NAIST-IS-DT0161030

博士論文

大語彙連続音声認識を基盤技術とする
実用指向音声インタフェースに関する研究

西村 竜一

2004年3月24日

奈良先端科学技術大学院大学
情報科学研究科 情報処理学専攻

本論文は奈良先端科学技術大学院大学情報科学研究科に
博士 (工学) 授与の要件として提出した博士論文である.

論文番号： NAIST-IS-DT0161030

提出者： 西村 竜一

審査委員： 鹿野 清宏 教授
小笠原 司 教授
木戸出 正繼 教授
猿渡 洋 助教授

大語彙連続音声認識を基盤技術とする実用指向音声インタフェースに関する研究*

西村 竜一

内容梗概

音声認識の進歩とともに機械との対話を実現する音声インタフェースへの期待が高まっている．電話応対システムやカーナビの音声インタフェースの一部では，今や実用に足る有用性を得たと言えよう．しかし，音声インタフェースが日常社会に普及した例はまだ少なく，その応用が多岐にわたるのに対し，利用実態調査は十分でない．本研究の目的は，音声インタフェースを備えたアプリケーションを開発し，そのフィールドテストの中で普及の障害となる課題を検証することである．要素技術の開発による利便性向上や開発のコスト削減も目的である．

はじめに，開発の前段階としてタスク適応言語モデルの構築法を検討する．大語彙連続音声認識では，必要な認識精度を得るのに認識対象と合致した単語 N-gram 言語モデルが必要である．その構築には大規模なテキストコーパスが必須で多大な人手を要した．本手法は，半自動に言語モデルを構築でき，開発のコスト削減に貢献するものである．その手順は，(1) コーパスの自動作成とトピック依存 N-gram モデルの構築，(2) モデル融合でのタスク操作，(3) 文法の適用によるモデル高精度化，の三段階から構成される．

研究プラットホームとして最初に開発した受付案内ロボット ASKA は，音声のみでなく画像処理等も応用した対話機能を持つ．ASKA は大学の受付案内が仕事であり，合成音声と手と頭のジェスチャを使って来訪者の応対をすることができる．本論文では，音声インタフェース部を中心に構成を述べ，開発プロジェクト

*奈良先端科学技術大学院大学 情報科学研究科 情報処理学専攻 博士論文，NAIST-IS-DT0161030, 2004 年 3 月 24 日.

の活動をまとめた．音声認識における言語モデルと文法の性能を ASKA のタスクで比較し，その併用手法の有効性も確認した．

音声情報案内システム「たけまるくん」は，音声インタフェースの利用実態調査を実現するシステムである．一問一答形式の音声インタフェースを持ち，誰でも気軽に利用できる受付案内サービスを提供する．システムを生駒市コミュニティセンターに常設し，約五ヶ月間，利用者の発話を収録した．時間にして 1,362 分の男女幅広い年齢層の発話から，大人と子供で発話の内容に傾向の違いがあるが，システムは有効に利用されていることを確認した．音声インタフェースの開発では，継続的運用による経験とデータの蓄積は不可欠である．本システムはそれを可能にした．評価実験では，大人で 86%の単語認識率と 76%の応答正解率を得ることができた．しかし，子供話者に対する性能不足が明らかになった．

音声インタフェースが家庭や公共施設に普及することを考えると，子供の存在は無視できない．従来のシステムでは不十分であった子供に対する利便性の改善を目指して，年齢層に即した柔軟な対話を可能にする音声インタフェースを新たに検討した．その実装に必要な大人・子供の識別方法として，音声認識の対数尤度から導出する音響的特徴と言語的特徴を併用した機械学習に基づく話者識別法を提案する．SVM（Support Vector Machine）を用いた実験では 94.6%の識別率を得た．これは音響的特徴のみを含む混合正規分布モデルを使った従来手法から 8.2%の改善である．また，収集発話から言語モデルと音響モデルを再構築して子供発話認識率の 12.5%の向上を得た．

キーワード

ロボット音声インタフェース，公共型音声インタフェース，大語彙連続音声認識，タスク適応言語モデル，フィールドテスト，大人・子供話者識別

Practical Speech Interfaces Based on Large Vocabulary Continuous Speech Recognition*

Ryuichi Nisimura

Abstract

Since spoken language is considered to be one of the most effective means for humans to communicate, it is for this reason that a speech interface is favorable to realize a novel human-machine interface. Improvements of large vocabulary continuous speech recognition technology have brought certain usefulness to speech interfaces. However, there are only few speech interface systems running in our daily life. To investigate the reasons that prevent a speech interface from being used as a utility, this study introduces two research platforms, "ASKA" and "Takemaru-kun". By using them, field tests of speech interfaces are performed to observe actual human-machine interactions in practical environments.

Firstly, as a preparation of this study, an automatic building procedure of task-specific N-gram language models is proposed. This is necessary in order to develop the field test system with the sufficient speech recognition accuracy. This procedure consists of collecting text corpus for model learning, building topic-dependent N-gram models, task decision operation by merging the models and applying network grammars to the merged model.

The next step, this thesis describes an overview of ASKA, a humanoid robot with a spoken dialogue developed as a computerized receptionist in our university. ASKA can recognize user's question utterances, and answer the user's question by text-to-speech synthesized voice with hand and head gestures.

*Doctor's Thesis, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DT0161030, March 24, 2004.

An overview of the Takemaru-kun system and its speech related parts are also explained. The analysis result of 1,362 minute length utterances collected via its field test is shown. Takemaru-kun system is a speech-oriented guidance system located at the Ikoma-City North Community Center. It has been operated daily from November, 2002, and collecting utterances from actual users with a wide variety of ages and both genders. Although the tendency of utterance contents is different between adults and children, the analysis result shows that this system is used effectively. In the evaluations with extracted samples of adult voices, 86% word accuracy and 76% response correct rate are observed. However, inadequate accuracy of recognition rate and response rate for child voices are clarified.

Lastly, this thesis proposes an automatic approach discriminating speakers between adults and children, which is based on a statistical learning. It realizes a flexible spoken dialogue to both adult and child users. It becomes impossible to disregard the increase of child users when the system is installed in a public place. This proposal aims at an improvement in convenience of speech interfaces for child users. As for parameter vectors in machine learning, acoustic and linguistic properties extracted from speech recognition logarithm likelihood are adopted to discriminate user's age group. Although GMM-based recognition uses only acoustic properties, this method can also consider linguistic properties. In the experiments with the SVM-based screening, we obtained 94.6% discrimination accuracy to the actual users' utterances. 8.2% improvement of using linguistic properties is observed. Also, to improve child speech recognition, collected utterances are applied to train recognition models. The experimental result shows 12.5% improvement in child speech recognition accuracy.

Keywords:

robot speech interface, public use speech interface, large vocabulary continuous speech recognition (LVCSR), task suitable language model, field-test, adult and child speaker discrimination

目 次

図 目 次

表 目 次

第1章　序論

1.1.　まえがき

> *2003*年 *4*月 *7*日，科学省長官・天馬博士は，交通事故で死んだひとり息子・飛雄（とびお）に似た人型ロボットを科学省の総力をあげて作り上げる．世界最高の電子頭脳を持ち，人間同様の優しい心と大きな正義感を備えたスーパーロボット．ロボットと人間が友達でいられる平和な世界を目指し，小さな体で巨大な敵に立ち向かう．

もちろん，これは現実ではなく，手塚治虫氏原作の漫画「鉄腕アトム」の中での話である．「鉄腕アトム」は，1952 年連載が開始された 21 世紀の世界を描いた SF 漫画．今もテレビのアニメーションが製作されるほどの強い人気がある．話の中でアトムは，ジェット噴射で空を飛ぶ，60 の言語を自由に話す，人間の善悪を見抜く，千倍の聴力で聞く，サーチライトにもなる目で見る，お尻からマシンガンを出して戦う，十万馬力パワー，の七つの能力を持ち，日常生活を人と一緒に暮している．手塚は，そんなアトム誕生の年を 21 世紀初頭の 2003 年とした．

日本でロボットの研究や開発に携わる人は，アトムの影響を少なからず受けたのは間違いない．そのためか，各種マスメディアなどによって 2003 年は「ロボット元年」と位置づけられ，実際に数多くの人型ロボットが 2003 年を目標に開発された．その代表がホンダの ASIMO[1] やソニーの QRIO（SDR-4X）[2] である．これらのロボットが産業用ロボットと違うことは，活躍の場が工場ではなく，人が普段生活している日常社会だということである．そのため，人の活動空間内で動きやすく，人が親しみやすいように，多くのロボットが二足歩行機能を持ち，人に似た容姿をしている．この進化の中でコミュニケーション能力も人と同じであることが求められるのは必然であろう．ボタンやスイッチで構成されたリモコ

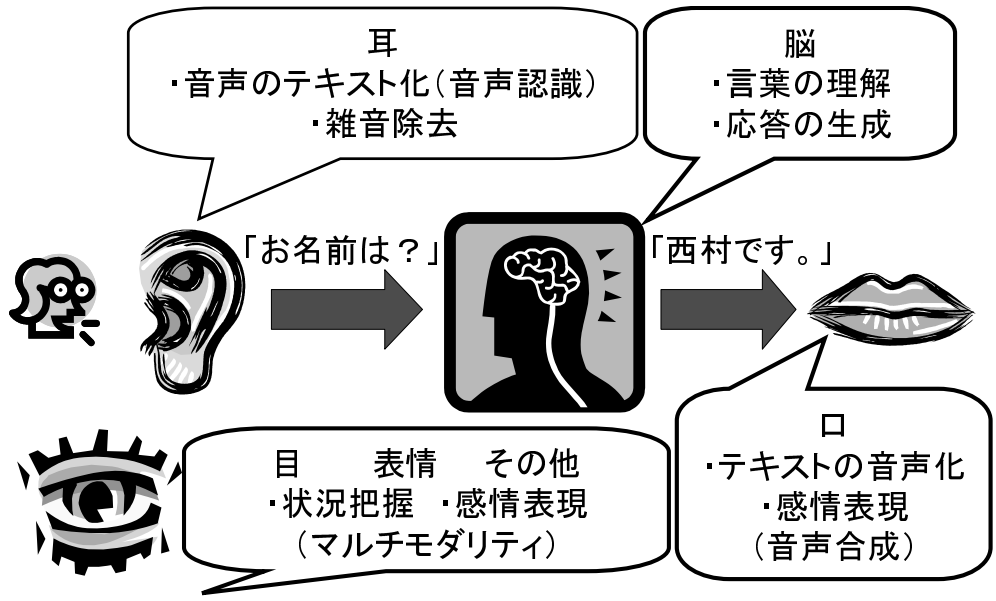


図 1.1　会話（音声コミュニケーション）の構成要素

ンによるコミュニケーションではなく，人と自然に会話（音声コミュニケーション）する能力が機械にも問われるようになってきた．

図 1.1 は音声コミュニケーションを抽象化した概略である．音声コミュニケーションは多くの要素が複雑に絡み合って成立するものだが，主な処理の流れは以下のようになる．

1. まず耳で聞く．計算機で処理する場合は音を文章（テキスト）に変換する音声認識がこれに相当する．このとき，不必要な雑音などは除去される必要がある．

2. 次に脳で考える．つまり，テキストの中身を解釈し，応答を作成する．

3. 最後に口で話す．作成された応答（テキスト）を音声に変換する音声合成技術がこれを担う．

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

さらに，音声コミュニケーションを円滑にすすめるために必要な能力に知能を忘れてはいけない．知能は，辞書によると「環境に適応し，新しい問題状況に対処する知的機能・能力」（広辞苑より）である．知識をもとに物事を解釈するだけでなく，随時新しい知識を吸収する能力をいう．しかし，未知の知識を新たに取得した後，その必要性を自分自身で判断，そして，体系的に吸収するような仕組みの実現は，現在の計算機システムでは非常に難しい．

そこで，知能を必要とするようなシステムは，過去の経験を統計的に抽象化したデータベースに基づく処理で実装されることが多い．図 1.1 の構成要素にある音声認識や音声合成は，その顕著な例であり，隠れマルコフモデルを代表とする統計モデルの導入により飛躍的発展を遂げた技術である．

ただし，統計に基づくアプローチは，抽象化されているとはいえ，過去の経験に依存している以上，本質的な意味での知能とは異なる．そのため，開発者は，利用状況を正しく想定し，必要な経験をあらかじめ計算機に学習させておく必要がある．人と計算機との音声コミュニケーションを成立するためには，計算機上の要素技術はブラックボックスであってはならず，人による面倒なタスク設定が必要であることを忘れてはならない．

ただでさえ複雑な会話，それを構成する要素技術も面倒なタスク設定を前提に成立する．このため，会話能力を計算機上で実装すると複雑な大規模システムになる．これらハードルを乗り越え，人と自然な会話ができるロボットは実現可能なのであろうか．

1.2. 研究目的と背景

本研究の最終的な目的は，人と機械との音声コミュニケーションを可能にする音声インタフェースの利便性向上と普及を達成することである．

これら音声インタフェースに関する研究はすでに多く提案されており，1980 年代から 1990 年代初頭には実際に数多くのシステムが構築された [3]．しかし，当時はハードウェアリソースなどの問題から音声認識の実時間処理は困難であり，実用化に成功したシステムはほとんど無い．ところが，その後の数年間での計算

機の爆発的発展により，実時間に近い処理スピードでの音声認識が可能となった．IBM の ViaVoice など商用音声認識プログラムの登場は，音声インタフェースを構築する際の音声認識の敷居を低くした．人型ロボットの登場と前後して，音声インタフェースに対する世間の期待が再び高まったのは，このような背景にもよる．

近年の音声インタフェース実用化の例としては，MIT によるコールセンターを想定した電話音声対話システムが先駆的である [4][5]．日本での同様の試みとして京都大学のグループによるバス運行情報案内システム [6] がある．カーナビゲーションでの音声インタフェースの利用も活発である [7]．これらのシステムは，利用環境がもたらす制限により，音声のみしか意思伝達手段を持たないことが共通する．電話は基本的に音声のみなのは当然として，車の運転中に手による操作は危険だからカーナビと音声インタフェースは相性が良いと考えられている．このとき，インタフェースの役割は，利用者に端的に正しい情報を伝えることが最重要であり，システムにとって都合が良いシステム主導の対話によって対話戦略が構成されることも多い．

一方で，会話ロボット [8][9] や擬人化エージェントシステム [10][11] では，応答の正確さも大事だが，会話の過程に注目をした研究がなされることが多い．松坂らが開発したロボット [12] は，対話相手の顔の向きなどの視覚情報を用いることで，複数の人間が話し合うグループ会話に参加することができる．ATR の音声対話ロボット Robovie[13][14] を用いた研究では，会話の中での非音声情報が人とロボットにもたらす相互作用を調査している点が興味深い．この中ではロボットの誤った応答もコミュニケーションの重要な要素と位置づけられる．

他にも，情報化家電のユーザインタフェースにおいて，音声対話による操作機能を付け加える検討もなされている [15]．

以上のような研究を総じて，音声インタフェースは実用化に近づきつつあると言えるのではないだろうか．確かに，解決しなければならない課題は多々残されているが，現状でも人にとって有用なシステムは十分に構築可能である．しかし，一般の人々が気軽に利用できる音声インタフェースが，我々の日常社会に導入された例はまだ少数にすぎない．これは何故だろう．

まず，社会が要求するインタフェースに対するレベルが高いことに理由がある．

確かに，現金自動預入れ支払い機（ATM）や自動販売機など厳密な正確さが必要とされるシステムへの音声インタフェースの導入は現状では難しい．

音声インタフェースを導入することによってコストを下げることができなければ，システムの開発・運用はビジネスとして成立しない．音声インタフェースの構築は，要素技術をただ組合わせれば良いものではない．利用実態に合わせた各要素技術の複雑なチューニングが必要だが，それを効率良く可能にする手段が確立されていないのが問題である．

機械に対して話しかけることに抵抗を持つ人がいることも理由として挙げられるだろう．ほとんどのケースでは慣れの影響が大きいが，マイクを前に意識すると急に会話ができなくなる人は多い．

その他にも理由は様々であるが，これら原因の究明には，社会学や心理学のアプローチも考慮したうえで，利用実態に基づく調査・検討が必要不可欠である．工学的には，人と機械のインタラクションのデータを収集し，それを分析することがシステム自体や要素技術の改良の過程では必須である [16]．

本研究では，実際に音声インタフェースのアプリケーションを開発したうえで社会に提供する．そして，システムのフィールドテストを通じて利用実態を調査，現在の音声インタフェースが抱える課題を検証することを第一の目的とする．その過程では必要に応じた要素技術の開発や改良を行い，インタフェースとしての利便性向上やシステム開発の低コスト化を目指すことを次の目的とする．同時に必要なデータ収集と整備を行う．

これまで同様の調査を試みた先例もあるが [5][9][17][18]，これから様々な分野で導入されることになる音声インタフェースの応用は多岐にわたり，その多様性に対しての調査の量は依然として十分ではない．むしろ積極的に，様々な状況を想定してシステムの開発をすすめ，調査することが必要になってくるだろう．中には自動販売機のように音声インタフェースの導入に慎重な検討を必要とするものもあるが，より気軽に音声インタフェースを導入できる場面も存在する．まずは，そのような気軽に利用できる音声インタフェースを対象にシステムを開発し，利用実態の調査を始めることにする．

1.3. 本論文の構成

本論文は7章から構成される．研究の主目的である音声インタフェースの利用実態調査については，4章の受付案内ロボットと5章の音声情報案内システムを通じて述べる．特に5章ではフィールドテストで収集したデータを実際に分析する．3章では開発のコスト削減のための技術開発を行い，6章の提案手法はシステムの利便性向上に寄与する要素技術である．

2章以降の各章の概要は以下の通りである．

2章の最初で，大語彙連続音声認識技術に関する基礎知識を導入する．また，音声インタフェースの実現可能性を探るために，人と機械の対話タスクを中心に現在の大語彙連続音声認識性能を調査した．その結果も報告する．

3章では，認識対象タスクに適した言語モデルを半自動に作成する手順を検討する．その手順は，(1) テキストコーパスの自動作成とトピック依存 N-gram モデルの構築，(2) モデル融合によるタスク操作，(3) ネットワーク記述文法の適用によるモデル高精度化，の三段階から構成される．

4章では，人との対話機能を持った受付案内ロボット ASKA について，システムの音声インタフェース部分を中心に具体的に解説し，その開発プロジェクトの活動を総括する．また，ASKA を使った実験で音声認識における言語モデルと文法の性能を比較，文法適用言語モデルの有用性を確認する．

5章では，音声インタフェースのフィールドテストを実施する．その実施に際しては，誰もがいつでも気軽に利用できる音声情報案内システム「たけまるくん」を開発した．その構成と収集発話の分析結果を報告する．評価実験では，システムが抱える問題点の一つとして，子供の利用者に対する性能不足を明らかにする．

6章では，話者年齢層に即した柔軟な対話を可能にした音声インタフェースを検討，5章で従来システムでの課題として挙げた子供に対する利便性向上を目指す．その実装で必要となる話者の年齢層（大人・子供）識別手法を音声認識スコアを素性とする機械学習に基づいて提案する．また，フィールドテスト収集発話を用いて話者年齢層に適応した言語モデルと音響モデルを再構築した．子供発話の音声認識性能を評価し，性能改善を得る．

最後に，7章で本論文を総括し，結論とする．

第2章　大語彙連続音声認識

2.1.　はじめに

本章では，はじめに本研究の基盤となる大語彙連続音響認識の原理とそのシステム構成について述べる（2.2 節）.

音声認識技術は，その認識対象とする単語辞書の大きさによって小語彙（数百語程度）と大語彙（数千語以上，数万語程度）に区分することができる．一般に，特定のタスクドメインを定めることによって，音声認識での認識の対象とする語彙を限定することはある程度可能であり，そのタスク操作は音声インタフェースの利便性に大きな影響を与える．音声インタフェース内で音声認識を利用することを考えると，ユーザが発話した自然な自由文章を認識できなければならない．しかし，ある特定の事柄を指し示す発話においても，その文章表現は多種多様に変化することが自然であり，その中で使用する単語を小語彙に限定することは難しい．また，システムの単語辞書に無い未知語を含んだ発話が入力されたとき，現在の音声認識では認識不可能な未知語が認識誤りを起こすのみではなく，その前後にまで影響が派生することが知られている．結果として，辞書に含まれる単語で構成される誤った文章が出力され，認識後の対話処理に悪影響を与えかねない．例えば，2 万語の単語辞書を用いることで 75 か月分の新聞記事に出現する全単語の約 97%を被覆することができる [19]．固有名詞など残りの単語の対処が必要になる場合もあるが十分な被覆率であると言えよう．つまり，大語彙単語辞書は，実際上必要なほとんどの語彙をカバーすることができ，これは音声インタフェースを構築する際には有利に働く.

一方，初期の音声認識には，単語単位などの離散発話を強制するシステムも存在したが，これでは利用者の負担が大きい．よって，自然文章発話を認識できる

連続音声認識は我々の目標とする音声インタフェースでは必須であると考える．

以上をふまえ，本論文においては大語彙連続音声認識を音声インタフェースシステムの核をなす基盤技術として扱うものである．本章は，その大語彙連続音声認識に関する基礎知識の導入を目的とする．

以下，2.3 節及び 2.4 節では，大語彙連続音声認識の重要な構成要素である音響モデル及び単語辞書と言語モデルについて概説する．現在，最も広く用いられている HMM 音響モデルと N-gram 言語モデルを取り上げる．

2.5 節では，音声認識の評価手法についてまとめる．音声インタフェースの性能評価では，システムの利便性やユーザの満足度，タスク達成時間などを評価することが求められる [3]．その手法としてアンケートによる主観評価がこれまで用いられてきた．しかし，5 章で実施するように，音声インタフェースのフィールドテストが本研究の主目的の一つである．フィールドテストの途中にアンケートを実施すると，利用者は気軽にシステムに接することを敬遠し，集めたデータにシステムの利用実態を正確に反映できなくなる恐れがある．このため，フィールドテストで収集した発話を使いシステムの音声認識や応答生成の精度を求めることで，音声インタフェースとしての性能も評価する必要がある．その評価尺度には，5 章で述べる応答正解率と併せて，本章の音声認識の評価尺度を用いた．

2.6 節では，大語彙連続音声認識を用いた音声インタフェースの実現性を探るために，現状の音声認識性能を調査する．特に人と機械の対話における音声認識の性能に注目する．

2.2. 音声認識の原理とシステム構成

入力音声をテキスト化する音声認識技術は，入力音声の特徴ベクトルの時系列パターン $\mathbf{X}$（フレーム数: n）を

$$\mathbf{X} = x_1, x_2, \cdots, x_n \tag{2.1}$$

としたとき，$\mathbf{X}$ を観測して最も尤度の高い単語列である

$$\mathbf{W} = w_1, w_2, \cdots, w_m \tag{2.2}$$

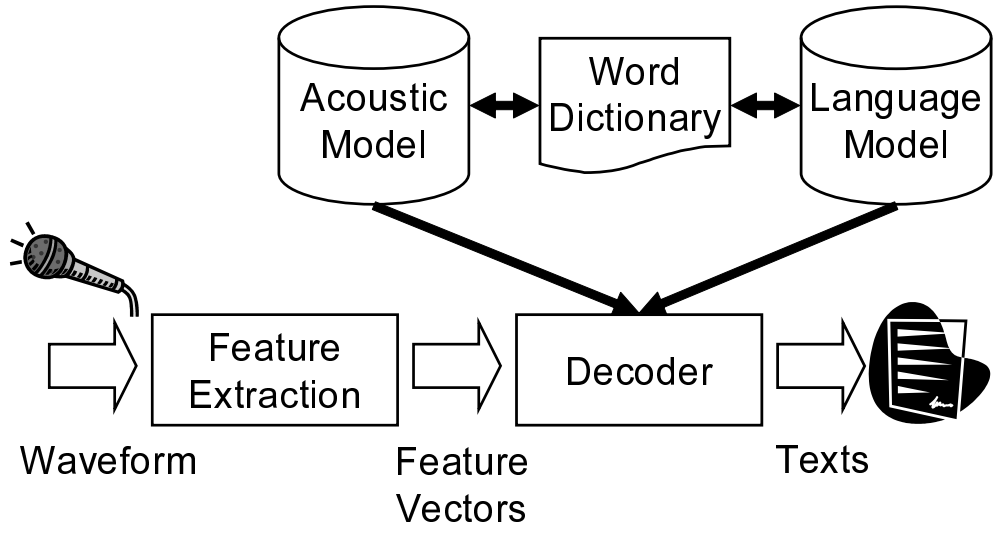


図 2.1 音声認識システム構成

を探索する問題として考えることができる（単語数: m）.

つまり，音声認識は事後確率 $P(\mathbf{W}|\mathbf{X})$ を最大にする単語列 $\mathbf{W}$ を探す問題となる．ここで $P(\mathbf{W}|\mathbf{X})$ に対してベイズの定理を用いると次のように変形ができる.

$$P(\mathbf{W}|\mathbf{X}) \quad = \quad \frac{P(\mathbf{W}) \cdot P(\mathbf{X}|\mathbf{W})}{P(\mathbf{X})} \tag{2.3}$$

(2.3) 式の分母 $P(\mathbf{X})$ は，入力パターン自体の生起確率であり，単語列 $\mathbf{W}$ には無関係である．よって，音声認識は (2.3) 式より,

$$P(\mathbf{W}) \cdot P(\mathbf{X}|\mathbf{W}) \tag{2.4}$$

を最大にする $\mathbf{W}$ を求める問題と考えることができる.

(2.4) 式の $P(\mathbf{W})$ は，単語列の事前確率であり，入力 $\mathbf{X}$ とは無関係な確率である．この単語列の出現確率を与えるモデルが言語モデルである.

$P(\mathbf{X}|\mathbf{W})$ は，単語列 $\mathbf{W}$ を発生したときに，特徴ベクトル時系列 $\mathbf{X}$ が観測される確率で，この計算に用いるモデルは音響モデルと呼ばれる.

