修　士　論　文

ロボット・プレゼンテーション エージェントの 音声インタラクティブ性向上の研究

平成20年2月4日提出

指導教員　石塚 満 教授

東京大学大学院 情報理工学系研究科

66442　簑津 真一郎

内容梗概

ロボットやアニメーションエージェントの応用の一つとして，自動プレゼンテーションが注目されている．その実現のために，ユーザからの音声入力を受け付けてプレゼンテーションの内容を変えるプレゼンテーション／エージェントシステムの開発を行った．

インタラクティブなプレゼンテーションを容易に作成することを目指し，MPML-HR に対して音声インタラクション機能を導入した．インタラクション機能の導入にあたり，説明の省略，前に話した内容の再度の説明，およびあらかじめ想定した説明に答えるということを目的とした．音声認識誤りへの対処が必要となるため，音声認識結果に対する信頼度を用いた機構を導入した．信頼度が低い場合は音声入力受付時に棄却，聞き返し，確認，受理のいずれかを行う．

インタラクティブなプレゼンテーションを行う際に，視聴者がどのような発話内容で質問や要求を行えばよいかがわからないという問題が発生する．これに対処するため，プレゼンテーションを行うエージェントのほかにもう一つのエージェントを用いて，質問や要求をデモンストレーションさせることにより，暗示的に視聴者に発話の仕方を教える方法を提案した．

音声認識率の向上のために bag-of-words をベースにしたトピック認識技術を導入した．音声認識システムは単語の出現頻度パターンとして文章のカテゴリを学習する．この方法により，音声認識誤りや音声認識の辞書に登録されていない単語に対しても頑健な学習と理解が可能となる．

目 次

図 目 次

表 目 次

第1章　序論

1.1　研究の背景と目的

近年ロボット技術の発展がめざましく，一般家庭へのロボットの普及も現実味を帯びてきた．ロボットはマルチモーダルな入出力を行うことができ，中でも操作手段としては，音声が有力である．ロボットによるインタラクティブなプレゼンテーションの技術は今後必要になっていくと考えられる．

当研究室では，マルチモーダルな情報提供手段として，プログラミングスキルを必要とせず，誰でも容易にマルチモーダルなコンテンツを作成，視聴可能とする記述言語 MPML (Multimodal Presentation Markup Language) の開発を行ってきた．MPML は，XML 規格に準拠したマークアップ言語で，CG アニメーションキャラクタによるプレゼンテーションを，Web ページを作成するように容易に記述することができる．一方，ロボット技術の発達により，これまで生産作業中心に利用されて来たロボットが，人間のパートナーとして用いられるようになってきた．また，プレゼンターとしてヒューマノイドロボットを利用できるように MPML を拡張し，MPML-HR を開発してきた．これまで，それぞれ専用の制御言語で操作されてきた様々なロボットを，MPML-HR の簡易なスクリプト記述で自由に操作することが可能となっている．

1.2　問題

インタラクティブなプレゼンテーションは，ロボットのプレゼンテーションにおいて音声による割り込みを受け付ける事で実現できるが，様々な問題も存在する．

MPML-HR を用いることでロボットからの一方的なプレゼンテーションコンテンツの作成は可能となったが，インタラクションを含むコンテンツの作成はできなかった．

また，インタラクティブなプレゼンテーションエージェント/ロボットシステムでは，視聴者が，どのような発話内容やタイミングで質問や要求を行ってよいかがわからないという問題がある．

さらに，インタラクティブなプレゼンテーションにおいて，視聴者の発話に対するシステムの反応をより適切なものとするには，発話の内容からトピックを認

識することが必要である．例えば，文法によるトピック認識ではフィラーによって認識精度が左右されるという問題が存在する．

1.3 本研究のアプローチ

MPML-HR を用いることでロボットからの一方的なプレゼンテーションコンテンツの作成は可能となったが，インタラクションを含むコンテンツの作成はできなかった．インタラクションはプレゼンテーションの重要な要素であり，プレゼンテーションに関する質問に答えることでプレゼンテーションを分かり易くさせ，充実したコンテンツとすることができると考えられる．そこで，これまで開発してきた MPML-HR を拡張し，プレゼンテーションコンテンツのインタラクションを対象とした記述言語を提案する．

インタラクティブなプレゼンテーションエージェント/ロボットシステムでは，視聴者が，どのような発話内容やタイミングで質問や要求を行ってよいかがわからないという問題がある．この問題に対して，プレゼンテーションを行うエージェントのほかに，もうひとつのエージェントを用い，質問や要求をデモンストレーションさせることにより，暗示的に，視聴者に発話の仕方を教える方法を提案する．この方法により，視聴者が質問・要求を行う際に，エージェントの発話と類似した発話を行う可能性が高くなり，結果として発話の認識・理解の精度が高まることが期待される．

インタラクティブなプレゼンテーションにおいて，視聴者の発話に対するシステムの反応をより適切なものとするには，発話の内容からトピックを認識することが必要である．文法によるトピック認識では，フィラーによって認識精度が左右されるという問題が存在する．この問題に対処するため，SVMV(Single random Variable with Multiple Value) 確率モデルに基づくトピック認識法を提案する．この方法では，学習させる文章全てを bag-of-words で一つの文書とみなし，トピック認識を行う．このため，単語の順序には影響されず，またトピックとの相互情報量をもちいて索引語を制限することで，トピックの認識に関与しない単語を取り除くことができる．

1.4 本論文の構成

本論文は 6 章で構成されている．

第 2 章では，ロボットやエージェントとのインタラクションにおける先行研究について述べる．

第 3 章では，インタラクティブなプレゼンテーションの実現とその課題について述べる．

第 4 章では，プレゼンテーションを行うエージェントのほかに，もうひとつのエージェントを用い，質問や要求をデモンストレーションさせることにより，暗示的に，視聴者に発話の仕方を教える方法について述べる．

第 5 章では，トピック認識技術の導入による音声認識率の向上について述べる．

第 6 章は，本論文の結論である．

第2章　ロボットやエージェントを用いたインタラクティブなシステム

2.1　はじめに

本章では，先行研究の紹介を行う．ロボットやエージェントはマルチモーダルなインターフェースを持っている．ここではそれを用いた対話システムについて記述する．

2.2　対話システム

対話システムは，人とエージェント (ロボットやスクリーンエージェントなど) のコミュニケーションの実現を目的としている．

中野らの報告によるとコミュニケーションをするためには入力が必要であるが，音声による入力またはキーボードによる入力を採用している研究が主流である [5].

一般的に，対話システムは，発話理解と発話生成の二つの機能からなると考えられる．この二つのモジュールは，対話状態 (Dialogue State) と呼ばれるデータを介してつながっている．対話状態には，その時点までの対話の履歴と，ユーザの意図の推定結果やシステムの発話プランなどの情報が書き込まれる．発話理解は，ユーザの発話と現在の対話状態を入力として，新しい対話の状態を出力する関数とみることが出来る．同様に，発話生成は，現在の対話状態を入力とし，システム発話と新しい対話状態の二つを出力する関数と見なすことが出来る．

対話状態に含まれるべき情報として以下のようなものが提案されている．

- **対話履歴** その時点までのユーザ発話の理解結果とシステム発話の内容
- **ユーザ意図の推定結果** ユーザがどのようなことをしたいかをシステムが推定した結果.
- **グラウンディング情報** ユーザ意図の推定結果に含まれる情報のうち，どの情報は確認済み (とシステムが思っている) かの情報.

- **データベースの検索結果** ユーザの意図の推定結果に基づいてデータベースを検索した結果.
- **談話オブジェクト** 今までの対話に出てきた,「もの」や「こと」. 代名詞によって表現されることもある.

対話の進行を司るのが対話制御部である. データベースの検索に至るまでどのようにユーザの要求を聞き出すか, そしてデータベース検索の結果をどのようにユーザに提示すればより使いやすい対話システムになるかが対話制御の研究課題である.

プレゼンテーションのようにストーリーのある表現をエージェントにさせる場合, 主導権はエージェントにある. しかし, 実際の人同士の対話では, 相づちや割り込み発言といったインタラクションが生じるので, それを考慮にいれた実装を行う必要がある.

ドメインの限定された対話では, 人が行う発話がある程度限定されるため, 意図した対話が出来る. しかし, 雑談やチャットのような自由な対話はすることが出来ない. ドメインの限定されていない対話では, 自由な対話が可能となるが, 必ずしも意図した対応が取れるとは限らず, 的外れなことを話す確率が多くなる. よって, これらの技術を複合するような研究が必要であるだろうと考える.

2.3 エージェントからの発話中心の戦略

発話行為をするときに, エージェントからの説明を中心とする戦略である. 販売する商品のプレゼンテーションや講習会の講師役などをエージェントに行わせる場合, あらかじめ原稿を作っておけばきちんとした発表を行わせることが可能である. 人が行う場合と同様に, 発表の途中で質問があったときも, 一時的にそれを処理したあとに元のストーリーに戻れば一貫した内容を発表することが出来る.

実際にストーリーを記述するとき, ロボットやスクリーンエージェントといったマルチモーダルなインターフェースでは, 視線や動作, 感情の動きといった様々な情報を記述する必要がでてくる. この場合, それぞれの実装につきそれぞれ異なる記述方法があると, あまり習熟していない人には大きな障壁となってしまう. ロボットで言うと, 手を振る動作一つを記述するにも, 手をある座標から座標に移すという表現や, 一連の動作としてあらかじめプログラムされているものを呼び出したりと言ったように様々な記述方法がある. ここではこのような記述方法の標準化に関する記述言語について紹介する.

2.3.1 MPML(Multimodal Presentation Markup Language)

MPML(Multimodal Presentation Markup Language) は Microsoft Agent を用いて簡単にマルチモーダルなプレゼンテーションを作成することができる [9].

PC 上の Web ブラウザで表示することを前提に開発がすすめられてきたが，携帯電話上でもキャラクタ付きプレゼンテーションを行うコンテンツ記述が可能な MPML-Mobile Version も開発された．また，3D の全身を持ったキャラクタが 3 次元空間で動作し，プレゼンテーションを行う MPML-VR の開発も行われた．

最新の研究では，Fig. 2.1 のようなヒューマノイドロボットのプレゼンテーション制御が可能な，簡単なインタラクション機構を備えた MPML-HR の開発が行われている [2].

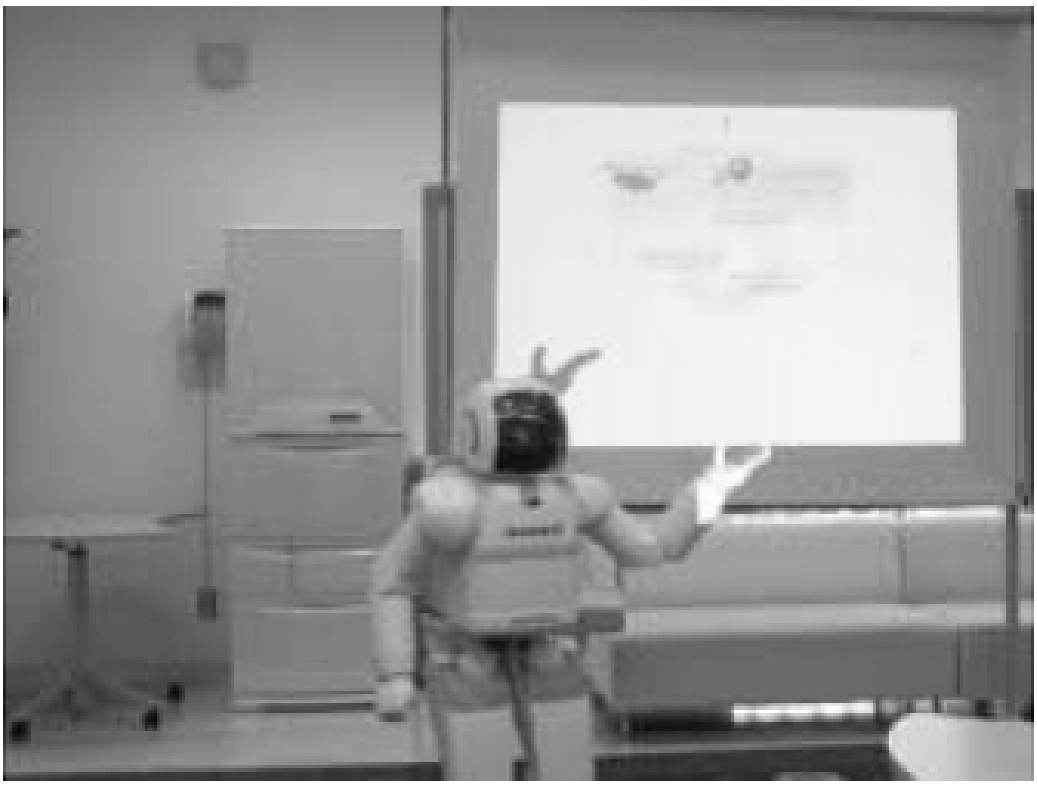

図 2.1: MPML-HR(Asimo 上で動作)

2.3.2 APML(Affective Presentation Markup Language)

対話を想定したマルチモーダルキャラクタの記述言語である．表情豊かな顔を持つキャラクタエージェントであり，基本的な表情を組み合わせて複雑な表情を作ることも可能になっている．記述は Web 上の標準として普及しつつある XML(Extensible Markup Language) のようなタグ付けで行うようになっている [10][11].

