音声翻訳構成技術の
機能拡充と統合化の研究

Advancement and Integration of
Component Technologies for Speech Translation

2010年3月

早稲田大学大学院 国際情報通信研究科
国際情報通信学専攻 音声言語情報処理研究II

中嶋 秀治

目 次

図 目 次

表 目 次

第1章　序論

1.1　背景

言語の壁を越えて様々な国とのグローバルな情報通信を実現することは人類長年の夢であり，重要な研究課題である．その中でも，人と人との即時のコミュニケーションにおいて，音声は最もよく用いられている媒体であり，音声翻訳の研究が精力的に行われている．例えば，音声翻訳の実現は内閣府のイノベーション 25 （2025 年までを視野に入れた成長に貢献するイノベーションの創造の為の長期的戦略指針）の研究テーマの 1 つ（多国間スーパーコミュニケーションの実現）にも採択されている．また，音声翻訳技術の社会還元を一層加速するための新しい音声翻訳の内閣府社会還元加速プロジェクト，総務省のユビキタス特区プロジェクト，音声・言語処理の研究開発拠点として情報通信研究機構（NiCT）の MASTAR プロジェクトが現在進められている [40]．これらのことからも音声翻訳の研究の社会的な必要性の高さを窺うことができる．

音声翻訳は，音声認識，言語翻訳，音声合成，といった要素技術を必要とする複合的なシステムであり，各要素技術の構築とそれらの統合化が必要となる．各要素技術は実データから人間によって，または，計算機によって自動的に規則やモデルを作成することにより構築される．構築時に参照されたデータと各要素技術が適用される時のデータとの間の不一致は不可避であり，この問題を克服することが課題となる．すなわち，音声翻訳システムの設計に用いた情報や知識の範囲で完結するわけではなく，適用時に遭遇する未知の語や未知の発話スタイルや未知のタスクが音声翻訳の処理性能に常に影響を与える．よって，学習していない未知の状況においても正しく推定できる未知情報の推定技術の構築が重要な研究課題となる．統合化に関しては，従来，音声翻訳の要素技術をそのままつなぎ合わせれば，あたかも，音声翻訳システムを構築できると考えられていたかのように，これらを統合する際に顕在化する課題への取り組みが少なかった．それらの課題の 1 つとして，各要素技術間，および，要素技術と人間との間における様々な情報の統一化課題がある．これは各要素技術の単純な結合では解決されない課題であり，また，各要素技術の観点から見れば，その外側で生じる課題であるため，見過ごされる可能性が高く，研究が少なかった．

本論文は，自然な音声翻訳通信の実現を目的として，音声翻訳構成技術のうちの音声認識と音声合成に関わる音声言語処理を中心に，未知情報の推定という課題を解決する要素

技術の機能拡充と，要素技術間での情報統一化課題を解決する統合化技術を提案する．

1.2　研究課題

音声翻訳通信の実現には様々な要素技術の搭載が必要である．円滑快適なコミュニケーションを支援するためには，実時間で動作するシステムを構成するための端末機器やマイクやスピーカーの設計等も重要な課題ではあるが，本論文では音声言語処理（音声認識と音声合成）に関わる部分に焦点を当てる．

本論文では，言語の壁を越えた人と人とのコミュニケーション（または，対話）を支援する音声翻訳の実現を目的とするため，読み上げ口調の発話ではなく，対話口調の発話の音声言語処理に関係する課題の解決を行なう．要素技術の一般的な改良だけではなく，対話口調を対象として扱うことは音声翻訳の実現には必須となる．現在は，コーパス観察やコーパスからの統計的学習に基づいて構築された処理規則やモデルを用いることが主流になっている．読み上げ口調で発話される文は書き言葉に近く，インターネット上の文書や新聞記事のアーカイブなどからのコーパス作成が可能である．一方，対話のコーパスは収集が困難であり，対話コーパスが未整備のタスクや言語が生じる問題は免れない．よって，少量または未開の言語やタスクでの統計的アプローチの適用は本質的な大きな課題である．

音声翻訳システムの構成図に関係づけて，音声翻訳実現にとって解決が必要な課題の例を図示する（**図 1.1**）．以下，音声翻訳にとっての入力側である発話者から出力側である聴取者に向かう順に，すなわち，音声認識，音声合成の順に課題の概要を述べる．

A　発話者と音声認識との間に位置するもの

1. 指示語と指示動作のそれぞれで指された実体情報間の情報の統一
   言語翻訳において，翻訳先の言語がドイツ語のような場合には，指示動作で指された物の性別に対応したドイツ語の指示語を用いることが必要となる．すなわち，指示語とそれが指す指示対象物との対応付けが必要となる．また，翻訳に要する時間がゼロではないことと，翻訳の前後で語順が必ずしも同じではないことから，翻訳後の指示語に同期して指示動作を表出するには，指示対象物の認識がやはり必要となる．
2. 話者の（性別などの）属性と表現との間での統一
   例えばタイ語では，文末表現が発話者の性別に応じて使い分けられている．タイ語への言語翻訳では発話者の性別の認識が必要となる．発話者が日本語を発話した場合，日本語文には性別の差が必ずしも現れない．しかし，この課題は簡単な性別登録でも解決できる．また，音声認識性能の改善への期待から，性

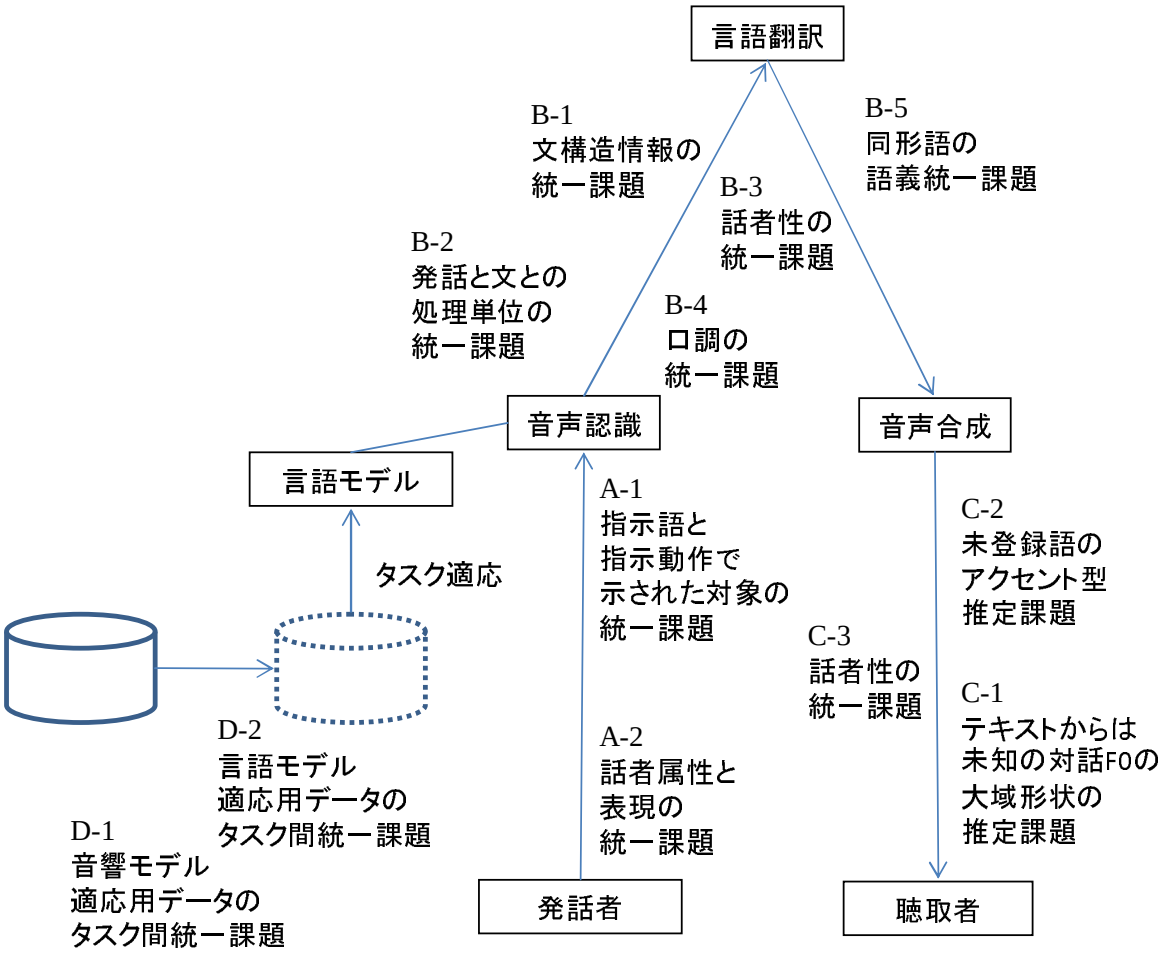

図 1.1: 音声翻訳の全般的な課題

別ごとに用意された音響モデルを備え，尤度の高いモデルを自動選択する等の研究がおこなわれている．そのモデル選択の結果を用いた性別判定も可能である．

B　音声認識と言語翻訳（や後段の音声合成）との間に位置するもの

1. 音声中の韻律で表現された文構造情報の統一
   従来の言語翻訳過程では，文構造の情報が必要である場合には，音声認識結果の文字列から独自に解析を行なってきた．最近では，音声認識性能の改善への期待から，言語モデルに構造情報を取り込む試みが既に多数行われており，そのようなモデルの適用結果から文構造の情報を得ることも可能である．
2. 情報処理単位の統一
   音声認識では発話単位，言語翻訳は文単位という処理単位の違いがこれまでにあった．処理単位が異なると翻訳の結果が大きく異なり，間を埋めるべき重要な課題であるが，あまり研究が無かった．
3. 話者性の統一
   翻訳先言語での合成音声が発話者自身による外国語での発声のように聞こえるようにするには必要と想像される．話者性を保存し，自然で人間らしい合成音声に必要な特性が何であるのかは未知の状況にある．
4. 口調情報の統一
   言語翻訳の原言語音声内で強調された箇所の特定が，翻訳先の言語での音声合成において強調するべき箇所の決定に役立つ可能性がある．将来の興味深い課題であるが，対話口調の研究は始まったばかりであり，本博士論文では後述する C の 1 にまず焦点をあてる．
5. 同形語の語義の統一
   字面が同じ単語間で読みの情報が欠落している場合，単語の多義解消（曖昧性解消）が必要となる．例えば,「最中」が「さいちゅう」か「もなか」であるかの違いは言語翻訳において問題となるが,（「〜の」や「を食べる」などの）前後の単語との共起関係を用いた分類問題としての解決が既に多数存在した．また，読みを含めた音声認識結果を出力すれば問題にならない場合もある．品詞の（下位範疇の）曖昧性解消も必要な場合がある．例えば,「石川」が県名であるか人名であるかの決定であるが，音声認識の研究で既に多数扱われている問題である．

C　音声合成と聴取者との間に位置するもの

1. 対話的な韻律を生成するための情報の欠落
   既存の音声翻訳器で利用されている音声合成器は読み上げ口調のものであった．しかし，音声翻訳器は対話場面で用いられるため，対話的な合成音が望まれる．このとき，対話の場面や対象とする発話内容に応じた韻律の制御が必要かつ重要である．音声合成器に入力される文にはそれらの韻律情報が明示されてはいないので，文からの予測が必要となる．しかし，対話口調の音声合成の研究は始まったばかりであり，対話の韻律の制御方法は解明されていない．
2. 未登録語のアクセント情報の欠落
   アクセントは音声合成時の目標情報として重要な情報であり，音声翻訳器の辞書に登録された情報を参照する．しかし，音声翻訳器にとっての未登録語のアクセントは参照できない．また，翻訳元の言語のアクセントが翻訳先のアクセントと一致するとも限らない．アクセントが聴取者の記憶にあるアクセントと異なる場合，聴取において違和感を与え，理解の障害にもなる．これらの点から，未登録語のアクセントの予測は，音声合成という要素技術単独でも，重要な課題である．
3. 合成音の話者性の欠落
   話者性のうち性別の違いは簡単な設定で克服できるが，全く別人の音声合成器が用いられることが多く，声の高低をはじめとする話者性は現状保存されていない．低（高）そうな声の話者の声を翻訳して，低（高）そうな合成音で合成するとその話者性の一部が保存されると予想される．

D　適応技術を介した適応元と適応先との間に位置するもの

1. 音響モデル適応用の適応先データの欠落
   教師なし学習などの適用により，音声認識器を利用し続ける間に適応用のデータを自動的に収集する研究が存在する．
2. 言語モデル適応用の適応先データの欠落
   新聞や WEB のニュース記事に比べて，対話の書き起こしデータは希少であり，作成には時間とコストがかかる．最近では WEB 上の blog などに砕けた書き言葉表現が大量に現れ出したが，日記であり，対話の書き起こし文とは必ずしも性質が同じものではない．話し言葉（対話）用の言語モデルの適応先データを用意することは重大な課題である．

これらの例が課題の全てではないが，これらのうち本博士論文で詳しく論じる課題を以下にまとめる（**図 1.2**）．すなわち，本論文では，

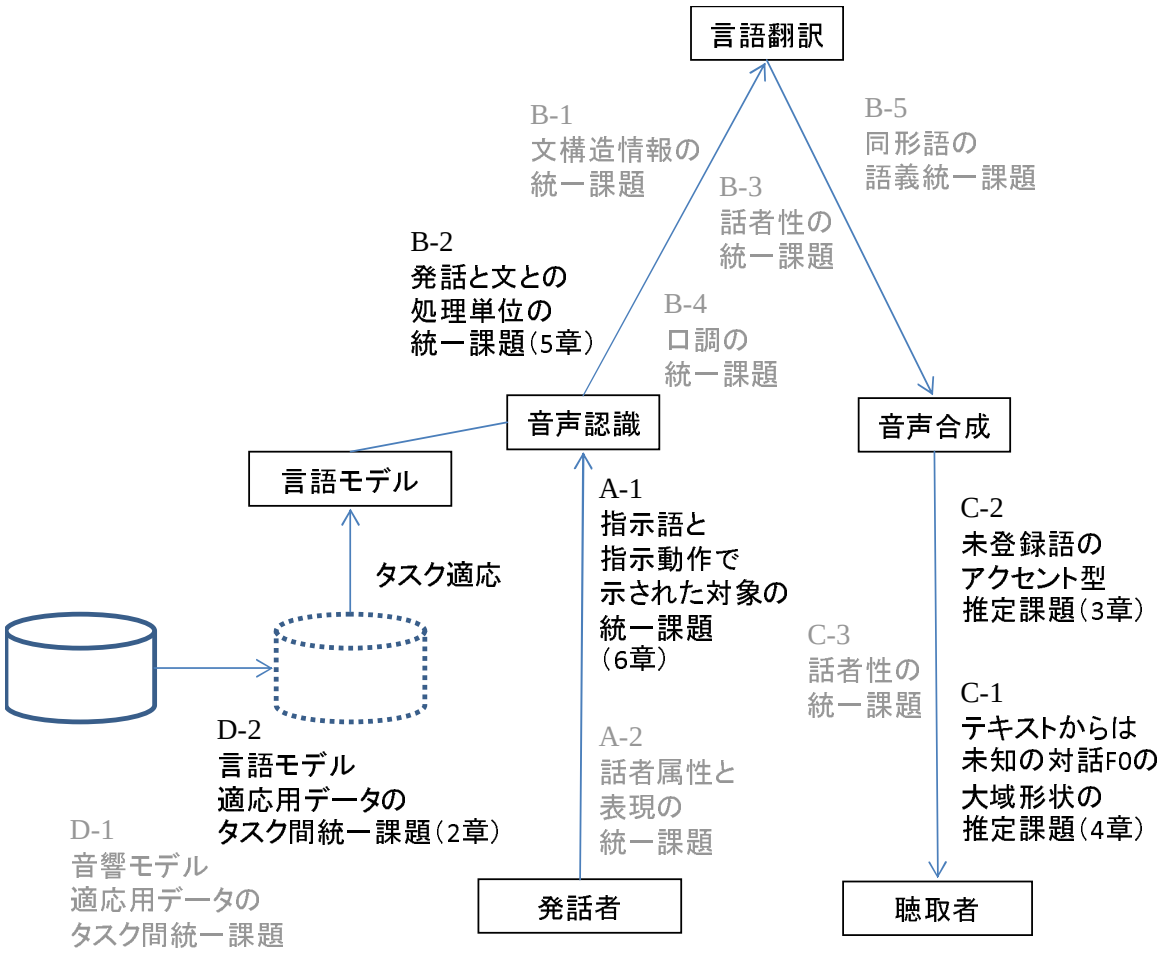

図 1.2: 本論文の課題と章番号との対応

1. 音声翻訳技術を実際のシステムとして稼働させる上で必要な要素技術そのものの機能を拡充する技術として

   (a) 他言語データの利用による統計的言語モデルの性能向上（第 2 章）
   前記 D-2 の，音声認識用統計的言語モデルの性能向上を目的とした，未知タスク・未知言語での未知の単語系列を推定するための，機械翻訳によって生成された追加テキストを使った統計的言語モデルの適応

   (b) 未登録単語のアクセント型推定（第 3 章）
   前記 C-2 の，音声合成器にとって未知である未登録語のアクセント推定

   (c) 対話韻律の実現に向けた F_0 大域形状の分析（第 4 章）
   前記 C-1 の，音声合成の対象文からは未知の対話音声の韻律自然性向上に向けた F_0 分析による制御要因の探索

2. 要素技術間での情報統一化課題を解決する統合化技術として

   (a) 音声認識と言語翻訳における処理単位の統一のための発話分割（第 5 章）
   前記 B-2 の，話し手，音声認識，言語翻訳の間に存在する処理単位統一への対処である，音声認識と言語翻訳の処理単位を揃えるための発話の文への分割

   (b) 指示語と指示動作との間での指示対象物情報の統一（第 6 章）
   前記 A-1 の，指示語と指示対象のそれぞれが指し示す情報の統一への対処である，指示動作と指示語との対応付け

の 5 つの課題に取り組み，解決技術の提案を行なう．これらの課題解決方法を提案することを通して，音声認識，言語翻訳，音声合成といった要素技術同士がさらに密接に連携した音声翻訳器の実現に本論文が貢献すると信じる．

1.3　論文の構成と各章の概要

以下，各章ごとに概要を述べる．

第 1 章では音声翻訳実現にとって必要となる課題のうち,（1）未知情報の推定という要素技術の機能拡充の必要な課題と,（2）要素技術間での情報の統一化という統合化の課題との 2 つの観点から，次章からの研究課題についてまとめた．さらに，本論文の構成，各章の概要について述べた．

まず，第 2 章と第 3 章と第 4 章において，音声翻訳器を適用する際に生じる未知情報の推定課題の解決を行なう要素技術の機能拡充について述べる．

音声認識のモデルを新規タスクに適用するためには，新規タスクのデータを用いた適応化によって精度が改善されることが従来から知られている．適応化のためには，適応先

タスクの少量データが必要となる．アプリケーションによっては，データ収集の問題は深刻で，時間的にも経済的にも実現上の大きな課題である．音声翻訳では，言語データが存在しない新言語等に利用する上でも，他言語のデータが使用できると，データ収集の問題をある程度回避できる場合が考えられる．第 2 章では，適応先タスクの適応用データとして，別言語の同タスクの文を機械翻訳にかけて得られた結果を用意し，それを用いて言語モデルの適応化を行う方式の検討結果を述べる．研究の結果，言語モデルの予測性能を示す平均単語分岐数（パープレキシティ）の向上が実験的に確認された。

第 3 章では，言語翻訳された結果を音声合成する際に問題となる，未登録単語のアクセント型の推定方式を述べた．一般語のアクセントや読みは，言語翻訳のリソースである辞書内に登録されている可能性が高く，それらを音声合成処理に用いることで自然なアクセントを実現することが可能である．一方，固有表現は数が膨大で，さらには，新たに作られる可能性も高く，あらかじめ登録しつくすことが困難である．読みは音訳により言語間での変換がある程度可能である．ところが，アクセントの位置は言語間で必ずしも同じではない．そのため，アクセント型の推定が必要となる．本研究では，アクセント型の推定問題をアクセント型の複数の候補のうちの 1 つへの分類問題として設定し，現在高い分類精度を示しているサポートベクトルマシンを分類器として適用し，効果を確認した。

言語翻訳器を介したコミュニケーションは対話であるので，読み上げを対象とした従来の音声合成では不十分であるが，世界的にも対話を対象とした音声合成は研究が始まったばかりである．第 4 章では，自然な対話音声のイントネーションの実現を目的として，対話種別ごとの実際の音声の基本周波数を重畳モデルの観点から分析を行なった結果について述べる．その結果，対話種別ごとの違いを大域的な F_0 変動が担い，対話種別ごとに特徴的な表現において読み上げ音声の基本周波数変動と比べた場合の差異が大きいこと，アクセント句内の局所的な変動は読み上げ音声との類似性が高いことが確認された．これらの結果から，対話種別ごとに特徴的な表現を手がかりとして大域的な F_0 変動を制御することで，対話口調の F_0 の制御方法を確立すればよいことが分かった．

次に，第 5 章と第 6 章で，音声翻訳を構成する各要素技術の結合では解決されない統合化の課題として情報統一化の課題の解決を行なう．

従来，音声認識器は発話を単位として処理を行うよう設計・研究されてきた．一方で，言語翻訳は文を単位として処理を行うよう設計・研究されてきた．言語翻訳においては，複数の文を含みうる発話そのものを単位として処理することは困難であり，音声認識過程内またはその後処理において言語処理の単位に統一化することが必要である．そこで，第 5 章では，音声認識と言語翻訳との間での処理単位を統一するため，すなわち，発話を文に分割するために，句点を 1 つの単語として組み込んだ統計的言語モデルを提案し，評価実験をおこなった．その結果，音声認識の過程において，文への分割を行わない従来の音声認識器での単語認識精度を損なうことなく，高い精度で発話を文に分割できることが明

表 1.1: 各章と原著論文との対応

章	内容	原著 (業績一覧参照)
2	他言語データの利用による言語モデルの性能向上	主論文（査読付学術論文）3
3	未登録単語のアクセント型推定	主論文（査読付学術論文）4
4	対話 F_0 大域形状の分析	紀要論文（査読付紀要論文）
5	処理単位統一のための発話分割	主論文（査読付学術論文）2
6	指示対象物情報の統一	主論文（査読付学術論文）1

らかとなった。

自然なコミュニケーションでは，手で対象物を指さしながら，同時に言葉では指示語（これ、それ、あれ等）を使って会話がなされる．指示語だけが翻訳されれば，言語翻訳の目的が達成されるわけではない．会話が続く中で，実際にどの対象物が指されていたのかを認識しておく必要がある．例えば，対象物の属性（男性名詞，女性名詞の別など）が指示語の訳語選択に影響を与える言語では重要となる．また，翻訳に要する時間がゼロではないことと，翻訳の前後で語順が必ずしも同じではないことから，翻訳後の指示語に同期して指示動作を表出するには，指示対象物の認識がやはり必要となる．第 6 章では，マウスを使った指示動作と指示語との対応付け手法について行なった研究をまとめた．指示語と指示動作との対応付けに焦点を当て，これら 2 種のモーダルを用いて WWW ブラウザを操作する場面での対応付けの精度の点で確認した．既存の Graphical User Interface（GUI）での操作様式に変更を加えることなく自然な指示動作を導入することができるので，離れた 2 地点に居る人間が共通の画面（GUI）を見ながら（リモート・アシスタンスの場面等），かつ，音声翻訳器を介した対話を行なう場面での指示動作と指示語とを用いた円滑な対話の実現が可能となる．

第 7 章では，本論文を総括し，さらに正確で自然性の高い音声翻訳器を実現するために必要と考えられる将来への課題を示した．なお，各章と原著論文との対応は**表 1.1** の通りである．

第2章　他言語データの利用による統計的言語モデルの性能向上

2章では，音声翻訳器を適用する際に生じる未知情報の推定課題の１つである，未知のタスクでの単語並びを予測する課題を解決する方式を提案する．

2.1　はじめに

統計的な言語モデルの作成には，そのモデルを利用するタスク（ターゲットタスク）における言語情報を反映した大規模コーパスが必要となる．大規模コーパスの入手が困難な場合には，言語モデルのタスク適応という手法がしばしば利用される．このタスク適応は，まず，ターゲットタスクとは必ずしも一致しない様々なタスクからなる大規模コーパス（以後,「一般コーパス」と呼ぶ）を使ってモデルを推定する．つぎに，ターゲットタスクについての小規模コーパスを利用して，そのモデルがターゲットタスクに対してうまくマッチするように「適応化」が行なわれる.（適応についての議論のため，以後,「ターゲットタスク」は「適応先タスク」と同じである.）言語モデルが多言語の話し言葉翻訳器のためのものである場合には各言語モデルでの適応が必要となるため，各言語の適応先タスクの小規模コーパスが必要となる．ところが，単一言語の話し言葉の小規模コーパスでさえコストの点で収集が困難であり，多言語のコーパスはなおさら収集が困難である．また，ある言語のあるタスクに対する認識システムを他の言語に移植する場合においても，同様に移植先言語におけるコーパス収集の問題が生ずる．

そこで本論文ではこれらの問題の解決を試みる．すなわち，ある言語で書かれた適応先タスクのコーパスを，タスク適応が必要な言語モデルの言語に機械翻訳し，その翻訳結果をタスク適応のための小規模コーパスとして利用して，言語モデルの適応化を行なう方法を提案する．また，本手法によって適応化を行なった言語モデルが，単語の予測性能を示す尺度であるテストセットパープレキシティーを大幅に改善することを実験によって示す.

