 節での対話例に沿って対話を行ない，システム発話が出現する度に，以下に示す評価尺度に従って評価を行なってもらった．

重要な語： システム発話のうち，どの語がユーザに伝えるべき重要な語であるか（複数選択可）

伝わりやすさ： 選んでもらった「重要な語」が，実際に「重要な語」に聞こえるか

自然さ： 文全体を通して，イントネーションやアクセントが自然かどうか

このうち,「自然さ」に関しては，4.7.1 節の聴取実験と同様の意図によるものである.「重要な語」に関しては，4.7.1 節の聴取実験の際に得られた被験者のコメントから,「重要度あり」と設定する語の決定方法について検討するために新たに評価尺度として採り入れた．また，現時点の手法により「重要度あり」とされている語が実際に「重要な語」として伝わりやすいかどうか，という評価のために「伝わりやすさ」という評価尺度を採り入れた．

実験手順としては，システム発話が出現する度に，まず音声を聞く前に「重要な語」をチェックしてもらった．チェックが終わった後に 3 種類の合成音声を聞き比べ，被験者（22 名）に 10 文 ×3 個の音声それぞれに対して 5 段階評価（1：悪い〜5：良い）で評定してもらった．音声の提示方法については，4.7.1 節の聴取実験と同様にランダムな順序で提示し，被験者にとってはどの音声がどの合成方法によるものかの事前知識は与えていない．

4.7.2.2　実験結果・考察

「重要な語」の実験結果・考察

「重要な語」については，各 10 文中のそれぞれの単語に対して,「重要である」とみなされた単語を 1 点，そうでない単語を 0 点として平均点を取る．これらの単語が「重要な語である」と有意にみなされているかどうかを検証するために,「0，1 でランダムに評価された」という帰無仮説を用意する．この時，有意水準 1%で帰無仮説を棄却できる平均点は 0.77 以上となる．この平均点 0.77 以上を満たす，有意に「重要な語である」とみなされた語は，以下の文中の下線部である．また，現時点でのシステムにおいて「重要な語である」とみなされる語を太字で示す.

S1: **どこへ** 行きたいですか

S2: **ボーリング場**，**ビリヤード**，**カラオケ**，**雀荘**，**漫画喫茶** があります

S3: では，**東へ** 行ってください

S4: **左へ** 曲がって **ボーリング場まで** 行ってください

S5: **右へ** 曲がって **郵便局まで** 行ってください

S6: **郵便局**はまだ先ですね

S7: まっすぐ **郵便局まで**行ってください

S8: **郵便局を** 通りすぎた ようですね

S9: では，**左へ** 曲がって **カフェまで** 行ってください

S10: **左へ** 曲がって **漫画喫茶まで** 行ってください

S6 では，どの語も「重要である」という結果が得られなかった.「まだ」が平均点 0.68,「先ですね」が 0.64 であったが，共に有意な結果とはならなかった.

また，S7「まっすぐ」，S8「通りすぎた」は，現時点のシステムでは「重要な語」であるとはみなしていなかった．これは，現時点では「方向・場所を表す単語」を「重要である」とみなしているためであるが，それ以外にも上記のような語については強調すべきであることがわかった.

「伝わりやすさ」・「自然さ」の実験結果

続いて，各 10 文における 3 種類の合成音声それぞれの評価点の平均点について，表 4.4 に「伝わりやすさ」の平均点を，表 4.5 に「自然さ」の平均点を示す.

統計的有意性を検証するために，t 検定を行なった．t 検定を行なうにあたり，帰無仮説を「提案手法と（JUMAN+KNP 解析/談話情報なし）の平均が同じである」と設定し，有意水準 5%の片側検定を行なった.

今回の聴取実験では，10 文のうち，提案手法と「JUMAN+KNP 解析」が（音素記号列+韻律制御記号列として）同じものとなったのが S1，S2，S3，S6，S8，S9，提案手法と「談話情報なし」が同じものとなったのが S1 である.

表 4.4: 「伝わりやすさ」の平均点

文 No.	1	2	3	4	5
提案手法	3.77	3.91	3.59	4.00	3.64
JUMAN+KNP 解析	3.95	3.82	3.50	3.23	3.18
談話情報なし	3.77	3.18	2.45	3.14	2.36
文 No.	6	7	8	9	10
提案手法	3.73	2.64	3.41	4.09	3.95
JUMAN+KNP 解析	3.55	2.73	3.23	4.09	3.32
談話情報なし	3.18	2.41	3.32	2.45	3.23

表 4.5: 「自然さ」の平均点

文 No.	1	2	3	4	5
提案手法	2.77	3.36	3.68	3.73	3.50
JUMAN+KNP 解析	2.77	3.50	3.55	3.05	2.73
談話情報なし	2.95	2.91	2.64	3.09	2.55
文 No.	6	7	8	9	10
提案手法	3.77	2.41	3.41	3.86	3.82
JUMAN+KNP 解析	3.50	2.59	3.45	3.95	3.09
談話情報なし	3.18	2.36	3.45	2.73	3.05

「伝わりやすさ」の考察

「伝わりやすさ」については，上記の「記号列として差がなかった音声間」では，有意な差が認められなかった．

この「伝わりやすさ」については，特に「重要度」がきちんと反映されているかどうかを検証するための項目であった．しかしながら，提案手法と「JUMAN+KNP 解析」とでは同じ「重要度」（及び「新規性」）を設定しているにも関わらず，S4，S5，S10 では全て有意な差が見られた．これは，これらの文章における構文解析結果の誤りが韻律に反映され，その結果「伝わりやすさ」にも影響したものと考えられる．また，S7，S8 においては，提案手法と「談話情報なし」との間でも有意な差が認められなかった．これは，システムの想定している「重要な語」と被験者が判断した「重要な語」の間に食い違いが生じていることが原因と考えられる．