2.3.3 TVML(TV program Making language)

TVML は，テレビ番組を記述することが出来るテキストベースの言語である．NHK の林らによって開発された [16]．実際のテレビ番組製作現場で用いられている番組台本の記述法に基づき，可読性があり，誰でも簡単に使いこなすことが出来る言語になるようにデザインされている．

TVML 台本は，スクリプト一行がある一つのイベントに対応するという簡潔なものになっている．基本的には時間の推移に従って何のイベントがどのように行われるかを列挙した物になっており，イベントを時間軸に沿って並べることで番組台本が作成される．イベントは例えば次のような物である．

character:talk(name=BOB,text=”こんにちは”)

この一行を記述し，これを TVML プレイヤーと呼ばれるソフトに入力することで，スタジオショットに CG キャラクタ BOB が登場して，合成音声で「こんにちは」としゃべる．TVML で使えるイベントは，character，camera，movie など約 10 種類で，コマンドは 70 種類ほどが用意されている．これらのイベントを時間軸に沿って並べることで番組台本が作成される．

TVML は手軽なテレビ番組製作ツールとして利用されており，実際に「英語でしゃべら Night」などの番組で使用された．また，マルチメディアの様々な研究の曲面において，動画というアウトプットをもったシステムを必要とするときに有用なツールとして使われている．

2.3.4 VHML(Virtual Human Markup Language)

VHML は，マルチモーダルなシステムにおけるデザインを容易にするための言語である [13]．VHML では，顔の表現や目の位置など，MPML や TVML と比べるとより細かく指定することが出来るが，そのため記述量が増えてしまっている．XML/XSL に基づいて実装されている．VHML を用いることによって，顔のアニメーション，身体アニメーション，感情表現およびマルチメディア情報を考慮した対話プランを記述することが出来る．

2.4 特定ドメインに対するインタラクション・対話実現への取り組み

あらかじめストーリーが決まっていない場合，エージェントには，ユーザの呼びかけに対して適切に応答することが望まれる．しかし，ユーザの問いかけの内容は様々であり，有限の情報しか持たないエージェントはそのすべてに対応することは出来ない．

この問題を解決するために，ユーザからの問いかけのドメインを限定することによって対応を行うという戦略が提案された．

2.4.1 Galatea

Galatea Project は，擬人化音声対話エージェントのツールキット Galatea Toolkit を開発し，オープンソース，ライセンスフリーで公開提供するプロジェクトである．サブシステムの一つである Dialogue Manager において，Linux 版では VoiceXML，Windows 版では XISL と呼ばれる対話記述言語を通して，対話制御を行うことが出来る．

VoiceXML

VoiceXML は，音声対話のアプリケーションを構築するための XML である．電話回線を使って既存のアプリケーションとのインタラクションを行うことを目的として開発された．VoiceXML によって，アプリケーション構築の際に必要だった，機器ごとに規定された仕様などの情報がなくとも，より抽象的なレベルにおいての構築が可能となる．実装例として，Galatea を用いた音メディアアーカイブシステムがある [17]．

2.4.2 カーナビ音声対話システム MINOS

MINOS は，EUROPA フレームワークを用いて音声対話システムとして実装されている．EUROPA フレームワークは奈良先端大学の笹島宗彦らが開発したものであり，音声対話システムの構築を目的とした汎用フレームワークである [18]．

以下は EUROPE システムの構成である．

- **音声認識部** EUROPA では音声認識にキーワードスポッティング法を用いる．対象タスクで使用される語彙を，単語，発音，品詞インスタンスの三つ組で表現し，その集合を語彙セットとして与える．認識結果はキーワードラティスの形で出力される．

- **構文解析部** 構文解析部では，音声認識部から受け取ったキーワードラティスをもとに該当する文型を求める．必要なデータは，認識対象語彙を品詞クラスに分類した品詞データと，その品詞クラスの並びで表現した文型データである．これらのデータをもとに，文テンプレートハッシュ法を用いてキーワードラティスの中から，与えられた文型で受理できるキーワード系列の集合をもとめる．

- **意図変換部** 意図変換部は構文解析によって得られた文型と認識単語系列から，意図表現と呼ばれるユーザの意図を表すフレームに変換する．意図変換部には，各文型に対応したテンプレートが知識として与えられる．各々のテンプレートには認識単語をどのスロットに代入するかを示す畳み込み知識が組み込まれており，意図変換部はこの畳み込み知識を解釈して，認識単語を意図フレームの適切なスロットに代入し，ユーザの意図を表すフレームを作成する．

- **問題解決部** 問題解決にかかる手続きはUSHI(Unification-based Script Handling Instruction set) 言語によって記述する．USHI は Pascal に似た表記法をもつスクリプト言語で，If 文，For 文などの制御構造や各種演算子をもち，フレーム表現のデータを直接操作する機能を備えている．また，手続きの定義と呼び出しが記述できるようになっており，USHI スクリプト中で定義された手続きと C++プログラム中で定義された手続きの両方を呼び出すことが可能である．

- **応答生成部** 文生成テンプレートを元に問題解決結果から応答文を生成する．テンプレートには特定のスロットの値の取得や範囲指定された複数のスロットの値を列挙するなどの機能が用意されており，これらを利用して応答文を生成する．応答文生成にどのテンプレートを利用するかはUSHI 言語による手続きで記述する．問題解決結果に応じたテンプレートを選択することで，適切な応答文を生成できる．

このシステムは，カーナビという範囲にドメインを限ることによってタスクに対する応答を生成している．

2.5 より自由度の高いインタラクションへの取り組み

2.5.1 マルチドメイン対話システム

マルチドメイン音声対話システムとは，種類の異なる複数のドメインを扱う音声対話システムである．ドメインとは音声対話システムが処理するタスク領域であり，例えばメールドメインといった場合にはメールの検索，送信，削除，表示といったタスクの集合を表す．理想的なマルチドメイン音声対話システムが持つべき性質として，以下の 3 つがあげられている [4].

拡張性 ドメインを容易に追加することが可能．それぞれのドメインの処理を独立に考えればよい．

スケーラビリティ 多くのドメインを扱う場合でも，妥当な速度で処理することが可能．ユーザに対する応答性が，扱うドメインの数に大きく影響されない．

ユーザビリティ あるドメインのみを利用する場合，あたかも単一ドメインの音声対話システムを扱うかのように利用可能

2.5.2 ドメイン知識から対話パターンを自動生成

京都工繊大の荒木らは，XML ベースのデータベースから，セミオートに対話シナリオを生成するシステムを開発した [19]．WWW を用いたアプリケーションには，データベース探索タスクとして扱える会話システムがある．データベースにアクセスする部分を基本として，どのような検索質問を投げかければいいのかという部分を自動的に判別し，VoiceXML ファイルを生成するシステムを開発した．これにより，コンテンツ製作者はただデータベースを構築すればよいことになり，対話パターンまで考える手間を省くことが出来るようになった．

対話タスクの分類

電話応答やオンライントレード，銀行取引などといった，具体的な対話のタスクごとにあらかじめ分類をし，それぞれについて対話のライブラリを作る．3 つの段階にわけて，対話ライブラリは作られている．

- **Top level library** インタラクションの構造を記述するレベル．
- **Middle level library** VoiceXML について詳細に記述するレベル．
- **Bottom level library** データについて記述するレベル．このライブラリはすべての対話タスクにおいて共有される．

2.6 自由な対話を実現するための取り組み (チャットボット)

チャットボットは弱い Ai と呼ばれており，対話の流れからの推論や，特定の分野に対する知識などは利用しないシステムである．日本では人工無脳と呼ばれており，愛好家らによって様々なものが創り出されているが，論文になっているものは少ない．ここでは，英語のチャットボットである Eliza と Alice を紹介する．

2.6.1 Eliza

Eliza はアメリカの MIT で開発され，チューリングテストにおいても好成績を残したプログラムである [1]．Eliza の会話機能の中核部分は分解 decomp と再構成 reasmb である．decomp の部分にはルールが列挙されており，マッチングを行う．

次に，reasmb に並んでいる候補を前から順に選択してユーザに返している．ユーザの発言を返事に組み込んでいるため，Eliza がこちらの言うことを聞いているような感覚を覚える．あまり望まない方向に進みそうな単語を見つけると，やんわりと話の流れを変えようとする．そして，疑問文を返す割合を多くし，ユーザに考えさせようとする特徴がある．このような質問をして絶対に答えの評価をしないという挙動は心理カウンセラーの基本的なルールである．

2.6.2 A.L.I.C.E

A.L.I.C.E は，Dr. Richard S. Wallace によって開発されたプログラムである．ユーザとの対話における発話パターンおよびレスポンスを格納するために，AIML(Artificial Intelligence Mark-up Language) と呼ばれる XML を拡張した言語を使っている．現在までに実装された発話応答パターンとしては，地理や自然，および人に対する発話までカバーするおよそ 24,000 のパターンが実装されている．これは人手で 10 年かけて書きためたものである [20]．2004 年度の Chatterbox Challenge において，優勝を飾った実績を持つ．

AIML

AIML は，1995 年から 2000 年にかけて Alicebot コミュニティによって開発された [20]．以下の要素によって記述されている．

- **CATEGORIES** AIML の知識の基本単位はカテゴリーと呼ばれる．カテゴリーはそれぞれ，入力質問 (パターン)，出力となる答え (テンプレート)，およびオプションからなる．それらの記述には，自然言語，空白およびワイルドカードが使用可能である．また，テンプレートタグの中ではスクリプト言語を用いて記述することも可能である．

- **RECURSION** AIML では再帰構造を用いてデータを記述することが出来る．例えば複雑な条件文を他の条件文に書くときに，定義しておいたものを新たなカテゴリーの中で使うことがあげられる．

- **TARGETING** チャットボットの人格は，TARGETING といわれる学習プロセスを経て形成されていく．ユーザからマッチしない質問が発せられたときに，BotMaster が正しい答えを入力していくという手順で行われる．

- **CONTEXT** AIML では，that という表現は，最後に発言した文章を示す．これはカテゴリーの中で使える表現の一種である．

2.7 まとめ

本稿では対話システムの概要とその対話戦略について述べ，それぞれの戦略について実際に研究された技術の例を示した．

様々なタスクを扱うことが可能であり，自由なインタラクションが可能なシステム開発への取り組みとして，マルチドメイン対話システムと，ドメイン知識から対話パターンを生成する研究を紹介した．

このようにロボットやエージェントとのインタラクションには様々なタイプがあるが，本研究ではインタラクティブなプレゼンテーションを主に扱う．

第3章　インタラクティブなプレゼンテーションの実現とその課題

3.1　はじめに

近年，ロボットが注目され，さまざまな種類のロボットが研究，開発されている．特にヒューマノイドロボットは多くのモダリティを持ち，人間の代わりとしての活躍が期待されている．ロボットは実空間上で動作するため，視聴者への印象が強く，興味をひきやすい．ヒューマノイドロボットによるプレゼンテーションコンテンツを実現するためには，ロボットの制御用プログラムを用いることが必要である．制御用プログラムは下位レベルの操作を行うため，記述が複雑であり，専門的な知識を必要とする．そこで我々は，ヒューマノイドロボットによるプレゼンテーションコンテンツを容易に作成することを目指し，MPML-HR[22][23] の開発を行った．MPML-HR は下位レベルのインタフェースをあらかじめ構築しておき，ユーザに中位レベルで記述させることで，誰でも簡単にプレゼンテーションコンテンツを作成することを可能とする．

MPML-HR を用いることでロボットからの一方的なプレゼンテーションコンテンツの作成は可能となったが，インタラクションを含むコンテンツの作成はできなかった．インタラクションはプレゼンテーションの重要な要素であり，プレゼンテーションに関する質問に答えることでプレゼンテーションを分かり易くさせ，充実したコンテンツとすることができると考えられる．MPML-HR はアニメーションエージェントをプレゼンタとしたコンテンツを作成する記述言語 MPML[24][23] をヒューマノイドロボット用に拡張したものである．MPML は，特定の箇所において音声インタラクションを受け付ける機能を有していたが，任意の時点での音声入力を受け付けることはできなかった．

プレゼンテーションコンテンツに音声インタラクションを付加するための一つの方法として，各プレゼンテーションごとに特定ドメインに依存する音声対話システム (e.g. [27, 28]) を結合することが考えられる．しかし，専門的知識を持たないユーザがプレゼンテーションに対する答えを導き出せるようにシステムを調整することは容易ではない．それゆえ，VoiceXML[25] や XISL[26] などのインタラクションを対象とした記述言語の開発がされている．VoiceXML は電話によるインタラクションを想定した記述言語であり，音声とプッシュボタンを入力とし，音声を用いた出力を行うインタラクションの記述が可能である．しかし，プレゼンテー

ションでは音声のみならず，スクリーンへの資料の表示や動作などによる出力も必要であるため，VoiceXML では不十分である．XISL[26] は音声，マウス，キーボードによる入力が可能であり，音声合成，ブラウザ，アニメーションエージェントを用いた出力を行う．ヒューマノイドロボットと同様のジェスチャや移動，発話というモダリティを持つアニメーションエージェントによるインタラクションが可能であるため，XISL を用いることで目的とするインタラクションの記述は可能であるとも考えられる．しかし，現段階の XISL はアニメーションエージェントのみを対象としているため，ロボットへの拡張が必要である．また，XISL はプレゼンテーション用ではなく，インタラクション一般を目的として作られているため，記述量が多くなりがちであるといった問題もある．音声入力による質問も想定したロボットによるプレゼンテーションを実装した例もある [29]．この研究では，ヒューマノイドロボットを用いたプレゼンテーションを行う枠組みについて実装が行われているが，記述言語のような簡単な機構でプレゼンテーションコンテンツを作成するための実装には触れられていない．

これまでに，MPML-HR を拡張した，プレゼンテーションコンテンツのインタラクションを対象とした記述言語が提案されている．その際，従来の MPML-HR のロボットによるプレゼンテーションの生成とインタラクションの生成を合わせて MPML-HR ver. 3.0 と呼ばれるようになった．MPML-HR ver. 3.0 は従来の MPML-HR の「簡単な記述が可能」というメリットを受け継ぎ，充実したコンテンツを専門的な知識なしに誰でも簡単に記述できることを特徴とする．MPML-HR ver. 3.0 の割り込み機構には，対話ロボットのエキスパートモデルである RIME (Robot Intelligence based on Multiple Experts) [1][30] を用いた．それぞれのエキスパートは，対話の特定のタスクまたは物理的な動作命令に対して応答可能である．MPML-HR によるプレゼンテーションを実現するため，RIME に MPML-HR 用エキスパートを開発し，追加した．MPML-HR のスクリプトは，エキスパート知識に変換され，MPML-HR 用エキスパートを通してプレゼンテーションが実行される．インタラクションの際の音声認識に用いる言語モデルは，MPML-HR のスクリプトごとに作成する．MPML-HR エキスパートのための音声インタラクション機能は他のエキスパートと共通のものである．本システムの実装により，RIME の制御機構が MPML-HR ver. 3.0 の実現に適していることが判明した．また，RIME を用いることで，MPML-HR によるプレゼンテーションとタスク指向または非タスク指向の対話システムとを簡単に接続することが可能となった．

[1]RIME は文献 [30] では MEDBP と呼ばれていた．

3.2 インタラクション機構実現のためのプレゼンテーション記述言語

本節では，アニメーションエージェントを用いたプレゼンテーションコンテンツを作成するための MPML と，ヒューマノイドロボットを用いたプレゼンテーションコンテンツを作成する MPML-HR について概説する．次に，インタラクション機構を含む MPML-HR ver. 3.0 について説明する．

3.2.1 MPML および MPML-HR

はじめに

MPML(Multimodal Presentation Markup Language)[35] は，石塚研究室で開発を進めてきた，CG キャラクタを利用したマルチモーダルプレゼンテーションコンテンツ作成言語である．Microsoft Agent を用いて簡単にマルチモーダルなプレゼンテーションを作成することができる [9].

