



音声翻訳のための連続音声認識の研究

脇田, 由実

(Degree)

博士 (工学)

(Date of Degree)

2002-03-31

(Date of Publication)

2014-04-21

(Resource Type)

doctoral thesis

(Report Number)

甲2532

(URL)

<https://hdl.handle.net/20.500.14094/D1002532>

※ 当コンテンツは神戸大学の学術成果です。無断複製・不正使用等を禁じます。著作権法で認められている範囲内で、適切にご利用ください。



博士論文

音声翻訳のための連続音声認識の研究

平成14年1月

神戸大学大学院自然科学研究科

脇田 由実

音声翻訳のための連続音声認識の研究

—目次—

第1章 序論

1.1 はじめに

1.2 連続音声認識技術の現状と問題点

1.3 音声翻訳システムの現状と問題点

1.4 本研究の目的とアプローチ

1.4.1 音声認識性能の向上

1.4.2 音声翻訳のための認識誤りの校正

1.5 各章の構成

第2章 音声認識性能の向上

2.1 音節／状態継続時間制御を用いた連続音声認識

2.1.1 はじめに

2.1.2 音節継続時間予測

2.1.3 音節継続時間予測法を用いた音節継続時間制御

2.1.4 音節継続時間制御法が認識性能に及ぼす効果

2.1.5 音節継続時間予測法を用いた状態継続時間予測

2.1.6 状態継続時間制御

2.1.7 状態継続時間制御法が認識性能に及ぼす効果

2.1.8 まとめ

2.2 複数音素にまたがる認識誤り特性を用いた連続音声認識

- 2.2.1 はじめに
- 2.2.2 現在の音響モデルによる認識誤り分析
- 2.2.3 認識誤り特性を用いた音素候補系列追加モデル
- 2.2.4 認識結果への音素候補系列追加
- 2.2.5 音素候補系列追加が認識性能に及ぼす効果
- 2.2.6 認識誤り特性と品詞との関係について
- 2.2.7 品詞の違いを考慮した状態系列の認識誤り特性抽出
- 2.2.8 認識誤り特性を用いたマルチ発音辞書構築
- 2.2.9 マルチ発音辞書の認識性能に及ぼす効果
- 2.2.10 まとめ

第3章 言語翻訳のための認識誤りの校正

3.1 単語間の意味的類似性を用いた正解認識部分特定

- 3.1.1 はじめに
- 3.1.2 自由発話文解析法を用いた正解部分特定
- 3.1.3 正解部分のみ翻訳する部分翻訳
- 3.1.4 部分翻訳の翻訳性能に及ぼす効果
- 3.1.5 考察
- 3.1.6 まとめ

3.2 単語間の依存関係を用いた類似用例選択

- 3.2.1 はじめに
- 3.2.2 音声翻訳のための類似用例選択
- 3.2.3 類似用例文選択のためのキーワードクラス化法
- 3.2.4 まとめ

第4章 まとめ

—参考文献—

—謝辞—

—発表文献—

第1章 序論

1.1 はじめに

通信ネットワークの急激な発達とともに、多種多様な情報の取り扱いが可能な情報端末の必要性が急速に高まってきている。このような情報端末を実現するためには、複雑化する機器操作を簡便に使いやすくするインターフェースが望まれており、一つのアプローチとして、音声関連技術を情報端末に導入し、操作性のよいインターフェースを実現することが考えられる。一方、音声認識技術の研究開発も、コンピュータ技術や通信技術などの周辺技術の急速な進歩を背景に加速されてはいるが、現実には音声認識技術を応用した実用化の市場はまだ小さい。これは、音声コミュニケーション技術の基礎となる「連続音声認識技術」を本格的に実用化するには、まだ解決されなければならない技術的課題が残されていることによる。

人間の発声は、単語の単独発声においても、その音響的な特徴は話者や発声環境などで千差万別である。連続発声となるとさらに、隣接単語との相互関係や表現内容によって様々な音響的変動が生じる。この変動はあまりにも多様であり、機械による音声認識が誤りを起こす主な原因となっている。これに対して人間は、音響特徴の識別能力が高いだけでなく、言語知識などの情報を活用することにより、正しく発声を認識することができる。また、全ての単語が正しく認識できなくともコミュニケーションをとっていくことができる柔軟な訂正能力を持ちあわせている。

連続音声認識における認識誤りを最小限にとどめ、実用化に耐えうる認識性能を保证するためには、(1) 連続発声の際に見られる多様な発声変動に耐え得る認識性能を如何に実現するか、(2) 仮に音声認識誤りを起こしても正しく認識された単語を特定して適切な校正処理ができるか、の二点が、重要なポイントとなる。この二点に対する解決策が見出せれば、実用化に向けての使用用途はより大きく開けるものと思われる。本論文は、連続音声認識技術を応用した音声翻訳システムの研究開発を通じて、実用化に特に重要な上記二点に対する研究成果を中心にまとめたものである。

1.2 連続音声認識技術の現状と問題点

連続音声認識は二つの過程からなる。一つは連続的な発声を音素、音節などの音声単位列に変換する過程である。音声単位毎のモデルの音響的特徴パターンと入力音声から得られた音響的特徴パターンとの類似性を調べることにより、音声単位系列候補を特定する。他の一つは得られた音声単位系列から単語列を探索する過程である。単語辞書や単語連結規則などに従いながら、音声単位系列候補から最も妥当であると考えられる単語列を出力する。

最初の音声単位系列への変換過程では、同じ発声内容でも、発声環境の違い、話者の違い、発声スピードの違いなどにより入力音声特徴パターンは様々であるため、予め設定された音響モデルから大きく異なる場合も多く、度々誤った音声単位系列が出力される。また第二番目の単語列探索過程でも、入力される語彙や発声表現は様々であり、予め設定された単語辞書、単語連結規則に表されていない場合は、誤った単語列が出力されてしまう。連続音声認識技術の研究は、これらのような様々な要因における音声言語の変動を如何に吸収し、最終的に正しい単語列結果を出力するか、という課題への取り組みである。

近年、音声認識処理において、発話文の音響的、言語的特徴を音声やテキストのデータベースから統計的に自動学習する手法が広く用いられている。この手法では、各音声単位の音響的特徴パターンを記述する音響モデルとして隠れマルコフモデル (HMM) などの確率モデル、単語列を探索する際の単語連接制約として N-gram と呼ばれる単語連接の統計量が広く使用されている。このような確率的・統計的な手法や変動要因を考慮したデータベースを使用することにより、認識性能は飛躍的に向上し、連続音声認識技術は大きく進歩した。

これらの技術の進展にもかかわらず、連続音声認識の技術を実用化した事例は未だ数少ない。これは、限りなく多数の語彙の組み合わせと多種多様な発声時の変動に対して、ユーザーが期待する確実な認識性能に達していないためであると思われる。十分な性能が得られない大きな原因の一つは、文単位での音声やテキストを、音素、音節、単語などの小さい音声単位に分割したモデルでは、連続している音響的特徴量を完全に記述できないことにある。たとえば、同一音素でも音響的な現われは音韻環境、

発話様式などによって大きく異なり、従来の音声単位毎のモデル化では、コンテキストに依存したモデルを用いる方法をとっている。たとえば、最も顕著な音響的ばらつきが見られる音素環境コンテキスト特有の発声の相違を吸収するためにコンテキスト依存型 HMM が提案され有効性が報告されている (Schwartz1985、Lee1990、Takami1992)。また、HMM の構造を最適に自動設計する方法 (Takami1993、Hori1997、Singer1996) や、音素や音節などの音声学や音韻学上で定義された音声単位ではなく、音響的に類似している部分 (セグメント) を音声単位としてその統計量を HMM の入力とする方法 (Ostendorf1989、Franzini1990、Hanna1999) など、コンテキストの違いによる多様な音響的変動を吸収するため、HMM 構造の決定や音声単位を自動推定する提案も多くなされている。しかし、これらの工夫を行っても、各モデル内での閉じた学習では、モデル間にまたがる特性の記述は不十分であり、完全に音声認識誤りをなくすことは難しい。

1.3 音声翻訳システムの現状と問題点

連続音声認識技術の進歩とともに、応用したシステムの実用化開発も加速されている。この好例が、音声翻訳システムである。音声翻訳は、音声認識技術、言語翻訳技術、音声合成技術が統合されて実現される技術である。従来の言語翻訳技術では、まず書き言葉の文法規則に従って文解析を行い、解析された語彙の意味や文構造に従って目的言語に言語変換を行うしくみが多く採用されていた。このため、書き言葉の文法から逸脱した文章が入力された場合、文解析ができなくなるという問題を抱えていた。音声翻訳システムでは、話し言葉が正しい単語列として認識されても、言語翻訳における文解析部で書き言葉しか解析できないのでは、翻訳は不可能である。近年では、話し言葉特有の言い回しなどを話し言葉コーパスから獲得し、限定された使用目的には話し言葉を翻訳することが可能となってきた (Sato1991、Sumita1992、Maruyama1992、Gtishman1992、Furuse1994)。このように、話し言葉データベースから得られたモデルや規則を駆使した連続音声認識技術や言語翻訳技術を用いることで、音声翻訳技術もまた実用化に大きく近づいてきている。

しかし、連続音声認識を応用した音声翻訳システムも、実用化のためには別の大き

な課題を抱えている。これまでのシステムには、音声認識誤りを起こし、発声内容とは意味が異なった認識結果が出力されたとしても、その誤りを誤りであると判断したり正しい結果に修正したりする処理機能が備わっていない。この処理機能の不足から、従来の音声翻訳システムは多くの場合、認識誤り結果に対応して誤った翻訳結果を出力するか、翻訳できずに再入力を待つことしかできない。

音声翻訳システムの最も必要な機能は、原言語における発声文の意味を目的言語に変換して正しく伝えることにありと考えられる。言い換えれば、発声文を一字一句誤りがないように認識を行うことが目的ではなく、目的言語に変換された際に発声文の意味が正しく伝わるように変換することが重要であり、この条件を満たせば、認識誤りを起こしても目的は達成されることになる。従って、先に述べた誤りを判断したり修正したりする処理は、発声文の意味を正しく理解することを目的とした処理である必要があると思われる。

一方、従来の連続音声認識技術のしくみは、単語接続頻度を反映した統計的言語モデルを用いて、適切な後続単語のみが探索されるように局所的な言語的制約を与えることで、正しい単語列候補の出力を可能にしている (Bahl1983)。しかし、このような統計的言語モデルでは単語接続のみの局所的制約しか与えていないため、部分的な単語接続が適切であっても、文全体から見ると不自然で翻訳が不可能な認識誤り文を出力してしまう問題が残る。

これらを総合的に考えると、音声翻訳システムの実用化のためには、認識誤り部分の特定や修正を行うしくみが必要であり、特に、音声翻訳システムの目的と連続音声認識のしくみとを考慮すると、誤り部分の特定や修正には、得られた認識結果全体の意味的な整合性を評価尺度とした判断が適切であると考えられる。

1.4 本研究の目的とアプローチ

本研究は、従来の統計的手法の利点を活かしながら上記のような問題を解決するために、従来手法では用いられていない音響的特徴や言語的制約を従来システムに導入することで、実用可能な連続音声認識技術の確立とさらにそれを応用した音声翻訳技術を実現した。音声翻訳に必要な技術の中で、グレーブロックで示した部分の処理が

新たに導入された本研究の位置づけである。図 1-1 に示したように、音声認識性能の向上と認識誤りの校正について、各々二つずつからなる計四つの処理を考案した。各々の詳細を以下の 1.4.1 節、1.4.2 節で説明する。

1.4.1. 音声認識性能の向上

音声認識性能が十分でない大きな原因が、話者の違いや使用環境の違いによる多種多様な音響的変動にあることを 1.2 節で述べた。この課題の解決のために、二つの方策を考案した。一つは音声単位の継続時間の利用であり、この新たな韻律特徴の利用により、音響特徴に相補的な情報に基づく性能向上が期待できる。他の一つは、音声単位にまたがる認識誤り傾向の利用であり、単一の音声単位毎に独立した学習では得られない情報による性能の向上をねらったものである。

音声単位の継続時間は、通常の HMM では、状態間の遷移確率の形で間接的に反映されてはいるが、認識箇所に対してダイナミックな形で反映されていない。つまり、各 HMM の単位毎に独立に継続時間を制御しているために、発声速度や話し方の違いなどに起因する継続時間の変動が考慮されていない。そこで、一発声内の他の音声単位部分との相互関係を考慮した継続時間制御を施すことで認識性能の向上を図る。

また、音声単位に起因する認識誤りを低減するために、従来から話者適応や環境適応といった技術が提案されてきた。しかしこれらの技術は、少量の適応用データベースを用いて音声単位毎にパラメータの再学習を行うアプローチがほとんどであり、単位構造自身の不適切さを解決するためにモデルの構造、単位、状態数などの条件を最適に設定しなおすものでは少ない。いくつかの試みはなされているが、実際には、少量のデータベースからこの最適条件を設定するのは難しい。従ってこれらの技術を用いた適応学習が行われた後も、音声単位構造を最適に変更できなかったことによる認識誤りが残り、特に音声単位にまたがった部分での音声認識誤り対策が不十分のままになることが考えられる。そこで、複数の音声単位にまたがった認識誤り特性を分析し、この認識誤りを解決することにより認識性能の向上をめざした。

1.4.2. 音声翻訳のための認識誤りの校正

認識誤りが生じた場合に全く意味の理解できない結果が出力される問題の所在を 1.3 節で述べた。音声翻訳システムの最も必要な機能が、発声文の意味を異なる言語に変換して伝達することであることを考慮して、上記問題の解決策として、認識結果文全体の意味的な整合性を評価尺度とすることで、従来の統計的に求めた単語連接制約のみでは特定できない認識誤り部分を特定し、これを校正した結果を提供できる処理を考案した。この処理ために、二つの方法を考案した。

一つは、音声認識誤り部分を推定することで、正解部分のみを部分翻訳する方法である。意味を理解する上であまり必要ではない部分文が認識誤りを起こした場合には、認識誤り部分を除去するだけで正しい意味伝達できる翻訳結果を提供できる可能性があると考えた。そこで、特定された認識誤り部分を除去し、正解部分のみを翻訳することで意味を正しく伝達できるしくみを考案した。

もう一つは、直接発声文を翻訳する代わりに、発声文と意味が類似した用例を選択し、その対訳を翻訳結果とみなす方法である。認識誤り部分が意味を理解する上で必要な単語（以後キーワードという）である場合には、認識誤りであると推定された部分を除去するだけでは、正しく意図が伝わる翻訳結果は望めない。この課題への対処として、認識結果に含まれるキーワードのみを用いて意味が類似した用例を選択し、その対訳を翻訳結果とみなして出力することで、正しく理解できる翻訳文の出力を可能にすることを考えた。

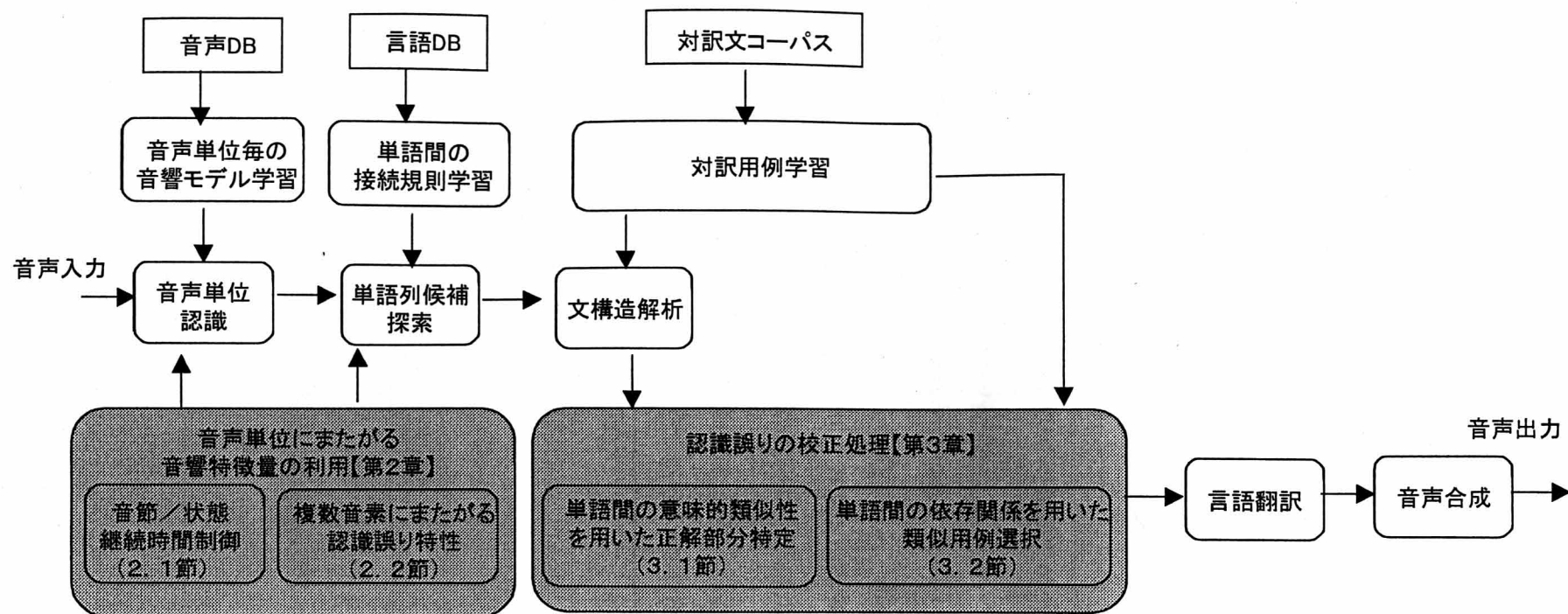


図1.1 音声翻訳に必要な技術

1.5 各章の構成

本論文では、以下次の内容について述べる。

以下第2章では、認識性能自身を向上させるための取り組みについて述べる。ここでは、従来の音声単位毎の統計的音響モデル化が必然的に持つ問題点を救済するために、音声単位にまたがる範囲での音響的特徴の利用法を述べる。2.1節では、音声単位の相互関係に注目した音節及び HMM の状態の継続時間制御法、2.2節では、複数音素にまたがる認識誤り特性を利用した認識手法について述べる。