機械翻訳において利用される知識として，辞書，統計的言語モデル，用例などが挙げられるが，これらの中に統計的言語モデルにとって最も重要と考えられる単語の連接制約に関する情報が保持されていることが期待できる．仮に機械翻訳の結果が訳文全体として不

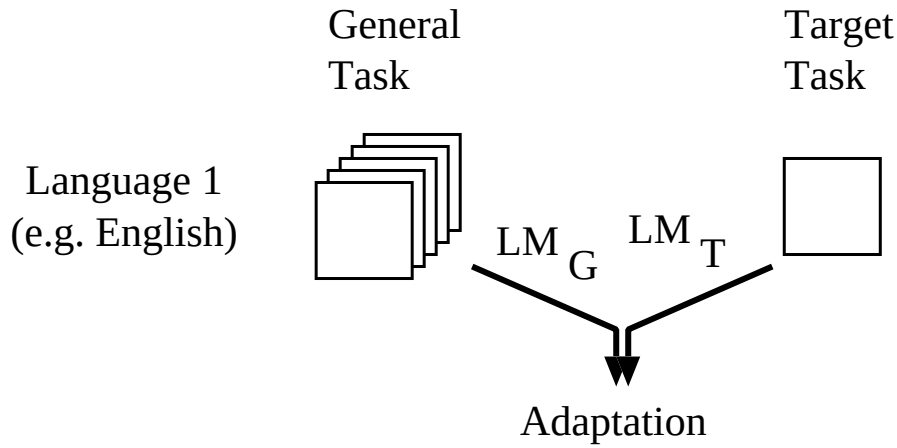

図 2.1: 適応という課題

自然であったり間違いを含んでいても，数単語からなる単語列のような局所的な文脈では適切な単語の連接制約が反映されていることが多いと考えられる．また，本提案法においては，適応先タスクのコーパスを翻訳するので，その翻訳結果も適応先タスクの話題や文のスタイルを反映していると考えられ，翻訳結果を言語モデルの適応用のコーパスとして利用できることが期待できる．

本論文では上記の方法の提案と評価とを行なう．まず **2.2** で従来と本研究での言語モデル適応の状況と方法とを述べ，用いる機械翻訳器の概要を **2.3** で説明する．そして，**2.4** で評価実験の条件と結果とを述べる．本手法と従来の研究との関係については **2.5** で述べ，まとめを **2.6** で行なう．

2.2　言語モデルの適応

はじめに言語モデルの適応という課題を明らかにする．**図 2.1** のように，適応という課題は，適応先タスクの小規模コーパス (図 2.1 の $Target\ Task$) を集めて，これを一般タスクの大規模コーパス (図 2.1 の $General\ Task$) と一緒に利用して，ターゲットタスクの性質を反映した（適応した）新たな言語モデルを作成することである．

一方，本研究での適応の状況は**図 2.2** のようになる．すなわち，図 2.2 で $Language\ 1$ の言語モデルの適応において，適応先タスクの小規模コーパス T_{L_1} が存在しない場合に，その代わりとなる言語コーパス $T_{L_1'}$ を，そのタスクでの他の言語 $Language\ 2$ で書かれた小規模コーパス T_{L_2} から機械翻訳によって生成する．そして，これを大規模の一般コーパス（図 2.2 左側の $Language\ 1$ 側の $General\ Task$ コーパス）とともに使って，適応化された言語モデルを作成する．

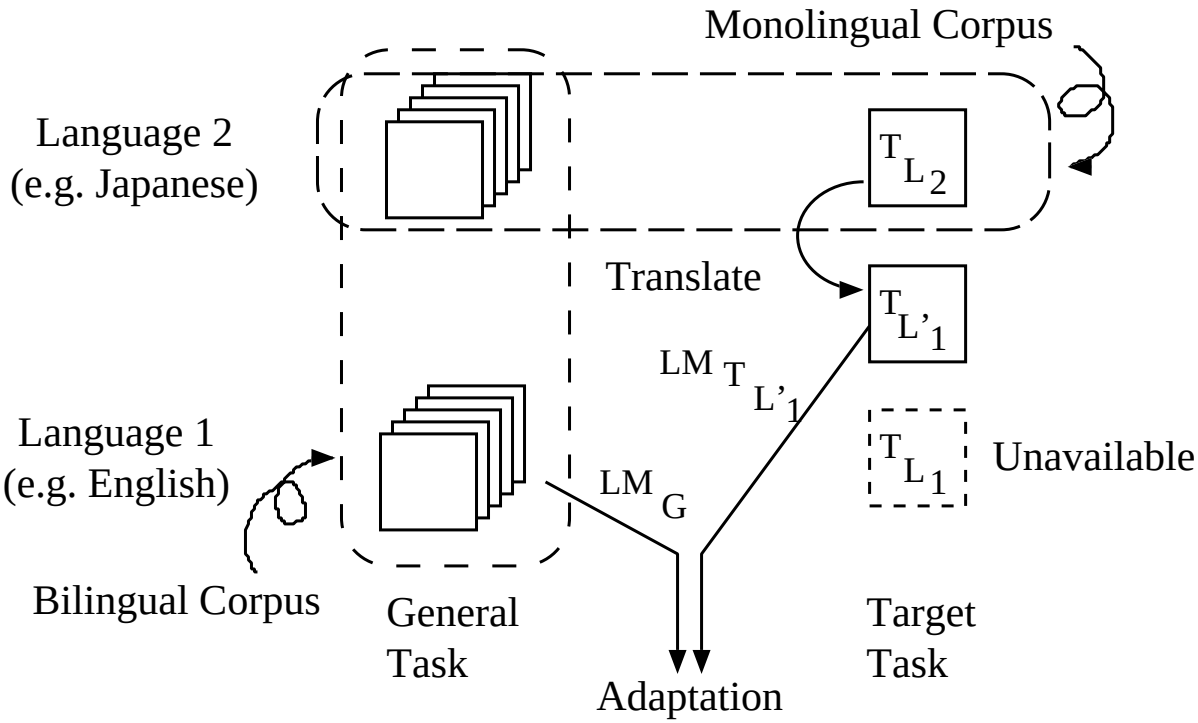

図 2.2: 翻訳結果を使った適応

適応後のモデルの推定手法として，MAP 推定による適応 [37] や，モデルを線形結合する方法 [48] などの様々な手法があるが，本研究ではモデルの線形結合を利用する．つまり，一般タスクの大規模コーパスから作成した言語モデル LM_G と適応先タスクの小規模コーパスから作成した言語モデル $LM_{T_{L'_1}}$ の各確率値を線形結合する．例えば，単語 w_1 と w_2 に引き続いて w_3 が出現する確率 $P(w_3|w_1, w_2)$ の適応後の値を計算する場合，図 2.2 の LM_G での確率を $P_{LM_G}(w_3|w_1, w_2)$ と $LM_{T_{L'_1}}$ での確率を $P_{LM_{T_{L'_1}}}(w_3|w_1, w_2)$ とすると，両者を $0 < \alpha < 1$ なる α を使って，

$(1 - \alpha)P_{LM_G}(w_3|w_1, w_2) + \alpha P_{LM_{T_{L'_1}}}(w_3|w_1, w_2)$

のように線形結合し，適応化された後の確率とする．

本研究における言語モデルの適応の手順をまとめると，以下のフローの通りとなる．

Step 1 翻訳器を用意する．

Step 2 一般タスクの言語モデル（図 2.2 の LM_G）を作成する．

Step 3 Step 1 の翻訳器を使って，一般タスクの言語モデルと同じ言語の適応先タスクのコーパス（図 2.2 の $T_{L'_1}$）を生成する．

Step 4 Step 3 で生成されたコーパスを使い，適応先タスク特定の言語モデル（図 2.2 の $LM_{T_{L'_1}}$）を作成する．

Step 5 Step 2 の言語モデル LM_G と Step 4 の言語モデル $LM_{T_{L'_1}}$ の各確率値を線形結合し，適応先タスクに適応化した言語モデルを作成する．

2.3　機械翻訳器

本提案手法には，様々な翻訳器の利用が考えられるが，評価実験ではオープンなテストセットによる評価実験を実現しやすいコーパスベースの翻訳器のひとつである統計的機械翻訳器を利用した．

本研究の評価実験で用いられる翻訳器の統計的なモデルは Brown らの IBM Model 4[7] に基づいている．このモデルでは，翻訳単位として句への考慮が含まれていることから，位置の対応のみを考慮した IBM Model 3 よりも高い翻訳精度が期待でき，さらに，モデルのパラメータの総量が少ないためモデルの推定が IBM Model 5 よりも速く精確であることが期待できるので，実験に向いていると考え，IBM Model 4 を選択した．翻訳モデルのパラメータの推定には，Och らの GIZA++[43, 44] を用い，そして，翻訳結果の探索には，Tillmann らのビームサーチ [64] と同様の方法を用いた．翻訳器のモデルのパラメータは全て学習用のバイリンガルコーパス（図 2.2 の General Task 側の Bilingual Corpus）から推定される．

ここで，統計的翻訳について，本論文に必要な事項についての概要説明を行なう．統計的翻訳では，翻訳の問題を雑音のある通信路での復号問題と見て，モデリングを行う．例えば，日本語（J）から英語（E）への翻訳を考えると，日本語の文から英語の文への翻訳で最も尤もらしい翻訳結果（ここではこれを E^* とする）を得るという問題は

$E^* = argmax\ P(E|J)$

と表わされる．通常，ベイズの公式で変形し，

$E^* = argmax\ \{P(J|E)P(E)/P(J)\}$

とし，日本語のある 1 文を入力に定めた場合には分母が定数項となるので，分子のみで翻訳結果 E^* の決定が行われる．ここで，分子の $P(J|E)$ は「翻訳モデル」，$P(E)$ は「言語モデル」と呼ばれる．言語モデルには隣接単語の N-gram がしばしば用いられる．

ここの言語モデル $P(E)$ が図 2.2 の LM_G に相当する．この言語モデルを新たなタスク (図 2.2 の Target Task) に合った言語モデルに適応させることが本研究の狙いである．

翻訳対象が学習用コーパスの外の新たなタスクの場合（T_{L_2}）でも，モデルのパラメータが正しく推定されていれば，適応前の言語モデル（$P(E)$）の制約が働いて，言語としては比較的正しく翻訳されると期待される．そして，出力された翻訳結果に対しては $P(E)$ と同時に入力文 J による制約が掛っていると考えられ，それゆえに，この翻訳結果を言語モデル適応用コーパスとして利用できることが期待できる．

表 2.1: 文の分類カテゴリ

基本	空港	飛行機	ホームステイ
両替	宿泊	研究	レストラン
軽食	飲物	移動	ビジネス
買物	観光	美容	トラブル
連絡	帰国	留学	コミュニケーション

表 2.2: 一般タスク，適応先タスク，及び，評価コーパスのサイズ（英語で計算）

コーパス名	文数	単語数
G	152,857	1,197,691
a1000	1,000	7,269
a2000	2,000	15,415
a4739	4,739	36,737
$test_A$	4,739	36,191
b1000	1,000	7,894
$test_B$	3,720	28,974

2.4　評価実験

以下，本手法の有効性を，言語モデルの単語予測性能を示すオープンなテストセット文での単語パープレキシティーによって確認する．

2.4.1　一般タスクおよび適応先タスクのデータ

本実験では日英の対訳コーパスを利用する．文の内容は旅行時の会話表現である．実験ではおよそ 16 万文を用いる．

このコーパスの各表現（各文）は，あらかじめ人手によって,「空港」,「飛行機内」,「レストラン」などの，場面を主とした複数のカテゴリに分類されている．分類カテゴリを**表 2.1** に示す．この表 2.1 の「空港」と「ビジネス」場面での会話表現を評価実験の適応先タスクに設定し，残りを一般タスクとして設定する．これらの内訳は**表 2.2** の通りである．表 2.2 の 「G」 は一般タスクを意味する．適応先タスクのデータの規模と適応の効果との関係を調べるために，サイズの異なる適応先タスクコーパスを,「空港」のカテゴリに対して 3 通り用意した．それらは，表 2.2 の “a1000”， “a2000”， “a4739” である．ま

た，タスクの違いによる適応の効果の違いを調べるために,「ビジネス」のカテゴリから “b1000” を用意した．

これらとは別に，適応の評価用コーパスとして,「空港」タスクに “$test_A$”,「ビジネス」タスクに “$test_B$” を用意した．なお，文と単語の総数は英語で数えた値である．

2.4.2　実験の手順

本実験では,

- 日本語から英語への翻訳（日英翻訳）を行なって得られた英語文を利用して英語の言語モデルの適応を行なう場合,
- 英語から日本語への翻訳（英日翻訳）を行なって得られた日本語文を利用して日本語の言語モデルの適応を行なう場合

の 2 つの場合において，適応の前後でのパープレキシティーを比較し評価する．言語モデルには単語トライグラムを利用した．

適応化された言語モデルは 2.2 節の最後にまとめた手順に従って作成する．すなわち，まず，一般タスクのバイリンガルコーパス（図 2.2 の general task の両言語コーパス，表 2.2 の “G” の両言語）だけを使って，統計的機械翻訳器（翻訳モデルと言語モデル）を作成する．この実験では，この言語モデルは一般タスクの翻訳先言語側のコーパスを使って作成される．これが LM_G に相当する．そして，適応先タスクのコーパス（本実験では，表 2.2 の a1000，a2000，a4739，b1000 のそれぞれの翻訳元の言語側の文）を統計的翻訳器に入力し，翻訳結果を得る．本実験では，探索された翻訳結果の第 1 位候補のみを，翻訳先言語で書かれた適応先タスクのコーパスとして利用し，その生成されたコーパスだけから作られた言語モデル LM_T を作る．このように本提案手法では人手で翻訳されたコーパスのほうは利用しない．最後に，2 つの言語モデル LM_G と LM_T とを線形結合することにより，適応先タスクに適応化された言語モデルを作成する．評価結果は，評価用コーパスに対して最適となる線形結合の重みの場合に得られるパープレキシティー値を示し議論する．

比較のために，人手によって翻訳された適応先タスクの各言語の小規模コーパス（対訳コーパスの訳文）を使った言語モデルの適応によって達成されるパープレキシティーも調査する．

本実験では，一般タスクのコーパスのみで作られた辞書の未知語率は，テストセット $test_A$ では 0.79% であり，$test_B$ では 0.0% であった．人手で翻訳された適応用コーパス a1000，a2000，a4739 を追加した場合も，$test_A$ での未知語率は，それぞれ 0.77%，0.76%，0.75% と

表 2.3: 英語のトライグラムの数

空港	人手翻訳		機械翻訳		R[%]
G	259,171	-	-	-	-
G+a1000	260,235	(2,995)	259,997	(2,797)	22.9
G+a2000	261,657	(6,000)	261,077	(5,322)	20.9
G+a4739	264,809	(12,337)	263,275	(10,224)	20.7
ビジネス	人手翻訳		機械翻訳		R[%]
G	259,171	-	-	-	-
G+b1000	261,160	(5,256)	260,622	(4,427)	21.1

表 2.4: 日本語のトライグラムの数

空港	人手翻訳		機械翻訳		R[%]
G	280,500	-	-	-	-
G+a1000	281,481	(3,089)	281,314	(2,877)	23.5
G+a2000	282,983	(6,366)	282,278	(5,371)	20.5
G+a4739	286,184	(13,030)	284,532	(10,468)	20.6
ビジネス	人手翻訳		機械翻訳		R[%]
G	280,500	-	-	-	-
G+b1000	282,563	(5,698)	281,999	(4,731)	19.9

なり，大きな変化はなかった．本実験では言語モデルの適応の効果を調べるので，語彙サイズの変化に伴う未知語への確率の配分の影響を排除するため，常に言語モデルの適応前の辞書を用いて語彙サイズを一定とした．実験に用いた辞書の語彙サイズは，日本語が 21,854 語，英語が 15,056 語であった．

適応の前後の言語モデルから得られるトライグラムの数を**表 2.3** と**表 2.4** に示す．“G” の行は人手で作成された一般タスクのコーパスからなる言語モデルのトライグラムの総数，“G+a1000” などの行は a1000 などに対して人手（または機械）によって翻訳された適応用コーパスと一般タスクのコーパスの両方を使って適応化された言語モデルのトライグラムの総数，同じ行の括弧内は適応用コーパスだけから得られるトライグラムの総数である．「人手翻訳」の列は人手翻訳結果つまり対訳コーパスの文を，「機械翻訳」の列は機械翻訳結果を，それぞれ適応用コーパスに使った場合の結果である． また，R は再現率で，人手翻訳から得られるトライグラムのうちの何%を機械翻訳結果から得られるかを示す値である．

表 2.5: コーパスサイズとパープレキシティ（英語の言語モデルの適応）

空港	PP_1		結合重み	削減率 [%]	PP_2		結合重み	削減率 [%]
G	32.0	-	-	base	32.0	-	-	base
G+a1000	23.5	(50.8)	0.47	26.7	27.8	(84.1)	0.26	13.1
G+a2000	21.8	(37.0)	0.61	31.9	27.9	(72.3)	0.31	12.8
G+a4739	19.8	(27.1)	0.79	38.1	27.9	(61.9)	0.39	12.8
ビジネス	PP_1		結合重み	削減率 [%]	PP_2		結合重み	削減率 [%]
G	55.2	-	-	base	55.2	-	-	base
G+b1000	50.0	(155.9)	0.04	9.42	53.2	(196.3)	0.01	3.62

機械翻訳の性能は，音声認識と同じ単語誤り率でしばしば測られる．単語誤り率は次式で定義される．

$WER[\%] = 100.0 \times (Sub + Ins + Del)/T$

ここで，T は翻訳結果の正解文の中の全単語数，Sub，Ins，Del は，それぞれ，置換誤り数，挿入誤り数，削除誤り数である．以上の実験条件下では,「空港」カテゴリでの単語誤り率は約 80%,「ビジネス」カテゴリでは約 74%であった．この計算は，対訳コーパスでの機械的照合の結果，翻訳器への入力側の言語の同一の文に対して，出力側の言語の文が複数存在する場合には，それらをどちらも正解と設定した状況で測った．例えば，日英翻訳において，対訳コーパス内の同じ日本語文 J に対して，異なる英語文 E_1 と E_2 がある場合には，J の翻訳結果として E_1 と E_2 の両方を正解とした．そして単語誤り率はそれぞれに対して計算し小さい方を採った．また，翻訳結果が大量であるため人間による主観評価は行なわなかったが，翻訳先の言語としては局所的な文脈は概ね正しい結果が比較的多く見られた．

2.4.3　結果

日英翻訳を行ない英語の言語モデルの適応を行なった場合

日英翻訳結果を使って英語の言語モデルの適応を行なった場合の評価コーパスに対するパープレキシティーを**表 2.5** に示す．表 2.5 の ‘G’ の行は一般タスクのコーパスだけで作られた適応前の言語モデルでのパープレキシティーを意味する．“ +a1000” や “+b1000” などで追加される文の数を示す．PP_1 の列は理想的な状況として人手で作成された対訳コーパスの訳（英文）を使って適応した場合のパープレキシティーを，PP_2 の列は本手法で統計的機械翻訳によって生成されたコーパスを使って適応をおこなった場合のパープ

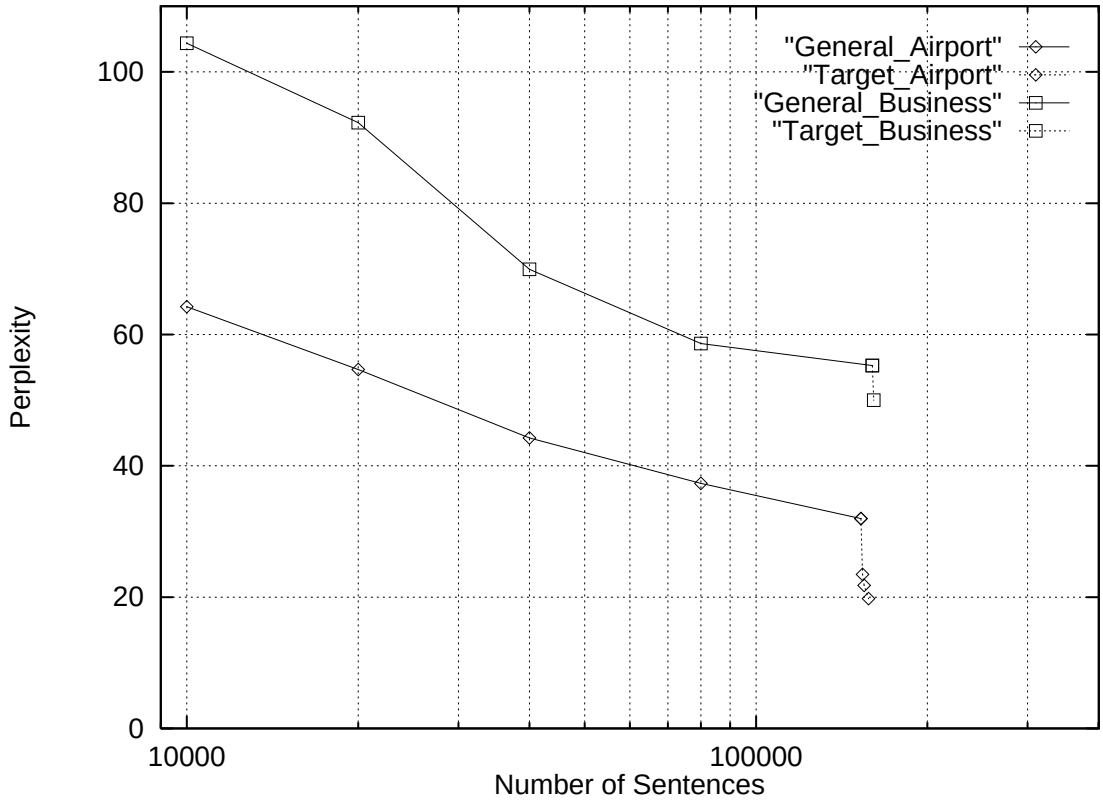

図 2.3: コーパスサイズとテストセットパープレキシティ

レキシティーである．それぞれの括弧の中の数値は，適応先タスクの小規模コーパスのみで作られた言語モデルによるパープレキシティーである．削減率は適応前と比べた適応後のパープレキシティーの削減率である．また各適応における線形結合の重み α の値も示した．

同様に，**図 2.3** の実線は一般タスクのコーパスの規模とパープレキシティー値との関係を示すグラフである．右側で急に下降する破線は，一般タスクのコーパスに適応用コーパスを追加した場合のコーパスの総量と，それらを使って適応を行なった後のパープレキシティー値 (表 2.5 の PP_1 の列のパープレキシティー) との関係を示すグラフである．ここでの追加コーパスは人手で翻訳された文だけからなる．上側の線が「ビジネス」タスクの場合を，下側の線が「空港」タスクの場合を示す．

表 2.5 の PP_1 の変化のように，一般タスクのコーパスだけから作られた言語モデルでのパープレキシティー（空港タスクでの 32.0，ビジネスタスクでの 55.2）に比べて，一般タスクのコーパスと適応先タスクのコーパスとを使って適応化された言語モデルでのパープレキシティー（空港タスクでの，a1000，a2000，a4739 のそれぞれで適応した場合の 23.5，21.8，19.8，ビジネスタスクでの 50.0）のほうが，値が大幅に（削減率で 9.42%（ビジネスタスク），および，38%（空港タスク））小さくなっている．これらの結果は，仮に人間の

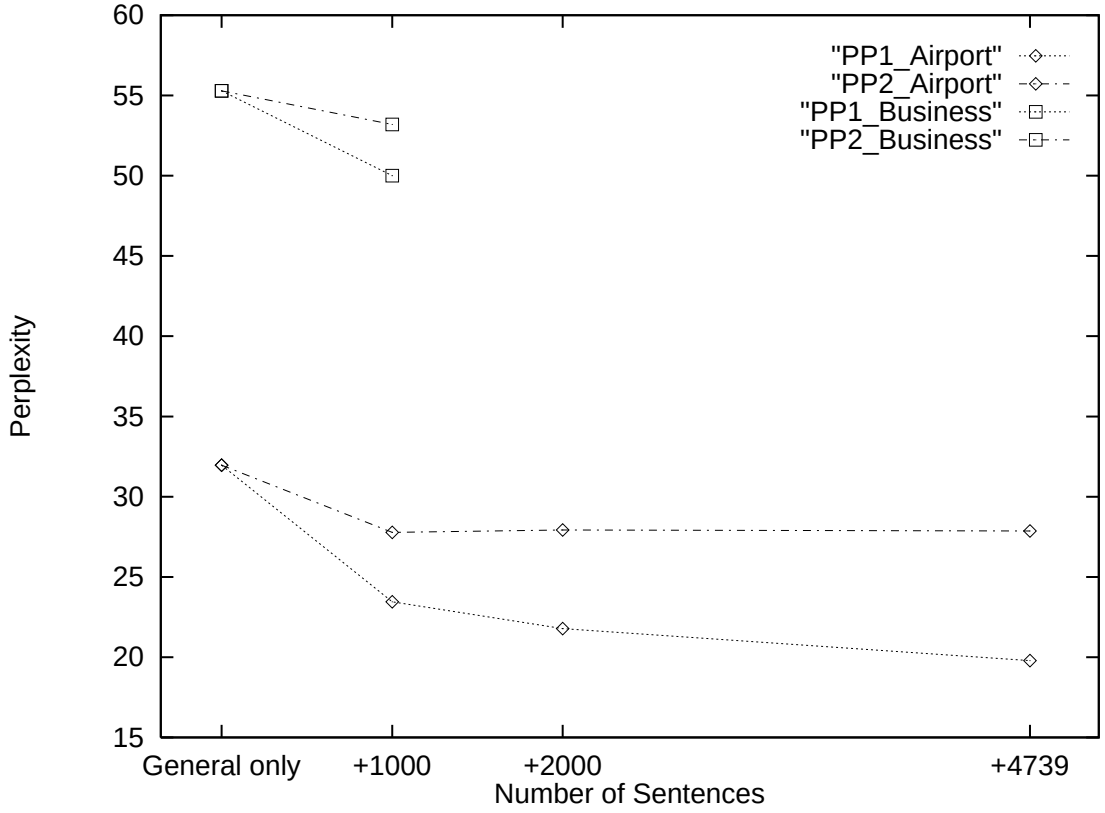

図 2.4: 本提案手法で生成された擬似コーパスを用いる場合と従来の人手で集められたコーパスを用いる場合のパープレキシティの比較

ような翻訳結果を出力する完全な機械翻訳器を使った場合に得られるパープレキシティー削減の限界を示している．すなわち，適応後のパープレキシティーの削減率は,「空港」タスク，および「ビジネス」タスクにおいて，それぞれ相対量にして高々38%，9.42%である．

また，図 2.3 から，一般タスクのコーパスをさらに数倍の規模だけ集める (実線の右側への延長) よりも適応先タスクのコーパスへモデルを適応させたほう（破線）が PP 値の削減が大きいと予想される．故に，本データにおいては言語モデルのタスク適応が有効であることがわかる．さらに，何らかの方法によって適応先タスクのコーパスを準備すべきであることが分かる．

この適応先タスクのコーパスを，本研究では統計的翻訳器を使って生成した．そしてそれを使って適応を行った結果が表 2.5 の PP_2 から右の列である．また，追加されるコーパスのサイズとパープレキシティーとの関係を**図 2.4** に示す．

機械翻訳によって得られた適応用のコーパスを使った言語モデル適応によって,「空港」タスクでは適応前に比べて相対値にして最大で 13.1%のパープレキシティーの削減が得られた.「ビジネス」タスクでは 3.62%であった．しかし，図 2.4 の「空港」タスクでの結果（$PP2\_Airport$ の線）のように，機械翻訳によって生成されたコーパスを使った適応では，

表 2.6: コーパスサイズとパープレキシティ（日本語の言語モデルの適応）

空港	PP_1		結合重み	削減率 [%]	PP_2		結合重み	削減率 [%]
G	23.0	-	-	base	23.0	-	-	base
G+a1000	17.7	(39.1)	0.46	22.8	20.0	(64.3)	0.27	12.8
G+a2000	16.6	(29.1)	0.59	27.6	20.1	(54.0)	0.32	12.6
G+a4739	15.4	(21.7)	0.77	32.9	20.2	(48.6)	0.38	11.9
ビジネス	PP_1		結合重み	削減率 [%]	PP_2		結合重み	削減率 [%]
G	43.5	-	-	base	43.5	-	-	base
G+b1000	39.5	(137.1)	0.05	9.25	41.3	(170.3)	0.02	4.9

1000 文追加以後は性能変化が横ばい状態になっている．この原因は後の 2.4.4 で議論する．

ところが，今回のような高くない機械翻訳精度の状況でさえ，本手法により得られたパープレキシティーの削減率は，人手による翻訳結果を使うという理想の場合のパープレキシティーの削減率のおよそ 30 から 50%を得ており，効果が大きい．

英日翻訳を行ない日本語の言語モデルの適応を行なった場合

前節と同じ設定で同様のデータを使って，日英翻訳とは逆向きの英日翻訳の結果を使って日本語の言語モデルの適応を行った場合の結果を**表 2.6** に示す．表 2.6 の書式は表 2.5 と同様である．

この場合にも，機械翻訳で得られた適応先タスクのコーパスを用いることにより，適応前に比べて,「空港」タスクでは最大で 12.8%,「ビジネス」タスクでは 4.9%のパープレキシティー削減率が達成されている．これは人手で翻訳されたコーパスを用いた理想的な場合のおよそ 56%にあたり，大きな効果が得られている．性能改善が 1000 文追加以後横ばいになっている点は前節と同様である．

以上の表 2.5 や表 2.6 のように，翻訳方向に関係なく，本提案手法の効果が確認された．

2.4.4　議論

本研究では「言語モデルと同じ言語の適応先タスクのデータがない場合に，適応先タスクのデータとして機械翻訳の結果を使えるかどうか」について研究を行っている．これまでに示した結果から，現状の翻訳性能にも関わらず

1. パープレキシティーの削減量の絶対値はタスクによって異なるが,

表 2.7: データ量と被覆率の変化（空港タスクで計算）

空港	人手翻訳での被覆率 [%]	機械翻訳での被覆率 [%]
G	54.3	54.3
1000	56.1	54.5
2000	58.2	54.8
4739	60.4	55.2

2. 機械翻訳によって生成されたコーパスを適応に用いることにより，人手で翻訳されたコーパスを使った適応のおよそ 50%前後のパープレキシティーの削減効果が得られ，その結果，

3. 本手法が適応先タスクのコーパスの生成に有効であること

の 3 点が明らかとなった．

上の２つ目の項目のように，適応先のタスクのコーパスが存在するという理想的な場合に比べると，機械翻訳結果を使った適応の効果がおよそ 50%前後に留まっている．この原因の１つとして，高くない翻訳精度の他に，尤度によって順序付けられた翻訳結果の第 1 位候補だけを利用していることが考えられる．2 位以下の文も確率的には 1 位よりも低いが対訳になる（すなわち，コーパスを集めたらその中に出現する）可能性もあるが，本実験では無視してしまっている．例えば，統計的翻訳器の学習用対訳コーパス内に，日本語の単語列 J に対応する異なる 2 つの英語の単語列 E_1 と E_2 がある場合に，英語の両単語列へのゼロではない確率が学習されていたとしても，機械翻訳の第 1 位候補だけを利用するので，高い確率を持つ一方の単語列（例えば E_1）だけが翻訳結果に現れやすくなり，他方は現れにくくなる．実際に表 2.3 や表 2.4 に示したように機械翻訳結果から得られるトライグラムは人手による翻訳で得られるトライグラムのおよそ 20%しか再現できていなかった．このため，出現しなかったトライグラムの確率値が小さくなり，パープレキシティーの改善が小さくなったと予想される．

また，機械翻訳結果の文数が増加してもパープレキシティーの変化は横ばいであった．この原因は上記の第 1 位の翻訳結果のみを用いていることの他に，適応によって得られるトライグラムのテストセットに現れるトライグラムに対する被覆率の増加が小さいためと考えられる．そこで，この被覆率を計算した（**表 2.7**）．人手翻訳を使った場合の被覆率は適応用コーパスの文数が 1000，2000，4739 と増えるにつれて，およそ 2 ポイントずつ大きくなるが，機械翻訳を使った場合のテストセット被覆率はおよそ 0.3 ポイント程度ずつ大きくなるのみで，ほとんど変化がなかった．