PC 上の Web ブラウザで表示することを前提に開発がすすめられてきたが，携帯電話上でもキャラクタ付きプレゼンテーションを行うコンテンツ記述が可能な MPML-Mobile Version も開発された．また，3D の全身を持ったキャラクタが 3 次元空間で動作し，プレゼンテーションを行う MPML-VR の開発も行われた．

このように，MPML はその機能や用途から多岐に渡って開発され，様々なバージョンがあるが，ここでは共通の特徴を中心に言及する．

概要

計算機の処理能力の向上と低価格化により利用者が増加し，専門家のみならず，主婦や低年齢の人々まで利用者層が拡大している．それに伴い，GUI 中心のインタフェースから，聴覚，視覚，触覚など様々な認知処理様式を利用するマルチモーダルなインタフェース技術への需要が高まっている．マルチモーダルなインタフェース実現のために，擬人化エージェントを利用したインタフェース，コンテンツの研究が進められてきている．計算機の処理能力向上でテキスト情報の処理に加え，音声，動画等の大容量の情報を扱うことが可能になったため，研究レベルに止まらず，実際に擬人化エージェントを利用したコンテンツ，インタフェースを個人用計算機で扱うことができる．モーションキャプチャやビデオカメラ等のインタフェース機器の充実化により，分野を限定した場合に関しては人工知能技術を利用したエージェントとの自然言語による対話も可能で，マルチモーダルなインタフェースを利用できる機会は多くなってきていると言える．擬人化エージェントには，文書情報のみならず音声，身振りや手振りを利用して伝達する情報量を増やすだけでなく，情報取得者の興味を引き付ける役割や，人間に近い姿で親しみ

を感じさせることで情報取得者の心理的負担を軽減する役割がある．しかし，擬人化エージェントを利用したコンテンツ作成には，アニメーション合成，音声合成，アニメーションと音声の同期をプログラムしなければならないため，プログラミング熟練者以外のコンテンツ作成者が作成し，公開することは容易ではない．

そこで，“誰も” が “容易” に “魅力的” なマルチモーダルなプレゼンテーションコンテンツを作成できるようにすることを目的として MPML(Multimodal Presentation Markup Language) が開発されてきた．MPML は XML(Extensible Markup Language) に準拠したコンテンツ記述言語で，HTML で Web ページを記述するのと同様に，タグ付けで内容を記述することで擬人化エージェントを利用したプレゼンテーションを作成することができる．

3.2.2 MPML の特徴

MPML はこれまで，最初に開発された MPMLVer1.0[36] から使用できるキャラクタやその機能によって様々なバージョンが開発されてきた．MPMLVer2.0e[37] からは，感情制御のためのタグが導入され，キャラクタの動作や音声の感情制御を容易に記述することが可能となった．MPML では，主に Microsoft Agent[42] をプレゼンターとして用いているが，異なった CG アニメーションキャラクタを用いるバージョンも作られている．MPML-VR[38] は，MPML for Virtual Reality を表し，VRML の技術を使用して 3 次元空間の中で 3 次元のキャラクタを制御し，プレゼンテーションを行わせることができる．MPML-mobile[39] では，3 次元のキャラクタを使ったコンテンツを携帯電話上で実行することができる．MPML-mobile では，音声を出力する代わりにテキストが表示される．

このように，MPML には，様々なバージョンが存在し，それぞれ独自の特徴ももつが，共通する大きな特徴がいくつか存在する．ここでは，共通する特徴について言及する．

容易な記述性　MPML は，XML(Extensible Markup Language) の規格に準拠している．現在，MPML で用いられるタグの種類は 20 種類程度あるが，6 種類程度のタグでもキャラクタを利用したプレゼンテーションを作成することが可能である．このため，HTML で Web ページを作成した経験がある人はもちろん，無い人でも容易にマルチモーダルなプレゼンテーションコンテンツを作成できる．また，熟練者の場合は用意された様々なタグ，それらそれぞれに存在する属性を調整することで，詳細にわたりプレゼンテーションを設定できる．

システム非依存　擬人化エージェントを利用したマルチモーダルなコンテンツは，専用ブラウザや専用エージェントを利用し，動作環境によっては実行できたり，実行できなかったりしてしまう場合が多い．MPML は，ブラウザやエージェントの

種類，オーディオプレイヤー等に依存することがなく，ある環境で記述したプレゼンテーションコンテンツを異なる環境で実行することが可能である．

メディア同期　音楽，画像，動画等の様々なメディアを再生することは，プレゼンテーションを魅力的なものにする上で効果的である．この点に関し，W3C から，WWW 上で複数のメディアデータの表示を制御するための記述言語である SMIL(Synchronized Multimedia Integration Language) の勧告が公開されている．MPML はこの仕様を参考にし，背景画像が自動的に次々に切り替わっていくだけではなく，メディア同期を行うための機能を有している．

豊富な制御機能　MPML は，マルチモーダルなプレゼンテーションという限られた条件下での人とのコミュニケーションを対象としているため，プレゼンテーションで人間が使用する動作と同程度の種類の動作を擬人化エージェントに行わせることができれば事足りる．しかし，実際には，プレゼンテーションで有効であると考えられる “おじぎ”，“指差し”，“さよなら” といった身振り，手振りや “笑顔”，“困惑” といった表情の他にも多数の動作をサポートしている．プレゼンターとして登場するキャラクタは，画面上を自由に移動し，身振り，手振り，音声，吹き出しによる文書情報を用いて視聴者に何らかの紹介，説明等を行う．コンテンツ作成者は，移動，身振り，手振り，発話といったキャラクタの制御を，タグ付けによる記述を行うことで簡単に行うことができる．

メタデータ記述　MPML は XML に準拠しているため，意味のあるデータを記述できるという機能をもつ．著者名，コンテンツ作成日，プレゼンテーションのタイトル等のメタデータをコンテンツに記述しておくことができる．これらのメタデータはプレゼンテーションの進行には影響を与えない．しかし，公開されている情報が膨大になっている現在，膨大なコンテンツをデータベース化した際，メタデータによって検索が容易になり，必要な情報を効率的に取り出せるようにできると考えられる．

感情表現　MPML では OCC モデル [43][44] で定義された感情を使用することができる．OCC モデルは，感情誘発と関連付けされる多様な要素を考慮した感情モデルとして，現在多くの感情理論の基礎となっている．OCC モデルでは，イベントの結果，エージェントの行動，対象の様子を社会における三つの感情を誘発する要素とし，それに基づいて 22 種類の感情が定義されている．MPML にはこの 22 種類の感情が用意されており，感情を指定するだけで，プレゼンターとなるキャラクタの身振り，手振りや移動，発声の仕方が変化する．感情表現豊かなプレゼンテーションが，感情をタグ付け指定するだけで容易に実現可能である．

プレゼンターとなるキャラクタの表情，動作，発声の音量，ピッチ，強調などの組み合わせで感情表現が実現される．コンテンツ内の記述においては，命令を，感

情を指定したタグで修飾することによって感情が反映される．修飾できるタグは <play>，<speak>，<move> で，それぞれキャラクタの表情や動作を指定する命令，発声を指定する命令，画面上の移動を指定する命令である．感情は <emotion> で記述し，type 属性に OCC モデルで定義された 22 種類の感情に加えて “neutral” を指定することができる．“neutral” は何も感情による強調がない状態のことである．ここで挙げた 23 種類の感情以外にも自由に感情の種類を拡張できるように設計されている．

ページ概念　幾つかのページで全体が構成されるプレゼンテーションを扱うことができる．<page> によりページ毎にキャラクタの動作や発声の制御を行うことができる．<page> の ref 属性によって指定された一つ一つの背景画像は，OHP や PowerPoint によるプレゼンテーションにおける，一枚の OHP やスライドに相当する．ref によって指定された画像が背景として表示され，<page> </page> で囲まれた部分のキャラクタ制御を実行し終わると自動的に次の <page> の ref で指定された画像に背景が切り替わる．このページ概念の導入により，製作者は提供したい情報を，一つの画像に無理に詰め込む必要がなく，組織的に，前後のバランスを考えた，より分かりやすいプレゼンテーションを作成することができる．

3.2.3　MPML の言語仕様

ここでは，MPML の言語仕様について紹介する．

MPML のドキュメント

MPML の言語仕様は，様々なタグ要素の構造木で表すことができる．全ての要素の祖先要素，つまりルートとなるのが <mpml> である．要素 <mpml> は属性として id を持つ．id は要素同士の区別を行うために用いられる識別値であり，ほとんどの要素に記載することができる．また，子要素として <head> と <body> を持つことができる．以下で各要素について順次解説を行う．

ヘッダ

ドキュメントのヘッダは <head> と </head> によって囲まれる．この部分には，プレゼンテーションに関するメタデータ及び，レイアウトに関する情報を記述することが可能である．メタデータは要素 <rec> に，レイアウトは要素 <layout> に記述される．

メタデータ記述　プレゼンテーションコンテンツの作成者は，要素 <rec> に，作成したプレゼンテーション全般に関するデータを記載しておくことができる．要素 <meta> は空要素であるが，属性を用いてそのタグ一つに一組のメタデータを記載しておくことができる．要素 <abst> はテキストデータを要素内容に持ち，プレゼンテーションの概要が記載できる．

レイアウト記述　要素 <layout> は，プレゼンテーションコンテンツのレイアウト情報を記載するためのタグである．任意の書式でレイアウト記述ができるが，デフォルトでは，MPML 特有の書式になり，<root-layout> 及び，<region> を子要素に持つ．プレゼンテーションを行うルートウィンドウに関する性質を，要素 <root-layout> で設定できる．要素 <region> の属性を使用し，“点” もしくは長方形型の “範囲” としてリージョン情報を記載できる．

エージェント記述　要素 <agent> で，プレゼンテーションを行うキャラクタエージェントの記載，設定を行うことが可能である．要素 <agent> で定義したエージェントを，要素 <move>，要素 <speak>，要素 <play> などで参照する．つまり，定義されたエージェントが移動，ジェスチャ，発話を行う．

ドキュメントのボディ

キャラクタエージェントを利用したプレゼンテーションの実行に関わる内容が，要素 <body> と </body> で囲まれる範囲に記述される．要素 <body> の直下に要素 <page> が存在する．全ての記述はページ単位で行われる．要素 <page> は暗黙に要素 <seq> の定義を含むため，その子要素の実行はシーケンシャルに行われる．

エージェントの動作記述　要素 <move> でキャラクタエージェントを移動させることができる．すでに設定したリージョンへの移動に加え，移動位置を属性 x，y で座標指定して移動することも可能である．要素 <speak>，</speak> で囲む部分にはテキストを記述する．ブラウザまたはプレイヤは，音声合成エンジンを用いてキャラクタエージェントにそのテキストの内容を発話させる．

また，要素 <play> によって，キャラクタエージェントに様々な動作をさせることができる．動作は属性 act で指定する．指定可能な動作数はキャラクタによっても異なるが 60 前後の動作を指定することが可能である．

感情装飾　プレゼンターとして登場するキャラクタエージェントの動作を指定できる <play>，<move>，<speak> の 3 種類の要素がある．その中の <speak> と <play> の 2 種は，<emotion> で修飾すること，つまり，<emotion>，</emotion> で囲むことで，感情の指定，感情を反映した動作を行うことができる．

<emotion type = ”pride”>
　　<speak>
　　　hello!
　　<speak>
</emotion>
例えば上記のように，<speak> に “pride” の感情を修飾すると，キャラクタが手を振りながら，文書の先頭を強調して発話するように，通常の発話から感情が入った発話に変化する．

メディア同期　要素 <par> の子要素として記述された部分は，そのドキュメント中の順序に関係なく，並列に実行される．例えば，以下の記述では，キャラクタエージェントは，“greet” の記述によって指定された挨拶動作を行い始めてから 2 秒後に発話を行う．

<par>
　　<play act = “greet”/ >
　　< speak begin = “2s”>
　　　hello!
　　</speak>
</par>

　また，<seq> の子要素として記述された部分は，そのドキュメント中の順序にしたがって，シーケンシャルに実行される．例えば以下の記述では，キャラクタエージェントは挨拶動作を完了してから 2 秒後に発話を行う．

<seq>
　　<paly act = “greet”/ >
　　<speak begin = “2s”>
　　　hello!
　　</speak>
</seq>
また，メディアオブジェクトの再生は要素 <ref> を持ちいてロケーションを指定することによって行われる．

まとめ

　MPML はスクリーンエージェントを用いたプレゼンテーションコンテンツを誰でも簡単に記述することのできる記述言語である．中位レベルの言語であるため，下位レベルのエージェント制御プログラムとは独立である．したがって，ユーザは下位レベルの制御知識なくコンテンツを作ることが可能である．また，XML に

準拠しており，エージェントの位置，動作，ジェスチャ，感情表現などを制御する豊富な関数が用意されている．

3.2.4 MPML-HR の特徴

MPML-HR は MPML をヒューマノイドロボット用に拡張したものである．拡張には例えば point タグなどがある．これは，実空間に存在するロボットがスクリーンのある座標を指示するための命令である．スクリーン上に存在するアニメーションエージェントは移動命令である move タグを用いて特定の座標を指示可能であるが，ロボットでは別の命令が必要となる．このように，エージェントの存在する空間の違いを吸収するための拡張が MPML-HR ではなされている．

3.2.5 MPML-HR へのインタラクション機能の導入

本研究では，プレゼンテーションにおける視聴者からの音声割り込みとして，説明の省略や，前に話した内容の再度の説明，あらかじめ想定した説明に応える機構を実現することを目的とする．これらのインタラクションを実現するため，コンテンツ作成者が発話パターンや次のロボットの動作を記述できるよう，MPML-HR を拡張する必要がある．

本研究の目的とするインタラクションは，プレゼンテーションの説明箇所を遷移させることで実現できる．プレゼンテーション内の特定の箇所に遷移するため，コンテンツをいくつかのパートに分け，各パートの先頭に遷移できるようにする．コンテンツを分けるために，MPML-HR の page タグを用いた．page タグはスクリーンの 1 つのスライドごとに作られる．プレゼンテーションでは 1 つのスライドが 1 つの話題に対応すると考えられるため，視聴者からの割り込みに対して遷移するのに適した単位である．

視聴者からの割り込みの実現においては音声認識誤りへの注意が必要である．音声認識誤りが起こり，正解とは異なる発話割り込みと解釈してしまうと，プレゼンテーションは視聴者の意図しない場所に遷移する．このようなことが頻繁に起これば，使い勝手が悪くなり，インタラクション機能を導入するメリットが失われてしまう．そこで，音声認識誤りに対処するため，二つの機構を取り入れた．一つは，音声認識結果に対する信頼度が低い場合に，発話者に再度の発話を依頼する (聞き返し)，あるいはその解釈内容が正しいか Yes または No による回答を要求 (確認) するものである．もう一つは，間違った遷移が行われた場合，視聴者からの指摘により，遷移前のプレゼンテーションに戻す機構である．

図 3.1 に MPML-HR ver. 3.0 のサンプルスクリプトを示す．スクリプトは二つの部分からなる．一つは body で囲まれた範囲であり，プレゼンテーションコンテンツ本体を記述する．もう一つは interaction で囲まれた範囲であり (Fig. 3.1 (1))，インタラクションに関する記述をする．grammar タグ (図 3.1 (2)) は認識文法を記

述する．発話割り込みが認識文法の一つと一致すると，ロボットは grammar タグの内部に記述された内容を実行する．recog タグの内部には grammar タグを記述することができ (e.g.(3))，これらの認識文法を受け付ける範囲を page 属性によって指定することができる．つまり，ページ範囲の指定された認識文法はその範囲内でのみ有効であり，recog タグによって囲まれていない認識文法はプレゼンテーション全体で有効である．jump タグと do タグは発話割り込みを検出した際の実行内容を決定する．これらはどちらも，あるページに遷移するための命令である．また，複数の jump タグまたは do タグを記述することにより，一つの割り込みから複数ページを実行することも可能である．但し，複数ページを指定する際には，grammar 内には jump のみまたは do のみで記述する必要がある．jump タグでページが指定された場合，そのページに遷移し，遷移先ページが終了した際にはもとのプレゼンテーションには戻らない．一方，do タグでページが指定された場合，そのページに遷移し，遷移先ページが終了した際にもとのプレゼンテーションが続けられる．goback タグは，認識誤りによりプレゼンテーションコンテンツが間違った場所に遷移してしまった場合に，もとのコンテンツに戻るための命令である．この場合も，grammar タグを用いて，goback を受け付ける際の認識文法の記述を行う．

grammar タグでは，view 属性の記述により，受け付け可能な発話内容のスクリーンへの表示を行うことができる．MPML-HR のプレゼンテーションでは，画面にスライドが表示されるが，右サイドに受け付け可能な発話割り込みを，画面下にはロボットの発話内容を表示するスペースが設けられている．grammar タグで view 属性が記述された場合にはこの内容が右サイドに表示されるが，省略された場合には認識文法である recgram の内容が表示される．