第3章では、音声翻訳のための認識誤りの校正について述べる。特に、認識誤り文の意味的な整合性を評価することで、認識結果文に含まれる認識誤り部分を改善することを行った。3.1節では、認識誤り部分を除去し、正しく認識された部分のみを翻訳する手法、3.2節では、認識結果に含まれるキーワードから、発声文と意味が類似した用例を選択し、直接入力文を翻訳する代わりに、選択された用例の対訳を翻訳結果として出力する手法について、特に類似用例の選択性能を向上させるためのキーワードクラス化法を中心に述べる。

最後に第4章で、以上の結論を総括し、今後の課題を整理する。

これらの章間の関係をわかりやすくするために、図 1-2 に各章の構成図を示した。

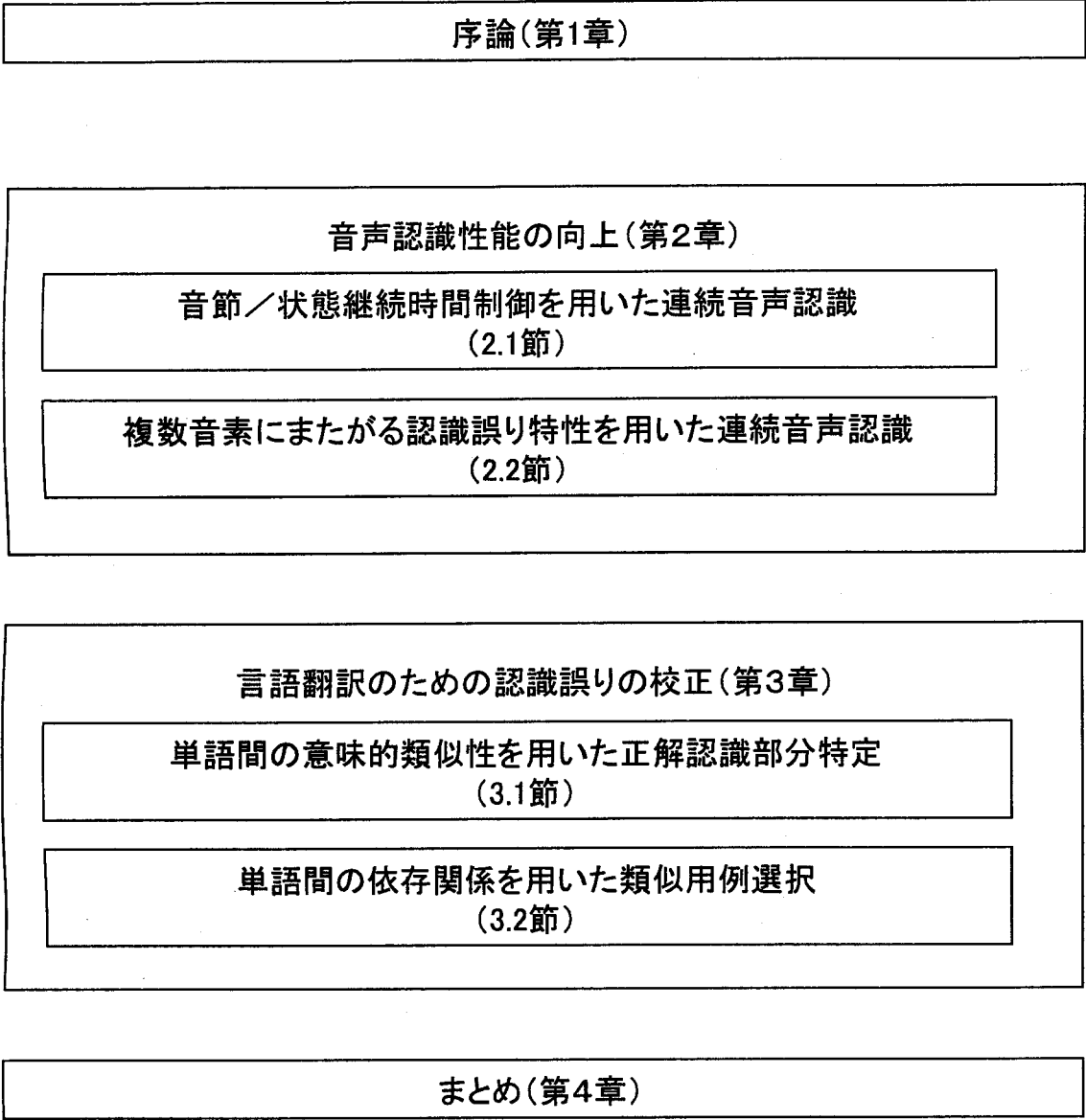


図 1-2 本論文の構成

第2章 音声認識性能の向上

2.1 音節／状態継続時間制御を用いた連続音声認識

2.1.1 はじめに

音声認識システムにおいて、スペクトル情報だけではなく継続時間情報を用いて各音声単位の存在する範囲を絞り込むことが、認識性能の向上や処理時間の削減に有効であることが知られている。HMM など確率的モデルを音響モデルとして利用した場合には、HMM の各状態にとどまっている確率が時間経過とともに単調減少し、実際の各状態の自然な継続時間を反映できない。各状態が自然な継続時間になるよう制御する方法として、ノンパラメトリックな継続時間制御を行う方法 (Ferguson1980, 中川 1989) や、正規分布、ポアソン分布、ガンマ分布、対数正規分布などの分布形状をもつパラメトリックな継続時間制御を行う方法 (Rabiner1989, Levinson1986, Russell1985, 高田 1991) などが提案されている。また、HMM の状態継続時間の直接的な制御としてではなく、学習データにおける音素や単語の継続時間を用いて、入力文の音素や単語の継続時間を制限し、その範囲で認識候補を探索する方法 (Lee1989, Rabiner1989, 松永 1989, Singer1992) も提案されてきた。しかし、これらの継続時間制御法は、各々の HMM の状態、音素、音節、単語を独立にしか制御できず、一発声内の隣接した HMM の状態、音素、音節、単語の継続時間の相互関係を考慮した制御法にはなっていない。

一方、音節 HMM を用いた連続音声認識システムにおいてその認識誤り結果を比べて見ると、一発声内にもかかわらず、音節の継続時間が不自然にばらついている認識誤り結果が多いことがわかった。このことから、従来のように個々の単位 (状態、音素、音節、単語など) の継続時間を独立に制御するだけでなく、一発声内における他の音響単位の継続時間を考慮しながら継続時間を制御することを新たに考える必要があると思われる。

本 2.1 節では、一発声内における音声単位の相互関係を積極的に利用して各単位の継続時間を制御する方法を提案し、このような制御法を用いることによる認識性能の

向上への効果を報告する。

まず、音節を音声単位とした left-to-right に探索が行われる認識システムにおいて、上記で述べた継続時間の相互関係を利用する方法として、既に認識された音節の継続時間から次に認識しようとする音節の継続時間を予測する方法を提案し、この予測値を用いて音節の探索区間を制限する方法について述べる。特に、音節の継続時間を左右する要因を調べ、これらの要因を用いることで高性能に音節継続時間の予測を行うことが可能となった。この音節継続時間予測に基づく継続時間制御法を、さらに小さな単位であり認識処理の最小単位である HMM の状態継続時間の制御に用いることができれば、認識処理と同時に逐次的に継続時間制御が可能となり、認識率、処理時間の面で、より高性能な認識システムの実現が可能となることが考えられる。そこでさらに次に、先に提案した音節継続時間予測法を用いた状態継続時間予測法を提案し、これを用いて状態毎に継続時間制御をする方法の認識性能に対するさらなる効果を報告する。

2.1.2 音節継続時間予測

音節の継続時間の制御を行うためには、基本的に、認識の対象となる音節継続時間を予測し、予測値をもとに適切な継続時間範囲を設定する必要がある。日本語においては、音節毎に継続時間がほぼ一定に保たれるように、先行する子音の継続時間に合わせて母音継続時間が調整される音節補償現象が認められている（比企 1967、佐藤 1977）。従って、一発声内の発声速度があまり大きく変化しない場合は、認識処理が left-to-right に音節認識候補を決定する手順に従い、次に認識しようとする音節の継続時間を、既に認識された音節認識候補の継続時間とほぼ同じものとして予測することが可能であると思われる。

一方、音声合成処理における継続時間制御の研究において、日本語の音素継続時間は、発声速度が同じでも、様々な要因、当該音素の種類、その近傍音素の種類、呼気段落内での音素の位置などに左右されることがこれまでに報告されている（匂坂 1980、海木 1990、Takeda1990）。これらの継続時間に関する研究では、2 つ前後の音素の種類、モーラ数、呼気段落内での位置などに継続時間が左右することが認められており、

これらの要因を考慮して音節継続時間を正規化しておくことによりより制度よく継続時間を予測する方法が実現できる可能性があると思われる。

(I) 音節継続時間予測法

一発声内にある音節継続時間に対し、継続時間を左右する要因を考慮して正規化を行うと、継続時間はほぼ等しくなると考えられる。このことは言い換えれば、二つの音節の継続時間の比、たとえばはじめから t 番目の音節継続時間 $dur(t)$ と i 番目の音節継続時間 $dur(i)$ との比は、音節継続時間を左右する要因毎に t 番目の音節と同じ条件にある音節の平均継続時間と及び i 番目の音節と同じ条件にある音節の平均継続時間との比に等しいということである。継続時間を左右する要因が複数、たとえば J 個あると仮定した場合、これを式で表すと式 2-1 のようになると考えられる。

$$\frac{dur(t)}{dur(i)} = \sum_{j=1}^J \alpha_j \frac{f_j(t)}{f_j(i)} \quad (2-1)$$

$dur(t)$: t 番目の音節の継続時間

α_j : j 番目の要因に対する重み

$f_j(t)$: 継続時間を左右する要因 j において t 番目の音節と同じグループに属する音節の平均継続時間

J : 上記要因の総数

α_j は j 番目の要因に対する重みであり、継続時間を左右する影響度の強さを反映して与えられるべきものである。既に認識された i 番目の音節の継続時間 $dur_o(i)$ を用いて、今から認識しようとする t 番目の音節継続時間を予測するために上記式 2-1 を利用すると、式 2-2 が得られる。

$$\overline{dur(t)} = dur_o(i) \sum_{j=1}^J \alpha_j \frac{f_j(t)}{f_j(i)} \quad (2-2)$$

$\overline{dur(t)}$: t 番目の音節の継続時間の予測値

$dur_o(i)$: i 番目の音節で観測された継続時間

各音節のばらつきを吸収して予測性能を向上させるため、 t 番目の音節の近傍の複数 m 個の音節を用いて平均することにより、式 2-3 が得られる。なお、式 2-3 において、もし音節の継続時間が様々な要因に左右されないと仮定すると、 $f_j(t)$ は、一発声内であれば t 及び j によらず一定値となる。

$$\overline{dur(t)} = \frac{1}{m} \sum_{i=t-m}^{t-1} dur_o(i) \sum_{j=1}^J \alpha_j \frac{f_j(t)}{f_j(i)} \quad (2-3)$$

以下、継続時間を左右する要因を明らかにし、上記予測式の有効性の評価結果を報告する。

(II) 音節継続時間を左右する要因

上記予測式 2-3 を用いて予測するためには、まず、音節継続時間を左右する要因を明らかにする必要がある。予測をなるべく少数のデータで効率良く行うために、要因の必要条件として次の 2 条件が考えられる。

(条件 1) 音節継続時間を大きく左右する要因であること

(条件 2) 話者や文が違っていても継続時間への影響の大きさに差がないこと

先の日本語の音素継続時間に関する研究結果では、当該音素の種類、その近傍音素の種類、呼気段落内での音素の位置などが要因として挙げられている。これらの要因の中からさらに、認識時に利用可能な情報であることを考慮して、当該音節の種類及び近傍音節の種類の二つを要因の候補として取り上げ、これが要因として上記二つの必要条件を満たすかどうかを調べた。まず、当該および近傍音節の種類が、音節継続時間を大きく左右する要因であること（上記（条件 1））を確認するために、（要因 1）当該音節の種類、（要因 2）先行音節の種類、（要因 3）2 つ前の音節の種類、の三つを要因とした場合に、各要因と音節継続時間との重相関係数および偏相関係数を求めた。次に、上記 3 要因における各音節の種類と音節継続時間との関係が話者や文章に依存しないかどうか（上記（条件 2））を調べるために、各要因毎に、音節の種類ごとの音節継続時間の平均値を算出し、この平均値に対して、話者間及び文セット間の相

関係数を求めた。

上記分析では次の音節データベースを用いた。音節境界が既に調べられている十分な量の音節を含むデータベースが存在しなかったため（研究開始 1992 年当時）、正しく認識された結果から信頼性の高い結果を選択し、選択された認識結果における音節の継続時間をデータとした。認識手法としては音節スポッティングを用い、次の 2 条件をともに満たす音節を信頼性の高い音節とみなした。

- (1) スポッティングされた音節どうしが重なり合う、もしくは隙間が空くことなく連なって文を形成していること。但し、多少の重なりや隙間は許容した。ここでは、正しく認識した文の平均音節スポッティング区間の半分の長さ以下を許容範囲とした。
- (2) 一発声内の全ての音節において、認識スコアが一定以上であること。つまり、対数化された遷移確率、出力確率、継続時間確率を加算することで求められる累積対数確率値（大黒 1988）が、十分高いこと。ここでは、正しく認識された文に含まれる全音節の 90%の音節における累積対数確率値が閾値以上になるように閾値を設定した。

データベースとしては、日本音響学会研究用連続音声データベースの中から、A 及び B セット 100 文を男声 6 名が発声した計 600 文を用いた。本章での実験は全てこのデータベースを用いている。スポッティングは、音節 HMM を用いて、入力音声側の始末端フリーの viterbi アルゴリズムにて行った（大黒 1988）。

当該および近傍音節の種類が音節継続時間を大きく左右する要因であることを確認するために求めた相関係数を表 2-1 に示す。これにより次の結果を得た。

- ・ 重相関係数の平均値は 0.764 であり、話者間の値のばらつきは少ない。
- ・ 偏相関係数の平均値は全て 0.4 以上である。最も係数値が大きいのは当該音節の種類であり、次いで先行音節、二つ前の音節の順であり、この順序は話者に依存しない。

これは、話者に関係なく、音節継続時間に及ぼす影響も当該音節の種類が最も大きいことを示している。

次に、上記 3 要因における各音節の種類と音節継続時間との関係が話者や文章に依存しないかどうか調べるために求めた話者間及び文セット間の相関係数を表 2-2、表 2-3

に示す。表 2-2 は、6 話者において全ての 2 話者間の相関値を求め平均した値、表 2-3 は、セット A B 間の相関値を話者ごとに求めて平均した値を示している。これにより、次のことがわかった。

- ・ 話者間および文章セット間の当該音節の種類における相関値は 0.8 以上の値である。これは、当該音節の種類ごとの平均継続時間の大小の傾向が、話者や文に寄らず同傾向であることを示している。
- ・ 当該音節の種類の次に、先行音節、二つ前の音節の種類、の順に話者間及び文章セット間の相関が高い。

以上の二つの実験により、認識時に用いることを考えた要因の中で、その寄与度の大きさは、当該音節の種類、先行音節の種類、二つ前の音節の順であることが確認できた。この傾向は音声合成を意図した分析結果とも符合し、話者に依らない傾向であることもわかった。

表 2-1 当該または隣接する音節の種類が音節継続時間に及ぼす影響

話者	重相関係数	偏相関係数		
		当該音節	先行音節	二つ前の音節
1	0.816	0.795	0.557	0.484
2	0.723	0.616	0.484	0.407
3	0.733	0.719	0.458	0.3
4	0.764	0.732	0.556	0.489
5	0.784	0.765	0.538	0.485
6	0.766	0.683	0.514	0.373
平均	0.764	0.718	0.518	0.423

表 2-2 音節の種類が音節継続時間に及ぼす影響の話者非依存性

[話者間の相関係数]

当該音節	先行音節	二つ前の音節
0.812	0.521	0.324

表 2-3 音節の種類が音節継続時間に及ぼす影響の文セット非依存性
[文セット間の相関係数]

当該音節	先行音節	二つ前の音節
0.817	0.404	0.239

(Ⅲ) 音節継続時間予測性能の評価

(Ⅲ－Ⅰ) 予測法の評価

音節継続時間を左右すると確認された三つの要因を用いて、前節で提案した継続時間予測式 2-3 により音節継続時間を予測し、予め音節スポッティングにより求められたスポッティング区間との誤差を調べることで予測性能を評価した。ここでは先に求められた要因の予測が、どれだけ有効かを確認するため、以下の2点を評価した。

- (a) 音節継続時間予測要因の種類と予測性能の関係。
- (b) 要因毎の平均音節継続時間（式 2-3 の f_j ）を求める際に用いたデータ（同話者または同じ文）で予測した場合と異なるデータ（異話者または異なる文）で予測した場合との予測性能の違い。