今後，翻訳結果の N-Best，ラティス，および，原理の異なる複数の翻訳器からの出力結果を使うなどの方法を取り，それらの中から多様な表現が取り出されることにより，性能

が改善されると予想される．これらは今後の課題として挙げられる．しかし，人手でコーパスを作成した場合と同じ効果を得るには，同じ意味を持つ適切な複数の結果を出力したり，未知語に対して十分な対策を備えるなどのさらなる翻訳器の改善が必須であるが，これは今後の統計翻訳の課題と考える．

本研究での統計的翻訳器のモデル学習には，適応先タスク以外の大規模コーパスを用いた．適応先タスクの表現に類似した表現がこの学習用の大規模コーパスに含まれている可能性があるが，その中のどの表現が適応先タスクの表現と類似するかは未知の設定である．すなわち本研究では，適応に有効な表現を機械翻訳によって取出し利用することで，言語モデルの適応化を達成した．

実際の言語モデルの利用場面では，例えば，レストランでのレストラン従業員と客との会話がレストランでの注文や支払い等の話題に集中したり，あるいは，ニュース解説ではニュース毎に話題が一貫しているように，話題が一貫して現れることが多い．しかし，しばしばその他のタスクの話題が混ざる恐れもある．これに備えるために，タスクの事前認定を行なった後で認定されたタスクに適応化された言語モデルを利用したり，タスクに適応化された言語モデルの混合 [24] が行なわれる．本研究では，事前のタスクの認定後に使われる言語モデルや混合の要素となる言語モデルに相当する，タスク適応された言語モデルの作成に焦点を置いた．特に，そのような適応化された言語モデルの作成に必要な適応先タスクの小規模コーパスを機械翻訳によって生成する方法について議論した．事前認定や混合等は重要ではあるが，別の課題と考える．

2.5　関連研究

本来，統計的言語モデルの作成には，そのモデルを利用するタスクでの大規模コーパスが必要であるが，その作成がしばしば困難である．それに代わる方法として，従来の研究では，限られたサイズの小規模コーパスを利用する [48] か，World Wide Web (WWW) から関連する文書を収集し利用する [5] ことによって，既存のモデルの適応先タスクへの適応を行なった．これらの多くは，ディクテーションタスクの性能改善を目的としており，多く存在しうる書き言葉のコーパスを使った研究であった．一方で，話し言葉のコーパスはそもそも少なく，新しいタスクの小規模コーパスの作成すら困難な場合がある．

また，コーパスの作成が同様に困難な，医療所見のディクテーションや機械への入力インタフェースの分野では，システム開発者がそのタスクで話される言語を文脈自由文法（CFG）として記述し，それを使って人工的な言語データを生成し，適応に利用した [22, 69]．ところが，話し言葉の文法を十分に記述することも難しい．

これまでのところ，我々の提案方法のような，第 1 の言語で書かれたコーパスを第 2 の言語に機械翻訳し，その翻訳結果を使って第 2 の言語の言語モデルの適応を行なうような

研究は例が無いようである．

2.6　2 章のおわりに

本論文では，ある言語での言語モデルの新しいタスクへの適応をその言語での適応先タスクのコーパスが存在しない状況下でも可能とするために，その適応先タスクの別の言語で書かれたコーパスを機械翻訳し，生成された適応先タスクのコーパスを使って，適応を達成する手法を提案した．

旅行用会話文を対象としたテストセットの単語パープレキシティーを評価尺度とする実験において，適応前にくらべて，大幅なパープレキシティー削減が確認された．適応化を行ないたい言語の適応用のコーパスがない場合には，翻訳結果であっても利用したほうが，言語モデルの改善が得られることが確認された．また，この削減量は人手によるコーパスが存在する理想的な状況のおよそ 50%前後の削減量にあたり，これらの結果から，本手法による言語モデル適応の有効性が確認された．また，このときの翻訳器の性能は単語誤り率で 80%と芳しくない状況であったが，このように訳文全体としての翻訳性能が低くても，言語モデルのタスク適応のためのデータ生成には十分に利用可能であることが確認された．

第3章　未登録単語のアクセント型推定

3章では，音声翻訳器を適用する際に生じる未知情報の推定課題の１つである，音声翻訳器にとって未知の単語の（音声翻訳器に未登録の単語の）アクセント型を予測する課題を解決する方式を提案する．

3.1　はじめに

音声合成においてアクセントは意味を正しく伝えるために重要である．そのため，現在のテキストからの音声合成（TTS）では，テキストを単語に分割し，辞書を参照して，単語に読みだけでなくアクセント型も付与する．この単語毎のアクセント型を利用して，アクセント結合規則 [51] 等でアクセント句全体での理想的な音の並びを決め，それに基づいて合成音を作成する．

新聞記事や WEB ページの文書等の任意のテキストを TTS の対象とする場合は，辞書に載っていない単語，即ち未登録語（未知語）の発生が避けられない．合成音を作るためには，それらの未登録語の読みとアクセント型の推定が必要となる．読みの推定は書記素（字面）と音素（読み）間の変換問題としてこれまでに様々な研究がある [14, 35]．一方，アクセント型推定の研究は少なく，方式を確立することが重要である．

アクセント型推定の対象は未登録語になりやすい固有名詞が重要と考えられる．それらには数が膨大で TTS 等の解析用辞書に登録し尽くせない姓名，企業名（組織名）がある．企業名はその構成パターンの多くが創業者名やグループ名等の固有名詞を要素とする複合語であると報告されている [23]．そこで，利用場面が多いと考えられる姓名を対象にして推定法を検証する．

本論文ではアクセント推定を以下のように識別問題と考える．単語のアクセントは “particular prominence attached to one syllable of a word or phrase by some phonetic means such as stress or pitch” と記されている [65]．日本語（東京方言あるいは標準語）のアクセントでは，syllable の代わりにモーラを単位として示されることが多い．例えば（「ハナ」のような）2 モーラの単語では，各モーラのトーンの高低を ‘H’ と ‘L’ で表すとすれば，トーンの変化が ‘LH（H）’ であるものが 0 型（平板型），‘HL’ が 1 型（頭高型），‘LH（L）’ が 2 型（尾高型）と呼ばれる 3（=2+1）通りのアクセント型の可能性がある．0 型以外の型の番号の数字は，トーンが ‘H’ から ‘L’ に変化する ‘H’ の位置を単語の先頭から

数えた数値に対応している．一般には，単語が N モーラから構成されていれば，アクセント型の可能性は N+1 通りある．そのため，アクセント型推定は N+1 個のクラス間での識別問題とみなせる．

本論文ではアクセント推定に用いる情報として読みとその並びに着目し，その推定性能を明らかにすることに焦点を当てる．読みだけからの推定が可能であれば，推定法をよりシンプルに構成できるからである．応用の面では，漢字を想定しないことによって，本研究の推定方法は TTS だけでなく音声翻訳の合成部分への適用が可能となる．日英のような言語間ではアクセントの表現方法も位置も異なり，言語独特のアクセント情報を割り当てる必要が有る．しかし過去の研究では，姓名は音訳されて翻訳先言語での読みになるが，アクセント型は送られてこなかった [45].

本論文では，上記のように，読みとその並びを入力として統計的な識別問題として未登録語のアクセント推定をおこなう．安定した識別のためには裏付けの有る機械学習が望まれる．そこで本研究では，統計的な機械学習に基づく識別器の 1 つとして，近年様々な分野の識別問題において高い性能が報告されている Support Vector Machine (SVM) を用いることにした．以下，**3.2** では，アクセント推定に統計的アプローチを用いる理由と妥当性を示す．そして提案法の詳細を **3.3** で述べる．**3.4** で評価実験の設定と結果を，**3.5** で関連研究との比較を行い，SVM を使った未登録語のアクセント推定法の有効性を示す．

3.2　統計的アプローチの利用

本研究では未登録語のアクセント型推定に統計的な手法を用いる．なぜなら，未登録語になりやすい固有名詞は時代とともに変化する性質を持つため，新たな未登録語に追随する際に規則の更新が必要となるが，その更新にかかる負担を軽くしたいためである．例えば，保険会社による調査結果[1]から名（forename）には次のような特徴が見受けられる．

- 男性の場合，明治から大正では「正」や「〜郎」を使った名前が，またその後昭和 40 年頃までは漢字 1 字の名前が多く，その後から今日にかけては「大」や「洋」などの広がりを感じさせる名前が多い．
- 女性の場合，大正初期から昭和 30 年頃まで「〜子」が多いが，以後は「〜子」は少ない．

また，最近では ‘Julia’ に似た「樹理亜」のように，外国においても認知されやすい名前が日本人の名前として用いられてきている．このように名の分布は変化する特徴が有る．

これまでに土田らは，名の末尾のモーラパターンとアクセント型との対応関係を規則化した [66]．例えば，3 モーラの人名の場合の判定規則は，

[1] http://www.meijiyasuda.co.jp/profile/etc/ranking/

1 型　語尾が，キ，コ，ゴ，シ，ジ，タ，ト，ヤ，イチ，ヘイ，ロー の場合，
0 型　語尾がその他の場合

と記されている．上記の規則は 1983 年以前に作成されたと考えられ，この規則では比較的新しいと予想されるアクセント型が 1 型の「サヤカ」や「アヤカ」などの名前のアクセント型の推定で失敗する．

姓の分布は名のように変化しないようであるが，未知の姓は少なからず出現する．未知の姓のアクセント型を推定できるように規則を変更する作業は名と同様に必要となる．人手で作成された規則はその適切な適用のために細かな規則の集合になっている場合が多く，更新には規則の細部にまで注意の及ぶ専門家が常に必要となる．

一方，規則をデータから獲得する統計的な機械学習に基づく方法では，新たに出現した固有名詞とそのアクセントとを学習データに追加すれば良いので，専門家は必ずしも必要ではない．機械学習を使って，少ないコストで，かつ，高いアクセント型推定精度を達成することができるなら，そちらの方が望ましい．

また，文献 [16, 17] の指摘のように，人手で作られた姓名の規則にも「〜型が多い」と記されており曖昧性がある．例えば,「転成語の 3 拍語は多く頭高型．但し『林』等は平板型」[28] とあるが，どのような規則に適合する語が「林」と同じように例外となるかも不明でもある．さらに，上記の転成語かどうかの判定のように，様々な条件判定のルーチンを要するので，推定法が複雑になる．このように人手で作られた規則をそのまま推定法に利用することは難しい．

機械学習を用いるには，データ間に矛盾の無いことが望ましい．即ち，合成を目的とする本研究の場合，同じ読みを持つ姓名が異なるアクセントを持つ場合が無いか，少ないことが望ましい．そこで，以下のようにアクセントの曖昧性の調査を行った．

東京アクセントが付与されている約 6 万 5 千語の姓名を用意した (**表 3.1**)．観察によれば，姓の「名波 (ナナミ)」が ‘HLL’ のトーンパターンでアクセント型は 1 型，名の「奈々美 (ナナミ)」が ‘LHH’ のトーンパターンでアクセント型が 0 型であるように，同じ読みでも姓と名というクラスが異なればアクセント型が異なる場合が見られた．そのため以下では姓と名とに分けて分析や実験を行う[2]．

姓名のそれぞれの中で，読みとアクセントの組（発音）が同じものを 1 つとして数えると，**表 3.1** の「発音形の数」の列のように大幅に数が減る．(例えば,「コウイチ」の読みを持つ「耕一」と「幸一」を 1 つとして数えると，名の数は 7,995 に減る．) さらに，読みが同じでアクセントが異なるものの割合を「アクセント曖昧率」として**表 3.2** に示した．

[2] このような前提を置くと，アクセント推定前に姓と名への分類が必要となるが，この分類は固有表現抽出と同様の技術である程度実現できると考えられるし，音声認識においては姓と名とは結果として分類した形で得られている [45]．そのためあまり問題にならないと考えられる．

表 3.1: 姓と名のデータ数

	数え方	
	出現形の数	発音形の数
姓	28,337	19,745
名	37,177	7,995

表 3.2: アクセント曖昧率

	アクセント	
	単一	曖昧　(曖昧率 [%])
姓	19,437	308　(1.56)
名	7,931	64　(0.80)

名では 7,995 個のうち 0.8%の 64 個に複数のアクセントがあった（曖昧であった）.「家康」での 0 型（LHHH）／ 2 型（LHLL）がその例である．これらには，音声合成という目的から 1 つのアクセント型に集約し一貫性を持たせることも考えられる.

このように，実際の姓名においてアクセント型に多少の曖昧性が存在するが数が少ないこと，また，合成の目的上 1 つのアクセント型に統一しても問題がほとんどないと考えられることから，アクセント型の推定問題は識別問題として解くことが可能のようである.

3.3　アクセント型推定手法

本稿では，以下に説明する，CRA，CCA，CCA+の 3 つのアクセント型の推定方法を提案し比較する．全て単語を構成するモーラ単位の読みとその並びの情報をアクセント型の推定に利用する．以下の説明の中の識別器には SVM を用いる．3 つの方法の概略を**図 3.1** に示す．この図では，全て，姓の読み {/ko/，/shi/，/ba/} からそのアクセント型を推定する様子を示している.

3.3.1　CRA

第 1 の方法では，最初に単語の各モーラのトーンの高低を単語の末尾まで推定し，得られたトーンの高低変化の中で高から低に変化する箇所をアクセントの位置とし，無ければ 0 型と推定する．このように前後の音の並びとトーンの高低との間のマッピング関数

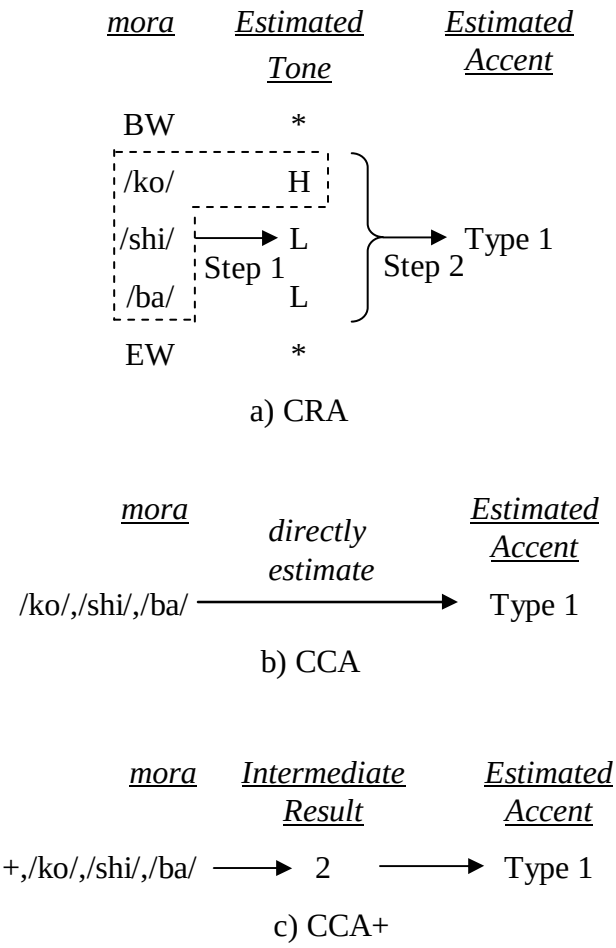

図 3.1: アクセント型推定への提案手法の概要図

を識別器が学習するため，本論文では「回帰に基づく識別アプローチ（Classification by Regression Approach: CRA）」と呼ぶ.

図 3.1 の a) では，Step 1 が/shi/のトーンの高低を推定する様子を示している．トーンは単語の前方または後方から順に推定できるが，この例では単語の前方から推定する様子を示している．ここでは，点線で囲まれた情報，即ち，/shi/の前後の 1 モーラの読み（/ko/と/ba/）と推定済の/ko/のトーン（‘H’）から，識別器を使って，/shi/のトーンを推定する様子を示している．全ての音のトーンの推定が終った後に，Step 2 で，全てのトーン ‘HLL’ から，トーンが高から低へ変化する部分を探索し，その位置からアクセント型を推定する．この例では，１モーラ目でトーンが高から低へ変化しており，アクセント型が１型（Type 1）と推定される過程を示している.

3.3.2　CCA

第 1 の方法では “単語全体の中でアクセントが高々 1 つである” という日本語の単語のアクセントが持つ性質を表現できていない．そこで，第 2 の方法では「単語の読みを構成するどのモーラにアクセントが有るか」を直接識別する．これは単語が何型のアクセントを持つ単語群に分類されるかという問題と同じであるため,「分類に基づく識別アプローチ（Classification by Classification Approach: CCA）」と呼ぶ.

ここでは識別器を単語の長さ毎に構成する．この識別器への入出力情報の例を**図 3.1** の b) に示す．単語を構成するモーラがどの位置に（先頭から何モーラ目に）どのような音として配置されているかをベクトル表現し，識別器の入力とする．出力は対応するアクセント型である．**図 3.1** の b) では，姓の読み {/ko/，/shi/，/ba/} を入力としてアクセント型，ここでは 1 型（Type 1），を直接推定する様子を示している．

3.3.3　CCA+

第 2 の方法では，単語の長さ毎に識別器を用意した．第 3 の方法では，識別器を，単語の長さによって分けること無く，1 つ構成する．これを前記 CCA の変形版として「CCA+」と呼ぶ.

予め単語の長さの最大値（N_{max}）を定めておき，それより短い長さの単語には先頭側にモーラの読みが無いことを示す記号を付与する．そして，それらから CCA と同じようにして中間結果を推定し，その結果をアクセント型に変換する．中間結果は，単語の末尾から数えたアクセントの位置で表現する．よって，変換では単語の長さから中間表現の数値を引き算する.

例えば，N_{max} を 6，モーラが無いことを示す記号を’+’ すると，姓の読み/ko//shi//ba/ が N_{max} に比べて 3 モーラ分先頭に記号が少ないので，入力は**図 3.1** の c) のように先頭に 3 つ ’+’ の記号を追加した表現となる．元々単語のアクセントの位置は/ko/に有る．これは単語の末尾を起点にして最初を 0 番として数えると，‘2’ の位置に有るので，正しく推定されると中間結果は ‘2’ となる．アクセント型の推定では，識別器がまず**図 3.1** の c) のように中間結果の ‘2’ を推定する．その後，この数値を単語の長さ（‘3’）から減算して，もとの単語の頭から数えたアクセントの位置（‘1’）に変換して，アクセント型の推定結果とする.

3.3.4　SVM を使う各処理の解釈

SVM の解説は文献 [67] 等に詳しく書かれている．ここでは本研究に係わる部分に注目し，上記提案法の処理の意味について説明する．

サポートベクトルは識別境界面付近に存在し，識別にとって重要な学習データの部分集合である．SVM による識別では，入力情報はベクトル表現され，それと全サポートベクトルとの類似度を計算し，類似度の大きなクラスに識別される．類似度の計算には内積またはカーネル関数が用いられる．識別問題が線形分離可能ではない場合には，非線形のカーネル関数を使うことで問題を効果的に解けることが知られている [67]．本研究ではカーネル関数として多項式カーネルを用いる．多項式の次元を高くすると，次元数までの入力情報の組合せが用いられることを意味する．

よって，CRA では，未登録語の当該のモーラのトーンを，そのモーラの読み，および，前後の数モーラの読みと推定済のトーンを入力として，それらとの類似度の高い学習データ（サポートベクトル）と同じトーンになるように識別が行われる．そして最後に，トーンが高から低に変化する所をアクセントの位置と推定する．CCA と CCA+では，未登録語の読みを構成するモーラの読みの列を入力として，それらとの類似度の高いサポートベクトルと同じ型になるよう推定が行われる．

3.4　評価実験

実験を行い，提案法の性能と有効性を明らかにする．以下の点での比較をおこなう．

1. カーネル関数の次元と精度
2. 従来法としての決定木による精度との比較
3. 実際の WEB 文書に現われる未登録語での精度

3.4.1　学習データと評価データ

本研究では，**表 3.2** の単一のアクセント型を持つ姓名のデータのうち，その量が十分に存在する，姓では 2 から 6 モーラ，名では 3 から 6 モーラのデータを利用した．交差実験を行なうため，姓と名において，長さごとにデータ数とアクセント型の分布がほぼ同じとなるようにして，全体を 5 つのグループに分けた．5 つのうちの 1 つを評価データとし残りを学習データとする，5 セットの交差実験用のデータセットを作成した．未登録語に対するアクセント型の推定が目的であるので，評価データと学習データとの間の重複が無いように設定した．このうちの 1 セット分のデータ量を例として**表 3.3** に示す．

表 3.3: 学習および評価データのサイズの一例

	単語長	学習用	評価用
姓	2	611	152
	3	4,655	1,163
	4	7,639	1,909
	5	1,082	270
	6	120	30
名	3	2,135	533
	4	3,094	773
	5	606	151
	6	290	72

表 3.4: WEB 上のニュースに現われた姓名の未登録語の数

	既知発音	未知発音		合計
		対象	対象外	
姓	34	33	0	67
名	113	29	3	145

このデータは，カーネル関数の次元と精度との関係を調べる実験と従来法との比較に利用する．CCA と従来法の決定木にはこのデータを単語の長さ毎に利用し，単語の長さ毎の CCA の SVM と決定木を構成する．CRA と CCA+では全ての長さのデータを学習データ側と評価データ側でそれぞれマージして，それぞれ単一の SVM を構成する．

さらに，実際の WEB 文書に現われた未登録語を評価に用いた．2003 年の 6 月と 7 月にウェブサイトの goo のニュース記事（社会面）に出現した語のうち，前記分析に用いた全データ（**表 3.1** の全データ）に入っていない姓名を未登録語とした．1 名の作業者がアクセント型を付与し，他の 1 名の作業者が検査を行った．抽出された未登録語の分量などを**表 3.4** に示す．得られた単語は姓が 67 個，名が 145 個の合計 212 個であった．このうち，分析に用いたデータに同じ発音の語が存在した姓は 34 個，名は 113 個であった（**表 3.4** の 「既知発音」の列）．また，学習データが十分に無く，今回の実験の対象外とした長さの単語が名の中に 3 個あった（**表 3.4** の 「未知発音」の「対象外」の列）．よって，実験では姓の 33 個と名の 29 個（**表 3.4** の「未知発音」の「対象」の列）に対してアクセント型の推定を行う．それらの内訳は，姓では，2，3，4 モーラがそれぞれ 3，6，22 個，5，6 モーラが各 1 個で，名では，3，4，5 モーラがそれぞれ 13，15，1 個であった．

3.4.2　評価尺度と実験条件

N モーラから成る単語でアクセント型が「N 型」の場合，N モーラ目が ‘H’ で，次に付く接辞等が ‘L’ となる．即ち 0 型と N 型との違いは，各単語の次の接辞等でのトーンが高いままか低くなるかの違いである．単語だけが持つ情報には 0 型か N 型かの区別を行うための情報（例えば接辞の情報）がないので，決定ができない．そのため，本論文では 0 型と N 型は同じ物とみなして，全部で N 個のクラスの間での識別問題と考えて評価をおこなう．

CRA では出力結果のトーンパターンが正解のトーンパターンと完全一致したもののみ正解とする．CCA と CCA+では単語のアクセント型に一致したものを正解とする．精度は全単語中の正解単語の割合とする．

カーネル関数の他に，学習データでの誤りを許して汎化誤差を下げることを目的とするソフトマージンの考え方を使うことによって，線形非分離な識別問題を効果的に解けることが知られている [67]．本研究のアクセント型推定も線形非分離な問題と考えられるので，ソフトマージンを利用する．このための制御変数の値によっては評価データに対する誤り率が最小となる点が見つかるが，その値を単調に増加させても，単調に評価データでの誤り率が減少するとは限らない．また，単語の長さやデータ数毎に最適な値が異なる．そのため，本研究ではこの値を 1 に固定して実験を行い，カーネル関数の次元と精度との関係に焦点を当てることにした．

SVM は 2 値分類器であるので多クラスの識別には One-vs-Rest 法や Pairwise 法が用いられるが，どちらもあまり性能が変わらなかったので，結果は One-vs-Rest 法のものを示す．実験には SVM を利用した汎用のチャンカーである Yamcha を利用した [32]．

従来法の決定木は C4.5[46] を使ってデータからの情報量基準に基づいて構成した．推定に用いられる情報（属性）は，提案法と同様に，各位置でのモーラの種別である．その他の様々な設定のうち，SVM との比較のため，出来上がる決定木の葉の最少サンプル数，属性値のグループ化の有無，および，汎化性能を高めるための木の枝刈りの有無を様々に試し，ここでも 5 回の交差実験を行なう．例えば仮に，属性値のグループ化を行わない設定で，単語の末尾から先頭に向かう順番で各モーラがアクセント判定に有効であれば，決定木の根の直近の高さのノードには単語の末尾のモーラに関する質問が配置され，ノードの高さが葉に向かうにつれて単語の先頭側の各モーラがどの音かという質問が各ノードに配置された決定木ができることになる．

WEB に現われた未登録語のアクセント型推定実験においては，全データ（**表 3.3** を例

とする学習用と評価用のデータの両方）を用いて CCA や CCA+の SVM と決定木を構成した．WEB の未登録語と SVM や決定木の構成に使われたデータとの間の重複は無い．

3.4.3　結果と考察

カーネル関数の次元と精度

カーネル関数の次元が 1，2，3 の場合の単語の長さ毎の評価データに対するアクセント型の推定精度を**表 3.5** に示す．CRA，CCA，CCA+のそれぞれの手法において，姓，名毎にまとめて，各単語長，各次元での精度を配置した．‘ℓ’ はモーラを単位として数えた単語の長さを示す．単語の長さ毎に，‘/’ の左右に学習データと評価データに対する精度を示した．精度は 5 回の交差実験の平均値である．下線は評価データに対する最高精度の箇所に示した．CRA での結果には，予備実験で最も精度が高かった前後 3 モーラに文脈長を設定した場合の精度を記した．右端の平均は上記の最高精度の姓毎，名毎での重み付き平均値である．これは，各長さでの最高精度に表 3.3 のモーラ毎のデータ数（学習用と評価用の和）の重みをつけて平均した．また表 3.5 の末尾には従来手法としての決定木での各単語長での最高精度と，それを得た設定と，提案法と同様の方法で計算した重み付き平均値を示した．

CRA，CCA，CCA+の全てにおいて，次元の増加に伴って学習データに対する精度は向上した．

CCA においては学習データに対する精度が飽和し始めた最小の次元（以下，飽和次元と呼ぶ）において，評価データに対してもほぼ最高の精度が得られた．名の 5 モーラでは 2 が飽和次元であるが，最高精度ではない．精度は 0.9 ポイントだけ落ちるが，これは 1.3 個の誤りに相当し，劣化も小さい．このように，CCA では適切なカーネル関数の次元を決めやすいと考えられる．

一方，CRA と CCA+において，評価データでの精度が最高となる次元は最大の次元でも飽和次元でもなく，様々であった．CRA は長い単語で大きく精度が劣化した．CCA+においては学習データに対する精度が 100%になった後も，評価データに対する精度が上がるものもあれば，姓の 5 と 6 モーラのように落ちるものも有る．また，CCA+で次元を 3 以上に増加させても，6 モーラの姓名で性能が 70%近くにまで劣化した．このように次元を決めにくい点で使いづらい．

CCA+では SVM が 1 つであるので，実際の推定時には 1 つの次元に固定となる．例えば次元を 3 にすると，5 と 6 モーラの姓では最高精度ではないが，CCA とあまり変わらない精度が得られている．よって，CCA+は次元さえ適切に選べれば最高精度は CCA とほぼ同じであり，識別器を 1 つにまとめて扱える点で CCA の代用として利用できる可能性が有る．