発話割り込みにより認識された音声認識結果はその信頼度の大きさに応じて，棄却，聞き返し，確認，受理のいずれかがなされる．信頼度が低かった場合，誤認識である可能性が高いため，発話内容は棄却され，ロボットは何ら動作を行わない．信頼度が高かった場合には受理となる．この場合，正しく認識できている可能性が高いため，ロボットは jump, do, goback で指定される動作を行う．聞き返し，確認は棄却と受理の間の信頼度であった場合になされる．その中でも信頼度が低い場合は聞き返し，信頼度が高い場合は確認となる．聞き返しでは，ロボットが “何かおっしゃいましたか？” と聞き返すことにより，発話者に対して再度の発話を要求する．何らかの発話割り込みが入った可能性が高いが，何と発話したかまでは正しく音声認識できていないような場合に有効である．確認では，音声認識内容に対して，その発話内容を行ったかどうかの聞き返しを行う．例えば，発話者が “明日の天気を説明して” といった場合，ロボットは “明日の天気を説明しますか？” と問い返す．確認の際は，常に Yes か No の答えが可能な形で問い返しを行うため，受け付け可能な認識文法が沢山ある場合と比べると二者択一の音声認識は正しくなされる可能性が高い．したがって，発話内容を正しく検出できている可能性が高い場合，聞き返しと比較して有効である．

棄却，聞き返し，確認，受理の境界を決定する閾値はあらかじめシステムで定義されている．しかし，より柔軟な割り込み機構を実現するため，この閾値を変更することも可能である．変更には，confidence-thresholds(図 3.1 (4)) を用いて行う．棄却と聞き返しの境界には reject，聞き返しと確認の境界には ask-back，確認と受理の境界には accept 属性を指定することにより閾値の変更がなされる．図 3.2 に示すように，reject, ask-back, accept の任意の値を同じにすることにより，聞き返し，確認を行わないよう設定することも可能である．また，確認を行う際には，grammar タグの属性として confirm を記述することが必要である．この記述は，確認の際にロボットが発話する内容を定義するものである．この confirm 属性が省略された場合には，確認は行われず，本来確認が行われるべき範囲の信頼度であった場合，聞き返し (聞き返しが行われない閾値設定では棄却) が行われる．閾値を変化させることで，さまざまな利用が可能である．例えば，認識文法が少ない場合には正しく認識できる可能性が高いため，聞き返しを省略することも有効である．また，コンテンツの最初の段階で，終了地点に遷移するような発話は想定し難いため，発話される可能性に応じて個別に閾値の設定を行うことも可能である．

図 3.3 に図 3.1 の MPML-HR ver. 3.0 サンプルスクリプトに関するインタラクション例を示す．

3.3 マルチエキスパートモデルに基づく実装

3.3.1 システム概要

図 3.4 に MPML-HR ver. 3.0 のシステム構成を示す．MPML-HR ver. 3.0 は対話ロボットの記号レベルの対話行動制御モジュールによって実現されている．対話・行動制御サブシステムはユーザからの発話を理解し，最適な次の行動を決定した後に動作を行う．本システムでは，出力形式の一単位を MADL (Multimodal Action Description Language) として表現する．この MADL には発話テキスト (e.g. “こんにちは ASIMO”)，身体を用いた動作 (e.g. “お辞儀” や “太郎に近づく”)，スライドの表示などが含まれる．そして，MADL の実行時には，全ての動作が同期して開始される．対話・行動制御サブシステムは，音声認識エンジン，音声合成器，ASIMO を制御するハードウェアコントローラと接続されている．

対話・行動制御サブシステムは対話ロボットの会話・行動のモデルである RIME (Robot Intelligence based on Multiple Experts)[30] を基に構築した．このモデルでは，エキスパートと呼ぶ，特定のドメインの対話や特定の物理行動など遂行するためのモジュールを用いる．MPML のプレゼンテーションを実現するため，MPML 用のエキスパートを開発した．(以後，MPML-HR ver. 3.0 の代わりに MPML と表記する．) MPML スクリプトは MPML コンパイラによって MPML エキスパートで用いる中間言語に変換される．

3.3.2 マルチエキスパートに基づく対話・行動制御サブシステム

RIME で用いるエキスパートはサブタスクごとに用意される必要がある．ロボットがサブタスクを実行する際には，対応するエキスパートは行動の決定を行う．ユーザからの発話が受け取られると，エキスパートはアクティベートされ，音声認識結果と文意を理解する．エキスパートはオブジェクト指向の枠組みのなかの一種のオブジェクトである．それぞれのエキスパートは内部状態を持ち，さまざまな情報を保持する．

システムには，複数のエキスパートを利用して動作するために，3 つのモジュールがある．*understanding* モジュールは音声認識結果をエキスパートに送信し，最適なエキスパートをアクティベートする．*action selection* モジュールは選択されたエキスパートに対し，次の動作の要求を行う．*task planning* モジュールはロボットが動作している間，どのエキスパートをアクティベートするか決定する．これら 3 つのモジュールは発話割り込みを扱うために並列で動作する．

それぞれのエキスパートには内部状態にアクセスするためのメソッドが必要である．ここでは MPML エキスパートに関連するもののみ説明する．*initialize* メソッドはエキスパートが呼び出されたときに生成され，内部状態を初期化する．*understand* メソッドは音声認識結果を受け取った際に *understanding* モジュールから呼び出され，音声認識結果に基づいて情報を更新する．*select-action* メソッドは，*action selection* モジュールから継続的に呼び出され，発話割り込み待ちの状態でなければコンテンツの情報に基づいて一つ動作を出力する．

3.3.3 MPML コンパイラ

MPML コンパイラは MPML スクリプトを中間記述に変換する．中間言語は複数の `block` 要素から構成され，一つの `block` 要素は MPML スクリプトの `body` 中の `page` に対応する．本編のコンテンツに加え，`jump` タグや `do` タグで指定されたページに対しても `block` が生成されるため，一つの `page` からは，最大で三つのの `block` が生成される．図 3.5 に図 3.1 の MPML スクリプトをコンパイルした中間記述の例を示す．

`handle-interruption` 要素は認識文法ごとに生成される．この中には，文法 ID，遷移先 ID および確認発話の内容が定義されている．遷移先 ID は複数の記述がなされるよう設計されており，遷移先 ID の境界はピリオド (.) によって表される．

`madl` 要素は一つの MADL を表し，対話や行動の制御を表現する．`show-slide` 要素はスクリーン上に指定されたスライドを表示する命令である．また，`show-text` 要素は指定されたテキストをスクリーンに表示する命令である．テキストを表示する位置は，受け付け可能な認識文法を表示するスライド右側 (side) と，ロボットの発話を表示するスライド下側 (bottom) の二種類が存在する．`utterance` 要素，

play 要素，go-to 要素はそれぞれ，発話，ジェスチャ，特定の位置への移動を表す．MPML での感情表現は，MADL の emotion 要素によって記述される．

中間記述への変換時には，文法リストファイルと文法記述ファイルが生成される．文法リストファイルは，認識可能な文法のリストが記述され，これを用いて認識文法が生成される．

3.3.4 MPML エキスパート

本来，MPML エキスパートはプレゼンテーションごとに作られる．しかし，コンテンツの内容以外の枠組みは共通であるため，MPML エキスパートはメソッドを共有し，中間記述と認識文法のみが異なるよう構成されている．MPML-HR はオブジェクト指向を用いることで，同じクラスとして表現されている．MPML エキスパートのクラスについて以下説明する．

initialize メソッドは最初に使われる中間記述を MADL キューに追加する．キューに使うデータの形式は，MADL と割り込みリスト (文法 ID と遷移先から構成される) をひとまとめにしたものである．

select-action メソッドは以下の処理を行う．まず，MADL キューからポップする．ポップされた割り込みリストを基に，アクティブな文法 ID リストをセットする．そして，ロボットの動作を行うため，MADL を返す．

understand メソッドは音声認識結果に対する信頼度の値が閾値を超えているかチェックする．閾値は 3.2.5 節で示した reject, ask-back, accept の 3 つがある．信頼度が reject 以下であれば，発話割り込みによる動作は何も行われない．信頼度が reject より大きく，ask-back より小さい場合には，聞き返しを行うため，聞き返し用の発話内容の MADL をキューの一番先頭にプッシュする．信頼度が ask-back より大きく，accept より小さい場合には，確認を行うため，確認用の発話内容の MADL をキューの一番先頭にプッシュする．なお，確認の際には，発話者の発話は Yes または No の意味合いを想定するため，その 2 つの認識文法のみを受け付けるよう，アクティブな文法 ID のセットを行う．信頼度が accept より大きな場合には，その発話内容により jump, do または goback の遷移が行われる．

3.4 インタラクティブなプレゼンテーションにおける課題

MPML-HR の拡張に際して認識の精度に応じて受理・確認・聞き返し・棄却を行う機構を取り入れた．これにより，システムはより柔軟な対応を行うことが出来るようになった．

しかしながら，視聴者が，どのような発話内容やタイミングで質問や要求を行ってよいかがわからないという問題がある．さらに，音声認識自体の精度向上も考えていかなければならない．

このように，インタラクティブなプレゼンテーションにおいては，割り込み音声の認識が問題となる．

3.5 まとめ

本章ではインタラクション機能を有するロボットのプレゼンテーション記述言語 MPML-HR ver. 3.0 の設計と実装を行った．MPML-HR ver. 3.0 を用いることにより，視聴者からの説明の省略，再度の説明，想定される内容の説明を行うインタラクションを専門的な知識なしに作成することが可能となった．MPML-HR ver. 3.0 の構築には，マルチエキスパートに基づくロボット対話システムコントローラが役立つと考え，これをベースにした開発を行った．現段階の実装はプレゼンタにロボットを用いたものであるが，アニメーションエージェントをプレゼンタとした同様の機構を実装することも可能である．

そして，インタラクティブなプレゼンテーションにおいては，割り込み音声の認識が問題となることを述べた．

```
<?xml version="1.0" encoding="UTF-8"?>

<mpml>
<body>
  <page ref=“./slide.ppt#1” id=“明日の天気">
    <synch>
      <speak>明日の東京は晴れでしょう</speak>
      <play act="victory" />
    </synch>
  </page>
  <page ref=“./slide.ppt#2” id=“明後日の天気">
    <speak>明後日の東京は雨でしょう</speak>
  </page>
  <end />
  <page ref=“./slide.ppt#3” id=“昨日の天気">
    <speak>昨日の東京は曇り空でしたね</speak>
  </page>
</body>
<interaction> ……(1)
  <grammar recgram=“[それは] ちがうよ“
           confirm=“遷移前に戻りますか？” view=“違うよ”> ……..(2)
    <goback />
  </grammar>
  <grammar recgram=“きのう [のてんき]“
          confirm=“昨日の天気ですか？” view=“昨日の天気”>
    <do page=“昨日の天気” />
  </grammar>
  <grammar recgram=“あさってときのう [のてんき]”
          confirm=“明後日と昨日の天気ですか？” view=“明後日と昨日の天気”>
    <do page=“明後日の天気” />
    <do page=“昨日の天気” />
  </grammar>
  <recog page=“明後日の天気“> ……..…(3)
     <grammar recgram=“あした [のてんき]“
              confirm=“明日の天気ですか？” view=“明日の天気”>
       <confidence-thresholds reject=“0.6”  ask-back=“0.8” accept=“0.9” />.....(4)
       <jump page=“明日の天気”  />
    </grammar>
  </recog>
</interaction>
</mpml>
```

図 3.1: MPML-HR ver. 3.0 Script の例

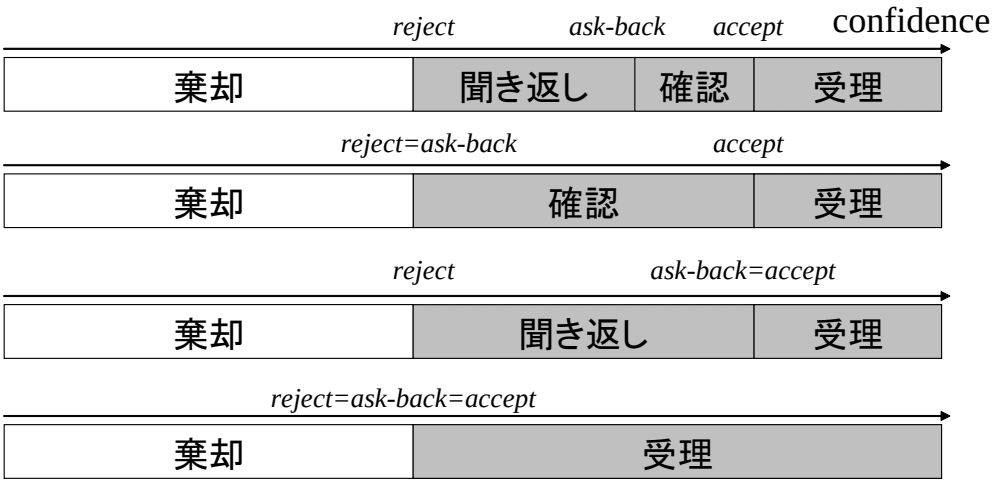

図 3.2: Confidence threshold の設定例

Robot: 明日の東京は…
User: 昨日の天気は？
Robot: はい, (Go to “明後日の天気”)
Robot: 明後日の東京は …
User: 違うよ
Robot: 明日の東京は… (Go back to “明日の天気”)
User: 昨日の天気は？
Robot: はい, (Go to “last”)
Robot: 昨日の東京は曇り空でしたね
Robot: 明日の東京は, … (Return to “明日の天気”)
…
Robot: 明後日の東京は…
User: 明日の天気
Robot: 何か言いましたか？ (Ask back)
User: 明日の天気
Robot: はい, (Jump to “明日の天気”)
Robot: 明日の東京は…
…….
Robot: 明後日の東京は雨でしょう

図 3.3: システムとユーザの対話例

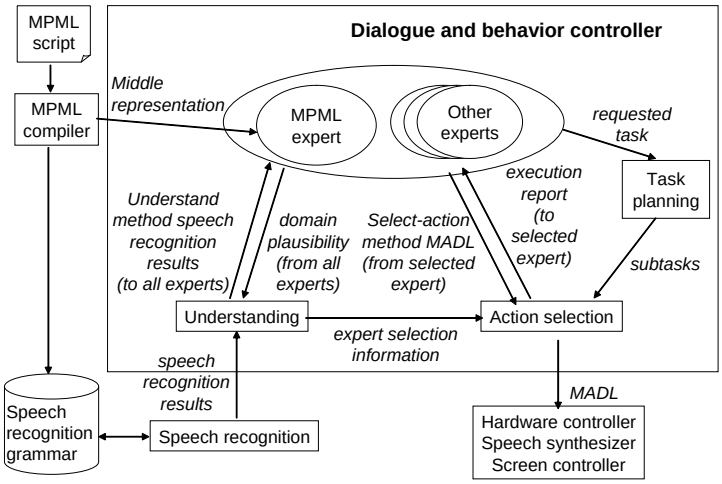

図 3.4: システム構成

```
<block id="1" next="2">
 <interruption-handle recgram="gram_0_1" goto="goback." confirm="遷移前に戻りますか？"></interruption-handle>
 <interruption-handle recgram="gram_0_2" goto="3_r." confirm="昨日の天気ですか？"></interruption-handle>
 <interruption-handle recgram="gram_0_3" goto="2_r.3_r." confirm="明後日と昨日の天気ですか？"></interruption-handle>
 <madl>
  <show-slide uri="./slide.ppt#1" />
  <show-text view="side">昨日の天気¥n明後日と昨日の天気¥n</show-text>
 </madl>
 <madl>
  <utterance>明日の東京は晴れでしょう</utterance>
  <show-text view="bottom">明日の東京は晴れでしょう</show-text>
  <play name="victory" />
 </madl>
</block>
```