さらに予測式において以下二つの事項（要因の重み α と使用音節数 m ）が予測性能に及ぼす効果を確認した。

- (c) 各要因の重み α の効果
- (d) 使用音節数 m の最適数の確認

(a) 音節継続時間予測要因の種類

認識時の音節継続時間予測に使用可能三つの要因が、予測性能に与える効果を調べた。式 2-3 において、音節継続時間が要因に左右されないと仮定し、要因による正規化を行わない、つまり既に認識された m 個の音節継続時間を正規化せずにそのまま平均した値を予測値とした場合、また、音節継続時間が要因に左右されると仮定し、使用する要因を一つずつ増やして用いた場合の予測値とした場合の各々の予測誤差を比較することで、各要因の予測に対する必要性を評価した。継続時間に及ぼす影響が大

きい要因を考慮するほど予測に有効であると考えられる。そこで、継続時間に対する影響度の大きい要因から順に増やして改善度を調べた。さらに、全話者、全音節の継続時間の平均値と各音節継続時間との差を予測誤差とした場合を「予測なし」の場合の誤差とし、予測することの有効性も合わせて評価した。

要因毎の平均音節継続時間を求める際（以後学習時という）には1名のAセットを、予測には学習話者を含む6名のA及びBセットを使用した。後に確認する要因の重み α と使用音節数 m については、ここではその影響の違いが反映しないように、要因の重み α は要因に寄らず一定とし、使用音節数 m は、同じ文章内の全ての音節を用いて評価した。予測誤差分布を図 2-1 に、また図 2-1 に示す予測誤差分布の標準偏差及び予測誤差の最大値を表 2-4 に示す。標準偏差および予測誤差の最大値が小さいほど、予測誤差が小さく予測性能が良い、と考えられる。図 2-1 に示した結果から次のことがわかる。

- ・ 「予測なし」場合が最も予測誤差が大きく、要因が一つ増える毎に予測誤差は改善される。但し、2 要因と3 要因の差は僅かである。
- ・ 当該音節を用いて予測した場合の改善度は著しく、要因を考慮する効果は高い。

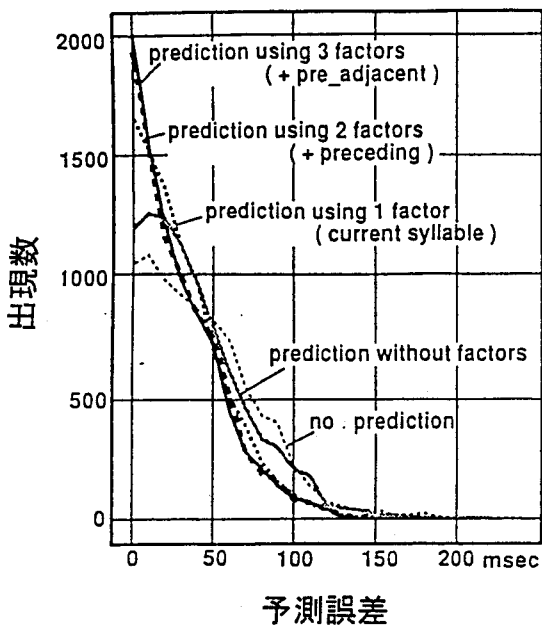


図 2-1 予測方法の導入による予測誤差の軽減効果

表 2-4 予測誤差分布の標準偏差および最大誤差値 [msec]

	標準偏差	最大誤差値
3要因にて予測	38.08	140
2要因にて予測	38.64	150
1要因にて予測	40.66	180
要因後との正規化なしで予測	49.56	220
予測なし	54.61	240

(b) 学習時と予測時とが同じ文である場合と異なる文である場合との予測誤差の違い

先の実験では、各要因と音節継続時間との関係が話者や文章に依存しているかどうかを調べたが、更にこれを予測誤差への影響という観点から確認した。ここでは、平均継続時間学習時のデータセットと予測時のデータセットとが、(1) 同話者、同じ文、(2) 同話者、異なる文、(3) 異話者、同じ文、(4) 異話者、異なる文、の場合の予測誤差を比較した。この四つの条件の詳細を表 2-5 に、予測誤差分布を図 2-2 に示す。結果から次のことがわかる。

- ・ 学習時と同話者、同じ文を用いた場合に、最も予測誤差が小さかったが、異話者、異なる文を用いた場合の予測誤差との差は少ない。学習時と異なる話者や異なる文が入力されても、継続時間の予測が適切に行われたと考えられる。平均継続時間を学習するためのデータは、認識されるデータと話者や文に関して一致している必要がないと考えられる。
- ・ 平均音節継続時間は、学習に用いた話者が 111ms、予測に用いた話者の中で最も長い話者で 125ms、短い話者で 100ms であり、この範囲の発声速度の違いであれば、予測性能には影響がないことがわかる。但し本データにおいては、発声速度による継続時間の差より音節の種類の違いによる継続時間の差の方が大きい。たとえば、6 名分の平均継続時間が最長である音節は「しよ」で 170ms、最短の音節は「う」で 76ms であった。発声速度が極端に異なる入力に対する有効性は、データを変えて実験する必要がある。

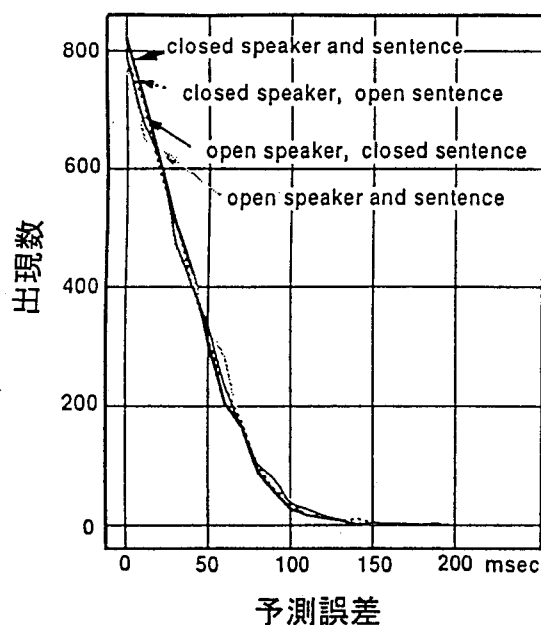


図 2-2 様々なデータセットにおける音節継続時間予測誤差

表 2-5 学習及び予測に用いたデータセット

条件	学習時	予測時
1	話者5名、文セットA	学習時と同話者・同セット
2	話者5名、文セットA	学習時と同話者・異なる文セットB
3	話者1名、文セットA	学習時と異なる話者5名・同セット
4	話者1名、文セットA	学習時と異なる話者5名・異なる文セットB

(c) 各要因の重み α を考慮する効果

一般には、重みを一定とせず偏相関係数値などを用いた方が予測性能が高くなると予測される。全て重みを同じにした場合と偏相関係数を用いて重み付けした場合との予測誤差分布の標準偏差および最大誤差値を表 2-6 に示す。その結果、予想とは逆に重みを要因に寄らず一定にしたほうが、僅かに予測誤差は小さかったが、その差は僅かであり、予測性能にほとんど差がみられない。実際の偏相関係数の割合は 0.43 : 0.31 : 0.26 となっており、予測性能への寄与がそれほど違わないためと理解される。

表 2-6 各要因の重み α と予測誤差との関係 [msec]

α 値	標準偏差	最大誤差値
要因に寄らず一定	38.08	140
偏相関係数を使用	38.34	140

(d) 使用音節数 m の最適数の確認

式 2-3 における予測式は、予測性能の向上のため、既に認識された複数音節の継続時間を用いて音節継続時間を予測する。しかし、使用音節数を増やし続ける場合の性能の限界や予測しようとする音節と時間的に遠い音節の情報ひずみの問題などから、使用音節数にも最適値があると考えられる。そこで、使用音節数と予測誤差との関係を調べた。使用音節数と予測誤差分布の標準偏差との関係を図 2-3 に示す。

- ・ 使用音節数が 4 以下の場合は、音節数が増えるほど、予測誤差が減少する。
- ・ 使用音節数が 4 以上では、予測誤差にほとんど差はない。

以上より、4 音節前後が最適数と思われる。予測式 2-3 に、以上で確認した最適条件を適用すると、予測式は式 2-4 となる。

$$\overline{dur(t)} = \frac{1}{4} \sum_{i=t-4}^{t-1} dur_o(i) \frac{1}{3} \sum_{j=1}^3 \frac{f_j(t)}{f_j(i)} \quad (2-4)$$

$f_1(i)$: i 番目の音節と同じ種類の音節の平均継続時間

$f_2(i)$: 先行音節が $i-1$ 番目の音節と同じ種類である音節の平均継続時間

$f_3(i)$: 二つ前の音節が $i-2$ 番目の音節と同じ種類である音節の平均継続時間

文頭から数えて m 番目の音節の継続時間を予測する場合、直前の四つの音節継続時間 $dur_o(i)$ をそれぞれ三つの要因を考慮して補正し、補正されたマッチング区間長の平均値を予測するものである。三つの要因とは、(1) 各音節の種類、(2) 先行音節の種類、(3) 二つ前の音節の種類、である。予め学習データを用いて、音節の種類ごと

の平均継続時間、それ自身の音節の種類が異なっても先行音節が同じ種類である音節の平均継続時間、同様に二つ前の音節が同じ種類である音節の平均継続時間をそれぞれ求めておき、 f_1 , f_2 , f_3 とする。予測したい n 番目の音節の平均継続時間との比を上記三つの要因毎に求め、これを全要因で平均した値が、継続時間を補正する際の補正係数となる。

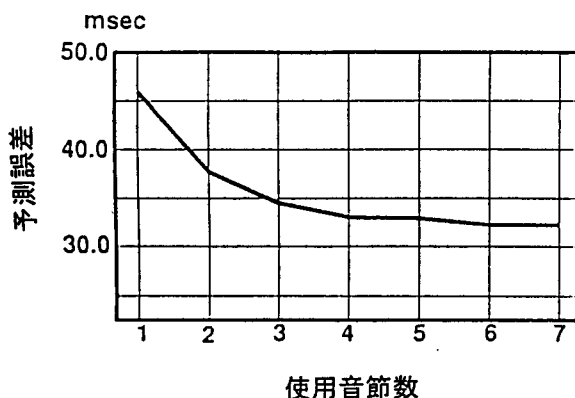


図 2-3 使用音節数と予測誤差との関係

2.1.3 音節継続時間予測法を用いた音節継続時間制御

上記の制御を連続音声認識システムに組み込み、認識性能への効果を確認した。

(I) 音節 HMM を用いた連続音声認識システム

ここでは、音節を音声単位とした連続音声認識システムを用いて、本音節継続時間制御法の効果を確認した。このシステムではまず、音節スポッティングにより音節ラティスを出力し、次に単語辞書を用いて音節ラティスを連結して単語ラティスを作り、最後に単語間 bi-gram を用いて、文認識結果を出力する (大黒 1988)。音節から単語、単語から文を作る過程は left-to-right の処理になっている。なお、音節 HMM には、文献 (中川 1989) で提案された離散分布継続時間制御を用いている。この継続時間は、HMM 学習に用いたデータベースから各状態に留まった継続時間の頻度分布を求めておき、これを用いて制御されるものである。

(II) 従来の認識結果における音節継続時間分布特性の分析

本節の冒頭で、認識誤り結果において、音節継続時間が不自然にばらついていることを述べたが、この状況を上記システムによって確認した。各文毎に、認識結果における音節スポッティング区間のばらつきを求め、正しく認識した文と認識誤り文との比較を行った。本実験では、各文毎に求めた音節スポッティング区間の標準偏差をばらつきの指標とした。たとえ、正しく認識した文であっても、先に述べた音節継続時間を左右する要因の影響により、継続時間はばらついていると予想される。そこで、継続時間そのものの標準偏差と要因によるばらつき分を除去するために、継続時間を前節で求めた要因を用いて補正し、補正後の継続時間の標準偏差を調べてみた。補正には以下の式 2-5 を用いた。補正前の各認識結果における音節継続時間の標準偏差を図 2-4 に、補正後の標準偏差を図 2-5 に示す。

$$dur'_o(i) = dur_o(i) \frac{1}{3} \sum_{j=1}^3 \frac{f_j(all)}{f_j(i)} \quad (2-5)$$

$dur_o(i)$: 既に認識された i 番目の音節の継続時間

$dur'_o(i)$: 補正後の i 番目の音節継続時間

$f_j(all)$: 要因 j における全音節データの継続時間の平均値

$f_j(i)$: 要因 j における i 番目と同音節の継続時間の平均値

その結果、次のことがわかった。

- ・ 補正前に比べ、正認識文の標準偏差はより小さくなっている。しかし、認識誤り文では、誤った音節で補正されるため、補正前に比べ小さくならない。従って、正しく認識された文と認識誤り文とのばらつきは補正前に比べより大きくなっている。
- ・ 音節継続時間を正しく設定することによる認識率が改善の可能性を調べるために、標準偏差により正しく認識された文と認識誤り文とが分離できるかを調べる。仮に 40ms に閾値を設定した場合、閾値以上の標準偏差を持つ文、つまりリジェクトされる文は、正しく認識された文では 12 文、認識誤り文では 72 文あった。この差

の文数だけ、文認識率が改善される可能性があると考えられる。

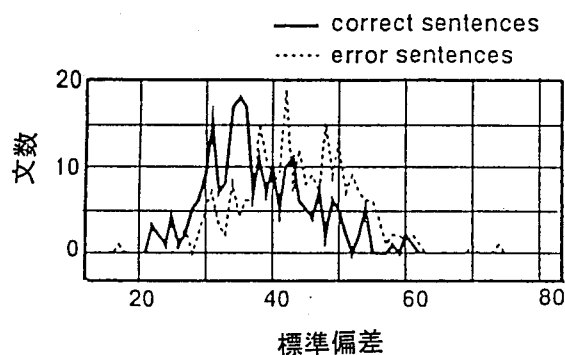


図 2-4 各文における音節継続時間の標準偏差（補正前）

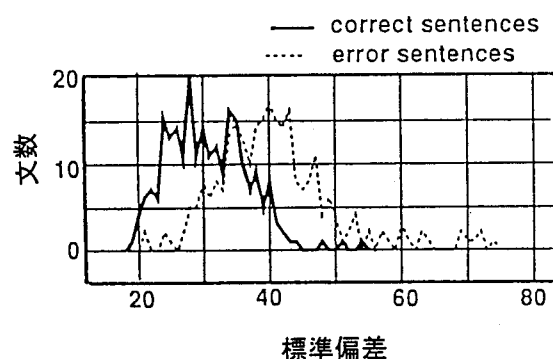


図 2-5 各文における音節継続時間の標準偏差（補正後）

（Ⅲ）音節継続時間予測を用いた認識システム

認識誤り文における音節継続時間の不自然なばらつきを軽減するため、前節で述べた連続音声認識システムをもとに、本音節継続時間予測法を用いたシステムを構築した。まず、音節ラティスから単語候補を構成する際に、左から順に音節継続時間予測を行い、互いに適切な音節継続時間を持つラティスを選択して連結していく。また単語を構成している全音節の平均継続時間の総和を単語平均継続時間とみなすことで、

単語ラティスから文候補を構成する際にも単語継続時間制御を行う。すなわち、音節の場合と同様に左から順に単語継続時間を予測し、互いに適切な単語継続時間を持つラティスを選択して文候補を出力する。実際の予測式、選択される音節及び単語の継続時間の範囲は式 2-6、式 2-7 で与えられる。本実験では式中の β 値としては、正しい認識文における補正後における音節継続時間の標準偏差の平均値である 31msec を用いた。システムのブロック図を図 2-6 に示す。

[音節継続時間予測式と音節継続時間の範囲]

$$\overline{dur(t)} = \frac{1}{4} \sum_{i=t-4}^{t-1} dur(i) \frac{1}{3} \sum_{j=1}^3 \frac{f_j(t)}{f_j(i)} \quad (2-6)$$

$$\overline{dur(t)} - \beta < \overline{dur(t)} < \overline{dur(t)} + \beta$$

[単語における継続時間予測式と単語継続時間の範囲]

$$\overline{dur^w(t)} = \frac{1}{4} \sum_{i=t-4}^{t-1} dur^w(i) \frac{1}{3} \sum_{j=1}^3 \frac{f_j^w(t)}{f_j^w(i)} \quad (2-7)$$

$$\overline{dur^w(t)} - \beta \times K(t) < \overline{dur^w(t)} < \overline{dur^w(t)} + \beta \times K(t)$$