表 3.5: 5 分割交差実験におけるアクセント型推定精度 [%]

model		l	次元			平均
			1	2	3	
CRA	姓	2	68.4 / 68.3	96.7 / 80.5	100.0 / 79.5	84.2
		3	73.4 / 72.3	97.9 / 79.1	100.0 / 78.5	
		4	84.5 / 83.6	99.6 / 87.0	99.9 / 87.3	
		5	88.8 / 87.4	99.5 / 86.6	99.6 / 88.0	
		6	72.2 / 63.5	96.4 / 58.1	97.1 / 65.4	
CRA	名	3	95.0 / 94.2	100.0 / 95.1	100.0 / 95.3	95.4
		4	96.1 / 95.1	99.8 / 95.7	99.8 / 95.9	
		5	94.4 / 93.9	100.0 / 91.7	100.0 / 92.7	
		6	92.3 / 90.6	99.2 / 93.1	99.2 / 93.9	
CCA	姓	2	84.7 / 79.6	100.0 / 80.2	100.0 / 80.0	86.1
		3	76.8 / 75.7	98.3 / 79.2	100.0 / 80.0	
		4	87.9 / 86.7	99.8 / 87.8	100.0 / 89.1	
		5	98.1 / 93.4	100.0 / 93.6	100.0 / 92.8	
		6	100.0 / 88.9	100.0 / 86.4	100.0 / 84.6	
CCA	名	3	96.0 / 94.6	100.0 / 95.5	100.0 / 95.8	96.0
		4	97.2 / 95.7	100.0 / 96.2	100.0 / 96.4	
		5	96.8 / 93.9	100.0 / 93.0	100.0 / 92.9	
		6	99.0 / 98.1	100.0 / 97.8	100.0 / 98.1	
CCA+	姓	2	82.7 / 78.5	100.0 / 80.0	100.0 / 80.1	84.8
		3	72.3 / 71.9	97.5 / 79.0	100.0 / 79.2	
		4	84.3 / 83.5	99.5 / 86.8	100.0 / 87.5	
		5	94.4 / 93.0	99.9 / 91.0	100.0 / 92.4	
		6	96.9 / 83.6	100.0 / 82.9	100.0 / 82.3	
CCA+	名	3	92.5 / 90.9	100.0 / 95.3	100.0 / 95.7	95.7
		4	96.1 / 94.9	100.0 / 95.8	100.0 / 96.1	
		5	95.1 / 92.2	100.0 / 92.8	100.0 / 93.1	
		6	97.9 / 95.9	100.0 / 96.4	100.0 / 97.5	

		l	最高精度	その場合の設定	平均
				最少事例数, 枝刈り, グループ化	
決定木	姓	2	100.0 / 79.1	1 , 無, 無	83.3
		3	100.0 / 76.6	1 , 無, 無	
		4	94.7 / 86.5	2 , 無, 無	
		5	98.6 / 91.5	2 , 有, 有	
		6	98.2 / 86.0	2 , 無, 有	
決定木	名	3	100.0 / 94.4	1 , 無, 無	94.5
		4	99.1 / 94.5	2 , 無, 有	
		5	99.7 / 92.9	2 , 有, 無	
		6	99.0 / 98.1	2 , 無, 無	

最も性能が高かった CCA での誤りに付いて述べる．最も多い誤り先は正解の型よりも学習データ数の多い型がほとんどであった．これは SVM の学習データ数がクラス間で大きく異なるために生じる．データ数の偏りを考慮に入れた学習にするなどの改善が必要である．例えば，4 モーラの名では 2，0，1，3 型の順にデータ数が多く，1，3，0 型は 2 型へ，2 型は 0 型への誤りが最も多かった．事例としては,「家康」が 2 つの型を持ったのに似て，2 型の「カツキヨ」が 0 型に，0 型の「シゲキヨ」が 2 型に誤るというような許容可能にもみえる誤りが多くみられた．

従来法との比較

CCA と CCA+と決定木（従来法）とを比較する．決定木での重みづけ平均は姓では 83.3%，名では 94.5%であった．一方，CCA と CCA+では**表 3.5** の右端の値のように，姓名の順に，CCA で 86.1%と 96.0%，CCA+で 84.8%と 95.7%であった．このように，提案法の CCA と CCA+は決定木を上回っている（CCA と決定木との間で 5%水準で有意）．また，CCA での各長さでの最高精度は決定木の場合よりも一貫して高い値が得られている．また，最高精度が得られる決定木の設定が様々であるため，CCA のほうが扱いやすい．

決定木のノードには各モーラの位置の読みが何という音かという質問が置かれている．推定時には，決定木の根から葉に向って未登録語の各読みを決定木の質問と照合していき，アクセント型が決定される．未登録語で，かつ，どの葉にもたどり着けない場合，根から葉に向かう経路上で合致しなくなったノードの 1 つ手前の根に近いノードでの最頻値がアクセント型として与えられる [46]．よって，決定木による推定では全ての読みが推定に使われる訳ではない．一方，SVM では，入力と複数のサポートベクトルとの類似度をカーネル関数で計算し，類似度の高いアクセント型に推定される．複数のサポートベクトルとの類似度を取る点で，未知の入力の読みのあらゆる部分がアクセント型推定に使われる可能性がある．実験結果に見られる決定木から CCA への性能改善の割合は小さいが，若干でも改善が得られるのはこのような仕組みの違いに有る．

WEB 文書の未登録語に対する精度

CCA と CCA+と決定木を用いて**表 3.4** の WEB 文書に現われた未登録語のアクセント型の推定を行った．結果を**表 3.6** に示す．

姓 33 個，名 29 個は評価データとしては規模が小さいが，表 3.3 を用いた実験結果と同様に，提案手法が決定木での精度を上回る．既知発音の姓名も合わせると，WEB 文書に出現した未登録の姓名の 94 から 95.5%に正しくアクセント型を付与できる．

表 3.6: WEB 文書の未登録語に対するアクセント型推定精度 [%]

	CCA	CCA+	決定木
姓	91	87	82
名	86	82	79

3.5　関連研究

土田らは名（forename）のアクセント型の推定に語尾のモーラを使って記述された規則を用いた [66]．**3.2** でも述べたように最近の名前では推定を誤る．最近の名前を含んでいる本論文の評価データに対する精度を調査したところ，精度は 80%台であった．従って，明らかに規則の更新が必要となる．一方，我々の方法では機械学習した CCA では 96.0%の精度が得られている．

広川らはカナ文字系列のみからなる日本人の姓のアクセント形を推定する方法を提案している [18]．この方法では，漢字で表した場合に漢字 2 字からなる姓のみを対象とし，平板型か起伏型かのみの推定に留まっている．起伏型で何モーラ目にアクセントの位置があるかまでは推定されていない．またクローズド評価のみで，未知の姓に対しての評価結果は無い．よって，音声応答装置において合成対象文中で様々に変化する人名部分のアクセント形をコンパクトに表現することが前提で，未知語を対象とする本論文とは目的が異なると推察される．辞書外，学習データ外の語は少なからず現われるので，未登録語のアクセント型の推定と精度評価が必要である．**表 3.4** のように辞書外の語は姓では約 50%現われている．仮に彼らの方法を未登録語に適用する場合，未登録語が学習済みの漢字の分割パターンにマッチしない場合が有り得るが，その場合の方法が明記されていない．2 つの漢字ではなく 3 つ以上の漢字に分解可能な場合についても不明である．姓とは異なり，名では，本論文の**表 3.1** からも読み取れるように，(喜代子と清子のように) 同じ読みに対して様々な漢字が対応しうるので，読みが同じでも未知の漢字を使った名のスコアは，既知の漢字と読みのペアのスコアのみから判定され，残った読みの情報が使われる仕組みになっていない．また「服部（はっとり）」,「大和（やまと）」,「岩動（いするぎ）」のような漢字と読みが対応しない場合の対処も不明である．一方，我々の手法では読みに基づくので未登録語に広く対応できる．

3.6　3 章のおわりに

単語の読みとその並びの情報から，Support Vector Machine を使って単語のアクセント型を推定する方法を提案した．

単語の長さ毎に構成された SVM を用いる CCA が提案法のなかでは最も性能が高く，姓名を使った未登録語の実験では，**表 3.5** のように，学習データに対してはほぼ 100%，未知の姓に対しては平均 86.1%，未知の名に対しては平均 96.0% という高い精度でのアクセント型推定が可能であることが分かった．さらに，実際の WEB 文書に出現した未登録の姓名に対してアクセント型推定を行ったところ，精度は姓名それぞれ 91%，86%の結果が得られた．また，従来の情報量基準で構成された決定木の精度を上回った．これらから提案法の有効性が確認された．

第4章　対話韻律の実現に向けたF_0大域形状の分析

4章では，音声翻訳器を用いて対話を支援する際に生じる未知情報の推定課題の１つである，対話口調での韻律を予測する課題を解決するための特徴分析について述べる．

4.1　はじめに

4.1.1　研究の背景

コーパスと統計的手法の導入により，現在のテキストからの音声合成器（Text-to-speech synthesis: TTS）は肉声感のある「読み上げ口調 (reading style)」韻律を持つ合成音を生成できるようになった [29, 47, 50, 49]．その結果，人間との対話的なやりとりが望まれるようなさまざまな応用場面においても音声合成を利用することへの期待が高まった．それらの応用場面には，娯楽の為の物語の語り聞かせ，電話コールセンターの自動応答，音声支援付きの電子商取引場面での商品宣伝，対話ロボットからの音声応答等の場面がある [[62] 等]．しかし，従来の TTS は新聞などを対象とした読み上げ口調，すなわち，感情や強調を伴わない淡々とした韻律を持つ合成音の開発が進められてきたため，それらの応用場面に対して十分な合成音を提供できなかった．この結果，人に語りかける場面を想定した「対話口調 (communicative style)」を実現する韻律制御の研究の必要性が高まった [31, 52, 62]．口調は，話し手が与える話し方や聞き手が感じる印象を指すが，これには声の大きさ，高さ，長さに関する「韻律」と声質に関する「音源スペクトル特性」の２つの音響的特徴が関係する．本研究では韻律，特に，声の高さ（基本周波数：fundamental frequency: 以下 F_0 と表記する）に焦点を当てる．音声合成においては，この F_0 をテキスト等の入力言語情報から予測することが必要となる．

4.1.2　用いるF_0制御モデル

本研究では F_0 のモデルとしてこれまでに提案されている重畳モデル（superpositional model）[1, 11, 50, 53] を念頭に置き，同族のモデルの確立を目指して検討を行なう．重畳

モデルでは F_0 を時間範囲の異なる複数の F_0 時間変化曲線（成分）を対数領域で足し合わせて表現する．すなわち，アクセント句（minor phrase）に基づく言語単位に対応した局所的な F_0 変化を表すアクセント成分と，全体的な下降特性を示すイントネーションに対応するより大きな区分（intermediate phrase，または，major phrase と呼ばれる）に対応するフレーズ成分，および，発話全体に及ぶ平均的な高さを示す成分等を重ね合わせることにより全体の F_0 時間変化曲線を構成する．このような重畳モデルは，F_0 時間変化曲線を生成機構を反映した形で数理的に記述するため，工学的にも音声科学的にも有用性が認められている．

一方，重畳モデルとは対照的に，F_0 時間変化特性を直接予測する方式がある [63, 75]．この方式は F_0 の生成機構を考慮せずに，F_0 時間変化自体を当該箇所周辺の多くの言語情報から統計モデルにより直接予測する．このモデル化では直接観測できない構成要素成分への分解という重畳モデルの推定問題の困難さを回避できる．この反面，モデル化に大量の学習用のデータを必要とし，分析モデルとしての理解が難しいなど，工学的，科学的双方に不都合な点がある．このため，本検討では，重畳モデルを採用することとした．表情豊かな対話音声では F_0 は多様に変化するが，単語のアクセントは読み上げ口調の時とほぼ変わらない．重畳モデルを用いることにより，対話口調と読み上げ口調との間で差異が発生する成分と発生しない成分を区別した記述の優位性が期待できる．これにより，各成分での差異の存在する箇所と無い箇所，その差異の程度，差異の発生する言語要因とを明らかにし，精度が高く，制御要因の明らかな F_0 制御モデルを効率的に獲得することが期待できる．

4.1.3　関連研究

対話口調韻律実現のために，本研究では読み上げ口調と対話口調との比較を通して，対話 F_0 の特徴を明らかにする．口調間の比較を行なった従来研究として，Abe らの研究がある [2, 3]．この研究では，小説，宣伝，事典の場面の口調を対象として，2 階層モデル [7] でアクセント句に亘る local モデルと，それより大きな範囲に亘る global モデルが提案されている．local モデルには口調間で共通のモデルを利用し，global モデルのみ口調ごとにモデルを構築することで対話口調（原典の表現では,「発話様式」）を実現できると述べられている．しかし，この実験で用いられた文章は場面に無関係の共通の文であるため，実際には場面毎に変わる語彙と F_0 変化との対応については未解明のままである．

一方，Sagisaka らは副詞からなるアクセント句と形容詞からなるアクセント句が連接している簡単な場合を例に，語彙の属性に応じて F_0 時間変化特性を推定できることを示した [52]．このモデル化では，指令応答モデル [11] を用いたアクセント成分の制御により対話口調の F_0 を再現できることを明らかにしているが，対象が極めて限定されている．こ

のため，対話用途を対象とした音声合成の実現のためには，種々の語彙や表現からなる多様な対話音声を対象とした分析が必要となる．

4.1.4　本章の構成

このような背景から，本研究ではさまざまな場面で実際に現れる対話音声を対象として，観測された対話口調と別に収録した対応する読み上げ口調との間での F_0 の比較をおこなう．分析には重畳モデルを用い，F_0 を構成する各成分の差異の大きな箇所，小さな箇所，および，それらと実際の発話に現れる様々な表現（語彙）との関係を明らかにする．以下，**4.2** では，本研究で用いる F_0 モデルとその各成分の抽出法を述べる．口調間での F_0 の成分の比較の方法を **4.3** で述べる．**4.4** では本研究で用いる音声コーパスについて説明する．成分ごとの比較結果を **4.5** で述べ，読み上げ口調に比べて対話口調において大きく異なる成分を明らかにし，関連する表現を考察する．**4.6** で結論と展望を述べる．

4.2　F_0 モデルと各構成成分の抽出方法

本研究では，F_0 制御モデルとして，時間範囲の異なる複数の構成成分の重ね合わせとして表現する重畳モデルを使用する．このモデルでは，

1. 発話文全体にわたる utterance 成分
2. アクセント句間の大まかな変動を示す phrase 成分
3. アクセント句内でのアクセントに対応する変動を示す local 成分

の 3 成分によって F_0 時間変化曲線を記述する．各成分の定義と抽出方法について以下説明する．

utterance 成分

各発話文の全体にわたる F_0 の特徴量として，各発話文での F_0 の平均値と標準偏差とを utterance 成分として抽出する．発話文ではなく，それよりも小さい単位であるポーズで挟まれた句を単位とする方法も考えられるが，発話文全体の内容を確認した後に発話を行なう場合，ポーズの後の内容も見たうえで，全体の声の高さが決定されると考えられるので，発話文を抽出の単位とした．

phrase 成分と *local* 成分

前記の utterance 成分を除去したあとの F_0 時間変化曲線をアクセント句ごとに時間軸上で分割する．そして，各アクセント句の平均値と標準偏差を phrase 成分とした．そして，アクセント句毎に各アクセント句の phrase 成分をさらに除去した後の F_0 時間変化曲線を local 成分とした．

各成分の除去には，除去する成分の平均値と標準偏差とを用いて，次式の正規化処理を適用することによりおこなう．

$$y = (x - \nu)/\sigma$$

ここで，x と y は正規化前後の変数，ν と σ は x の平均値と標準偏差である．

4.3　口調間比較の方法

比較の対象とする音声データは，口調間で同じ位置にポーズが置かれており，かつ，同一の音素ラベルが付与されている発話（片方だけでの無声化や長音化がない発話）であり，後述のアクセント句境界の位置が口調間で同一の発話のみを比較分析の対象とした．そのような音声データから F_0 を抽出し，前記の各成分を抽出し，口調間比較を行って差異のある箇所の特定や，差異と発話内容との対応を検討する．

比較の方法を説明するために，F_0 構成成分の例を**図 4.1** に示した．発話内容は「[会社帰りは][携帯で][ショッピング！]」であり，商品宣伝場面の音声データから抜粋したものである．”[ ] ” で囲まれた区間が韻律の単位である「アクセント句」である．アクセント句は，韻律語，または，minor phrase とも呼ばれ，その中では声の高さが急激に落ちる箇所（アクセント）が高々１つとなっており，従来から韻律の研究で利用されている単位である．図 4.1 の３枚のパネルに亘って縦に引いた点線はアクセント句の境界を示す線である．全てのパネルにおいて，青色（白黒印刷の場合，濃い灰色）が対話口調，ピンク色（白黒印刷の場合，淡い灰色）が読み上げ口調に対応する．図 4.1 の最も上のパネルには，それぞれの音声から抽出した基本周波数（F_0）を示した．このパネルの段階では抽出したままであるので縦軸の単位は Hz，ただし表示上は log スケールで表示している．発話全体にわたって水平に引いた点線がそれぞれの口調での発話全体にわたる F_0 の平均値，すなわち，utterance 成分（のうちの平均値）である．この utterance 成分を除去した後の F_0 変動パターンを時間軸上でアクセント句ごとに分割し，アクセント句ごとに平均値と標準偏差，すなわち phrase 成分，を計算し，平均値のみを水平な直線で記した結果が真中のパネルである．縦軸は utterance 成分で正規化後の値であるので倍率になる．ここから，さらに phrase 成分を除去して残った結果が最も下のパネルの変動パターンであり，local 成分の変動を示している．各点は有声の母音の中心部に対応する．アクセント句ごとにまとまるよう点を繋いで記した．縦軸は phrase 成分で更に正規化を行なった後の倍率になっている．

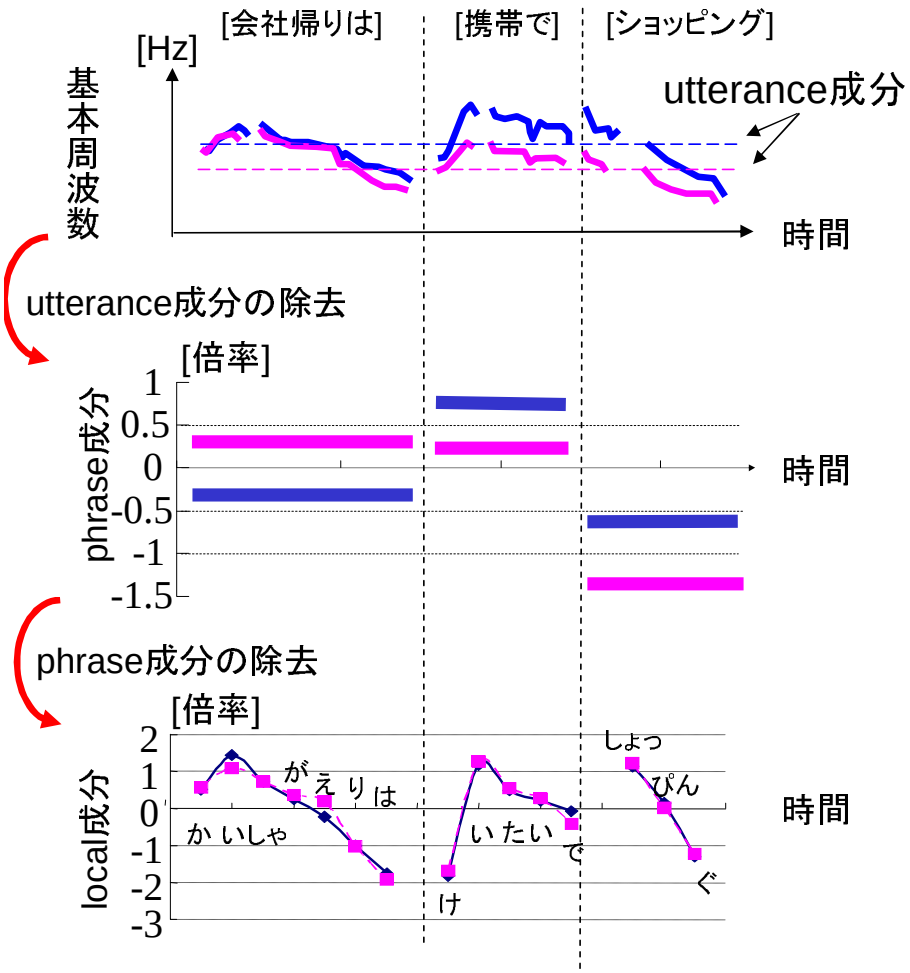

図 4.1: F_0 構成成分の抽出・除去と口調間比較の例

表 4.1: 音声データ諸元

	AP	FT	OP
文数	152	64	104
アクセント句数	1,550	684	1,061
形態素数	3,613	2,802	3,132
総発声時間 [分]	20	13	14

例えば，図 4.1 のような分解結果の比較から,「local 成分が口調間でほぼ一致し，phrase 成分では，対話口調での 1 番目のアクセント句と 2 番目のアクセント句の平均値の変化が読み上げ口調に比べて大きく増大し，その発話箇所の語彙内容が副詞句であった」等の結果を得ることが期待される．

4.4　音声コーパス

本研究では，従来研究 [2, 3] とほぼ同様のジャンルである，商品宣伝 (AP)，童話の語り聞かせ (FT)，電話対話 (OP) の 3 種類の場面を想定し，音声コーパスの作成を行なった．それぞれの場面での発話内容に対して，女性声優 3 名による以下の口調での発話を収録した．

- 商品宣伝（products **Ap**peal: AP）
  CM 等で客に向かって商品を売り込む口調
- 童話（**F**airy **T**ale: FT）
  親が子供等に童話を語り聞かせる口調
- 電話オペレータ対話（telephone **Op**erator: OP）
  電話センター等で従業員であるオペレータが客と話す口調

本論文ではこれらの口調をまとめて「対話口調」と呼ぶ．比較用に，対話口調と同じ発話内容を同じ数だけ強調等をつけず淡々と素読させた音声を収録した（「読み上げ口調」と呼ぶ）．このコーパスの諸元を**表 4.1** に示す．アクセント句数と総発話時間は 3 名の話者での平均値である．これらの音声に対して，音素ラベル，アクセント句の境界，アクセント句の境界にポーズがある場合のその始終端時刻，各母音の中心の F_0 値 (Hz) を付与した．

１発話あたりの平均のアクセント句数は場面を問わず，ほぼ 10 であるが，１アクセント句あたりの形態素数は AP が 2，FT が 4，OP が 3 というように，場面ごとに異なっている．FT は語り部分が多く，会話部分は少数であった．

4.5　各構成成分の口調間比較

4.5.1　utterance 成分の比較

発話全体にわたる平均的な F_0 の高さ，すなわち，utterance 成分が口調と場面に応じてどのように異なるかを調べるために，対話と読み上げの 2 つの口調と AP と FT と OP の３つの場面からなる合計６つの組における utterance 成分 (のうちの平均値) の分布の比較を行なった．

例として，3 人のうちのある 1 話者の utterance 成分の分布を**図 4.2** に示す．横軸が前記の組である．AP，FT，OP が場面を表わし，左側の 3 つが対話口調での分布，右側の 3 つが読み上げ口調での分布である．縦軸方向に各組での分布を Box-Whisker-plot として表現している．縦軸の単位は Hz であるが，表示は log スケールで行なっている．縦の点線の下端が分布の最小値であり，上端が最大値である．その上下に丸印がある場合は外れ値を意味する．点線の間に置かれた矩形の底辺部分の高さが 25 パーセンタイル，上辺部分の高さが 75 パーセンタイル，真ん中の水平線の高さが 50 パーセンタイルの各 F_0 の値 [Hz] である．図が縦に長ければ裾野の広い分布，短ければ裾野の狭い集中した分布となっていることを意味する．

50 パーセンタイルを中心として上下にほぼ均等な形状であるので，発話を単位としてみると，データは概ね偏りなく収集されている．同一の場面では，全ての話者において，右側の読み上げ口調に比べて左側の対話口調の分布が高く位置した．また，読み上げ口調に比べて，対話口調では長方形の上辺と下辺の間が広い．すなわち，読み上げ口調に比べて対話口調では高い F_0 帯で，広い F_0 範囲を使い分けて発話されている．

対話口調では，分布は AP と OP が高く，最後に FT が来るという順であった．AP と OP の順序は話者によって異なった．このような対話口調における場面間での分布の高さの違いの生じる理由の 1 つとして，合成対象の文からは抽出が困難な要因の関与が考えられる．例えば，AP や OP では聞き手が遠くに居るが，FT ではすぐ近くに居るという話し手と聞き手の間の距離との相関や，AP や OP では強く FT では弱いという聞き手への訴えかけの度合いとの相関が挙げられる．これは場面ごとのようにモデルを細かく使い分ける方式であれば，場面ごとのデータで学習した各モデルが差異を吸収する．

一方，差異をより明らかにするために，各場面の語彙との関係を調べた．対話口調の AP では,「発売」や発話末の価格のような威勢よく話される語彙を含む発話の平均値が高く，商品説明のような落ち着いた内容の発話では平均値が低いという違いや，OP では謝意を述べる部分では高いことが観察された．この観察結果は，対話口調の TTS においては，合成対象の発話内容の語彙に応じた utterance 成分の調整が必要であることを示唆する．発話内容や場面に応じて平均を変えることは「声を張る」というように直感にも合い，必要な要素と考えられる．一方，従来の読み上げの TTS では，この utterance 成分に

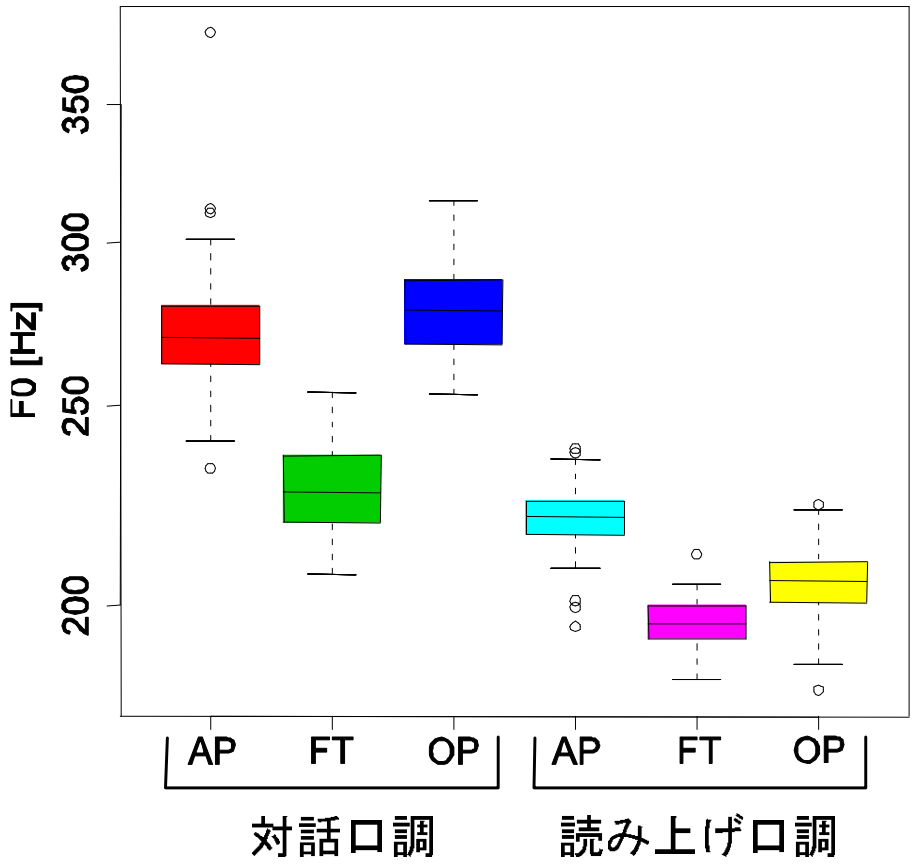

図 4.2: utterance 成分の分布の事例

相当する値は固定値の場合が多かった．

4.5.2　phrase 成分の比較

相関係数の比較

phrase 成分の口調間比較の為に，発話毎に，口調間の phrase 成分系列の間の相関係数を算出した．図 4.1 の真ん中のパネルの平均値系列の口調間比較である．口調間で同様の上下変動をおこなっていれば，相関係数が 1 に近い値となる．

結果を**表 4.2(a)** に示す．表 4.2(a) では，相関係数のレンジを，0.4 以上 0.7 未満の「中程度の正の相関」を示す区間と，0.7 以上の「強い正の相関」を示す区間に入る発話の割合 [%] のみを記載した．その他の区間，すなわち，0.4 未満の「弱い正の相関」や「負の相関」を示す発話はまとめて，0.4 未満とした．

口調間で 0.7 以上の強い正の相関を示す発話が AP では約 67%に留まった．FT と OP では 90%近く存在した．

変動量の比較

相関係数は高い結果であったが，両口調の phrase 成分（図 4.1 の真ん中のパネルのような変動パターン）がきれいに重ならない場合が見られたので，変動量の口調間での差異の程度を確認するため，隣り合うアクセント句間での変動量を求め，読み上げ口調を基準として口調間での比を算出した．この量が 1 に近い場合には口調間での違いが少なく，1 から離れる場合には，相関係数が同じであっても，より大きな変化が存在することになる．

本研究ではこの変動量の比が，0.8 未満であれば「縮小」，0.8 以上 1.2 未満であれば「ほぼ同じ」，1.2 以上であれば「拡大」と定義して，この 3 つの領域に含まれる，アクセント句間の変動を分類した．結果を**表 4.2(b)** に示す．