今回の実験において,「伝わりやすさ」の項目は,「重要度あり」とみなす語の決定方法を検討するための項目であり,「重要な語」の項目で述べた「現時点のシステムにおいて『重要な語である』とみなす語」には,「新規性」のあるものとないもの両方が含まれている．しかしながら，詳細な数値に関しては付録 A.3.2 節で示してある通り,「重要度あり」とみ

なされる語については，アクセント指令の大きさは「新規性」の有無でほとんど変わらない．そのため，今回の実験については,「新規性」の有無に関する検討は行なっていない．

「重要度なし」の語に関しては，新規性の有無で指令の大きさは大きく異なる．そのため,「重要度」・「新規性」の有無による 4 種類の指令の大きさは,「重要度あり・新規性なし」≃「重要度あり・新規性あり」＞「重要度なし・新規性あり」＞「重要度なし・新規性なし」となる．ここで，現時点のシステムでは,「重要度なし・新規性あり」という設定がなされる語は存在しない．そのため，指令の大きさの序列は事実上「重要度あり」＞「重要度なし」となる．このような理由から,「重要度」に関しての項目を「重要な語」・「伝わりやすさ」という形で聴取実験に採り入れた．

「自然さ」の考察

「自然さ」については，上記の「記号列として差がなかった音声間」では，有意な差が認められなかった．

その他の音声間で今回の実験で有意な差が認められなかったのは，S7 の提案手法と両手法との間，及び S8 の提案手法と「談話情報なし」との間である．S7，S8 においては，システムと被験者との間で「重要な語」に食い違いが生じ,「伝わりやすさ」の評価において有意な差が得られなかったが,「自然さ」の評価においてもそれが影響されてしまっている可能性が考えられる．また，実際の合成音声として比較した場合にほとんど差がわかりづらかったことも原因として考えられる．特に S8 に関しては，4.6 節での聴取実験における S6「公園を通りすぎたようですね」と非常に良く似た文章であり，この形の文章では差が現れにくくなっていると考えることもできる．

今回の実験における S8「郵便局を通りすぎたようですね」では，提案手法と「JUMAN+KNP 解析」で同じ音声となったが，ほぼ同様の構造を持つ 4.6 節の聴取実験 S6「公園を通りすぎたようですね」では，両者では違う音声が生成されていた．これは,「郵便局/公園を通りすぎたようですね」という文章には ICRLB 境界が存在しないが,「JUMAN+KNP 解析」では「郵便局/公園を」の後に ICRLB 境界が存在する，という構文解析結果を出している．それに対し，提案手法では 3.4.2.1 節で述べたルールから,「公園」と「郵便局」のモーラ数の違いによって，フレーズ指令を前者は挿入せず，後者では挿入した音声が生成されている．その結果，今回の聴取実験での S8 では，音素記号+韻律制御記号列としては同一のものとなっているが,「郵便局を」の直後に置かれるフレーズ指令は異なった意味を持っている．現在の手法ではそのような違いまでは考慮していないが，フレーズ指令挿入規則構築にあたり，単語の意味的要因も考慮する必要性があることが示唆される結果となった．

全体を通してみると，改めて本提案手法の有効性が示されるとともに，新たな知見も得ることができた．

4.8　まとめ

本章では，エージェント音声対話システム構築の際に提案した応答生成手法のさらなる検証のために構築した道案内音声対話システムについて，その概要を述べると共に，言語

情報の取り扱い手法，韻律制御手法の両提案手法について検証を深めると共に，これらの提案手法の改良について述べた．

言語情報の取り扱い手法については，テンプレートの単位を文から文節とした，定型フレーズを用いる応答文生成手法を提案した．従来の定型文を用いる手法では，応答が複雑になればなるほど必要なテンプレート数が増大していったのが，定型フレーズを用いる手法を導入することによって，同じ状況を実現するために必要なテンプレート数の増加分を大幅に減らすことができた．この新提案手法によって，より少ないテンプレート数でより多様な応答生成を実現できると共に，タスクに依らない汎用的な応答文生成手法を構築することができたと考える．

韻律制御手法については，これまでの韻律制御規則に加えて文節間結合規則を導入したり，それに伴って韻律制御記号の設定手順を変更したり，という改良を行なった．エージェント音声対話システムまでの韻律制御手法による音声と新提案手法による音声との比較を聴取実験によって行なった結果，新たに導入した韻律制御規則の妥当性が確認された．

また，提案手法そのものの有効性を検証するために，2 種類の比較音声を用いた聴取実験を行なった．1 つ目の比較音声は，統語構造の取り扱いについて検証するために用意したもので，テキスト音声合成の際に必要な構文解析情報を基に合成した音声を想定し，形態素・構文解析ツールとして一般的な JUMAN+KNP による構文解析情報を基に合成した音声である．2 つ目の比較音声は,「重要度」や「新規性」といった談話情報を用いることの有効性の検証のために用意したもので，これらの談話情報を用いずに合成した音声である．この 3 種類の音声の比較による聴取実験を行なった結果，提案手法の有効性が示された．

また，この聴取実験の際に得られた知見を基に，さらなる検証を行なうために道案内音声対話システムのタスク拡張を行なった．そして同様の聴取実験を行ない，提案手法の有効性の再確認をすると共に新たな知見を得ることができた．

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音声出力方式として望ましいと考えられる概念音声合成の枠組を実現することを目的とし，そのための応答生成手法の構築を行なうとともに，音声対話システムに組み込むことで，既存のテキスト音声合成による合成音声に比べて，ユーザにとってシステムの意図がつかみやすい音声出力ができていることを示した．具体的には，音声対話システムの構成のうち音声出力に関わる言語生成・音声合成の両モジュールについて，手法を提案し，その評価を行なってきた．