$\overline{dur^w(t)}$: t 番目の単語の継続時間予測値

$\overline{dur^w(j)}$: j 番目の単語で観測された単語継続時間

$f_j^w(t)$: 下記式 2-8 で定義される t 番目の単語の平均継続時間

$$f_j^w(t) = \sum_{i=1}^{K(t)} f_j(i) \quad (2-8)$$

$K(t)$: t 番目の単語の総音節数

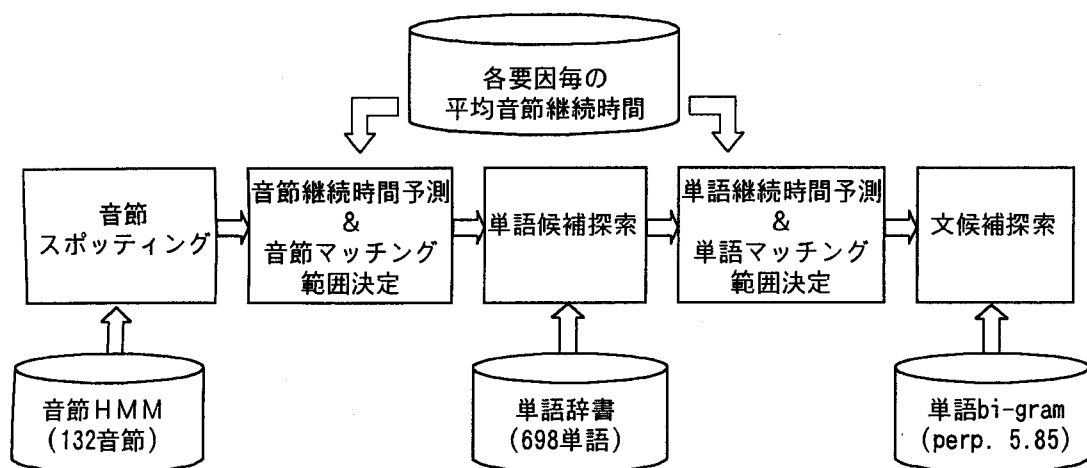


図 2-6 音節継続時間予測を用いた認識システム構成

2.1.4 音節継続時間制御法が認識性能に及ぼす効果

表 2-7 に示した音節 HMM と音声データを用いて、本音節継続時間制御法を導入した認識システムの性能を評価した。2.1.2 節における実験で、直前の音節継続時間を用いて現在の音節継続時間を予測する際に、複数の要因を用いて音節継続時間を補正する方が、要因をただ一つだけ用いて補正するよりも予測性能が向上することを示した。ここでは、要因数の違いにより認識性能に及ぼす効果をみるため、1 要因のみ使用した場合と、3 要因全て使用した場合との認識率を比較した。

従来、学習データにおける音声単位の継続時間を用いて入力における HMM の継続時間を制御する方法として、学習データの音素の種類ごとに継続時間の最小値、最大値を調べておき、この最小値と最大値を用いて音素の継続時間を限定する方法が提案されている（松永 1989、Singer1992）。ここでは、学習データのみから求められた音節継続時間の最小値と最大値を用いて継続時間を制御する従来方法との比較も合わせ、以下 4 種類の制御法の比較を行った。

- (a) 音節継続時間制御をしない場合 [従来法]
- (b) HMM の学習に用いた全データから、音節の種類ごとの最大値、最小値を求め、マッチング区間を左記の最大値から最小値の間に限定した場合 [予測なし制御法]

- (c) 本音節継続時間予測法において、用いる要因を当該音節の種類のみにした場合。
つまり、式 2-4 において、 $i = 1$ とした場合 [本制御法 1]
- (d) 本音節継続時間予測法において、3 要因すべて用いた場合。つまり、式 2-4 の全ての要因を利用した場合 [本制御法 2]

単語の認識誤り率表 2-8 に示す。本音節継続時間予測法により音節継続時間を制御することで、使用する要因数に無関係に認識性能は向上した。3 要因使用した場合で、文認識率が 4 %改善された。

表 2-7 認識実験条件

分析条件	サンプリング周波数	12KHz
	分析窓長	21.3msec
	フレーム周期	10msec
	パラメータ	10次LPCメルケプストラム 10次 Δ LPCメルケプストラム Δ パワー
音節HMM	モデル数	132音節
	モデル構成	4状態、単一正規分布
	継続時間	離散確率分布制御(中川1989)
音声データ	モデル学習用	男声10名、Aセット50文 男声6名、Bセット50文
	平均音節継続時間学習用	男声1名、Aセット50文
	認識用	男声10名、Aセット50文中で 音声区間検出が正しい442文
認識システム	単語辞書	689単語
	単語パープレキシティー	5.85
	平均単語数／文	14.4

表 2-8 本音節継続時間制御法が認識誤り率改善に及ぼす効果
[単語認識誤り率%]

話者	従来法	予測なし制御	本制御(要因1)	本制御(要因3)
1	33.4	31.2	26.7	24.4
2	36.3	36.3	34.1	31.8
3	31.8	34.1	31.8	27.1
4	31.8	31.8	31.8	31.8
5	17.8	17.8	13.3	15.6
6	43.3	41.0	38.7	36.4
7	27.1	27.1	27.1	25.0
8	45.5	47.8	47.8	45.5
9	29.5	27.1	27.1	27.1
10	31.8	31.8	27.1	24.6
平均	32.8	32.6	30.6	28.9

2.1.5 音節継続時間予測法を用いた状態継続時間予測

認識処理と同時に逐次的に継続時間の制御を行うために、2.1.2 節で述べた方法で求められた音節継続時間予測値を用いてHMMの各状態の継続時間を予測し、各状態の認識マッチング区間を制限する方法を提案する。

(I) 状態継続時間と音節継続時間の関係

音節の継続時間を用いることで認識率の向上を確認できたが、HMMの状態の継続時間を制御することでさらに詳細に制約を与えることができる。音節継続時間を予測する場合には、音節継続時間が発声速度に大きく影響される、言い換えれば、一発声内の発声速度が一定であれば音節継続時間はほぼ一定には保たれることを前提に、一発声内の既に認識された音節の継続時間からまだ認識されていない音節の継続時間を予測することを基本方針とし、さらに各々の継続時間に影響を及ぼす様々な要因を考慮することで予測精度を高めることを行った。しかし、音節よりさらに小さな単位となる音素の継続時間では、発声速度の変化により伸縮が異なる（比企 1967）。また、文中の韻律的構造や韻律句内の位置による音素継続時間の変化の存在も指摘されてい

る (勾坂 1980、海木 1990、Takeda1990)。従って、状態継続時間は、状態を含んでいる音節の種類や各状態の音節内での位置などの違いで、発声速度に受ける影響の大きさは異なると思われる。

状態継続時間制御ためのモデル化を行うために、状態継続時間と音節継続時間との関係を調べた。ある状態が音節継続時間との相関が強い場合は、その状態継続時間は発声速度により変化しやすい。逆の場合は、発声速度にかかわらず状態継続時間が決定されることになる。1 音節が 4 状態で構成されている音節 HMM を用いて認識を行い、正しく認識した結果を対象に、各状態のマッチング区間と音節マッチング区間との相関を調べた。発声様式の違いを考慮するために、単語発声のデータと、文発声のデータとを別々に調べた。単語データベースとしては、電子協 100 都市名の男声 48 名分を、文データベースとしては、ASJ 連続音声データベース A セット 50 文の男声 6 名分を用いた。

表 2-9 は、状態と音節の継続時間の相関が 0.7 以上であった音節の全音節に対する割合を、状態毎に示している。文発声と単語発声とは同傾向であったので、単語発声の結果は全話者を対象にした場合のみについて示す。表 2-9 より、次のことがわかる。

- ・ 最終の第 4 状態目は、発声速度と相関の高い音節が多く、第 2、第 3 状態目は、発声速度との相関が低い傾向があり、この傾向は発声様式 (単語か文か) に依存しない。第 4 状態目でも、単母音、長母音や、／ふ／、／す／など、無声化しやすい音節に関しては、相関が低かった。
- ・ 上記の状態継続時間と発声速度との関係は、どの話者にも共通した傾向である。

図 2-7 は、相関が最も高かった第 4 状態と、最も低かった第 2 状態の、全音節の相関値の度数を示している。予測されたように、状態により発声速度の依存性は異なっていた。

表 2-9 状態と音節との継続時間相関が 0.7 以上であった音節の割合

発声形式	話者	最初の状態	2番目の状態	3番目の状態	最後の状態
文発声	1	0.44	0.2	0.31	0.74
	2	0.43	0.26	0.41	0.64
	3	0.45	0.29	0.43	0.68
	4	0.39	0.28	0.32	0.61
	5	0.51	0.21	0.36	0.63
	6	0.49	0.21	0.35	0.66
	全話者	0.39	0.13	0.27	0.56
単語発声	全話者	0.22	0.21	0.37	0.53

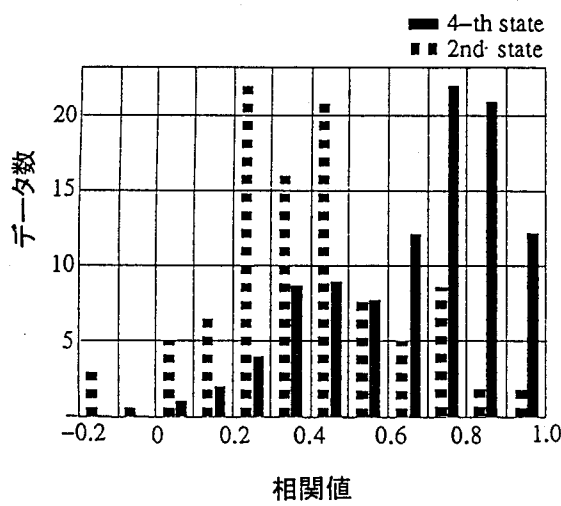


図 2-7 状態継続時間と音節継続時間との相関値

(II) 状態継続時間予測式

従来、HMMを用いず学習音声から音素毎の音響特徴量を学習し、DPマッチングにて認識を行う認識方式において、各音素の継続時間が発話速度の変化により伸縮が異なることを考慮し、学習音声における平均母音継続時間の一次回帰式として音素継続時間を推定し、音素継続時間を制御する方法が提案されている(大坂 1993、松尾 1989)。2.1.2 節で提案した音節継続時間予測法により予測された音節継続時間を

与え、同様に次式を用いて状態継続時間を予測することを試みた。

$$\overline{dur^{st}(t,s)} = \overline{dur^{sy}(t)} \times a(t,s) + b(t,s) \quad (2-9)$$

$\overline{dur^{st}(t,s)}$: t 番目の音節における s 番目の状態の継続時間予測値

$\overline{dur^{sy}(t)}$: t 番目の音節の継続時間予測値

$a(t,s), b(t,s)$: t 番目の音節の s 番目の状態における重み係数

$a(t,s), b(t,s)$ は、予め学習データを用いて、最小 2 乗法によって状態毎に決定する。 $\overline{dur^{st}(t)}$ は、2.1.2 節で提案した音節の継続時間を用いて式 2-10 のように予測される。たとえば、孤立単語発声の場合などは、認識する前に単語毎の音節の種類と音節数とが明らかであるので、全音声区間と各音節の種類と全音節数を用いて式 2-11 のように予測することもできる。

$$\overline{dur^{sy}(t)} = \frac{1}{m} \sum_{i=t-m}^{i-1} (\overline{dur_o(i)} \sum_j \alpha_j \frac{f_j(t)}{f_j(i)}) \quad (2-10)$$

$\overline{dur_o(i)}$: 既に認識されている i 番目の音節継続時間

$$\overline{dur^{sy}(t)} = L \frac{\sum_j \alpha_j f_j(t)}{\sum_{i=1}^N \sum_j (\alpha_j f_j(i))} \quad (2-11)$$

$f_i(t)$: j 番目における t 番目の音節と同じ条件の音声継続時間の平均値

α_j : j 番目の要因の重み

L : 音声区間

N : 音声区間に含まれる音節数

(III) 状態継続時間予測法の評価

本状態継続時間予測における予測性能を単語発声、文発声データを用いて評価した。ここでは、各状態に対して観測された継続時間と式 2-9 を用いた予測値との誤差を、観測された状態継続時間と各音節の各状態毎の平均値との誤差と比較することで評価

した。状態の観測された継続時間としては、各状態の正しい継続時間つまり継続時間予測値の目標値を設定する目的で、正しい認識結果における状態継続時間を用いた。また、式 2-9 における重み a, b は、これらの観測された状態継続時間を用いて算出した。

単語発声の場合は、モデルの学習および重み a, b の学習データとして、2-1-2 節で用いた電子協 100 都市名を用い、評価データには、学習時と同データセット（TWセット）、Tセットとは異なる 100 単語、男声 9 名のセット（EWセット）の 2 種類のセットを用い、学習セットと異なるデータにおいての有効性も合わせて評価した。文発声の場合も、同じ趣旨で、学習データとして 2.1.2 節で用いた ASJ 連続音声データの Aセット 50 文を用い、評価データには、学習時と同セット（TSセット）、ASJ 連続音声データの Hセット 50 文、男声 6 名分（ESセット）の 2 種類のセットを用いた。

表 2-10、表 2-11 に、各々、単語発声および文発声における予測誤差分布の標準偏差と最大誤差値を示す。その結果、発声様式、データセットに無関係に、本予測方法を用いた方が、予測誤差の標準偏差、最大誤差値が小さくなることわかる。最大誤差値が小さくなるということは、認識の際に、スポッティング範囲をより制限できるということであり、認識の際の効果が期待できる。

表 2-10 単語発声における予測誤差分布の標準偏差と最大誤差値 [msec]

データセット	予測方法	標準偏差	最大誤差値
TWセット	各状態の平均値を利用	33.6	340
	本予測方法	27.1	220
EWセット	各状態の平均値を利用	35.1	310
	本予測方法	30.9	240

表 2-11 文発声における予測誤差分布の標準偏差と最大誤差値 [msec]

データセット	予測方法	標準偏差	最大誤差値
TSセット	各状態の平均値を利用	21.8	220
	本予測方法	16.6	150
ESセット	各状態の平均値を利用	27.4	310
	本予測方法	20.3	240

2.1.6 状態継続時間制御

2.1.5 節で述べた状態継続時間予測法を用いて、認識時に各状態継続時間を制御する方法について提案した。予め、音節HMMモデル学習後に、モデル学習用のデータを用いて正解既知の状態マッチングを行い、そのマッチング区間を各状態の正しい継続時間とみなし記録しておく。さらに、同データを用いて、式 2-9 と式 2-10、または式 2-9 と式 2-11 に従って各状態の継続時間を予測し、予測値と上記の正しい状態継続時間との誤差が最小になるように、最小 2 乗法にて、重み a , b を求めておく。認識時には、フレーム毎に、次に認識しようとする音節の各状態の状態継続時間を学習時と同様に予測し、予測値に一定区間前後した範囲を各状態のスポッティング範囲として認識を行う。

2.1.7 状態継続時間制御法が認識性能に及ぼす効果

(I) 認識実験条件

上記状態継続時間制御法が認識性能に及ぼす効果を調べるために、認識実験を行った。ここでは、認識システムとして、オートマトン制御型の認識システムを用いた。これは、予め認識対象となる単語および文を、音節HMMを単位としたオートマトンで表現し、辞書化しておく。認識時には、この単語および文辞書を読み込み、Viterbi アルゴリズムに従って、入力フレームに同期した認識処理を行うものである。システムの構成図を 図 2-8 に、システムの条件を表 2-12 に示す。

なお、学習データには、状態継続時間予測性能を評価した際の学習データと同セッ

ト (TWセットおよびTSセット) を用い、評価データには、単語発声のものは2.1.5章の評価セットと同セット (TWセットおよびEWセット) を、文発声のものは、オープンセットであるESセットのみを用いた。また、ESセット (ASJ 連続音声データHセット) については、各文の各文節の終端で発声が途切れ、文が最後まで発声されていなくても、発声された部分までの認識が可能となるようにオートマトンで表現された文辞書を作成した。作成された辞書の例を表 2-13 に示す。このように、文が最後まで発声されていない文を1文と数えると、ESセット 50 文は 354 文になり、前部分がほとんど同じである類似文が多く含まれる辞書が作成される。今回の認識結果は、上記の 354 文辞書を用いた場合のESセット 50 文に対する認識率である。また、各状態の継続時間の範囲は、正しく認識した結果における各状態継続時間の 90%以上が範囲内になるように設定した。

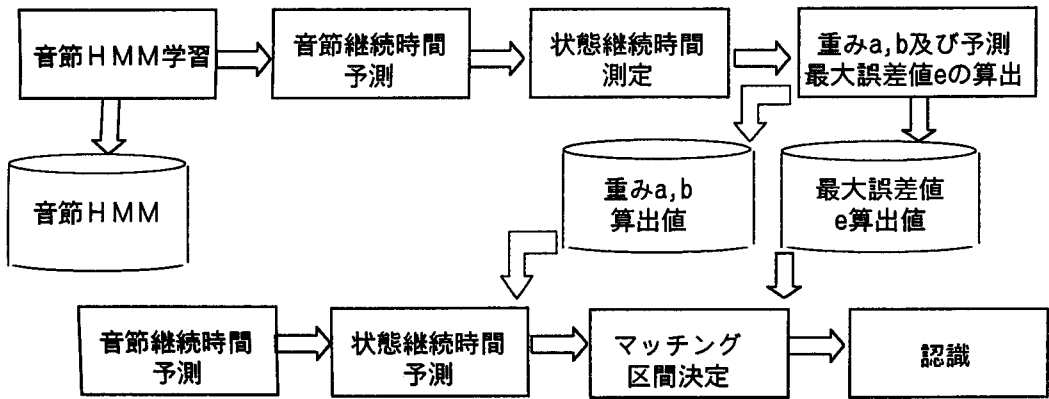


図 2 - 8 システム構成図

表 2-12 認識実験条件

音節HMM	モデル数	132音節
	モデル形式	FVQ-HMM (坪香1991)
	モデル構成	4状態、単一正規分布
	コードブックサイズ	LPCメルケプストラム256 LPC Δ メルケプストラム256 Δ パワー64
単語データ	重みa,b、予約誤差e学習用	電子協100都市名 48男声(TWset)
	認識用(closed)	学習用と同データ(TWset)
	認識用(open)	異なる9男声(EWset)
文データ	重みa,b、予約誤差e学習用	音響学会研究用データAセット 50文、6男声(TSset)
	認識用(open)	音響学会研究用データHセット 50文、6男声(ESset)