図 2.1 に音声認識システムの構成を図示する．音声認識システムは音声分析部（Feature Extraction）とデコーダ（Decoder）から構成される．音声分析部では

表 2.1 本論文で用いる音響分析パラメータ

量子化ビット数	16 bit
サンプリング周波数	16 kHz
フレーム長	25 ms
フレーム周期	10 ms
分析窓	ハミング窓
プリエンファシス	0.97
特徴ベクトル	MFCC 12 次元 +Δ MFCC 12 次元 +Δ パワー （計 25 次元）

入力された音声波形から短時間周波数分析によって特徴ベクトルを抽出する．近年の音声認識では，特徴量として MFCC（Mel-Frequency Cepstrum Coefficient）を用いることが多い．本論文で用いた音響分析パラメータを音声のサンプリング条件と併せて表 2.1 に示す．

デコーダは抽出された特徴ベクトルを入力として，言語モデル（Language Model），音響モデル（Acoustic Model），単語辞書（Word Dictionary）を用いた尤度計算により入力音声のテキスト化を行うプログラムである．デコーダは膨大な認識候補の中から最も尤度の高い解を探索するアルゴリズムによって構成されている．連続音声認識では，単語音声認識などの連続ではない離散発声音声認識と比較すると，仮説の時間方向への曖昧性から出現する仮説数が莫大になる．その探索は複雑かつ困難なものになり，実装には高度なアルゴリズムが要求される．本研究では，このデコーダに大語彙連続音声認識エンジン Julius[20][21] を用いた．単純に 1 パスの処理で認識をすると高精度モデルの適用や複雑な仮説管理に計算が多くなりかねない．Julius では，認識を 2 段階マルチパスに分けて処理することで実時間に近い実行時間での高い認識性能を得ている．Julius の第 1 パスでは，簡易な単語 2-gram モデルを用いてフレーム同期ビーム探索を行う．第 2 パスでは，第 1 パスの認識スコアを先読み情報（ヒューリスティック）として利用した高精

度な単語 3-gram モデルによる A* 探索を行っている．

Julius はオープンソースプログラムとしてそのソースコードとともに仕様は完全に公開されており[1]，その利用実績は多く信頼性も高い．アプリケーションの実装に必要な API 等も整備されており，音声インタフェースの研究の基盤を成すプログラムとして適したプラットホームであると言えよう．

2.3. HMM 音響モデル

音響モデルは，音素の並びを統計的に学習したものであり，HMM（Hidden Markov Mode; 隠れマルコフモデル）によるモデル化が広く使われる [22]．HMM は特徴ベクトル時系列の確率モデルであり，自己遷移を持つ複数の状態間を遷移することで，音声のような長さの一定しない時系列信号を効率良くモデル化することが可能である．HMM の学習には EM（Expectation Maximization）アルゴリズムと呼ばれる最尤パラメータ推定手法が用いられる．

音響モデルは，通常，単純に音素ごとにモデル化を行ったモノフォン（monophone）モデルである．しかし，音素は同じ音素であっても調音結合の影響により後続する音素によって変化する．この対処として前後の音素環境を考慮した三つ組み音素をモデル単位として利用する．その音素環境依存モデルのことをトライフォン（triphone）モデルと呼ぶ．前後の音素環境を考慮するためトライフォンモデルは高い認識精度を得ることができるが，音素の組合わせにともなって HMM のモデル数が増大するため，認識に必要な計算量は大量になる．また，膨大な数のすべてのトライフォンに対して十分な学習用音声データを確保することは困難である．そこで，状態ごとに音響的特徴が近いトライフォンを共有し，モデル数を削減した状態共有トライフォンが用いられる．一方，HMM では状態ごとに混合正規分布を用いて出力確率分布を構成するため，異なるモデルや状態間で混合正規分布のための正規分布を共有することで効率的なモデルを構成することができる．これを Tied-Mixture モデルと呼ぶ．特に中心音素が同一であるトライフォンの間だけで分布を共有する手法は，PTM（Phonetic Tied-Mixture; 音素内タイ

[1] http://julius.sourceforge.jp/

ドミクスチャ）モデル [23] と呼ばれる．PTM モデルは，モノフォンモデルから出力確率分布を，状態共有トライフォンモデルから状態共有構造を抽出し，出力確率分布を状態共有構造に重みづけを行い作成される．

以後，本研究で用いる音響モデルは，特に断りがない限りこの PTM モデルで統一するものとする．使用した音素数は 5 種類の母音と約 20 種類の子音に長母音，拗音などを加えた合計 43 音素である．音響モデルの学習には HTK[24] を使用し，ファイルフォーマットは HTK 形式のものに準ずる．

2.4. 単語辞書と N-gram 言語モデル

音響モデルが話者性や音声入力環境などの音声認識における音響的特徴を担うものであるに対し，言語モデルと単語辞書は，言い回しなどの文章表現や認識対象単語などの言語的特徴を定めるものである．言い換えれば，言語モデルは音声認識システムにおいて認識対象となるタスクを決定する要素である．言語モデルと単語辞書の中で定義されていない文章表現や単語を音声認識では受理することは困難なので，音声インタフェース内で使用することもできない．多様な言語的特徴を柔軟に受理できる言語モデルを使用することが求められる．

単語辞書は，単語のエントリ表記と出力上の表記及び音素記号列から構成される．単語辞書ファイルの例を図 2.2 に示す．ファイルのフォーマットは HTK 形式である．言語モデルの学習元であるコーパス中に出現する異なり単語数は，膨大な数になってしまうため，計算量や記憶量などシステムの実装上の問題から使用する単語数を制限する必要がある．一般にコーパスに出現する頻度が高い単語を上位数千から数万のオーダで限定して単語辞書に使うことが多い．コーパスに出現した単語でもこの辞書に登録されてない単語やその読み（音素記号列）は未知語として扱われ，音声認識することはできない．このため，単語辞書の内容は認識性能に大きな影響を与える．

音声認識の言語モデルには，文脈自由文法などの有限状態ネットワークで記述された記述文法やコーパスから統計的な手法によって確率推定を行う統計的言語モデルが用いられる．通常，自動販売機などの狭い認識対象に限定しても良いタ

```
</s>     []       silE
<s>      []       silB
、+、+79/0/0     [、]     sp
日報+ニッポー+2/0/0      [日報]  n i q p o:
日没+ニチボツ+2/0/0      [日没]  n i ch i b o ts u
日本+{ニッポン/ニホン}+12/0/0   [日本]  n i q p o N
日本+{ニッポン/ニホン}+12/0/0   [日本]  n i h o N
```

図 2.2 単語辞書ファイルの例

スクの場合には，文法記述型の音声認識を用いることが多い．しかし，記述文法では，システムはあらかじめ想定された文法内の発話のみしか受理できず，発話の文章表現や語尾などの発話様式までも限定されてしまう．認識対象語彙が比較的小さくなる事や複雑な文法を開発者が記述する必要があるなどの問題点も知られている．一方，統計的言語モデルを用いた大語彙連続音声認識では，認識結果を開発者があらかじめ決定的に定義することは難しく，一見するとアプリケーションに組み込むのには向かない．しかし，柔軟に様々な発話を受理することが可能であるため利用する価値は高い．そこで，本研究で開発する音声インタフェース内では統計的言語モデルを採用することにした．

現在，音声認識において最も良く利用される統計的言語モデルは単語 N-gram モデルである．単語 N-gram モデルは単語連鎖のマルコフモデルで構成され，単純なモデルでありながら効果が大きい [25]．

言語モデル $P(\mathbf{W})$ による m 単語からなる単語列 $w_1 w_2 \cdots w_m$ の生起確率は以下の (2.5) 式で表すことができる．

$$P(w_1 w_2 \cdots w_m) \;=\; \prod_{i=1}^{m} P(w_i | w_1 w_2 \cdots w_{i-1}) \tag{2.5}$$

しかし，この確率を推定するのは現実的には不可能であるため，N-gram モデルでは，ある単語の生起が直前の N-1 単語の生起にのみ依存するという近似によって，単語列の生起確率を推定する．つまり，(2.5) 式は以下のように近似すること

ができる．

$$P(w_1 w_2 \cdots w_m) \fallingdotseq \prod_{i=1}^{m} P(w_i | w_{i-N+1} \cdots w_{i-1}) \tag{2.6}$$

各々，N=1 のときは，1-gram (unigram)，N=2 のときは，2-gram (bigram)，N=3 のときは，3-gram (trigram) と呼ばれ，現在の音声認識の言語モデルでは 2-gram または 3-gram モデルを使用することが多い．Julius では第 1 パスで 2-gram モデル，第 2 パスで逆向き 3-gram モデルを用いる．

(2.6) 式の条件付き確率は，コーパス中に出現する N 個の単語列と N-1 個の単語列の出現回数から最尤推定により算出できる．例として 3-gram 確率の推定方法を考えよう．単語列 $w_{i-2}w_{i-1}w_i$ を w_{i-2}^i と書き，そのコーパス中の出現回数を $C(w_{i-2}^i)$ で表すと，3-gram 確率 $P(w_i|w_{i-2}^{i-1})$ は，

$$P(w_i | w_{i-2}^{i-1}) = \frac{C(w_{i-2}^{i-1})}{C(w_{i_2}^i)} \tag{2.7}$$

となる．

この推定方法では，コーパスに出現しなかった N-gram（N 個の単語の単語列）の出現頻度が 0 となると，その N-gram の単語列の生起確率も 0 になる．しかし，これはたまたま学習に使用したテキスト中に出現しなかったものであり，値をなんらかの方法で推定する必要がある．この推定には確率の平滑化の一種であるバックオフ平滑化 [26] が用いられる．バックオフ平滑化では，コーパスに出現しなかった N-gram 確率を (N-1)-gram 確率から推定する．最尤推定での確率を

$$f(w_i | w_{i-2}^{i-1}) = \frac{C(w_{i-2}^{i-1})}{C(w_{i_2}^i)} \tag{2.8}$$

とするとき，バックオフ平滑化後の確率は次式で推定される．

$$P(w_i | w_{i-2}^{i-1}) = \begin{cases} \lambda(w_{i-2}^i) f(w_i | w_{i-2}^{i-1}) & \text{if } C(w_{i-2}^i) > 0 \\ (1 - \lambda_0(w_{i-2}^{i-1})) \alpha P(w_i | w_{i-1}) & \text{else if } C(w_{i-2}^{i-1}) > 0 \\ P(w_i | w_{i-1}) & \text{otherwise} \end{cases} \tag{2.9}$$

ここで λ はディスカウント係数である．言語モデルの学習時にすべての確率値を出現 N-gram に割り振らずに，あらかじめ確率値を割り引いておき，その確率値

を未観測 N-gram 集合の確率の推定値として割り当てるために使われる．ディスカウント係数 λ は，経験的もしくは統計的に求める様々な方法が提案されているが，本研究では Witten-Bell 法 [27][28] を用いた．また，

$$\lambda_0(w_{i-2}^{i-1}) = \sum_w \lambda(w_{i-2}^{i}) f(w_i|w_{i-2}^{i-1}) \tag{2.10}$$

である．α は，確率の総和を 1 にするための正規化係数であり，

$$\alpha = \left(1 - \sum_{C(w_{i-2}^{i})>0)} P(w_i|w_{i-1})\right)^{-1} \tag{2.11}$$

となる．

前述のように単語 N-gram モデルの学習はコーパスに出現した単語を数え上げることで推定する．しかし，一般に日本語のテキストは単語ごとに分割して記述されていないので，分かち書きされた学習用テキストへの変換が必要である．本研究では，形態素単位の分かち書き処理に形態素解析システム ChaSen[29] を用いた．同時に「日本語ディクテーション基本ソフトウェア（99 年度版）」[30] に含まれる読み付与変化プログラム ChaWan 及び数字読み付与プログラムを使用して，単語辞書に必要な発音文字の付与を行う．分かち書き学習用テキストからの単語 N-gram モデルの学習には Palmkit[31] を用いた．作成するモデルは ARPA 形式の単語 2-gram モデルと逆向き単語 3-gram モデルである．Julius ではこの ARPA 形式の言語モデルをそのまま扱えるが，認識処理の高速化のために Julius 独自バイナリ形式の言語モデルへの変換も行った．

2.5. 評価手法

本節では，本論文内で用いる音声認識の性能評価法について述べる．以下で述べる評価法はすべて評価用のテストセットに対する性能尺度であるため，テストセットは音声インタフェースの実際の利用状況を十分に反映した発話の集合でなければならない．

2.5.1 未知語率

テストセットに対する言語モデルの未知語の割合が未知語率（OOV; Out Of Vocabulary rate）である．前述のように単語辞書に含まれない未知語は認識不可能なため，未知語率は十分に小さい事が求められる．未知語率（OOV）の算出式を (2.12) 式に示す.

$$\text{未知語率}（OOV） = \frac{\text{テストセットに対する未知語の出現回数}}{\text{テストセットの総単語数}} \quad (2.12)$$

2.5.2 テストセットパープレキシティ

言語 L における単語列 $w_1 \cdots w_m$ の生成確率を $P(w_1 \cdots w_m)$ とすると，言語 L の単語あたりのエントロピーは,

$$H(L) = -\frac{1}{m} \sum_{w_1 \ldots w_m} P(w_1 \ldots w_m) \log_2 P(w_1 \ldots w_m) \quad (2.13)$$

と表すことができる．(2.13) 式は，言語から生成される単語を特定するために必要な情報量であり，ある時点での単語の後に等確率で接続する $2^{H(L)}$ 個の単語の候補があることを示している．よって,

$$PP = 2^{H(L)} \quad (2.14)$$

は，情報理論的な意味での単語の平均分布数を表しており，パープレキシティ（Perplexity）と呼ばれる．

音声認識の言語モデルの評価にはテストセットの書き起こしテキストに対するパープレキシティが用いられる．これを言語 L をテストセットにした場合のテストセットパープレキシティと呼ぶ．パープレキシティが低いとテストセットに含まれる単語列が出現する確率が高く，テストセットに対して高い性能を持つ言語モデルであると言える．

しかし，パープレキシティによる評価は必ずしも音声認識の性能には結び付かない．これはパープレキシティがある時点に生起した単語に等確率に接続する単

語候補数を表したものであり，そのある時点での単語自体の間違いやすさという基準が含まれないためである．一般に単語数が小さいほど一つの単語に割当てられる確率は大きくなるからパープレキシティは低下する．一方，未知語率が小さい言語モデルではパープレキシティが増大することが多い．厳密には単語数や未知語率が異なる言語モデルをそのまま比較することは難しい．この問題を対処した未知語率を考慮する補正パープレキシティも提案されている [26].

2.5.3 音声認識率

実際に大語彙連続音声認識実験を行うことで性能評価する．評価尺度には，(2.15) 式で表す単語正解率（Word Correct）と (2.16) 式で表す単語正解精度（Word Accuracy）を用いる．ここで W は単語数，S は置換誤り，D は脱落誤り，I は挿入誤りの単語数を表す.

$$\text{単語正解率}（Corr.） = \frac{W - S - D}{W} \tag{2.15}$$

$$\text{単語正解精度}（Acc.） = \frac{W - S - D - I}{W} \tag{2.16}$$

単語正解率と単語正解精度では，無音区間などにわき出した単語による挿入誤りを算出に含めるか否かが異なる．純粋に音声認識の評価をする場合，単語正解精度を用いることが多いが，音声インタフェースで良く使われるキーワードマッチの枠組みでは，挿入誤りが存在してもキーワードさえ正しく認識できれば良いと考えることもできる．よって，単語正解率と単語正解精度の両方を求めることが望ましい.

なお，単語正解率と単語正解精度は，認識結果と正解ファイルとの自動比較の結果より求めることができる [32].

2.6. 人と機械対話における大語彙連続音声認識

表 2.2 は，四種類の異なった状況下において発話された成人音声を Julius で音声認識した際の結果をまとめたものである．PP はテストセットパープレキシティ

表 2.2 大語彙連続音声認識の性能評価（単語数 20k）

タスク	文体	発話様式	PP (3-gram)	OOV [%]	Corr. [%]	Acc. [%]
新聞記事 [26]	文語体	読み上げ	50.5	4.0	91.1	89.7
医療相談	口語体	読み上げ	26.5	0.3	89.9	88.0
車の商談（人と人の対話）	口語体	自然発話	51.2	1.6	52.8	42.3
ASKA（人と機械の対話）	口語体	自然発話	10.1	1.1	86.3	83.5

（3-gram），OOV は未知語率，Corr. は単語正解率，Acc. は単語正解精度を示す．使用した音響モデルは，性別非依存の PTM モデルである．言語モデルには各タスクドメインのコーパスから作成した単語数 2 万語の単語 N-gram モデルを用いた．

表中の “新聞記事” は，新聞記事の文章を朗読調に読み上げた発話を評価した際の結果である．“医療相談” は，健康に関する質問（相談）のテキストを用意し，それを読み上げたときのものである．このとき，テキストは新聞記事と違い口語調で書かれている．“車の商談” は，店員と客との車の購入に関する対話の音声データの評価結果である．疑似対話音声とは大きく異なり日常の会話と同様な自然な人と人の対話音声で構成される．話者によっては発話が明瞭ではなく，「えーと」「あのー」といったフィラーや言い間違い等の認識が困難な発話も多く含むのが特徴である．実験には RWCP 音声データベース [33] の対話音声データを用いた．“ASKA” は，我々が開発した奈良先端科学技術大学院大学情報科学研究科の受付案内ロボット ASKA（4 章参照）の利用を想定した受付案内に関する利用者の質問発話である．発話者が話す相手はロボットであると意識して発話した，人と機械の対話音声である．

この結果から新聞記事の読み上げは 90%程度認識できることがわかる．口語体でも読み上げ音声（医療相談）なら新聞読み上げとほぼ同程度認識可能である．読み上げ音声には言い間違いや言い淀みが含まれず，比較的はっきりとした明瞭

な発話であるので認識することができる．一方，人と人の完全な自然対話（車の商談）では，著しい認識率の低下が見られた．確かに認識結果には言い間違いや言い淀みによる誤りも発生している．さらに，発話速度のゆらぎや発話の明瞭性など音響的な影響も大きく，それら要因が複雑に影響することで人の本当に自然な発話の認識は難しいものにしている．これは今回使用したデータに限った話しではなく，南條ら [34] は，自由発話である学会等の講演音声認識の難しさを報告している．

それでは，人と機械の対話音声の認識精度はどの程度だろうか．結果（ASKA）は，新聞読み上げと比較して認識率 5%程度は低いが，高い精度を示している．実際にシステムを使用する場面においては，ユーザはある目的を達成したいからシステムと対話する．人は機械と対話するとき，自分の音声を認識して欲しいし，理解してほしいので，機械に対して協調的な態度になる傾向がある．発話は読み上げに近い比較的丁寧な発話になるのだろう．丁寧な話し言葉は，大人が子供に話しかけるような明瞭な発話になるので音声認識は比較的簡単だと言える．また，現状の音声インタフェースはどのような対話にも答えることができるわけではないので，人と機械のやりとりがある程度パターン化され，多様性が少なくなるのも理由である（パープレキシティが小さくなる）．この結果は，大語彙連続音声認識が人と機械の対話である音声インタフェースに現実的に利用可能であることを示している．

なお，この ASKA を想定した人と機械の対話の評価実験は予備的実験であり，あまり多くのデータを収集していない（計 580 文）．データが少ないと，テストセットは，システムの利用実態を十分に反映していない可能性が高い．5 章では，フィールドテストで大規模に収集した発話を使って人と機械の対話認識の評価を行い，さらに検討をすすめる．

2.7. 本章のまとめ

本章では，大語彙連続音声認識の原理を述べ，音声認識システム構成の解説を通じて本論文で必要になる基礎的知識の導入を行った．同様に音声認識の評価手

法（未知語率，テストセットパープレキシティ，単語正解率，単語正解精度）について述べた．

本研究で開発する音声インタフェースでは，統計的言語モデルを用いた大語彙連続音声認識を採用することで，自然発話が持つ文章表現の多様性に対する頑健さを確保する．その統計的言語モデルには，現在最も広く用いられている単語 N-gram モデルを使用する．

大語彙連続音声認識の性能を四種類の異なった状況下の発話に対して調査し，現状をまとめた．この結果，人の本当に自然な発話の認識は依然として難しいが，人と機械の対話のような丁寧な話し言葉に対しては，ある程度の高い精度を得ることができるのがわかった．つまり，音声インタフェース内で大語彙連続音声認識が現実的に利用可能であるのを確認することができたと言える．

第3章　N-gram言語モデルにおけるタスク適応手法

3.1.　はじめに

2.4節で述べたようにN-gram言語モデルはコーパスから単語の並びを統計的に学習し，大語彙を容易に扱うことができる．このため，多様な文章表現を柔軟に受理可能である．しかし，コーパスに含まれる語彙や語尾様式，発話表現などの言語的特徴は作成された言語モデルにも反映される．音声認識で高い認識精度を得るには，入力音声と言語モデルの言語的特徴が近くなければならず，一つの言語モデルであらゆるタスクの語彙や表現をすべて網羅することは難しい．

例えば，IPA（情報処理振興事業協会）の「日本語ディクテーション基本ソフトウェアの開発」プロジェクト[30]では，言語モデルの学習に新聞記事コーパスを用いた．新聞記事は容易に大量のデータを収集できるが，完成した言語モデルは新聞記事の書き言葉を反映したものになり，日常会話で使用するような話し言葉を対象としたタスクでの認識には適さない．話し言葉を認識するためには話し言葉のテキストコーパスからの言語モデルの学習が必要であり，現在も話し言葉テキストコーパスの整備が求められている[35]．

また，音声インタフェースで想定されるタスクは多種多様であるため，それら個々のタスクに適した言語モデルを提供できることが望ましい．ところが，認識対象に合致する大量の学習元コーパスが常に得られるとは限らない．言語モデルの学習用テキストの作成には，対象タスクの話題に応じたコーパス収集の他にも，収集テキストの整形などコストが高い作業を要する．そのため，想定されるあらゆるタスクごとに言語モデルを準備するのは量的に限界があり現実的ではない．

以上のような理由から，効率良く低コストに認識対象タスクに適したN-gram

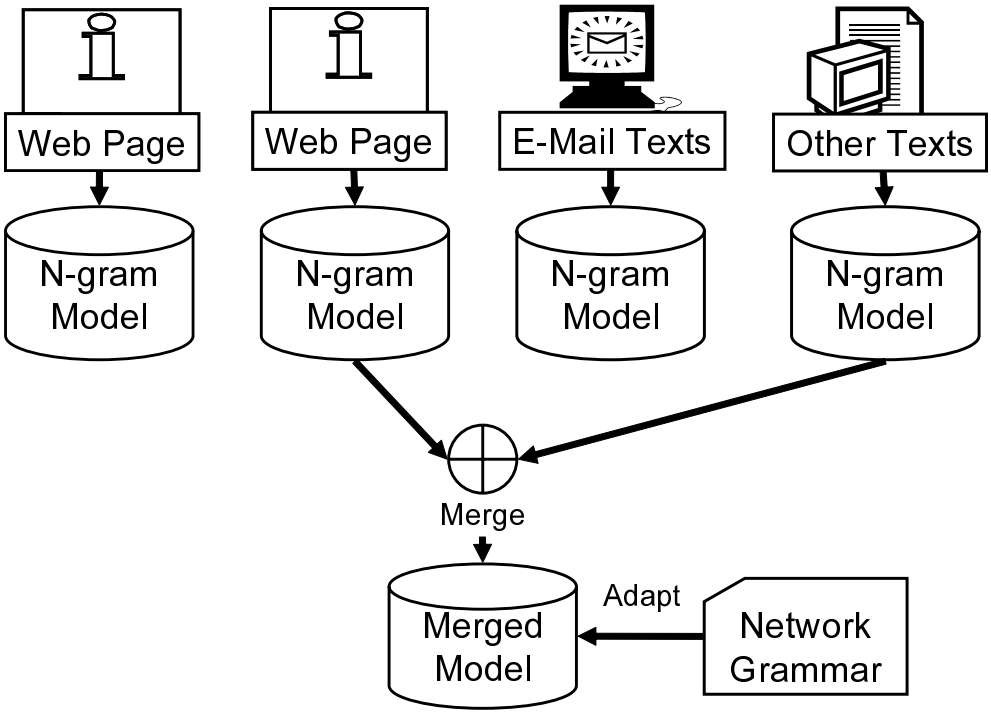


図 3.1 タスク適応 N-gram 言語モデルの構築手順

言語モデルを作成する手段の確立が求められてきた．そこで，本研究では，タスク適応 N-gram 言語モデルの作成過程のほとんどを自動化可能なモデル構築の手順を開発する．その手順は以下の三段階から構成される．概略を図 3.1 に示す．

1. N-gram モデルの学習に必要なテキストを自動収集する．このとき，テキストはインターネットの代表的なアプリケーションである World Wide Web（以下，Web と略）の Web ページやメール，その他のテキストコーパスから抽出する．収集した学習用のテキストが，ある特定のトピックに偏ったコーパスになるなら，構築される N-gram モデルもそのトピックに依存したモデルとなる．

2. 認識対象タスクに関連するトピックを持つモデルを作成した前述のものの中から選択，互いに融合する．以上の手順により言語モデルのタスク操作を実現する．

3. 高い認識精度を得るために，別途作成したネットワーク記述文法を融合モデルに適用し，タスク適応 N-gram 言語モデルは完成する．

次節以降では，トピック依存コーパスのためのテキスト収集と N-gram モデルの構築（3.2 節），モデルの融合（3.3 節），ネットワーク記述文法の適用技術（3.4 節）を中心に，その詳細を述べる．

3.2. Web 検索を用いたトピック依存 N-gram モデルの作成

3.2.1 トピック依存 Web ページの収集

N-gram 言語モデルの学習には大量のテキストコーパスを用意する必要がある．3-gram の可能性は単語数の三乗個存在するが，大語彙になるとその数は膨大になる．コーパスに出現しない未観測の 3-gram 確率に対するバックオフ平滑化を用いた補正も難しくなるため，コーパスには十分な数の 3-gram が含まれる必要がある．しかし，コーパスの整備は多大な労力が必要であり簡単にできるものではない．ある特定のトピックに依存したテキストを大量に集めるのは特に困難である．本研究では Web の検索サービスを用いた Web ページの大規模自動収集でこの解決を試みる．

図 3.2 に，その大規模自動収集の概念を示した．インターネット上には数多くの Web ページの検索サービスが提供されている．これらは入力されたキーワードと関連の高い Web ページへのリンクを提供するサービスである．作成する N-gram モデルのトピックに関連するキーワードで検索サービスを利用，検索結果を集めることで，そのトピックに関連する Web ページを効率良く収集することができる．そして，それらを学習元コーパスとしてトピック依存の N-gram モデルを構築する．実際の検索サービスでは，検索結果の数に上限があることが多いが，検索結果の各 Web ページのリンク先をたどって取得することで関連する Web ページをさらに大量に取得できる．よって，N-gram モデルの学習に十分な量のテキストを含むコーパスの自動作成が可能である．

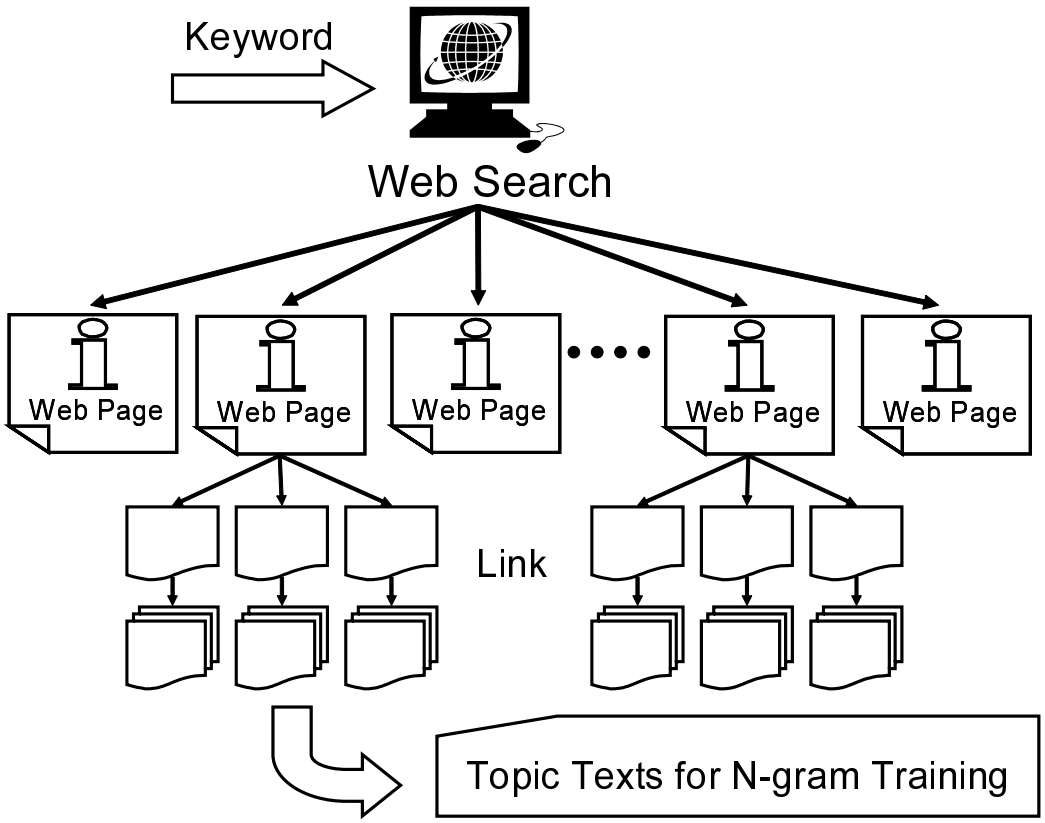


図 3.2 Web 検索を利用したトピック依存 Web ページの大規模自動収集

Web ページは，新聞記事などに比べて話し言葉に近いテキストを多く含むので，本手法による言語モデルは話し言葉の音声認識にもある程度対応できると期待する．また，本手法はテキストの収集から言語モデルの作成までの自動化を可能とし，容易な言語モデルの更新を実現するものでもある．

Web リソースを用いた N-gram 言語モデルの高精度化手法には，Zhu ら [36] による手法が挙げられる．Zhu らは，Web 検索サービスに直接 N 個組みの単語列（N-gram）をキーワードとして入力し，その結果ヒットした Web ページ数を根拠にした N-gram 確率の推定を定式化している．しかし，この手法は Web 検索サービスの仕様に強く依存するためツールには向かない．さらに，N-gram をキーワードにしたため Web 検索の範囲が限定されてしまい，トピックに関連する Web ページを広くカバーすることは困難である．一方，本手法は単純ではあるが，関連す

るだろう Web ページを広く網羅できることに利点がある．

3.2.2 統計的テキストフィルタによる整形

学習用テキストを作成する際は，作成したコーパスから必要とするテキストを抽出する必要がある．従来，新聞記事などの比較的整ったコーパスからの学習用テキスト作成においても，テキストの抽出過程でテキスト整形処理は必要とされてきた [37]．実際には発話されないヘッダや記号などの削除を行うことで，未知語率の上昇を防ぎ，言語モデルの性能向上に効果があるためである．

従来のルール記述型のテキスト整形フィルタでは，正規表現の列挙などで記述したルールに基づいて処理を行う．ヘッダやタグなどの明確な定義があるものについては，容易にルールを記述できる．新聞記事のような整った文章であれば，ある程度，普遍的なルールも作成できるため，従来型フィルタによる整形処理は可能であった．しかし，Web ページからのテキストの整形に従来型のフィルタを使用することは困難である．実際の Web ページテキストには，例えば，メールや投稿記事のシグナチャ（署名），注釈文（括弧による挿入），宣伝文（コンテキストを無視した単語の羅列，記号の多用），箇条書きや羅列（文章の区切りが明確でない）などの整っていない部分が多く含まれる．このような部分を適切に削除，整形する普遍的なルールの作成は限界があり，対象テキストに依存したアドホックな作業に基づかざるを得ない．これは，学習用テキストの作成を困難にする主な原因の一つである．さらに，Web ページのテキストの記述は多様性に富み，記述言語や文字コードも多様であるため，従来型のフィルタで対応するのは不可能である．

この対策として，収集した Web ページの整形処理に統計情報に基づいて日本語文章を判別，抽出する統計的テキストフィルタを導入する．統計的テキストフィルタでは，入力テキストと選別の基準となる言語モデルの言語的特徴の類似度を評価する．つまり，新聞記事など十分に整備された日本語文章らしいテキストから学習した文字レベルや単語レベルなどの統計量で入力テキストの日本語文章らしさを評価し，閾値によって選別することでフィルタリングする．

フィルタリングの手順を図 3.3 に示す．まず，取得した入力テキストを行ごと

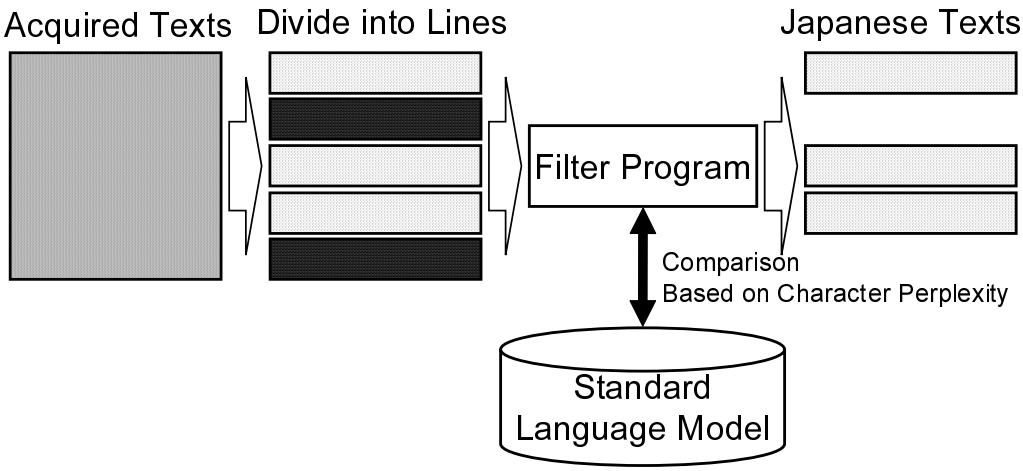


図 3.3 統計的テキストフィルタの概要

に分割する．次に基準言語モデルとの比較から入力行の日本語文章らしさを求める．設定した閾値に適合した行は，日本語文章らしい行であると判別し，学習用テキストに含める．

基準言語モデル（Standard Language Model）には，新聞記事一年分（容量: 95.2MB，異なり文字数: 4890 文字）から作成した文字 3-gram モデルを用いた．文字単位のモデル化を適用したのは，形態素解析にともなう単語の分かち書きの影響を防ぐためである．文字に分解することで新聞記事固有の表現の学習の影響を小さくすることもできる．