比が 0.8 以上 1.2 未満という「ほぼ同じ」とみなす領域に属する割合は高々30%程度であることから，読み上げ口調を基準とすると，対話口調の phrase 成分 (のうちの平均値) には拡大または縮小の大きな変化があると解釈できる．音声信号からの F_0 の自動抽出の研究における誤差評価では，抽出値が正解の上下 5 %の幅にある場合に正解と定義する Gross error 評価が行なわれる．本研究の「ほぼ同じ」とみなす範囲の幅の基準は基本周波数抽出の Gross error 評価の基準よりも緩い基準であるが，それさえ外れる点で大きな変動が生じていると考えることは可能である．

「読み上げ口調は『朗読』と同等である」との見方を持つ場合には，FT での口調間差が小さいと予想するかもしれない．しかし，本研究の比較結果では大きな差が表れた．こ

表 4.2: phrase 成分，および，local 成分の口調間での比較

(a) phrase 成分相関係数による発声の分類

相関係数のレンジ	AP	FT	OP
0.4 未満	8.8	0.0	1.3
0.4-0.7	24.5	12.5	9.2
0.7-1.0	66.7	87.5	89.5

(b) phrase 成分変動量比による発声の分類

変動量比のレンジ	AP	FT	OP
0.8 未満	51.2	47.6	37.4
0.8-1.2	25.7	23.1	30.2
1.2 以上	23.1	29.3	32.3

(c) local 成分相関係数による発声の分類

相関係数のレンジ	AP	FT	OP
0.4 未満	8.1	5.5	6.4
0.4-0.7	7.3	6.7	4.8
0.7-1.0	84.6	87.8	88.8

れは，本研究での対話口調での音声収録では文章の実際に発話される場面を想定して起伏をつけて発話されているのに対し，本研究の読み上げ口調の音声収録では強調等をつけずに淡々と発話（声読と呼ばれる）しているためと考えられる．すなわち，本研究の読み上げ口調は朗読ではないので，大きい拡大または縮小の変動が生じている結果は自然な結果と思われる．

口調間での大きな差異の生じている箇所の発話内容の語彙は次のようであった．

AP
: 商品の感想や特徴に関する表現，感動詞，相手に納得や同意を求める表現，勧誘や発売販売等の語，および，ポーズ前や文末の体言

FT
: 擬態語，擬音語，台詞の部分

OP
: 敬語部分

この結果から，対話口調の TTS での phrase 成分の制御においては，発話内容，特に，語彙の属性に応じた差異化が必要であることが示唆される．一方，従来の研究 [1, 19, 49] では，品詞や構文情報を制御要因とする個別のモデル（場面や感情ごとのモデル）が用いられてきた．

4.5.3　local 成分の比較

local 成分（図 4.1 の最下のパネルに示した 2 回の正規化後の F_0 変動パターン）の口調間の相関係数をアクセント句ごとに算出し，その大きさによるアクセント句の分布を調査した．

全分布の内で，相関係数が 0.4 未満のアクセント句の割合はまとめて 0.4 未満とし，0.4 以上の分布，すなわち「中程度の正の相関」を示すアクセント句の割合と,「強い正の相関」を示すアクセント句の割合とを**表 4.2(c)** に示す．

全場面において，0.7 以上の強い相関を示すアクセント句が約 85 %以上存在した．発話毎に local 成分（図 4.1 の最も下のパネルのような変動パターン）を目視で比較し確認したところ，強い相関を示すアクセント句では両口調の変動パターンの軌跡がほぼ重なるか一致していた．

相関の低いアクセント句では，従来の談話研究で言及されてきたような phrase 末尾の助詞付近での F_0 の再上昇があった．これらについては，従来研究 [2, 63] と同様に，末尾の助詞付近に上昇パターンを対応付けることで制御できる．他には，強い負の相関を示す

アクセント句が存在した．例えば,「[お二人とも][すっかり](ポーズ)[この][フットマッサージャーの][虜みたいですね]」の「すっかり」のように，１つの副詞からなるアクセント句の直後にポーズがある場合に，対話口調では phrase boundary tone に乗って F_0 が上昇するのに対し，読み上げ口調では下降し，逆相関となった．local 成分の生成において，アクセント句末尾に向かって上昇する変動を与えるか，phrase 成分で同様の効果を加えるかといった方法の適用が必要となる．これは正規化の逆変換の適用 [例えば [62]] だけでは十分ではないことを意味する．

相関係数が 0.4 未満になったアクセント句を語彙との関係でみれば，AP では発話末やポーズ前後，FT では擬態語や擬音語や台詞部分で，OP では，敬語表現部分で生じていた．しかし，少数であった．

4.5.4　比較結果のまとめ

以上の 3 成分の比較から，一部には local 成分の中には負の強い相関を示す場合もあり，正規化の逆変換の適用 [62] だけでは十分ではないことが明らかとなった．一方で，読み上げ口調の local 成分を対話口調の phrase 成分と対話口調の utterance 成分に重畳させることで，対話口調の F_0 をほぼ生成できる見込みがあることがわかった．このとき下記の新機能が必要となる．

1. utterance 成分を発話内容から予測する機能
2. phrase 成分の，特に，変動量を，従来の構造情報に加えて，発話内容の語彙から精度高く予測する機能

従来は，多様性の発生要因として，文法，話し手の心的態度，談話，スタイルというイントネーションの 4 機能 [70]，および，それらの一部を細分した情報（例えば，心的態度の細分化と考えられるパラ言語情報からの音声分析，および，感情からの制御（ともに [19] の 2.3 節と 4 章]）について研究が行われてきた．これらの研究のうち，本研究に特に関連するものとしては，従来研究 [70] の「スタイル機能」の言及部分が対応するが，発話の場面ごとの発話スタイルの全般的な特徴の記述に留まっている．音声合成を目的とした場合，合成対象の文に含まれる語彙と発話スタイルとの関係の解明が必要と考えられる．

本章で行った比較分析，特に，口調間差異と語彙との関係を明らかにすることで，上記のイントネーションの４機能の解明につながることが期待される．このためには，工学的には汎化やデータ量との点で，言語科学的には機能理解の点で，予測に用いる発話内容は語彙そのものよりは，各語彙を表現する概念的なベクトル表現が妥当であると考えている [31, 55]．

4.6　4 章のおわりに

種々の対話に用いる音声合成のための対話 F_0 予測方式の確立を目的として，複数の成分の足し合わせとして F_0 をモデル化する重畳モデルを用いて，対話口調と読み上げ口調との間での F_0 の各成分の比較を行った．比較においては，商品宣伝，童話の語り聞かせ，電話応対の各場面の実際の発話内容を用いて収録した音声を利用した．その結果，utterance 成分と phrase 成分といった大域的な成分に場面間や口調間での大きな差異が存在することを確認した．差異の大きな箇所は電話応対での敬語部分のように場面に特徴的な表現で生じることが多いことを観察した．local 成分といった局所成分では，phrase 境界前後などのように個別に対処が必要なアクセント句が存在するものの，口調間で強い相関を示す発話が 8 割から 9 割存在することがわかった．

今後は，大域的な成分での差異およびその差異の程度と語彙との対応関係を明らかにする．そして，その対応関係を定量的に記述し，それを大域的な成分の予測のための統計モデルの入力素性として導入し，精度の高い対話口調の F_0 の予測を実現する予定である．

第5章　音声認識と言語翻訳における処理単位の統一のための発話分割

5章では，音声翻訳を構成する各要素技術の結合で解決されない課題である情報統一化の課題，すなわち，音声認識と言語翻訳との間での処理単位の統一化課題の解決を行なう．

5.1　はじめに

自然な話し言葉による対話においては，1回に，あるいはひと息に複数の文が発話(発声)されることがしばしば起こる．この様子を**図 5.1** に示す．図5.1の話者Aから話者Bへと話者交替が起こるまでの区間（turn1）の中の第1番目の音声区間，すなわち，発話 utt1 の中の文 sent1 と文 sent2 や、話者Bから次の話者へ話者交替が起こるまでの区間（turn2）の中の第2番目の音声区間，すなわち，発話 utt4 の中の文 sent5 と文 sent6 がその例である．従来の音声認識では，発話（発声，図5.1の utt）を単位として認識処理が行なわれる．

しかし，音声対話システム内の理解処理や，音声翻訳処理や，話し言葉での対話を要約する処理においては，複数の文を含んだ発話そのものを単位として処理することは困難であり，音声認識において，あるいは，その後の言語処理のどこかで，発話を文などに分割する処理が必要となることが指摘されている [12, 34, 60, 76]．例えば，話し言葉を対象とした翻訳では処理単位は発話よりも小さく，文相当であった [13]．つまり，翻訳単位の区切りとしての句点の位置（図5.1の target）を認識する処理が必要となる．そして，音声翻訳対話システムにおいて，待ち時間の少ない円滑な会話を実現させるためには，発話の中の文の境界（図5.1の target）を，音声認識の過程で発話の終了（utt の右端）を待たずに発見し，発話を分割して，分割された時刻までの音声認識結果の単語列を翻訳処理へ送ることが必要となる．以上のような目的から，本研究では音声認識過程で発話内に含まれる文の境界（図5.1の target）を検出することを問題とする．

発話を分割する従来の手法として，隣り合う単語間での分割の有無の判定を，その単語の並びの生起頻度とその単語並びの中で分割が生じる頻度と人手で決められた閾値からなる識別関数を用いて行なう方法 [34, 60]，文の先頭の語であるか否かという2つの状態を導入した確率モデルを用いる方法 [57]，ニューラルネットを使って隣り合う単語間での分

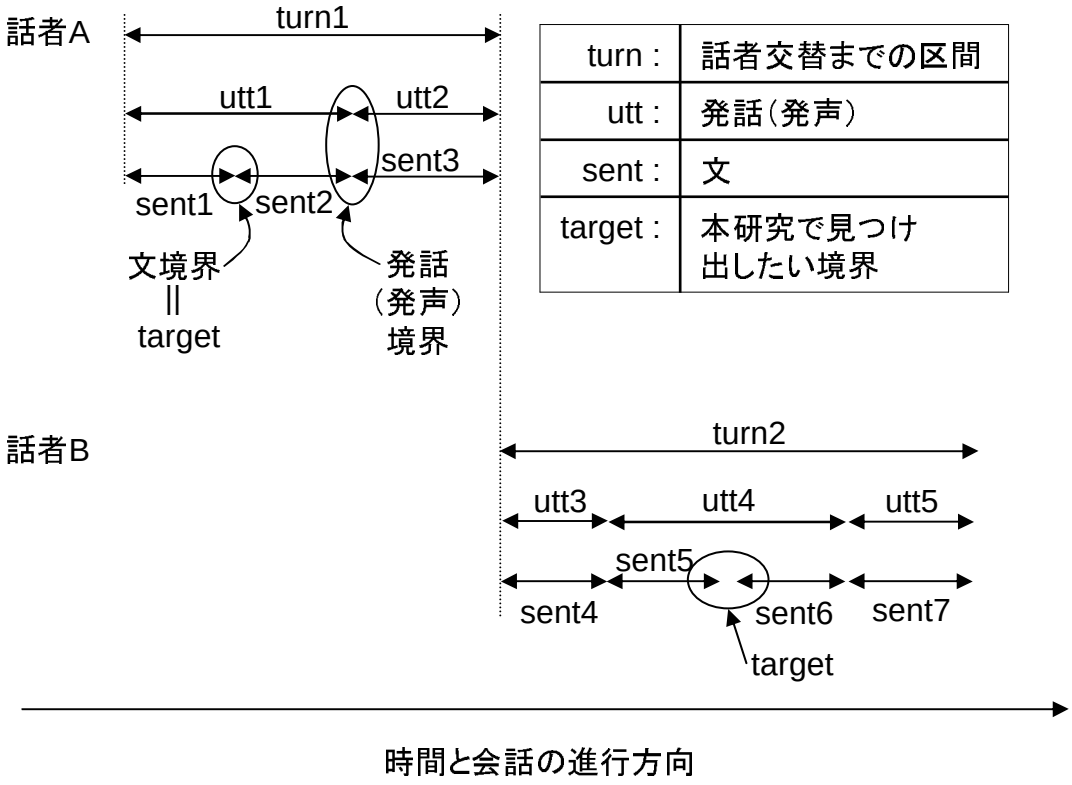

図 5.1: 対話の中の 1 回の発話で複数の文が話される様子

割の有無を判定する方法 [15, 76] が研究されている．これらは，a) 音声認識用の言語モデルとは異なる知識として，分割のためのモデルが作成され利用されていること，b) 音声認識の後処理として分割が行なわれること，を特徴としている．しかし，音声の認識と分割のための統合されたモデルを使って，音声認識過程で分割を行なう方法とその評価を行なった研究は例があまりない．ディクテーション結果の読みやすさの向上を目的として，英語や日本語の言語モデル（単語 3gram）に，ピリオド，カンマ，セミコロンまたは句点などを組み込んで，音声認識時にそれらを認識する研究は行なわれている [8, 26]．それらの研究では認識対象文は対話文ではない．文献 [8] は分割の評価を行なっているが対象が英語であり，文献 [26] では発話の分割の点での評価は行なわれていない．また，句点，読点，息継ぎを同一の記号にして学習された言語モデルが研究されているが，目的が単語認識精度の向上にあり，分割の観点からの評価はない [21]．このように，我々が認識の対象としている対話音声での分割とその評価は行なわれていない．

そこで本研究では，日本語の対話音声を対象とし，複数の文が含まれる発話内での文の切目である句点の認識を問題とする．その解決のために，音声認識用の統計的言語モデルに分割の情報を統合する．本稿では，言語モデルの構成を述べ，分割の性能と，句点以外

の単語の認識性能の点で議論する．以下，**5.2** では，分割が必要となる発話の事例とその頻度について簡単に説明し，本手法を **5.3** で説明する．**5.4** では評価実験について述べ，**5.5** で評価実験のまとめを行ない，結論を **5.6** で述べる．

5.2　分割が必要となる発話

自然な会話では 1 回の発話の中に複数の文が含まれている場合があり，そのような発話は文への分割が必要となる [60]．本章ではそのような発話の例を示す．例えば，ATR から公開されている ATR 自然発話音声言語データベース [61] を考える．これはホテルの予約やサービスの問い合わせに関するホテルの従業員と宿泊客との会話を想定して集められたデータベースであるが，その音声データの中には次のような発話音声がある．

例１：複数の文を含む発話音声

宿泊客：　もしもし交通手段についてちょっと教えて頂きたいんですが

ホテル：　はい畏まりましたどちらへお出かけでしょうか

この例では，ホテル側の 1 つの発話に，“はい畏まりました。” と “どちらへお出かけでしょうか。” のように 2 つの文が含まれており，分割の必要な発話となっている．このように，発話全体が必ずしも 1 文にはなっていない．従って，音声認識の後または翻訳などの言語処理の前に，そのような発話を文へ分割する処理が必要となる．

このデータベースの書き起こしテキストデータは，発話を単位として作成されている (例えば，ホテル側の発話は “はい” から “しょうか” までを 1 つとして扱われている．)．発話の末尾には句点が付けられている．さらに，その発話の中に複数の文があるかどうか (句点を置くべきかどうか) については，予め定められた一定の基準 (話し始めの決まり文句の後や挨拶文の後以外の箇所のうち，一般的な文法において文末になる可能性のある箇所と，話し始めの決まり文句や挨拶文と同じ単語列でも決まり文句や挨拶文とは異なる意味を持つ場合にはその末尾の箇所に句点を置く，という意味まで考慮した上での判定基準) に基づいて専門の作業者によって判定され，句点が付与されている．翻訳でもまた句点を単位として翻訳処理が行なわれてきた [13]．

ATR 自然発話音声言語データベースでの発話中の句点の頻度を**表 5.1** に示す．1 会話がおよそ 12 発話からなり，その 1 発話がおよそ 16 単語を含んでいる．そのような全発話のうちのおよそ 36%が発話中に句点を含んでいる．

このデータベースでは，文間の無音区間の長さは様々であり，無音区間に関する物理量 (音響的特徴量) のみに基づいて文を定義し，発話を分割することは難しい [60]．

表 5.1: 発話中の句点の分布

片側会話数	7,262
発話数	89,152
発話内句点数	32,258
総単語数	1,392,557

そのため本研究では，5.3 以降で説明と評価を行なうように，音響的特徴と言語的特徴の両方を利用して尤もらしい分割が行なえるよう，音声認識を利用する．

本研究では，モデルの学習及び評価用のデータとして，会話の書き起しデータである ATR 自然発話音声言語データベースを利用する．そして，この中の句点を翻訳単位の境界と定義し，この句点で区切られている単位を翻訳単位，すなわち，文と定義する．このため文は，発話の終端または句点までの単語列となる．本稿では発話を上記のような文へ分割する問題を研究する．

5.3　実現手法

本研究では，音声認識の過程で発話の分割を行なう．つまり，音声認識の過程で，音響モデルから得られる音響尤度と以下に説明する言語モデルから得られる言語尤度とを使って，一般の単語と同時に，分割の位置を示す句点を認識する．本研究では，音声認識と分割のための統一された言語モデルとして統計的言語モデル (単語 Ngram，多重クラス複合 Ngram) を用いる．従来の音声認識では句点の認識は考慮されていなかったため，句点を取り除いて言語モデルが作成されていた．しかし，本研究では，発話中の句点を学習データの中に残し，発話中の句点への遷移確率，および，発話中の句点からの遷移確率の推定を行なうことによって，モデルの統合を行なう．さらに，音声認識用の辞書には句点とその発音記号を登録する．

そして，入力された音声に対して，音声認識過程で一般の単語と同様に句点の認識を行うことによって，単語グラフ内の尤もらしい位置に句点が埋め込まれる．発話の分割は選ばれた単語グラフ内のパスの中の句点の位置で行なう．単語グラフのパスに句点が無ければ，分割しないものと判定する．

全体の単語グラフをリアルタイム処理で文毎の単語グラフに分割し，早期に後段の処理に送ることが可能となるが，その実装手法は通常の探索の問題であり，ここでは議論しない．

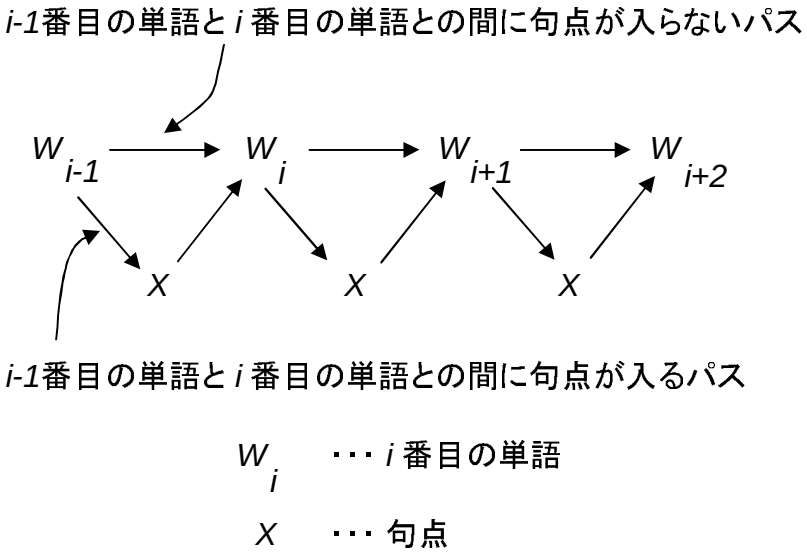

図 5.2: テキスト入力に対する分割実験での全単語間に句点が入る場合までを想定した単語グラフの一部分

5.4　評価実験

音声認識過程で本統計的言語モデルを使うことによって，発話の中の文境界位置での句点の認識と句点以外の単語の認識とが正確に行なえるかどうかを確かめるために，以下の実験を行なう．

5.4.1　実験用データ

まず最初に実験に用いるデータを説明する．本研究では ATR 自然発話音声言語データベースを用いる．本実験ではそれを以下に説明するように 4 つのサブセットに分けてモデルの学習と評価に利用した．

まず最初に，句点を含まないテキストを入力として，本言語モデルによって文への分割を行なう (文に句点の付与を行なう) 性能を検査する．句点を含まないテキストを入力とすることにより，音声認識において句点以外の単語の系列が完全に認識された場合を模擬する．すなわち，単語の位置が既知であるが句点の位置は未知とし，すべての単語間に句点が入る場合までを想定して単語グラフを作り，その中の各単語と句点に本言語モデルによって言語尤度付けを行ない，尤度のもっとも高いパスを選択する．その結果に含まれる句点の位置で発話の分割を行なうものとする．例を**図 5.2** に示す．W_{i-1} と W_i との間に句点 X が入りうるか否かを “W_{i-1}, X, W_i” と “W_{i-1}, W_i” との尤度から判定する．これ

表 5.2: 発話分割実験で用いるデータセット

	片側会話数	発話数	発話内句点数	総単語数
ED1	18	216	73	2,437
ED2	42	551	89	4,990
ED3	7,202	88,385	32,096	1,385,130
ED4	1,218	15,891	5,585	244,049

により，句点以外の単語が完全に認識される場合だけではなく，句点の音響的特徴 (ポーズ) が完全で音声認識の候補に含められる場合を模擬した状況下での発話の分割性能を評価する．

そのために以下で説明するデータセット ED1 と ED4 を用意した．ED1 は評価用のデータで ATR 自然発話音声言語データベースの一部である音声翻訳研究用 ATR 音声言語データベース [38] の 9 会話 (通常の 2 人による会話を，話者の役割 (ホテル側／客) 毎に区別してそれぞれを「片側会話」と呼ぶことにすると，18 片側会話) である．そして，音声翻訳研究用 ATR 音声言語データベース [38] の残り (609 会話 (1,218 片側会話)) はモデルの学習用のデータ ED4 とした．

次に，音声入力に対する発話の分割性能を評価するためにデータセット ED2 と ED3 とを用意した．ED2 は音声認識性能の評価用のデータで，提案の言語モデルに句点を含めることによって句点以外の他の単語の認識性能への影響を評価するために利用する．ED2 は ED1 とは別の 42 片側会話である．また，このデータの話者は音響モデルの学習データには含まれていない．ED3 は，前述の ED2 以外の ATR 自然発話音声言語データベースの残りのデータであり，言語モデルの学習に用いられる．音声認識用の言語モデルの学習にはできるだけ多くのデータが必要であるため，このデータセットを用意した．なお，ED4 は完全に ED3 に含まれている．

以上のデータセットのそれぞれの片側会話数，発話数 (発話末の句点の数はこれに一致)，発話内の句点の総数，および，総単語数を**表 5.2** に示す．

5.4.2　評価実験の設定

次に、実験で用いる言語モデル，音声認識時の探索，音響モデル，音響分析条件について説明する．

評価対象の言語モデルと音声認識時の探索

本実験では，句点を単語に含めた統計的言語モデルとして，句点を単語に含めた単語 3gram と単語 4gram，および，多重クラス複合 2gram[72] を用いる．多重クラス複合 2gram は，音声認識の探索の第 1 パス内での計算コストの高い高次の言語モデル (単語 Ngram ($N \geq 3$)) の利用を避けるために用いる．

多重クラス複合 2gram では次式で単語の予測確率が計算される．

$$P(w_i|w_{i-1}) = P(w_i|C^t_{w_i})P(C^t_{w_i}|C^f_{w_{i-1}}) \tag{5.1}$$

ここで，w_i や w_{i-1} は単語，または単語系列である．$C^t_{w_i}$ は w_i が属するクラス，$C^f_{w_{i-1}}$ は w_{i-1} が属するクラスである．右辺第 1 項はクラスから単語または単語系列が出現する確率，右辺第 2 項は発声の頭側のクラス $C^f_{w_{i-1}}$ から 末尾側のクラス $C^t_{w_i}$ への遷移確率である．そして，発話中の句点 (x) は，C^t_x は発話終了記号と同じクラスとして登録し，C^f_x は発話開始記号と同じクラスとして登録した．その他の点では多重クラス複合 2gram[72] と同じ方法で，単語の自動クラスタリング，単語系列抽出，パラメータ推定を行なった．

全てのモデルにおいて，発話中の句点は発話末の句点とは別の単語として登録した．本実験では，最短 20 ミリ秒のポーズと同じ発音記号を句点に付与した．

次に音声認識時の探索の方法であるが，発話中の句点が発話開始記号や発話終了記号とは別の単語として登録されているので，従来の認識システムをそのまま用いることができる．本実験では ATRSPREC[39] を用いる．

前節で述べた学習データと上記のモデルとの組合わせを変化させることにより数種の言語モデルを作成して，以下のような評価実験を行なった．

まず，句点を含まないテキスト入力に対する分割性能の評価のために，ED4 を学習データとして用いた発話中の句点を含む言語モデルとして，単語 3gram(3-SPLT-S) と単語 4gram(4-SPLT-S) と多重クラス複合 2gram(M-SPLT-S) との 3 種類のモデルを作成した．全モデルの語彙サイズは約 5,600 で，M-SPLT-S のクラス数は 700 とした．また，獲得された単語系列数はおよそ 960 であった．

続いて，音声入力に対する性能を確認する目的で，発話中の句点を含む多重クラス複合 2gram(M-SPLT) を作成した．さらに，M-SPLT が句点以外の単語の認識性能において劣化がないかどうかを確認するために，発話中の句点を含まない多重クラス複合 2gram(M-BASE) を作成した．これら 2 つのモデルの学習には，ED3 の全部と ED1 の半分のデータを用いた．評価には ED1 と ED2 とを用いる．M-SPLT と M-BASE との間で語彙サイズは同等で約 14,000 であり，句点を含むか否かだけが異なる．また，多重クラス複合 2gram

表 5.3: 発話分割評価実験で用いる言語モデルと学習及び評価データの組合せ

モデル名	学習データ	評価データ	モデルタイプ	句点	実験対象
3-SPLT-S	ED4	ED1	単語 3gram	有	テキスト
4-SPLT-S	ED4	ED1	単語 4gram	有	テキスト
M-SPLT-S	ED4	ED1	多重クラス複合 2gram	有	テキスト
M-SPLT	ED3，ED1 の半分	ED1, ED2	多重クラス複合 2gram	有	音声
M-BASE	ED3，ED1 の半分	ED1, ED2	多重クラス複合 2gram	無	音声
3-SPLT	ED3，ED1 の半分	ED1	単語 3gram	有	音声
4-SPLT	ED3，ED1 の半分	ED1	単語 4gram	有	音声

の M-SPLT と M-BASE の，クラス数は 700 クラスのモデルを採用した．また単語系列数はおよそ 4,700 であった．

さらに，音声入力に対する分割性能において，発話中の句点を含む多重クラス複合 2gram(M-SPLT) と他の単語 Ngram モデルとを比較するために，M-SPLT の学習で用いたものと同じ学習データ (ED3 の全部と ED1 の半分) を用いて，発話中の句点を含む単語 3gram(3-SPLT) と同単語 4gram(4-SPLT) とを作成した．評価データには ED1 を用いる．単語 3gram と単語 4gram に関しては M-SPLT で得られた単語グラフをリスコアリングすることにより評価した．

以上の言語モデルとその学習データと評価データとをまとめて **表 5.3** に示す．

また本認識実験で用いられる辞書内の各単語の発音記号の末尾には，一部の動詞の活用形の末尾 (例えば，動詞 “頂く” の未然形，“頂か” の末尾など) を除いて，ポーズが選択的に付与されている．

音響モデルと音響分析条件

音声認識実験で用いられる音響モデルの設定は，**表 5.4** の通りで，ATR 自然発話音声言語データベースに含まれる音声で学習されたモデルである．

表 5.4: 音響モデル学習条件

音響分析	
	サンプリング周波数 16kHz，preemphasis 0.98
	フレーム周期 10 ms，フレーム長 20 ms (Hamming 窓)
	logpower，Δlogpower，12 次-MFCC，12 次-ΔMFCC
	ケプストラム平均，パワーを正規化
音響モデル	
	男性モデル：音声 1400 状態 5 混合，無音 3 状態 10 混合
	(学習データ 167 話者，総発話時間 約 2 時間)
	女性モデル：音声 1400 状態 15 混合，無音 3 状態 10 混合
	(学習データ 240 話者，総発話時間 約 3 時間)

5.4.3　テキスト入力に対する分割性能の評価

句点なしのテキストを入力として文への分割実験を行なった．

本実験では，評価データの中の話者が交替するまでの間の形態素解析済の単語列 (図 5.1 の turn1 などの単語列に相当) を対象として，単語既知，しかし，句点未知のもとで，隣り合う全ての単語間に句点が入る場合までを想定して単語グラフを構成し，これに本言語モデルを使って言語尤度付けを行ない，そのうちの最も尤度が高いパスを選択した (図 5.2)．上のように句点が入りうる全ての可能性を想定することにより，句点の認識に必要な音響的要素であるポーズが与えられている場合を仮定した上での分割性能の評価を行なった．

ここでは，ED4 を用いて作成された言語モデル (3-SPLT-S，4-SPLT-S，M-SPLT-S) を使って，ED1 のテキストを評価対象の入力として，実験を行なった．評価対象の句点は，話者が交替するまでの区間に話したすべての文の境界を示す句点のうち，話者が交替する直前の最後の句点 (図 5.1 の turn1 の右端の句点など) を除いた句点 (図 5.1 の sent1，sent2 の各末尾の句点など) とした．これに該当する句点は ED1 の中に 123 個有った．

実験結果を**表 5.5** に示す．数値は左から再現率，適合率，およびこれらの調和平均である F 値の順で書かれている．F 値は

$$\frac{2.0 \times \text{再現率} \times \text{適合率}}{\text{再現率} + \text{適合率}} \tag{5.2}$$

で計算した．

テキスト入力は，音声入力において発話内の全単語が認識され，句点の音響的条件が満たされ，言語的尤度によって句点が採用されるか否かを決めるだけが残された状態を模擬している．この状況では，単語ベースの統計的言語モデルである単語 3gram(3-SPLT-S)

表 5.5: テキスト入力に対する発話分割の性能

方式	再現率 [%]	適合率 [%]	F 値
3-SPLT-S	77.9	88.8	83.0
4-SPLT-S	76.2	89.4	82.3
M-SPLT-S	86.9	46.5	60.6

および単語 4gram(4-SPLT-S) が F 値の点で高い性能を示した．多重クラス複合 2gram(M-SPLT-S) では，再現率は高いが，適合率が低かった．

単語ベースの言語モデル (3-SPLT-S や 4-SPLT-S) では，固有の単語列に引き続き句点が現れる場合を直接表現できるため，精度が高く適合率が高いが，学習データに現れた組合わせしか記憶されないためスパースネスからカバレッジは低く再現率が低くなる．一方，クラス Ngram に基礎をおく多重クラス複合 2gram では，精度は低いもののカバレッジは高いため，単語 Ngram の場合とは逆の傾向となる．このため表 5.5 のような結果になったと考えられる．

また，この結果から，探索ではまず多重クラス複合 2gram を用い，句点の候補が得られたならば、局所的に単語 4gram を用いることにより，高い再現率と適合率を得ることが期待できる．

以上の実験では，話者が交替するまで (図 5.1 の turn1 などの右端まで) を単位として処理を行なった．一方，音声認識では，話者が次の話者へ交替するまでの区間を単位とするのではなく，それよりも小さな単位である発話 (図 5.1 の utt1 など) を単位としている．それは，比較的長いポーズが入る箇所で，話者が次の話者へ交替するまでの区間 (図 5.1 の turn) が複数に分けられた区間である．この対話における発話境界，または発話という音声区間の開始と終了は高い割合で正しく認識することが可能である [73].