第2章では，音声対話システムについて，その一般的構成を概観すると共に，音声対話システムにおける諸研究について，各研究が着目しているモジュール別に分類して紹介した．

音声対話システム研究は，そのほとんどが「対話音声特有の現象」にどのように対処するか，についてのものである．しかしながら，日本国内外を問わず，音声対話システム研究のうち最も多いのが音声認識・言語理解に関する研究である．このような研究としては，頑健な言語理解手法や韻律情報に基づくユーザの態度の認識，雑音下での音声認識と非常に研究内容が多岐に渡っている．また，対話管理に関する研究として，ユーザ発話途中で認識・応答を行なえるシステムや適切なタイミングで相槌を打つといった研究もなされている．

このような中で，音声出力に関する研究はほとんどなされていない．特に，概念音声合成の枠組を実現・実装した研究例は日本国内では皆無であり，海外に目を向けても近いものがわずかに存在する程度である．ほとんどの音声対話システムでは，テキスト音声合成と呼ばれる手法によって音声出力を行なってきた．しかしながら，音声対話システムに求められる「対話音声」を考えた場合，テキストのみから音声を合成するテキスト音声合成では,「対話音声特有の現象」を全くと言っていいほど反映させることができない．

このような背景が動機となり，本研究では概念音声合成の実現を目指して研究を進めてきた．

第3章では，エージェント音声対話システムについて述べた．このエージェント音声対話システム構築において，本論文における提案手法の基礎が確立された．

言語生成（応答文生成）では，統語構造が応答音声の韻律のうちイントネーションに大きく関わってくることから，常に構文木構造を保持したまま扱う，という方針が決定された．それを実現するために，テンプレートとして構文木構造を持たせた定型文を用意し，対話状況に応じて適切な定型文を用意する，という手法を採用した．また，定型文の中にタグを挿入しておくことにより，同じ属性の単語は同じタグを用いることができるようにした．そして，このタグに単語を挿入する際,「重要度」・「新規性」という談話情報を同時に挿入しておき，韻律制御の際に用いることのできるような手法を提案した．

音声合成（韻律制御）では，基本周波数パターン生成過程モデルに基づくフレーズ指令・アクセント指令について，統語構造や談話情報によって指令の大きさや挿入位置を制御する，という手法を提案した．特に，付属語アクセント結合規則を導入することにより，より自然な合成音声が生成できることが聴取実験から確認された．

第4章では，道案内音声対話システムについて述べた．道案内音声対話システムは，エー

ジェント対話システムでは実現できなかったような多様な応答音声の種類の実現を目的として構築された．この道案内音声対話システムにおいて，本論文における提案手法がさらに改良された．

言語生成（応答文生成）では，エージェント音声対話システムにおける手法では，修飾語を付ける等のわずかな応答文の変化に対しても異なる定型文を用意する必要があり，タスクの拡張等に伴い応答の種類が多様化すればするほど，必要な定型文の数も増大する，という問題点があった．そこで，道案内音声対話システムにおいて，新たに文節単位でテンプレート（定型フレーズ）を用意し，定型フレーズを適切に組み合わせることで応答文を生成する，という手法を提案した．定型フレーズを組み合わせる際，構文木構造や談話情報は常に保持されたまま組み合わされる．その結果，応答の種類が多様化すればするほど，それらを実現するのに必要なテンプレートの数が，定型フレーズを用意する手法の方が定型文を用意する手法に比べて圧倒的に小さい増加量で済んだ．そのため，定型フレーズを用意する手法の方が少ないテンプレート数でより多様な応答生成を実現できることがわかった．また，定型フレーズを用意する手法は，タスクに依らず汎用的な手法であることも示唆された．

音声合成（韻律制御）では，新たに文節間結合規則を導入した．これにより，エージェント音声対話システムの時点での手法に比べてより自然な応答音声が合成できていることが聴取実験から確認された．

また，本研究における提案手法の有効性を評価するために，比較音声を用いての評価実験を行なった．統語構造の観点からは，比較音声として一般的な形態素・構文解析ツールである JUMAN+KNP を用いて得られた構文情報を基に合成した音声を用意した．また，談話情報の観点からは,「重要度」「新規性」という概念を持たない音声を用意した．提案手法とこれら 2 種類の音声を聴取実験によって比較してもらった結果，提案手法による合成音声が最も自然な音声であるという結果が得られ，提案手法による構文情報の取り扱い・談話情報による韻律制御が有効であることが示された．

さらにより詳細な検討を加えるために，より多様な応答生成を可能にする必要性があったため，対話の種類や場所の種類を増加させるというタスク拡張を行なった．その結果を基に再び聴取実験を行なった．この実験では，先行する聴取実験と同様の比較音声を用意し，音声の自然性の他に「システム発話のうちどの語が伝えるべき語か」「伝えるべき語がきちんと伝えられているか」という評価尺度も導入した．その結果，音声の自然性に関しては先行実験と同様の結果が得られ，また新たな評価尺度から，対話制御に関する新たな知見を得ることができた．

5.2　課題と今後の展望

5.2.1　課題

概念情報を韻律に反映させた応答生成としては，本研究において一定の成果を挙げることができたと考えられる．しかしながら，より「わかりやすい」応答生成という観点から

考えると，テキストレベルでの対話制御をより詳細に行なう必要がある．

語順の観点から言えば，例えば「右に曲がるとコンビニがあります」というような文章を考えた場合，強調すべき語は「コンビニ」となる．この時，本研究で提案した「重要度」による韻律制御の他にも，語順を入れ替えることによって「コンビニが，右に曲がるとあります」のように，強調すべき語を文頭に持ってくることでもよりユーザにとって意味の伝わりやすい応答となると考えられる．