入力テキストの日本語文章らしさの尺度には文字パープレキシティを用いる．パープレキシティの値が小さいテキストは基準言語モデルに近いテキスト，つまり，日本語文章らしいテキストとなる．文字 3-gram モデルでは，文字あたりのエントロピーは以下の式で求められる．

$$H(L) = -\frac{1}{n}\sum_{i=1}^{n} p(c_i|c_{i-1}c_{i-2}) \log p(c_i|c_{i-1}c_{i-2}) \tag{3.1}$$

よって，文字パープレキシティは，

$$P = 2^{H(L)} \tag{3.2}$$

として算出することができる．

3.2.3 トピック依存 N-gram モデルの評価

ここまで説明した手法で作成したトピック依存 N-gram モデルの評価を行う．まず，Web ページの大規模収集の結果を示す．比較として手動で商用インターネットプロバイダの Web 掲示板から集めたテキストでの結果も示す．ここで作成する言語モデルの想定トピックは医療相談である．

実験では，一般に広く利用されている Web ページ検索サービス[1]を用いて Web ページの大規模自動収集を行った．検索キーワードにトピックを連想する単語として「医療」を用いて，検索結果からリンクをたどり二階層先までの Web ページを収集した．整形処理では，従来型フィルタにより HTML（Hyper Text Markup Language）のタグを除去した後に，統計的テキストフィルタを利用して文字化けや記号などの削除を試みた．統計的フィルタのパラメータ閾値 P の値には，50，100，200，400，600 を使用し，算出された文字パープレキシティが閾値より大きくなる行を除外した．フィルタをかけずに取得した全テキストを利用した学習用テキストも作成した（P=∞）．

掲示板からの人手収集では，医療相談をトピックとする掲示板を探し出し，保存されているログを従来型フィルタを用いて整備した．人手収集では，適切な掲示板を探す必要があり，自動化には向かない．存在するテキストの量は Web ページ全体に比べてはるかに少なく，N-gram モデルの学習に必要なだけの大量のテキストを収集，整備するには多大な労力を要するのも欠点である．

大規模自動収集したテキストの諸元を表 3.1 に，人手収集したテキストの諸元を表 3.2 に示す．参考として従来から学習元コーパスとして用いられる新聞記事テキスト一年分の値を表 3.3 に示す．

実験の結果，大規模自動収集では総合計で新聞記事二年分相当量のテキストを収集することができた．表 3.1 からわかるように，統計的テキストフィルタの閾値 P の値を小さくすることで除外されるテキストが多くなる．同様に異なり単語数や文章数も少なくなる[2]．閾値 P を大きくすると，異なり単語数の増加量は少なくなる傾向があり，P=200 の時に学習用テキストはすべての収集テキストの語

[1] http://www.google.co.jp/

[2] 閾値 P=∞ の時に異なり単語数や文章数で他の閾値のものより小さい値になるのは，整形してないテキストの場合，プログラムが正しく処理を完了できなかったためである．

表 3.1 大規模収集した Web ページテキストの諸元

閾値 P	50	100	200	400	600	∞
容量（MB）	97.6	148.0	179.6	195.6	200.9	212.0
異なり単語（個）	138287	209623	254077	277200	283986	272005
文章数	2093310	3402167	4161743	4681117	4874011	4473410

表 3.2 人手収集した Web 掲示板テキストの諸元

容量（MB）	11.5
異なり単語（個）	41973
文章数	319498

表 3.3 新聞記事 1 年分テキストの参考値

容量（MB）	92.0
異なり単語（個）	142202
文章数	906106

彙の約 90%の語彙を含むことがわかった．一方で，人手で収集できたテキストは新聞記事一年分の約十分の一相当量であった．

収集したテキストから 2-gram 言語モデル及び逆向き 3-gram 言語モデルを作成し，評価する．学習に使用した単語はテキスト中の出現頻度上位 2 万語とした．

テストセットとして，医療相談に関する話し言葉テキストを使用した．例文を図 3.4 に示す．これらは音声による医療相談対話システムの利用を想定して，実際の対話例を参考に作成した比較的丁寧な話し言葉の文章である．文章数は 150 文，単語数は 1,189 個である．トピック外での傾向を調べるためにグルメ・レシピに関する問合せのテキストも合わせて用意した．こちらは図 3.5 の例のような単語数 2,046 個の 200 文である．

単語 3-gram でのテストセットパープレキシティを図 3.6，未知語率を図 3.7 に示

このへんの内科を教えてください。
お腹が痛いんですけど。
風邪気味なので病院を教えてください。

図 3.4 トピック内（健康相談）テストセットの例

何かおいしいものを教えてください。
煮物料理について教えてください。
ちょっとしたおひたしを作りたいんですけれど。

図 3.5 トピック外（グルメ・レシピ）テストセットの例

す．図中の値は，大規模自動収集したテキストから作成した自動作成（Automatic）モデルと人手収集テキストから作成した人手作成（Manual）モデルでの結果である．新聞記事（Newspaper）は，表 3.3 で示した新聞記事一年分から構築したモデルでの結果である．

結果から，自動作成モデルは新聞モデルよりパープレキシティの値が小さく，高い性能を示すことを確認できた．Web ページに含まれる大量の話し言葉風テキストを学習した結果，話し言葉の特徴をある程度含んだ言語モデルが構築できたためと推測できる．人手作成モデルと比較しても，わずかながらパープレキシティの値を抑えることができた．トピック外での評価ではトピック内と比べて一様にパープレキシティの値が大きくなっており，トピックに対する依存性を確認できた．

未知語率においても自動作成モデルは新聞モデルより低く抑えられている．しかし，自動作成モデルの未知語率は 0.87%（P=200, 400 の時）であり，人手作成モデルとの比較において 0.6%の増加が見られた．収集では，検索結果のリンク先二階層までの Web ページを集めたため，トピックと関連の薄い Web ページを必要以上に収集してしまった可能性が高い．その結果，トピック外の単語も多く含むことなり，異なり単語が増加する傾向がある．よって，単語数 2 万語の N-gram

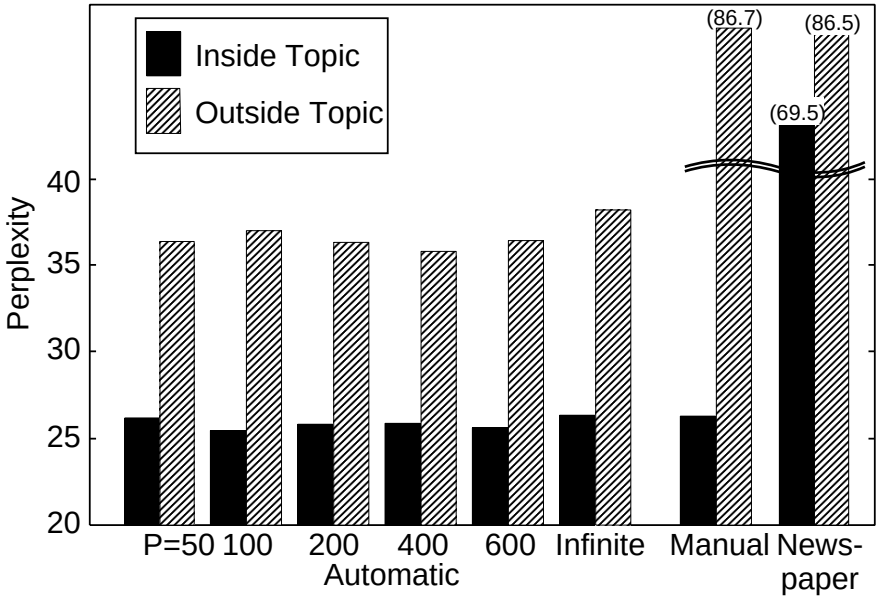


図 3.6 トピック依存 N-gram モデルの評価（3-gram 単語パープレキシティ）

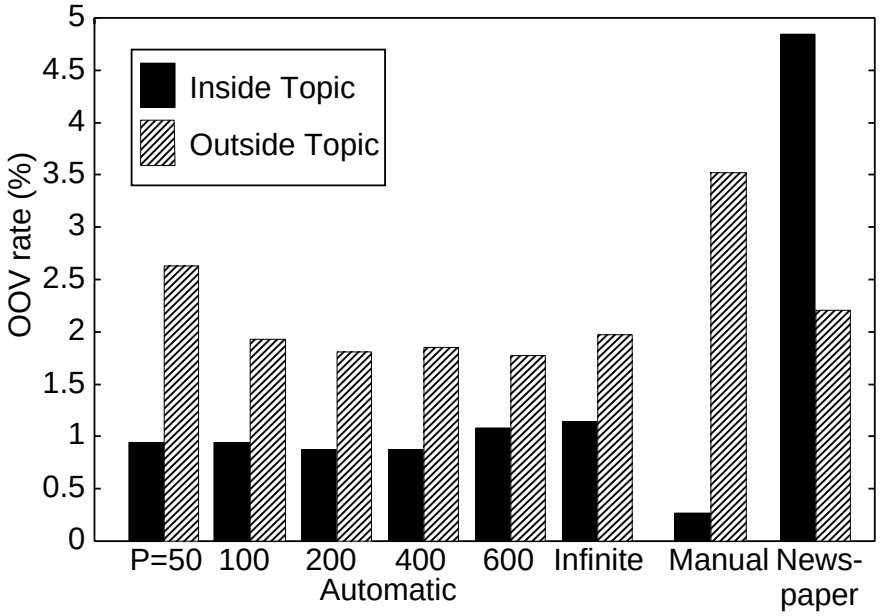


図 3.7 トピック依存 N-gram モデルの評価（未知語率）

モデルでは，学習用テキスト中の全単語に対する被覆率が不十分であり，未知語率の上昇が生じる．

統計的テキストフィルタを適用することで，P= ∞ と比べて若干の未知語率低下が見られた．これはフィルタによって，記号や文字化け文字などの音声認識に不必要な単語が除外され，必要な単語が正しく N-gram モデルの学習に使われたためと考えられる．しかし，この結果ではフィルタの効果は小さく，アルゴリズムの改良が必要であることがわかる．

Julius を用いた大語彙連続音声認識実験を行う．実験に使用した音響モデルは，NEDO「シニア支援システムの開発」プロジェクト [38] において作成された高齢者向け音響モデル [39] の性別非依存モデルである．言語モデルのうち，自動作成モデルの統計的テキストフィルタの閾値には前述の未知語率の評価で最も良い結果を得た P=200 を用いた．評価用音声には，音響モデル同様に高齢者音声を用いた．話者は 60 から 90 歳の高齢者女性 50 人，男性 51 人で，各話者が前述のテストセットの中からトピック内（医療相談）30 文，トピック外（グルメ・レシピ）40 文を読み上げたものである．

単語正解精度を図 3.8 に示す．トピック内評価において，自動作成モデルの単語正解率は 85.6%であり，新聞モデルとの比較で 17.1%の向上である．人手作成モデルとの比較でも 2.4%の低下で抑えることができた．自動作成モデルのトピック外発話での精度は，トピック内より低下しており，トピック依存の傾向を得た．ただし，その差は手動作成モデルよりも小さい．

以上の結果より，本手法により一応のトピック依存 N-gram モデルを構築できたと言える．人手作成モデルの精度には若干及ばないが，自動作成によって手軽にトピック依存 N-gram モデルを構築できる本手法の有効性は高い．トピック依存低下の原因は，大規模収集によりトピックと関連の薄いテキストも学習に含んだためである．しかし，トピック外発話に対してでも，ある程度認識できたほうが音声インタフェースでは都合が良いと考えることもできる．この場合の有効性の検討は，統計的テキストフィルタの改良と合わせて今後の検討事項とする．

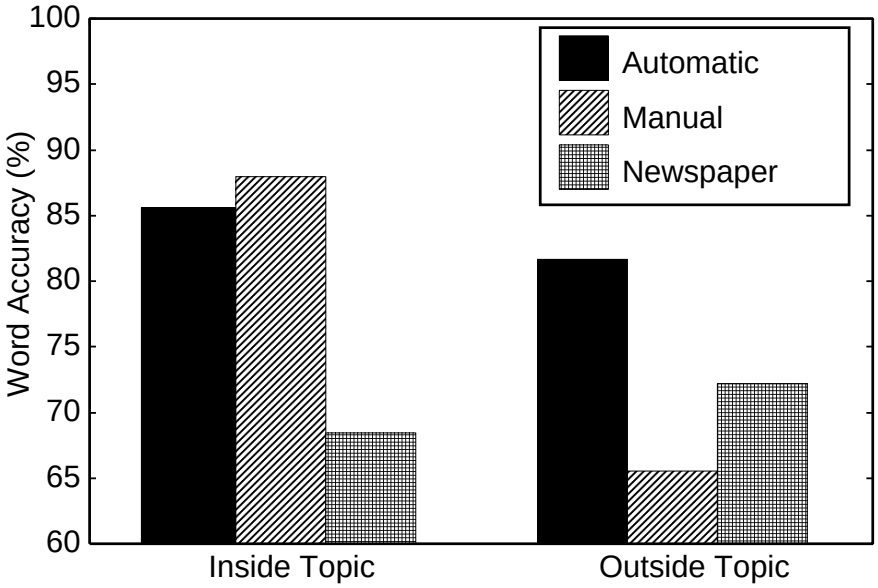


図 3.8 トピック依存 N-gram モデルの評価（単語正解精度）

3.3. 相補的バックオフを用いたN-gramモデル間融合

3.3.1 N-gramモデルのモデル間融合

さまざまな異なるトピックを持つ N-gram 言語モデルを任意に組み合わせることで言語モデルのタスク操作は実現される．ここで重要となる技術が複数の異なる N-gram モデルの融合手法である．最も単純な融合手法は，各々の学習元コーパスをつなぎ合わせて再学習する方法（コーパス結合）である．しかし，コーパス結合は，巨大なコーパスを保持し，それをもとにモデルの再学習をする必要があるため，利便性に欠ける．より扱い易い方法としては，各モデルにおける N-gram 確率を重み付きで内挿する方法が研究されており，特に少量コーパスによる学習データ不足の補間や言語モデルのタスク適応などに用いられている [40][41].

しかし，単純な N-gram 確率の融合では，言語モデル間の未観測な N-gram エントリの不整合の問題が生じる．ある N-gram モデルにとって特徴的な N-gram は，他方にとっては未観測であることが多い．特に固有名詞などの一方にしか現われない語は他方にとっては未知語であり，それらの N-gram 確率も他方にとって

未観測である．そのような未観測 N-gram 確率に基づいて融合を行うことで，融合前のモデルが持つ N-gram 確率の分布が平坦化されてしまい，結果として，トピックに依存したモデルの特徴が薄らいでしまう．これら未観測の N-gram に対して，通常のバックオフ手法を用いて (N-1)-gram から推定することは一応可能であるが，融合相手が持つトピックに特有な単語に対して，自身のモデル内の情報のみから正しい確率を割り当てることは難しい．このような信頼性の低い確率に基づいて融合を行うことは，融合後のモデルの精度低下を引き起こす．

3.3.2 相補的バックオフの原理

本研究では，長友ら [42] が提案した相補的バックオフによる言語モデル間融合手法を用いる．本手法は，言語モデルの融合において不整合を起こす未観測な N-gram を融合相手のモデルから相補的に推定することで，トピックごとの特徴を反映しつつ，より高精度な融合を行なうことができる．以下，その原理を述べる．

N-gram モデルの融合は，各モデルに含まれる各々の N-gram の出現頻度の重み付き和をとることで行われる．以下の議論では，簡単のため，融合する N-gram モデルを二つに限定する．もとになるコーパスを F，G と表し，各々から構築された二つの N-gram モデルを L_f，L_g とする．この時，ある N 単語の組 w_{i-N+1}^{i} の融合後のモデルにおける出現頻度 $C(w_{i-N+1}^{i})$ は，

$$C(w_{i-N+1}^{i}) = \lambda_f C_f(w_{i-N+1}^{i}) + \lambda_g C_g(w_{i-N+1}^{i}) \tag{3.3}$$

となる．ここで $C_f(w_{i-N+1}^{i})$ はコーパス F における出現頻度，$C_g(w_{i-N+1}^{i})$ はコーパス G における出現頻度である．λ_f 及び λ_g は，$\lambda_f + \lambda_g = 1$ の重み係数である．

N-gram モデルでは，ある N 単語の組 w_{i-N+1}^{i} において w^i の出現する条件付き確率 $P(w^i|w_{i-N+1}^{i-1})$ をコーパスに出現する N-gram の頻度 $C(w_{i-N+1}^{i})$ を用いて次式のように求める（2.4 節参照）．

$$P(w^i|w_{i-N+1}^{i-1}) = \frac{C(w_{i-N+1}^{i})}{C(w_{i-N+1}^{i-1})} \tag{3.4}$$

これより，融合後のモデルにおける出現確率は

$$P(w^i|w_{i-N+1}^{i-1}) = \frac{C(w_{i-N+1}^{i})}{C(w_{i-N+1}^{i-1})} = \frac{\lambda_f C_f(w_{i-N+1}^{i}) + \lambda_g C_g(w_{i-N+1}^{i})}{\lambda_f C_f(w_{i-N+1}^{i-1}) + \lambda_g C_g(w_{i-N+1}^{i-1})} \tag{3.5}$$

となる．

通常，言語モデルには出現確率とバックオフのための情報のみが保持され，頻度情報は含まれない．この場合でもコーパスに含まれる総単語数 $C(*)$ が分かれば，$C(w_{i-N+1}^{i})$ は，

$$C(w_{i-N+1}^{i}) = C(*)P(w^{i-N+1})\ldots P(w^{i-1}|w_{i-N+1}^{i-2})P(w^{i}|w_{i-N+1}^{i-1}) \tag{3.6}$$

として計算できる．(3.5) 式及び (3.6) 式から，融合においては各コーパスの大きさ $C_f(*), C_g(*)$ が必要となるが，実際にはそれらの比を考慮した重み λ_f 及び λ_g を与えればよい．よって，N-gram モデルが頻度情報を保持してなくてもモデル間の融合が可能になる．

次にある言語モデルが与えられた時，そのコンテキスト w_{i-N+1}^{i-1} の未観測確率値を $P(*|w_{i-N+1}^{i-1})$ と表す．あるコンテキストに属するすべての N-gram の出現確率の和は 1 であるから，この値は

$$P(*|w_{i-N+1}^{i-1}) = 1 - \sum_{C(w_{i-N+1}^{i})>0} P(w^{i}|w_{i-N+1}^{i-1}) \tag{3.7}$$

として算出できる．ある二つのモデルの融合を考えるとき，一方のモデルで観測されているが他方では未観測である N-gram 及び双方で未観測である N-gram を考慮する必要がある．融合時には，前者の互いに未観測な N-gram については確率の推定を行いその推定値をもとに融合を行うが，後者は融合時には未知であるので，融合後のモデルにおいて確率を推定することになる．このため，融合時に互いに未観測 N-gram の確率を推定する際には，融合後に未観測な N-gram 集合に対して割り当てる確率値を残しておく必要がある．すなわち，融合において推定した確率の合計が，上記の未観測確率値 $P(*|w_{i-N+1}^{i-1})$ よりも小さくなければならない．この推定するすべての未観測 N-gram の確率の総和が，もとの未観測確率値に占める割合を $\beta(w_{i-N+1}^{i-1})$ で表す．

未観測 N-gram の確率 $\hat{P}(w^{i}|w_{i-N+1}^{i-1})$ を求める際には，$\beta(w_{i-N+1}^{i-1})P(*|w_{i-N+1}^{i-1})$ を統計情報に従って分配する．その分配率を $\gamma(w^{i}|w_{i-N+1}^{i-1})$ とすると，この未観測 N-gram の出現確率は次式で表わされる．

$$\hat{P}(w^{i}|w_{i-N+1}^{i-1}) = \gamma(w^{i}|w_{i-N+1}^{i-1})\beta(w_{i-N+1}^{i-1})P(*|w_{i-N+1}^{i-1}) \tag{3.8}$$

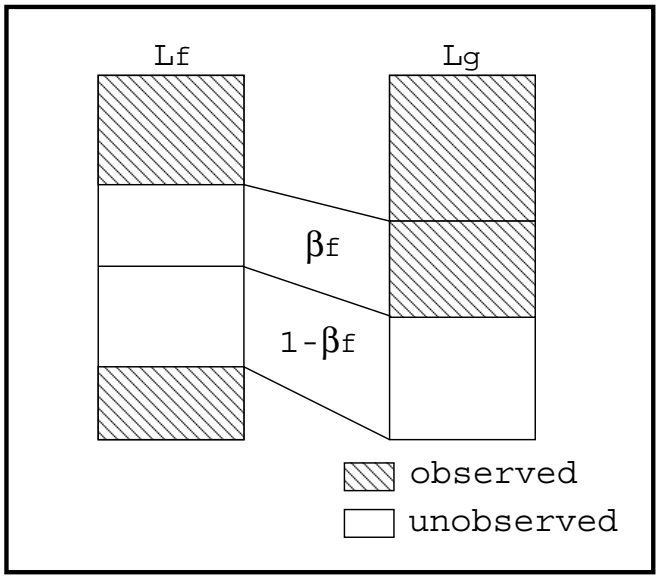


図 3.9 β の推定

よって，未観測 N-gram 確率の推定は，二つのパラメータ β，γ の設定問題に置き換えられる．

相補的バックオフにおけるパラメータ β の推定法を図 3.9 に示す．言語モデル L_f の未観測確率集合 $P_f(*|w_{i-N+1}^{i-1})$ に占める未観測 N-gram の確率の総和の大きさ β_f は，モデル L_g の当該 N-gram の確率の総和とモデル L_g の未観測確率集合の比で，以下の式を用いて推定できる．

$$\beta_f(w_{i-N+1}^{i-1}) = \frac{\sum_{C_g(w_{i-N+1}^{i})>0} P_g(w^i|w_{i-N+1}^{i-1})}{\sum_{C_g(w_{i-N+1}^{i})>0} P_g(w^i|w_{i-N+1}^{i-1}) + P_g(*|w_{i-N+1}^{i-1})} \tag{3.9}$$

ただし，この時，$C_f(w_{i-n+1}^{i}) = 0$ である．

パラメータ γ の推定法を図 3.10 に示す．言語モデル L_f のすべての未観測 N-gram の確率の総和に占めるある未観測 N-gram w_{i-N+1}^{i} の出現確率の割合は，モデル L_g における当該 N-gram の総和に占める w_{i-N+1}^{i} の割合に等しいと仮定する．すなわち，

$$\gamma_f(w^i|w_{i-N+1}^{i-1}) = \frac{P_g(w^i|w_{i-N+1}^{i-1})}{\sum_{C_f(w_{i-N+1}^{i})=0} P_g(w^i|w_{i-N+1}^{i-1})} \tag{3.10}$$

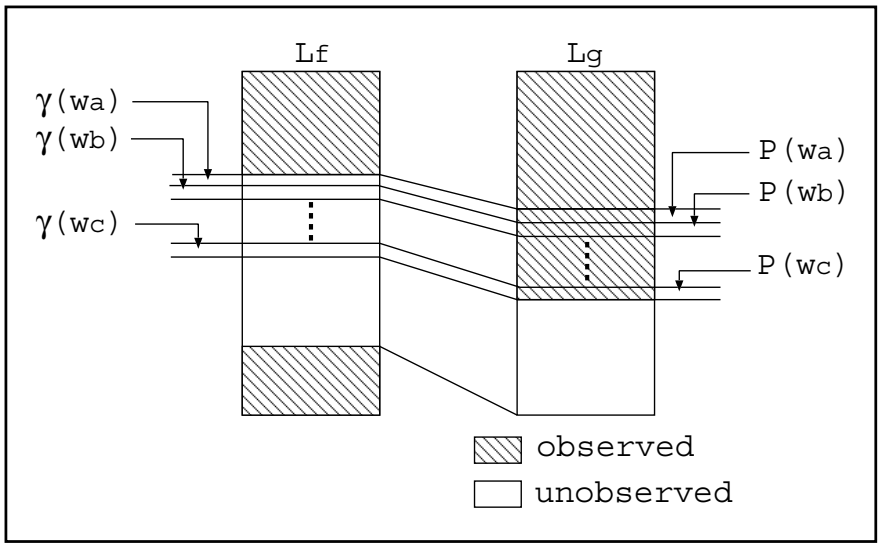


図 3.10 γ の推定

となる.

以上をまとめると，ある N-gram w_{i-N+1}^{i} の融合確率の計算の手順は以下のようになる.

1. モデル L_f で観測されず，モデル L_g でのみ観測される N-gram をすべて調べ，推定パラメータ $\beta_f(w_{i-N+1}^{i-1})$ を求める.

2. 同様に，推定パラメータ $\beta_g(w_{i-N+1}^{i-1})$ を求める。

3. モデル L_f における確率 $P_f(w^i|w_{i-N+1}^{i-1})$ を求める．もし N-gram が L_f にとって未知ならば，推定パラメータ $\gamma_f(w^i|w_{i-N+1}^{i-1})$ を求め，(3.8) 式より推定確率 $\hat{P}_f(w^i|w_{i-N+1}^{i-1})$ を求める.

4. モデル L_g における確率 $P_g(w^i|w_{i-N+1}^{i-1})$ を同様に求める．もし N-gram が L_g において未知ならば，推定パラメータ $\gamma_g(w^i|w_{i-N+1}^{i-1})$ を求め，(3.8) 式より推定確率 $\hat{P}_g(w^i|w_{i-N+1}^{i-1})$ を求める.

5. 両モデルにおける確率あるいは推定確率を用いて，融合後の確率 $P(w^i|w_{i-N+1}^{i-1})$ を (3.6) 式及び (3.5) 式より求める.

6. 手順 (3) から手順 (5) までをコンテキストに属するすべての観測 N-gram について繰り返す.

いずれのモデルでも未観測な N-gram については，融合後のモデルも通常のバックオフを用いて出現確率値を計算することになる．ただし，1-gram の場合は，融合後のすべての 1-gram 確率値の総和をとり，これを 1 から引いて言語モデル中の未知語クラスの確率値に割当てられる.

3.3.3 N-gram モデル融合ツール

以上のアルゴリズムを実装した言語モデル融合ツールを「情報処理学会音声言語情報処理研究会連続音声認識コンソーシアム（CSRC）」[43] のソフトウェアとして配布した．ツールの特徴としては相補的バックオフアルゴリズムにより高精度な融合が行えることや既存のツールのように学習元コーパスや頻度ファイルを用意する必要がないことが挙げられる．語彙は元の二つのモデルで共通である必要はなく，任意の ARPA 標準フォーマットの N-gram モデルを簡単に融合することができる．本研究でも同ツールを用いて N-gram モデル間融合を行い，言語モデルのタスク操作を実現した.

3.4. ネットワーク記述文法の N-gram 言語モデルへの適用

続けて，認識対象タスクに対してさらに高精度な言語モデルを得るために，ネットワーク記述文法の単語間制約をタスク操作を行った融合言語モデルに適用して，言語モデルの構築を完了する.

表 3.4 は記述文法と統計的言語モデルの特徴をまとめたものである．本研究では，多様な発話の表現を受理するために音声認識に統計的言語モデルを用いるが，そのままでは文レベルでの細かい制約を決定的に定義することはできない．例えば，音声インタフェースで頻出する「はい」「いいえ」のみや「今は何時ですか?」などの常套句を認識させたい場合は，文法を用いることで高い精度が得られる.