また，この発話と発話の境界の認識に失敗した場合，例えば，図 5.1 で文 sent2 と文 sent3 とが別の発話 (utt) に分かれているが実は 1 つの文であるような場合，つまり，音声区間の切り出しによって元々 1 つの文が複数の発話に分割されている場合でも，本モデルによって誤って認定された発話末が文末であるかどうかを，すなわち，その位置に句点が来るかどうかを高い精度で判定できる．それが可能であることは，話者が交替するまでの区間を単位として，句点を含まないテキストを入力とした場合の句点の認識結果である表 5.5 から確認されている．

このため，以下では，発話を入力単位として発話の中の文境界としての句点 (図 5.1 の target) を探索する．以下では，上記の評価対象であった句点のうちの発話の末尾の句点を除き，発話内の句点のみ (図 5.1 の target) を対象として実験を行なう．

表 5.6: 音声入力に対する発話分割の性能

	再現率 [%]	適合率 [%]	F 値
3-SPLT	64.4	100.0	78.3
4-SPLT	94.5	100.0	97.2
M-SPLT	78.1	90.5	83.8

5.4.4　音声入力に対する分割性能の評価

できるだけ多くのデータを音声認識の言語モデルの学習に用いることによって，認識精度を高くできる見込みがある．そのため ED3 と ED1 の半分を学習データとして作られた言語モデルである M-SPLT を用いた．そして，分割の評価データに ED1 を用いて発話を入力単位とした音声を入力として，音声認識過程での発話分割実験を行なった．さらに，M-SPLT を使った音声認識で得られた単語グラフのリスコアリングを，単語 3gram(3-SPLT) と単語 4gram（4-SPLT）とを用いてそれぞれ行なうことによって，単語 3gram(3-SPLT) と単語 4gram（4-SPLT）とを用いた音声認識過程での分割を模擬した．

ここでは発話内の句点が対象であり，その総数は 73 であった．分割の評価については，認識結果の第 1 位候補での句点の再現率と適合率と F 値の観点から評価を行なった．

結果を**表 5.6** に示す．表 5.6 から，単語 3gram よりも単語 4gram のほうが F 値の点で分割性能が高かった．単語 Ngram の N が大きいモデルの方の性能が高い結果となった．

発話内の句点に関しては多様性が少ないため再現率が高くなり，単語 4gram での結果が他の 2 つのモデルよりも良くなったと考えられる．また，多重クラス複合 2gram では，やはり適合率が単語 3gram や単語 4gram に比べて低くなったけれども，辞書に登録された単語系列によって，単語 3gram よりも再現率の高い結果となったと考えられる．

5.4.3 の後半で，句点を含まないテキスト入力に対して，まず多重クラス複合 2gram，次に単語 4gram を利用することによって性能向上を期待したように，M-SPLT の出力結果を 4-SPLT でリスコアリングすることによって再現率 94.5%，適合率 100.0%という結果が得られた (表 5.6)．

分割誤り

ED1 の音声入力に対する M-SPLT を使った分割結果のうち，分割誤りの事例の幾つかを挙げる．

削除誤り（分割漏れ）には例 2 のような事例があった.「×」が分割位置であるにも関わらず正しく分割されなかった分割位置である.

例 2

削除誤： ‥ 申し訳ございません × シングルは ‥

削除誤： 東京シティーホテル御滞在 × 零三の …

削除誤： 調べます × しばらくお待ち下さい

“申し訳ございません” の後にはポーズが認識されていなかった．次の単語である “シングル” の頭の音 /sh/ に句点の発音が吸収されたか，または 20 ミリ秒以下のポーズであったため捉えられなかったと考えられる.

“御滞在” のような体言止めの例は少ない恐れがある．そのため，その語の後の句点に対して多重クラス複合 2gram から与えられる言語尤度が高くなかったことが考えられる．本実験用の認識用の辞書の各単語の発音記号の末尾には，動詞の一部の活用形以外にはポーズが選択的に付与されている．句点の言語尤度が十分に高くない場合，句点のポーズが単語の末尾のポーズとして単語に吸収されてしまうことが考えられる．実際に、“御滞在” の後の句点の削除誤りでは，句点の直前の単語 (御滞在) の方に，ポーズ (615 ミリ秒) が吸収されてしまっており，句点の発音としてのポーズが認識されていなかった.

“調べます” の後でも，直前の “ます” の方にポーズ (170 ミリ秒) が吸収されてしまっていた.

一方，ポーズが短い場合でも，句点の認識ができているものもあった．このため，句点を認識できるか否かはポーズの長短に必ずしも依存するものではなく，句点に与えられる言語尤度が重要と考えられる.

同様に，挿入誤り（過分割）には，例 3 のような事例があった.「※」が誤って挿入された分割位置を示す.

例 3

挿入誤： そうですか※料金はそれぞれおいくら ….

挿入誤： そうですか※じゃバス付の方でお願いしたい ‥

データベース内の「そうですね」や「そうですか」は話し始めの決まり文句に相当し，それらの後の位置には，句点ではなく読点がおかれているが，一般に終助詞のあとに句点が置かれることが多い．その結果挿入誤りとなったと考えられる．これは M-SPLT のようなバイグラムでは正しく捉えることは困難である.

表 5.7: 句点以外の単語認識率 [%]

	ED1	ED2
M-SPLT	92.9	85.6
M-BASE	93.1	85.3

単語認識率の比較

句点を組み込まない従来の言語モデルを句点を組み込んだ言語モデルに変更することによって，句点以外の単語の認識性能の劣化がないかどうかを確認するために，M-SPLT と M-BASE を使った場合の認識率の比較を行なう．これまで用いてきた評価データ ED1 に加えて，主に音声認識の評価用として使われるデータである ED2 に対する実験も行なった．音声認識結果の第 1 位候補での単語認識率（%Accuracy）を**表 5.7** に示す．表 5.7 の「M-SPLT」の行は，M-BASE と比べるために算出した，認識結果の第 1 位候補と正解との間での句点以外の単語の認識率である．表 5.7 のように，M-SPLT と M-BASE とのそれぞれの結果の比較によれば，M-SPLT モデルは M-BASE と句点以外の単語の認識性能においてほとんど違いが無く，句点を学習することによる性能の劣化は無い．

5.5　評価実験結果のまとめ

5.5.1　分割の観点から

句点を含まないテキストの分割においては，オープンな評価実験の結果 (表 5.5)，単語 3gram(3-SPLT-S) および単語 4gram(4-SPLT-S) が F 値の点で高い性能を示した．多重クラス複合 2gram(M-SPLT-S) では，再現率は高いが，適合率が低かった（単語 3gram との間で，再現率は 5%水準で、適合率は 1%水準で有意）.

音声入力に対する分割においては，評価実験の結果 (表 5.6)，多重クラス複合 2gram(M-SPLT) では，8 割近くの再現率と 9 割の適合率が得られた．多重クラス複合 2gram(M-SPLT) を使った場合の分割の削除誤り (例 2) や挿入誤り (例 3) の事例の分析から，句点の認識には句点に与えられる言語尤度が重要であると考えられる．また，多重クラス複合 2gram では，話し始めの決まり文句の末尾の終助詞の後に，句点が来る確率を正しく推定することが困難であり，実際よりも大きな確率が与られ挿入誤りが生じたと考えられる．

M-SPLT の単語グラフのリスコアリングによって，単語 3gram(3-SPLT) や単語 4gram(4-SPLT) を用いることにより，認識過程の第 1 パスで計算コストの高い高次 Ngram($N \geq 3$) の使用を避けながら，単語 4gram で最高 94%の再現率と 100%の適合率 (F 値で 97%) と

いう再現率と適合率とがともに高い結果が得られた（M-SPLT からの改善は，再現率は 1%水準で、適合率は 5%水準で有意）.

5.5.2　音声認識率の観点から

句点以外の単語認識率の点では，表 5.7 のように，句点を含むモデルと含まないモデルとの間での性能の差は無い．すなわち，句点を言語モデルに含めることによる句点以外の単語の音声認識への悪影響は無い（検定の結果，句点を含める前後では有意差が無いことを確認した）.

5.6　5 章のおわりに

本稿では，発話内の文境界の記号としての句点を他の単語と同様に音声認識することによって発話の分割を行なうための統計的言語モデルを提案し，音声対話データでの分割性能の評価と音声認識性能の評価とを行なった．従来の音声認識のための統計的言語モデルとの違いは，句点を単語として扱うこと，句点にはポーズの発音を与えることだけであるが，その結果,

- 対話音声に対して高い再現率と適合率で発話の分割を行なえること
- 句点以外の単語の音声認識性能を劣化させないこと

が確認された．また，本方法には

- 音声認識の過程で句点を認識し，そこで発話の分割を行なうので，リアルタイムで認識結果を出力でき，早期に後段の言語処理に多くの情報を出力できること
- 分割のための情報が音声認識の統計的言語モデルつまり Ngram の枠組みに統一されているので，モデルの維持管理が容易であること

という利点があり，有効なモデルとなっている.

第6章　指示語と指示動作との間での指示対象物情報の統一

6章では，音声認識された指示語と指示動作によって指示された対象物との対応付けにより，対象物についての情報を統一化するという課題の解決を行なう．

6.1　はじめに

計算機GUIの構築においては，人間の日常的コミュニケーションに近い自然なインタフェースの構築が重要であると指摘されている[33]．人間同士の日常のコミュニケーションでは，特に装置や地図のような物を人間同士の間に挟んで行なう会話では，指示語を含んだ音声が発声され，それと並行して装置や地図などへの指先等を使った指示動作が行われる[74]．この場合，音声中の指示語が参照する物の同定は話し手の指示動作に基づいて行なわれ，話し手から聞き手に意図が伝達される．このような音声と指示動作といった複数のモードを同時に使った会話（マルチモーダル会話）は，人間にとって自然である．故に，このような会話を扱えることが自然なインタフェースの構築のための1つの要件となっており[68]，マルチモーダルインタフェースの研究で盛んに取り上げられている[4, 6, 20, 58]．

我々の研究の目的は，人間が人間に話す場合のような自然なマルチモーダル会話の様式を，スムーズにGUI上での直接操作と融合させることにある．従来のマルチモーダルインタフェースの研究では，計算機ディスプレイ上の対象への指示動作としてマウスクリックやパネルへの接触が用いられていた[4, 20]．ところが，これらの指示動作の様式は，従来から存在するGUIの操作の慣習の１つである，GUIのボタンを押す為の直接操作と同じ様式である．そのため，これらの指示動作の様式を単純にGUIに導入すれば，指示動作と直接操作との間での曖昧性（単に指さしたのか押したのかの曖昧）が生じる．このような曖昧性の存在は望ましくない．

本研究では， 曖昧性が生じないようにするために，直接操作とは異なり，自然でしかも従来の操作慣習と矛盾のない指示動作の様式として，クリックを使わずにマウスを動かすだけという様式を採用した．このように本研究では，指示動作として従来のマウスクリックやパネルへの接触のような明確な様式を用いない．そこで，マウスの動きから得ら

れる情報と指示語を含む音声から得られる情報とに基づいて，指示された対象の候補を抽出し，音声中の指示語と指示された対象とを対応付けるマルチモーダル入力方法を実現した．そして，本方法により指示語と指示対象との対応付けにおいて高い精度が得られることを確認した．これらを GUI に導入する結果，ボタンなどの対象への直接操作は従来の様式のままで操作でき，さらにマウスを動かすだけの様式を使うことにより指示動作と音声とを併用した自然なマルチモーダル入力が可能となる．さらに本研究では，これらを現在利用者が急増し操作法が広く知られている GUI である WWW ブラウザへの入力部分に適用し，WWW ブラウザとのインタラクションのマルチモーダル化を行なった．

WWW ブラウザは World Wide Web 上の HTML 文書を表示するソフトウエアで，そのページ上には別の HTML 文書へのリンクを表すアンカーが一般に表示される．アンカーをクリックすれば，そのアンカーが示すファイルの中味が WWW ブラウザ 1 つのページ上に複数のアンカーが表示されていることは一般的で，それらの複数のアンカーに対して同じ命令（ファイルの取得やアンカーの登録）を繰り返す場合がある．そのような複数のアンカーに対して同じ操作を繰り返す代わりに，指示動作と「これとこれを〜して」という音声とを併用しマルチモーダルに入力できれば，ユーザと WWW ブラウザとの間でのインタラクションが自然で柔軟なインタラクションになると考えられる．

本論文では，マウスと音声とを使ったマルチモーダル入力について述べる．**6.2** では，GUI の直接操作との曖昧性を避けるために，（クリックを使わずに）マウスを動かすことだけによって行われる指示動作を導入する理由を説明する．**6.3** で，この指示動作によって指された指示対象と音声中の指示語とを対応付ける方法を説明し，その方法の有効性を確認するために行なった実験について述べる．そして，本方法の適用によってユーザから WWW ブラウザへのマルチモーダル入力を可能にするシステムと，WWW ブラウザ上での対応づけの精度とを **6.4** で述べる．**6.5** では，ユーザとシステムとのインタラクションの事例を挙げ，従来と本システムとの間でのインタラクションの比較を行なう．また関連研究との比較を行なう．

6.2　クリックを使わない指示動作の導入

6.2.1　人間の指示動作と直接操作

人間の日常のコミュニケーションに見られる指示動作と直接操作は，以下に述べるように本来異なるものである．

指示動作は対象を「選択」するために機能する．選択された対象に対して何を行うか（以後,「命令」と呼ぶ）は音声で伝えられる．音声には命令のための言葉以外に指示語が含まれる．例えば，飲食店などで客がメニュー上の商品を指し示しながら「これとこれを

下さい」と店員に向かって言う会話がある．このように指示動作と指示語を使って対象の「選択」が表現され（「これとこれを」），それらの対象に対して何を行うかという「命令」はその前後の音声（「下さい」）によって表現される．以上のように分離している「選択」と「命令」とのさまざまな組み合せを，音声と指示動作とを使って作り出すことによって，自然で柔軟なマルチモーダル会話が行なわれている．

また，この場合の指示動作は指先等のポインタを使って行なわれる．対象を指すとき，そのポインタは，しばしば（１）対象の上に重ねられたり，（２）対象を囲むように動かされたり，（３）対象の下側で動かされる．このようにして，指示動作は対象を「選択」するために機能する．そして，指示動作自体は「命令」の実行の意味を持たない．

ところが，もう一方のボタンを押すなどの直接操作は，対象の「選択」と対象に対して固定された「命令」の実行の２つ機能をもつ．例えば，自動販売機で或る商品のボタンを押すという操作は，（別の商品ではなく）その商品を選択することと，その商品を取り出すという命令とが 1 度に実行されることを意味する．

このように，人間の日常のコミュニケーションにおける指示動作は「選択」だけの機能を持ち，直接操作は「選択」と「命令」の実行の 2 つの機能を持ち，両者は異なるものである．

6.2.2　設計方針

人間の日常のコミュニケーションにおいて異なる指示動作と直接操作の両方に，GUI 上での同じ操作様式を割り当てると，その操作の解釈において曖昧性が生じる．例えば，実世界の物体を模した GUI 上のボタンは，直接操作と指示動作の両方の対象になりうる．すなわち，処理の終了を待って逐一ボタンを押すという直接操作の繰り返しの代わりに，ボタンを指示しながら「これとこれを実行して」とマルチモーダルに命令する場合に，ボタンは指示対象になる．このような対象への指示に，直接操作と同じマウスのシングルクリック（従来研究での指示動作）を用いると，その動作が指示動作なのかボタンを押すという直接操作なのかが曖昧となる．この曖昧性を回避しマルチモーダル入力を実現するためには，人間の動作のように対象の「選択」だけが可能な様式を指示動作として導入することが必要となる [1]．

本論文では GUI への指示動作と直接操作にマウスを使う場合について議論している．この場合，実世界の物体を模した GUI 上の「ボタンを押す」という直接操作には，実世界での動作と対応のあるシングルクリックが自然である．また，実世界に物がなく対応が取れないが，WWW ブラウザのアンカーはボタンのように機能する．“WWW ブラウザ上

[1] 選択と命令の分離という点では，シングルクリックで対象を選択し，メニューから命令を選んで実行するという方法が存在する．しかし，この方法はボタンのような対象の指示には使えない．

でアンカーをクリックしてページを変更する” ことは公知の操作になっている．実際，シングルクリックでのページの変更はブラウザの種類を問わず共通な操作である．そのため，WWW ブラウザではアンカーへの直接操作として従来からのシングルクリックが自然である．このように，ボタンのように機能する対象への直接操作にも従来からのシングルクリックが自然である．

従って，直接操作にシングルクリックを利用し，指示動作にはそれ以外の様式を割り当てることが自然である．本研究では，その為の様式として,（クリックを用いずに）マウスの動きだけによる指示動作を採用した．この種の指示動作は前節で述べた人間の動作のうちの最初のもの，つまり対象の真上にポインタを重ねること（（１）の動作）に相当し，自然と考えられるからである．

6.3　指示語と指示対象との対応づけ

6.3.1　方法

指示動作にはクリックを用いない．そのため，指示動作の為に行なわれた一連のマウスの動きから指示された対象を判定する．本方法では，マウスカーソルが対象の領域に入ってから出るまでの間に，マウスカーソルの移動速度が予め決めておいた閾値よりも遅くなった場合に，その対象が指示された可能性があると判断され，その対象の領域に入ってから出るまでの時間帯が，対象を指示していた時間帯（今後，指示時間帯と呼ぶ）として抽出される．マウスカーソルが閾値以上の速度で通過した対象は，指示されなかったものと判断され抽出されない．一方，音声にもその語が発声されている時間帯（発声時間帯と呼ぶ）が存在するので，これを抽出する．

本方法では，音声入力と指示動作との時間軸上での同期に基づいて指示語と指示された対象とを対応付ける．しかし，文献 [30] でもデータが示されているように，指示時間帯と発声時間帯のどちらか一方が完全に他方に含まれるような，2 つの時間帯の間での完全に近い同期は期待できない．

そこで本方法では，発声時間帯と指示時間帯の 2 種類の時間帯が重なる時間帯の長さを測り，指示語の発声時間帯ともっとも長い重なりを作る指示時間帯を持つ（GUI 上の）対象を指示された対象であると判断し，指示語と対応付ける．この様子を**図 6.1** に示す．この図 6.1 では，マウスカーソルの移動速度が閾値以下になり obj-1 と obj-2 とが抽出され，それと並行して発声された「これと」が認識された場合の対応づけを説明する．マウスカーソルが指示対象の領域に入った時刻が指示開始時刻で，その対象から出た時刻が指示終了時刻である．その 2 つの時刻の間が指示時間帯である．また，発声開始時刻から発声終了時刻までが発声時間帯である.「これと」の発声時間帯は，obj-1 の指示時間帯と t1

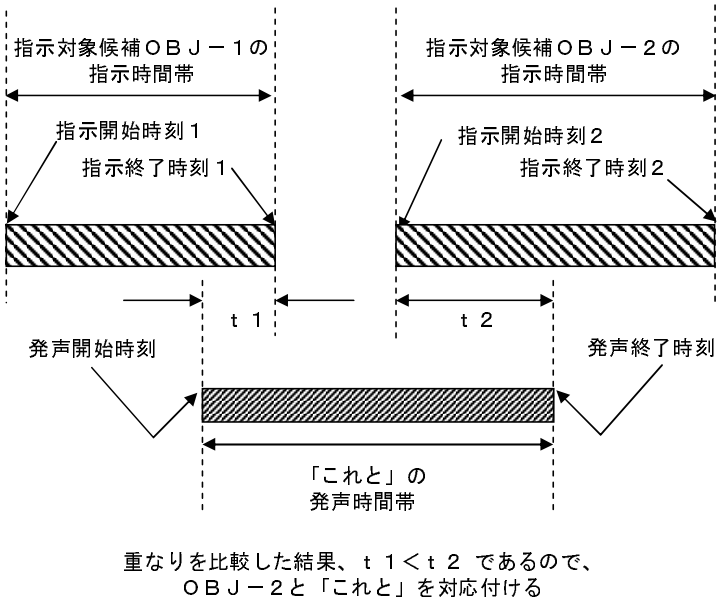

図 6.1: 指示語と指示対象との対応付けアルゴリズム

の長さの重なりを作り，obj-2 の指示時間帯と t2 の長さの重なりを作る．これら t1 と t2 の長さを比較した結果，t2 の方が長いので,「これと」という指示語は obj-2 と対応づけられる．

この入力方法ではユーザが指示対象の上でポインタの速度を遅くすることを仮定している．また音声面では，ユーザが 1 つの発声時間帯に指示語を 2 つ以上含まないように発声することを仮定している[2]．後者の仮定は，従来の研究 [6] で指示語を含む音声と音声の間にポーズがおかれている [54] ことに相当する．

6.3.2 対応づけの精度

本入力方法が自然であることと，指示語と指示動作とを対応づける本方法が有効であることとを確かめるために，ワークステーションに実験環境を設定し，被験者 5 人による簡単な実験を行なった．

1 辺が 1.5cm の同色の正方形を 9 行 9 列の格子状に配置した．正方形相互の間隔を 1cm とした．そのうちの 3 つをランダムに選び，それぞれ異なる色をつけた．これを 10 パターン用意した．各パターンの中の 3 つの正方形を，指示動作によって指し示す正方形とした．被験者に，色の付いた 3 つの正方形を一定の色の順番で指し示しながら,「これとこれを

[2]1 回の発声時間帯だけを知ることができる音声認識装置を利用するという前提のもとで必要となった仮定である．1 回の発声に含まれる指示語等の部分の発声時間帯を得ることができる音声認識装置を用いることができる場合にはこの仮定は必要ではない．

表 6.1: 指示語と指示動作との対応づけの精度 [%]

	実験 1	実験 2
全発声	79.3	86.7
文節発声	86.4	93.9

ここに移動する」と発声させた．ただし，最初（以後，実験 1 と呼ぶ）は発声や指示動作のためのマウスの動かし方について特に教示しなかった．マウスカーソルの形状は丸型に設定した．

以上の入力実験の後，指示動作では指示対象の領域にマウスカーソルを入れて指示を行なうこと，指示と移動との間でマウスカーソルの速度に違いをつけること，及び，音声と動作とのタイミングを合わせることを被験者に教示した．その後，被験者は実験 1 と同じ数だけ入力を行なった（以後，実験 2 と呼ぶ）．

指示された対象と指示語との対応づけの精度を**表 6.1** に示す．

指示動作の回数は，両実験ともに，1 人あたり 30 回 (=3 × 10)，5 人で 150 回 (=30 × 5) である．実験 1 では，これらの指示動作のうちのほとんどが対象の領域にカーソルを入れる動作であった．クリックしたり，丸く囲むような動作は同じ被験者の中でも数回しかなかった．全 150 回のうち 79.3%が正しく対応づけられた．これは，すなわち発声とマウスの動かし方のどちらにも教示を行わない場合の精度である．

また，音声が切って発声された（以後，文節発声と呼ぶ）のは 150 回の指示動作のうち 126 回であった．この 126 回のうち 86.4%が正しく対応づけられた．この結果から，文節発声されれば，特に教示を行わないマウス動作のうちの 86.4%を本手法で対応づけることができる．

実験 2 の場合，全 150 回のうち 86.7%が正しく対応づけられた．教示の結果，7.4 ポイント (=86.7-79.3) の向上が見られた（統計的に有意傾向があると確認された）．

教示の結果，文節発声の数が増え，文節発声されたのは 150 回の指示動作のうち 132 回であった．この 132 回のうち 93.9%が正しく対応づけられた（5%水準で有意）．

5 人の被験者のうち 2 名はワークステーションを，特に実験に用いたマウスと同じマウスを，毎日使っていた．この 2 人の精度はほぼ 100%であった．

以上の結果から，本手法の有効性が確認できた．

また，本方法によって何も教示を行なっていない入力の 8 割近くを正しく対応づけることができる．この高い精度から，何も教示を行なわなかった場合の指示動作でも，指示対象の上でマウスのカーソルが遅くなるという傾向があり，他方の音声には「これと」「これを」「ここに移動する」のように切って，しかも指示動作とある程度の同期をとって話す傾向があるといえる．従って，これらの傾向を仮定した本方法によって対応づけることができる入力は，おおむね自然な入力方法であるといえる．

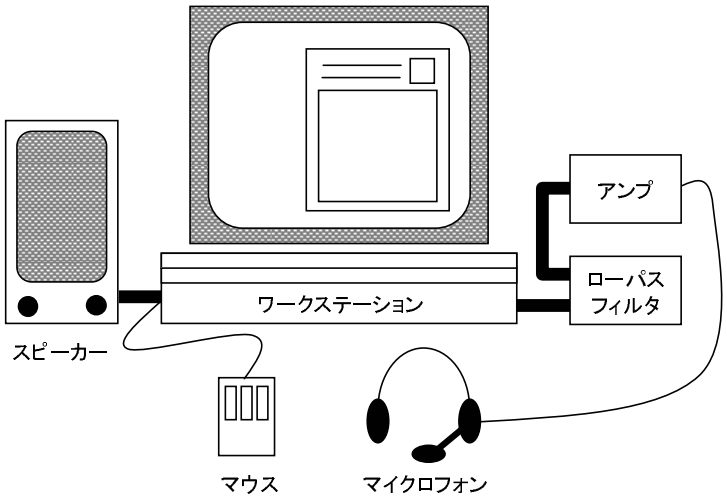

図 6.2: マルチモーダル制御システムの全体の構成

6.4　WWW ブラウザのマルチモーダル制御システムの構成

WWW ブラウザ上でユーザが指示した対象とその指示動作に伴って発声された指示語とを，前節の方法に基づいて対応づけるシステムの実装について説明する．

6.4.1　全体の構成

システムを構成する装置の概略を**図 6.2** に示す．

本システムは，1 台のワークステーション，マイク，アンプ，ローパスフィルタ，及びスピーカーで構成されている．今回の実装には，クライアントサーバ型の音声認識システム，およびテキストからの音声合成システムとして，それぞれ NTT ヒューマンインタフェース研究所が開発した音声認識システム [71] と音声合成システム [10] を用いた．また，WWW ブラウザとして NCSA の Mosaic[42] を用いた．本システムは市販のワークステーション（HP-9000-C110）上で動作している．