また，より複雑な対話を行なえるような状況を想定すると，適切な省略・照応表現の使用も重要な課題として挙げられる．現時点で道案内音声対話システムにおいて可能な対話は比較的単純なものであるが，タスク拡張等を考えた場合には，全く省略・照応表現を用いないとかえって冗長な応答となることが考えられる．かといって，過度な省略・照応表現の使用はユーザにとって「わかりづらい」応答となるため，どの程度省略・照応表現を用いるかを含めて検討する余地がある．

本研究では，音声合成器として，提案手法に則した波形接続型の音声合成器を用いている．イントネーションやアクセントといった韻律の面からは,「対話音声の合成」という目的に対する一定の成果を挙げているが，音質の面から考えると，機械音に聞こえる感が否めない．音質に関しては，近年音声合成の分野において盛んに研究が行なわれている HMM 音声合成方式 [50, 51, 67] を基に，本研究における統語構造・談話情報の取り扱い手法を踏襲した音声合成手法の構築が最も有効であると考えられる．HMM 音声合成方式では，学習データの質・量が合成音声の品質に関わってくる．現在，道案内音声対話を多数収録したコーパス作成プロジェクトが進められており，早期の完成が望まれる．

5.2.2　今後の展望

既に述べた通り，音声対話システムにおける音声出力に関する研究は非常に少ない．その中で,「対話音声」の合成を目指し，一定の成果を挙げた本研究は，今後の対話音声合成研究の方向性について 1 つの道を示したものであると考えている．

「対話音声」の合成に関しては，本研究で取り上げた語の「重要度」・「新規性」以外にも，パラ言語情報・非言語情報と呼ばれるテキストには現れない情報の韻律への反映が重要となる．特に「人間らしさ」という見方をした場合，意図や感情といったパラ言語情報の合成音声への反映が挙げられる．このような観点からの研究は既に [68, 69] 等によって行なわれているが，これらの研究はコーパスベースによる音声合成手法である．そこで，本提案手法と組み合わせることによって，意図や感情といった「概念表現」から感情豊かな応答生成を行なうことのできる音声対話システムが実現できるのでは，と見込まれる．また，このような考え方は，年齢や性別といった非言語情報についても当てはまる．この観点から考えると，ユーザによって話し手が様々に変化するような音声対話システムも実現できるのでは，と思われる．

音声対話システムの究極の目標は,「人」との対話と同じことが「システム」との間で行なえることである．本研究は「音声出力」に関する研究であったが，最終的には「音声入力」や「対話制御」の観点から行なわれている諸研究の技術と組み合わせることで，この究極の目標に近付いていけるのでは，と考えている．

謝辞

本論文を執筆するにあたり，指導教員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，前秘書の光永悦子さん，峯松研秘書の笠島恵さんに，深く感謝いたします．

音声対話システムの研究を私の前に行なっておられた桐山伸也氏，多胡順司氏には，本研究を進めていく上で様々な助言を頂きました．また，高田靖也氏とは共同研究者として，3 年間にわたり議論を交わしたり共同でシステムを開発したりと充実した研究生活を送ることができました．深く感謝いたします．

本研究の一部は，科学研究費補助金（学術創成研究）13NP0301 「言語理解と行動制御」（2001 年 4 月〜2006 年 3 月）の援助により行なわれました．プロジェクト内での会議において，代表の田中穂積氏（プロジェクト当時東京工業大学教授）を始め，非常に多くの方々に助言を頂きました．本プロジェクトに関わっておられた全ての方々に感謝いたします．

日頃の研究室生活を行なう上で，様々な面で多くの方の助力をいただきました．西澤信行氏，毛利太郎氏には，計算機管理の仕事で非常にお世話になりました．また，先輩であられる成澤修一氏，渡辺美知子氏，修士課程時の私の同期であられる鄭聖曄氏，佐藤賢太郎氏，浜野紘一氏，高橋力矢氏，学部時代からの友人である朝川智氏，後輩であられる古山悠介氏，村上隆夫氏といった方々とは，研究に関する議論を交わしたり研究の合間に雑談をしたり時には私生活に関する話題にも及んだりと，5 年間の広瀬（啓）・峯松研生活を送る上で誰一人欠かすことのできない方ばかりでした．残念ながら全員の名前を挙げることはできませんが，ここに名前を挙げた以外の方も含め，私の 5 年間の研究室で関わった全ての皆様に深く感謝いたします．

博士課程時において，東京大学 21 世紀 COE「未来社会を担うエレクトロニクスの展開」に特別補助研究員という形で関わらせて頂きました．博士 1 年次の年度末報告会として，私と共に力を合わせてフォーラム [70] を開催した神尾正太郎氏，川西直氏，鈴木康文氏，Chaminda de Silva 氏には，深く感謝いたします．このフォーラム立ち上げは，非常にやりがいのあるものでした．その他にも，COE に関わった教職員・学生の皆様に深く感謝いたします．

2006 年 11 月〜12 月に行なわれた「日米音響学会ジョイントミーティング」に参加するにあたり，東京大学学生生活委員会より東京大学学術研究活動等奨励事業学術奨励費を頂きました．深く感謝いたします．

謝辞

電気系事務室の方々には，学部 3 年次からの 7 年にも渡り，履修関係や事務手続き等で非常にお世話になりました．深く感謝いたします．

本研究を行なうにあたり，聴取実験に協力して下さった研究室内外の被験者の方にも，深く感謝いたします．また，本研究の発表の際に様々な助言を下さった，国内外の全ての研究者の方々に感謝いたします．

中学・高校時代からの友人，教養学部時代からの友人，サークル活動における仲間，電気系進学からの友人，これら日頃から多岐にわたり私を支えて下さった全ての友人に感謝いたします．

最後に，27 歳の今まで私の博士課程進学というわがままを聞き入れ，援助・助力を尽くして下さった家族に深く感謝いたします．

2006 年 12 月 15 日
八木 裕司

参考文献

[1] 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.