表 3.4 文法と統計的言語モデルの特徴比較

ネットワーク記述文法	人手で記述 ○ 想定内発話に関しては高精度で認識可能 × 想定外（タスク外）発声に対しては認識に失敗 × 表現パターンの記述の網羅は困難
統計的言語モデル	コーパスから統計的に学習 ○ タスク外の発声も柔軟に受理可能 × 学習には大量のコーパスが必要 × 出力単語列を決定的に定義できない

よって，文法と統計的言語モデルの互いの利点を残しつつ補完する併用手法が求められる．今回は，文法の単語間制約を統計的言語モデルに適用することでこれを試みる．

適用は融合言語モデル中の N-gram エントリに含まれる二単語対が用意した文法により受理可能な場合，そのエントリが持つ確率値を強制的に上げて単語間制約を強化することで行う．具体的には，2-gram エントリに関しては，その単語対が文法で受理可能なもの，3-gram エントリは後ろの単語対が受理可能なものに対して，それらの対数確率値を 0.55 倍した．図 3.11 の例では，3-gram エントリ「の　図書館　は」の後ろ単語対「図書館　は」が文法により受理できるので，3-gram エントリの持つ対数確率値を 0.55 倍している．1-gram（単語）に関しては，文法にその単語が定義されているものを対象とし，その出現確率の値を上げる．ここで使用した倍率（0.55）は，鶴身ら [44] による先行研究の実験結果を参考に設定した．

適用に用いる文法は BNF（Backus Normal Form）風に記述された Julian[45] フォーマット文法である．

また，用意した文法では定義されているが，適用先の言語モデルには含まれない単語に関して，単語辞書中の未知語クラスのエントリに対して，その単語の出力表記と読み（音素記号）を与えることで文法適用言語モデルに追加した．単語

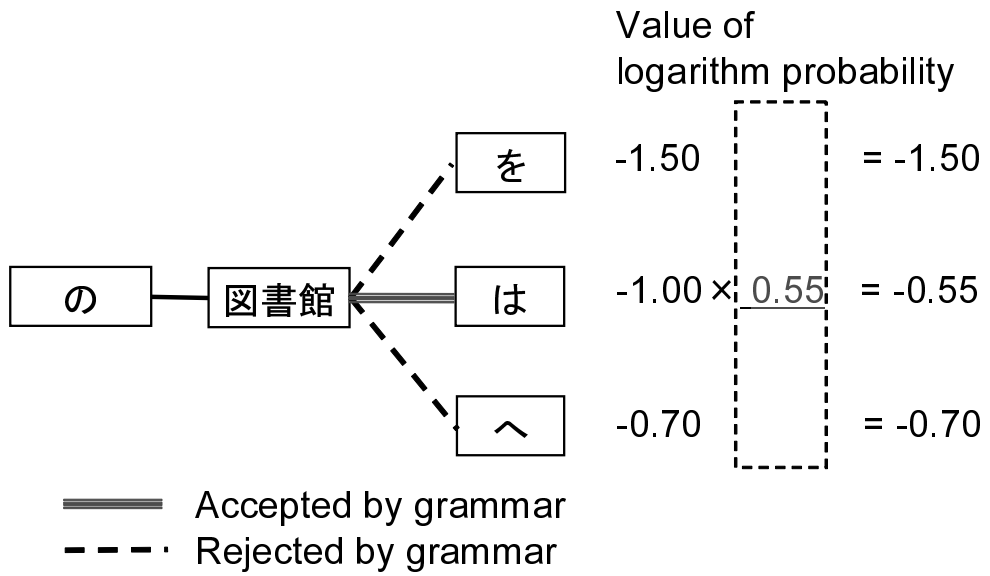


図 3.11 文法適用における単語間制約強化

```
<UNK>   [くろんど池]    k u r o N d o i k e
<UNK>   [竹林寺]        ch i k u r i N j i
<UNK>   [長弓寺]        ch o: ky u: j i
<UNK>   [ＮＡＩＳＴ]    n a i s u t o
```

図 3.12 文法適用によって単語が追加された単語辞書例

追加された単語辞書ファイルの例を図 3.12 に示す．図中の “<UNK>” は言語モデル中で未知語クラスを示す特殊エントリで，ディスカウントで生じる未知語のための確率値が N-gram モデル学習時に割当てられている．この追加によって，学習元コーパスに現れない特殊な固有名詞などの未知語も認識可能になる．

3.5. 本章のまとめ

タスク適応 N-gram 言語モデルを半自動に構築する手法を開発した．その手順は，(1) テキストコーパスの自動作成とトピック依存 N-gram モデルの構築，(2) モデル間融合によるタスク操作，(3) ネットワーク記述文法の適用による N-gram モデル高精度化，の各技術で構成される．

本章では，まず，Web 検索を中心とする Web リソースを利用してトピック依存 N-gram モデルの作成を自動化する手法を提案し，評価した．評価の結果，労力をかけて人手で作成したモデルと同程度の精度を得るモデルを作成できることを確認した．

続いて，相補的バックオフアルゴリズムによる高精度な N-gram モデル間融合について述べた．本研究では同アルゴリズムを実装したモデル融合ツールを用いて N-gram モデルのタスク操作を行っている．

最後に，ネットワーク記述文法の N-gram モデルへの適用手法を検討した．これは文法に含まれる単語対制約に基づいて，モデル内の N-gram 確率を強化することで実現している．この過程は N-gram モデルの高精度化と未知語対策に寄与するものである．

これら一連の技術を組合わせたタスク適応 N-gram 言語モデル構築法は，音声インタフェースを実装するうえでの開発コストの削減をもたらすものである．本手法は，4 章の受付案内ロボット ASKA や 5 章の音声情報案内システム「たけまるくん」の開発において言語モデル構築に実際に利用している．後述の 4 章において，本章で述べたモデル間融合と文法適用 N-gram モデルの有用性を確認する．5 章では，大語彙連続音声認識の性能評価を通じて，本章の手順で作成した言語モデルの性能を検証する．

第4章　実環境研究基盤としての ロボット音声インタフェース

4.1.　はじめに

音声インタフェースのアプリケーションとして，まず，人との対話機能を持つ人型の受付案内ロボットを具体例に考える．

従来のロボット研究においては，機構，制御をはじめとした主に機械系の要素技術の開発が注目されてきた．しかし，ロボットの人間社会への進出が近づくにつれ，インタフェース寄りの研究に注目が集まっている．ロボットのインタフェースに関連する研究分野は幅が広い．音声処理は当然として，ロボティクス，画像処理，センシング工学，自然言語処理，知識処理，情報通信，コンピュータグラフィックスなど，様々な情報科学の分野の要素技術が必要となる．このとき，ロボット開発は様々な要素技術を集めて実装するというシステム統合化中心の作業になる．ただし，既存の研究成果を寄せ集めて，ロボットに組み込んだとしても，実用化はそう単純ではない．各要素技術内でのシステムへの適応や改良，統合に向けた要素技術間の調整が必要である．その過程には，ロボットを研究プラットホームとして見据えた複合的な分野の見地に立つ研究・開発が求められる．

これらの要素技術は，筆者が所属する奈良先端科学技術大学院大学情報科学研究科の複数の研究室において，いずれも従来から研究されているテーマである．このため，ロボットは本研究科で開発する題材として適していると考えられる．受付案内ロボットは，このような背景から本研究科の「共通の研究プラットホーム」として開発が始められた．ベースとなる一台のロボットに「受付案内」という共通の目標となるタスクを設定し，研究室の枠を越えて様々な研究成果を統合する．そして，実環境でサービスを提供可能なロボットシステムを構築すること

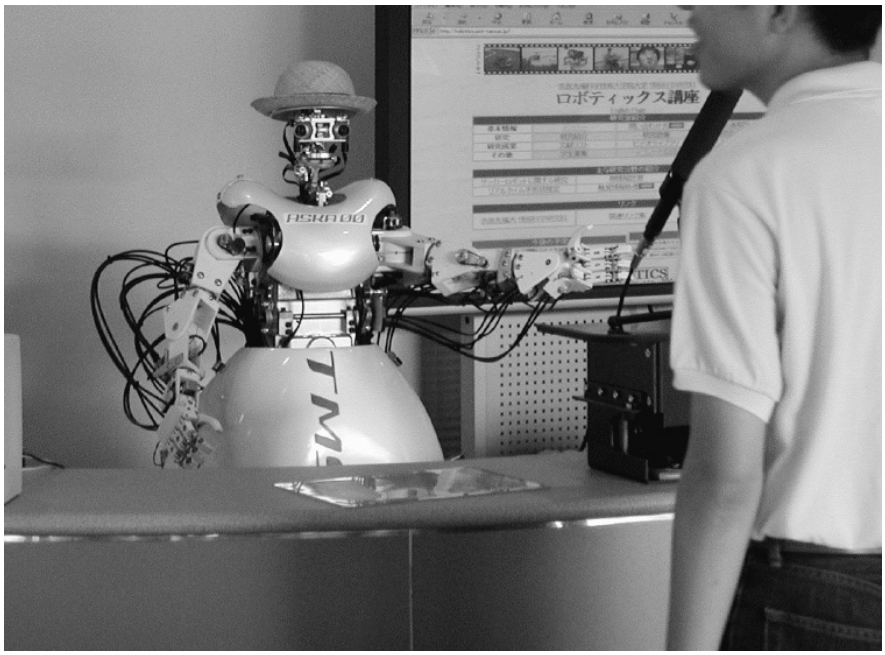


図 4.1 受付案内ロボット ASKA（アスカ）

を目指す.

開発は現在も継続中ではあるが，本章では，音声インタフェース部を中心に ASKA のシステムについて詳説したのち，本学の複数の研究室で構成された ASKA プロジェクトのこれまでの活動成果を総括する．また，ASKA の想定タスク下の実験を行い，統計的言語モデルとネットワーク記述文法の音声認識性能を比較する.

4.2. 受付案内ロボット ASKA

受付案内ロボット ASKA（アスカ）は，筆者も参加する ASKA プロジェクトによって開発された．同プロジェクトは，本研究科のロボティクス講座や音情報処理学講座などの教員，学生から構成される．ASKA は，本研究科の研究科棟一階フロアの入り口に設置される．図 4.1 にその外観を示す．想定されるタスクは来客の案内であり，以下のような質問に応対できるように設計されている.

- 教官及び研究室の場所と内線番号

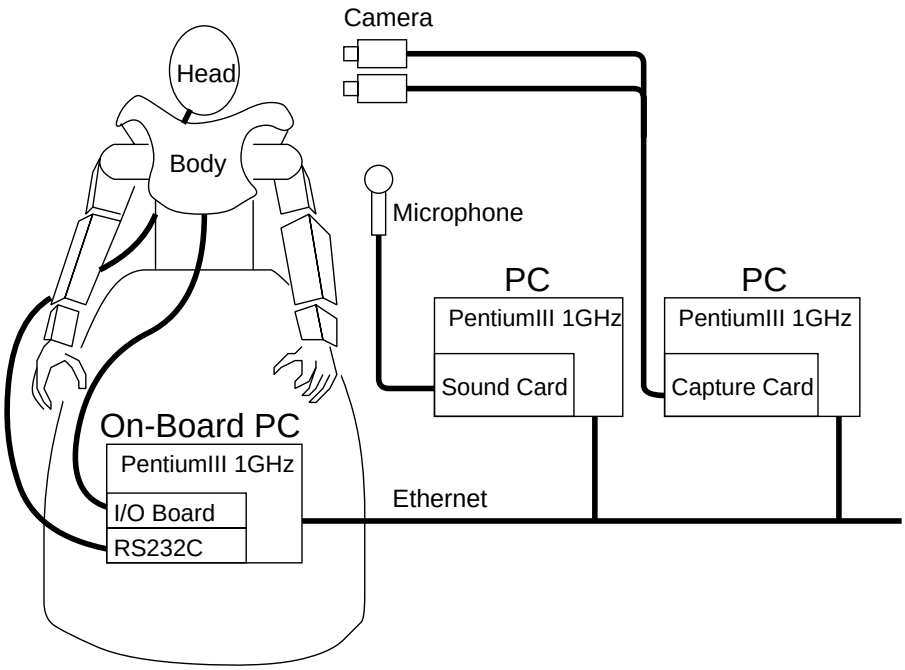


図 4.2 ASKA のハードウェア構成

- 学内及び周辺の施設
- ASKA 自身に関する事柄
- その他，いくつかの挨拶

4.3. ハードウェアとソフトウェア構成

図 4.2 は ASKA のハードウェア構成図である．ベースとなっているのは，テムザック社[1]の人間型遠隔操作ロボット TmsukIV である．TmsukIV は，専用コックピットからの PHS 経由での遠隔操縦用に特化されており，そのままでは自律ロボットとしての利用はできない．そこで，内部に搭載するコンピュータを自律動作に必要な性能を持つ Linux ベースの組み込み PC に載せ替えた．内部 PC 同士

[1] http://www.tmsuk.co.jp/

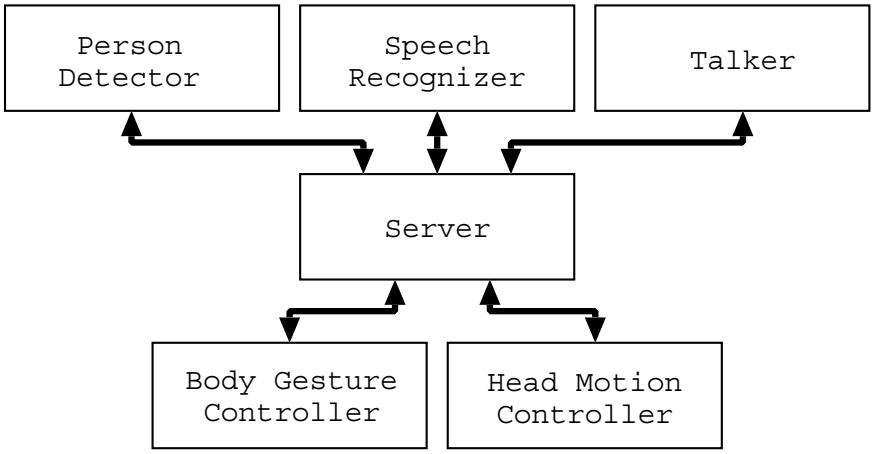


図 4.3 ASKA のソフトウェア構成

と外部には通常のイーサネットの LAN 経由で接続している．

頭部には，人間と対面したり，画像による環境認識を行うために，カメラを組み込んだヘッドを製作した．このハードウェアは，通信総合研究所によって開発された Infanoid Robot[46] とほぼ同一のものである．首はパン・チルトだけでなく，前後に傾けることができ，首振り動作が可能である．目（眼球）も首とは独立にパン・チルトが可能であり，寄り目をすることもできる．唇は，開閉と上げ下げの自由度を持っており，発話や感情の表現に利用可能である．目に埋め込まれた計 4 個（各目に広角と望遠）の小型 CCD カメラは，ステレオ画像処理 [47] に利用できる．ASKA は，そのステレオカメラを応用したユーザとのアイコンタクト機能を持ち，発話者の自動検知をすることができる [48]．

ソフトウェアの構成としては，図 4.3 に示すモジュールが組み込まれており，サーバプロセスを介して TCP/IP によるソケット通信により相互に情報をやりとりしている．モジュールには，音声認識理解（Speech Recognizer），音声合成（Talker），胴体ジェスチャ（Body Gesture Controller），頭部ジェスチャ（Head Motion Controller），ユーザ検知（Person Detector）がある．このうちの音声インタフェース関連のモジュール（音声認識理解，音声合成）については音情報処理学講座が，それ以外のモジュール（胴体ジェスチャ，頭部ジェスチャ，カメラによるユーザ検知）はロボティクス講座が開発を担当した．

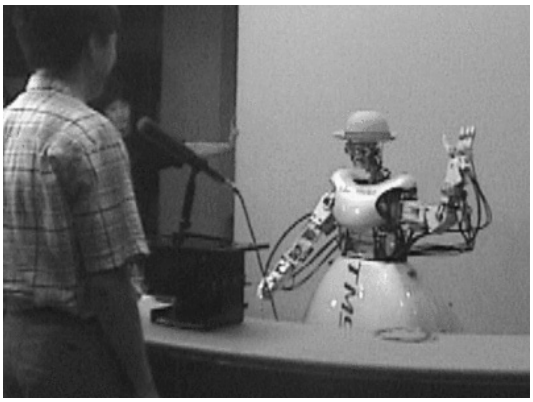
モジュール間の強調動作は，定められたプロトコルを用いてサーバプロセスとメッセージ交換をすることで実現している．すべてのモジュールの ON/OFF 動作状況及びモジュール上の処理で得られた戻値をサーバが保持しており，各モジュールはそれを常に参照しながら動作する．モジュールが互いに独立に動作するため，複雑な同期処理には向かないが，システム全体の開発が容易になる利点を持つ．基本的にモジュール内に閉じた開発ができるため，新たな技術の組み込みを手軽に行えるのもメリットである．

本システムで用いる計算機の OS はすべて Linux である．サーバとの通信を担う C 及び Perl のライブラリを用意しており，モジュールの記述は簡単化されている．計算機の数，モジュールの数は任意に増減が可能になっている．

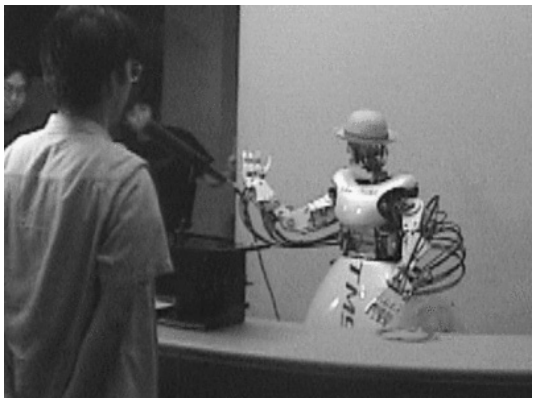
4.4. 対話機能の概要

対話機能しては，一問一答形式により，合成音声及び手と頭のジェスチャを用いてユーザの発話に応答する機能を持つ．ASKA の対話例を図 4.4 に示す．対話処理の概略は以下のようになる．

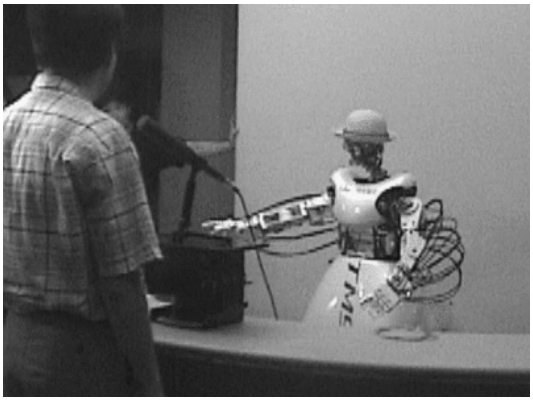
1. 目の中の CCD カメラを用いたステレオ画像処理によって，ASKA の前に立つ発話者を検知する．
2. 検知の後，音声認識理解部が音声の入力を開始する．同時に顔をユーザに向け，質問の受け付けが可能な状態であることを示す．
3. ユーザが ASKA に質問をする（音声入力）．入力には ASKA の前の据え置き型のマイクロホンを用いる．
4. 音声認識理解部が入力音声に対する応答文を作成，結果をサーバに送信する．
5. 音声合成部は，応答文から TTS（Text To Speech）プログラムにより合成音声を作成，発話待ちの状態で待機する．
6. 胴体と頭のジェスチャ部は，応答文等の必要なパラメータがサーバに蓄えられたことを検出，ジェスチャの定義パターンに基づいて動作を開始する．



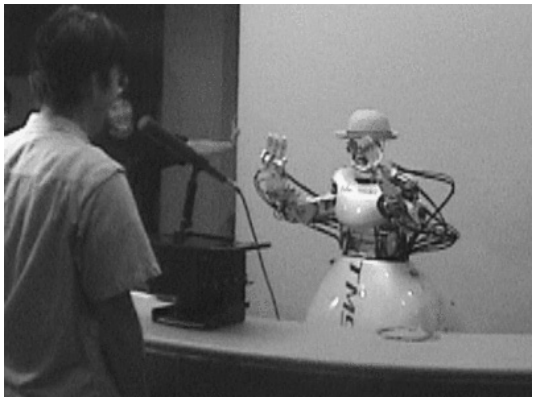
（人）こんにちは.
（ASKA）こんにちは.



（人）公衆電話はどこですか？
（ASKA）公衆電話は私の後ろにあります.



（人）内線電話をかけたいんですが.
（ASKA）内線電話はこちらです.



（人）写真を撮ってもいいですか？
（ASKA）ヤメテクダサイ！

図 4.4 ASKA と人の対話例

7. 音声合成部は，ジェスチャと同時に発話を開始する.

8. すべてが終了すると，ユーザからの発話待ち状態に戻る.

ジェスチャは，胴体と頭部の動作の組合わせで作成する．胴体には，前後左右をはじめとした様々な方向を指し示す動作約 30 種類が登録されており，これらの動作から応答の出力内容に合うものが選択され，音声出力に同期して実行され

る．応答の内容と動作の対応は，あらかじめ定義され固定である．頭部は，通常の待機時，ステレオ画像処理で人を発見しやすいように首を左右に振っており，人を検知し質問を聞くときにはユーザに顔を向ける．応答のときには首振りジェスチャをしながら合成音声の出力時間長に合わせた口の開閉を行い発話行為を表現している．

4.5. キーワードマッチを用いた応答生成

ASKA の音声インタフェースは，前述の音声認識理解部と音声合成部モジュールで構成される．音声認識理解部では，最初に入力音声を音声認識処理する．音声認識エンジンには Julius を用いる．そして，あらかじめ用意した応答文の候補の中から認識結果を基に適切なものを選択する．応答文の選択では，認識結果とキーワードの一致回数の数え上げを基準とする．以下では，具体例をまじえながら詳細を述べる．

ASKA では，来客の質問に対する応答文の候補をあらかじめ用意している．この応答候補は，ASKA に答えてほしい事項に関する学内アンケートを実施し，その結果から必要性が高いと思われる質問を選びだし，その質問に応じた応答を用意するといった順で作成した．図 4.5 に登録されている応答文の例を挙げる．文頭の三桁の数字は，応答文ごとに付けられたインデックス番号であり，他のモジュールとの通信では，このインデックス番号を送受信することでメッセージ交換を行う．

応答文の候補には，定型なものと他のデータベースからデータを検索してスロットに挿入できるもの（スロット型）の 2 種類がある．挨拶や場所の案内には，定型のものを用いる．教職員の居室や内線番号案内などには，名前，電話番号などのパラメータを別に記憶しておき，そこからデータを検索してスロットに挿入することで応答を生成する．この結果，応答文の候補数を少数にまとめることができる．図 4.5 の例の中で，以下の文章はスロット型応答候補の例であり，<is-staff:3> と <is-staff:5> がデータを挿入するスロットを示す．

<is-staff:3>先生の部屋は、<is-staff:5>です。

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

図 4.5 応答候補の例

073 李 晃伸 リ アキノブ B613 B 6 5282 音情報処理学 鹿野 オトジョウホウショリガクコウザ シカノケン

図 4.6 is-staff ファイルの例

<is-staff:3> は，図 4.6 に示すようなファイル名 is-staff のデータベースから，登録されているパラメータが音声認識結果と一致する回数の多いエントリを検索し，そのエントリの三番目のパラメータ（この例では名字の読み）をスロットに挿入する処理を表す．なお，is-staff 中の先頭のインデックス番号は 0 番目として数える．

応答文は，容易に追加削除が可能である．基本的な応答候補として登録されているものは 61 個であるが，必要に応じて適時追加している．

応答の選択に用いるキーワードのリストは，図 4.7 のように応答候補ごとにキーワードを定義して作成する．行頭の番号は対応する応答候補のインデックス番号である．

図 4.7 の例中で，インデックス番号の後ろの “k” に続く文字列が登録するキーワードであり，形態素単位に記述されている．応答文は，音声認識結果の形態素

```
100 p おはよう
302 k 先生
302 k 教授
302 k 助教授
302 k 助手
302 k 部屋
302 k 番号
```

図 4.7 キーワードリストの例

単語列とキーワードが一致した数を数え上げて，最も一致数の多い応答候補を選択して生成される．音声認識の結果出力には N-best を用いる．キーワードリスト中で “p” が指定されている文字列は，パターンマッチワードとして登録する文字列である．認識結果の一部がこの文字列と一致する場合はキーワードより前にカウントされ，キーワードの数え上げは行わない．キーワードリストはシステムを止めることなく随時追加することができる．

4.6. 性能評価

4.6.1 テストセット

本システムの構築に必要なデータ作成と評価のため，質問文の収集を行った．

評価に用いたテキストは，学内の学生（男性 51 名，女性 7 名）のアンケートから作成した ASKA に対する質問文，各 10 文の計 580 文（総単語数 5,591 個）である．それらテキストをヘッドセットと指向性マイクを用いて話し言葉調に読み上げた音声（クリーン環境）を評価用音声とする．

また，各々のテキストを (a) 学内及び周辺施設，(b) 情報棟内の設備，(c) 教員の居室と内線番号，(d) 研究分野からの講座検索，(e) 電車とバスの発車時刻，(f) 天気予報，(g) 今日のニュース，の問合わせに分類した（タスク内）．これら

は ASKA の想定するタスクに含まれるが，応答生成が実装が完了しているのは主に (a) 〜 (c) に関してである．また，ASKA のタスク想定外の 219 文をタスク (z) とした（タスク外）.

4.6.2 言語モデルの作成

言語モデルの作成には以下のテキストを用いた．このうちアンケートで集めた質問文は，前述の評価用に収集したアンケートとは内容が一致しないように指示したうえで別途に集めたものである.

(1) 学内の学生に対して行ったアンケート結果である ASKA への質問文テキスト（769 文，総単語 10k 個，異なり単語 1.1k 個）

(2) Web 検索エンジンを用いて収集した本学関連の Web ページテキスト（22,078 文，総単語 600k 個，異なり単語 26.8k 個）

(3) 過去約 2 年間の学内の学生連絡用メーリングリストに流れたメールテキスト（8,183 文，総単語 253k 個，異なり単語 9.5k 個）

作成した言語モデルは，以下の 4 種類になる．これらの言語モデルは，固有名詞（教官名，講座名，場所など）や専門用語などのキーワードとして利用することの多い特有な単語を含み，本学の受付案内のためのモデルとなっている.

- QA … テキスト (1) のみから学習.
- Web+Mail … テキスト (2) と (3) をコーパス結合した状態から学習．結合後の単語数は 20k に制限.
- QA+Web+Mail … テキスト (1) と (2) と (3) をコーパス結合して学習．結合後の単語数は 20k に制限.
- QA+Web+Mail (merge) … QA モデルと Web+Mail モデルをモデル間融合し（3.3.1 節参照），単語数を制限せずに作成．融合重みはパープレキシティが最小となるように 0.6（QA の割合）を用いた.

表 4.1 言語モデルの性能結果（3-gram）

言語モデル	単語数	PP	OOV [%]
QA	1,107	6.9	4.8
Web+Mail	20,000	49.8	2.1
QA+Web+Mail	20,000	16.6	1.2
QA+Web+Mail (merge)	20,143	10.1	1.1

各言語モデルによるテストセットパープレキシティ（PP）と未知語率（OOV）を表 4.1 に示す．言語モデル QA は，タスク依存度が高いためパープレキシティは最良である．しかし，単語数が少なく未知語率が高い．Web+Mail モデルは本学関連の単語を多く含み未知語率を下げているが，QA の表現をカバーできておらずパープレキシティが高い．QA+Web+Mail モデルは未知語率，パープレキシティともに下げることに成功している．QA+Web+Mail (merge) モデルではパープレキシティの値をさらに改善することができた．

4.6.3 音声認識実験

音声認識で評価する．音響モデルには,「連続音声認識コンソーシアム 2000 年度版ソフトウェア」[43] に含まれる性別非依存の新聞読み上げ音声モデルを用いた．

実験の結果を表 4.2 に示す．認識率は，QA+Web+Mail (merge) モデルが最も高い精度を示した．単純にコーパス結合した QA+Web+Mail モデルよりも 4%程度の向上を得ており，モデル間融合の有用性を示すことができた．マイクの仕様の違いは 1%前後の認識率の差となった．

続く表 4.3 は，ヘッドセットで収録した音声に対してタスク内外の音声認識精度を比較したものである．全体を通してタスク外よりもタスク内が高い値を得ており，作成したモデルのタスク依存性を確認した．

表 4.2 言語モデルの性能比較（音声認識率）

言語モデル	指向性マイク		ヘッドセット	
	Corr. [%]	Acc. [%]	Corr. [%]	Acc. [%]
QA	83.7	79.2	84.2	79.6
Web+Mail	72.7	68.7	72.9	68.6
QA+Web+Mail	80.1	77.6	82.0	79.0
QA+Web+Mail (merge)	85.7	83.1	86.3	83.5

表 4.3 音声認識率のタスク内外比較

言語モデル	タスク内（361 文）		タスク外（219 文）	
	Corr. [%]	Acc. [%]	Corr. [%]	Acc. [%]
QA	89.8	86.1	74.7	68.7
Web+Mail	75.1	71.1	69.0	64.3
QA+Web+Mail	86.6	84.2	74.1	70.2
QA+Web+Mail (merge)	90.9	88.6	78.6	74.7

4.6.4 応答性能実験

ASKA の音声インタフェースが発話をどの程度正しく受理できるか調べた．実験条件は音声認識の評価のものと同一である．言語モデルには，QA+Web+Mail (merge) モデルを用いた．

表 4.4 は，評価用音声の認識結果を ASKA の応答生成プログラムに入力して得た応答の調査結果である．同表の () 内には，音声認識を通さずに書き起こし文を入力した場合の結果を示す．

タスク内発話に対して，質問にキーワードが含まれ適切な応答を生成したものが 35.1%，キーワードが存在せず「わかりません」と応答したものが 35.1%であった．未実装部分を除いたタスク (a) から (c) に対しては，61.7%が適切な応答を返せた．また，時間を扱う部分が未実装であるため，発車時刻を問うタスク (e)

表 4.4 応答性能実験の結果

タスク	キーワード無し	キーワードマッチ		
		正解	不十分	誤り
(a)	29 (31)	32 (33)	7 (7)	3 (0)
(b)	12 (12)	52 (59)	4 (2)	8 (3)
(c)	2 (2)	40 (45)	11 (7)	1 (0)
(d)	14 (14)	3 (4)	1 (1)	1 (0)
(e)	12 (12)	0 (0)	47 (49)	2 (0)
(f)	29 (34)	0 (0)	0 (0)	11 (6)
(g)	29 (36)	0 (0)	0 (0)	11 (4)
(z)	146 (159)	11 (14)	19 (20)	43 (26)

は不十分な回答しか得られなかった．書き起こし文と認識結果との差はあまり見られなかった．

このようにタスクに対する実装が適切になされているか否かが応答能力に影響することは明らかである．キーワードマッチに基づく応答生成では，必要なキーワードが登録されてなければ「わかりません」と答えることになり，今回の実験のみならず実際の運用でもその頻度が目立つ．しかし，利用者の意図を全く理解できなかったとしても「わからない」ではなく，なんらかの情報を返した方が有益であろう．次章の音声情報案内システムではこの問題の対処を試みる．

4.7. 言語モデルと文法の性能比較

3.4 節では，音声認識における統計的言語モデルとネットワーク記述文法の特徴の違いをまとめ，その併用として，単語間制約に基づく N-gram 確率強化による言語モデルへの文法適用手法について述べた．本節では，ASKA を実験環境として用いて，言語モデルと文法の音声認識性能の比較と適用手法の効果を確認する．

実験に用意した言語モデルは，4.6 節と同じ以下のテキストをコーパス結合し

た状態から学習した単語 N-gram モデルである．言語モデルに含まれる単語数は 2 万語である．

1. 学生に対して行ったアンケート結果である ASKA への質問文テキスト（769 文，総単語 10k 個，異なり単語 1.1k 個）

2. Web 検索エンジンを用いて収集した奈良先端科学技術大学院大学の関連 Web ページテキスト（22,078 文，総単語 600k 個，異なり単語 26.8k 個）

3. 過去約 2 年間の学内学生連絡用メーリングリストに流れたメールテキスト（8,183 文，総単語 253k 個，異なり単語 9.5k 個）

適用に用いる文法は，形態素単位に人手で記述した単語数 433 の Julian[45] フォーマット文法である．適用の手順は 3.4 節の説明に準ずる．

テストセットには，本学の学生 58 人から ASKA への質問発話を計 384 個を集めて用いた[2]．テストセットに対する言語モデルの未知語率は 1.2%，テストセットパープレキシティは 28.0 であった．また，用意した文法に含まれる語彙のテストセットに対する未知語率は 18.0%である．テストセットの発話内容のうち，文法で受理可能な文は 282 文，受理不可能な文は 102 文であった．

音響モデルには，日本音響学会新聞記事読み上げ音声コーパス（JNAS）[49] から学習した性別依存モデルを使用した．

認識実験の結果を表 4.5 に示す．表中の数字は単語正解率である．テストセットを用意した文法で受理可能なものと不可能なものに分け，各々について結果を示した．

文法のみを用いた場合，文法受理可能な文章に対しては 96.3%の非常に高い認識精度を得ることができるが，受理不可能なものに対しての著しい精度不足が明らかである．言語モデルを用いると 90%には達しないが常に精度を維持することができる．以上は，3.4 節で述べた言語モデルと文法の特徴と一致している．また，実験結果は文法適用言語モデルが最も良い精度を示したことから，統計的言語モデルをもとに文法適用を施すことの有用性を確認することができた．