図 3.5: 中間記述の例

図 3.6: ASIMO とのインタラクション

第4章　インタラクティブなプレゼンテーションでのユーザ発話の自然な制限のための複数エージェントの利用

4.1　はじめに

ロボットやアニメーションエージェントの応用のひとつとして，自動プレゼンテーションが注目されている．我々はこれまで，ユーザからの音声入力をうけつけてプレゼンテーションの内容を変える，インタラクティブなプレゼンテーションエージェント/ロボットシステムを開発してきた [32]．しかしながら，視聴者が，どのような発話内容やタイミングで質問や要求を行ってよいかがわからないという問題があった．視聴者の発話を認識・理解するためには，あらかじめ設計者が構築した知識を元にするため，それにマッチしなければシステムがうまく理解できない．

この問題に対処するために，音声理解をロバストにし，設計者の予測からはずれた発話内容やタイミングでも理解できるようにすることや，予測からはずれた発話を検出してヘルプを出す方法 [33] などが試みられている．

本章では，これらとは異なるアプローチとして，プレゼンテーションを行うエージェントのほかに，もうひとつのエージェントを用い，質問や要求をデモンストレーションさせることにより，暗示的に，視聴者に発話の仕方を教える方法を提案する．この方法により，視聴者が質問・要求を行う際に，エージェントの発話と類似した発話を行う可能性が高くなり，結果として発話の認識・理解の精度が高まることが期待される．

4.2 インタラクティブなプレゼンテーションシステムにおける複数エージェントの利用

4.2.1 MPML-HR ver.3 システム

提案手法は，これまで開発してきたインタラクティブプレゼンテーション記述言語 MPML-HR (Multi-modal Presentation Markup Language for Humanoid Robots) ver.3（以下 MPML-HR3）とその実行システムを拡張することで実現している．MPML-HR3 実行システムは，MPML-HR3 で記述されたスクリプトに基づいて，ロボットやエージェントによるインタラクティブなプレゼンテーションを行うことができる．スライドを表示し，音声とジェスチャーを用いて，プレゼンテーションを進めて行く．視聴者は，インタラクション中に,「... を説明して」,「前に戻って」,「終わっていいよ」など，別の内容に遷移したりプレゼンテーションを終了したりするような発話をすることができる．どのような発話に対して，どのような変更，すなわちプレゼンテーション内容の遷移を行うかは，あらかじめ，スクリプトの中に書かれている．発話パターンはネットワーク文法で記述する．質問・要求発話の認識結果の信頼度に応じて，そのまま遷移したり，問い返したり，無視したりすることができる．

4.2.2 複数エージェントの利用

視聴者に暗示的に発話の仕方を教えるために，エージェントをもう一体用い，質問・要求発話の実演を行わせる．これは, プレゼンテーションを行っている方のエージェントの発話に，もう一体のエージェントが反応して質問・要求を行うような規則を記述することで可能にする．エージェントによる質問・要求が入力されると，視聴者からの質問・要求へ反応するのと同様に，プレゼンテーション内容の遷移を行う．このような拡張を行ったプレゼンテーション記述言語を MPML-HR ver.4（MPML-HR4）と呼ぶこととした．

4.3 実験

提案法の有効性の確認を目的に，予備的に被験者実験を行った．被験者がプレゼンテーションを見る際に，説明をするエージェントのみの場合と，それに加えて質問の手本を見せるエージェントを用いた場合に，被験者の質問・要求発話にどのような差異が生じるかを以下に示すとおりの方法で調べた．

4.3.1 方法

被験者は，日本語を母国語とする男性 12 名，女性 11 名 (年齢 20 歳-62 歳) である．実験では，2 体のマイクロソフトエージェントを用いた．プレゼンテーションを行うエージェント (説明エージェント) と，質問の手本を見せるエージェント (質問エージェント) について，以下のような 2 条件を設定した．

質問エージェントなし 説明エージェントのみが登場してプレゼンテーションが行われる

質問エージェントあり 説明エージェントおよび質問エージェントが登場してプレゼンテーションが行われる

実験はそれぞれ 2 回ずつ行い，被験者 23 名を [質問エージェントなし]，[質問エージェントあり] の順に行うグループと [質問エージェントあり],[質問エージェントなし] の順に行うグループに男女比及び人数がほぼ等しくなるよう配分した．両条件とも提示する情報の総量（具体的には説明するページの数）が同じになるようにした．用いたプレゼンテーションの内容は，インターネットのウェブサイト (goo,google) の説明である．質問エージェントありの条件では，質問エージェントが「〜について教えてください」という文型で 3 回の質問を行った．この文型は，視聴者の発話の認識文法に含まれている．

4.3.2 手続き

被験者に，対話システムから情報を聞き出すという目的と，音声発話により対話システムが反応することを教示した．エージェントは音声発話を完全に理解することはできず，認識されない場合は話す内容やタイミングを変えるとよいということも教示した．また，質問エージェントの発話を聞いてから質問してもらうため，一通り説明が終わってから質問をするように教示した．

被験者に与えた課題は以下に示す通りである．

1. サイト名の由来を聞き取る
2. サービスの内容を 2 つ聞き出す

実験では，被験者の質問と考えられる発話が質問エージェントの発話と文型が似ているかどうか，及び音声認識に成功したかどうかを観察した．

4.3.3 仮説と予測

質問エージェントがある場合，質問エージェントの発話に影響され，質問エージェントの発話と同じ文型で話す被験者が多いことが予測される．2 回のプレゼ

ンテーションのうちの初めのプレゼンテーションの最初の質問に注目することでエージェントの影響を最もよく観測することができると考えられる．また，被験者は自分の発話がうまく認識されたかどうかによって，発話パターンを変えると考えられる．うまく認識された場合は，同じ発話パターンを繰り返し，それ以外の場合は，別のパターンを試すと考えられる．このとき，複数のエージェントを用いた場合，エージェントの質問発話パターンに移る可能性が高いと考えられる．

4.4 結果

2 回のプレゼンテーションのうちの初めのプレゼンテーションの最初の質問と，それがうまく認識されなかった場合のその後の質問を分析した結果を表 5.1 に示す．また，システムとユーザの対話例を図 4.1 に示す．

User: グーグルの名前の由来について教えてください
Robot: はい(Go to “グーグルの由来”)
Robot: グーグルという名前は、10 の 100 乗を表すグーゴルに由来します
…
User: グーグルのサービス
Robot: サービス例についてですか
User: グーグルの、乗り換え案内のサービスについて教えてください
Robot: はい(Go to “乗り換え案内”)
Robot: Google の乗り換え案内は，地図と連動しています
Robot: 路線が地図上に表示されます

図 4.1: システムとユーザの対話例

ここで,「〜について教えてください」と被験者が発話した場合を「当該文型」とした．それ以外（「〜を教えて」も含む）を「別文型」とした．また被験者の発話に対し，その意図どおりにシステムが応答した場合を「認識」，その他を「誤認識」とする．

「当該文型・認識」は，最初の質問が当該文型であり，うまく認識された場合である.「別文型・認識」は，最初の質問が別文型であったがうまく認識された場合である.「別文型→当該文型・認識」は，最初の質問が別文型だった場合，3 発話目までの質問が当該文型で，うまく認識できた場合である.「その他」は，2 回とも 4 発話目まで別文型で 4 発話目でうまく認識できた場合である．

表 4.1: 被験者の発話とエージェントの影響

	質問エージェントなし	質問エージェントあり
当該文型・認識	5	5
別文型・認識	3	4
別文型 →当該文型・認識	1	3
その他	2	0
合計	11	12

4.5 考察

実験の結果，半数近くの被験者は質問エージェントのありなしに関わらず，質問エージェントと同じ文型で最初の質問を行った．これは，質問エージェントの質問の文型が極めて一般的なものだったからと思われる．最初の質問が質問エージェントと異なる文型で，うまく認識できなかった場合（「別文型→当該文型・認識」および「その他」），質問エージェントありの条件では，3 人とも 3 発話目までに質問エージェントの文型で質問した（「別文型→当該文型・認識」）のに対し，質問エージェントなしの条件では，はじめて認識する前に質問エージェント文型で質問したのは 1 人にとどまった．これは，質問エージェントの発話が,「〜を教えてください」の文型に誘導した可能性がある．被験者の人数が少ないため，確定的なことは言えないが，提案手法の有効性を示唆する結果と言える．

4.6 まとめ

本章では，プレゼンテーションをエージェントが行う際に，質問の仕方の手本を見せるもう一体のエージェントを用いることで，発話の自然な制限を行うことにより，視聴者の発話の認識・理解を容易にする手法を提案し，実装した．予備実験の結果，質問エージェントの発話が被験者の質問パターンを，質問エージェントの発話の文型に誘導した可能性が示された．

第5章　トピック認識技術の導入による音声認識率の向上

5.1　はじめに

5.1.1　トピック認識

トピック認識とは，対話の話題を認識することである．発話に対するシステムの反応をより適切なものとするには，トピックの認識が必要である．

トピック認識技術は，様々な応用が考えられている．例えば映像データの分類や分割，議事録の音声データからの要約文作成などに利用することができる．プレゼンテーションタスクでは，質問の内容を聞き分けることにあたる．

本章では，音声認識エンジンを利用したトピック認識について述べる．

5.1.2　トピック認識の問題点

トピック認識で一番問題となるのは，トピック認識誤りである．例えば，ユーザは「営業時間」について音声入力を行ったにもかかわらず，システムが「ラストオーダー」についての発話だと認識するといった具合である．

従来手法の文法ベースのトピック認識では，文法に沿った発話が行われた場合には比較的良い認識結果を得ることが出来る．しかし，フィラーが入ったり，文法には無い発話の仕方をされた場合にはトピック認識率が著しく減少する．

5.1.3　本研究のアプローチ

前に挙げた問題に対して，本研究ではBWG(Bag of Words in Graph)とSVMV(Single random Variable with Multiple Value)法を用いて対処する．

提案する音声トピック認識手法は，語彙や文法を制限されること無しに自由に発生された音声のトピックをロボットが理解できるようにする物である．まず，文法データを用意する代わりに，トピックに対応するワードグラフを用意する．ワードグラフは単語をエッジとした非循環であり，人が作成した例文や音声認識の結果などから作成する．

BWG 法はワードグラフを文書と見なし，これに統計的な文書トピック認識の手法を適用する物である．本研究では文書トピック認識の手法として SVMV 法を用いた．この手法は，トピックが文法や単語の出現位置，順序に関係なく単語の出現頻度のパターンで定義される bag-of-wrds モデルに基づいたものである．

5.2 音声認識の原理

基本原理

音声認識は入力音声 X に対する事後確率 $p(W|X)$ が最大となる単語列 W を見つける問題として定式化される．事後確率 $p(W|X)$ を直接計算することは非常に困難であるので，ベイズ則により以下のように書き換える．

$$p(W|X) = \frac{p(W)p(X|W)}{p(X)} \tag{5.1}$$

この分母は W の決定に影響しない正規化係数と考えられるので，無視することが出来る．

音声認識における言語モデル

$p(W)$ は音声 X が入力される時点でのある単語列 W のパターンが生起する確率であり，音声 X とは無関係の言語的な確からしさである．言語モデルの適用は通常先頭の単語から逐次敵に行われ，単語列 $W = \{w_1, w_2, \ldots, w_k\}$ に対して，

$$p(W) = \prod_i p(w_i|w_1 \ldots w_{i-1}) \tag{5.2}$$

のようになる．記述文法の場合は，単語列が受理されるか否かで判定するオートマトン/パーサにより実現されるが，統計モデルの場合は $p(w_i|w_1, \ldots, w_{i-1})$ を直近の N 単語連鎖 $p(w_i|w_{i-N+1}, \ldots, w_{i-1})$ で近似して用いることが一般的である．これを単語 N-gram モデルと呼ぶ．

音声認識における音響モデル

$p(X|W)$ は単語列 W から音声のパターン X が生起する確率であり，音響的なモデルによるマッチングに基づいて評価が行われる．単語列 $W = \{w_1, \ldots, w_k\}$ は音素列 $\{m_1, m_2, \ldots m_l\}$ に展開されるので，$p(X|W)$ は以下のように計算できる．

$$p(X|W) = \prod_i p(x|m_i) \tag{5.3}$$

ここで $p(x|m_i)$ は通常音素単位の音響的特徴を表した HMM(Hidden Markov Model) を入力音声の一部 x とマッチングすることにより計算する．

デコーダの動作

音声認識の過程において式 (5.1) を様々な単語列 W の仮説について計算し，その中で最も事後確率の高いものを選択する．分子のみを考慮し，対数スケールに直すと次のように書き換えられる．

$$\begin{aligned}\hat{W} &= \arg\max p(W|X) \\ &= \arg\max\{\log p(W) + \log p(X|W)\}\end{aligned} \tag{5.4}$$

さらに，この式 (5.4) を次のように設定することが一般的である．

$$\log p(X|W) + \alpha \log p(W) + \beta * N \tag{5.5}$$

ここで，α は言語モデル重み，β は単語挿入ペナルティと呼ばれるパラメータであり，N は仮説 W に含まれている単語数 (単語列の長さ) である．

5.3 文法ベースの音声認識を用いたトピック認識

5.3.1 音声認識エンジン Julian

ディクテーションなどの大語彙連続音声認識を主な目的として李らが開発したフリーソフトウェアである Julius/Julian[50] を用いて実験を行った．音響モデルや言語モデルなどのインタフェースが公開されており，それらを置き換えたり修正したりすることが容易である．

5.3.2 MPML-HR における音声認識文法

MPML-HR コンテンツに記述された文法は，コンパイラにより Julian 用の音響モデル (voca ファイル) および言語モデル (grammar ファイル) に変換される．

文法はトピックごとに定義される．文法中の記号の意味は次の通りである．

parenthesis ( ) 通常の数式と同じ

bracket [ ] どちらの (またはどの) 単語が来ても良い

vertical bar | その中の単語が出現するかしないかのどちらかである

例えば，トピック「予算」に対して文法「(予算 [を教えて | が知りたい | はいくら])」を定義するといった具合である．このとき文法は,「予算」,「予算を教えて」,「予算が知りたい」,「予算はいくら」に展開される．入力された音声データに対して，この四通りの音素の繋がりに対する事後確率が他の文法のものより高い場合，音声データのトピックは予算と認識される．

5.3.3 認識精度の算出方法

音声認識システムが出力する仮説のそれぞれについて，認識システムがどれだけの確信を持ってその結果を出力したかの尺度を数値として付与することで，後段の音声認識処理でその値を考慮した処理が行える．この信頼度算出方法の一つとして，単語の事後確率に基づく信頼度計算法がある．

事後確率を用いた単語の信頼度計算

認識処理の結果得られた単語グラフ，あるいはN-bestの候補のリストにおいて，その中のある単語仮説 w が入力のフレーム τ から t に存在するとき，その単語仮説 $[w;\tau,t]$ の入力音声系列 X に対する事後確率 $p([w;\tau,t]|X)$ は，その仮説をパス上に含む全ての文仮説の出現確率の和から求められる．すなわち，

$$\begin{aligned} p([w;\tau,t]|X) &= \sum_{W \in W_{[w;\tau,t]}} \frac{p(X|W)p(W)}{p(X)} \\ &= \sum_{W \in W_{[w;\tau,t]}} \frac{e^{g(W)}}{p(X)} \end{aligned} \tag{5.6}$$

ただし，$W_{[w;\tau,t]}$ は単語仮説 $[w;\tau,t]$ をパス上に含む全文仮説の習合であり，$g(W)$ はデコーダより得られる文化説 W の言語モデル上および音響モデル上の出現確率の対数尤度である．$p(X)$ はそのN-best上の候補リストもしくは単語グラフにおける全ての文仮説の出現確率の和として計算できる．この事後確率から，単語仮説 $[w;\tau,t]$ の信頼度 $C([w;\tau,t])$ は以下のように定義される．

$$C([w;\tau,t]) = \sum_{W \in W_{[w;\tau,t]}} \frac{e^{\alpha \cdot g(W)}}{p(X)} \tag{5.7}$$