システムを構成するプロセスとそれらの相互関係を**図 6.3** に示す．

3 節で述べたように音声認識結果と発声時間帯が必要となる．本システムでは，音声が入力される毎に，音声認識プロセスによって，その音声の発声開始時刻，その音声の発声終了時刻，音声認識結果の文字列，及び，その認識結果の属性からなる 4 つ組が作成され，キュー 1 に格納される．また，指示対象候補と指示時間帯が必要となるが，本システムでは，マウスの動きから，ある対象が指示された対象の候補であると判定された場合には，マウスカーソルがその対象に入った時刻，出た時刻，及び，その対象の名前からなる 3 つ組が作成され，キュー 2 に格納される．MM 入力解析プロセスは，キュー 1 の音声認識結果に指示語が含まれている場合に，その指示語とキュー 2 の中の指示対象候補とを対

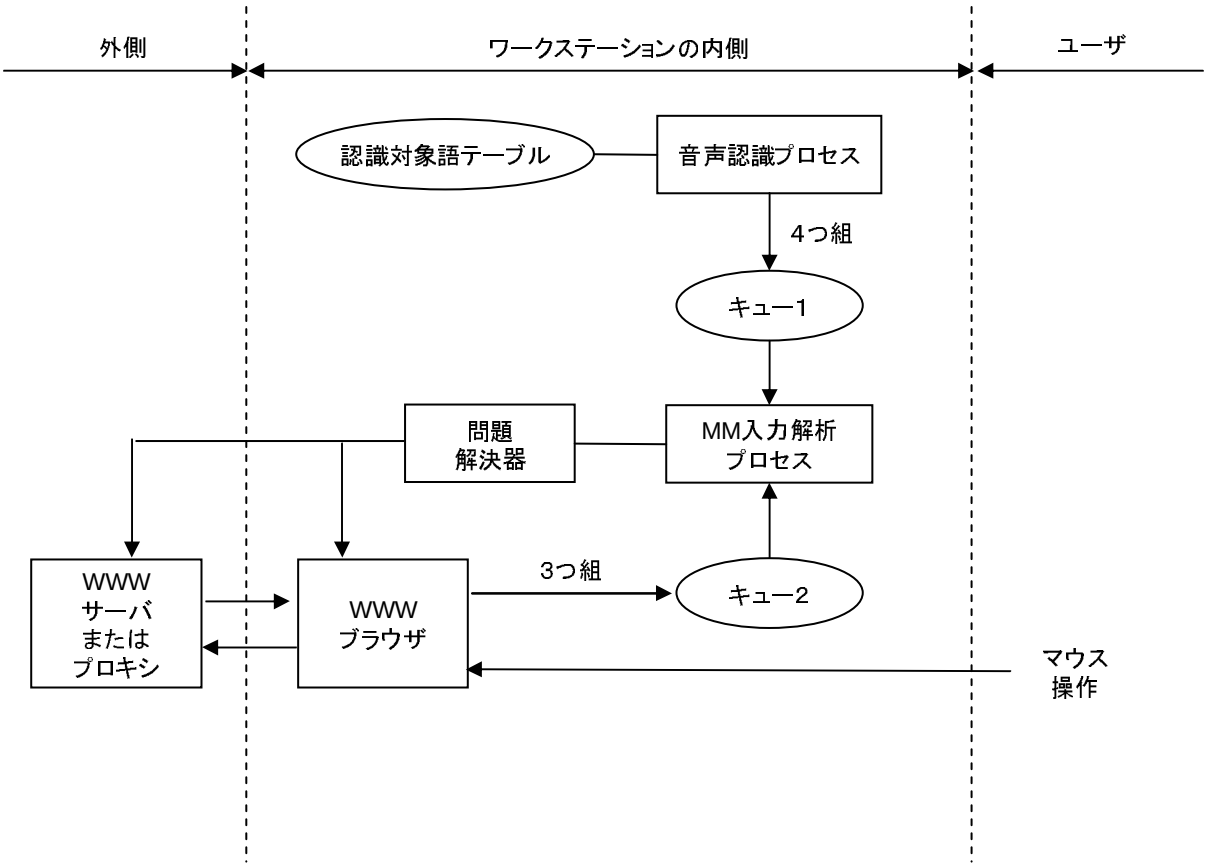

図 6.3: システムを構成する全プロセス

応付ける．問題解決器は，ユーザが入力した命令の内容に応じてブラウザにアクセスし，ユーザからの命令を実現する．

本システムでは指示動作と音声とを使ったマルチモーダル入力が可能である（詳細は 6.4.5 の 1 から 3）．また，一般に WWW ブラウザでは，メニューをプルダウンするか，その代わりにキーボードから数文字（ホットキーまたはアクセラレータキー [3]）を入力することによってコマンドを選択することができる．このコマンド選択を，従来の研究と同様に，音声によって行なえるよう実装した（詳細は 6.4.5 の 4）．ページを変える操作は，本システムではアンカー上でクリックを行なうという通常の操作により実行できる．

この他，何がマルチモーダル入力によって選択されたかをユーザに示すための確認用のウィンドウの提示機能も実装されている．このウィンドウには，指示された対象ごとに，チェックボタン，選択されたファイル名，セーブする場合のファイル名（ユーザによる書き換え可能）が横一行に表示される．ウィンドウの一番下には，命令全体をキャンセルす

[3]WWW ブラウザ上にマウスカーソルが置かれた状態で，キーボードの数個のキーを打つことによって，プルダウンメニュー内のコマンドの実行やページの前後移動が可能となっている．これらのキーがホットキーやアクセラレータキーと呼ばれる．

表 6.2: 認識対象語句とその語句に与えた属性値との対応関係の例

認識対象語句	属性値
これと	R1
これを取ってきて	V1
検索して下さい	V0

るか実行するかの確認用のボタンがそれぞれ表示される．指示動作からは指示された対象の名前として “*http* : //･･･” などの URL が取り出されるが，これはユーザにとって確認しにくい．これを避けるため，その時 WWW ブラウザに表示されている HTML 文書を CCI[41] を用いて取得し，URL と対になっているアンカー文字列をファイル名として表示し，確認を促すように実装されている．指示されたと誤って認識されたファイルは，チェックボタンをチェックすることで命令の適用対象から外される．

6.4.2　音声の処理:4 つ組の作成

音声認識プロセスが作成する 4 つ組の構成要素の 1 つである認識結果の属性値には，Rx と Vx がある．Rx の x はその語句に含まれる指示語のさす対象の数が x 個であることを示し，R はその語句が動詞句を含まないことを意味する．Vx の x はその語句に含まれる指示語のさす対象の数が x 個であることを示し，V はその語句が指示語を x 個以上含む動詞句であることを意味する．例えば，認識対象語句と属性値との間には**表 6.2** の対応関係がある．

本システムでは認識対象語句として，表 6.2 のはじめの 2 行のようなマルチモーダル入力用の言葉を登録した．この他に，“これを登録して” と “これを読んで” を V1 と対応づけて登録した．さらに，プルダウンメニューを引き出さずにコマンドを実行するためのホットキーやアクセラレータキーに割り当てられたコマンドを音声で投入できるようにするために，それらを起動するための語句を表 6.2 の 3 行目のように V0 と対応づけて登録した [4].

6.4.3　指示動作の処理：3 つ組の作成

WWW ブラウザでは，マウスカーソルがアンカーの領域に入ったときに，WWW ブラウザの左下隅に URL が表示され，一方そのアンカーから出た時には表示されていた URL

[4]WWW ブラウザのメニューに登録されている search，forward，back，OpenURL などの英語の 21 個を，“検索してください”，“次に進む”，“前に戻る”，“ページを開けたい” などの日本語に訳し登録した．

の表示は消える．本システムではこれを用い，指示開始時刻，指示終了時刻，および指示対象候補の名前である URL からなる 3 つ組を作成する[5]．

6.4.4　MM入力解析プロセス：モード間での対応付け

指示語のみ（Rx）または指示語を含む動詞句（Vx）が発声され，かつ指示動作が行われた場合には，システムは 2 種類の時間帯をもつ情報を取得している．すなわち，アンカーが指示されている指示時間帯（3 つ組）と，音声が話されている発声時間帯（4 つ組）である．4 つ組には R1 または V1 が含まれる．3 章で述べたように，4 つ組と時間軸上での重なりを持つすべての 3 つ組の中で，最も長い重なりを作る 3 つ組を持つ指示対象（アンカー）が指示されたと判断され，指示表現と対応付けられる．

6.4.5　問題解決器：コマンドへの変換

今回の試作では，ユーザは認識対象語句毎に区切って，動詞句まで省略せずに発声することを仮定している．そのため，V1 または V0 が得られた後でそれまでの入力をすべて用いてコマンドに変換する．音声認識結果の文字列の違いに基づいて，システムはコマンドを選択する．例えば,「これと」「これを取ってきて」と言いながら 2 つの対象を指示した場合，後者の “これを取ってきて” に対応する属性値 V1 が得られたときにコマンドへの変換を開始する．文字列 “これを取ってきて” をキーとして変換すべきコマンドを取りだし，そのコマンドを適用する対象として，“これと” と “これを取ってきて” のそれぞれに対応づけられた対象を指定する．

現在の実装では，マルチモーダル入力によって起動できるコマンドとして，次のものがある．

1. **複数のファイルのダウンロードのための一括命令**
   例えば，ftp サイトから複数のファイルをダウンロード（get）したい場合がある．ftp サイト上のファイルは英単語に由来するものであったり，それにバージョン番号がついていたりして，声に出して読みづらいものが多い．このような場面で，指示動作を伴う「これと」「これを取ってきて」という命令で複数のファイルの取得を１度に命令できる．これを実現するために，MM 入力解析の結果として得られた 2 つの URL から取得（ftp での get）すべきファイル名を取りだし，WWW ブラウザ自体のプロセスとは別に取得を行なうよう実装されている．

[5]今回の実装では WWW ブラウザとして NCSA の Mosaic を用いたためソースプログラムの入手が可能であるので，関数 pointer_motion_callback と TrackMotion に変更を加え，速度が閾値以下である場合に URL と時刻とを送信した．

2. **複数の URL のホットリストへの一括登録**
   例えば，サーチエンジンでの検索の後，結果のリストの前後の文を見て，自分が探しているものにより近いと思われる幾つかのアンカーだけを記録する場合がある．このような場面で，指示動作を伴う「これと」「これを登録して」という命令で複数のアンカーを 1 回の命令で登録できる．これを実現するために，MM 入力解析の結果として得られた 2 つの URL を順番にホットリストに登録するよう実装されている[6]．

3. **アンカーの指し示す HTML 文書の読み上げ**
   例えば，新聞社が WWW に新聞記事を提供している．通常それらの記事のタイトルのページで各記事のアンカーをクリックし，記事を目で読む．本システムでは，ユーザがアンカーを指し示しながら「これと」「これを読んで」という命令を行なうことにより，音声での読み上げが開始される．これを実現するために，MM 入力解析の結果として得られた 2 つの URL が指し示す HTML 文書を CCI[41] を用いて読み込んで，そこから HTML のタグを取り除いたテキストを作成し，音声合成プログラムに入力する．そして音声合成プログラムが合成音声でユーザに出力するよう実装されている．このように WWW ブラウザに本来ないコマンド実行のためにアンカー（またはそれが指す内容）を渡すことが可能である．

   さらに音声だけによって次のコマンドの実行が可能である．

4. **「検索」や「ページのアップ／ダウン」や「ページのフォワード／バックワード」などのアクセラレータキーとホットキーの実行**
   例えば「検索してください」という入力の場合，“検索してください” に対応する属性値は “V0” で登録されているから，MM入力解析部では対応付けを行なわず，即座に問題解決器に送られ，認識結果の文字列からコマンドへの変換が開始され，検索用のウィンドウが表示されるよう実装されている[7]．

6.4.6　評価

WWW ブラウザの状況での対応づけの精度を確認する目的で，被験者 5 人による簡単な実験を行なった．被験者は 3 節の実験とは別の被験者である．ftp サイトのページを想定し，622 ドット × 528 ドットの領域に 10 行 2 列に並べて 20 個のアンカーを表示した．そ

[6] ホットリストのウィンドウに，URL を構成する文字列を一文字ずつ X-Window の KeyPressEvent として送ることによって実現される．

[7] ホットキーである文字 “s” を X-Window の KeyPressEvent として送ることによって，検索用のサブウィンドウの表示が実現される．

のうちの 3 つのアンカーを順番に指示しながら「これと」「これと」「これを取ってきて」と被験者に発声させた．3 つのアンカーは乱数で選んだ．これを 10 パターン用意した．すなわち被験者は 1 人あたり 30 回の指示を行なった．全被験者分をあわせて指示の回数は 150 回である．

被験者毎の精度は 80.0%から 93.3%までであり，平均すると 88.0%であった．

使用感については，実験後に簡単なアンケート（項目は付録参照）を行った．その結果によれば，本方式における「一括命令の便利さ」と「指示動作の自然さ」については支持する（5 段階評価で 4 点以上）被験者が多かった（それぞれ 4 人と 3 人）．

一方，指示動作が「やり易いかやり難いか」については「どちらとも言えない」という評価が最も多かった．その理由は,「やり易さ」の評価においてはシステムの全般的な使用感が評価されているからと考えられる．特に，十分な feedback をシステムからユーザに返せていないことがこの評価結果の原因と考えられる．実際，アンケートで得られた感想の中には,「自分が，ある物を指示しているということを確認できるような反応を（システムから）返して欲しい」,「指示をしたという feedback がほしい」,「音声を入力中であるという feedback を，目線をアンカーに向けたままでも得られるよう設計した方が良い」があった．

6.5　考察

従来の慣習に沿って操作される GUI では直接操作が多用される．その操作にはマウスのシングルクリックが用いられる．この直接操作では対象の選択と命令の実行という 2 つの機能が一体となっている．そのため，この動作はボタンのような対象の指示には使えない．そこで本研究では，クリックを使わないマウスカーソルの動きを指示動作として採用した．これは，対象の選択と命令の実行が一体となった直接操作とは明確に異なり，対象の選択だけが可能な様式である．また，その指示動作は，人間の指さしのように対象の上にポインターを重ねるだけという自然な様式である．これらを導入することにより，ボタンのような対象を指示の対象として扱えるようになり，6.2 で述べたような自然で命令の内容を音声で柔軟に変更できるマルチモーダル入力が従来の GUI に対しても可能となった．

その例として，WWW ブラウザとユーザとのマルチモーダルインタラクションを可能にするシステムを実装した．ボタンとして機能する WWW ブラウザのアンカーを指示の対象に含めることが可能となった．その結果，ユーザは指示動作と一緒に「これとこれをダウンロードして」のように発声するマルチモーダル入力によって，自然な様式で WWW ブラウザに命令を伝達できる．

ここでさらに，6.4.5 の 1 のような用件を果たす場合のインタラクションについて考え

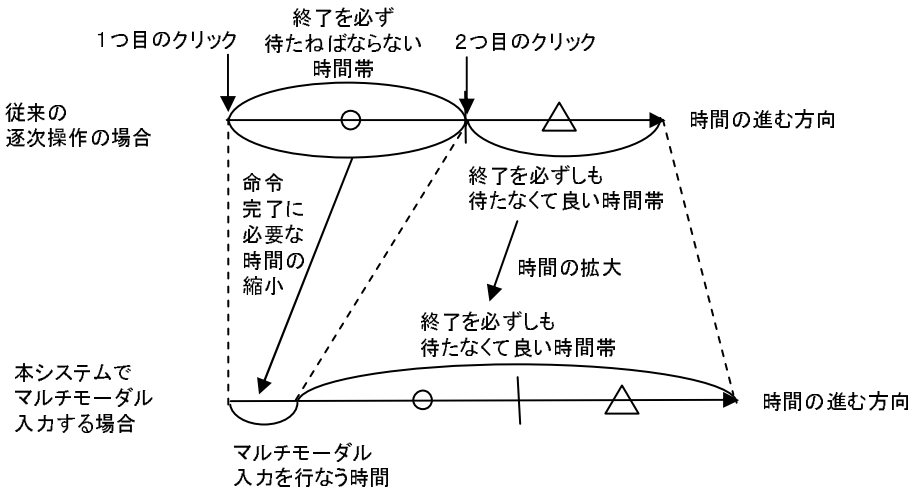

図 6.4: 従来と本システムの間でのインタラクションの比較

る．或る WWW ブラウザでは，新たに複数のブラウザを開くことで連続したファイルのダウンロードが可能である．しかし，そのたびにブラウザが開かれるのは煩わしい．別の方法として１つのブラウザ上で順番に１つずつ取得することも可能である．しかし，その場合でも，接続が確立されそのセーブ先の名前を入力するまで次のダウンロードにはとりかかれないので，ユーザは先に行なった処理の結果を待たなければならない．

しかし，本システムのように対象の選択と命令の実行の分離と音声と指示動作の導入によって自然な操作が可能となる．WWW ブラウザにマルチモーダルインタラクションが可能な本システムでは，上と同じ事態で，ユーザにとって必要なことは，各アンカーを指示しながら「これとこれをダウンロードして」と言うことだけである．その後は，実際のファイル取得や接続の確立の終了を待たなくて良い（**図 6.4**）．また，URL とアンカーとの対応表を用いて,「ここからここまでをダウンロードして」という範囲を指定する入力を解釈できるように改良すれば，さらに効率の高い入力が可能となる．このように，マルチモーダル入力を受け付ける本システムは自然で柔軟で効率的なインタラクションを実現する．すなわち，我々の設計方針である，指示対象の選択と命令の実行の分離と，音声と指示動作の導入によって一般の GUI での直接操作の回数を減らすことも可能となる．

本システムの完成度という観点からは，6.4.6 で明らかとなったように，インタフェース全体としての使用感を満足させるようなシステムからユーザへの feedback の面での改良が今後必要である[8]．

WWW ブラウザへの命令に音声を利用する研究はこれまでにも存在した [9, 36, 59]．し

[8]マルチモーダル入力の一般的な課題とも共通する課題という観点からは，対話の文脈の考慮などを含む言語処理や知識処理の導入が必要である．

かし，これらの研究では，各チャネル，つまり音声とマウスの操作というそれぞれの入力チャネルの間には関連性がない．すなわち従来の研究では単にマウスの操作を音声に置き換えただけである．このような置き換えは我々のシステムによって得られるような柔軟性を生み出さない．また 6.4.5 の 1 の事態では，読みづらいアンカーに対してさまざまな読みを仮定して音声入力に備えなければならなくなる．

さらにマルチモーダルインタフェースの従来の研究では，参照物への指示動作として，マウスクリックかパネルへの接触が使われていた．WWW ブラウザの状況では，これらの方法はページを切り変えるために機能する．これらの方法は，我々のシステムによって実現されるような自然で柔軟で効率的なマルチモーダル入力には役立たない．

6.6　6 章のおわりに

本研究では，慣習になっている GUI の直接操作との間で曖昧性を生じさせることなく，かつ，人間にとって自然な指示動作として，クリックを使わないマウスの動きによる指示動作を導入した．そして，この指示動作と，それと並行して発声された指示語との間での対応付けを行うマルチモーダル入力方法を説明し，対応づけの精度を報告した．さらに，その方法が適用された WWW ブラウザへのマルチモーダル入力システムの構成，及び，インタラクションの事例，精度を挙げ，考察した．

クリックを使わない指示動作と音声とを用いたマルチモーダル入力により，GUI への命令入力に柔軟性，自然性がもたらされる．さらに範囲表現を扱えるようにすれば，設計の方針から論理的に効率化が図られる見込みがある．

6 章の付録: 被験者に提示したアンケートの詳細

以下に，被験者に提示したアンケートの文面を示す．尚，Q1 から Q3 の 5 段階評価の結果をまとめる際に，各尺度の左端から 5 点，4 点，3 点，2 点，右端に 1 点を与えた．

Q1) １つの ftp サイトの１つのページから，2 つのファイルを取得する場合についてお尋ねします．通常の WWW ブラウザでは，次の方法でファイルを取得できます．
1) Mosaic の場合，1 つ目のアンカー上でマウスクリックを行ない，そのファイル取得が終わるのを待つ．それが終わったら，2 つ目のアンカー上でマウスクリックを行なう．
2) Netscape の場合，1 つ目のアンカー上でマウスクリックをして，ファイル取得用のプログラムとアンカーが示す URL との接続が完了するのを待つ．それが終わったら，2 つ目のアンカー上でマウスクリックを行なう．
このように，1 つ目の取得または接続が終わるまで，2 つ目の入力ができません．
こういった「手作業の繰り返し」と本システムで実現した「音声と指示による一括入力」とでは，どちらが便利だと思いますか？
下の 5 段階で評価してください．(＋に○をつけてください)
どちらが便利か？
一括入力 ＋―＋―＋―＋―＋ 手作業の繰り返し

Q2) 本システムの指示動作は，マウスのカーソルをアンカーの上に重ねるというものでした．指示は日常の指使いと同じように自然だと感じましたか？
自然 ＋―＋―＋―＋―＋ 自然ではない

Q3) 本システムでの指示動作はやり易かったですか？
やり易い ＋―＋―＋―＋―＋ やり難かった

Q4) 感想を書いてください．

第7章　結論

7.1　本研究のまとめ

音声翻訳は，音声認識，言語翻訳，音声合成，といった要素技術を必要とする複合的なシステムであり，各要素技術の構築自体とそれらの統合化が必要となる．各要素技術は実データから人間によって，または，計算機によって自動的に規則やモデルを作成することにより構築される．しかし，構築時に参照されたデータと各要素技術が適用される時のデータとの間の不一致は不可避であり，構築時に参照していない未知情報の推定技術の実現が重要な課題となる．統合化においては，各要素技術の単純な結合では解決されない課題が顕在化するが，それらは各要素技術の観点から見ればその外側で生じる課題であり，見過ごされてきた．その１つとして要素技術間での情報の統一化課題が存在する．本論文では，自然な音声翻訳通信の実現を目的として，音声翻訳構成技術のうちの音声認識と音声合成に関わる音声言語処理を中心に，未知情報の推定という課題を解決する要素技術の機能拡充と要素技術間での情報統一化課題を解決する統合化技術の提案を行なった．

まず，第 2 章から第 4 章において，音声翻訳を適用する際に生じる未知情報の推定課題の解決を行なう方式の提案と評価を行なった．

第 2 章では，データ未整備の新規タスクへの適応用データを用意することを課題とし，機械翻訳を用いて同タスクの別言語のデータから生成したデータでの適応方式を研究した．その結果，言語モデルの予測性能を示す平均単語分岐数（パープレキシティ）の向上が実験的に確認された．最近では，より希少なアイスランド語の音声認識器の適応の手法として本研究が参考にされている [27].

第 3 章では，言語翻訳された結果を音声合成する際に問題となる，未登録単語のアクセント型の推定を研究した．本研究では，アクセント型の推定問題を複数の候補のなかの分類問題として設定し，現在高い分類精度を示しているサポートベクトルマシンを適用して，高い精度でのアクセント型推定が可能であることを確認した．

第 4 章では，言語翻訳結果を音声合成して人間に聞かせる際に重要となる対話的なイントネーションの実現を問題として，対話種別ごとの実際の音声の基本周波数を，重畳モデルの観点から分析を行なった結果について述べた．その結果，対話種別ごとの違いを大域的な F_0 変動が担い，対話種別ごとに特徴的な語句において読み上げ音声の基本周波数変動と比べた場合の差異が大きいこと，アクセント句内の局所的な変動は読み上げ音声との

類似性が高いことが確認された．この結果から，対話種別ごとに特徴的な表現を手がかりとして大域的な F_0 変動を制御することで，対話口調の韻律の制御方法を確立すればよいことが分かった．

次に，第 5 章と第 6 章で，音声翻訳を構成する各要素技術の結合では解決されない統合化の課題として情報統一化の課題の解決を行なう方式の提案と評価を行なった．

第 5 章では，話し手，音声認識，言語翻訳の間に存在する発話区分不一致という課題への対処として，発話の文への分割を実現するための統計的言語モデルについての研究結果を述べた．発話境界を組み込んだ統計的言語モデルを提案し，評価実験をおこなった．その結果，音声認識の過程において，従来の単語認識精度を損なうことなく，高い精度で発話を文に分割できることが明らかとなった．句読点による境界情報の明示の無い音声認識結果は人間にとって読みにくい．最近では音声認識結果の表示を見やすくするための手法としても本研究の結果が参考にされている [56]．

自然なコミュニケーションでは，手で対象物を指さしながら，同時に言葉では指示語を使って会話がなされる．訳語の選択のみならず，翻訳先言語での指示語と指示動作とが対応したマルチモーダル出力のためにも入力された指示語と指示動作との対応付けが必要である．第 6 章では，マウスを使った指示動作と指示語との対応付け手法の研究について述べた．音声と指示動作という 2 種のモードを用いて WWW ブラウザを操作する場面での指示語と指示動作との対応付けの方式の検証を行った．既存の Graphical User Interface（GUI）での様式に変更を加えることなく自然に指示動作を導入することができるので，離れた 2 地点に居る人間が共通の画面（GUI）を見ながら（リモート・アシスタンスの場面等），かつ，音声翻訳器を介した対話を行なう場面での指示動作と指示語とを用いた円滑な対話の実現が可能となる．

これら本論文の第 2 章から第 6 章の提案法・分析により**図 7.1** の太字で示した課題を解決に近付けることができた．発話の文への分割（第 5 章）と未登録語のアクセント型の推定（第 3 章）は実際に利用されている．機械翻訳結果を用いた言語モデル適応（第 2 章）の発想はリソース未整備の言語やタスクでの方策研究のきっかけとなった [27]．対話 F_0 の予測（第 4 章）と指示語の認定（第 6 章）は，現段階では，まだ将来の技術という位置づけである．しかし，これらの課題解決方法の提案により，要素技術においては未知の状況下での推定精度が改善され，要素技術間においては情報が統一され，全体的な自然性とパフォーマンスの高い音声翻訳器の実現に貢献する可能性が高い．

7.2　今後の課題

各章で提案した方法を適用することによって，本論文で焦点を当てた課題の解決は可能であるが，限界も予想される．以下では，さらなる発展に向けた今後の課題を章毎にまと

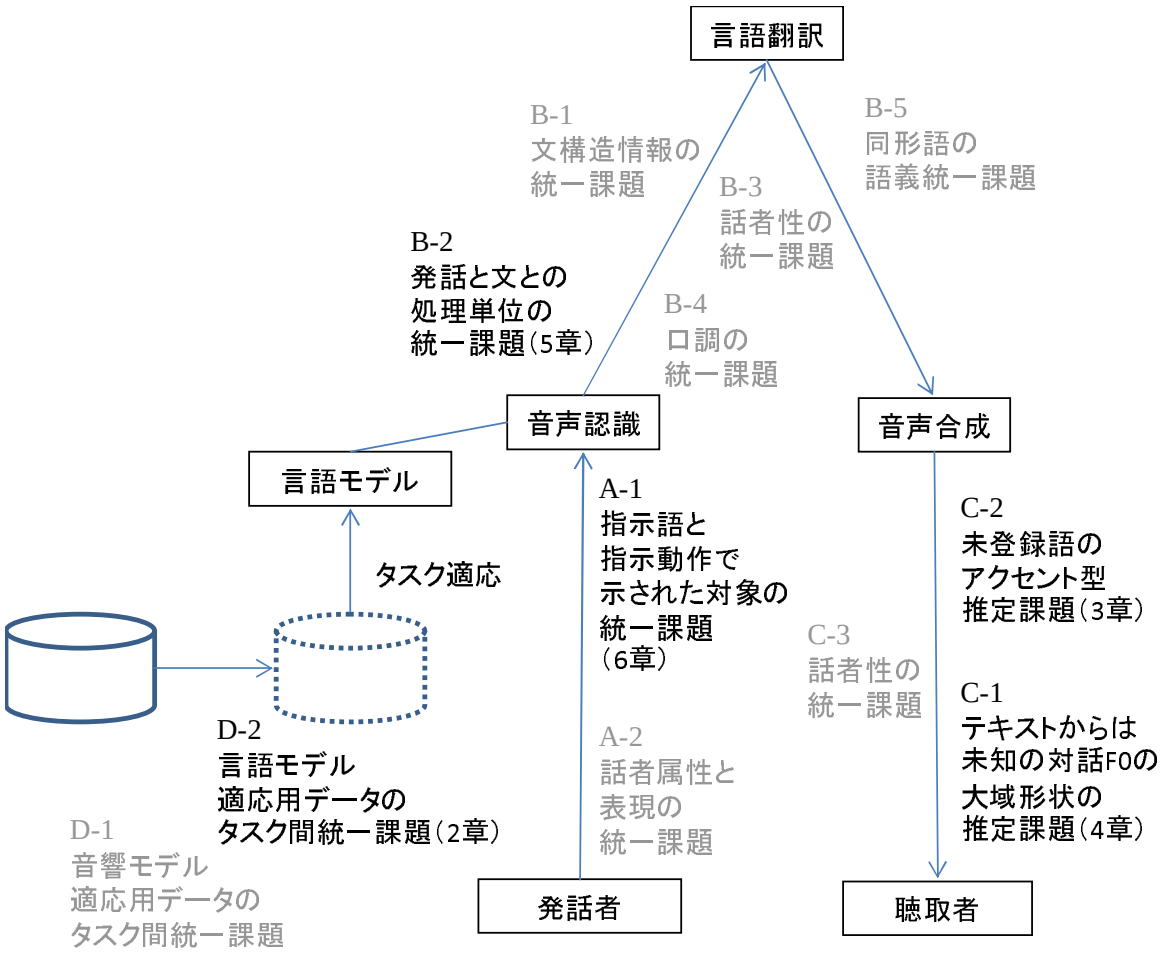

図 7.1: 本論文で解決を試みた課題

める．

現在の統計的機械翻訳では尤度の高い順に翻訳結果が出力されるので，第 1 位の翻訳結果の中にみられる単語系列のパターンが限定される．その結果，データ未整備・未開拓の言語での単語系列種別（バリエーション）の増加に限界があり，パープレキシティの改善に飽和が生じていた．今後，翻訳結果の N-Best，ラティス，および，用例に基づく翻訳器やルールベースの翻訳器等の原理の異なる複数の翻訳器からの出力結果を使う等の方法を取り，それらの中から多様な表現が取り出されることにより，第 2 章で取り組んだ適応の性能がさらに改善されると予想される．