[2] 独立行政法人新エネルギー・産業技術総合開発機構 NEDO: "接客ロボット Actroid," http://www.nedo.go.jp/expo2005/robot/work/page007.html

[3] 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.

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

[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] 広瀬啓吉: "音声合成技術," 情報処理学会誌, Vol.38, pp.11, pp.984-991, 1997.

[7] S. Bennacef, L. Devillers, S. Rosset and L. Lamel: "Dialog in the RAILTEL Telephone-Based System," Proc. ICSLP'96, Vol.1, pp.550–553, 1996.

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

[9] X. D. Huang, Y. Ariki and M. A. Jack: "Hidden Markov Models for Speech Recognition," Edinburgh University Press, 1990.

[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. Eurospeech'93, pp.2063–2066, 1993.

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

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

[14] S. Seneff: “TINA : A Natural Language System for Spoken Language Applications,” Computational Linguistics, Vol.18, No.1, pp.61–86, 1992.

[15] 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.

[16] Shinya Fujie, Daizo Yagi, Yosuke Matsusaka, Hideaki Kikuchi and Tetsunori Kobayasi: “Spoken Dialogue System Using Prosody as Para-Linguistic Information,” Proc. Speech Prosody 2004, pp.387–390, 2004.

[17] 藤崎博也: “韻律研究の諸側面とその課題,” 日本音響学会講演論文集, Vol.1, 2-5-11, pp.287–290, 1994.

[18] 江尻康, 松坂要佐, 小林哲則: “対話中における頭部ジェスチャの認識,” 電子情報通信学会技術報告, Vol.102, No.218, pp.31–36, PRMU202-61.

[19] K. Kurakata, K. Matsushita and Y. Kuchinomachi: “Database of Domestic Sounds for Evaluation of Auditory-signal Audibility : JIS/TR S 001,” 日本音響学会誌, Vol.24, No.1, pp.23-26, 2003.

[20] 小川峰義, 高橋玲: “環境要因の評価に用いる騒音データベースの構築,” 日本音響学会 1996 年秋期研究発表会講演論文集, 1-P-13, pp.187-188, 1996.

[21] K. Yao and S. Nakamura: “Sequential noise compensation by sequential Monte Carlo method,” Proc. NIPS’01, pp.1205-1212, 2001.

[22] M. Fujimoto and S. Nakamura: “Particle filtering and Polyak averaging-based non-stationary noise tracking for ASR in noise,” Proc. ASRU05, pp.337-342, 2005.

[23] M. S. Arulampalam, S. Maskell, N. Gordon and T. Clapp: “A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking,” IEEE Trans. SP, Vol.50, No.2, pp.174-188, 2002.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

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

[31] 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.

[32] 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.

[33] （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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

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

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

[42] 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.

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

[44] 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/.

[45] （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.

[46] 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.

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

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

[49] 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.

[50] 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.

[51] 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.

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

[53] 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.

[54] 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.

[55] 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.

[56] 山下洋一, 水谷直樹, 角所収, 溝口理一郎: “汎用音声出力インタフェースにおける概念表現からの音声合成,” 電子情報通信学会誌, vol.J76-D-II, no.3, pp.415-426, 1993.

[57] “OpenGL,”
http://www.opengl.org/

[58] 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.

[59] 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. (CD-ROM)

[60] “The MBROLA Project,”
http://tcts.hpms.ac.be/synthesis/mbrola.html

[61] “Extensible Markup Language (XML),”
http://www.w3c.org/XML/

[62] “Netwek,”
http://www.netwek.com/

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

[64] 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, vol.70, no.4, pp.378-385, 1987.

[65] 日本語形態素解析システム JUMAN ver5.1,
http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html

[66] 日本語構文解析システム KNP ver2.0,
http://www.kc.t.u-tokyo.ac.jp/nl-resource/knp.html

[67] 小林隆夫, 徳田恵一: “コーパスベース音声合成技術の動向 [IV]-HMM 音声合成方式-,” 電子情報通信学会誌, vol.87, no.4, pp.322-327, 2004.

[68] K. Hirose, K. Sato, Y. Asano and N. Minematsu: "Synthesis of F_0 contours using generation process model parameters predicted from unlabeled corpora: applicatoin to emotional speech synthesis," Speech Communication, vol.46, no.3-4, pp.385-404, 2005.

[69] M. Tachibana, J. Yamagishi, T. Masuko and T. Kobayashi: "A style adaptation technique for speech synthesis using HSMM and suprasegmental features," IEICE Trans. Information and Systems, vol.E89-D, no.3, pp.1092-1099, 2006.

[70] Frontier Young Researchers' Forum〜新領域若手の会〜
http://www.mlab.t.u-tokyo.ac.jp/fyrf/

[71] H. Fujisaki: "From information to inotation," Proc. 1993 International Symopsium on Spoken Dialogue, pp.7-18, 1993.

発表文献

[1] Yuji Yagi, Seiya Takada, Keikichi Hirose and Nobuaki Minematsu: "Concept-to-Speech Conversion for Reply Speech Generation in a Spoken Dialogue System for Road Guidance and Its Prosodic Control", 4th Joint Meeting of ASA(Acoustical Society of America)/ASJ(Acoustical Society of Japan), 2006-11.