[2]ここで用いる発話は 4.6 節のものとは別に集めた．

表 4.5 言語モデルと文法の音声認識性能比較（単語正解率 [%]）

	文法受理可能（282 文）	文法受理不可能（102 文）	合計（384 文）
文法のみ	96.3	57.4	83.7
言語モデルのみ	87.7	78.0	84.6
文法適用言語モデル	92.3	81.2	88.7

表 4.6 ASKA のメディア出演リスト（主なもののみ）

タイトル	メディア名	内容
特集 ROBODEX2002	米 CNN TV（2002/4/4）	アメリカ人キャスタと挨拶
コミュニケーション能力を得た ASKA	雑誌ニュートン（2002/6 月号）	ASKA の紹介
奈良ウィーク「新・まほろば宣言」	NHK ニュース関西（2002/9/11）	番組のアシスタント
暮らしにロボット 人生のパートナー「ロボデックス 2003」来月 3 月 横浜で開幕	朝日新聞朝刊（2003/3/25）	ROBODEX2003 の紹介
スームイン!!SUPER「今日のイチオシ ナマやねん!!」	讀賣テレビ （2003/4/24）	本学のロボット研究を紹介

4.8. ASKA 開発プロジェクトの成果

ASKA は，その完成度が高い対話ロボットとして，新聞やテレビなどのマスコミ報道にも多数取り上げられた（表 4.6）．ロボットの総合博覧会である ROBODEX でのデモンストレーションも成功を果たし（図 4.8），社会から大きな関心を集めた．音声インタフェースと画像処理技術などの異なる分野の要素技術を統合し，実環境下での検証を可能にしたのみでなく，それを世間に広くアピールできたという意味で ASKA プロジェクトの成果は大きい．今後も，より発展的に様々な要素技術との統合を成し得ながら，人にとって有用性の高いロボットを開発することが期待される．

しかし，ロボットのハードウェア保守などの理由で，一般ユーザが ASKA を実際に利用できる機会は本学のオープンキャンパスのデモンストレーションなどの

図 4.8 ASKA デモンストレーション風景（ROBODEX2002 にて）

少ない場に限られているのも事実である．ASKA の運用時には説明員が必要など，ASKA は実環境下で手軽に利用できるシステムとは言い難い．補助員が近くにいるとユーザが気軽に ASKA に話しかけることはできず，音声インタフェース研究の観点から見ると，ASKA を用いた日常的な利用実態調査は難しいことが次第に明らかになった．そのため，音声インタフェースのフィールドテストの目的に特化した新しい研究プラットホームの開発が必要となった．その開発では，ASKA で得たシステム構築の経験は大きな利点となり，ASKA 開発プロジェクトの成果の一つと言えよう．

4.9. 本章のまとめ

本章では，情報科学研究科共通の研究プラットホームとして開発された受付案内ロボット ASKA について述べ，性能を評価した．ASKA の開発プロジェクトの成果を統括したうえで今後の展望についても述べた．

また，ASKA の対話タスクにおいて，統計的言語モデルとネットワーク記述文法の音声認識性能を比較し，単語間制約に基づく言語モデルへの文法適用手法の効果を確認した．

今後の ASKA 開発における音声インタフェース分野の課題としては，5 章で述べる音声情報案内システムを通じて得ることになる改良を継続的に ASKA に統合していくことが挙げられる．例えば，後述する用例テキストを使った応答生成は ASKA への実装も完了しており，現在のシステムでは「わかりません」の数は少なくなっている．本論文では扱っていないが，音声認識の前処理としてマイクロホンアレーなどによるハンズフリー音声入力の導入も望まれる．

ASKA を日常的に運用することは困難であり，音声インタフェース研究では新たなシステムが必要となったが，ASKA の開発が終わったわけではない．開発に参加する研究者は，ASKA プロジェクトで得た問題点を各自持ち帰り，研究課題として取り組まなければならない．この受付案内ロボットが，研究用のプラットホームとして今後も研究者に活用されることを望む．

第5章　利用実態調査を目的とした公共型音声インタフェース

5.1.　はじめに

従来の音声インタフェース研究の多くでは，被験者が少人数であったり，利用状況を想定した疑似対話だったりする実験において，音声インタフェースに関する調査や検証がなされてきた．一般向けに実用化されたサービスを用いた調査の報告 [5] もあるが，その観察は短期間や少人数利用者に対してのみにとどまっていることも少なくない．今後多くの分野で利用されるであろう音声インタフェースの応用は多岐にわたり，それら多様性に対してユーザとシステムとのインタラクションの観察は依然として十分ではない．筆者らの研究においても，前章で述べた受付案内ロボット ASKA は，その運用上の問題からフィールドテストには不向きであり，その利用実態の調査は不十分なものになってしまった．

そこで，本研究では，ユーザがいつでも気軽に利用できる音声インタフェースとして，公共スペースでサービスを提供することができる音声情報案内システムを新たに開発した．そして，長期間フィールドテストを通じてシステムの利用実態を調査，同時にインタラクションの観察に必要な発話の大規模収集を行った．不特定多数の人を利用者として想定した本システムにおいて，多くの人が訪れ，実際に触れることのできる公共の場でのフィールドテストの実施は意義が大きい．長期間の調査によって，システムの利用状況やユーザのシステムへの接し方の変化などを観察することもできる．

以下，本章では，開発した音声情報案内システムの構成を，その中核を成す音声認識部分を中心に解説する．次にフィールドテストで収集したユーザ発話の分析結果を報告する．実験では収集データを使った性能評価を通じて，今後の課題

事項を検討する．また，これまでの音声インタフェースではあまり考慮されてこなかった子供に対する音声認識性能の不足について問題提起を行う．

5.2. 音声情報案内システム「たけまるくん」

音声情報案内システム「たけまるくん」は，奈良県生駒市の協力のもと,「生駒市北コミュニティセンター ISTA はばたき」[1]で受付案内のサービスを提供するために開発したシステムである．

「生駒市北コミュニティセンター ISTA はばたき（以下，北コミュニティセンターと略)」は，市民向け多目的コミュニティセンターであり，456 人収容できる大ホール（はばたきホール）や小ホール，セミナー室，和室，図書館などを備える．市民サービスコーナーでは市役所業務の一部も行っている．

本システムは，北コミュニティセンターの館内施設や生駒市の観光情報，周辺情報などの各種案内を行うものである．システムは同センター内に常設され，一般の来訪者がいつでも気軽に利用できる音声インタフェースによる館内案内サービスを提供することを目指している．そのため，安定した継続動作を最優先にシステムは実装されている．システムの長期間フィールドテストを通じて来訪者（ユーザ）の発話音声を大規模に収集し，そのデータベース化をするのも本研究の目的である．

本システムが想定しているタスクは，以下のようなトピックに関するものである．

1. 北コミュニティセンター館内の案内（部屋や施設の場所など）
2. 業務の内容（手続きの方法や開館時間の案内など）
3. 周辺の案内（駅，バス停，郵便局の場所など）
4. 奈良・生駒の観光情報
5. たけまるくん自身に関する情報

[1]奈良県生駒市上町 1543 番地

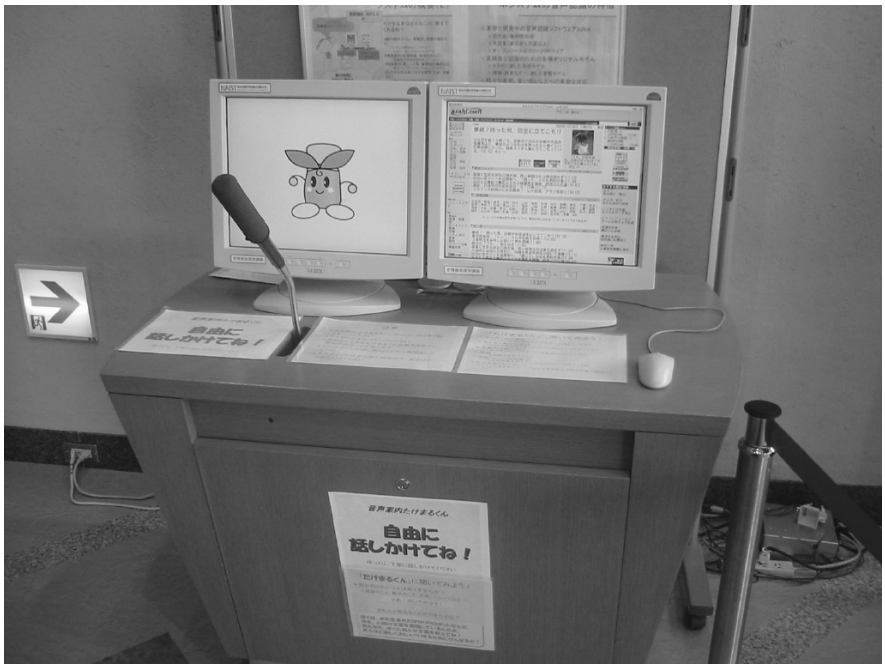


図 5.1 音声情報案内システム「たけまるくん」

6. その他，挨拶など

北コミュニティセンターに設置された本システムの外観を図 5.1 に示す．システム構成は，ロボットのハードウェアの代わりにソフトウェアによるビジュアルエージェントを利用することを除いては ASKA のものとほぼ同一である．

機能としては，一問一答形式の音声インタフェースを持ち，来訪者の質問に対して合成音声とアニメーションを用いた応対ができる．ユーザが机上マイクに向かって発話した内容に二個のディスプレイとスピーカから応答を出力する．左のディスプレイに表示されるのは，生駒市のイメージキャラクタ「たけまる」のビジュアルエージェントである．たけまるエージェントはマクロメディア社[2]の Flash で作成したアニメーションを用いたジェスチャを行う．今のところ，ジェスチャは 38 の動作パターンを持つ．図 5.2 にそのアニメーションの例を示す．たけまるエージェントは音声入力の開始時にうなずき動作を開始することで，ユーザへ発

[2]http://www.macromedia.com/jp/

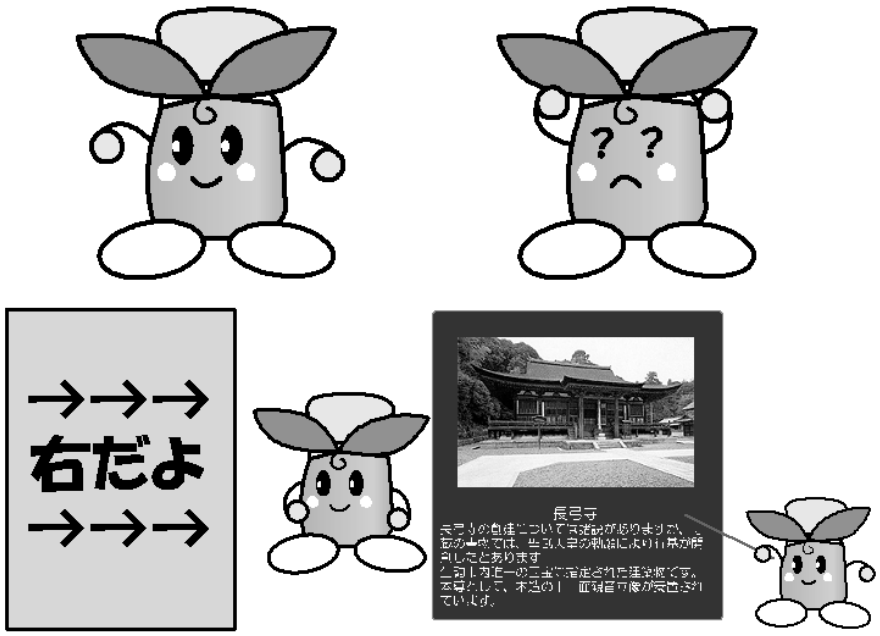


図 5.2 たけまるエージェントの例

話の検知を知らせることもできる．同時に Web を利用した関連情報の提示が可能である．右ディスプレイには，応答内容に関連する Web ページや図 5.3 に示す例のような案内図や時刻表などの関連情報を提示する．

本システムで一問一答形式の音声インタフェースを採用したのは，ユーザが気軽に利用できるシステムを提供するためである．複数の話者が入れ替わりに使用しても対話処理が破綻することなく利用でき，公共の場において多数の人に使ってもらえるシステムを目指して設計した．これはフィールドテストを円滑にすすめるためにも重要である．

対して，複数のインタラクションで構成され，対話状態の遷移が生じる深い対話は現状では実現できない．しかし，本システムのような受付案内タスクでは気軽に利用できる方が重要であると考えた．深い対話が必要となるシチュエーションは今後のフィールドテストで明らかになるだろう．

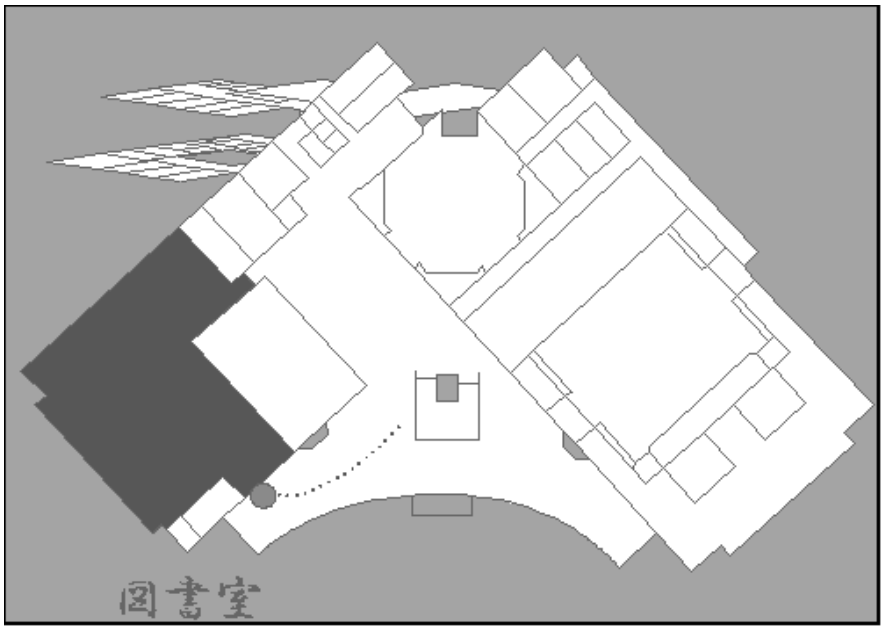


図 5.3 場所案内図（図書館の案内例）

5.3. 音声インタフェースの構成

本システムの音声インタフェースは，これまで述べてきた統計的言語モデルを用いる大語彙連続音声認識を基盤に開発された．統計的言語モデルを使うことで，人手では記述を網羅することが困難な文法ベースの音声認識と比較して，言い回しなどの文章表現の多様性を吸収しながら柔軟に認識できるのは前述の通りである．システムの運用で収集したログを言語モデルの再学習に利用することができ，音声認識の精度向上を効率的に実現できるのも利点である．

音声インタフェースのプログラム構成は音声認識部とその認識結果を用いた応答生成，応答音声を発声する応答合成部から成る．応答合成部はクリエートシステム開発社の「Linux版日本語音声合成ライブラリー」[3]のTTS（Text To Speech）プログラムを使って実装した．以下では，音声認識と応答生成の詳細を述べる．

[3]http://www.createsystem.co.jp/linux.html

5.3.1 音声認識部

大語彙連続音声認識プログラムには Julius を用い，認識対象のタスクに適した言語モデルを得るため，3 章の手順に従ってタスク適応 N-gram 言語モデルを構築した．以下に手順を述べる．

まず，下記の 2 種類のテキストをモデル学習用に収集し，各々について 2-gram 及び逆向き 3-gram モデルを作成した．

1. Web 検索を用いて収集した生駒市関連及び生駒市ホームページ内の Web ページテキスト（3.2.2 節の統計テキストフィルタを用いて整形），1,080,272 文，総単語 31,265k 個，異なり単語 218.7k 個

2. 人手で収集した本システムを想定した質問文テキスト，6,488 文，総単語 56k 個，異なり単語 3.2k 個

各モデルの語彙は，「1. Web ページテキスト」に関しては出現頻度の高いものから上位 4 万語であり，「2. 想定質問文テキスト」では出現したすべての単語（3,231 個）である．

次に生成した各モデルを N-gram モデルの融合ツール（3.3.3 節参照）を用いて融合した．融合重み λ_f，λ_g には，0.5 を用いた（融合割合 1:1）．以降，このモデルを融合言語モデルと呼ぶ．

続けて，融合言語モデルにネットワーク記述文法の単語間制約を適用して N-gram 確率を強化した文法適用言語モデルを作成した（3.4 節参照）．適用に用いた文法の異なり単語数は 441 である．文法では定義されているが融合言語モデルには含まれない 43 単語に関して，単語辞書中の未知語クラスのエントリに対して 43 単語の出力表記と読み（音素記号）を与えることで文法適用言語モデルに追加した．

音響モデルには，日本音響学会新聞記事読み上げ音声コーパス（JNAS）[49] のクリーン音声に 25 dB SNR で電子協騒音データベース [50] の展示会場の雑音を重畳した音声から学習した性別非依存 PTM モデル（以降，展示会場モデル）を使用した [51]．

101 こんにちは。
208 今は、<<hour>>時<<min>>分です。
212 バスの時刻表を表示します。
301 トイレは、左の奥か、はばたきホール、入り口の近くにあります。
305 図書館の入り口は、市民サービスコーナーの横です。

図 5.4 応答候補文の例

5.3.2 応答生成

音声インタフェースの応答生成は，ユーザの一発話ごとの音声認識結果をもとに，あらかじめ用意された応答候補文から一文を選択し，応答として返すものである．今回，システムの設計に際して，ユーザに時間遅れ無しに応答が返せることを重視し，簡潔なプログラム構成になるように心がけた．応答内容に関しては，ユーザの関心を維持することを重視して，なんらかの情報を持った応答を返すことを主眼とした．ユーザがシステムに対する関心を失うとフィールドテストやデータ収集には好ましくないためである．具体的には,「わかりません」や「もう一度お願いします」などのインタラクションを阻害する応答を極力生成しないようにするとともに，質問に関連した話題を広くカバーできるよう応答候補文を作成した．運用開始後もログなどを参考に，必要な応答候補文の追加を行っている．

以下では，応答生成の具体的な手順について説明する．図 5.4 は，前述のようにあらかじめ用意した応答候補文の例である．各文にはインデックス番号が付けられている．応答候補文には，定型文の他に図 5.4 の二行目のようにスロットにパラメータを代入できるスロット型がある．現在，登録されているすべての応答候補文数は 202 であり，そのうちスロット型の応答候補は，時間や日付などの問い合わせに関する三つである．

これら候補からの応答の選択には，音声認識結果とあらかじめデータベースに登録した用例テキストとの形態素マッチングによるスコア計算を用いた．用例テキストは，本システムの過去ログや生駒市役所の業務記録などから作成した質問

トイレ+トイレ+2　は+ワ+65　、+、+79　どこ+ドコ+14　です+デス+74/56/1　か+カ+70　？+？+77#301
食堂+ショクドー+2　は+ワ+65　、+、+79　あり+アリ+47/17/5　ます+マス+74/58/1　か+カ+70　？+？+77#332

図 5.5　形態素単位に分割された用例テキストの例

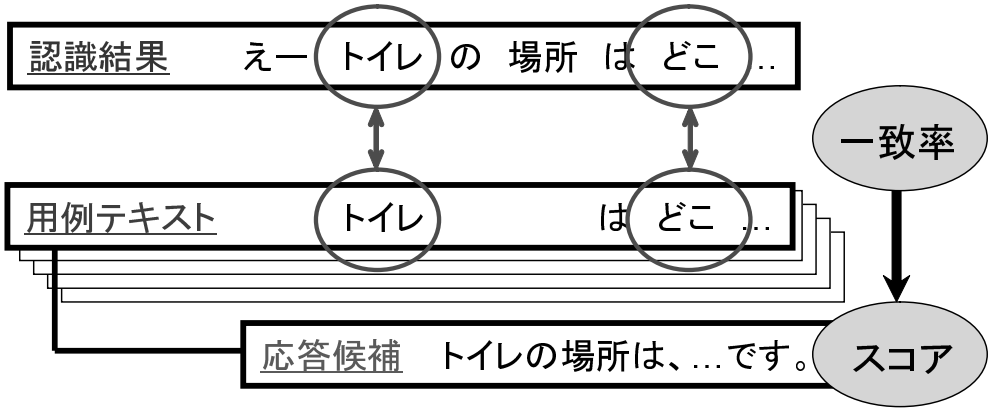


図 5.6　用例ベースのスコア計算

文の形態素列である（図 5.5）．現在，2,309 文の用例テキストが登録されている．すべての用例テキストには，前述の応答候補文の中からその質問の応答として相応しいものへの対応が定義されている．具体的には，図 5.5 中の # の後に書かれた数字がその用例テキストに対応づけられた応答候補文のインデックス番号である．

図 5.6 に用例テキストを用いたスコア計算の概略を示す．入力音声認識結果と用例テキストとの形態素の一致数を求め，用例テキストの単語数で割り一致率を求める．その一致率をスコアとする．このとき，カウントに用いるのは質問の意図理解に重要な名詞，動詞，形容詞などの自立語形態素のみである．すべての用例テキストに対して一致率を算出し，最も高いスコアに対応する応答候補文を応答とする．複数の候補のスコアが同じ場合は，応答はそれら候補の中からランダ

ムに選択される．

本手法では，キーワードリストを必要とする応答生成とは違い，用例テキスト数を増やすことで言い回しや語句の違いなどの発話の多様な表現様式に対処できるようになるのが特徴である．作成した用例テキストは言語モデルの学習にも流用でき，統計的言語モデルを用いた音声認識との相性も良い．

応答生成プログラムの入力である音声認識結果にはN-best出力結果を用いる．このため，正解形態素が1-best結果に出現しない場合でも2-best以降に出力されればスコア計算に含めることができる．実際の運用では，十分な候補数を用いるために音声認識プログラムの100-best出力を応答生成プログラムに入力し，100-bestすべての認識結果に対する全スコアを合計した．

5.4. フィールドテストによるデータ収集と分析

本システムは，北コミュニティセンターのオープン初日である2002年11月6日から運用を開始した．本研究では，ユーザとシステムのインタラクションログとして，認識エンジンが音声入力として切り出した音声区間データをすべて記録している．以下，これまでのデータの収集と分析の結果について報告する．

5.4.1 収集結果

フィールドテストとデータ収集は現在も継続しているが，本研究で用いるデータは，オープン2日後の2002年11月8日から2003年3月31日までの休館日などを除く125日間分のデータである．記録されたデータは，雑音のみや不明瞭な発話も含めて46,754個であり，ファイルの総容量は2493.2メガバイトになる．これは単純計算で約1,362分の長さに相当する．

収集と並行して人手による収集データの整備を行った．この作業では，切り出しによる無音などの非音声区間の除去，発話内容のテキストへの書き起こし，音声のみからの主観による話者の性別及び年齢層のラベル付けを行っている．この整備の結果，雑音が多少は含まれるが，内容を明瞭に聞き取れる比較的クリーン

表 5.1 収集データの年齢層と性別ごとの分類結果

上段: 全収集データにおける分類数
下段: クリーン発話における分類数

年齢層		男性	女性	性別不明	合計
(a)	幼児	115	1940	1044	3099
		76	1421	585	2082
(b)	低学年子供（小学校 3 年生ぐらいまで）	2709	9793	4579	17081
		1920	7961	2843	12724
(c)	高学年子供（中学生ぐらいまで）	1142	1290	673	3105
		934	1154	498	2586
(d)	大人	5867	2778	103	8748
		5520	2496	70	8086
(e)	高齢者	10	12	0	22
		8	12	0	20
(x)	年齢層の判断ができなかったもの	2	6	14691	14699
		0	0	16	16
合計		9845	15819	21090	46754
		8458	13044	4012	25514

性別不明: 声から性別を判断できなかったもの

な発話（以下，クリーン発話）は，25,514 個であった．

収集データの年齢層及び性別の分類結果を表 5.1 に示す．表中上段の数字は，雑音のみや不明瞭な発話も含むすべての収集データを分類したときの数である．下段には，前述のクリーン発話（25,514 個）のみを分類した場合のデータ数を示す．表 5.1 から，本システムでは男女を問わず広い年齢層の利用者の発話を大規模に収集できたことがわかる．

全収集データのクリーン発話の割合は，54.6%であった．年齢層と性別がともに不明な 14,691 個の収集データに関しては，その多くがマイクと手の摩擦音や物

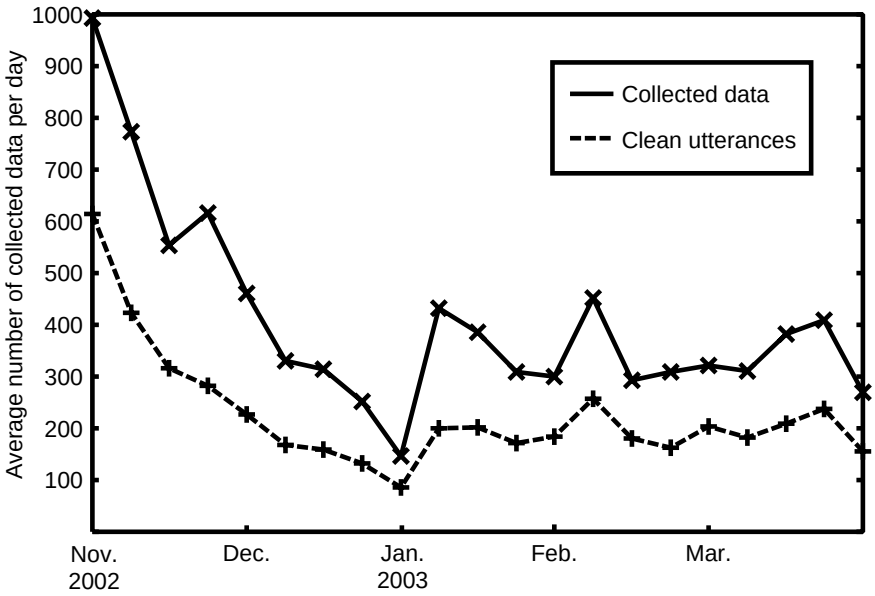


図 5.7 フィールドテストによるデータ収集数（一日平均）

の衝突音，離れた場所での会話音などの雑音のみで発話を含まない入力であった．これら雑音のみの入力でも，現状のシステムは音声区間の切り出しのミスにより反応する．なお，クリーン発話の割合は大人と子供で異なり，大人の 92.4%に対して，子供が 74.7%と低い．子供の発話には，入力レベルのオーバフローやはっきりしない不明瞭な発話などが多く含まれていた．特にオーバフローが発生している発話は，大人は 50 個であったが，子供は 2,683 個と多い．この結果は，大人はシステムとのコミュニケーションにおいて丁寧に発話する傾向があるが，子供は荒い発話をすることが多いことを示唆している．

図 5.7 は，週ごとにデータ収集数の推移を示したものである．各折れ線グラフは，すべての収集データ（Collected data）とクリーン発話（Clean utterances）の一日平均の収集数を示す．このグラフから，オープン当初からは大きく減ったが，2003 年の 1 月以降，システムは定常的に利用されていることがわかる．なお，2003 年 1 月から 3 月末までの一日平均の収集データ数は，335 個であった．

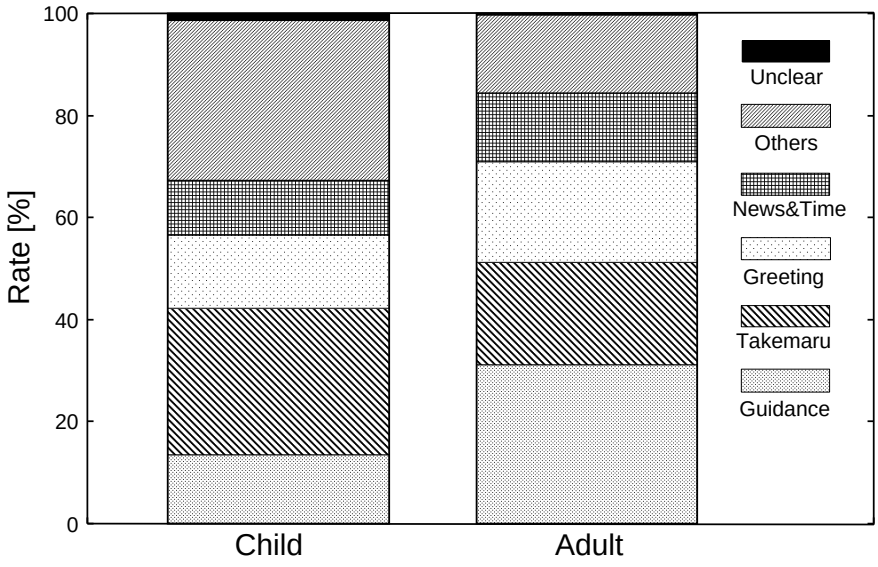


図 5.8 年齢層ごとの発話トピックの割合

5.4.2 発話内容に関する分析

収集データの発話の内容に関する分析を行った．全収集データの中から話者の年齢層を判別できた発話（表 5.1 中の (a) 〜 (e) の全収集データ）に関して，トピックごとの発話数の割合を集計した．トピックの分類は，整備した書き起こしテキストをもとに人の主観によって行った．図 5.8 に，大人と子供における分類結果を示す．表 5.2 は，図 5.8 で用いたトピック名の一覧である．

“Guidance” は，本システムの主目的である案内に関する問合わせである．バスや電車の時刻の問合わせもここに含む．“Takemaru” は，名前や年齢,「好きな食べ物は何？」など，たけまる君自身に関する質問である．より楽しいコミュニケーションを実現するため，たけまるエージェントのキャラクタ設定を行っており，本システムでは答えることができる．“Greeting” は,「こんにちは」「おはよう」などの挨拶である．ユーザがシステムに話かけるときに発話されることが多い．“News&Time” は，主に時事ニュースや天気予報に関する質問であり，新聞の Web ページを表示することでユーザに関連情報を提示する.「今，何時ですか？」などの時間の問合わせも含む．“Others” は，本システムでは想定していな

表 5.2 発話内容分類に用いたトピックの一覧

トピック名	概要
Guidance	館内や周辺，業務，観光などの問合わせ
Takemaru	たけまるくん自身に関する質問
Greeting	挨拶
News&Time	ニュース，天気予報や時間の問合わせ
Others	その他
Unclear	叫び声などの不明瞭な発話

い発話である．人が聞いても意味理解が困難な発話も多い．ただし,「わかりました」や「もう一度言って」などのシステムの応答に対するユーザの反応もここに含む．これらユーザの反応発話は，”Others” の全 14,320 発話中，1,539 発話を占めていた．現在の一問一答の応答生成プログラムでは，これらの反応に対処することができない．これら発話には，応答生成時に一つ前の発話履歴も用いることで対応できると考えており，今後のシステム改良時に導入を予定している．最後に，“Unclear” は，叫び声などの発話の内容が意味的に不明瞭な発話である．オーバフローなどの発生により聞き取りにくいものでも，発話内容を理解可能なデータは，”Unclear” 以外のトピックに分類した．

図 5.8 から，本システムの主目的である”Guidance” の問合わせでは，子供が 13.5%であるのに対して，大人が 31.2%と多いのがわかる．大人の方が本システムを本来の目的で利用していると言える．一方で，”Takemaru” の発話に関しては，子供の方が 8.7%多く，子供はキャラクタ設定に関する興味が強いことがわかる．この結果，大人と子供の間には，発話内容に異った傾向があることがわかった．

5.5. 性能評価

収集した発話に対して音声認識性能と本システムがどの程度正しく応答できるかの応答性能を調べるため，評価実験を行った．

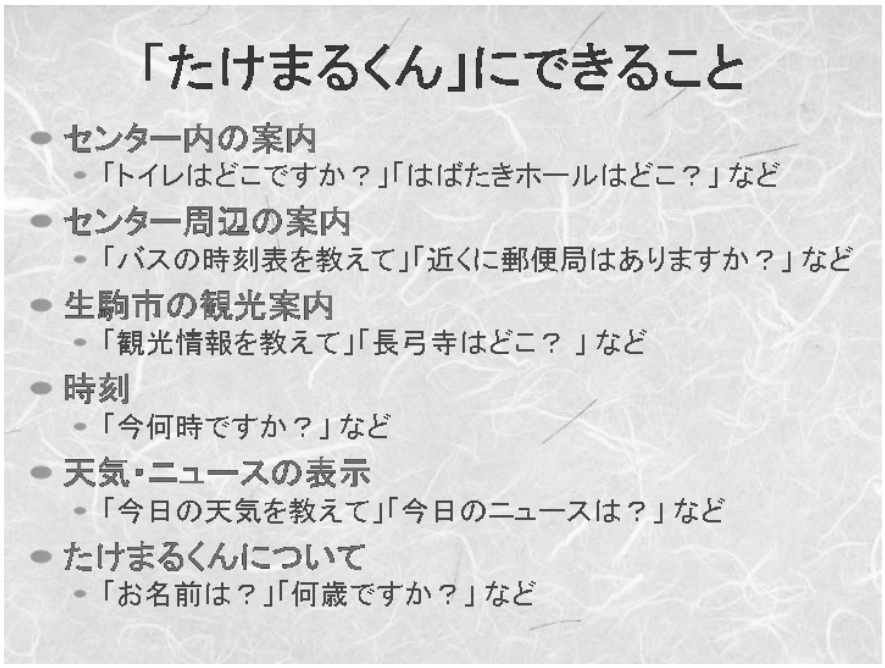


図 5.9 ユーザへのガイド

5.5.1 テストセット

実験のテストセットに使用したのは，5.4.1 節で述べたクリーン発話 25,514 個から抽出した，大人（大人及び高齢者）500 発話と子供（幼児，低学年及び高学年子供）500 発話である．なお，本システムの周辺には，ユーザへの発話を促すため発話例を示した図 5.9 のようなガイドポスターを掲示している．このため，ユーザの発話はガイドに大きく影響されており発話内容に偏りが生じる．掲示する発話例を適時変更して偏らないよう対処しているが，今回収集した発話でも発話内容の偏りは確認した．テストセット作成の際には，この偏りを防ぐため，書き起こしテキストをキーにしてソートを行い，同じ発話内容のデータを間引いて除いた後の発話（大人 2,833 個，子供 5,599 個）から，その発話内容の出現頻度に基づく上位 500 個のデータを選んだ．

作成したテストセットに含まれる発話内容の例を図 5.10 及び図 5.11 に示す．テストセットデータの緒元及び 5.3.1 節の融合言語モデルに対する 3-gram テスト

元気ですか？
高山サイエンスプラザはどこですか？
市役所はどこにありますか？
バスの時刻を教えてください。
図書館の利用時間教えてください。

図 5.10 大人用テストセットの例

あなたは誰？
図書館はどこですか？
何ができんの？
男の子ですか？女の子ですか？
んなとこ聞いてない。
トイレはどこでしゅか？

図 5.11 子供用テストセットの例

セットパープレキシティ，未知語率，文法で定義された 43 単語を追加した文法適用言語モデルに対する未知語率を表 5.3 に示す.