表 2-13 受理可能となる部分文の例

h-49-1	老人は
h-49-2	老人は漁夫として
h-49-3	老人は漁夫としてすばらしい
h-49-4	老人は漁夫としてすばらしい人だと
h-49	老人は漁夫としてすばらしい人だと思ふ

(II) 認識実験

上記システムを用いて、単語発声および文発声データを用いた認識実験を行い、本状態継続時間制御法を評価した。実験では、本状態継続時間予測法の優位性を明らかにするために、継続時間制御を行わない場合と、発声速度を考慮せずに各状態毎に共通の継続時間を設定して制御を行う場合と、発声速度を考慮した本予測法を用いて継続時間制御を行う場合の下記三つの方法を比較した。各状態に共通の継続時間を設定する場合のマッチング範囲は、各状態毎の継続時間の最小値から最大値の範囲とした。また本予測法におけるマッチング範囲は、学習データにおける予測された各状態継続

時間を中心に、予測値と実際の継続時間との差の最大値分だけ前後した範囲を用いた。
認識結果を表 2-14 に示す。

表 2-14 本状態継続時間制御法が文認識率に及ぼす影響 [認識誤り率%]

	制御なし	状態毎に制御	本制御
TWset(単語)	2.1%	2.0%	1.6%
EWset(単語)	8.9%	8.2%	7.8%
ESset(文)	5.0%	4.7%	2.7%

結果から次のことが明らかとなった。

- ・データベース、発声様式に無関係に、継続時間制御を行うことで認識率は向上した。
本制御法を用いた場合が最も向上した。
- ・本制御法の場合には、他の制御法に比べて、従来認識が比較的困難であった類似語
(単語発声の場合は、「きりゅう」と「ちりゅう」、または「ごぼう」と「むこう」
など、文発声の場合は、表 2-13 のおける h-49-4 と h-49 など) に対し、認識率が
向上した。
- ・本制御法により、認識率を低下させずに継続時間を制御できた。

また、本状態継続時間制御により、認識処理時間の削減効果も期待できる。各継続
時間制御を行った場合の、一状態あたりの平均継続時間範囲を調べた結果を表 2-15 に
示す。この結果が示すように、各状態毎に共通の継続時間を設定して制御を行う場合
と比較して、単語発声の場合で 25%、文発声の場合で 50%に、継続時間を短縮してい
る。

表 2-15 1 状態あたりの平均継続時間範囲 [msec]

	制御なし	状態毎に制御	本制御
EWset(単語)		80msec	60msec
ESset(文)		60msec	30msec

2.1.8 まとめ

本論文では、音節HMMを用いた連続音声認識において、既に認識された部分の音節の継続時間と音節の継続時間を左右する要因とを用いて、次に認識する音節の継続時間を予測する方法を提案した。また、予測された音節継続時間を用いて音節マッチング区間を制限することにより、認識性能の向上を確認した。さらに、この予測法を用いて、状態の継続時間を適応的かつ逐次的に予測する状態継続時間予測法を提案した。この予測法を利用して状態継続時間制御を行うことが、認識性能の向上と認識時の探索範囲の削減に有効であることを確認した。

2.2 複数の音声単位にまたがる認識誤り特性を用いた連続音声認識

2.2.1 はじめに

連続音声認識において、音素環境コンテキストによる発声の相違を吸収するためにコンテキスト依存型の HMM が提案されている (Schwartz1985, Lee1990, Takami1992)。これらを用いてもなお、コンテキストに依存した認識誤りは多く、その認識誤りは話者により異なることが見受けられた。また、単語列を探索する際に探索幅を広げても、正解単語が結果候補に入っていない。これらの原因として次の二つが考えられる。

まずは、音響モデルである HMM を統計的手法で学習する際に考慮される発声変形要因が、認識時に発生する変形要因を全て満たしていないことである。たとえばコンテキスト依存型の学習では、考慮しているコンテキストは音素環境だけであり、話者、発声速度、発声文の長さ、といった要因が考慮されていない。認識誤りを軽減するためには、発声変形の要因を分析し、各要因毎に様々な環境のデータを準備し、これを用いて HMM を学習する必要があるが、実際にはこれらの要因を網羅することも、大規模なデータを準備することも容易ではない。

次に考えられる原因としては、HMM の精度の不完全さである。使用時の様々に異なる発話条件に応じて HMM の学習条件を最適に設計するのは困難である。通常の連続音声認識では、予め、発声中の小さな単位 (例えば、音素、音節) ごとに HMM を作成しておき、この HMM に与えられたラベル系列が認識結果として出力される。学習時と異なる音声が入力されても、それに対応して、HMM の構造、単位、状態数は変更されない。このため、学習時と異なる発声の入力に対して正しいラベル系列を出力しようとする、音響的特徴の変化の少ない部分が必要以上の数の HMM に対応したり、特徴が大きく変化している部分が一つの HMM に対応するような不自然なマッチングが起こり得る。このような発声変形を吸収するためには HMM の構造、単位、状態数などの条件を独立に考慮した HMM の設計 (Singer1996) が必要であるが、全学習データ量、学習に要する時間などを考えると、さらに工夫が必要である。

以上のように様々なコンテキストや発話条件の違いに起因する問題全てを、音響モデルの学習問題として扱うのは得策ではない。本 2.2 節では、既に学習された音響モ

デルの不備を救済する方法、すなわち、HMM ラベルとしては正解候補に入っていない音素系列を新たに正解候補に加えることが可能な方法を提案し、認識できなかった音素系列を認識可能にする機能について述べる。ここでは、新たに音素系列を追加する二つの方法を提案した。一つは、認識後に認識結果のラベル系列上で正解音素候補系列の追加を行う方法である。この方法では、認識結果の状態系列と認識誤り特性テーブルに保管されている認識誤りの状態系列とを比較し、認識誤り状態系列と同じ状態系列が認識結果に含まれていた場合に、その部分に対応する正解状態系列に置き換えたものを結果候補に追加する。他の一つは、発音辞書上で正解音素候補系列を追加する方法である。発音辞書上に登録してある単語状態系列とテーブルに保管されている認識誤り特性の正解状態系列を比較し、正解状態系列と同じ状態系列を含む語彙があった場合、その部分に対応する認識誤り状態系列で表現したものを、その語彙の可能性のある発音と追加して発音時書上に追加登録する方法である。以下各々の方法の提案とその効果を評価した。

2.2.2 現在の音響モデルによる認識誤り分析

まず、音響モデルを使用状況に合わせて適応的に変形する学習や、単語探索における探索幅を増加する工夫を行っても、なお残る認識誤りに対応できるモデルを提案するために、現在の音響モデルが抱える認識誤りの特徴を調べ、新たなモデルが解決すべき問題点を探る。

従来、ある発声変形の要因のみに注目して認識誤りを改善するために、音響モデルを再学習する方法が提案されている。たとえば話者の違いによる認識性能を改善する技術として話者適応技術が提案されている (Okura1992, Lee1991, Leggeth1994)。これらの話者適応技術では、各HMM単位毎に各話者のスペクトル特性を適応させることで、認識性能の向上に対する効果が確認されている。しかし、特定の話者の限られた発声から、HMM の構造や状態数まで話者に合わせて変更することは難しいため、話者適応後のモデルを用いても、特定話者の性能には及ばない。

現行の話者適応が救済できる認識誤りの範囲を見定め、音響モデルの問題点を明らかにするために、話者適応により不特定話者用HMMを改良した後の認識誤りの特徴

を話者適応前の認識誤りの特徴と比較した。話者適応としては、少数データでの適応性能が既に確認されている移動ベクトル場平滑化法（VFS）（Okura1992, Takami1992）を採用した。さらにコンテキスト環境を考慮した平滑化をこれに加え、コンテキストに依存した発声変形に対応できるようにした。実験条件、話者適応条件を表 2-16 に示す。図 2-9 は様々なビーム幅に対する話者適応前後の話者 6 名の平均音素認識率を比較したものである。話者適応により約 20% 前後の認識率の改善が確認されたが、ビーム幅が一定値以上になると、認識率は飽和することが認められた。

表 2-16 認識実験条件

タスク	国際会議参加問い合わせ
音響モデル	不特定話者音素HM-Net 延べ状態数201、混合分布数10、継続時間制御なし
認識方式	One-pass DPによるN-best候補探索
HMM構造決定用データ	孤立発声された5240重要語の偶数番目、1名
再学習データ	文節発声、音素バランス50文、285名
話者適応法	移動ベクトル場平滑法(VFS)
話者適応条件	近傍数6、平滑化係数20
話者適応用データ	32文(754音素)／話者
評価用データ	33文(637音素)／話者
話者	男女3話者ずつ

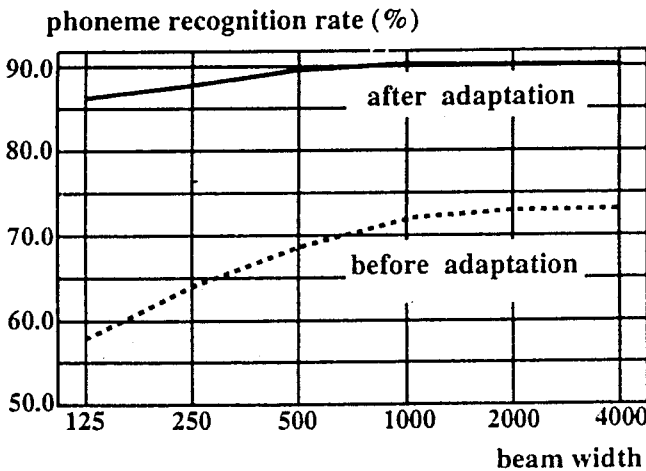


図 2-9 ビーム幅と音素認識率の関係

次に、話者適応技術によりなお解決できない認識誤り特性を明らかにするために、話者適応前後の認識誤り特性を比較した。まず、認識誤り特性を正確に抽出するため、Viterbi-alignment による正解音素系列の各音素のマッチング区間と認識結果が与える音素区間とを比較することで、認識誤り部分を決定した。具体的には、正解音素系列と認識結果の音素系列との各音素の境界となるフレームを比較して、音素の種類に関係なく両者の境界フレームが同じ値になるところに印を付け、印にはさまれた部分を一区間とした。同じ区間内の正解と認識結果との音素系列の種類が異なっている場合、その区間を認識誤り区間とみなし、正解と認識結果との音素系列ペアを抽出した。出現数の多い認識誤り特性を上位 6 位までを表 2-17 に示す。話者適応前には、/u/→/e/ や/o/→/u/や/t/→/k/のように単音素間の認識誤り特性の頻度が上位を占めているが、話者適応後には、/o g o/→/o/や/sh i ts/→/s/や/i g i/→/i/のような複数音素にまたがる認識誤り特性の頻度が相対的に高くなることがわかる。

さらに、全認識誤りの中の、複数音素にまたがる認識誤り特性の割合が話者適応前後でどのように変化しているかを調べた。表 2-18 が示すように、話者適応により複数音素にまたがる認識誤り特性の絶対数(表中の括弧内の分子)は減少しているが、複数音素にまたがる認識誤り特性の割合は増加している。

表 2-17 出現数の多い認識誤り特性
[/正解系列/ ⇒ /認識誤り系列/ (総出現数)]

話者適応前	話者適応後
/u/ ⇒ /e/ (14)	/o g o/ ⇒ /o/ (11)
/o/ ⇒ /u/ (12)	/d/ ⇒ /r/ (10)
/t/ ⇒ /k/ (11)	/sh i ts/ ⇒ /s/ (9)
/wa/ ⇒ /a/ (10)	/i g i/ ⇒ /i/ (7)
/i/ ⇒ /j/ (9)	/e/ ⇒ /a/ (7)
/w/ ⇒ /r/ (9)	/o/ ⇒ /u/ (3)

表 2-18 複数音素にまたがる認識誤り系列の全認識誤りに対する割合

話者適応前	話者適応後
59.7% (350/586)	76.4% (158/206)

以上の分析結果が示すように、話者適応により認識誤りは減少するが、適応に用いた文と同じ内容の文を認識した場合でも認識誤りの修復には限界がある。特に各音素が一对一对応した認識誤りに対する有効性に比べ、複数音素にまたがる認識誤りに対しては更に対策が必要である。

2.2.3 認識誤り特性を用いた音素候補系列追加モデル

前節で述べた認識誤りの救済方法として、本節では、与えられた音響モデル（ここでは音素 HMM）とのマッチングで得られた認識結果の音素候補系列を入力として、音素系列を認識結果候補に新たに追加するモデルを提案する。ここで追加する音素系列は、認識誤り特性を用いて予測する。この音素候補系列追加モデルは、学習データを用いて認識誤り特性を抽出する部分と、学習時に抽出された認識誤り特性を用いて音素系列を追加する部分とから成り立つ。

（Ⅰ）複数音素にまたがる認識誤り特性の抽出

音素HMMでは対応できない発声変形は、誤った音素系列を認識結果として出力する。さらに、話者特有の発声変形は、同じ認識誤り音素系列を頻繁に出力する可能性がある。従って、この認識誤りの傾向を把握できれば、音響モデルでは吸収できない発声変形が生じた場合でも、認識結果の音素系列に対応した正解音素系列を予測できる。

認識誤りの傾向を把握するために、認識誤りラベル系列とそれに対応する正解系列とのペアを抽出する。従来にも、認識誤りの特性を用いて、認識誤りを訂正する研究がなされており（新津 1979, Araki1994, Ito1992, Ringer1996）、これらの研究におけ

る認識誤り部分の抽出法は、誤った音素または音節と、正解の音素または音節とをなるべく一対一に対応させ、場合によっては2音素以内の挿入、脱落を考慮しながら、認識誤り系列と正解系列とのペアを抽出している。しかし、前節の実験結果から明らかなように、誤ったHMMラベル系列と正解音素系列とを比べた場合、正解の音素と認識誤りラベルとが一対一対応しているとは限らない。複数HMMで成り立つ認識誤りラベル系列と正解音素系列とが対応しており、各々に含まれる音素数は様々である。そこで、音素という単位に拘束されずに認識誤り特性を抽出するために、1つの認識誤り部分に含まれるHMMラベル数に制限を与えず、複数の認識誤りラベル系列が複数の正解音素系列に対応するように抽出した。また、発声変形の特徴をできるだけ正確に認識誤りに反映させるには、系列の表現は、なるべく小さい単位であることが良いと考えられる。そこで認識誤りを、音素ではなくHMMの状態系列として表現した。

上記の方法に従い、認識誤り特性を次の手順で抽出した。学習データを用いて認識を行い、予め Viterbi-alignment を用いて調べた正解のHMM状態系列と認識結果の状態系列とを比較することで認識誤り部分を決める。この方法は、2.2.2 節で認識誤り部分を抽出した場合と同じであるが、違いは 2.2.2 節では系列を音素で表現していたものを、状態として表現したところにある。例えば、正解音素系列 /m o sh i/ に対して /o sh u/と認識誤りを起こした場合の抽出例を図 2-10 に示す。図 2-10 では、正解状態系列 $\{s_0, s_1 \dots s_{13}\}$ と認識結果状態系 $\{s_0, s_1, \dots s_{20}\}$ とを比較し、両者の音素系列の境界点と同じになるところ（図中の破線）に印をつけ、印を付けられた間の区間を一区間とする。一区間内の両者の状態系列を比較し、状態系列が等しくない場合に、対象となる正解状態系列と認識結果状態系列とをペアにして、テーブルに登録する。例では、/mo/→/o/及び/shi/→/shu/に相当する状態系列が登録される。全学習データについて、認識誤り状態系列と正解状態系列のペアを抽出した後、各認識誤り状態系列の信頼度を表すものとして、認識誤り状態系列の出現確率を付与し、上記ペアとともに、テーブルに登録される。ある認識誤り状態系列 $\{E_i=e_1, e_2, \dots, e_n\}$ と正解状態系列 $\{C_i=c_1, c_2, c_3, \dots, c_m\}$ (e, c は各状態)のペア (E_i, C_i) の出現確率 $P(E_i, C_i)$ は、式 2-12 で定義する。

$$P(E_i, C_i) = \frac{\text{ペア}(E_i, C_i) \text{ の出現数}}{\text{認識結果上の系列 } \{E_i\} \text{ の出現数}} \quad (2-12)$$

認識結果の状態系列{S_j}が正解であった場合は、ペア(C_i, C_j)として上記ペア(E_i, C_j)と同じように出現確率が計算されテーブルに登録される。図 2-10 のテーブルでは、pair no.が 2 と 4 のペアは正解状態系列と認識結果状態系列が同じものであり、それぞれ pair no.が 1 と 3 のペアの認識誤り状態系列が正解であった場合として登録されている。