[2] Yuji Yagi, Seiya Takada, Keikichi Hirose and Nobuaki 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.

[3] Yuji Yagi, Seiya Takada, Keikichi Hirose and Nobuaki 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), vol.2, pp.703-706, 2005-10.

[4] 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," 情報処理学会研究報告（音声言語情報処理研究会）, 2005-SLP-57-16, pp.93-98, 2005-7.

[5] 八木裕司, 多胡順司, 峯松信明, 広瀬啓吉: "エージェント対話システムにおける対話管理と応答生成," 情報処理学会研究報告（音声言語情報処理研究会）, 2003-SLP-47-13, pp.65-70, 2003-7.

[6] 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音声の評価," 日本音響学会 2007 年春期研究発表会講演論文集, 1-9-5, pp.???-???, 2007-3.（発表予定）

[7] 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, pp.123-124, 2006-3.

[8] 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, pp.121-122, 2006-3.

[9] 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.

[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

[11] 八木裕司, 広瀬啓吉, 峯松信明: “韻律に着目した対話システムにおける応答生成の改良,” 日本音響学会 2004 年春期研究発表会講演論文集, vol.1, 3-8-7, pp.135-136, 2004-3.

[12] 八木裕司, 広瀬啓吉, 峯松信明: “エージェント対話システムにおける応答生成の改良,” 日本音響学会 2003 年秋期研究発表会講演論文集, vol.1, 1-6-25, pp.49-50, 2003-9.

[13] 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 Spoken Dialogue System,” Annual Project Report Grant-in-Aid for Creative Basic Research “Language Understanding and Action Control,” pp.219-227, 2006-3.

[14] Keikichi Hirose, Nobuaki Minematsu, Yuji Yagi and Seiya 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.

[15] Keikichi Hirose, Yuji Yagi, Seiya Takada, Yasufumi Asano and Nobuaki 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.

[16] Keikichi Hirose, Yuji Yagi, Kentaro Sato and Nobuaki Minematsu: “Generation of Speech Reply in an Agent Dialogue System and Synthesis of Emotional Speech,” Annual Project Report Grant-in-Aid for Creative Basic Research “Language Understanding and Action Control,” pp.240-249, 2004-3.

[17] Keikichi Hirose and Yuji Yagi: “Generation of Reply Speech in the Agent Dialogue System,” Proc. International Symposium on Electronics for Future Generations, pp.191-196, 2004-3.

[18] 八木裕司: “エージェント対話システムにおける応答音声生成手法の改良,” 修士論文, 東京大学大学院工学系研究科, 2004-1.

付録A

基本周波数パターン生成過程モデル

A.1　基本周波数パターン生成過程モデル

音声の韻律的特徴を表現する客観的・物理的な量として，基本周波数（F_0）パターンは言語の構文や意味の伝達に重要な役割を果たしている．F_0 パターンは日本語を含む多くの言語の音声の抑揚を表し，一般に単語レベルのアクセントに対応する局所的で急激な起伏と，句・節・文レベルの，より広い範囲にわたる緩やかな起伏とから成るが，この F_0 パターンが生成される過程のモデルとそのモデルを用いる分析手法は，藤崎らによって初めて考案された [58]．このモデルでは，発話の言語学的情報と密接に関係のある少数のパラメータで，実際に観測される F_0 パターンを極めて良く近似できることが知られている [36]．

A.2　F_0 パターンとその生成過程モデル

音声の F_0 パターン生成過程モデルは，対数軸上で表現した F_0 パターンが 2 種類の成分の和により表されるとしている．その 1 つは，句頭から句末に向かう緩やかな下降に対応するもので，これをフレーズ成分と呼ぶ．2 つめは，個々の単語また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 パターンを話者ごと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，単語及び文音声の F_0 パターンの特徴を統一的に把握することができる．

F_0 パターンを測定することにより，声の主観的な高低の型を，それと対応する客観的な物理量として表すことができる．

A.2.1　フレーズ成分

F_0 パターンを構成している成分の 1 つであるフレーズ成分は，およそ全ての発話に共通なもので，声帯振動の開始よりもおよそ 200〜400ms 以内から準備され始め，やや上昇しながら最大値に達した後，緩やかに下降してある一定の値に漸近し，発話の終端近くで急激に下降する成分である．

これは単独発話では 1 個であるが，文の発話では複数個存在し得る成分で，その形は質量とバネ定数とを持つ 2 次の力学系が瞬間的な外力（インパルス引力）を受けた場合の運動とよく似ている．これを数学的に表現するため，フレーズ成分を質量・バネ定数・摩擦抵抗を持った何らかの力学系のインパルス応答を用いて近似し，かつ仮想的な力学系は線形性を持ち，臨界制動系であると仮定すれば，フレーズ成分に相当する系のインパルス応答 $G_p(t)$ は次式で表される．

$$G_p(t) = \begin{cases} \alpha^2 t \exp(-\alpha t) & t \geq 0 \\ 0 & t < 0 \end{cases}$$

ここで，α はフレーズ指令に対する系の速さを定める係数であり，日本人の話者の平均的な値として，経験的に 3.0 を用いると良いことがわかっている [71]．

A.2.2　アクセント成分

F_0 パターンを構成しているもう 1 つの成分であるアクセント成分は，個々の単語または連続した単語に付随するもので，主観的に高い拍の発音にやや先行して始まり，始めはゼロから緩やかに上昇し，途中はかなり急激に上昇し，その後また緩やかに一定のレベルに漸近するもので，高い拍が続けばそのレベルを保ち，高い拍から低い拍への移行に際しては，上記とは逆に低い拍の発音にやや先行して緩やかな下降を始め，途中は急激に下降し，その後またゼロとなる成分である．