5.5.2 音声認識実験

作成したテストセットに対して，Julius による大語彙連続音声認識実験を行った．言語モデルには，5.3.1 節で述べた文法適用言語モデルを用いた．音響モデルには，実際の運用で使用した展示会場モデルの他に JNAS モデル，CSRC 女性モデル及び CSRC 子供モデルを用いた比較実験を行った．JNAS モデルは，JNAS コーパスのクリーン音声から構築した性別非依存 PTM モデルである．CSRC 女性モデルは，連続音声認識コンソーシアム（CSRC）[43] の 2001 年度版ソフトウェアに含まれる女性音声モデルである．CSRC 子供モデルには，同じく CSRC の

表 5.3 テストセットデータの緒元

	大人	子供
男性発話数	377	99
女性発話数	123	401
総単語数	2767	2497
異なり単語数	369	379
テストセットパープレキシティ (3-gram)	16.8	32.0
ベースラインモデルに対する未知語率 [%]	1.01	1.52
文法適用モデルに対する未知語率 [%]	0.76	1.16

2002 年度版ソフトウェアに含まれる小児モデルを用いた．

実験の結果を表 5.4 に示す．表中の Corr. は単語正解率，Acc. は単語正解精度を示す．JNAS モデル，展示会場モデルともに大人に比べ，子供発話の認識精度は大きく低下しており，十分な精度が得られていないことがわかった．原因の一つとして，JNAS データベースが成人男女の音声で構成されているため，子供声の音響的特徴に合致しなかったことが考えられる．子供音声コーパスを用いて音響モデルを再学習することで認識精度を向上できることが報告されている [52][53]. 今回は，子供モデルを用いることで JNAS モデルに比べ 8.0%の認識精度の向上を得ることができた．しかし，その精度は依然として低い．また，自由発話による子供発話の認識では，大人と比較して単語正解率と単語正解精度の差が大きく，挿入誤りが多いことがわかる．これは，子供はスムーズに発話することができず，その発話には言い直しや不要語が多いことが原因である．よって，音声とモデルの音響的ミスマッチに関する対策の他に，子供声音声認識には，言い直しへの対応などがさらに必要であることがわかる．

雑音に関しては，展示会場モデルを用いることにより，JNAS モデルに比べて認識精度が向上しており，環境雑音への対応が必要であることを確認した．

表 5.4 大語彙連続音声認識実験結果

	大人		子供	
音響モデル	Corr. [%]	Acc. [%]	Corr. [%]	Acc. [%]
展示会場	85.6	79.7	55.2	38.8
JNAS	83.9	79.1	50.1	35.2
CSRC 女性	-	-	52.5	37.9
CSRC 子供	-	-	60.1	43.2

5.5.3 応答性能実験

ユーザの発話に対して，システムがどの程度正しく応答できるかを調べ，本システムの応答性能の指標を明らかにする．実験結果を表 5.5 に示す．表中の数字は，テストセット 500 発話での正解応答文の割合（応答正解率）である．応答正解率は，満足な応答結果が得られた応答文の割合である．応答内容が満足なものかは人の主観により判別した．実験に使用した音響モデルは展示会場モデルと CSRC 子供モデルである．また，音声認識の誤りを除外した応答性能も検証するために，音声認識結果ではなく書き起こしテキストを入力した際の応答正解率も示す．

結果は大人に対しては 73.4%の高い応答正解率を示した．しかし，音声認識結果と同様に子供の正解率が大きく低下している．子供モデルを用いることで，5.8%の正解率の回復を得ることができた．

次に応答生成の入力に用いる音声認識出力の N-best 候補数の応答正解率との関係を調べた．図 5.12 は大人発話（展示会場モデル使用），図 5.13 は子供発話（CSRC 子供モデル使用）での結果である．横軸は使用した N-best 候補数，縦軸は応答正解率を表す．1-best や 5-best を使用するより，50 や 100 などのある程度大きい候補数を用いた方が精度が高くなることを確認した．しかし，大人発話に対して 100-best で若干ではあるが精度低下が見られる．これは低位の候補に出現した誤り単語の影響であり，パラメータの調整等により防ぐことができる．

最後に，応答生成の誤りの原因について分析を行う．表 5.5 の応答正解率におい

表 5.5 応答正解率 [%]

音響モデル	大人	子供
展示会場	73.4	37.4
CSRC 子供	-	43.2
書き起こしテキスト	78.8	62.0

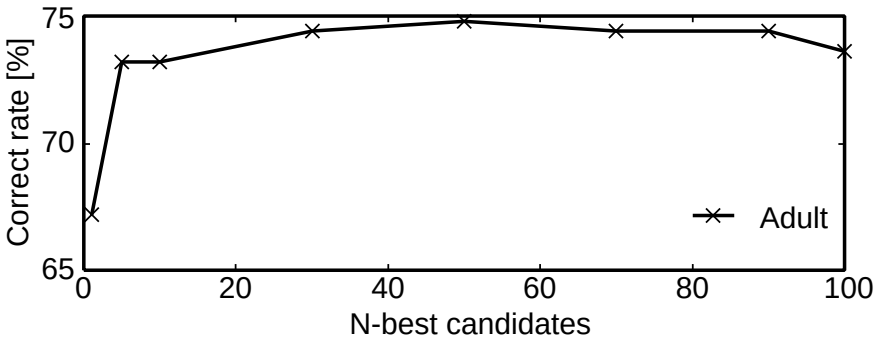


図 5.12 N-best 候補数に対する応答正解率の推移 （大人）

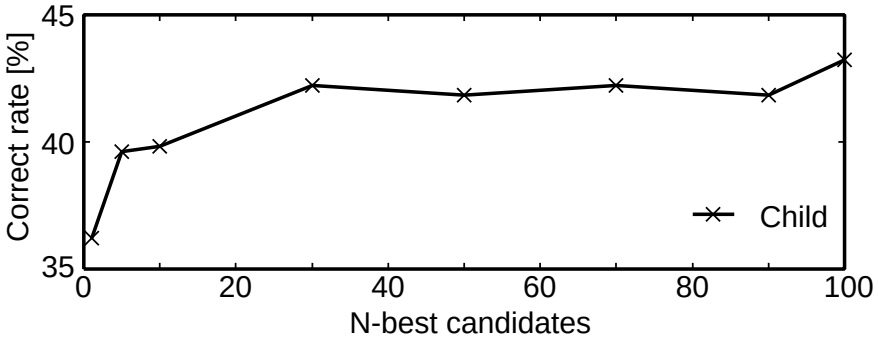


図 5.13 N-best 候補数に対する応答正解率の推移 （子供）

て，書き起こしテキストと実際の音声認識結果を用いた時の差は，大人で 5.4%，子供で 18.8%である．表 5.4 のように子供発話認識の精度は大人のものと比べて著しく低いことから，この大人と子供の違いは主に音声認識精度の影響であると

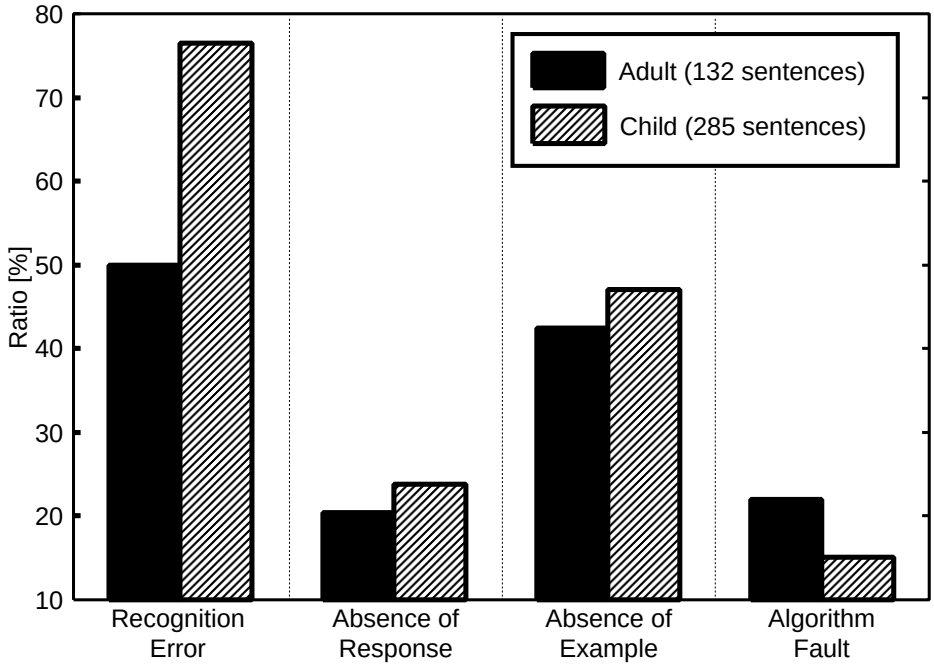


図 5.14 応答生成の誤り原因の分析

考えられる．しかし，大人に関しては音声認識結果と書き起こし使用時の差が小さいことから音声認識以外に応答誤りの原因があることは明らかである．

そこで，先の実験で応答生成が誤ったものに関して，その誤りの原因を調査した．調査の結果を図 5.14 を示す．

今回の調査では原因を大きく以下の四種類に分類した．図中の “Recognition Error” は，音声認識の誤りが原因で応答生成を誤ったものを示す．“Absence of Response” は，発話に対する応答候補が存在しなかったもの，“Absence of Example” は，用例テキストが存在しなかったものの割合である．必要な単語が正しく音声認識結果中に存在して，応答候補や用例テキストも登録されているが，応答を選択するアルゴリズムが原因で誤った応答を出力したものに関しては，“Algorithm Fault” とした．なお，上記の複数が原因と考えられるものは，そのすべてにカウントしている．例えば，応答候補が存在しないものは，対応する用例テキストは登録されていないことがほとんどであり両方にカウントされる．また，後に音声認識誤りが解消されたとしても，新たにアルゴリズムが原因で応答を誤ることも

ありえるが，それらは現状では予測が困難なのでカウントしていない．

結果を見ると，大人と子供ともに音声認識誤りを原因とするものが最も多い．特に子供の場合は顕著であり，音声認識精度の改善は必要であることを確認した．次に用例テキストの不足，応答候補の不足と続くが，これらも子供の方が多い．これは子供の発話にはシステム設計時に想定していなかった内容のものが多いためである．これらに関しては用例テキストや応答候補の登録を追加することで改善が見込まれる．一方で，アルゴリズムに起因する誤りは少ない．今回，テストセット作成時に発話内容で出現頻度の高いものを選択したため，比較的単純な発話が多くなってしまったことも考えられる．例えば，テストセットには含まれていなかった否定的な表現に現状では対処できない．また，テストセットの総単語数（表 5.3）が子供の方が少ないことからもわかるように，子供は比較的発話が単純であり，アルゴリズムに起因する誤りは子供の方が少ない結果となった．

5.6. 収集データを用いた音声認識モデルの再構築

本節では，収集した発話を音声認識モデル（音響モデルと言語モデル）の学習データとして利用することで，フィールドテストの有用性を示す．

まず，音響モデルを作成する．学習に用いたデータは，JNAS 読み上げ音声（40,086 文）と収集発話のうちクリーン発話（20,119 文）である．学習に必要な音素ラベル等が原因で一部の収集発話に関しては学習から除外した．

言語モデルの学習に用いたテキストデータは，5.3.1 節で準備した Web ページテキスト（Web）と想定質問テキスト（QA）に加えて，収集発話の書き起こしテキスト（24,498 文，単語数 129k，異なり単語数 4.0k）である．はじめに Web と QA テキストからベースとなる単語数 40k の言語モデルを作成する．次に書き起こしテキストから言語モデルを作成して，ベース言語モデルにモデル融合ツールを用いて重み 0.8 で融合する．そして，生成された言語モデルに別途人手で作成した異なり単語数 441 のネットワーク記述文法を適用した．

大語彙連続音声認識及び応答性能の評価実験結果を表 5.6 に示す．表中の Corr. は単語正解率，Acc. は単語正解精度，Res. は応答正解率を表す．ベースラインは，

表 5.6 音声認識率と応答正解率（再構築モデルを使用） [%]

	大人			子供		
モデル	Corr.	Acc.	Res.	Corr.	Acc.	Res.
ベースライン	85.6	79.7	73.4	60.1	43.2	43.2
再構築	91.7	89.8	75.4	73.4	66.8	52.0

表 5.4 と表 5.5 で示した値のうち，大人は展示会場，子供は CSRC 子供音響モデルを用いた際のものである．これらはフィールドテスト前に準備した音声認識モデルの中で最も高い精度を得た条件での結果である．

表から単語正解率，単語正解精度ともにモデルを再構築することでベースラインに対する認識精度の改善を確認することができた．また，音声認識精度の向上にともなって応答正解率も改善した．これらは収集データから構築した音声認識モデルが有効に働いていることを示すものであり，結果としてフィールドテストの有用性を確かめることができたと考える．

5.7. 本章のまとめ

本章では，音声インタフェースのフィールドテストを目的に開発した生駒市北コミュニティセンターの音声情報案内システム「たけまるくん」について述べた．本システムは同センター内に常設されており，来館者がいつでも気軽に利用できる一問一答形式の音声インタフェースを使った館内案内サービスを提供している．

約五ヶ月間に渡るフィールドテストで収録したユーザ発話の収集状況を報告した．その結果，25,514 個のクリーン発話を含む幅広い年齢層の自然発話を大規模に収集することができた．2003 年 1 月以降，平均して一日 335 個のデータが収集できており，本システムが定常的に利用されていることを示した．これらの結果を総じて，音声インタフェースの利用実態調査に必要なユーザ発話を大規模に収集することに成功したと言える．また，発話トピックの分類から大人と子供では発話の内容に傾向の違いがあり，主に大人が本システムを実用的に利用している

ことがわかった．

実験では，収集データを用いてシステムの性能指標を評価した．その結果，本システムは大人に対してフィールドテストに必要な基本性能を備えているが，子供に対する改善が必要であることを確認した．

本システムとフィールドテストより得た主たる成果は，音声インタフェース研究における新たなる課題の発見とその検証に必要なデータを収集したことであり，システムの開発は今後も継続される．また，フィールドテストの継続も当面の最優先課題である．5.6 節で述べた音声認識モデルの構築は，収集データの利用事例の一つに過ぎず，その利用価値は多岐にわたる．今後の展開として筆者らが検討を行っている課題事項を以下に挙げる．

次章の 6 章では，主に子供に対する利便性向上を得ることを目的に，大人と子供の発話内容の違いに柔軟に順応することができる音声インタフェースを検討する．また，子供発話の認識性能不足の対策として収集データを用いた音声認識モデルの適応を試みる．

5.4.1 節で述べた摩擦音や衝突音等の雑音の入力に関する対策としては，中村ら [56] による混合正規分布モデルの尤度比較に基づく音声と非音声入力の自動識別手法が，音声認識の前処理段階で雑音を廃棄することを可能にしている．中村の実験では，本研究のたけまるくん収集データを用いて 90%以上の精度の音声・非音声識別を実現しており，システムへの実装も予定されている．

自由発話の認識精度向上には，モデル適応以外に言い直しや不明瞭な発話に関する対策が必要である．これを応答生成時の不明瞭部分をシステムが聞き返し応答をすることで補う予定である．その実装に必要な応答生成の確信度 [55] の導出を検討している．

発話データの収集は継続中であり，追加データを使った詳細な分析も必要である．図 5.15 に 2003 年末までのデータ収集数の月ごとの推移を示した．小中学校が長期休業の八月はオープン当初よりも収集数が多く，本システムが一般市民に受け入れられている様子がわかる．音声インタフェースのフィールドテストが実施されること自体これまで少数であったが，その中でもこれほど長期的なものは皆無であり，今後の収集データの整備とその分析が急がれる．

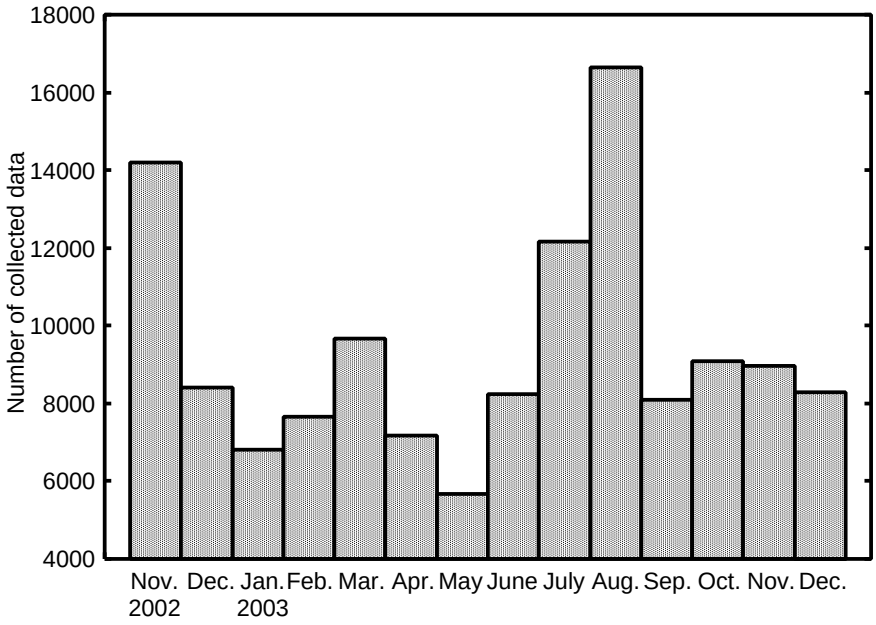


図 5.15 フィールドテストによるデータ収集数（月単位）

ユーザのシステムの扱いは非常に雑であり，ロボットでなくてもハードウェアの故障は頻発する．公共の場で安定したサービスを提供し続けることは，多大なる労力を必要とする．しかし，音声インタフェースの利便性向上を得るには，長期に渡る継続的な運用での経験の蓄積とその運用を通じた実態調査とデータ収集は必要不可欠である．たけまるくんシステムは，それを可能にした．ここに述べた範囲に決してとどまることなく，様々な観点に立った検討を今後も続けていきたい．

第6章　話者年齢層識別能力を持つ音声インタフェース

6.1.　はじめに

大語彙連続音声認識で高い認識精度を得るには認識対象タスクに適した言語モデルが必要であることをここまで述べた．音響モデルに関しても同様のことが言え，音声インタフェースの使用環境や収録系に起因する雑音への適応技術，特定個人に適した音響モデルを提供する話者適応技術は，音声認識分野において最も盛んに研究がなされている重要課題である．言語モデルと音響モデルから構成される音声認識モデル作成時の認識対象の把握が，現在のモデルベースの音声認識を基礎とする音声インタフェースの利便性を決定する一つの大きな要因であり，システム開発において特に重要な行程である．

また，対話状況によって変化する認識対象をシステムが正確に把握し，状況に応じて使用する音声認識モデルを切り換えることで，はじめて高精度音声認識は実現される．ここでユーザに応じた柔軟な対話戦略を実現するアプローチの一つであるユーザモデルを例に挙げて考えてみる．ユーザモデルは，ユーザの持つシステムに関する習熟度，タスクドメインに関する知識レベル，利用時の性急度などの性質をモデル化したものであり，駒谷ら [57] は，不特定利用者向けの音声対話において，その性質に応じた対話戦略の決定に用いた．ユーザモデルは音声対話における応答精度向上をもたらす．同様に個々のユーザの性質を音声認識モデルの選択に反映できれば認識精度の向上を得ることができるだろう．

これら適応に関する研究を含め，従来のほとんどの音声インタフェースは大人を対象として設計されてきた [9]．それらシステムが中核とするのは大人音声から作成されたモデルを用いる音声認識である．そのため，前章 5 章の評価実験でも

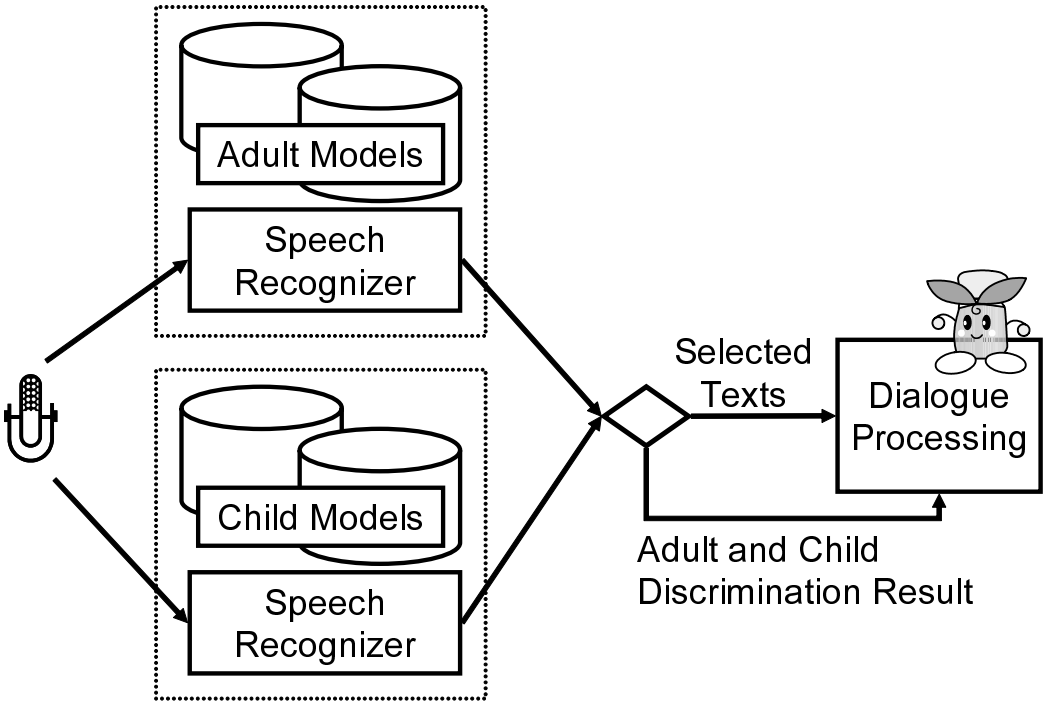


図 6.1 大人・子供識別能力を備えた音声インタフェース

明らかなように子供の音声認識精度が不足していた．しかし，今後，家庭や公共施設などに音声インタフェースが導入されることを考えると，ユーザに子供が多く含まれるのは当然であり，その数は無視できない．音声インタフェースの普及を見据え，大人のみならず子供の発話も扱うことができる音声認識の実現は重要な検討課題と言える．

そこで，本研究では，大人・子供の両利用者に対して柔軟な対話を可能にする技術を新たに音声インタフェースに導入する．ユーザモデルのアプローチを大人・子供別に適応した音声認識モデルと対話戦略の導入という形で音声インタフェース上に実装，主に子供の利用者に対する音声インタフェースの利便性向上を目指す．開発するシステムの概要を図 6.1 に示す．その実現に必要となる具体的な技術を整理すると，大人・子供別音声認識モデルとそれらを用いた並列音声認識，話者が大人または子供かを識別する能力，識別結果による認識出力の選択，識別結果を考慮した応答生成手法が挙げられる．

表 6.1 6 章の実験で用いる収集発話

年齢層		男性	女性	性別不明	合計
(a)	幼児	76	1421	585	2082
(b)	低学年子供	1920	7961	2843	12724
(c)	高学年子供	934	1154	498	2586
(d)	大人	5520	2496	70	8086
(e)	高齢者	8	12	0	20
合計		8458	13044	3996	25498

本章では，これらの中から特に大人・子供の話者識別能力に着目する．音声認識スコアを素性とする機械学習に基づいた話者識別手法を提案し，その評価実験を行う．また，大人と子供各々に特化した音声認識モデルの提供を目的として，大人と子供のフィールドテスト収集発話で適応した音声認識モデルを構築した．認識精度に向上が見られたので，その結果も報告する．

6.2. フィールドテスト収集データ内の子供発話

本研究は，5 章の音声情報案内システム「たけまるくん」をプラットホームにして，大人・子供の両利用者に柔軟な順応力を備えた音声インタフェースを実装することを目標とする．よって，表 5.1 に示したたけまるくんのフィールドテスト収集発話を用いて以下の議論をすすめる．また，雑音の影響を除外して効果を明確にするために，収集発話の中から雑音が比較的少なく明瞭，さらに収録音のみから人の主観によって話者の年齢層が判別できた 25,498 発話のみを用いることにする．これら発話の分類結果を表 6.1 に示す．年齢層の区分に使用したクラスは，(a) 幼児，(b) 低学年子供（小学校 3 年生ぐらいまで），(c) 高学年子供（中学生ぐらいまで），(d) 大人， (e) 高齢者の五つである．本研究では，(a) から (c) を子供と見なす．この表の 68.2%が子供による発話であり，音声インタフェースで子供発話を扱うことの重要性がわかる．

5 章 5.4.2 節で述べた収集データの発話トピック分析では，大人と子供で音声情報案内システムを利用する際の興味の対象が異なることを明らかにした．よって，大人・子供別の対話戦略を適用し，これら発話傾向の違いを考慮することで，ユーザに適した応答が実現できる可能性が高い．この発話傾向の違いは，発話中の使用単語や言い回し表現やコンテクストなどにも影響する．つまり，大人・子供の話者識別に際しては，発話に含まれる言語的な特徴の考慮が識別精度向上をもたらすことを示唆している．

6.3. 年齢層別音声認識モデルによる子供発話認識性能の改善

従来の大人をターゲットとした音声認識では，子供発話に対して十分な精度を得ることができなかった．しかし，子供の発話から子供用に音声認識モデル（音響モデルと言語モデル）を学習することで，ある程度の精度回復を見込める [52][53][54]．この確認のために Julius による大語彙音声認識実験を行った．