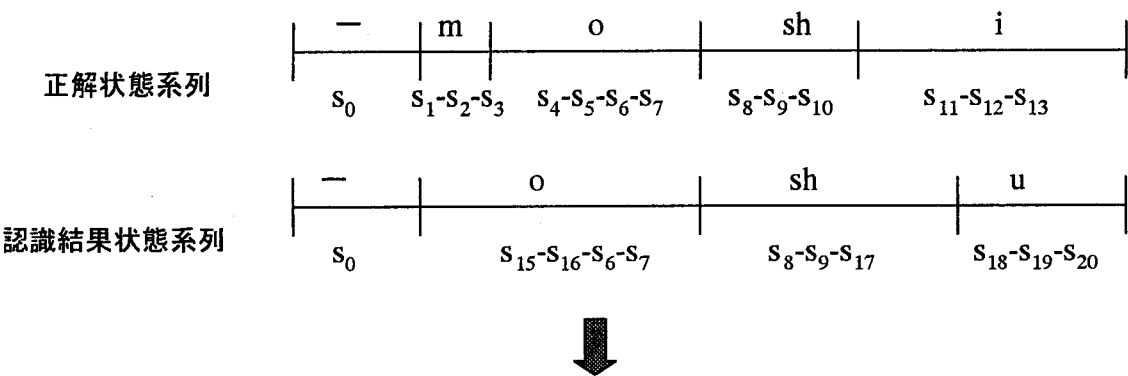


図 2-10 正解音素系列 /m o sh i/ に対して /o sh u/と認識誤りを起こした場合の抽出例

2.2.2 節で話者適応に用いた 32 文を用いて、実際に話者 6 名分の認識誤り特性を抽出した。抽出された認識誤り特性の傾向は 2.2.2 節の表 2-17 に示した話者適応後の音素表現による認識誤り特性の傾向と似ており，出現確率の高い認識誤り特性は，6 名とも複数音素にまたがるものが多かった．しかしその内容は，各話者の認識誤り特性の約半分は各話者特有の認識誤り特性であった。たとえば表 2-19 は，6 話者のうちの 2 名（男女 1 名ずつ）における上位 5 位までの認識誤り特性例である。認識誤り特性は状態系列で表されているが、傾向をわかりやすくするために表 2-19 では音素系列で表現する。この 2 名については，認識誤り特性 /o g o/→/o/は両者に共通しているが，他の認識誤り特性は各話者特有のものである。

表 2-19 出現確率の高かった上位 5 位の認識誤り特性例
[/正解系列/ ⇒ /認識誤り系列/ (出現頻度)]

話者1	話者2
/o g o/ ⇒ /o/ (66.6%)	/o g o/ ⇒ /o/ (66.6%)
/a w a/ ⇒ /a/ (60.0%)	/sh i ts/ ⇒ /s/ (60.0%)
/g o/ ⇒ /k o/ (37.5%)	/i g i/ ⇒ /i/ (40.0%)
/- d/ ⇒ /- t/ (37.5%)	/r e/ ⇒ /t e/ (40.0%)
/m/ ⇒ /d/ (18.2%)	/w a/ ⇒ /m a/ (33.3%)

(II) 音素候補系列の追加

次に、抽出された認識誤り特性を用いて音素系列の追加を行なう。ここでは二通りの追加手段を考えた。一つは、認識後に認識結果のラベル系列上で正解音素候補系列の追加を行なう方法である。この方法では、認識結果の状態系列と認識誤り特性テーブルに保管されている誤認識状態系列とを比較し、誤認識状態系列と同じ状態系列が認識結果に含まれていた場合に、その部分に対応する正解状態系列に置き換えたものを結果候補に追加する。もう一つは、発音辞書上で正解音素候補系列を追加する方法である。発音辞書上に登録してある単語状態系列とテーブルに保管されている認識誤り特性の正解状態系列を比較し、正解状態系列と同じ状態系列を含む語彙があった場合、その部分に対応する誤認識状態で表現したものを、その語彙の可能性のある発音と追加して発音時書上に追加登録する。認識時には、追加登録された発音辞書を用いて認識を行う。

どちらの音素系列追加方法においても、認識誤りを状態で表現する場合には、追加登録時に前後の音素の種類を考慮する必要がある。音素環境コンテキスト依存型のHMMでは、同じ種類の音素を表すHMMであっても、前後の音素の種類に依存して状態系列は異なる。但し、全ての前後の音素の違いに対して異なった状態系列が用意されている訳ではなく、前後の音素が異なっても、類似した状態または状態系列は共有されるように設計される場合が多い。従って、認識誤りを状態系列で表現している場合は、認識誤りと同じ音素系列が、認識結果の音素系列または発音辞書に記述さ

はなく、前後の音素に依存して状態系列が異なっていれば、置き換えは行なわれない、という条件のもとに音素系列の追加を行う。以下 2.2.4 節及び 2.2.5 節では認識結果のラベル系列上で正解系列の追加を行う方法、2.2.6 節から 2.2.8 節では発音辞書上で追加を行う方法を述べる。

2.2.4 認識結果への音素候補系列追加

本節では、認識誤り特性を用いて認識後に認識結果のラベル系列上で正解音素候補系列の追加を行なう方法の有効性を評価するために構築した、音素タイプライター型の正解音素系列追加システムについて述べる。音素タイプライター型認識システムでは通常、認識の途中で、限られた候補から正解音素をなるべく落さないようにするために、スコアの高い複数の途中結果を探索空間（ビーム）上に残しながら、認識処理を行っている。提案する音素候補系列追加モデルは、ビームの幅を大きくしてもなお落ちてしまう正解音素系列を結果に追加することで、従来限界に達していた認識性能をさらに向上させるのに有効であると考えられる。そこで、音素タイプライターを用いて音素系列探索を行い、予め学習されている認識誤り特性の頻度が大きい特性を優先的に用いて、探索された音素系列から新たな音素系列を生成し追加する。追加された音素系列が正しいかどうかを判断することで本方法の有効性を評価する。

本モデルが認識率の向上に有効であることを確認するためには、認識結果の第 1 候補の音素系列に対して正解系列を追加できることが望ましい。そこで、2.2.2 節で用いた音素タイプライターの第 1 候補の認識結果を入力し、本モデルを後処理的なアプローチで用いることで正解音素系列を探索する。認識誤り特性の出現頻度のみに基づいた N-best 探索を行った際の探索結果に正解系列が含まれているかを評価した。システム構成を図 2-11 に示し、探索手順を次に説明する。

まず認識を行う前の学習作業として、認識誤り状態系列と正解状態系列とのペアを抽出し、ペアの出現頻度とともにテーブルに登録しておく。これは前節で説明した通りである。次に正解音素系列の探索する際には、通常の認識処理を行った後、認識結果の第 1 候補の状態系列と認識誤り特性とを比較し、認識誤り状態系列と同じ状態系

列が認識結果の状態系列に存在した場合、その部分を認識誤り状態系列とペアである正解状態系列に置き換えたものを新たな認識結果候補として追加する。この作業を left-to-right に行う。一つの認識誤りに含まれる音素数に制限を与えていないことから、追加の際に使用される認識誤り状態系列長はまちまちで、お互いに一部が重なっている、または包含関係にある。そのため、同じ入力音素系列から、様々な認識誤り特性の組合せからなる音素系列が複数個作成され、探索結果候補は指数関数的に増加する。そこで、認識誤り特性の出現確率に基づいたスコアを各々の認識結果候補に付与し、スコアに基づいた N-best 探索を行うことで、多くの候補を絞りながら最終的に最も信頼できる探索結果を第 1 候補として出力する。各認識結果候補に付与されるスコアは認識誤り特性の出現確率に基づいて次のように算出される。

今、ある結果状態系列を $\{s_1, s_2, \dots, s_n, E_i, \dots\}$ として、この部分状態系列 $\{E_i\}$ が誤っており、正解は状態系列 $\{C_j\}$ である確率が $P(E_i, C_j)$ であるとする。この場合、状態系列 $\{s_1, s_2, \dots, E_i\}$ まで探索処理が進んだ段階で、新たに追加された結果状態系列 $\{s_1, s_2, \dots, C_j\}$ に与えられスコア $S(s_1, s_2, \dots, s_n, C_j)$ は、

$$S(s_1, s_2, \dots, s_n, C_j) = S(s_1, s_2, \dots, s_n) \times P(E_i, C_j) \quad (2-13)$$

となる。 $S(s_1, s_2, \dots, s_n)$ は、部分的な状態系列 $\{s_1, s_2, \dots, s_n\}$ に与えられているスコアであり、状態系列 $\{s_1, s_2, \dots, s_n\}$ の中の全ての状態が認識誤り特性により置き換えられていない場合には、その状態系列の出現確率は 100% であるため、状態系列の長さに関係なく $S(s_1, s_2, \dots, s_n) = 1$ となる。一部が置き換えられている場合は、 $S(s_1, s_2, \dots, s_n)$ は式 2-13 を用いて同様に算出された値をもつ。上記のスコアリングでは、誤認識状態系列が追加される度に認識候補のスコアが更新される。同じ回数だけ追加が行なわれた他の結果系列候補とのスコアを比較することにより、N-best 候補が決定される。

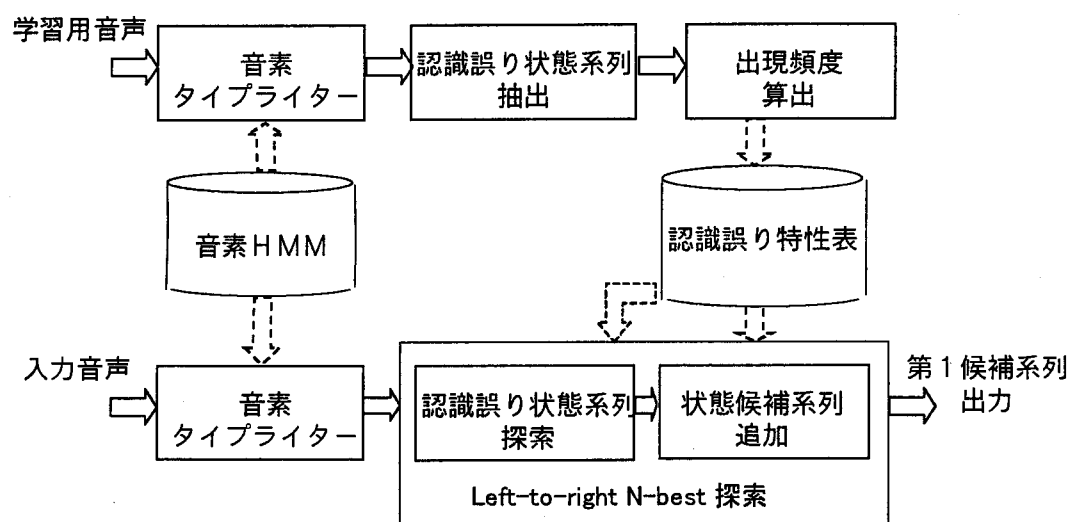


図 2-11 音素タイプライターを用いた left-to-right の N-best 探索による正解音素探索システム

2.2.5 音素候補系列追加が認識性能に及ぼす効果

(I) 正解音素系列探索モデルによる認識誤りの減少

提案した音素系列候補追加モデルの有効性を確認するために、探索結果の第 1 候補と従来の認識結果第 1 候補との音素認識率とを比較した。

実験条件は、表 2-16 とほぼ同じであり、追加条件のみを表 2-20 に示す。話者適応を用いても、また広いビーム幅を用いてもなお解決しない認識誤りが、本モデルで解決できるかどうかを確認するため、音響モデルは話者適応を行った後のモデルを用い、ビーム幅は 2.2.2 節の実験で認識率が既に飽和している際のビーム幅 4000 を用いた。話者適応法も、2.2.2 節で用いた VFS 法を用いた。

評価データは、既知文（学習データと、文は同じ発声は異なる）として、2.2.3 節で用いた評価データ 32 文を用い、文も発声も学習文とは異なる文（以降未知文という）として 18 文を用い、既知文と未知文とを別々に評価した。探索を行う前後の話者 6 名の音素認識誤り率を比較した。結果を表 2-21 に示す。既知文で平均 26.6%(9.84%→7.22%)、未知文で平均 7.5%(17.93%→16.59%)認識誤りが減少した。本正解音素系列

探索法が、評価データの種類に関係なく認識誤りを減少させ、認識結果に入らなかった音素系列を探索できることを確認した。ただ、6名の話者を比較すると、認識誤率が大きく減少した話者（例えば話者1）とあまり減少しない話者（例えば話者6）があり、効果の度合いは話者により異なっていた。

表 2-20 評価実験条件

音響モデル	VFSによる話者適応後のモデル (探索前、探索後ともに)
ビーム幅	4000
認識誤り抽出用データ	32文(754音素)／話者
評価用データ	[既知文] 33文(637音素)／話者 [未知文] 18文(629音素)／話者
話者	男女3話者ずつ

表 2-21 正解音素系列探索後の音素認識誤り率の減少

	既知文探索前	既知文探索後	未知文探索前	未知文探索後
話者1	6.83%	2.58%	13.44%	8.22%
話者2	12.97%	9.56%	21.12%	21.12%
話者3	11.95%	11.10%	17.09%	16.51%
話者4	10.76%	10.21%	17.15%	16.26%
話者5	8.19%	2.32%	15.05%	14.01%
話者6	8.37%	7.55%	23.42%	23.42%
平均	9.84%	7.22%	17.93%	16.59%

(II) 学習データ量と探索性能の関係

学習データ量と正解音素系列の探索性能との関係を調べた。未知文データを表 2-16 の条件に固定し、学習データに用いた文数を徐々に増やして、正解音素系列の探索を行った場合の音素認識率を図 2-12 に示す。実際に各々の学習文に含まれる音素数は 267,474,754,1123,1577 と変化し、図 2-12 の横軸はこれを表している。本実験条件で

は、学習データに含まれる音素数が少なくても認識率は改善されており、学習音素数が増えるとともに改善率も高くなる傾向にある。表 2-21 の実験で扱った学習データは 754 音素を含んでいるが、さらにその約 2 倍(1577 音素)の学習データを用いても性能はまだ限界に達しておらず、さらに多くの学習データを用いることでより高い効果が期待される。図 2-12 に見られるように認識率の改善は学習データ量の増加に対して単調ではなく、学習音素数が 300 から 500 の間では、認識率の向上はみられない。このように、認識誤り音素系列に相当する正解音素系列の出現頻度が少ない場合には、探索時の N-best スコアリングとのトレードオフで局所的に認識率が向上しない場合も生ずる。

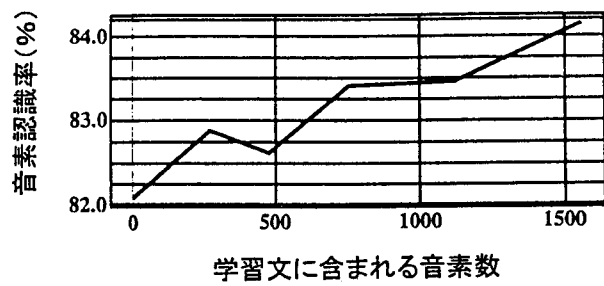


図 2-12 認識誤り特性の学習データ量と効果との関係

(Ⅲ) 従来の認識誤り抽出法との性能比較

2.2.2 節で述べたように、従来の方法では、音響的なマッチング区間を調べずに、文字系列上で正解系列と結果系列のマッチングを行い、なるべく音素と音素、音節と音節を一対一対応させながら、認識誤りを抽出していた。本提案の認識誤り特性抽出法における従来法との違いは以下の点である。

- 1. 認識結果における状態系列の各状態のマッチング区間が正しい区間と一致しているかどうかを考慮しながら、ラベル長の制限無しで認識誤り特性を抽出する。
- 2. 認識誤り特性がHMMの状態系列として表現されている。

これらの違いの優位性を上記項目別に確認する。

(Ⅲ－Ⅰ) 状態表現と音素表現との認識誤り特性記述力の比較

本モデルで、認識誤り特性をHMMの状態で表現したことの優位性を確認するために、従来のように音素で表現した場合との誤認識減少率を比較した。ここでは次の3つの表現を比較した。

(a) 認識誤り系列を状態系列として表現。(状態表現)

(b-1) 認識誤り系列を音素系列として表現。(音素表現)

(b-2) 認識誤り系列をその前後の1音素を含めた音素系列として表現。(拡張音素表現)

これらの表現の違いは、たとえば、/o ch i r u/ が /o i r u/ と認識された場合、次のようになる。

(a) /chi/の状態系列 → /i/の状態系列

(b-1) /chi/ → /i/

(b-2) /ochir/ → /oir/

我々が実験に用いている音素環境コンテキスト依存型HMM(Takami1992)では、先に述べたような状態系列の共有化が行なわれているため、同じ音素において、その前後の音素が異なる場合でも、音響的特徴が類似していれば同じ状態系列になる。従って、認識誤り系列における前後のコンテキストに対する拘束は、(b-1)の音素系列で表現した場合が最も緩く、次いで (a) (b-2) の順である。音素系列を探索する際には、(b-1)による音素表現の場合が、最も多くの音素系列が探索される可能性があるが、逆に誤った候補も増えるため、効率がわるくなる危険も予想される。

実験結果を表 2-22 に示す。表 2-22 にも見られるように、既知文では、全ての方法で誤認識が減少しており、(a)の状態表現が最も認識率が高く、次いで、拡張音素表現(b-2)、音素表現(b-1)の順であった。未知文の場合も(a)の状態表現が最も認識率が高く、次いで、拡張音素表現(b-2)、音素表現(b-1)の順であり、音素表現、拡張音素表現の場合は、正解音素系列探索前に比べて認識率は低下しており、探索により改善が見られたのは、本モデルが提案している状態表現(a)のみであった。

コンテキストを考慮したHMMの状態の共有化が認識誤り特性の共有化にも有効であり、音素表現を用いた場合の過剰な置き換えや、拡張音素表現を用いた場合の必要

な置き換えの不足を補う効果があることがわかる。

表 2-22 認識誤り系列の状態表現と音素表現との音素認識誤り率の比較 [%]

	探索前	状態表現(a)	音素表現(b-1)	拡張音素表現(b-2)
既知文	9.84%	7.22%	8.57%	7.61%
未知文	17.93%	16.59%	18.52%	18.20%

(Ⅲ－Ⅱ) シンボルDPによる抽出法との比較

本モデルの特徴の優位性を確認するために、音響的なマッチング区間を考慮せずに、なるべく音素と音素を一対一に対応させながら認識誤り特性を抽出する場合と比較した。ここでは、上記の抽出法の違いのみの比較を行うために、認識誤り系列の表現は状態ではなく、ともに音素で表現したものを用了。