ただし，α はスムージング係数 $(0 < \alpha \leq 1)$ である．係数 α は尤度のダイナミックレンジを補正するために用いられる．

しかし，Frankら [46] によると，十分な性能を得るには $N = 100$ 程度が必要で，多くの計算量が必要となる．また，この計算は認識処理が終了した後に行わなければならないため，システム全体の応答速度にも遅延をもたらす．

探索過程での近似的な信頼度計算法

デコーダーの解探索過程において，その探索時のスコアから簡便に信頼度の計算を行う手法が提案されている [47]．次単語を展開する際に，その辞典で展開仮説が持つ推定尤度から，展開単語の事後確率を近似的に求めるというアプローチである．

トリートレリス探索は，木構造化辞書とスタックでコーディングを用いた 2 パス探索法の一つである [48]．Julius では大語彙連続音声認識のために拡張されたトリートレリス探索手法が用いられている [49]．第 1 パスでは木構造化辞書を用いたビーム探索を行い，各入力フレームにおいてビーム幅内に単語終端が残った単語について，その始終端時刻と入力先頭からの累積尤度を保存する．第 2 パスでは逆向き探索を行うが，その際に部分文仮説の最終単語の尤度を未探索部のスコアとして探索を行う．

第 2 パスにおいてある部分文仮説 $w_1^{n-1} = w_1, w_2, \ldots, w_{n-1}$ に対して単語 w_n を接続する際に，新たな仮説のスコア $f(w_1^n)$ は次のようにして求められる．

$$f(w_1^n, [w_n; \tau, t]) = g(w_1^{n-1}, t) + \hat{h}(w_n, t) \tag{5.8}$$

$$f(w_1^n) = \max_{0 \le t < T} f(w_1^{n-1}, [w_n; t]) \tag{5.9}$$

ただし $g(w_1^{n-1}, t)$ は時刻 t における第 2 パスの部分文仮説 w_1^{n-1} の先頭部分の前向き尤度，$\hat{h}(w_n, t)$ は時刻 t における接続する単語 w の後ろ向き尤度である．

尤度の話の計算においては最尤の確率が最終的な和の値を支配する場合が多いことから，単語 w_n を通る全ての仮説パスの出現確率の合計を求める代わりに，近似的に単語 w_n を通る最尤パスの出現確率を用いることが出来る．$f(w1^n)$ は探索途中の部分文仮説の評価スコアであるが，これは式 (5.8) から探索済みの部分の尤度と未探索部分のヒューリスティックな尤度の合計になっており，これをその時点での単語 w_n を通るパスの最尤スコアとみなすことができる．以上から，式 (5.6) の $g(W)$ を $f(w_1^n)$ で近似することが出来る．

$p(X)$ は，その展開単語 w_n と同じフレーム上に展開される全てのトレリス上の単語を展開候補として，その仮説 $[w; \tau, t]$ について $f(w_1^{n-1}, [w_n; t])$ を計算して足し合わせることで近似することが出来る．

以上から，単語信頼度を以下のような式で求めることが出来る．

$$\begin{aligned} C([w; \tau, t]) &= \hat{p}(w_n|X) \\ &= \frac{e^{f(w1^n)}}{\sum_{W_c} e^{f(w_1^{n-1}, [w; \tau, t])}} \end{aligned} \tag{5.10}$$

5.4 BWG とSVMV 確率モデルに基づくトピック認識

5.4.1 BWG(Bag of Words in Graph)

BWG とは，文章を単語の集合とする表現のことである．語順を完全に無視しても大丈夫な問題の際に用いられる．船越らにより，bag-of-words をベースとしたトピック認識技術が提案されている [45]．トピック認識にかける文章には，手動で作成した訓練データおよび Julius で音声認識を行った結果の文章を用いる．Julius を用いた場合は，出力された N-best の全てを bag-of-words で一つの文書とみなし，訓練データの作成を行う．

5.4.2 SVMV(Single random Variable with Multiple Value) 法

文書トピック認識の手法として，SVMV(Single random Variable with Multiple Value) 法 [52] が提案されている．この手法は，トピックが文法や単語の出現位置，順序に関係なく単語の出現頻度のパターンで定義される bag-of-words モデルに基づいた物である．トピック認識にかけるテキスト d からランダムに選択された索引語が t_i である事象をあらわす確率変数 $T = t_i$ とすると，$p(c|d, T = t_i)p(T = t_i|d)$ はテキスト d の中から索引語 t_i を選んだ時にそのトピックが c である確率を表す．全ての t_i について $p(c|d, T = t_i)p(T = t_i|d)$ を合計すれば，テキスト d がトピック c である確率を求めることが出来る．すなわち

$$p(c|d) = \sum_{t_i} p(c|d, T = t_i)p(T = t_i|d) \tag{5.11}$$

ここで索引語 $T = t_i$ が与えられたときに，テキストのトピックが c である事象とテキストが d である事象が独立であると仮定すると，すなわち，$p(c|d, T = t_i) = p(c|T = t_i)$ であると仮定すると次の式を得る．

$$p(c|d) = \sum_{t_i} p(c|T = t_i)p(T = t_i|d) \tag{5.12}$$

これにベイズ則を適用すると次の式を得る．

$$\begin{aligned} p(c|d) &= \sum_{t_i} \frac{p(T = t_i|c)p(c)}{p(T = t_i)} p(T = t_i|d) \\ &= p(c) \sum_{t_i} \frac{p(T = t_i|c)p(T = t_i|d)}{p(T = t_i)} \end{aligned} \tag{5.13}$$

ここで，式 (5.13) の各要素確率はあらかじめカテゴリを付与されたテキスト集合を訓練データとして使うことによって，以下のように推定することが出来る．

- $p(T = t_i|c)$ は，カテゴリ c に属するテキストからランダムに索引語を選択したときに，それが t_i になる確率であるので，カテゴリ c 中のテキストの索引語中の索引語中の t_i の相対頻度で推定できる．

- $p(T = t_i|d)$ は，分類すべきテキスト d からランダムに索引語を選択した際に，それが t_i となる確率であるので，テキスト d 中の索引語 t_i の相対頻度で推定できる．

- $p(T = t_i)$ は，ランダムに選択した索引語が t_i となるかくりつであるから，全ての訓練データ中の t_i の相対頻度で推定できる．

- $p(c)$ は，ランダムに選択したテキストがカテゴリ c に属する確率であるから，訓練データのテキスト中の c に属するテキストの相対頻度で推定できる．

このようにして求められた $p(c|d)$ のうち最大のものを与える c をこの入力テキスト d のトピックとする．すなわち，

$$topic(d) = \arg\max_{c \in C} p(c|d) \tag{5.14}$$

5.4.3 相互情報量による索引単語の制限

訓練に用いるデータが大量になると，索引語の数も膨大になってくる．訓練データに含まれる単語のうち，トピックとの相互情報量が大きい物を選択することによって，トピックの識別に貢献しない索引語を排斥できる．

トピックと索引語の相互情報量 $I(T_i; c)$ は次式で計算される．

$$I(T; c) = H(c) - H(c|T) \tag{5.15}$$

$H(c)$ は確率変数 c のエントロピーを，$H(c|T)$ は索引語が t_i であるという事象 T のもとでの c の条件付きエントロピーを表す．

訓練データはカテゴリを付与された文書の集合，つまり例文の集合である．ここで，索引語 t_i が訓練データ内のトピックを付与された文書データの中に存在する/存在しないという事象に対して，存在する $(= T_1)$/存在しない $(= T_0)$ の二値を取る $T \in \{T_0, T_1\}$ という確率変数を定義する．さらに，トピックの集合 C に対して，確率変数 $c \in \{c_1, \ldots, c_n\}$ を定義すると，$H(c)$ および $H(c|T)$ は次のように表される．

$$H(c) = -\sum_{c_i \in C} p(c_i) \log p(c_i) \tag{5.16}$$

$$\begin{aligned} H(c|T) &= -\sum_{j=0,1} p(T_j) H(c|T_j) \\ &= -\sum_{j=0,1} \sum_{c_i \in C} p(T_i) p(c_i|T_j) \log p(c_i|T_j) \\ &= -\sum_{j=0,1} \sum_{c_i \in C} p(T_j, c_i) \log p(c_i|T_j) \end{aligned} \tag{5.17}$$

ここで，式 (5.16) と式 (5.17) の各要素確率は以下のように推定することが出来る．

- $p(c_i)$ は全単語数に対するカテゴリ c_i に属する単語数の割合で推定できる．
- $p(T_i, c_i)$ は，全単語数に対する T_j でありかつカテゴリが c_i である単語の数として推定できる．
- $p(c_i|T_j)$ は，T_j である単語の数に対する T_j でありかつカテゴリが c_i である単語の数として推定できる．

5.4.4 認識精度の算出方法

SVMV 法によって求められた $topic(d)$ は，入力テキストが d であるという事象が発生した元でのトピックが c であるという事象の事後確率である．事後確率は，候補群の中における競合する候補同士の相対的な尤度の比率を表すことができ，信頼度として有効に働く事が知られている [51].

よってここでは，SVMV 法での認識精度を次の式で計算する．

$$\begin{aligned} C(topic) &= topic(d) \\ &= \arg\max_{c \in C} p(c|d) \end{aligned} \tag{5.18}$$

5.5 トピック認識実験とその結果

5.5.1 実験に用いるデータ

前章と同様に，プレゼンテーションエージェントを用いたシステムで収録を行った．被験者は，日本語を母国語とする 13 名である．被験者に，対話システムから情報を聞き出すという目的と，音声発話により対話システムが反応することを教示した．被験者に与えた課題は以下に示す通りである．

1. 営業時間とラストオーダーの時間を聞き出す

2. パーティーは開催できるかどうか聞き出す

次に，収録された音声について，書き起こしを行った．その際，各発話ごとに話題となっているトピックを付与した．これを正解データとして扱う．トピックのうち,「はい」「いいえ」「その他」という発話になっている物に関しては，人によっても判断が分かれる物であったため除外した.

これらにより，表 5.1 のようなデータを得た.

5.5.2 文法ベースの音声認識を用いたトピック認識

レストラン案内タスクにおける文法の例を表 5.2 に挙げる．これはレストラン案内のコンテンツの作成者が，予想される被験者の反応に対応するために記述した文法である．これを用いて，文法を用いたときのトピック認識率の測定を行った.

そして図 5.1 のような結果を得た．正解率は，話者 13 人の平均で 61.6%となった.

誤り率の高い話者 3 の，正解データの例を表 5.3 に，不正解データの例を表 5.4 に示す.

このように，文法ベースの音声認識を用いたトピック認識では，発話にフィラーが入っていたり，未知語が入っていたりする場合に認識率が急激に下がってしまうという特徴がある.

表 5.1: 被験者と発話数

被験者番号	全データ	抽出された適切なデータ
1	41	23
2	53	41
3	112	78
4	85	67
5	86	68
6	145	78
7	65	51
8	60	42
9	84	49
10	92	76
11	61	52
12	35	31
13	58	51
合計	977	707

表 5.2: 文法の例

トピック	文法
一つ前に戻る	[ひとつ] まえ [にもどって]
おわる	(おわり [にして]| もう (せつめー | しょーかい)[わ] いいよ)
予算	(ねだん | よさん) わ [いくら [くらい][ですか]]
営業時間	えーぎょーじかん (わ | おおしえて | なんじから [なんじまで])
:	:

表 5.3: 正解データの例

正解タグ	話者の発話
お勧め料理	おすすめりょーりおおしえてください
場所	おみせのばしょおしえてください
営業時間	えーぎょーじかんおおしえてください
パーティー	えーとぱーてぃーができるかおしえてください

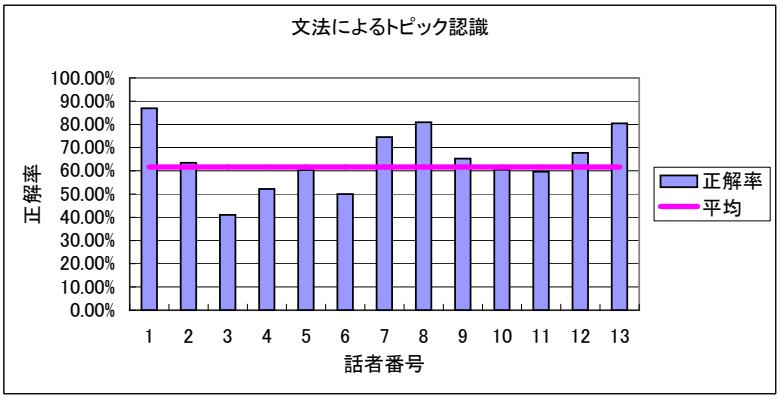

図 5.1: 文法ベースの音声認識を用いたトピック認識

表 5.4: 不正解データの例

正解タグ	話者の発話
お勧め料理	あえっともっぺんおすすめりょーりおおしえてください
場所	ともーいっかいおみせのばしょおしえてくれまくれますか
営業時間	なんじからやってますか
パーティ	えとぱーてぃーすぱーてぃーするとひとりどのぐらいのよさんでできますか

5.5.3 学習例文を作成した場合の SVMV 法によるトピック認識実験

手動で作成した訓練データによる学習

以下のような条件で測定を行った．なお，条件 1 と 2 は著者が例文を作り，条件 3 は研究に関係の無い第三者に作成を依頼した．

1. 各トピックごとに例文を 10 用意する
2. 各トピックごとに例文を 10 用意し，さらにプレゼンテーションコンテンツからトピックに対応する本文を追加する
3. 各トピックごとに例文を 20〜50 用意する

条件 1 では，比較的簡単に作成された例文でどのくらいトピック認識率が上がるのかを調べた．

条件 2 では，条件 1 の結果レストラン案内タスクの中に出現する単語を話者が発話するという現象がみられたため，コンテンツ本文を学習データに含めることによりトピック認識率があがるのでは無いかと考え実験を行った．

条件 3 では，実験結果をみたことがない大学生に，例文の作成を依頼した．これにより，一般の人が例文を考えたときにどのていど認識されるのかの概算ができると考えた．この条件での結果は図 5.2 のようになった．正解率は，条件 1 で 62.3%，条件 2 で 64.3%，条件 3 で 59.5%となった．

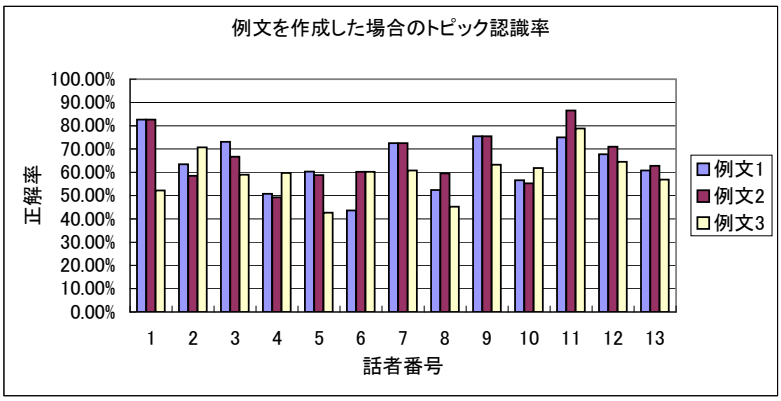

図 5.2: 文例作成によるトピック認識

自立語のみを含む訓練データによる学習

人が作成した例文を見ると,「営業時間」や「予算」といった単語がトピックを決める際に大きく影響しているのではないかと考えられる．ここから，例文 1 に含まれる単語の中から自立語 (名詞，副詞，動詞，形容詞，連体詞，感動詞) のみを取り出し，学習データとして実験した．そして，図 5.3 のような結果を得た．正解率は，13 人の平均で 58.8%となった．