まず，大人と子供に別個の音響モデル（性別非依存）と N-gram 言語モデルを収集発話から作成する．

6.3.1 音響モデル

音響モデルの学習データの諸元を表 6.2 に示す．フィールドテストの収集発話では含まれる音素に偏りがあり，トライフォンの学習に必要な音韻バランスが取れていない．音韻バランスの確保には，音韻バランスを考えて設計された日本音響学会新聞記事読み上げ音声コーパス（JNAS）[49] の音声データも学習に含める必要がある．そのため，大人向けの音響モデル学習には表 6.2 の JNAS と大人収集発話，子供モデルには JNAS と子供収集発話のデータを用いた．

さらに，大人・子供別の依存性を高めるため，MAP（Maximum *A Posteriori*）適応 [58] もしくは MLLR（Maximum Likelihood Linear Regression）適応 [59] を施した適応音響モデルも作成した．適応元モデルは単純に学習データを結合して

学習で作成した大人・子供別音響モデルである．生成モデルに対して，大人モデルには大人収集発話のみ，子供モデルは子供発話のみで適応を施し，各モデルの年齢層依存性を高めている．収集発話には収録系の雑音なども含まれているので，環境適応も同時に施されたと考えることができる．

MAP 適応とは HMM パラメータの再推定を最大事後確率推定（MAP 推定）法を用いて行う適応法である．MAP 推定の原理を概説する．確率変数 x の確率密度関数を $f(x|\theta)$ として，x の観測値 $\mathbf{X} = x_1, x_2, \cdots, x_N$ を用いてパラメータ θ を推定することを考える．このとき，最尤推定ではパラメータ θ は，

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \ f(\mathbf{X}|\theta) \tag{6.1}$$

で推定される．ここで θ も確率変数とみなし，その確率密度関数を $g(\theta)$ とする．これを θ の事前確率と呼ぶ．また，データ $\mathbf{X}$ を観測した後のパラメータ θ の確率密度関数 $g(\theta|\mathbf{X})$ は事後確率と呼ばれ，ベイズの定理に従い以下の式で表される．

$$g(\theta|\mathbf{X}) = \frac{f(\theta|\mathbf{X})g(\theta)}{\int f(\theta|\mathbf{X})g(\theta)d\theta} \tag{6.2}$$

よって，MAP 推定は次式に示すように事後確率を最大にするようなパラメータを求める問題となる．

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \ g(\theta|\mathbf{X}) = \underset{\theta}{\operatorname{argmax}} \ f(\theta|\mathbf{X})g(\theta) \tag{6.3}$$

この $\hat{\theta}$ は MAP 推定量と呼ばれる．もし事前知識がない場合は，$g(\theta)$ は定数になり，MAP 推定量は最尤推定量と一致する．MAP 推定は最尤推定と比べ，推定に用いるデータが少量でも頑健なモデル推定ができ，データが多くなるにつれて最尤推定に近づくことになる．

一方，MLLR 適応は適応元モデルから適応に用いる発話に対して尤度が大きくなるように線形写像行列を求め，音響特徴量空間を一括変換する適応法である．ここでは HMM を構成する正規分布の平均と分散の線形写像行列を最尤推定により求める．MLLR 適応は，その取り扱いやすさと性能の高さにより音響モデルの話者適応手法として広く用いられてきた．

表 6.2 音響モデル学習データの諸元

大人	収集発話（男女，大人） 4,197 文
子供	収集発話（男女，子供） 15,922 文
JNAS	JNAS 読み上げ音声（男女） 40,086 文

表 6.3 言語モデル学習データの諸元

大人	収集発話書き起こし（大人） 7,606 文，単語数 40k，異なり単語数 1.7k
子供	収集発話書き起こし（子供） 16,892 文，単語数 89k，異なり単語数 3.4k
Web	生駒市関連 Web ページテキスト 1,080,272 文，単語数 31,265k，異なり単語数 218.7k
QA	人手で収集した想定質問文テキスト 6,488 文，単語数 56k，異なり単語数 3.2k

6.3.2 言語モデル

言語モデルの学習に用いたテキストデータの諸元を表 6.3 に示す．まず，Web と QA テキストからベースとなる言語モデルを作成する（単語数 40k）．次に大人・子供の書き起こしテキストから 2 つの言語モデルを作成し，年齢層依存モデルとする．年齢層依存モデルを先ほど作成したベース言語モデルにモデル融合ツールを用いて重み 0.8 で融合し，大人・子供各々の年齢層適応モデルとする．最後に，モデルの高精度化のため，別途人手で作成した異なり単語数 441 のネットワーク記述文法を適用，年齢層適応モデルが持つ N-gram 確率を強化した．具体的な手順は 5.3.1 節のたけまるくん用の言語モデル作成手順に準ずる．

表 6.4 単語正解率（年齢層別音声認識モデルを使用） [%]

音声認識モデルタイプ	音響モデルの適応	テストセット	
		大人 500 発話	子供 500 発話
大人モデル	適応なし	92.5	(62.7)
	MAP	94.9	(70.0)
	MLLR	94.8	(68.8)
子供モデル	適応なし	(88.1)	72.8
	MAP	(82.6)	82.5
	MLLR	(83.7)	80.8
年齢層依存	MAP	93.3	81.6

6.3.3 音声認識実験

大人 500 発話，子供 500 発話のテストセット（5.5.1 節参照）での大語彙連続音声認識の結果として単語正解率を表 6.4 に示す．大人・子供発話ともに各々対応した音声認識モデルを用いることで最も高い認識率を得ることができた．大人の音声認識モデルで子供発話を認識すると，最も良い MAP 適応音響モデルの場合でも 70.0%であり，既存の大人モデルベースのシステムで子供発話の認識が困難であることを再確認した．子供モデルを用いることで 12.5%（MAP 適応音響モデル使用時）の向上を得た．

音響モデルの適応手法間の比較に関しては，MAP 適応モデルが最も高い精度を示した．適応なし音響モデルと比べ，MAP 及び MLLR 適応後の子供モデルは大人発話に対して認識率低下，子供発話には認識率向上を得ている．つまり，適応することで子供発話への依存性を高めることができたことがわかる．

一方で，大人と子供の両収集データから単一の年齢層非依存のモデルも構築した．その音声認識率は，大人と子供の年齢層依存モデルを用いた時よりも，大人発話に対して 1.6%，子供に対して 0.9%劣る結果となった（音響モデルに対して MAP 適応済み）．以上の結果から，話者年齢層ごとの音声認識モデルの選択は認識精度向上に有用であると言える．

6.4. 音声認識スコアに基づく話者年齢層識別

以下では，本章の提案手法である音声認識スコアに基づく大人・子供の話者識別手法について述べる．

提案手法では，音声認識結果から得られる音響的特徴及び言語的特徴を素性とする話者識別を行う．従来，話者認識にはGMM（Gaussian Mixture Model; 混合正規分布モデル）に基づく尤度比較が用いられてきた[60]．峯松ら[61]は，このGMMによる話者認識のアプローチを用いることで，成人と高齢者に話者を識別している．これらの手法では発話の持つ音響的特徴のみが用いられており，言語的特徴は考慮されていない．これは，発話のコンテクストに依存しない任意文章で識別できる方が利便性が高いと判断されたためである．これまでの実験に提供されたデータは主に読み上げ音声であり，その発話内容にユーザの性質が反映されていなかったことも理由である．しかし，ここまで述べたように音声インタフェースに対する発話内容には，大人と子供で異なる特徴が含まれるのが自然である．このため，言語的特徴も考慮することで識別精度の向上が見込まれる．また，本手法は大人・子供の発話傾向の違いを考慮しても発話のコンテクストを限定するものでなく，利便性が失われることもない．

本手法では，音響的特徴（Acoustic Property; 以下，AP）としてJuliusの第2パスが出力する音声認識スコアのフレーム平均音響対数尤度，言語的特徴（Linguistic Property; 以下，LP）として単語平均言語対数尤度を利用する．各々は次式で導かれる．

$$AP = \frac{\text{音響対数尤度}}{\text{入力音声のフレーム数}} \tag{6.4}$$

$$LP = \frac{\text{言語対数尤度}}{\text{出力単語列の単語数}} \tag{6.5}$$

音声認識モデルや収録系に変更があっても，その違いを吸収できるように，実際に用いる音響的特徴，言語的特徴には大人音声認識モデルを用いた音声認識結果から求めたもの（AP_{adult}，LP_{adult}）から子供音声認識モデルによるもの（AP_{child}，LP_{child}）を引いた差を使う．

図6.2は，適応なし音響モデル使用した音声認識の結果による音響的特徴 $AP_{adult} - AP_{child}$ の頻度分布である．音声認識には6.3節作成の大人，子供用の各音響モデ

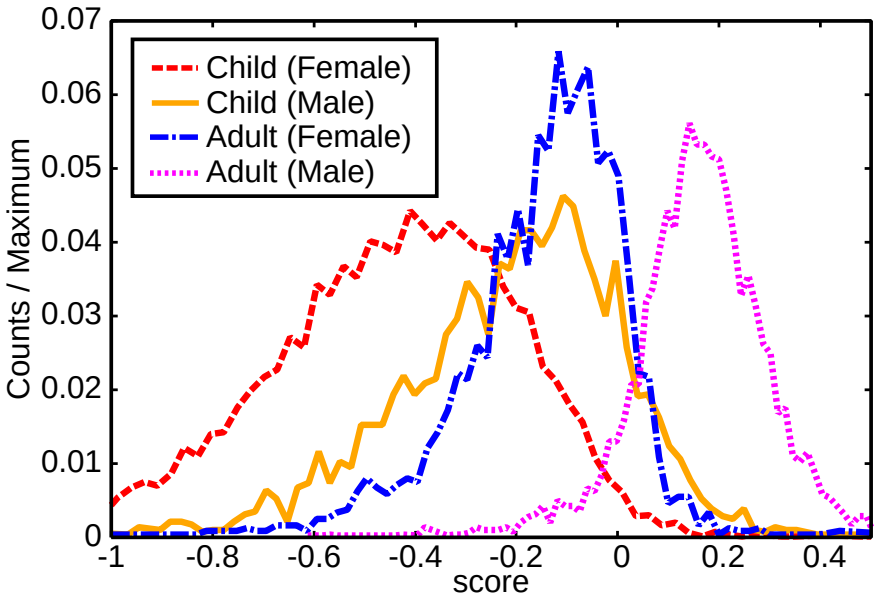


図 6.2 $AP_{adult} - AP_{child}$ の頻度分布 （適応なし音響モデル使用）

ル及び言語モデルを用いた．各グラフは，表 6.1 の子供女性，子供男性，大人女性，大人男性の収集発話を認識器に入力した時の出力結果である．大人男性の分布は明確に区別されるが，大人女性に関しては子供の分布と重なる部分が多い．これは大人女性の音響的特徴が子供に比較的似ているために生じた結果と思われる．次に MAP もしくは MLLR によって適応を施した音響モデル使用時の結果を示す（図 6.3，図 6.4）．大人・子供各々に適応がなされ，音響モデル学習に用いた JNAS 大人（男女）発話が持つ特徴が薄らいだ結果，大人女性の分布を子供から分離することができた．

同様に，言語的特徴 $LP_{adult} - LP_{child}$ の頻度分布を図 6.5 に示す（音響モデルには MAP 適応モデルを使用）．音響的特徴ほど大人と子供で分布傾向の違いは確認できない．しかし，若干ではあるが，分布の中心以外で，大人と子供の分布カーブに特徴的な違いが男女に共通して見られ，先の音響的特徴と組合わせた機械学習による識別精度向上が見込まれる．

識別のための機械学習アルゴリズムには，二値分類器である SVM（Support Vector Machine）[62][63] を用いた．SVM は，点在する事例で構成された多次元

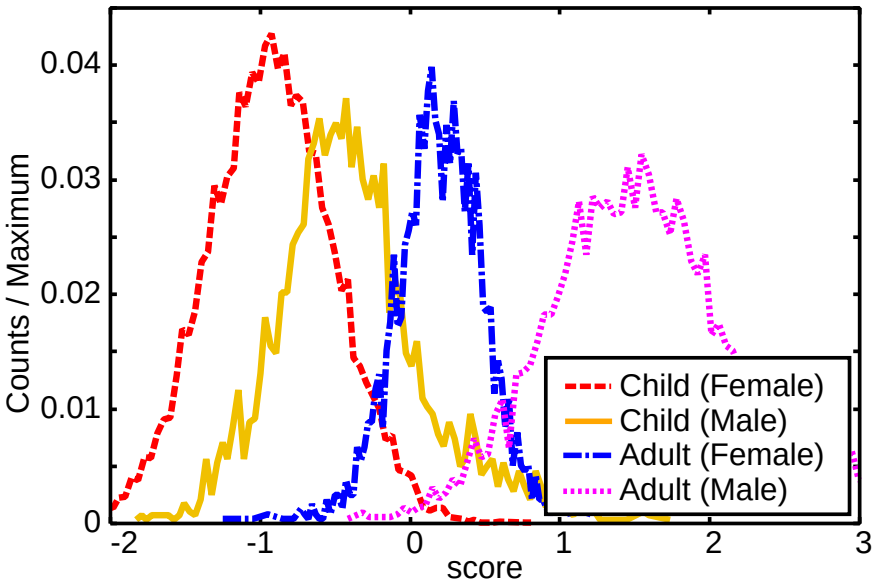


図 6.3 $AP_{adult} - AP_{child}$ の頻度分布 （MAP 適応音響モデル使用）

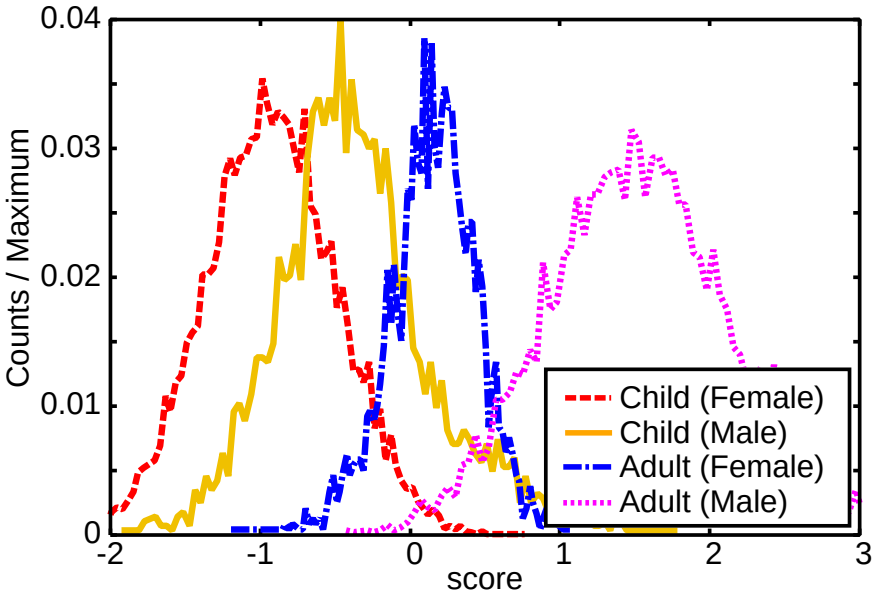


図 6.4 $AP_{adult} - AP_{child}$ の頻度分布 （MLLR 適応音響モデル使用）

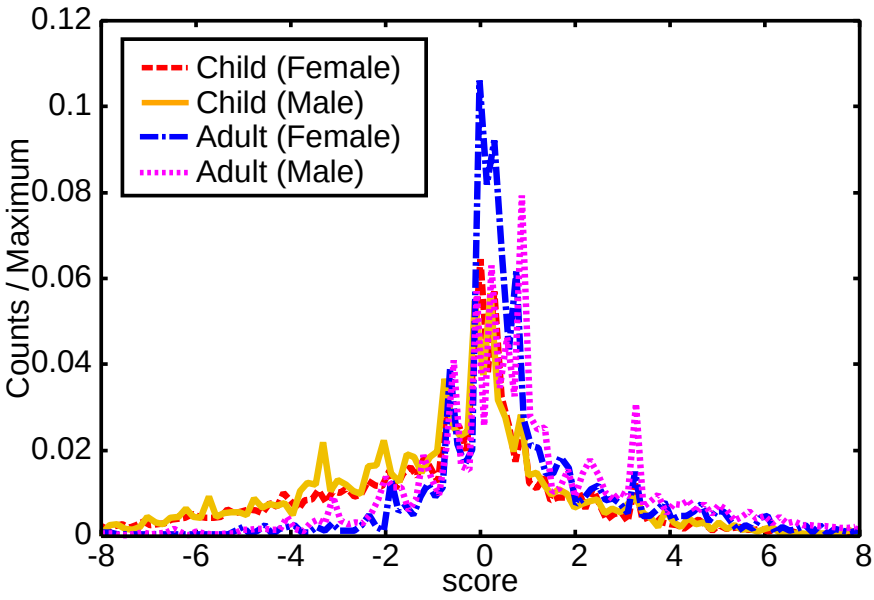


図 6.5 $LP_{adult} - LP_{child}$ の頻度分布

空間を二つに分類する境界面を，境界面をはさんで最も近い位置にある事例間の間隔（マージン）を最大化するように学習する．ここでカーネル関数を用いることで非線型な境界面を学習することも可能である．SVM は主に自然言語処理分野で広く使われるアルゴリズムであり信頼性は高い [64]．音声認識の研究では，複数の異なる音声認識の出力を統合する際に用いられ，小玉らの報告 [65] によると SVM は他手法より高い精度向上を得ることができる．本研究では，大人・子供別の音声認識モデルを用いて並列に音声認識を行い，その結果から得られる AP と LP で構成されるベクトルを SVM に与える事例とする学習を行った．SVM のカーネル関数には予備実験で最も良い結果を得た Gaussian 関数を用いた．

表 6.5 大人・子供識別率 [%]

特徴量タイプ	音響モデルの適応		
	適応なし	MAP	MLLR
音響的特徴のみ 1. $AP_{adult} - AP_{child}$	84.4	93.1	92.1
音響的特徴と言語的特徴 1. $AP_{adult} - AP_{child}$ 2. $LP_{adult} - LP_{child}$	87.1	94.6	93.3
音響的特徴と言語的特徴 （実測値を使用） 1. AP_{adult}　2. AP_{child} 3. LP_{adult}　4. LP_{child}	86.4	92.2	91.7
GMM 識別（ベースライン）			86.4

6.5. 評価実験

6.5.1 識別実験

提案手法を用いてテストセット（大人 500 発話，子供 500 発話）の話者識別実験を行った．SVM の学習に用いた音声は，表 6.1 の収集発話から抽出した大人，子供各 8,180 発話である．

実験結果を表 6.5 に示す．本手法は最高で 94.6%の識別率を得た．事前の検討の通り，特徴量に言語的特徴を加えた方が音響的特徴のみの場合より全体的に高い識別精度を示した．MAP または MLLR 適応した音響モデルの効果も確認でき，MAP 適応時が最も精度が良い．適応によって音響モデルの年齢層依存を高めることができた効果であると思われる．音響的特徴，言語的特徴において大人・子供音声認識モデル間の差ではなく実測値を用いると精度低下を起こす結果となった．

比較として従来法である GMM に基づく尤度比較を用いた識別実験も行う．GMM の学習に用いた音声はテストセットを除いた表 6.1 の全発話である．SVM

表 6.6 GMM による年齢層分類結果

正解		分類結果				
		(a)	(b)	(c)	(d)	(e)
(a)	幼児	45	8	1	-	-
(b)	低学年子供	65	182	114	2	-
(c)	高学年子供	1	22	43	17	-
(d)	大人	2	17	97	382	-
(e)	高齢者	-	-	1	1	-

による識別は大人・子供の二値分類だったが，GMM では (a) から (e) クラスの発話から各々64 混合のモデル計 5 個を作成する．学習に使用した音声分析パラメータは，音声認識のもの（表 2.1 参照）と同じ 16 bit，16 kHz 音声を窓シフト長 10 ms で分析した 12 次元の MFCC と Δ MFCC，Δ Power である．

GMM 尤度比較による分類結果を表 6.6 に示す．各結果は，5 モデル間で入力発話に対する尤度を比較して最も高い尤度を得たクラスに分類されたものである．表中の数字は，分類された発話の個数である．そして，(a) から (c) に分類された発話を子供，(d)，(e) に分類された発話を大人の発話と識別することにする．この結果，大人・子供の識別において正しく識別された発話は 86.4%（表 6.5 中ベースライン）であった．表 6.5 の提案手法の識別率は，この結果と比較して最大 8.2%の改善であり，提案手法の有効性が確認できたと言える．

6.5.2 識別結果を反映した音声認識率

提案手法の音声認識精度における影響を調べる．大人・子供テストセット計 1,000 発話に対して SVM を用いた提案手法によって大人・子供を識別する．そして，大人もしくは子供音声認識モデルが出力した認識出力（表 6.4 参照）を識別結果に従い選択し，認識率を算出した．単語正解率を表 6.7 に示す．選択方法が “SVM で選択” が，提案手法によって認識出力を選んだ時の結果である．一方，“

表 6.7 単語正解率（話者年齢層識別を反映） [%]

選択方法	音響モデルの適応	大人	子供	合計
SVM で選択	適応なし	92.3	72.6	83.1
	MAP	94.9	82.6	89.2
	MLLR	94.8	80.7	88.2
正解を選択	適応なし	92.5	72.8	83.3
	MAP	94.9	82.5	89.1
	MLLR	94.8	80.8	88.3
選択なし 年齢層非依存 単一モデル	適応なし	91.7	73.4	83.2
	MAP	93.3	81.6	87.9
	MLLR	93.4	80.2	87.3

正解を選択” は，自動識別には依らず，テストセット含まれる人の主観で作った年齢層ラベルを拠所に認識出力を選択した時のものである．“選択なし” は，大人・子供別の音声認識モデルを用いる並列デコーディングはせず，大人と子供すべての学習データ（表 6.2，表 6.3）から作成した単一の年齢層非依存モデルを用いて音声認識した際の結果である．

表 6.7 が示すように，単一モデルより年齢層依存モデルでの認識の方が高い認識率を得ることができた（適応音響モデルを使用時）．前説で最も高い年齢層識別率を得た MAP 適応音響モデル使用時は，提案手法で 89.2%，正解を選択で 89.1%の認識率である．これは選択なしの単一モデル使用時より 1.3%の精度向上である．また，本手法と正解を選択の認識精度はほぼ同等のものであり，提案手法による話者年齢層識別が有効に働いていることを確認できた．

6.5.3 応答正解率

最後に応答正解率での評価結果を述べる．選択された音声認識結果を用いてたけまるくんの応答を生成，その応答が満足なものかを人の主観によって判別した．

表 6.8 応答正解率（話者年齢層識別を反映） [%]

選択方法	大人（500 発話）	子供（500 発話）	合計（1,000 発話）
SVM で選択	79.6	59.8	69.7
正解を選択	79.4	60.0	69.5
選択なし	79.4	58.8	69.1
ベースライン	73.4	43.2	58.3

音声認識の音響モデルには，ここまで作成したモデルの中から MAP 適応のものを用いる．音声認識部分を除く，応答候補や用例テキストの数，応答生成のアルゴリズムは 5 章のものと同一である．

表 6.8 に結果を示す．選択なしの単一音声認識モデル使用時と比べ，0.6%ではあるが提案手法による正解率の改善を確認できた．なお，5.5.3 節の表 5.5 の結果（ベースライン）に対しての改善を確認することができた．今回の実験では，応答候補や用例テキストの追加は行っていないので，これは音声認識精度の向上がもたらした改善である．

6.6. たけまるくんシステムへの実装

筆者らの研究グループでは，以上の成果をたけまるくんシステムへ実装することで，利用者の年齢層に柔軟に順応できる音声インタフェースの開発を試みた．

西原ら [66] は，並列音声認識を大人・子供用のモデルを持つ Julius を 2 個独立にサーバモードで実行することで実装した．各 Julius には，録音プログラム（adintool）から TCP/IP 経由で入力音声の特徴量が渡され，結果として二つの認識出力のテキストを得る．続いて，システムは録音音声から話者が大人か子供かを識別し，その結果によって認識出力のどちらか一方を選択する．

開発したシステムでは，応答生成過程においても話者年齢層の識別結果を利用する．具体的には，大人と子供別の応答候補を収集したログを参考に追加する作業がすすんでいる．これは大人向けと子供向けに特化した応答を切り換えること

で柔軟な情報案内を可能にした．一例を挙げると,「タバコを吸うところはありますか?」という問いに対して，ユーザが大人なら「喫煙コーナーは，この壁の裏側にございます.」と場所案内を行うが，子供に対しては「タバコなんて吸ったら駄目.」と応答することができ，音声インタフェースの柔軟性を獲得し，利便性を向上することができたと考える．

同グループでは，開発したシステムのフィールドテストを実施している．

6.7. 本章のまとめ

本章では，大人・子供の両利用者に柔軟な順応力を備えた音声インタフェースを検討し，その実装に必要な音声認識スコアに基づく話者の大人・子供識別法を提案した．本手法は，機械学習のパラメータに音響的特徴と言語的特徴を併用することにより高い識別性能を有する．

また，子供発話の音声認識性能を調査し，既存の大人発話ベース音声認識の性能不足を確認した．子供発話から音声認識モデルを再構築することで一定の精度向上を得ることはできたが，その認識精度は大人と比べると低く，さらなる改良の必要性は依然として残る．

実験では，音声情報案内システムたけまるくんのフィールドテスト収集発話を用いた提案手法の評価を行い，GMM の尤度比較に基づく従来の識別法より 8.2%の識別率改善を得ることができた．音声認識精度と応答正解率においても性能向上を確認することができた．これら結果から，本手法は音声インタフェースの利便性向上に寄与するものであると判断し，結論とする．

話者識別手法に関する今後の検討事項としては，SVM 以外の機械学習アルゴリズムや音声認識結果の対数尤度以外の素性の導入が挙げられる．

大人と子供では音声情報案内システムを利用する際の興味の対象が異なる．この違いに即した対話戦略の実装も今後の課題である．単純なアプローチとして，大人と子供で別々に応答文章の候補を用意しておき，識別結果に基づき適した方の応答を返す方法を述べた．さらなる改良として識別結果を応答選択のスコア計算の過程に反映させる必要もあるだろう．

第7章　結論

7.1.　本論文のまとめ

本論文では，大語彙連続音声認識を基盤技術に人と機械の音声コミュニケーションを可能にする音声インタフェースについて考え，その利便性向上や普及のために解決が必要な問題点に着目した．この分野の先駆者たちは，音声インタフェースを現状で実用に足るレベルのものにしている．しかし，我々の日常社会に導入される機会はまだ少数であり，実環境における音声インタフェースの利用実態に関する調査は十分でない．このため，システムの開発や改良をすすめるうえでの方針となる評価や検証が従来の音声インタフェース研究に足りないことが問題であった．そこで，本研究では，4 章と 5 章で開発した受付案内ロボットと音声情報案内システムを用いて公共の場での音声インタフェースの利用実態調査を実現し，この結果，フィールドテストで収集した発話データは新たなる課題解決を可能にした．また，要素技術の開発にも取り組んだ．3 章で述べたタスク適応言語モデルの構築法は，音声インタフェース開発のコスト削減を成し，普及に貢献するものである．6 章で提案した話者年齢層識別手法は音声インタフェースの利便性向上に寄与するものである．

以下では 3 章から 6 章の各章を改めてまとめる．

3 章では，音声インタフェース開発の準備段階に必要なタスク適応 N-gram 言語モデルの構築手法について述べた．現在の大語彙連続音声認識では，経験を抽象化したモデルベースの統計処理を行うため，モデルが認識対象と合致してないと十分な認識精度を得ることはできない．しかし，N-gram モデルの学習には大量のタスク依存テキストコーパスが必要など，その構築は容易ではなかった．本手法によりタスク適応言語モデルは半自動に構築可能である．その手順は，Web

検索を利用したコーパスの自動作成とトピック依存 N-gram モデルの構築，モデル間融合によるタスク操作，ネットワーク記述文法の適用によるモデル高精度化の技術から構成される．

4 章では，人との対話機能を持つ受付案内ロボット ASKA を開発，その構成について述べた．ASKA は，情報科学における研究プラットホームとして，画像処理などの要素技術（例えばアイコンタクト）を統合して，対話機能の実用性を高めている．ASKA の開発を通じて今後も様々な新しい知見を得ることができるだろう．しかし，ハードウェア保守などが理由で ASKA を日常的に運用することはできず，音声インタフェースの利用実態調査の目的を果たすことはできなかった．また，4 章の実験では，統計的言語モデルとネットワーク記述文法の音声認識性能を比較し，文法に含まれる単語間制約を使った N-gram 確率強化による文法適用言語モデルの有効性を確認した．

5 章では，生駒市北コミュニティセンターの音声情報案内システム「たけまるくん」について述べた．本システムの長期間フィールドテストを通じて音声インタフェースの利用実態調査は実施された．五ヶ月間にわたる運用の結果，時間にして 1,362 分の男女幅広い年齢層の利用者の発話を収集した．その収集発話の分析から本システムは有効に利用されていることを示した．評価実験では，大人は 86%の単語認識率と 76%の応答正解率を得て，本システムは一定の実用レベルを備えていることを確認した．収集したデータは様々な新たなる課題を明らかにし，その解決も可能にした．その一例として，中村ら [56] による摩擦音や衝突音の雑音や不要発話の識別が挙げられる．

音声インタフェースが家庭や公共施設へ今後普及することを考えると，子供の存在は無視できない．しかし，たけまるくん収集発話の分析から，従来システムでは子供に対する性能不足が明らかになった．6 章では，収集発話をモデル学習に使うことで子供発話認識精度の改良を試みた．さらに，利用者の年齢層に即した柔軟な対話を可能にする音声インタフェースを検討して，その実装に必要となる大人・子供の話者識別手法を提案した．提案手法は音声認識結果から導出する音響的特徴と言語的特徴を併用した機械学習に基づく．二値分類アルゴリズムである SVM を用いた実験では，識別率 94.6%を得た．これは音響的特徴のみを含

む GMM を使った従来法から 8.2%の改善である．提案手法は主に子供を対象に音声インタフェースの利便性向上をもたらすものである．

7.2. 今後の課題

今後の課題として音声インタフェースの利用実態調査の継続は最優先である．しかし，本論文の冒頭で述べたように音声インタフェースの適用分野は広く，その多様性に対応するためには，現在のたけまるくんシステムにとどまるべきではない．検討すべき事項は数多くあるが，その中から主だったものを以下に挙げる．

まず，サービス対象の広がりに関する課題である．本研究ではコミュニティセンターでの情報案内タスクにおける調査は実現したが，他のタスクでも同様の試みが求められる．例えば，家庭や仕事場での個人的な仕事のための音声インタフェースを考えたい．本研究では検討を避けた自動販売機や ATM 等の高度な正確性を必要とするタスクでの調査も環境が整い次第始めるべきだろう．