音素と音素をなるべく一対一に対応させて認識誤り特性を抽出する際には、音素認識率を算出するために開発された、各音素の挿入、置換、脱落箇所を決定するツールを用了（たとえば文献（Young1996）などで用いているものと同方法）。この方法ではまず、正解音素ラベル系列と認識結果の音素ラベル系列とのDPを行う。正解音素系列の*i*番目までの音素系列 {Pc₁,Pc₂,...,Pc_i} と認識結果音素系列の*j*番目までの音素系列 {Pr₁,Pr₂,...,Pr_j} 間のDP距離D_{*i,j*}は式 2-14 のように算出される。

$$D_{i,j} = \min \left\{ \begin{array}{l} D_{i-1,j} + 1.0 \times f_{i,j} \\ D_{i-1,j-1} + 1.1 \times f_{i,j} \\ D_{i,j-1} + 1.0 \times f_{i,j} \end{array} \right\} \tag{2-14}$$

$$f_{i,j} = \begin{cases} 0, & \text{if } P_{ci} = P_{ri} \\ 1, & \text{if } P_{ci} \neq P_{ri} \end{cases} \tag{2-15}$$

上記DPを用了と、認識結果の各音素は系列に従ってなるべく同じ種類の音素とマッチングするように処理され、この場合距離値は加算されない。同種の音素がない場合には、異種の音素と対応し、置換誤りとして「1」が距離 D_{*i,j*} に加算される。正解と認識結果両者の音素系列に含まれる音素数が異なる場合にも、余まったり不足し

た音素は、挿入または脱落誤りとして距離値「1」が距離 D_{ij} に加算される。

認識誤りを抽出する際には、上記で置換誤りと処理された音素は、正解音素と認識誤り音素とが一对一に対応している認識誤り特性としてテーブルに登録される。挿入および脱落誤りの場合には、基本的には、挿入及び脱落している音素の前後の音素を認識誤り系列及び正解系列に含むように認識誤り部分を決定する。たとえば、正解音素系列が/aieuo/で認識結果の音素系列が/aieo/の場合、つまり/u/が脱落している場合には、正解系列/iue/、認識誤り系列/ie/のペアがテーブルに登録される。同じ正解系列に対し認識誤り系列が/aio/の場合は、正解系列/iueo/、認識誤り系列/io/となる。但し挿入及び脱落している前後の音素の一方が正解で、一方が置換誤りであった時には、例外的に正解音素に対応する方は系列に含めず、置換誤りしている音素のみを系列に含める。たとえば、正解系列/aiu/に対し認識誤り系列が/ao/であった場合には、正解系列/iu/、認識誤り系列/o/のペアが認識誤り特性に対応する。上記の処理を、以降、シンボルDPによる認識誤り特性抽出法と呼ぶ。

表 2-23 に、探索前後の音素認識誤り率を示す。"マッチング区間考慮"の結果は、(III-I) の実験結果の表 2-22 における音素表現(b-1)と同じである。

既知文、未知文ともに、マッチング区間が正しいかどうかを考慮しながらラベル長に制限を与えずに認識誤り特性を抽出した方が、探索後の音素認識率が高く、特に未知文で両方法の差が大きい。これは、シンボルDPがマッチング区間を考慮せずに音素と音素をなるべく一对一に対応させていることにより、実際の認識誤り部分に対応する正解系列がずれてしまうことが原因である。評価データが既知文の場合には、上記の学習データにおけるずれが同じ傾向で評価データ上でも起こるため、誤った探索を行いながらも、結局は正しい音素系列が探索できていた。しかし未知文では、上記のずれが原因で、正解している結果を誤って置き換えている場合があり、マッチング区間を考慮した場合との差はこれが原因であった。

表 2-23 音響的なマッチング区間の正しさを考慮した場合と
シンボルDBとの音素認識誤り率の比較 [%]

	探索前	マッチング区間考慮	シンボルDP
既知文	9.84%	8.57%	8.74%
未知文	17.93%	18.52%	20.71%

(Ⅲ－Ⅰ)における状態表現と音素表現との比較実験及び本節(Ⅲ－Ⅱ)におけるシンボルDBとの比較実験の結果をまとめると、次のようになる。

- ・ 本音素候補系列追加モデルにより、通常の認識システムでは得られなかった正解音素系列が探索可能となり、結果として、探索前に比べて誤りは減少する。
- ・ 本モデルがHMMの状態の共有化情報を用いて音素系列の追加を行うこと、認識結果の各状態のマッチング区間が正しい区間と一致しているかを考慮しながらラベル長に制限を与えずに認識誤りの抽出を行うことは、特に認識誤り特性学習用データにない文に対しての正解音素系列探索の効果をj得る上で、必要な条件である。

(Ⅳ) 音素候補系列追加モデルの課題

2.2.5 節の(Ⅲ)節に示した評価結果から、認識誤り特性をHMMの状態表現として抽出しておき、認識時にこの認識誤り特性と認識結果系列とを用いて新たな音素系列を追加するモデルが、ビーム幅を大きくしても従来の認識方法では結果に含むことができなかった正解音素系列を結果に含めることができることを確認した。

しかしこの探索方法は、次のような問題点と改良すべき点を抱えている。まず、上記方法では、認識終了後の認識結果から正解音素系列を探索しており、認識中に既に枝刈りされた結果を利用することができない。認識中に逐次的に探索を行うように改良することで、探索空間から落ちてしまう正解音素を探索空間に残すことが可能となり、探索性能をより向上させられると考えられる。また、認識誤り特性をより詳細に分析してみると、一部の認識誤り特性は品詞の種類にも依存していることがわかった。認識誤り特性の信頼性を上げ、より高性能な正解音素の探索に利用するためには、誤

り系列の属する単語の言語的性質の利用を検討する必要があると考えられる。

以上の二つのことを考慮した正解音素系列の探索を可能とするために、マルチ発音辞書を用いることを考えた。発音辞書は、各単語毎に単語の発音をHMM系列を用いて表現しているものであるが、通常は、辞書を様々な発声変動に対応させるために、一つの単語が複数の発音つまり複数のHMM系列またはHMMネットワークで表現されている。このマルチ発音辞書に、学習された認識誤り特性に従ったHMM系列を追加することで、追加する前には含まれなかった音素系列が、認識中に探索可能となる。また、各単語毎にHMM系列表現されているので、認識誤り特性の言語的性質の利用も容易である。そこで、認識誤り特性を利用してマルチ発音辞書を自動作成し、この辞書を用いて音声認識を行うことで、従来認識できなかった正解語彙を探索することを試みた。

従来、認識誤り特性を用いてマルチ発音辞書を作成する手法が提案されている(Wooters1994, Imai1995, Sloboda1995, Humphries1996, Neumeyer1996, Fukuda 1997, Ravishanker1997)。しかしこれらの手法は、音素系列で表現された認識誤り特性を用いたものであり、認識誤り特性の言語的性質も考慮されていない。そこでまず、状態系列を用いた認識誤り特性の抽出方法と抽出された認識誤り特性と言語的性質との関係について述べ、状態系列で表された認識誤り特性を用いて認識誤り特性の言語的性質を考慮して構築されたマルチ発音辞書の認識性能に及ぼす効果を報告する。

2.2.6 認識誤り特性と品詞との関係について

認識誤り特性が言語的性質に依存していることは、直観的に想像できる。たとえば、機能語と内容語という言語的性質が異なった2種類の単語において、調音が比較的明確な内容語と、調音が曖昧な機能語とでは、たとえ、周りの音素環境が同じであったとしても、その認識誤り特性は異なるであろうことは容易に想像できる。また、単語のアクセントやイントネーションなどは認識誤り特性に大きく影響すると思われるが、このアクセントやイントネーションもまた単語の言語的性質に強い関係があると思われる。そこで、認識誤り特性の言語的性質を用いることでより信頼性の高い認識誤り

特性を得るために、言語的性質を表すものとして品詞を取り上げ、品詞と認識誤り特性との関係を明らかにすることを試みた。

認識誤り特性の品詞依存性を明らかにするために、認識誤りを起こした単語の各品詞毎の出現分布を全単語（正認識単語＋認識誤り単語）の品詞毎の出現分布と比較した。もし両者の出現分布の傾向が違う場合は、認識誤りがある特定の品詞に偏って出現しているということになる。この傾向の違いをより定量的に明確にするために、両者の出現分布の χ^2 分析も合わせて行った。この χ^2 は式 2-16 で定義される (Mori1990)。

$$\chi^2 = \sum_{i=1}^N \frac{(M_i - E_i)(M_i - E_i)}{E_i} \quad (2-16)$$

N : 品詞数

M_i : 品詞 i における誤認識単語数

E_i は M_i の期待値であり、式 2-17 で表される。

$$E_i = W_i \frac{\sum_{i=1}^N M_i}{\sum_{i=1}^N W_i} \quad (2-17)$$

W_i : 品詞 i における全単語数

この χ^2 分析に用いた認識誤り特性は、表 2-24 に示した認識条件とデータベースを用いて抽出されたものである。品詞の種類は表 2-24 のデータベース内で既に定義されている一般的な品詞 13 種類を用いた (Morimoto1994)。図 2-13 に各品詞毎の単語出現分布と認識誤りを起こした単語の出現分布を示す。このデータベースでは、助詞の出現数が最も多く、次いで普通名詞、副詞と続く。しかし認識誤りを起こした単語の出現数は名詞で最も多く、単語の出現数が多かった助詞や副詞では名詞の約半分の出現数しかなかった。また、 χ^2 値は 144 であり、95%の信頼閾値よりも大きい。認識誤りはある特定の品詞に偏っており、品詞に依存して出現しているとみなすことが可能である。

表 2 - 24 認識誤り抽出条件及び使用データベース

ドメイン	旅行会話
音響モデル	話者適応後の音素HM-Net 延べ状態数401、混合分布数10
話者適応法	VFS法 近傍数6、平滑化係数20
音素認識手法	One-pass DPによるN-best候補探索
認識誤り特性表現	HMM状態系列
認識誤り特性抽出用データ	189文／話者（8183音素／話者）、
話者	男声3名
品詞	日本語13品詞 "名詞"、"固有名詞"、"代名詞"、"本動詞"、"副詞"、 "形容詞"、"助動詞"、"接続詞"、"助詞"、"感動詞"、 "接頭辞"、"接尾辞"、"連体詞"

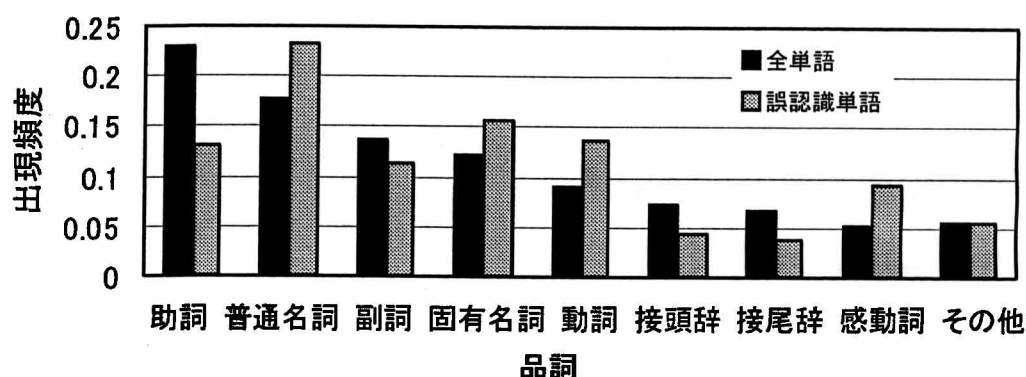


図 2-13 認識誤り出現頻度の品詞依存性 [%]

認識誤り特性の品詞依存性をさらに詳細に調べるために、各々の認識誤り特性毎に、その正解系列と認識誤り系列との品詞毎の出現分布を調べ、 χ^2 分析を行った。信頼できる χ^2 値を得るために、全話者の認識誤り特性データを合わせたデータを対象に分析を行うことでデータ数を確保し、また、2 種類以下の品詞にしか出現しない認識誤り特性を分析の対象から除外した。さらに信頼度の高い χ^2 分析を行うための、期待値 E_i に対する以下の 2 つ条件を満たす認識誤り特性のみを対象とした。

- ・期待値 E_i が 5 以上である品詞の割合が 80% 以上であること。
- ・全品詞における期待値 E_i が 1 以上であること。

本データにおいて、上記に示した様々な条件を満たす認識誤り特性は 26 個あった。

これら 26 個の認識誤り特性のうち、95%以上の信頼度で、特定の品詞に偏って出現したとみなされる認識誤り特性は 12 個 (45.3%) であった。表 2-25 に 95%の信頼度を閾値とみなした際に、特定の品詞に偏って出現していない認識誤り特性例 (前 2 例) と品詞に偏って出現している認識誤り特性例 (後 2 例) を示す。たとえば、2 番目や 3 番目の認識誤り特性の正解系列/m/は様々な品詞で用いられている状態系列であるが、認識誤り特性/m/→/n/は感動詞に偏って出現している。また、認識誤り特性/d/→/t/は、その正解系列/d/が多く含まれる接続詞では出現していない。一方、認識誤り特性/go/→/ko/や/m/→/g/は、正解系列の出現分布に似た分布で出現している。

もし、各品詞において、認識誤り特性が特定の単語に偏って出現していた場合は、認識誤り特性は“品詞”よりも“単語”との依存性が強い、という可能性がある。単語毎の出現分布をしらべることで、認識誤り特性が“品詞”との関係が強いのか、“単語”との関係が強いのかを明らかにしておく必要があると考えられる。そこで、先の実験で品詞との依存性が強いと判断された 12 の認識誤り特性を対象に、認識誤り出現数が最も多かった品詞について、さらに、認識誤りが単語に偏って出現していないかどうかを調べた。その結果、二つの認識誤り特性については認識誤りが一つの単語にのみ出現していたが、残りの 10 個の認識誤り特性については、認識誤りは複数の単語で出現していた。この結果は、ほとんどの認識誤り特性は単語よりも強く品詞に依存していることを示していると思われる。たとえば、表 2-25 にて品詞に依存しているとみなされた認識誤り特性/m/→/n/と/d/→/t/とについて、各々最も認識誤りの出現数が多かった「感動詞」と「助動詞」における単語毎の認識誤り出現分布を調べた。表 2-26 に示す。どちらの場合も、認識誤りはそれぞれ三つの単語にほぼ均等に出現していることがわかる。これら二つの認識誤り特性は単語に依存していないと思われる。

上記の分析を通して、今回使用した 13 品詞に依存している認識誤り特性があることがわかった。これは、品詞の違いを考慮して認識誤り特性を扱ったほうが、より精度の良い認識誤り特性を得ることができることを示唆している。しかし一方、約半分の認識誤り特性は、品詞との依存性がみられなかった。この品詞依存性の有無をまずは判断し、依存性があるとみなされた認識誤り特性のみに対し、品詞の種類を考慮して認識誤り特性を扱うことが必要である。

表2-25 品詞に偏って出現した誤認識特性及び品詞に偏らず出現した誤認識特性の例

	接頭辞	固有名詞	普通名詞	動詞
正解系列/go/の状態系列	53.6%	20.4%	20.4%	5.4%
上記系列から/ko/の状態系列に誤認識	64.0%	16.0%	12.0%	8.0%

$$\chi^2=1.87$$

	副詞	感動詞	動詞	接続詞	助詞	固有名詞	普通名詞
正解系列/m/の状態系列	44.1%	29.1%	10.6%	8.9%	3.0%	2.2%	2.1%
上記系列から/g/の状態系列に誤認識	48.0%	32.0%	12.0%	0.0%	4.0%	0.0%	4.0%

$$\chi^2=2.43$$

	副詞	感動詞	動詞	接続詞	助詞	固有名詞	普通名詞
正解系列/m/の状態系列	44.1%	29.1%	10.6%	8.9%	3.0%	2.2%	2.1%
上記系列から/n/の状態系列に誤認識	8.0%	60.0%	8.0%	24.0%	0.0%	0.0%	0.0%

$$\chi^2=26.35$$

	接続詞	副詞	助詞	動詞	固有名詞	普通名詞	接頭辞
正解系列/d/の状態系列	27.6%	24.1%	19.8%	7.8%	5.2%	4.3%	1.7%
上記系列から/t/の状態系列に誤認識	0.0%	32.4%	23.4%	17.6%	11.8%	11.8%	3.0%

$$\chi^2=19.8$$

表 2-26 認識誤りが多く出現した品詞における単語毎の認識誤り出現分布

認識誤り特性 (/正解系列/⇒/誤認識系列/)	品詞	認識誤りが出現した単語 (出現頻度)
/m/状態系列⇒/n/状態系列	感動詞	ありがとうございました(40.0%) かしこまりました(26.7%) すみません(33.3%)
/d/状態系列⇒/t/状態系列	副詞	どう (45.4%) どうも (27.3%) どうやって (18.2%) どうぞ (9.1%)

2.2.7 品詞の違いを考慮した状態系列の認識誤り特性抽出

2.2.3 節における認識誤り特性抽出法を用いて、表 2-23 に示したデータベースから認識誤り特性を抽出し、認識誤り特性毎にその出現頻度を算出する。出現頻度は、その正解系列の出現数に対する認識誤り系列の出現数の割合である。但し、一定数以下の品詞にしか出現しない認識誤り系列は、認識誤り系列が品詞に偏って出現しているとみなし、その品詞毎の出現頻度を算出する。各認識誤り特性と認識誤り系列が含まれる単語の品詞と上記で計算した認識誤り特性の出現頻度とを認識誤り特性表に記録する。認識誤り特性の信頼性を確保するために、出現頻度が一定数以下の認識誤り特性は除去した。