第 3 章での未登録語のアクセント型推定では，SVM(Support Vector Machine) が決定木に比べて高い性能を示した．自動学習のみで構築される点でも優れているが，まだ若干精度の改善の余地がある．SVM 以後にも，分類問題において高い分類精度を示すモデルが登場した．それらの適用も興味深い今後の課題である．音声翻訳用のアクセント型推定としては翻訳元言語では漢字が未知であるが，音声翻訳用以外の用途のためのアクセント型推定には漢字の分布に従ってバイアスをかけることも考えられる．

第 4 章では，対話口調と読み上げ口調との間での基本周波数を重畳モデルの各成分毎に比較した．その結果，対話種別ごとに特徴的な語句においてアクセント句間の大域的な成分の差異が大きいこと，アクセント句内の局所的な変動は読み上げ音声との類似性が高いことが確認された．これらの分析をさらに推進して，特徴を明らかにし，合成対象の文のもつ言語情報から各種の基本周波数成分の変動を予測するモデルを構築することが今後の課題である．また，変動の大きなところを集中的に少量のデータから学習できるモデルの効率的な構築法の検討も課題である．

第 5 章で取り組んだ日本語での発話から文への分割の成功には，境界を示す品詞・活用形（終助詞，終止形など）の役割が大きい．一方，英語はそのような情報の有効性は低く，難しい課題となっている．日本語においても，対話性が増し，断片的な発話が増えるにつれて，発話末の単語の品詞等が発話境界を強く示す情報として機能しなくなることが予想される．今後は品詞以外の言語情報，例えば，複数の単語からなる意味まとまりを考慮に入れた方式や，音響情報の関与の大きい方式の研究が必要になると考えられる．精度の高い識別モデルが自然言語処理に続いて音声認識においても用いられるようになってきた．それらの適用も今後の課題である．

第 6 章では，指示動作で示された指示対象と指示語との対応付け方式の提案を行なった．指示対象の認識には音声と指示動作とのオーバーラップする時間長の情報を用いた．文節発声になる場合には今回の提案法で高い精度が得られたが，連続発声の場合には速度の緩急等の別の情報の併用が必要である．

謝辞

本研究を進めるにあたり，多大なる御指導と御助言を賜りました早稲田大学大学院国際情報通信研究科 匂坂芳典教授に深く感謝致します．

本論文をまとめるにあたり，貴重な御意見を賜りました早稲田大学大学院国際情報通信研究科 山崎芳男教授，早稲田大学大学院国際情報通信研究科 河合隆史教授，早稲田大学スポーツ科学学術院 誉田雅彰教授，早稲田大学理工学術院 小林哲則教授に感謝致します．

本論文において行ないました研究は，早稲田大学大学院国際情報通信研究科博士後期課程に入学して行なった研究，日本電信電話（株）の研究所，および,（株）国際電気通信基礎技術研究所（ATR）における業務として行なった研究をまとめたものです．早稲田大学大学院博士後期課程への進学をご支援いただきました，高橋敏様（NTT サイバースペース研究所 音声言語メディア処理プロジェクト 音声対話インタフェースグループリーダー）に感謝致します．ATR にて音声言語処理の研究を開始するきっかけや研究の機会を与えてくださり，多大なるご支援を下さいました，坂間保雄様（現在，坂間技術翻訳事務所，筆者が ATR 出向時の日本電信電話（株）の研究部長），東田正信様（現在，NTT ソフトウエア（株），元 国際電気通信基礎技術研究所 取締役企画部長）をはじめとする，NTT と ATR の人材育成と研究マネージメントでお世話になった皆様に感謝致します．

本論文の中の各研究の共同研究者として，懇切丁寧に御意見，御討論，御助言を頂きました，加藤恒昭様（現在，東京大学），山本博史様（現在，近畿大学），渡辺太郎様（現在，情報通信研究機構（NiCT)），永田昌明様（現在，NTT コミュニケーション科学基礎研究所)、阿部匡伸様（現在，NTT サイバーソリューション研究所）に深く感謝致します．

本研究の一部では NTT 研究所に在籍した研究員の方々にも被験者としてご協力いただきました．また，ATR では代々作成・維持管理されてきた貴重な音声言語データベースを活用させて頂きました．韻律の研究用のデータ作成には NTT 研究所の音声対話インタフェースグループの多くの方々にご協力いただきました．お世話になった皆様に感謝致します．

在学中に様々な面でお世話になりました，早稲田大学大学院国際情報通信研究科 匂坂研究室に所属する学生の皆様にも感謝します．

最後に，様々な面で支え応援してくださいました両親と弟と友人に感謝致します．

参考文献

[1] M. Abe and H. Sato. Two-stage F0 control model using syllable based F0 units. *Proc. of ICASSP, IEEE*, vol. II, pp. 53-56, 1992.

[2] M. Abe and H. Sato. Statistical analysis of the acoustic and prosodic characteristics of different speaking styles. *Proc. of Eurospeech*, pp. 2170-2110, 1993.

[3] M. Abe and H. Mizuno. Speaking style conversion by changing prosodic parameters and formant frequencies. *Proc. of ICSLP*, pp. 1455-1458, 1994.

[4] H. Ando, H. Kikuchi, and N. Hataoka. Agent-typed Multimodal Interface Using Speech, Pointing Gestures and CG.. Proceedings of HCI International '95, ELSEVIER, pp.29-34, 1995.

[5] A. Berger and R. Miller. Just-In-Time Language Modeling. *Proc. of the ICASSP*, pp. 705-708 1998.

[6] R. A. Bolt. Put-that-there: Voice and Gesture at the Graphics Interface. ACM Computer Graphics, 14, 3, pp.262-270, 1980.

[7] P. F. Brown, S.A.D. Pietra, V.J.D. Pietra and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2), pp. 263–311,1993.

[8] C. J. Chen. Speech Recognition with Automatic Punctuation. *Proc. of the Eurospeech-99*, pp.447-450, 1999.

[9] H. Dohi and M. Ishizuka. A Visual Software Agent connected with the WWW/Mosaic. Proceedings of Multimedia Japan 96, pp.392-397,1996.

[10] K. Hakoda, T. Hirokawa, H. Tsukada, Y. Yoshida, and H. Mizuno. Japanese Text-To-Speech Software based on Wave Form Concatenation Method. Proceedings of AVIOS'95, pp.65-72,1995.

[11] H. Fujisaki and S. Nagashima. A model for the synthesis of pitch contours of connected speech. *Annual Report of the Engineering Research Institute, Faculty of Engineering, University of Tokyo*, pp. 53-60, 1969.

[12] O. Furuse et al.. Splitting Long or Ill-formed Input for Robust Spoken-language Translation. In *Proceedings of COLING-ACL-98*, pp.421-427,1998.

[13] 古瀬蔵, 美馬秀樹, 山本和英, M. Paul, 飯田仁. 多言語話し言葉翻訳に関する変換主導翻訳システムの評価. 言語処理学会第 3 回年次大会発表論文集, pp.39–42, 1997.

[14] L. Galescu and J. F. Allen. Bi-directional conversion between graphemes and phonemes using a joint n-gram model. *Proc of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, pp. 103–108, 2001.

[15] M. Gavalda, K. Zechner and G. Aist. High Performance Segmentation of Spontaneous Speech Using Part of Speech and Trigger Word Information. In *Proceedings of the 5th ANLP*, pp.12-15,1997.

[16] 蜂谷雅弘, 森山高明, 小川均, 天白成一, 橋本雅行. 言語情報を考慮したアクセント型推定手法. 情報処理学会 研究会報告 音声言語情報処理 17 ,pp. 103–108,1997.

[17] 蜂谷雅弘, 森山高明, 小川均, 天白成一, 橋本雅行. アクセント法則を取り入れたアクセント型推定手法. 日本音響学会秋季研究発表会講演論文集, pp. 241–242, 1997.

[18] 広川智久, 佐藤大和. 人名（姓）におけるアクセント形推定法. 日本音響学会春季研究発表会講演論文集, pp. 425–426, 1981.

[19] 広瀬啓吉 編著. 韻律と音声言語情報処理. 丸善, pp. 24-34, 2006.

[20] M. Hiyoshi and H. Shimazu. Drawing Pictures with Natural Language and Direct Manipulation. Proceedings of the Coling'94, pp.722-726,1994.

[21] T. Imai, Y. Saito, A. Ando and S. Furui. A language model for recognition of continuously uttered sentences. In *Journal of Acoustical Society of Japan.(E)*, 21, 2, pp.111-114,2000.

[22] 伊東伸泰, 荻野紫穂, 新島仁. 文法を利用した N-gram モデルのタスク適応. 言語処理学会第 4 回年次大会発表論文集, pp. 610–613, 1998.

[23] 岩瀬成人. 自然言語処理を用いた企業名解析方式. 電子情報通信学会論文誌 Vol. J82-D-II No.8, pp. 1305–1314, 1999.

[24] R. Iyer, M. Ostendorf, and J. R. Rohlicek. Language Modeling with Sentence-Level Mixtures. *Proc. of the ARPA Workshop on Human Language Technology*, pp. 82-87 1994.

[25] Java 技術研究会: HTML と JavaScript, 工学図書 (1996).

[26] 情報処理振興事業協会 (IPA). 日本語ディクテーション基本ソフトウエアの開発 仕様策定書および説明書-1997 年度-. 情報処理振興事業協会 (IPA), 1998.

[27] A. T. Jensson, E. W. D. Whittaker, K. Iwano, S. Furui. Language Model Adaptation for Resource Deficient Languages Using Translated Data. *Proc. of Interspeech*, pp. 1329-1332, 2005.

[28] 金田一春彦監修, 秋永一枝編, アクセント習得法則. 明解日本語アクセント辞典 第二版, 1996.

[29] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda. XIMERA: a new TTS from ATR based on corpus-based technologies. *Proc. of 5th ISCA Speech Synthesis Workshop, ISCA*, pp. 179-184, 2004.

[30] K. Loken-kim, S. Mizunashi and M. Tomokiyo. Analysis and Integration of Multimodal Inputs in Interpreting Telecommunications. 情報処理学会研究会報告 SLP-7-12, pp.73-78, 1995.

[31] Y. Kokenawa, M. Tsuzaki, H. Kato, and, Y. Sagisaka. F0 control characterization by perceptual impressions on speaking attitudes using multiple dimensional scaling analysis. *Proc. of ICASSP, IEEE*, vol.1, pp. 273-276, 2005.

[32] 工藤 拓. Yamcha :Yet Another Multipurpose CHunk Annotator.
URL http://chasen.org/~taku/software/yamcha .

[33] B. Laurel. Interface agents: metaphors with character. The Art of Human-Computer Interface Design (Laurel, B. (ed.)), pp.355-365, Addison-Wesley Publishing Company, Inc.,1990.

[34] A. Lavie, et al.. Input Segmentation of Spontaneous Speech in JANUS: a Speech-to-speech Translation System. In *Proceedings of ECAI-96 Workshop on Dialogue Processing in Spoken Language Systems*, 1996.

[35] R. Luk and R. Damper. Stochastic phonographic transduction for English. *Computer Speech and Language, 10*, pp. 133–153, 1996.

[36] 松井一郎. ソフトだけで音声を認識. 日経コンピュータ 1996.9.2, pp.116-117, 1996.

[37] H. Masataki, Y. Sagisaka, K. Hisaki, and T. Kawahara. Task Adaptation using MAP Estimation in N-gram Language Modeling. *Proceedings of the ICASSP*, pp. 783-786, 1997.

[38] T. Morimoto, et al.. A Speech and Language Databases for Speech Translation Research. In *Proceedings of the ICSLP-94*, pp.1791-1794,1994.

[39] 内藤正樹, SINGER Harald, 山本博史, 中嶋秀治, 松井知子, 塚田元, 中村篤, 匂坂芳典. 旅行会話タスクにおける ATRSPREC の性能評価. 日本音響学会秋季研究発表会論文集, 3-1-9, pp.113-114,1999.

[40] 中村哲. 音声翻訳システムの研究開発. 電子情報通信学会技術研究報告, SP, 音声, pp.31-36, 2009.

[41] NCSA Mosaic Common Client Interface (CCI 1.1):
URL http://www.ncsa.uiuc.edu/SDG/Software/XMosaic/CCI/cci-spec.html.

[42] NCSA Mosaic Home Page:
URL http://www.ncsa.uiuc.edu/SDG/Software/Mosaic/Docs/help-about.html.

[43] F. J. Och, and H. Ney. A Comparison of Alignment Models for Statistical Machine Translation. *Proc. of COLING-2000*, pp. 1086–1090,2000.

[44] F. J. Och, and H. Ney, Improved Statistical Alignment Models. *Proc. of ACL-2000*, pp. 440–447,2000.

[45] S. Ohnishi, H. Yamamoto, and Y. Sagisaka. Structured language model for class identification of out-of-vocabulary words arising from multiple word classes. *Proc. of Eurospeech Vol.1*,pp. 693–696, Sept. 2001.

[46] J.R. キンラン. AI によるデータ解析, トッパン, 1995.

[47] M. D. Riley. Tree-based modeling of segmental durations. *Talking Machines: Theories, Models, and Designs*, G. Bailly, C. Benoit, and T.R. Sawallis eds., North-Holland, pp. 265-273,1992.

[48] A. I. Rudnicky. Language Modeling with Limited Domain Data. *Proc. of the ARPA Workshop on Spoken Language Technology*, pp. 66-69 1995.

[49] A. Sakurai, K. Hirose, and N. Minematsu. Data-driven generation of F0 contours using a super positional model. *Speech Communication*, vol.40, no.4, pp. 535-549, 2003.

[50] Y. Sagisaka. On the prediction of global f0 shape for Japanese text-to-speech. *Proc. of ICASSP,IEEE*, pp. 325-328, 1990.

[51] 匂坂芳典, 佐藤大和. 日本語単語連鎖のアクセント規則. 電子通信学会論文誌 Vol. J66-D No.7, pp. 849–856, 1983.

[52] Y. Sagisaka, T. Yamashita, and, Y. Kokenawa. Generation and perception of F0 markedness for communicative speech synthesis. *Speech Communication*, vol.46, no.3-4, pp. 376-384, 2005.

[53] J. V. Santen, T. Mishra, and E. Klabbers. Estimating phrase curves in the general superpositional intonation model. *Proc. of 5th ISCA Speech synthesis workshop*, pp. 61-66, 2004.

[54] C. Schmant. コンピュータとのヴォイスコミュニケーション. サイエンス社,1995.

[55] M. Schroeder. Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions. Proc. of Workshop on Affective Dialogue Systems, Kloster Irsee, Germany, pp. 209-220, 2004.

[56] 下岡和也, 南條浩輝, 河原達也. 講演の書き起こしに対する統計的手法を用いた文体の整形. 自然言語処理, vol.11, No.2, pp. 67-83 ,2004.

[57] A. Stolcke, and E. Shriberg. Automatic Linguistic Segmentation. In *Proc. of ICSLP-96*, pp.1005-1008,1996.

[58] M. Streit. Active and Passive Gestures - Problems with the Resolution of Deictic and Elliptic Expressions in a Multimodal System. Proceedings of a workshop on referring phenomena in a multimedia context and their computational treatment organized by the ACL Special Interest Group on Multimedia Language Processing pp.44-51,1997.

[59] surftalk:URL http://www.surftalk.com/surftalk/index.html.

[60] 竹澤寿幸, 森元逞. 発話単位の分割または接合による言語処理単位への変換手法. 自然言語処理, Vol.6, No.2, pp.83-95,1999.

[61] T. Takezawa, et al.. Speech and Language Databases for Speech Translation Research in ATR. In *Proc. of the 1st International Workshop on East-Asian Language Resources and Evaluation(EALREW '98)*, pp.148-155,1998.

[62] H. Tang, X. Zhou, M. Odisio, M. Hasegawa-Johnson, and, T.S. Huang Two-stage prosody prediction for emotional text-to-speech synthesis. *Proc. of Interspeech*, pp. 2138-2141, 2008.

[63] P. Taylor. Rise / fall/ connection model of intonation. *Speech Communication*, 15 (1-2), pp. 169-186, 1994.

[64] C. Tillmann and H. Ney, Word Re-ordering and DP-based Search in Statistical Machine Translation. *Proc. of COLING-2000*, pp. 850–856,2000.

[65] R. L. Trask. *A Dictionary of Phonetics and Phonology*, 1996.

[66] 土田尚純, 大山実, 笹岡信. 日本人の名前アクセント型付与規則. 電子通信学会論文誌 Vol. J67-D No.5, pp. 625–626, 1984.

[67] V. N. Vapnik. *The Nature of Statistical Learning Theory*,Springer-Verlag,1995.

[68] W. Wahlster. User and Discourse Models for Multimodal Communication. Intelligent User Interfaces (Sullivan, J.W. and Tyler S.W. (ed.)),pp.45-67, ACM Press,1991.

[69] Y. Wang, M. Mahajan, and X. Huang. A Unified Context-Free Grammar and N-gram Model for Spoken Language Processing. *Proc. of ICASSP*, 2000.

[70] 渡辺和幸, 英語イントネーション論. 研究社出版, 1994.

[71] 山田智一, 野田喜昭, 井本貴之, 嵯峨山茂樹. クライアント・サーバ構成の HMM-LR 連続音声認識システムとその応用. 情報処理学会研究会報告 SLP-5-6,pp.39-46,1995.

[72] H. Yamamoto and Y. Sagisaka. Multi-class Composite N-gram Based on Connection Direction. In *Proceedings of ICASSP-99*, pp.533-536,1999.

[73] 山本博史, シンガー ハラルド. 母音および無音の HMM を用いた音声始端検出法. 日本音響学会春季研究発表会論文集, 1-Q-3, pp.137-138,2000.

[74] 谷戸文廣, キュンホ ローケンキム, ローレル ファイス, 森元逞. 道案内タスクにおけるマルチモーダル対話の会話文の特徴分析. 電子情報通信学会論文誌 D-II, Vol.J77, No.8, pp.1475-1483,1994.

[75] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Simultaneous modeling of spectrum, pitch, and duration in HMM-based speech synthesis. *Proc. of Eurospeech*, pp. 2347-2350, 1999.

[76] K. Zechner and A. Waibel. DiaSumm:Flexible Summarization of Spontaneous Dialogues in Unrestricted Domains. In *Proceedings of the Coling-2000*, pp.968-974,2000.

研究業績一覧

主論文

査読付学術論文

1. 中嶋秀治, 加藤恒昭: ”WWW ブラウザとのマルチモーダルインタラクション-クリックを使わないマウスの動きと音声を入力とするインタフェース”, 情報処理学会論文誌, Vol. 39, No. 4, pp.1127-1136, 1998.

2. 中嶋秀治, 山本博史: ”音声認識過程での発話分割のための統計的言語モデル”, 情報処理学会 論文誌, Vol. 42, No. 11, pp.2681-2688, 2001.

3. 中嶋秀治, 山本博史, 渡辺太郎: ”機械翻訳によって生成された追加テキストを使った統計的言語モデルの適応”, 電子情報通信学会 論文誌 D-II, Vol. J86-D-II, No.4, pp.460–467, 2003.

4. 中嶋秀治, 永田昌明, 浅野久子、阿部匡伸: ”SUPPORT VECTOR MACHINE を使ったモーラ列からの日本語姓名のアクセント推定”, 電子情報通信学会論文誌 D-II, Vol. J88-D-II, No.3, pp.480–488, 2005.

査読付国際会議

5. Hideharu Nakajima and Tsuneaki Kato: ”Multimodal interaction with WWW browsers”, Abridged Proceedings of HCI International 97 (San Francisco), pp.43, 1997.

6. Hideharu Nakajima, Hirofumi Yamamoto, and Taro Watanabe: ”Language model adaptation with additional text generated by machine translation”, Proceedings of COLING (Taipei), pp.716–722, 2002.

7. Hideharu Nakajima, Masaaki Nagata, Hisako Asano, and Masanobu Abe: ”Estimating Japanese word accent from syllable sequence using Support Vector Machine”, Proceedings of Eurospeech (Switzerland), pp.2681–2684, 2003.

8. Hideharu Nakajima and Yoshinori Sagisaka: ”F0 analysis for Japanese conversational speech synthesis”, Proceedings of 8th International Symposium on Natural Language Processing (SNLP2009), pp.137–142, 2009.

査読付紀要論文

9. 中嶋秀治, 匂坂芳典: ”対話音声合成を目指した対話音声の韻律分析”, 早稲田大学大学院 国際情報通信研究科 紀要, pp.134–139, 2009.

国内研究会

10. 中嶋秀治、加藤恒昭: ”クリックを使わないマウスの動きと音声を入力とするインタフェース”, 情報処理学会 音声言語情報処理 ヒューマンインタフェース 合同研究会報告 96(21), pp.15-20, 1996 年 2 月 3 月.

国内大会

11. 中嶋秀治, 加藤恒昭: ”これをここに置いて：バイモーダルインタフェースの提案”, 情報処理学会第 51 回平成 7 年後期全国大会講演論文集, 第 6 巻, 6U-2, pp.243-244, 1995.

12. 中嶋秀治, 山本博史: ”発話分割付き実時間音声認識”, 日本音響学会秋季研究発表会講演論文集, pp.147-148, 1999.

13. 中嶋秀治, 山本博史, 渡辺太郎: ”機械翻訳によって生成された追加テキストを使った言語モデルの適応”, 言語処理学会第 8 回年次大会発表論文集, pp.283-286, 2002.

14. 中嶋秀治, 永田昌明, 浅野久子, 阿部匡伸: ”SUPPORT VECTOR MACHINE を使った音韻系列からの日本語単語アクセントの推定”, 日本音響学会秋季研究発表会講演論文集, pp.223-224, 2003.

15. 中嶋秀治, 匂坂芳典, 宮崎昇, 水野秀之, 間野一則 : ”品詞とその下位分類とアクセント句間の結合型を使った対話 F_0 制御”, 日本音響学会春季研究発表会講演論文集, pp.367-368, 2008.

16. 中嶋秀治, 匂坂芳典: ”対話音声と読み上げ音声との間での F_0 の比較”, 日本音響学会春季研究発表会講演論文集, pp.471-472, 2009.

参考論文

査読付学術論文

17. 内藤正樹, 山本博史, シンガー・ハラルド, 中嶋秀治, 中村篤, 匂坂芳典: ”対話音声を対象とした連続音声認識システムの試作と評価”, 電子情報通信学会 論文誌 (D-II), Vol. J.84-D2, No.1, pp.31-40, 2001.

18. Atsushi Nakamura, Masaki Naito, Hajime Tsukada, Rainer Gruhn, Eiichiro Sumita, Hideki Kashioka, Hideharu Nakajima, Tohru Shimizu, and Yoshinori Sagisaka: ”A speech translation system applied to a real-world task/domain and its evaluation using real-world speech data”, IEICE Transactions on Information and Systems, Vol.E84-D No.1, pp.142-154, 2001.

査読付国際会議

19. Hideharu Nakajima and Masahiko Oku: ”An automated grouping of segments in a line-chart to explain its movement”, Artificial Intelligence - sowing the seeds for the future - (Proceedings of AI’94, Australia), ed. by Chengqi Zhang et al., World Scientific, pp.450-457, 1994.

20. Hideharu Nakajima, Yoshinori Sagisaka, and Hirofumi Yamamoto: ”Pronunciation Variants Description using Recognition Error Modeling with Phonetic Derivation Hypotheses”, Proceedings of ICSLP (Beijing China), pp.1093–1096, 2000.

21. Hideharu Nakajima, Izumi Hirano, Yoshinori Sagisaka, and Katsuhiko Shirai: ”Pronunciation Variant Analysis using Speaking Style Parallel Corpus”, Proceedings of Eurospeech, Vol.1, pp.65-68 (Aalborg, Denmark), 2001.

22. Hideharu Nakajima, Yoshihiro Matsuo, Masaaki Nagata, Kuniko Saito: ”Portable Translator Capable of Recognizing Characters on Signboard and Menu Captured by Built-in Camera”, Proceedings of ACL Interactive Poster and Demonstration Sessions, pp.61-64, 2005.

23. Tsuneaki Kato, Yukiko I. Nakano, Hideharu Nakajima, and Takaaki Hasegawa: ”Interactive Multimodal Explanations and their Temporal Coordination”, Proceedings of the 12th ECAI, pp.261-265, 1996.

24. Hisako Asano, Hideharu Nakajima, Hideyuki Mizuno, and Masahiro Oku: ”Long vowel detection for letter-to-sound conversion for Japanese sourced words transliterated into the alphabet”, Proceedings of ICSLP (Korea), pp.1917- 1920, 2004.

社内誌

25. 間野一則, 水野秀之, 中嶋秀治, 宮崎昇, 吉田明弘: ”顧客へのリアルな音声応答を実現するテキスト音声合成技術「Cralinet」”, NTT 技術ジャーナル, Vol. 18, No. 11, pp. 19–22, 2006.

26. Kazunori Mano, Hideyuki Mizuno, Hideharu Nakajima, Noboru Miyazaki, and Akihiro Yoshida: ”Cralinet – Text-to-speech system providing natural voice response to customers”, NTT Technical Review, Vol. 5, No. 1, pp. 28–33, 2007.

国内研究会

27. 加藤恒昭, 中野有紀子, 中嶋秀治, 長谷川隆明: ”対話的マルチモーダル説明とその時間的協調”, 情報処理学会研究会報告 自然言語処理研究会, Vol.1996, No.65, pp.135-142, 1996.

28. 奥田浩三, 中嶋秀治, 河原達也, 松井知子, 中村哲: ”講演音声認識のための音響モデル構築方法の検討”, ワークショップ「話し言葉の科学と工学」, pp.109-116, 2001.

29. Hideharu Nakajima, Izumi Hirano, Yoshinori Sagisaka: ”発声スタイル並行コーパスを用いた発音変形の分析”, 電子情報通信学会技術研究報告 音声研究会, Vol.101 No.155, pp.37-43, 2001.

30. 奥田浩三, 中嶋秀治, 河原達也, 中村哲: ”講演音声の音響的特徴分析と音響モデル構築方法の検討”, 情報処理学会研究会報告 音声言語情報処理研究会, Vol.2001, No.68, pp.73-78, 2001.

国内大会

31. 中嶋秀治, 大内幸雄: ”数値データ説明文生成に関する一考察”, 電子情報通信学会 春季大会, 第 6 巻, pp.110, 1993.

32. 平野泉, 中嶋秀治, 匂坂芳典, 白井克彦: ”発話スタイル並列コーパスを用いた発音変形の分析”, 日本音響学会春季研究発表会講演論文集, pp.15-16, 2002.

33. 浅野久子, 中嶋秀治, 水野秀之, 奥雅博: ”ローマ字表記単語の読み上げのための長音置換・追加位置判定法”, 言語処理学会 第 10 回年次大会発表論文集, pp.685-688, 2004.

34. 間野一則, 水野秀之, 中嶋秀治, 浅野久子, 磯貝光昭, 長谷部未来, 吉田明弘: ”コンタクトセンタ向け波形接続型コーパスベース音声合成システム「Cralinet」の開発”, 日本音響学会秋季研究発表会講演論文集, pp.347-348, 2004.

35. 中嶋秀治, 永田昌明, 松尾義博: ”「領域抽出不要型文字認識」に基づく景観中単語認識”, 電子情報通信学会総合大会講演論文集 情報システム (2), D-12-67, p.217, 3 月, 2005.

36. 李克, 中嶋秀治, 時岡洋一, 匂坂芳典: ”日本人学習者に見られる中国語声調制御難易度の分析”，電子情報通信学会 2009 総合大会講演論文集 情報・システム，D-14-11, 2009.