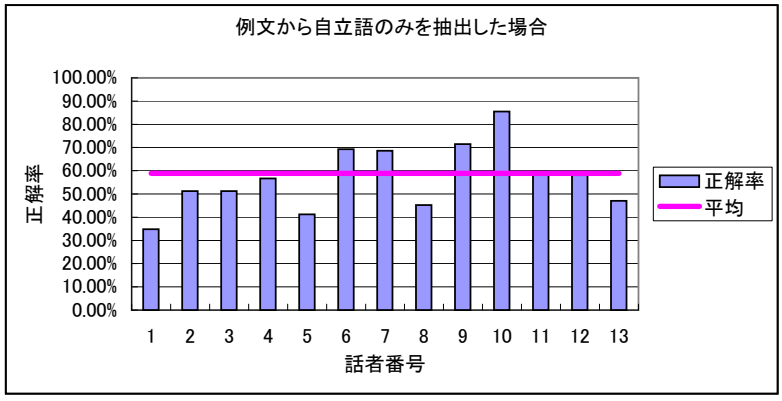

図 5.3: 自立語のみによる学習

相互情報量による索引語の絞り込み

学習に用いる例文の中で，全ての文に出現する単語などはトピック認識の精度に貢献しないと予想される．そのため，学習に用いる索引語を，トピックに対する相互情報量により絞り込むことを提案する．そして，相互情報量の順に上位何%を用いれば良いのかを求めるために，用いるデータ数を0%から100%まで5%刻みで調べた．そして図5.4のようなデータを得た．正解率は，条件3で索引語のうち上位45 %を使った場合の66.3%が最大となった．

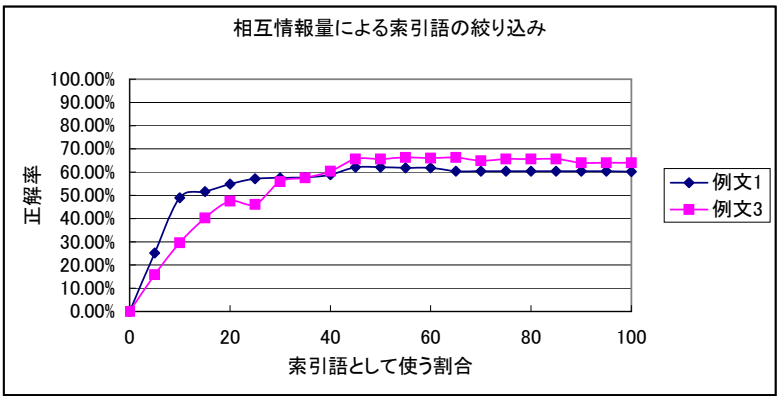

図 5.4: 相互情報量による索引語の絞り込み

5.5.4 音声を用いて訓練データを作成した場合のSVMV法によるトピック認識実験

音声による訓練データの作成

SVMV法において文例を人が作る場合は，音声認識エンジンが出力する結果を用いないため，その点において認識失敗する確率があがる．例えば「営業時間」というトピックに対しても，音声認識エンジンは「えぎょう」「えいぎょ」などとして出力する可能性があり，簡単には予想できない．

音声データを学習データとして用いる場合は，学習時に用いる例文を音声認識エンジンが出力するため，誤った認識をしていたとしても，トピック認識する際に同じく誤って認識することでトピック認識自体は成功することが出来る．

音声認識エンジンJuliusから出力された文章を学習データとして用いて実験を行った．出力される文はN-BEST=5(尤度の高い順から5番目まで)とした．そして，図5.5のような結果を得た．正解率は，12人分の音声を学習させ，索引語のうち上位70%を使用した場合が74.1%で最大となった．

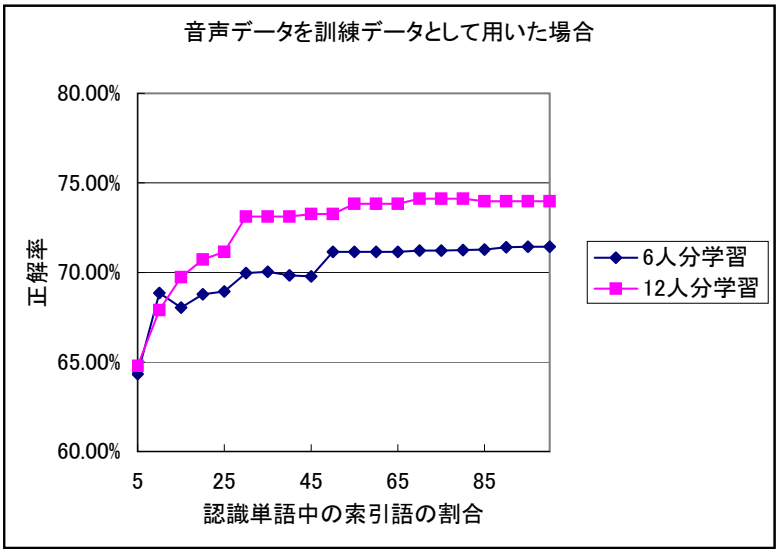

図 5.5: 音声による学習データの作成

音声による訓練データの数とトピック認識率

音声データによる学習において，学習に必要な人数が分かると実際上の応用もしやすくなる．そこで，話者数を 1 人から 12 人まで増やした場合において正解率を算出してグラフにした．この際，全ての話者に対して，その話者以外から 1，2，...12 人を選び評価するという Cross Validation を行った．その結果，図 5.6 のような結果を得た．正解率は，12 人分の音声を学習データとして用いた場合で，75.6%となった．

音声から訓練データを作成する際の N-BEST 数とトピック認識率

音声認識エンジン Julius を用いて訓練データを作成する際に，候補となる出力文を N-BEST 数だけ出力する．BWG では文書を単語の集合として扱っているので，文書に含まれる単語の数が増えるほど情報量が増え，正解率が上がると予想される．正解率は，使用する索引語の割合を 50 %，N-BEST を 10 とした場合の 76.1% が最大となった．

5.6 認識時に算出される信頼度の評価

トピック認識とは別に，システムがその認識に対する信頼度 (確信度) を出力することで，後段の処理をより適切に行うことが出来る．

MPML-HR システムも，Confidence Threshold を用いて応答や確認，聞き返しなどを行う機構を実装している．

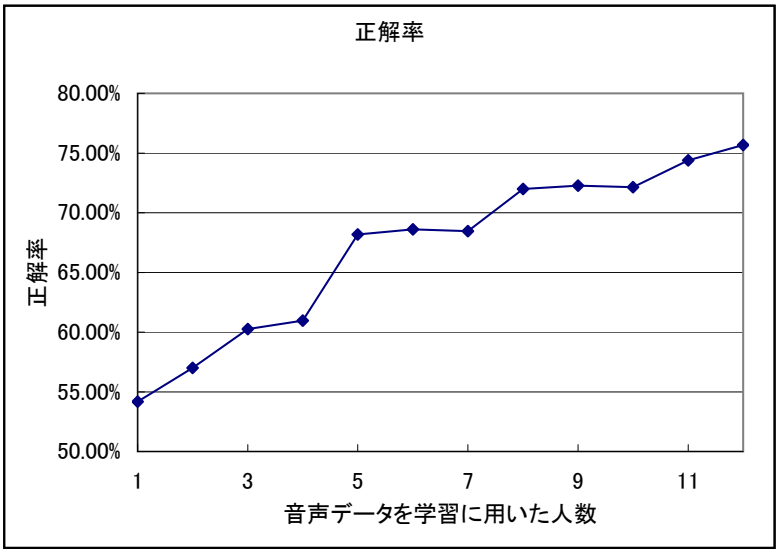

図 5.6: 音声データを学習に用いる人数とトピック正解率の関係

このため，それぞれの手法の信頼度と正解に対して，受信者動作特性曲線 (Receiver Operating Characteristic Curve) をによる評価を行った．受信者動作特性曲線においては，グラフが原点を通るときが最も良い信頼度指標であるとされる．

発話に対する Confidence 値の分布は図 5.8 のようになった．

図 5.9 中の FAR(False Acceptance Rate) および FRR(False Rejection Rate) は次の式で表される．

$$FAR = \frac{\text{トピック認識失敗したデータのうち } threshold \text{ 以上のスコアの数}}{\text{トピック認識失敗したデータの数}} \tag{5.19}$$

$$FRR = \frac{\text{トピック認識に成功したデータのうち } threshold \text{ 以下のスコアの数}}{\text{トピック認識成功したデータの数}} \tag{5.20}$$

ROC カーブ上の点は，threshold をパラメータとして表現される．曲線状の $FAR = FRR$ である点の FAR,FRR を Equal Error Rate と呼ぶ．この曲線上の，例えば $(FRR, FAR) = (0.4, 0.1)$ の点は，正解であるにもかかわらず不正解とされた割合が 40%，不正解であるにもかかわらず正解とされた割合が 10%という意味である．

5.7 まとめ

全体としてのトピック認識率は，音声を学習データとして用いた条件が最も良いが，その後の処理に用いるトピック認識の信頼度については文法ベースの音声

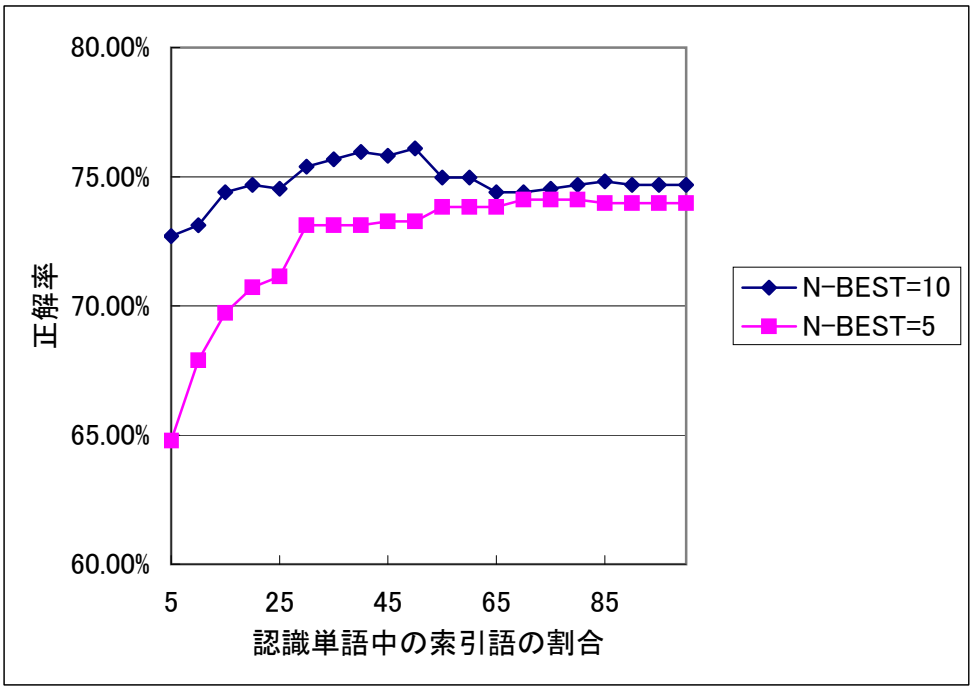

図 5.7: N-BEST を変化させた場合の正解率

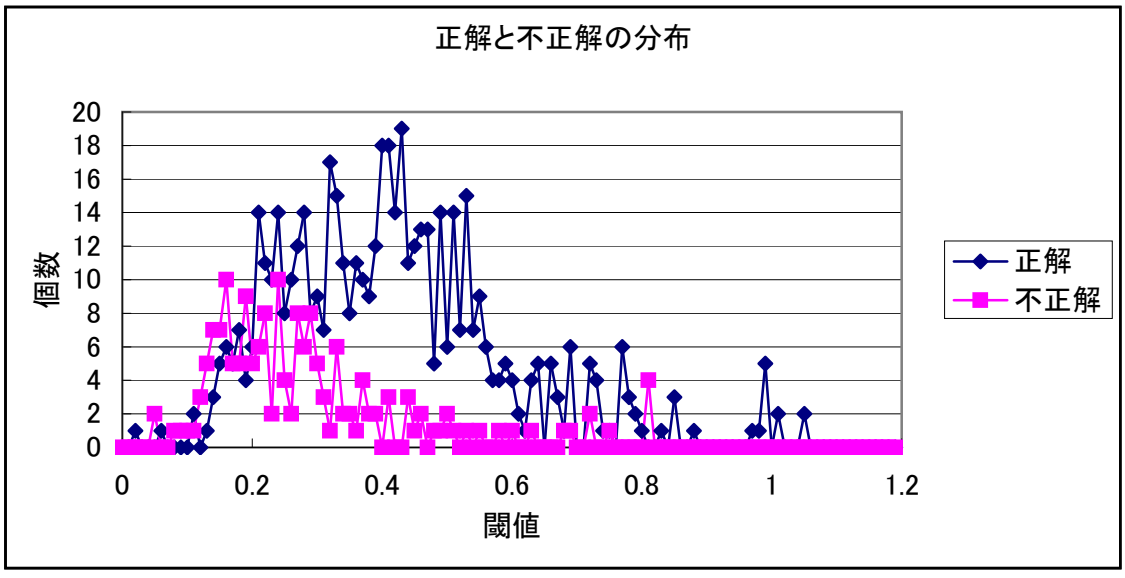

図 5.8: 発話に対する Confidence 値の分布

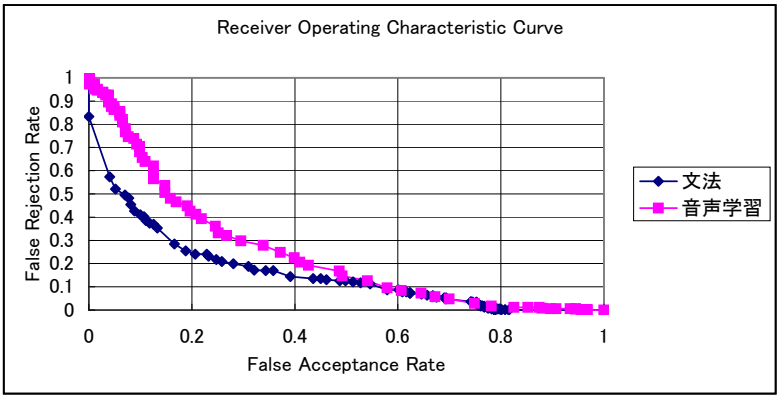

図 5.9: 受信者動作特性曲線

認識を用いたトピック認識が良い精度だった.

音声を学習データとして用いた場合には，音声認識エンジンの出力 N-BEST 数を 10，使用する索引語の割合を 50%，学習させる人数を 12 人とした場合がもっとも正解率が高く，76.1%となった.

また，音声を学習データとして用いた条件の際にトピック認識誤りをしているデータでも文法ベースのトピック認識法では正解しているデータが全体の 9%程度存在する.

このため，お互いの認識結果が異なる際にどちらを用いるべきかを分類器を用いて学習させることなどにより，最終的なトピック認識率を上げることが出来る可能性がある.