これは，質量とバネ定数とを持つ 2 次の力学系が，ある時間持続する一定の外力（ステップ入力）を受けた場合の運動とよく似ている．これもフレーズ成分と同様に数学的に表現すると，アクセント成分に相当するステップ応答 $G_a(t)$ は次式で示される．

$$G_a(t) = \begin{cases} \min[1-(1+\beta t)\exp(-\beta t), \gamma] & t \geq 0 \\ 0 & t < 0 \end{cases}$$

ここで，β はアクセント成分の立上りの速さを定める係数であり，平均的な値として，20.0 が用いられる [71]．記号 $\min[x, y]$ は，x と y のうち小さい方を取ることを意味する．実際の F_0 パターンでは $G_a(t)$ が有限の時間内に上限値 γ に達し，以後その値を保持することに対応する．γ の値は，通常 0.9 が用いられる [71].

A.2.3　基本周波数パターンの生成

これら 2 つの成分を用いて，F_0 パターンの特徴を非常に良く近似することができる．フレーズ成分を比較的時定数の長い線形 2 次系のインパルス応答，アクセント成分を比較的時定数の短い線形 2 次系のステップ応答で近似できるものとし，それらの和に非礼して対数 F_0 パターンが変形するものとしている．また，これらの表現を用いれば，F_0 パターンの特徴を良く近似できるだけでなく，発話の言語学的意図から F_0 パターンが生成される過程について，図 A.1 のようなモデルを考えることができる．

図 A.1 は，文音声の F_0 パターンを想定したもので，入力となる 2 種類の指令のうち，フレーズ成分の指令は正または負のインパルスとして，正のインパルスは文頭・文中のフレーズの先頭に，また，負のインパルスは文の終わりに生起してそれまでのフレーズ成分を下降させる役割を持っている．また，アクセント指令は正の方形波として，個々の単語または単語連鎖ごとに生起してアクセント成分を生ずる．最後にこの 2 種類の成分は相加され，声帯振動の基本周波数の対数値に変化を生ずる．ここで，時刻 t における基本周波数の値を $F_0(t)$ で表せば，その対数値は具体的には次式で表される．

$$\log_e F_0(t) = \log_e F_b + \sum_{i=1}^{I} A_{pi} G_p(t - T_{0i}) + \sum_{j=1}^{J} A_{aj}\{G_a(t - T_{1j}) - G_a(t - T_{2j})\}$$

ここで，F_b は F_0 パターンの基底値（基底周波数），I は文中のフレーズ指令の数，J は文中のアクセント指令の数，A_{pi} は i 番目のフレーズ指令の大きさ，A_{aj} は j 番目のアクセント指令の大きさ，T_{0i} は i 番目のフレーズ指令が生起する時点，T_{1j} は j 番目のアクセント指令の立上り時点，T_{2j} は j 番目のアクセント指令の立下り時点を表す．

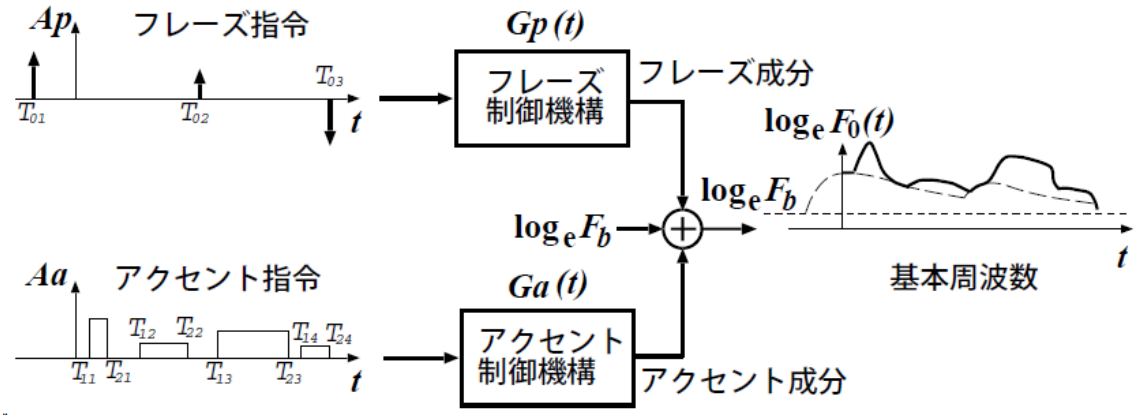


図 A.1: 基本周波数パターン生成過程のモデル [58]

A.3　音声合成器における F_0 モデル

A.3.1　F_0 モデルと韻律制御記号との対応

3.4.2 節や 4.5 節で述べた，本論文にて用いる音声合成器で用いられる韻律制御記号も，この基本周波数パターン生成過程モデルに基づくものである．A.2.3 節の式と合成器での韻律制御記号との対応は，以下のようになる．まず，フレーズ指令に関しては，時刻 T_{0i} に大きさ A_{pi} のフレーズ指令を立てることを，合成器では Ph/Pm/P0（ただし，Ph は文頭，P0 は文末のみ）を立てることで表す．また，アクセント指令に関しては，時刻 T_{1j} に大きさ A_{aj} のアクセント指令を立てることを，合成器では D$I_1I_2I_3$/F$I_1I_2I_3$ （D/F はアクセント型によって決定，I_1，I_2，I_3 は 3.4.2.2 節の表 3.2 参照）を立てることで表す．また，それに対応するアクセント立下げについては，時刻 T_{2j} に j 番目のアクセントを立下げることを，合成器では A0 を用いて表す．