次に対話の深さに関する検討である．今回使った音声インタフェースは，一問一答形式の非常に単純なコミュニケーションを前提としている．実際には一問一答形式でも十分に実用的な音声インタフェースを構築できるのだが，人同士の会話は複数のインタラクションで構成されるのが自然であり，人と機械のコミュニケーションにも深い対話が求められている．たけまるくんシステムからの拡張として，過去の発話の履歴も参照する応答生成アルゴリズムが考えられる．しかし，対話が途中で破綻しないような対話処理の実装は一般に困難である．話者交代の検知も必要となる．本研究で集めた発話の分析をすすめ，フィールドテストに耐えうる深い対話処理の実現を目指したい．

利用者の多様性の対処も検討課題である．6 章では，話者年齢層識別手法を開発したが，たけまるくんシステムへの実装は不完全である．また，年齢層以外にも性別，システムの扱いに関する習熟度，タスクドメインに対する興味，性急度等，利用者の性質を多方面にわたって考慮できるシステムが必要である．それら性質はお互いが複雑に作用するため，統合にも多くの検討を要する．

最後に，音声入力と音声認識に起因する課題を挙げる．雑音や話者性に端を発

する認識精度劣化の他に，複数話者による同時発声に対する解決も必要であろう．その対策としては音声認識の前処理としての音源分離技術が考えられる．

7.3. あとがき

いつの日か，人に優しい音声インタフェースが我々の日常社会にも普及する時が来ることを願っている．本論文がその実現に必要な議論のきっかけになれば幸いである．

以上をもって本論文の結論とする．

謝辞

本論文は，奈良先端科学技術大学院大学情報科学研究科の音情報処理学講座での筆者の研究成果をまとめたものです．日々の研究活動や私生活において学内，学外問わず数多くの方々にお世話になりました．ここに感謝を意を込め，皆様の御名前を紹介させて頂きます．

奈良先端科学技術大学院大学情報科学研究科 鹿野清宏教授には筆者が大学院在学中，主指導教官として終始御指導頂きました．大学院での研究活動を成し遂げることができたのは，鹿野先生の研究に対する深い御理解と的確かつ暖かい御助言に励まされたおかげです．鹿野先生は他では類を見ない恵まれた研究環境を筆者に提供してくださいました．また，学内外の研究者の皆様と出会え，活発な議論に参加できたのも先生のマネージメントによる研究プロジェクトに参加できたためです．鹿野教授の御尽力に敬意を表し，最初に深く感謝致します．

同研究科 小笠原司教授，木戸出正繼教授には副指導教官として本論文の執筆において御指導を頂きました．ASKA や学内研究プロジェクトの活動では，使いやすいインタフェースの在り方など，本研究をすすめるうえで指針となる重要なアイデアを御教示頂きました．ここに感謝致します．

同研究科 猿渡洋助教授には，脱線気味な筆者の研究を修正すべく，厳しくも的確な御助言を頂きました．深く感謝致します．

同研究科 助手の李晃伸博士には，先生が京都大学に在籍時から多くの御助言を頂きました．先生の作である音声認識プログラム Julius 無しでは本研究の実現は不可能でした．また，論文の添削や研究に関する討論に多くの貴重な御時間を割いて頂きました．心より御礼申し上げます．

同じく助手の川波弘道博士には研究室での生活を常日頃サポートして頂きました．

本研究の一部は，NEDO（新エネルギー・産業技術総合開発機構）の「シニア支援システムの開発」プロジェクトの支援を受けて行われました．黒田由香氏（TIS），小松久美子氏（イメージ情報科学研究所），馬場朗氏（松下電工，本学博士課程在学），芳澤伸一博士（松下電産）をはじめとするプロジェクト関係者の皆様に深く感謝致します．

受付案内ロボット ASKA の開発に携わってくださった ASKA プロジェクトの皆さんにお礼を述べます．特に活動の中心となって先導してくださったロボティクス講座 松本吉央助教授，同講座博士課程の怡土順一氏，小枝正直氏の高い技術力には多くのことを学ばせて頂きました．

音声情報案内システム「たけまるくん」のフィールドテストは「生駒市北コミュニティセンター ISTA はばたき」で実施しました．生駒市市役所 白本 和久氏，北コミュニティセンター 米田 秀一館長をはじめとする生駒市職員の皆様の御尽力に感謝致します．

筆者が音声インタフェースに興味を持つきっかけとなった学部卒業研究を指導してくださった名古屋大学工学研究科 板倉文忠教授に感謝致します．また，同大学情報科学研究科 武田一哉教授，情報連携基盤センター 梶田将司助教授には名古屋大学卒業後も研究を進めるうえで重要な数多くの御助言を頂きました．

音声認識に関する多くの知識と経験を与えてくれた（社）情報処理学会音声言語情報処理研究会連続音声認識コンソーシアムの活動に敬意を表します．代表の河原達也教授（京都大学），実行幹事の伊藤克亘助教授（名古屋大学）をはじめとする音声認識の発展と普及活動に関わっておられる研究者の皆様に感謝致します．

奈良先端科学技術大学院大学元教官の中村哲博士（現 ATR 音声言語コミュニケーション研究所第一研究室室長）と陸金林博士（現愛知県立大学助教授）には，本学在職時に御指導頂きました．

音声認識技術に携わる同期として仲間の京都大学情報学研究科博士課程の南條浩輝氏には多くのアドバイスを頂きました．

日々，有意義かつ楽しい研究生活をサポートしてくださった音情報処理学講座（鹿野研究室）の先輩，友人の皆さんに御礼を述べます．修士課程二年間ほとんどの時間を常に一緒に過ごした長友健太郎（NEC），山田実一（アドバンスト・

メディア）の両氏，筆者の属する研究グループの後輩になってしまったがために数多くのハードワークと困難に直面することになってしまった内田賢志君（富士通），鶴身玲典君（日立製作所），西原洋平君（本学修士課程），中村敬介君（本学修士課程）には特に感謝しています．先輩の西浦敬信博士（和歌山大学），榎本成悟氏（京都大学），立蔵洋介博士（静岡大学），廣瀬良文氏（松下電産）には，研究の立ち上げから日々の細かい御指導まで御支援に感謝致します．Panikos Heracleous 博士（本学 COE 研究員）と Randy Gomez 氏（本学博士課程）には，筆者の拙い英語文章の校正をお願いしました．そして，友人の戸田智基博士（名古屋工業大学），西川剛樹君（本学博士課程），阿部敬子さん（鹿野研秘書）をはじめ，個々にお名前を挙げることはできませんが，先輩，同期，後輩，そして，歴代の秘書さん，技術補佐員さん，皆様に御礼を申し上げます．

最後に，ここまで耐え忍んで僕を支え続けてくれた家族への感謝の意を表しつつ本論文を終えるものとします．

参考文献

[1] Y. Sakagami, R. Watanabe, C. Aoyama, S. Matsunaga, N. Higaki, K. Fujimura: "The Intelligent ASIMO: System Overview and Integration," *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2002)*, pp.2478–2483, 2002.

[2] M. Fujita, Y. Kuroki, T. Ishida, T. Doi: " A Small Humanoid Robot SDR-4X for Entertainment Applications," *Proc. IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM2003)*, 2003.

[3] 速水 悟, 菅村 昇: "音声対話システムの研究と実用化の動向," 日本音響学会誌, vol.50, no.7, pp.574–580, 1994.

[4] S. Seneff, E. Hurley, R. Lau, C. Pao, P. Schmid, V. Zue: "Galaxy-II: A Reference Architecture for Conversational System Development," *Proc. 5th International Conferences on Spoken Language Processing (ICSLP98)*, pp.931–934, 1998.

[5] V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T.J. Hazen, L. Hetherington: "JUPITER: A Telephone-Based Conversational Interface for Weather Information," *IEEE Trans. Speech and Audio Processing*, vol.8, no.1, pp.100–112, 2000.

[6] 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, 2002-SLP-40-23, 2002.

[7] 有田 正剛, 島津 秀雄: “カーナビゲーションシステム用音声対話インタフェース,” 人工知能学会研究会資料, SIG-SLUD-9502-1, 1995.

[8] 小林 哲則: “ROBISUKE: 新世代の対話ロボット,” 人工知能学会研究会資料, SIG-Challenge-0318-1, pp.1–6, 2003.

[9] L. Bell, J. Gustafson: “Child and Adult Speaker Adaptation during Error Resolution in A Publicly Available Spoken Dialogue System,” *Proc. 8th European Conference on Speech Communication and Technology (EUROSPEECH2003)*, pp.613–616, 2003.

[10] 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: “カスタマイズ性を考慮した擬人化音声対話ソフトウェアツールキットの設計,” 情報処理学会論文誌, vol.43, no.7, pp.2249–2263, 2002.

[11] 新山 祐介, 徳永 健伸, 田中 穂積: “自然言語を理解するソフトウエアロボット: 傀儡,” 情報処理学会論文誌, vol.42, no.6, pp.1359–1367, 2001.

[12] 松坂 要佐, 東條 剛史, 小林 哲則: “グループ会話に参与する対話ロボットの構築,” 電子情報通信学会論文誌, vol.J84-D-II, no.6, pp.898–908, 2001.

[13] 神田 崇行, 石黒 浩, 小野 哲雄, 今井 倫太, 前田 武志, 中津 良平: “研究用プラットホームとしての日常活動型ロボット “Robovie” の開発,” 電子情報通信学会論文誌, vol.J85-D-I, no.4, pp.380–389, 2002.

[14] 神田 崇行, 今井 倫太, 小野 哲雄, 石黒 浩: “人—ロボット相互作用における身体動作の数値解析,” 情報処理学会論文誌, vol.44, no.11, pp.2699–2709, 2003.

[15] 渡辺 裕太, 関口 芳廣, 鈴木 良弥: “ビデオ装置を例とした家電品の音声対話機能について,” 情報処理学会論文誌, vol.44, no.11, pp.2690–2698, 2003.

[16] A.L. Gorin, G. Riccardi, J.H. Wright: “How May I Help You?,” *Speech Communication*, vol.23, pp.113–127, 1997.

[17] E. Ammicht, A.L. Gorin, T. Alonso: “Knowledge Collection for Natural Language Spoken Dialog Systems,” *Proc. 6th European Conference on Speech Communication and Technology (EUROSPEECH99)*, vol.3, pp.1375–1378, 1999.

[18] J. Glass, J. Polifroni, S. Seneff, V. Zue: “Data Collection and Performance Evaluation of Spoken Dialogue Systems: The MIT Experience,” *Proc. 6th International Conferences on Spoken Language Processing (ICSLP2000)*, vol.4, pp.1–4, 2000.

[19] 広瀬 良文, 伊藤 克亘, 鹿野 清宏, 中村 哲: “日本語ディクテーションシステムにおける被覆率の高い言語モデル,” 電子情報通信学会論文誌, vol.J83-D-II, no.11, pp.2300–2308, 2000.

[20] 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.

[21] A. Lee, T. Kawahara, K. Shikano: “Julius – An Open Source Real-Time Large Vocabulary Recognition Engine,” *Proc. 7th European Conference on Speech Communication and Technology (EUROSPEECH2001)*, pp.1691–1694, 2001.

[22] 中川 聖一: 確率モデルによる音声認識, コロナ社, 1988.

[23] 李晃伸，河原達也，武田一哉，鹿野清宏: “Phonetic Tied-Mixture モデルを用いた大語彙連続音声認識,” 電子情報通信学会論文誌, vol.J83-D-II, no.12, pp.2517–2525, 2000.

[24] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland: *The HTK Book (for HTK Version 3.2.1)*, Cambridge University Engineering Department, 2002.

[25] F. Jelinek: “Self-Organized Language Modeling for Speech Recognition,” *Language Processing for Speech Recognition*, pp.450–506, Mercel Dekker, Inc., 1990.

[26] 鹿野 清宏, 伊藤 克亘, 河原 達也, 武田 一哉, 山本 幹雄: 音声認識システム, オーム社, 2001.

[27] I.H. Witten, T.C. Bell: “The Zero-Frequency Problem: Estimating The Probabilities of Novel Events in Adaptive Text Compression,” *IEEE Trans. Information Theory*, vol.37, no.4, pp.1086–1094, 1991.

[28] P. Placeway, R. Schwartz, P. Fung, L. Nguyen: “The Estimation of Powerfull Language Models from Small and Large Corpora,” *Proc. 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP93)*, vol.2, pp.33–36, 1993.

[29] 松本 裕治, 北内 啓, 山下 達雄, 平野 善隆, 松田 寛, 高岡 一馬, 浅原 正幸: 日本語形態素解析システム『茶筌』 version 2.3.3 使用説明書，奈良先端科学技術大学院大学，2003.
http://chasen.aist-nara.ac.jp/

[30] 河原 達也, 李 晃伸, 小林 哲則, 武田 一哉, 峯松 信明, 嵯峨山 茂樹, 伊藤 克亘, 伊藤 彰則, 山本 幹雄, 山田 篤, 宇津呂 武仁, 鹿野 清宏: “日本語ディクテーション基本ソフトウェア（99 年度版）,” 日本音響学会誌, vol.57, no.3, pp.210–214, 2001.

[31] 伊藤 彰則, 好田 正紀: “単語およびクラス N-gram 作成のためのツールキット,” 電子情報通信学会技術研究報告, SP2000-106, pp.67–72, 2000.
http://palmkit.sourceforge.net/

[32] 山本 俊一郎, 伊藤 克亘, 鹿野 清宏, 中村 哲: “ディクテーションにおける日本語の特質を考慮した単語正解率判定ツール,” 日本音響学会講演論文集, 3-Q-19, pp.155–156, 1999.

[33] K. Tanaka, S. Hayamizu, Y. Yamashita, K. Shikano, S. Itahashi, R. Oka: "Design and Data Collection for A Spoken Dialogue Database in The Real World Computing (RWC) Program," *Proc. Third Joint Meeting of Acoustical Society of America and Acoustical Society of Japan*, pp.1027–1030, 1996.

[34] 南條 浩輝, 加藤 一臣, 李 晃伸, 河原 達也: "大規模な日本語話し言葉データベースを用いた講演音声認識," 電子情報通信学会論文誌, vol.J86-D-II, no.4, pp.450–459, 2003.

[35] 竹澤 寿幸: "いまこそ話しことば処理技術の研究を," 情報処理, vol.42, no.2, pp.173–177, 2001.

[36] X. Zhu, R. Rosenfeld: "Improving Trigram Language Modeling with The World Wide Web," *Proc. 2001 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2001)*, 2001.

[37] 伊藤 克亘, 松岡 達雄, 竹澤 寿幸, 武田 一哉, 鹿野 清宏: "大語彙連続音声認識研究のためのテキストデータ処理," 日本音響学会講演論文集, pp.105–106, 1996-9.

[38] 鹿野 清宏: "2000 年代に何をすべきか – 研究課題と取り組み – NEDO シニア支援システムと音声研究の課題," 情報処理学会研究報告, 2000-SLP-32-5, pp.79–80, 2000.

[39] 馬場 朗, 芳澤 伸一, 山田 実一, 李 晃伸, 鹿野 清宏: "高齢者音響モデルによる大語彙連続音声認識," 電子情報通信学会論文誌, vol.J85-D-II, no.3, pp.390–397, 2002.

[40] S.F. Chen, K. Seymore, R. Rosenfeld: "Topic Adaptation for Language Modeling using Unnormalized Exponential Models," *Proc. 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP98)*, vol.2, pp.681–684, 1998.

[41] M. Weintraub, Y. Aksu, S. Dharanipragada, S. Khudanpur, H. Ney, J. Prange, A. Stolcke, F. Jelinek, E. Shriberg: “LM95 Project Report: Fast Training and Portability,” *Research Notes No. 1*, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD., 1996.

[42] 長友 健太郎, 西村 竜一, 小松 久美子, 黒田 由香, 李 晃伸, 猿渡 洋, 鹿野 清宏: “相補的バックオフを用いた言語モデル融合ツールの構築,” 情報処理学会論文誌, vol.43, no.9, pp.2884–2893, 2002.

[43] 河原 達也, 住吉 貴志, 李晃 伸, 板野 秀樹, 武田 一哉, 三村 正人, 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 2002 年度ソフトウェアの概要,” 情報処理学会研究報告, 2003-SLP-48-1, 2003.

[44] 鶴身 玲典，李 晃伸，猿渡 洋，鹿野 清宏，“タスク文法による N-gram 確率の部分強化を用いた認識アルゴリズムの評価,” 情報処理学会研究報告, 2003-SLP-45-13, pp.77–82, 2003.

[45] 李 晃伸, 河原 達也, 鹿野清宏: “記述文法に基づく高性能連続音声認識エンジン Julian,” 日本音響学会講演論文集, 3-1-10, pp.111–112, 2001-10.

[46] H. Kozima, H. Yano: “A Robot that Learns to Communicate with Human Caregivers,” *Proc. The First International Workshop on Epigenetic Robotics*, 2001.

[47] Y. Matsumoto, A. Zelinsky: “An Algorithm for Real-time Stereo Vision Implementation of Head Pose and Gaze Direction Measurement,” Proc. IEEE Fourth International Conference on Face and Gesture Recognition (FG2000), pp.499–505, 2000.

[48] J. Ido, Y. Myouga, Y. Matsumoto, T. Ogasawara: “Interaction of Receptionist ASKA using Vision and Speech Information,” *Proc. IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI2003)*, pp.335–340, 2003.

[49] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, S. Itahashi: "JNAS: Japanese Speech Corpus for Large Vocabulary Continuous Speech Recognition Research," *The Journal of the Acoustical Society of Japan (E)*, vol.20, no.3, pp.199–206, 1999.

[50] 板橋 秀一: "騒音データベースと日本語共通音声データ DAT 版," 日本音響学会誌, vol.47, no.12, pp.951–953, 1991.

[51] 山出 慎吾, 李 晃伸, 猿渡 洋, 鹿野 清宏: "雑音に頑健な音韻モデルと教師なし話者適応," 電子情報通信学会技術研究報告, SP2002-124, pp.19–24, 2002.

[52] 小川 厚徳, 山口 義和, 松永 昭一: "小学生音声データベースの構築とそれを用いた子供音声認識の一検討," 電子情報通信学会技術研究報告, SP2002-36, pp.1–6, 2000.

[53] K. Shobaki, J.P. Hosom, R.A. Cole: "The OGI Kid's Speech Corpus and Recognizers," *Proc. 6th International Conferences on Spoken Language Processing (ICSLP2000)*, vol.4, pp.258–261, 2000.

[54] 中村 敬介, 西村 竜一, 李 晃伸, 猿渡 洋, 鹿野 清宏: "実環境音声情報案内システムにおける子供発話の認識," 日本音響学会講演論文集, 1-6-28, pp.55–56, 2003-9.

[55] 駒谷 和範, 河原 達也: "音声認識結果の信頼度を用いた効率的な確認・誘導を行う対話管理," 情報処理学会論文誌, vol.43, no.10, pp.3078–3086, 2002.

[56] 中村 敬介, 西村 竜一, 李 晃伸, 猿渡 洋, 鹿野 清宏: "実環境音声情報案内システムにおける環境雑音および不要発話の識別," 電子情報通信学会技術研究報告, SP2003-172, pp.13–18, 2004.

[57] K. Komatani, S. Ueno, T. Kawahara, H.G. Okuno: "User Modeling in Spoken Dialogue Systems for Flexible Guidance Generation," *Proc. 8th European Conf. on Speech Communication and Technology (EUROSPEECH2003)*, pp.745–748, 2003.

[58] J.L. Gauvain, C.H. Lee: “Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observation of Markov Chains,” *IEEE Trans. on Speech and Audio Processing*, vol.2, no.2, pp.291–298, 1994.

[59] C.J. Leggetter, P.C. Woodland: “Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous-Density Hidden Markov Models,” *Computer Speech and Language*, vol.9, pp.171–185, 1995.

[60] D.A. Reynolds, R.C. Rose: “Robust Text-Independent Speaker Identification using Gaussian Mixture Speaker Models,” *IEEE Trans. on Speech and Audio Processing*, vol.3, no.1, pp.72–83, 1995.

[61] 峯松 信明, 広瀬 啓吉, 関口 真理子: “話者認識技術を利用した主観的高齢話者の同定とそれに基づく主観的年代の推定,” 情報処理学会論文誌, vol.43, no.7, pp.2186–2196, 2003.

[62] V.N. Vapnik: *The Nature of Statistical Learning Theory*, Springer, 1995.

[63] V. Wan, S. Renals: “SVMSVM: Support Vector Machine Speaker Verification Methodology,” *Proc. 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2003)*, vol.2, pp.221–224, 2003.

[64] T. Kudo, Y. Matsumoto: “Fast Methods for Kernel-based Text Analysis,” *Proc. 41st Annual Meeting of the Association for Computational Linguistics (ACL2003)*, pp.24–pp.31, 2003.

[65] 小玉 康広, 渡辺 友裕, 宇津呂 武仁, 西崎 博光, 中川 聖一: “機械学習を用いた複数の大語彙連続音声認識モデルの出力の混合,” 情報処理学会研究報告, 2003-SLP-45-16, pp.95–100, 2003.

[66] 西原 洋平, 西村 竜一, 李 晃伸, 猿渡 洋, 鹿野 清宏: “ユーザ層別並列モデルを用いた音声情報案内システム,” 日本音響学会講演論文集, 1-8-22, pp.49–50, 2004-3.

研究業績

学術論文

1. 西村 竜一, 西原 洋平, 鶴身 玲典, 李 晃伸, 猿渡 洋, 鹿野 清宏: “実環境研究プラットホームとしての音声情報案内システムの運用,” 電子情報通信学会論文誌, vol.J87-D-II, no.3, pp.789–798, 2004.

2. 西村 竜一, 梶田 将司, 武田 一哉, 板倉 文忠, 鹿野 清宏: “Web ベースコースウェアのための音声入力システムの開発,” 情報処理学会論文誌, vol.42, no.3, pp.605–613, 2001.

3. 長友 健太郎, 西村 竜一, 小松 久美子, 黒田 由香, 李 晃伸, 猿渡 洋, 鹿野 清宏: “相補的バックオフを用いた言語モデル融合ツールの構築,” 情報処理学会論文誌, vol.43, no.9, pp.2884–2893, 2002.

国際会議

1. Ryuichi Nisimura, Akinobu Lee, Hiroshi Saruwatari, Kiyohiro Shikano: “Public Speech-Oriented Guidance System with Adult and Child Discrimination Capability,” *Proc. 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004)*, Montreal, Canada, 2004.

2. Ryuichi Nisimura, Yohei Nishihara, Ryosuke Tsurumi, Akinobu Lee, Hiroshi Saruwatari, Kiyohiro Shikano: “Takemaru-Kun: Speech-Oriented Information System for Real World Research Platform,” *Proc. First Inter-*

national Workshop on Language Understanding and Agents for Real World Interaction, pp.70–78, Sapporo, Japan, 2003.

3. Ryuichi Nisimura, Takashi Uchida, Akinobu Lee, Hiroshi Saruwatari, Kiyohiro Shikano, Yoshio Matsumoto: "ASKA: Receptionist Robot with Speech Dialogue System," *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2002)*, pp.1314–1319, Lausanne, Switzerland, 2002.

4. Ryuichi Nisimura, Kumiko Komatsu, Yuka Kuroda, Kentaro Nagatomo, Akinobu Lee, Hiroshi Saruwatari, Kiyohiro Shikano: "Automatic N-gram Language Model Creation from Web Resources," *Proc. 7th European Conference on Speech Communication and Technology (EUROSPEECH2001)*, pp.2127–2130, Aalborg, Denmark, 2001.

5. Ryuichi Nisimura, Shoji Kajita, Kazuya Takeda, Fumitada Itakura: "Development of Speech Input System for WWW-based Courseware," *World Conference on Educational Multimedia, Hypermedia & Telecommunications (ED-MEDIA2000)*, Montreal, Canada, 2000.

6. Ryuichi Nisimura, Shoji Kajita, Kazuya Takeda, Fumitada Itakura: "Development of Speech Input System for WWW-based Courseware," Proc. First Annual WebCT Conference on Learning Technologies, Vancouver, Canada, 1999.

研究会

1. 西村 竜一, 中村 敬介, 李晃 伸, 猿渡 洋, 鹿野 清宏: "大人・子供に適応した音声情報案内のためのユーザ自動識別," 電子情報通信学会技術研究報告, SP2003-129/NLC2003-66, pp.97–102, 2003.

2. 西村 竜一, 西原 洋平, 鶴身 玲典, 李 晃伸, 猿渡 洋, 鹿野 清宏: “生駒市コミュニティセンター音声情報案内システムの開発と運用,” 情報処理学会研究報告, 2003-SLP-45-6, pp.35–40, 2003.

3. 西村 竜一, 内田 賢志, 李 晃伸, 猿渡 洋, 鹿野 清宏: “Julius を用いた学内案内ロボット用音声対話システムの作成,” 電子情報通信学会技術研究報告, SP2001-99/NLC2001-64, pp.93–98, 2001.

4. 西村 竜一, 長友 健太郎, 小松 久美子, 黒田 由香, 李 晃伸, 猿渡 洋, 鹿野 清宏: “Web からの音声認識用言語モデル自動生成ツールの開発,” 情報処理学会研究報告, 2001-SLP-35-8, pp.43–48, 2001.

5. 西村 竜一, 梶田 将司, 武田 一哉, 板倉 文忠, 鹿野 清宏: “音声入力を利用した Web コンテンツの作成支援環境,” 電子情報通信学会技術研究報告, SP99-32, pp.41–48, 1999.

6. 中村 敬介, 西村 竜一, 李 晃伸, 猿渡 洋, 鹿野 清宏: “実環境音声情報案内システムにおける環境雑音および不要発話の識別,” 電子情報通信学会技術研究報告, SP2003-172, pp.13–18, 2004.

7. 鹿野 清宏, 馬場 朗, 芳澤 伸一, 山田 実一, 西村 竜一, 小松 久美子, 黒田 由香, 李 晃伸: “高齢者音声の認識,” 電子情報通信学会技術研究報告, WIT2001-20, pp.25–30, 2001.

8. 長友 健太郎, 西村 竜一, 小松 久美子, 黒田 由香, 李 晃伸, 猿渡 洋, 鹿野 清宏: “相補的バックオフを用いた言語モデルの融合ツールの構築,” 情報処理学会研究報告, 2001-SLP-35-9, pp.49–54, 2001.

大会発表

1. 西村 竜一, 中村 敬介, 西原 洋平, 李 晃伸, 猿渡 洋, 鹿野 清宏: “ユーザの大人・子供を識別する音声情報案内システム,” 情報処理学会第 66 全国大会, 2004-3.

2. 西村 竜一, 李 晃伸, 猿渡 洋, 鹿野 清宏: “実環境データに基づく音声情報案内システムのための話者年齢層自動識別,” 日本音響学会講演論文集, 2-6-19, pp.97–98, 2003-9.

3. 西村竜一, 西原 洋平, 鶴身 玲典, 李 晃伸, 猿渡 洋, 鹿野 清宏: “音声対話エージェントによる生駒市コミュニティセンターの案内システム,” 情報処理学会第 65 回全国大会講演論文集, vol.2, pp.33–34, 2003-3.

4. 西村 竜一, 李 晃伸, 猿渡 洋, 鹿野 清宏: “音声対話機能を持つ受付案内ロボット ASKA の実装,” 日本音響学会講演論文集, 1-5-19, pp.37–38, 2002-3.

5. 西村 竜一, 怡土 順一, 李 晃伸, 松本 吉央: “情報科学研究の実環境プラットフォームとしての受付案内ロボット ASKA,” 情報処理学会第 64 回全国大会講演論文集, vol.4, pp.565–570, 2002-3.

6. 西村 竜一, 長友 健太郎, 小松 久美子, 黒田 由香, 李 晃伸, 猿渡 洋, 鹿野 清宏: “Web からの音声認識用言語モデルの自動作成,” 情報処理学会第 62 回全国大会講演論文集, vol.2, pp.121–122, 2001-3.

7. 中村 敬介, 西村 竜一, 李 晃伸, 猿渡 洋, 鹿野 清宏: “実環境音声認識システムのための GMM を用いた環境雑音および不要発話の自動識別,” 日本音響学会講演論文集, 1-8-21, pp.47–48, 2004-3.

8. 西原 洋平, 西村 竜一, 李 晃伸, 猿渡 洋, 鹿野 清宏: “ユーザ層別並列モデルを用いた音声情報案内システム,” 日本音響学会講演論文集, 1-8-22, pp.49–50, 2004-3.

9. 鹿野 清宏, 西村 竜一: “ロボット音声インターフェースにおける研究プラットフォーム,” 平成 15 年電気関係学会関西支部連合大会, S11-5, p.S60, 2003-11.

10. 中村 敬介, 西村 竜一, 李 晃伸, 猿渡 洋, 鹿野 清宏: “実環境音声情報案内システムにおける子供発話の認識,” 日本音響学会講演論文集, 1-6-28, pp.55–56, 2003-9.

11. 西原 洋平, 西村 竜一, 鶴身 玲典, 李 晃伸, 猿渡 洋, 鹿野 清宏: “生駒市コミュニティセンター音声情報案内システムの評価,” 日本音響学会講演論文集, 2-4-21, pp.99–100, 2003-3.

12. 内田 賢志, 西村 竜一, 李 晃伸, 猿渡 洋, 鹿野 清宏: “学内受付案内タスクにおける音声認識の検討,” 日本音響学会講演論文集, 1-5-24, pp.47–48, 2002-3.

13. 長友 健太郎, 西村 竜一, 小松 久美子, 黒田 由香, 李 晃伸, 猿渡 洋, 鹿野 清宏: “相補的バックオフを用いた言語モデル融合アルゴリズム,” 情報処理学会第 62 回全国大会講演論文集, vol.2, pp.119–120, 2001-3.

14. 小松 久美子, 黒田 由香, 長友 健太郎, 西村 竜一, 李 晃伸, 鹿野 清宏: “高齢者タスクにおける話し言葉言語モデルの構築,” 情報処理学会第 62 回全国大会講演論文集, vol.2, pp.117–118, 2001-3.

解説

1. 鹿野 清宏, 馬場 朗, 芳澤 伸一, 西村 竜一, 李 晃伸: “高齢者話者音声の認識,” 日本音響学会誌, vol.58, no.8, pp.512–517, 2002.

受賞

1. 情報処理学会第 65 回全国大会 大会奨励賞.