第6章　結論

ロボットやアニメーションエージェントの応用の一つである自動プレゼンテーションの実現のために，ユーザからの音声入力を受け付けてプレゼンテーションの内容を変えるプレゼンテーション／エージェントシステムの開発を行った．インタラクティブなプレゼンテーションは，ロボットのプレゼンテーションにおいて音声による割り込みを受け付ける事で実現できるが，様々な問題も存在する．

MPML-HR を用いることでロボットからの一方的なプレゼンテーションコンテンツの作成は可能となったが，インタラクションを含むコンテンツの作成はできなかった．インタラクティブなプレゼンテーションを容易に作成することを目指し，MPML-HR に対して音声インタラクション機能を導入した．インタラクション機能の導入にあたり，説明の省略，前に話した内容の再度の説明，およびあらかじめ想定した説明に答えるということを目的とした．音声認識誤りへの対処が必要となるため，音声認識結果に対する信頼度を用いた機構を導入した．信頼度が低い場合は音声入力受付時に棄却，聞き返し，確認，受理のいずれかを行う．

また，インタラクティブなプレゼンテーションエージェント/ロボットシステムでは，視聴者が，どのような発話内容やタイミングで質問や要求を行ってよいかがわからないという問題がある．これに対処するため，プレゼンテーションを行うエージェントのほかにもう一つのエージェントを用いて，質問や要求をデモンストレーションさせることにより，暗示的に視聴者に発話の仕方を教える方法を提案した．

さらに，インタラクティブなプレゼンテーションにおいて，視聴者の発話に対するシステムの反応をより適切なものとするには，発話の内容からトピックを認識することが必要である．例えば，文法によるトピック認識ではフィラーによって認識精度が左右されるという問題が存在する．音声認識率の向上のために bag-of-words をベースにしたトピック認識技術を導入した．音声認識システムは単語の出現頻度パターンとして文章のカテゴリを学習する．この方法により，音声認識誤りや音声認識の辞書に登録されていない単語に対しても頑健な学習と理解が可能となる．

謝辞

本研究を進めるにあたり，終始暖かい御指導を受け賜りました石塚満教授，土肥浩助教に感謝します．本研究は，(株) ホンダ・リサーチ・インスティチュート・ジャパンにおいて多くの方々の暖かいご支援に支えられて行うことができたものです．研究に関する助言のみならず，実験の手伝いや論文のチェックに至るまで面倒をみていただいた中野幹生氏，西村義隆氏，船越孝太郎氏，竹内誉羽氏，辻野広司氏に感謝申し上げます．実験方法の相談に乗っていただいたり，実験準備を手伝っていただいた長谷川雄二氏，中臺一博氏，中島弘史氏に感謝します．実験のためのプログラムを利用させていただいた (独) 情報通信研究機構の木村法幸氏，岩橋直人氏に感謝します．本研究を行うためのすばらしい環境を提供していただいた (株) ホンダ・リサーチ・インスティチュート・ジャパンの皆様，評価実験に参加していただいた皆様，実験データの書き起こしなどを手伝っていただいた皆様に感謝します．日頃から石塚研究室のメンバーにもお世話になりました．石塚研究室秘書の藤田メイコさんには，平時より楽しく快適に研究を行う環境を整えていただき感謝します．電気系事務室の池谷幸絵さんには，論文提出時の手続きなどでお世話になり，感謝します．同期入学でもあり，学生生活において励ましてくれた唐門準氏，岡嶋穣氏をはじめ石塚研究室のメンバー皆様に深く感謝申し上げます．本研究を進めるにあたり，励まして下さった方々に改めて心から感謝致します．

参考文献

[1] J. Weizenbaum. ELIZA–A Computer Program For the Study of Natural Language Communication Between Man and Machine. Commun. ACM 10[1] 36-45(1966)

[2] Kazutaka Kushida, Yoshitaka Nishimura, Hiroshi Dohi, Mitsuru Ishizuka, Johane Takeuchi and Hiroshi Tsujino: Humanoid Robot Presentation through Multimodal Presentation Markup Language MPML-HR, 4th Int'l Conf. on Autonomous Agents and Multi Agent Systems (AAMAS-05) Workshop 13 "Creating Bonds with Humanoids", pp.23-29, Utrecht, The Netherlands (2005.7)

[3] Y. Niimi, K. Ueda, T. Nishimoto, Y. Niimi: "Dialogue Scenario Generation from XML-based Database," Proc. of 1st NLP and XML Workshop, pp.9-14, 2001.

[4] マルチドメイン音声対話システムの構築手法, 長森 誠, 河口 信夫, 松原 茂樹, 外山 勝彦, 稲垣 康善, 自然言語処理,137-12, 音声言語情報処理,31-7,2000.6.2

[5] 音声対話システムの言語・対話処理, 中野 幹生，堂坂 浩二, 人工知能学会誌,2002.5.

[6] Subjective Evaluation of Spoken Dialogue Systems Using SERVQUAL Method,M Hartikainen, EP Salonen, M Turunen - Proc. of ICSLP, 2004

[7] Darpa Communicator: Cross-system results for the 2001 evaluation,ICSLP 2002

[8] Evaluation of the dutchtrain timetable information system developed in the ARISE project,Voice Technology for Telecommunications Applications, IVTTA, 1998

[9] Helmut Prendinger,San ti Saeyor,and Mitsuru Ishizuka,MPML and SCREAM: "Scripting the Bodies and Minds of Life-like Characters ",in Life-like Characters. Tools,Affective Functions and Applications,ed Predinger H. and Ishiduka M.,Springer, 2004

[10] B. De Carolis, F. de Rosis, V. Carofiglio and C. Pelachaud (2001): B. De Carolis, F. de Rosis, V. Carofiglio and C. Pelachaud, Interactive Information Presentation by an Embodied Animated Agent. International Workshop on Information Presentation and Natural Multimodal Dialogue. Verona, Italy, December 14-15, 2001

[11] B.D. Carolis, C. Pelachaud, I. Poggi, and M. Steedman: “ APML, a Markup Language for Believable Behavior Generation, ” Life-like Characters (H. Prendinger and M. Ishizuka eds.), Springer, pp.51-54, 2004.

[12] HAYASHI M. , UEDA H. , KURIHARA T. , YASUMURA M. ,“ TVML(TV Program Making Language) . Automatic TV Program Generation from Text-Based Script . “ , Proceedings of IMAGINA99, pp.122-133, (1999)

[13] S. Beard, D. Reid, and A. Marriott, ”MetaFace and VHML: A First Implementation of the Virtual Human Markup Language,” Curtin University of Technology, Perth, Australia, Workshop paper - to be published 2002.

[14] M. Araki, K. Ueda, T. Nishimoto, and Y. Niimi: Dialogue scenario generation from xml-based database; In Proc. of 1st NLP and XML Workshop, pp. 9-14, 2001.

[15] 河野恭之, 屋野武秀, 笹島宗彦. カーナビ音声対話システム MINOS の試作. 人工知能研資, SIG-SLUD-9901-4, 1999.

[16] 林正樹, テキスト台本からの自動番組製作〜TVML の提案,1996 年テレビジョン学会年次大会,S4-3,pp.586-592(1996)

[17] 音メディアデータの高度アーカイブシステムに関する研究－芸能音楽の記録と再現のための高度アーカイブシステムの研究－, 山下洋一, 立命館大学２１世紀ＣＯＥプログラム京都アート・エンタテインメント創成研究, 2003

[18] 河野恭之, カーナビ音声対話システム minos の試作，人工知能学会研究会資料,Vol.SIG-SLUD-9901-4 21-26,1999

[19] 安達史博, 河原達也, 奥乃博, 岡本隆志, 中嶋宏, VoiceXML の動的生成に基づく自然言語音声対話システム, 情報処理学会研究報告. SLP, 音声言語情報処理 2002(10),133-138,20020201(社団法人情報処理学会)

[20] Wallace, R.S. The Anatomy of A.L.I.C.E. in A.L.I.C.E. Artificial Intelligence Foundation, Inc., (online),http://www.alicebot.org/.

[21] A. Marriott: “ VHML - Virtual Human Markup Language, ” (Online), http://www.vhml.org/

[22] Y. Nishimura, K. Kushida, H. Dohi, M. Ishizuka, J. Takeuchi, and H. Tsujino, "Development and psychological evaluation of multimodal presentation markup language for humanoid robots," in *Proc. of IEEE RAS Humanoids*, 2005, pp. 393–398.

[23] J. Takeuchi, K. Kushida, Y. Nishimura H. Dohi, M. Ishizuka, M. Nakano, and H. Tsujino, "Comparison of a Humanoid Robot and an On-Screen Agent as Presenters to Audiences," in *Proc. of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2006, pp. 3964–3969.

[24] H. Prendinger, S. Descomps and M. Ishizuka, "MPML:A Mark p Lang age for Controlling the Behavior of Life-like Characters," Journal of Visual Languages and Computing, 2004, Vol.15, No.2, pp. 183–203.

[25] Voice XML Forum homapage: http://www.voicexml.org/

[26] K. Katsurada, Y. Nakamura, H. Yamada, T. Nitta, "XISL:A Language for Describing Multimodal Interaction Scenarios," *Proc. of ICMI*, 2003, pp. 281–284.

[27] J. Glass and E. Weinstein, "SpeechBuilder: Facilitating spoken dialogue system development," in *Proc. Eurospeech*, 2001, pp. 1335–1338.

[28] S. Sutton et al., "Universal Speech Tools: The CSLU Toolkit" in *Proc. ICSLP*, 1998, pp. 3221–3224.

[29] 酒井，西山，溝口，"人型二足歩行ロボットを用いたプレゼンテーションロボットシステムの設計，" 日本ソフトウェア科学会第 21 回大会論文集，2004, pp. 1–4.

[30] M. Nakano, Y. Hasegawa, K. Nakadai, T. Nakamura, J. Takeuchi, T. Torii, H. Tsujino, N. Kanda, and H. G. Okuno, "A two-layer model for behavior and dialogue planning in conversational service robots," in *Proc. IEEE/RSJ IROS*, 2005, pp. 1542–1548.

[31] T. Kawahara, A. Lee, K. Takeda, K. Itou, and K. Shikano, "Recent progress of open-source LVCSR engine Julius and Japanese model repository," in *Proc. ICSLP*, 2004, pp. 3069–3072.

[32] Y. Nishimura, et al. ,"A Markup Language for Describing Interactive Humanoid Robot Presentations," Proc. IUI. 2007.

[33] 福林他：音声対話システムにおける発話パターンを教示するヘルプの動的生成．人工知能学会 SLUD 研究会 A601-03，2006．

[34] 小松他: 人間と人工物との対話コミュニケーションにおける発話速度の引き込み現象．情処研報 ICS105, 2004.

[35] Helmut Prendinger, Santi Saeyor, and Mitsuru Ishizuka, MPML and SCREAM: "Scripting the Bodies and Minds of Life-like Characters", in Life-like Characters. Tools, Affective Functions and Applications, ed Predinger H. and Ishiduka M.,Springer, 2004

[36] T. Tsutsui, S. Saeyor and M. Ishizuka: MPML: A Multimodal Presentation Markup Language with Character Agent Control Functions, Proc.(CD-ROM) WebNet 2000 World Conf. on the WWW and Internet, San Antonio, Texas, USA, (2000)

[37] Y. Zong, H. Dohi and M. Ishizuka: "Multimodal Presentation Markup Language MPML with Emotion Expression Functions Attached", Proc. 2000 Int'l Symp. on Multimedia Software Engineering (IEEE Computer Soc.), pp.359-365, 2000.

[38] N. Okazaki, S. Aya, S. Saeyor, and M. Ishizuka: "A Multimodal Presentation Markup Language MML-VR for a 3D Virtual Space," Workshop Proc. (CD-ROM) on Virtual Conversational Characters: Applications, Methods, and Research Challenges (in conjunction with HF2002 and OZCHI2002), 4 pages, 2002.

[39] Santi Saeyor, Koki Uchiyama, Mitsuru Ishizuka, "Multimodal Presentation Markup Language on Mobile Phones", AAMAS Workshop Proc. (W10) - Embodied Conversational Characters as Individuals, (in conjunction with Second Int'l Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-03)), Melbourne, Australia, pp.68-71, (2003.7)

[40] http://www.miv.t.u-tokyo.ac.jp/mpml.html

[41] http://www.honda.co.jp/ASIMO/

[42] http://www.microsoft.com/msagent/

[43] Ortony, A., Clore, G.L., Collins, A."The cognitive structure of emotions". Cam-bridge University Press, Cambridge, MA. (1988)

[44] Justine Cassell, Joseph Sullivan, Scott Prevost and Elizabeth Churchill "Emodied Conversational agents" MIT Press, 2000

[45] 船越孝太郎, 中野幹生, 鳥井豊隆, 長谷川雄二, 辻野広司, 木村法幸, 岩橋直人, 語彙制限のない発話の頑健な学習と理解を行う家庭用ロボット, Human-Agent Interaction Symposium 2007

[46] Frank Wessel, Ralf Schluter, Klaus Macherey, and Hermann Ney. Confidence measures for large vocabulary continuous speech recognition. IEEE Trans. Speech & Audio Process., Vol.9, No.3, pp.288-298, March 2001.

[47] Akinobu Lee, Kiyohiso Shikano, and Tatsuya Kawahara, "Real-time word confidence scoring using local posterior probabilities on tree trellis search," In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2004), Vol.I, pp.793–796, May 2004.

[48] F.K.Soong and E.F.Huang. A tree-trelis based fast search for finding the N best sentence hypotheses in continuous speech recognition. In Proc. ICASSP, Vol.1, pp.705-708, 1991.

[49] 李晃伸, 河原達也, 堂下修司. 単語トレリスインデックスを用いた段階的探索による大語彙連続音声認識. 電子情報通信学会論文誌,Vol.J82-D-II No.1, pp.1-9, 1999.

[50] T. Kawahara, A. Lee, K. Takeda, K. Itou, and K. Shikano, "Recent progress of open-source LVCSR engine Julius and Japanese model repository," in *Proc. ICSLP-2004*, 2004, pp. 3069–3072.

[51] Frank Wessel, Ralf Schluter, Klaus Macherey, and Hermann Nry. Confidence measures for large vocabulary continuous speech recognition. IEEE Trans. Speech & Audio Process., Vol.9, No.3, 99.288-298, March 2001

[52] M.Iwayana abd T.Tokunaga: "A Probabilistic model for text categorization: Based on a single random variable with multiple values", In Proc. of the 4rth Applied Natural Language Processing Conference(ANLP), pp.162-167(1994)

発表文献

- 西村義隆，簑津真一郎，土肥浩，石塚満，中野幹生，船越孝太郎，竹内誉羽，長谷川雄二，辻野広司，インタラクション機能を有するプレゼンテーション記述言語の開発， HAI シンポジウム, Dec. 2006

- Yoshitaka Nishimura, Shinichiro Minotsu, Hiroshi Dohi, Mitsuru Ishizuka, Mikio Nakano, Kotaro Funakoshi, Johane Takeuchi, Yuji Hasegawa, Hiroshi Tsujino, A Markup Language for Describing Interactive Humanoid Robot Presentations, Proc. International Conference on Intelligent User Interfaces (IUI-2007), Hawaii, Jan. 2007.

- 簑津真一郎，西村義隆，土肥浩，石塚満，中野幹生，船越孝太郎，竹内誉羽，長谷川雄二，辻野広司，“プレゼンテーション記述言語 MPML-HR における音声インタラクション機能，” 第 69 回情報処理学会全国大会論文集 (CD-ROM), (2007.3).

- 簑津真一郎，中野幹生，船越孝太郎，竹内誉羽，長谷川雄二，土肥浩，石塚満，辻野広司，“インタラクティブなプレゼンテーションでのユーザ発話の自然な制限のための複数エージェントの利用，” 第 69 回情報処理学会全国大会論文集 (CD-ROM)，(2008.3)．(発表予定)