A.3.2　音声合成器における韻律制御記号の大きさ

本節では，音声合成器で用いられる記号と，パラメータごとの大きさを示す．

A.3.2.1　文頭フレーズ指令

文頭フレーズ指令とは，文の先頭に置かれるフレーズ指令である．文頭フレーズ指令におけるパラメータごとの大きさを表 A.1 に示す．これらの文頭フレーズ指令は，実際には合成器で置かれている場所（つまり発話開始時点）の 210ms 手前に立てられる．

A.3.2.2　文中フレーズ指令

文中フレーズ指令とは，フレーズ指令のうち文頭以外に置かれるものを表す．文中フレーズ指令におけるパラメータごとの大きさは，P22=0.15，P21=0.25，P12=0.27，P11=0.37

表 A.1: 文頭フレーズ指令のパラメータごとの大きさ

パラメータ	大きさ	パラメータ	大きさ	パラメータ	大きさ	パラメータ	大きさ
P222122	0.16	P212122	0.21	P222112	0.22	P122122	0.24
P221122	0.25	P112122	0.26	P212112	0.27	P122112	0.27
P222121	0.28	P221112	0.30	P211122	0.30	P121122	0.30
P222222	0.32	P112112	0.32	P222111	0.33	P212121	0.33
P122121	0.33	P111122	0.35	P211112	0.36	P121112	0.36
P221121	0.37	P212222	0.37	P122222	0.37	P222212	0.38
P212111	0.38	P122111	0.38	P112121	0.38	P221222	0.41
P111112	0.41	P221111	0.42	P211121	0.42	P112222	0.42
P121121	0.42	P212212	0.43	P122212	0.43	P112111	0.43
P222221	0.44	P221212	0.46	P211222	0.46	P121222	0.46
P211111	0.47	P121111	0.47	P111121	0.47	P112212	0.48
P222211	0.49	P212221	0.49	P122221	0.49	P111222	0.51
P221221	0.52	P211212	0.52	P121212	0.52	P111111	0.52
P212211	0.54	P122211	0.54	P112221	0.54	P121221	0.57
P111212	0.57	P221211	0.58	P211221	0.58	P112211	0.59
P211211	0.60	P121211	0.60	P111221	0.60	P111211	0.60

となっている．これらの文中フレーズ指令には，実際に合成器で置かれている場所の 30ms 手前に立てられる．

A.3.2.3　文末フレーズ指令

文末フレーズ指令とは，文の末尾に置かれるフレーズ指令である．文末フレーズ指令 P0 の大きさは-0.30 である．文末フレーズ指令も，実際に合成器で置かれている場所の 30ms 手前に立てられる．

A.3.2.4　D 型アクセント立上げ指令

D 型アクセント立上げ指令は，頭高型もしくは起伏型のアクセントを表す．D 型のアクセント指令における，パラメータごとの大きさを表 A.2 に示す．これらの D 型アクセント指令は，実際には合成器で置かれている場所の 20ms 手前に立てられる．

A.3.2.5　F 型アクセント立上げ指令

F 型アクセント立上げ指令は，平板型のアクセントを表す．F 型のアクセント指令における，パラメータごとの大きさを表 A.3 に示す．これらの F 型アクセント指令も，実際に

表 A.2: D 型アクセント指令のパラメータごとの大きさ

パラメータ	大きさ	パラメータ	大きさ	パラメータ	大きさ	パラメータ	大きさ
D422	0.24	D222	0.30	D412	0.30	D212	0.36
D421	0.38	D122	0.39	D322	0.41	D221	0.44
D425	0.45	D423	0.45	D411	0.45	D112	0.45
D424	0.46	D312	0.47	D211	0.50	D413	0.51
D225	0.51	D223	0.51	D415	0.52	D414	0.52
D224	0.52	D121	0.53	D321	0.55	D215	0.57
D213	0.57	D214	0.58	D111	0.59	D125	0.60
D123	0.60	D311	0.61	D124	0.61	D325	0.62
D323	0.62	D324	0.63	D115	0.64	D113	0.64
D114	0.65	D315	0.65	D313	0.65	D314	0.65

表 A.3: F 型アクセント指令のパラメータごとの大きさ

パラメータ	大きさ	パラメータ	大きさ	パラメータ	大きさ	パラメータ	大きさ
F422	0.17	F412	0.23	F222	0.23	F212	0.29
F421	0.31	F122	0.32	F322	0.34	F411	0.37
F221	0.37	F425	0.38	F423	0.38	F112	0.38
F424	0.39	F312	0.40	F211	0.43	F415	0.44
F413	0.44	F225	0.44	F223	0.44	F414	0.45
F224	0.45	F121	0.46	F321	0.48	F215	0.50
F213	0.50	F214	0.51	F111	0.52	F125	0.53
F123	0.53	F311	0.54	F124	0.54	F325	0.55
F323	0.55	F324	0.56	F115	0.59	F113	0.59
F114	0.60	F315	0.61	F313	0.61	F314	0.62

は合成器で置かれている場所の 20ms 手前に立てられる．

A.3.2.6　アクセント立下げ指令

D 型・F 型のアクセントの立下げ指令 A0 の大きさは 0 である．アクセント立下げ指令は，実際に合成器で置かれている場所の 10ms 手前に立てられる．

A.3.2.7　その他の韻律制御記号

本研究で用いている音声合成器には，フレーズ指令・アクセント指令の他にも様々な韻律制御記号がある．以下，それらについて述べる．

ポーズを表す記号として，Sn がある．n は 1〜5 の値をとり，S1=700，S2=300，S3=100，S4=40，S5=20（単位は ms）である．また，促音もポーズの一種として表現され，記号は SX，長さは 160ms となる．

また，基底周波数（A.2.3 節での F_b）は 120Hz である．