2.2.8 認識誤り特性を用いたマルチ発音辞書構築

状態系列で表現された認識誤り特性を利用して、認識を行いながら正解語彙の探索を可能にするために、抽出された認識誤り特性の正解系列と同じ状態系列が標準的発音に含まれる単語を探索し、正解系列を認識誤り系列に置き換えた発音を追加することで、マルチ発音辞書を作成した。認識誤り特性と品詞との関係を反映させるために、認識誤り特性が品詞の種類に依存している場合には、限られた品詞に絞って発音を追

加した。辞書サイズが膨大となってしまうことを避けるために、単語毎に追加される発音数に上限を定め、出現頻度が高い認識誤り特性を優先的に追加した。

2.2.9 マルチ発音辞書の認識性能に及ぼす効果

(I) 実験条件

認識誤り特性の品詞依存性を考慮して作成されたマルチ発音辞書が、連続音声認識の性能向上に効果があることを確認するため、2.2.7 節、2.2.8 節に示した手順で作成された認識誤り特性を用いてマルチ発音辞書を作成し、図 2-14 の連続音声認識システムを用いて単語 bi-gram による連続音声認識実験を行った。得られた単語認識率を、従来の 1 種類の発音のみを記述している発音辞書（単発音辞書という）を用いた場合と比べた。実験条件を表 2-27 に示す。

全話者に出現した認識誤り特性は全認識誤り特性の 10% しかなかった。この結果は、認識誤り特性は話者に強く依存していることを示している。そこで、本実験では、認識誤り特性の抽出を話者毎に行った。また、話者適応を行ってもなお残る認識誤りへの効果を調べるために、これまでの実験で用いていた話者適応法である VFS 法による適応後のモデルを用いて認識誤り特性の抽出を行った。さらに、信頼度の高い認識誤り特性のみを利用するため、出現数が 2 以下の認識誤り特性はマルチ発音辞書作成には用いなかった。辞書の過増を避けるため、一つの単語に対して辞書に追加する発音数の上限を 3 と設定し、認識誤り特性の出現頻度の高い方から優先的に追加した。品詞依存性については、異なる二つ以下の品詞のみにしか出現しない認識誤り特性を品詞に依存している認識誤り特性であるとみなした。本実験では、73% の認識誤り特性が品詞に依存しているとみなされた。ここでは、本認識誤り特性の扱い方の特徴である次の二つの事項について、その効果を評価した。

- ・ 認識誤り特性を HMM の状態系列で表現すること
- ・ 認識誤り特性の品詞依存性を考慮すること

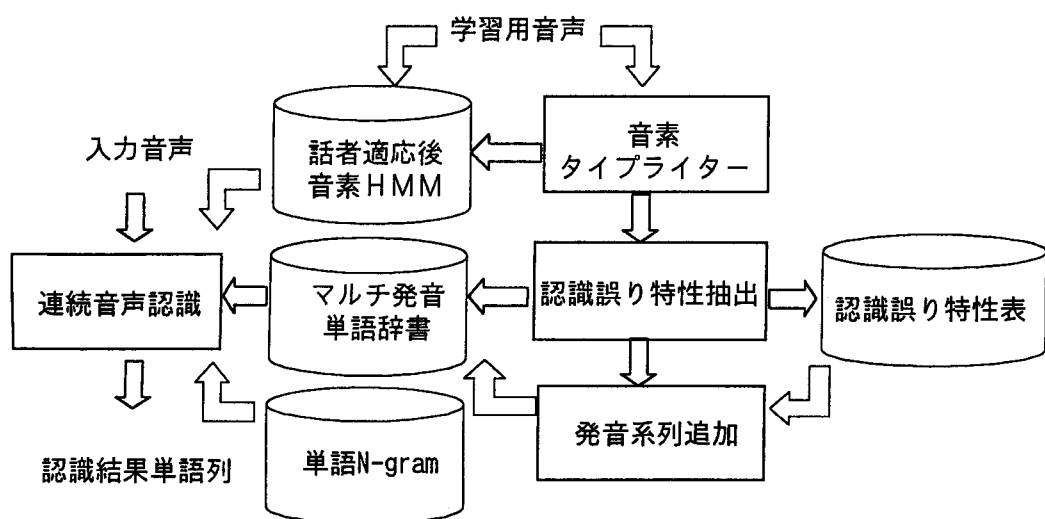


図 2-14 マルチ発音辞書学習システム構成

表 2-27 認識実験条件

ドメイン	旅行会話
音響モデル	話者適応後の音素HM-Net 延べ状態数401、混合分布数10
言語モデル	単語bi-gram、単語perplexity13.6
言語モデル学習用データ	3476文、44147延べ単語
連続認識手法	One-pass DPIによるN-best候補探索、ビーム幅1000
認識対象語彙数	2343単語
認識誤り特性抽出用データ	109文／話者（4854音素／話者）、
評価用データ	認識誤り特性抽出用データと同じ
話者	男声3名

(II) 認識誤り特性を用いたマルチ発音辞書の認識性能に及ぼす効果

(II - I) 認識誤り特性を HMM の状態系列で表現したことの効果

マルチ発音辞書を作成する際に、状態系列表現による認識誤り特性を用いる提案法の条件（状態表現・品詞考慮なし）と、音素系列表現による認識誤り特性を用いる方法（音素表現・品詞考慮なし）との単語認識誤り率と辞書サイズ比を比較した。辞書

サイズ比は単発音辞書のサイズを1とした場合のサイズ比で示している。結果を表 2-28、表 2-29 に示す。次のことがわかった。

- ・ どの話者についても、音素系列に比べて状態系列を用いた方が認識誤り率は減少している(13.13%→11.78% , $P \leq 0.05$)。認識誤り減少率は 10.28%であった。音素系列を用いた場合は、多くの話者で単発音辞書を用いた場合より認識誤り率が増加している。この結果は、認識誤り特性はその前後の音素環境に依存しており、この依存関係を考慮せずにマルチ発音辞書を作成した場合には、不必要な発音が追加されることで認識率が低下する場合があることを示唆している。
- ・ 上記結果に関連して、音素系列で認識誤り特性を表現する時に比べて、状態系列で認識誤り特性を表現することで、辞書サイズを小さく抑えることが可能である。辞書サイズは単発音辞書に比べて、音素系列の場合は 1.85 倍、状態系列の場合は 1.59 倍であった。

これらの結果から、前後の音素環境を反映していると考えられる状態系列を用いることにより、不必要な発音の追加が制限され、辞書サイズをコンパクトにすることができる上に、安定して認識率を向上させることが可能であることがわかった。

(II - II) 認識誤り特性の品詞依存性を考慮することによる単語認識誤り率の改善

認識誤り特性と品詞の関係を利用することの有効性を確認するために、3.2.8 節の手順通り、品詞を考慮して発音を追加した場合（状態表現・品詞考慮）と、考慮しない場合（状態表現・品詞考慮なし）との単語認識誤り率を比較した。この認識実験結果も、表 2-28、表 2-29 に示す。この結果より次のことがわかった。

- ・ 品詞情報を用いた方がより認識誤り率が減少することがわかる(11.78%→10.36% , $P \leq 0.05$)。品詞を考慮しない場合に比べて認識誤り減少率は 11.9%であった。
- ・ 品詞依存性を考慮しない場合に比べて、品詞依存性を考慮した場合は、辞書サイズを小さく抑えることが可能である。辞書サイズは単発音辞書に比べて、品詞依存性を考慮しない場合は 1.59 倍、考慮した場合は 1.36 倍であった。

これらの結果から、認識誤り特性の品詞依存性を考慮することにより、不必要な発音の追加が制限され、辞書サイズをコンパクトにすることができる上に、誤認識率を

減少させることが可能であることがわかった。

表 2-28 認識誤り特性の状態表現及び品詞依存性の考慮が
認識誤り率の減少に及ぼす効果 [認識誤り率%]

	認識誤り特性	話者1	話者2	話者3	平均
単発音辞書	(95%信頼区間)	10.38% ±1.8	5.79% ±1.73	22.21% ±3.08	13.13% ±1.27
マルチ発音辞書	音素表現・品詞考慮なし	11.40%	5.33%	23.40%	13.82%
	状態表現・品詞考慮なし (95%信頼区間)	9.18% ±1.7	5.20% ±1.64	19.96% ±2.61	11.78% ±1.22
	状態表現・品詞考慮	8.85%	3.81%	<u>17.31%</u>	<u>10.36%</u>

- ・太字の認識率は、 χ^2 判定により単発音辞書による認識率から 95%の信頼区間で有意差があるとみなされた認識率である。(単発音辞書に添付している 95%信頼区間を参照)
- ・下線の認識率は、 χ^2 判定により“状態表現・品詞考慮なし”における認識率から 95%の信頼区間で有意差があるとみなされた認識率である。 (“状態表現・品詞考慮なし”に添付している 95%信頼区間を参照)

表 2 - 29 認識誤り特性の状態表現と品詞依存性考慮とが
辞書サイズ削減に及ぼす効果

	認識誤り特性	辞書サイズ比
単発音辞書		1
マルチ発音辞書	音素表現・品詞考慮なし	1.83
	状態表現・品詞考慮なし	1.59
	状態表現・品詞考慮	1.36

2.2.10 まとめ

学習時に、音響モデルの不備に起因する認識誤り特性をHMMの状態表現として抽

出し、抽出された認識誤り特性を用いて新たな音素候補系列を追加する方法を提案し評価した。ここでは二つの追加方法を提案した。

一つの方法は、認識時に、この認識誤り特性と認識結果系列とを用いて認識結果には含まれない新たな音素系列を追加するモデルである。音素タイプライター型の認識システムを用いて上記音素候補系列追加モデルを評価した結果、従来の認識において、ビーム幅を大きくしても結果に含まれなかった正解音素系列を、本音素候補系列追加モデルを用いることで、結果に含めることができることを確認した。さらに、本モデルを認識誤り特性を用いた従来の認識誤り抽出法とを比較した結果、本モデルにおいて、誤認識系列の各状態のマッチング区間が正しい区間と一致しているかどうかを考慮しながら誤認識に含まれる音素数に制限を与えずに認識誤り特性を抽出していることや、認識誤り特性をHMMの状態系列で表現していることが、正解音素系列の探索性能を向上させるのに有効であることがわかった。

他の一つの方法は、認識しながら探索空間から落ちる語彙を探索空間に残すことが可能な正解語彙の探索を可能とするために、後処理的な正解音素の探索ではなく、状態系列で表現された認識誤り特性を用いてマルチ発音単語辞書を作成し、これを用いて連続認識を行う方法である。認識実験により、状態系列で表現された認識誤り特性を用いることが、従来の音素系列で表現された認識誤り特性を用いるより、高い認識率を示すことを確認した。また、認識誤り特性と品詞との関係を考慮してマルチ発音辞書を作成することで、さらに認識性能が向上した。

本論では、話者により認識誤り特性が異なることを考慮して評価実験を行ったが、今後は話者だけでなく他の様々な要因も含めて、条件の違いに依存しない認識誤り特性と条件に特有の認識誤り特性を区別し、条件に依存しない認識誤り特性は、HMMモデル作成時に予め抽出しておくなど、学習データの効率の良い使用法を工夫する必要があると考えられる。

第3章 言語翻訳のための認識誤りの校正

3.1 単語間の意味的類似性を用いた正解認識部分特定

3.1.1 はじめに

連続音声認識において、単語 N-gramと呼ばれる統計的手法に基づいた言語モデルが広く使用されており (Bahl1983)、限られた探索空間上で認識精度を向上させるためには、信頼できる単語連鎖統計値を得るための大量のデータを用いて大きなN値による単語 N-gram を用いるのが効果的である。しかし、大量のデータを用意することは非現実的であり、実際にはNを2または3とした bi-gram や tri-gram などを用いて、局所的な単語連鎖にのみ制約を与えて使用している。従って、単語 N-gram モデルを用いた音声認識誤りには、N単語以上、実際には3、4単語以上からなる長い単語連鎖部分から判断すると不自然なものが多い。たとえば、単語 bi-gram を用いた認識結果として「東京をしているのですか?」という文が出力された。単語 bi-gram で制御されている単語の2連鎖間の関係には不自然さはなく、文頭に「東京」と話すことも、「東京」+「を」の連鎖で話すことも不自然ではない。しかし、「東京」+「を」+「する」の3連鎖でみると、その内容は少し不自然である。このような認識誤りの特徴を考慮して、認識誤り文の不自然性を指摘する、あるいは認識誤り文の解析ができる手段が必要であると考えられる。

通常、文解析を行う手段としては、文脈自由文法のように構文構造の解析を目的とした文法が頻繁に用いられている。文脈自由文法では、構文的に誤った文は解析が不可能になることから、上記のような長い単語連鎖部分において文法的に不自然となるような音声認識誤り文について、解析不可能な文として指摘することが可能となる。しかしこの方法は、音声認識誤りさえなければ文全体の記述が可能であることを前提としている方法であり、文法記述が難しい実際の話し言葉に関しては、上記の前提が成り立たない。つまり、解析できなかった場合に、入力文自身が文法を逸脱した文なのか、音声認識誤りを起こしている文なのか、を判断ができないことが問題となる。この問題を回避するために、文脈自由文法に則って、非文法的な文を解析する手法も

提案され (Saitou1988、 Mellish1989), それを音声翻訳システムに導入した例も紹介されている(Lavie1996)。これらは, 入力文中に解析できない部分があったとき, その部分を削除、あるいは他の単語を挿入, 置換しながら解析を続行する方法である。しかし、実際の話し言葉における文法記述の困難の度合いが、単語の削除や挿入の処理のみでは解決できる範囲を超えており、この方法によっても、まだ多くの話し言葉及びその音声認識誤り文が解析できないのが現状である。

一方、上記で問題となった、文全体を文法記述することが困難な話し言葉の解析を可能にするために、発声の際のポーズで区切られた部分を一単位として、文全体を単位毎に分け、各々の単位部分を部分木で記述し、この部分木を列挙したもので文全体を記述する方法も提案されている(竹沢 1996)。この方法は話し言葉の解析を可能とする上で効果的な方法である。しかし、これらの部分木は局所的な一部分文の範囲で解析を行うものであり、これら部分木が N-gram の N 以下の単語で構成されている場合は、N 個以上の長い単語連鎖からなる大局的な部分の自然性を判断できる方法ではない。従って、問題となる音声認識誤り文を誤りであると判断するには不十分であり、これらの認識誤り文を誤ったまま解析してしまうという問題がある。

さらに、これら従来の解析方法は文全体にせよ部分文にせよ、文脈自由文法による文法的制約を基本としており、意味的な整合性を判断した解析ではない。そのため、文の「不自然さ」を判断するにはいずれにしても不十分であると考えられる。

話し言葉に対して、統計的言語モデルを用いて単語探索を行った場合の音声認識誤り文を解析するためには、文法以外の制約を積極的に用いて、音声認識誤り文から正しく認識された部分を特定するしくみを新たに導入し、特定された部分のみから文を解析することが必要である。そこで本節では、単語 N-gram の N の範囲を超えた大局的な部分を対象に、その意味的な自然性を判断することにより、認識結果部分から正しく認識された部分のみを特定する方法を提案する。

3.1.2 自由発話文解析法を用いた正解部分特定

(I) 単語間依存構造を基本とした自由発話文解析

話し言葉の翻訳をめざして、変換主導型機械翻訳 (Transfer Driven Machine Translation 以下 TDMT と記述する) と呼ぶ音声翻訳手法が提案されている(古瀬 1994)。TDMT では、話し言葉に頻繁に出てくる文法記述の困難な言い回しを取り扱うために、実際の話し言葉に現われる種々の単語間の言い回しを表現パターンとして記述し、これらの表現パターンとその対訳とを蓄積しておく。実際の翻訳の際には、入力された文に類似した表現パターンを選択し、その対訳パターンを用いて翻訳を行う。この表現パターンには次の特徴がある。

- ・ 表現パターンは表現を決定するマーカーと変数から構成されており、マーカーは機能語、変数は内容語であることが多い。たとえば「京都のホテル」という学習文における表現パターンは「XのY」となり、この際のマーカーは「の」、変数X及びYは「京都」と「ホテル」となる。このように多くの表現パターンは、変数に相当する内容語間の依存関係を、マーカーに相当する機能語で表していることになる。
- ・ 表現パターンの単位は、話し言葉の意味を理解する上での最小単位であることを基本とするため、パターンの長さは単語から句、文まで様々である。この単位を constituent boundary と呼ぶ。

一つの表現パターンの依存構造が複数通り考えられることがある。依存構造が異なると翻訳結果が異なるため、適切な依存構造を決定することは重要なことである。そこで最も適切な文構造を持つ表現パターンを決定するために、学習された表現パターンの変数に相当する単語と入力文の単語間との意味的な類似性を調べ、入力文に含まれる単語と最も類似した単語を含んでいる文構造を選択する(隅田 1994)。たとえば、「京都ホテルの一万円の部屋」という入力文については、(京都ホテルの(一万円の部屋))と((京都ホテルの一万円)の部屋)の二通りの構造が考えられる。どちらかふさわしい構造を決定するために、既に学習された表現パターン(先に述べた変数とマーカーで記述されたもの)を参照する。「XのY」という学習表現パターンには様々な目的言語表現があり、たとえば、「Y of X」と訳されるものには、それぞれXとYに((京都、ツアー)、(円、部屋),...)などの単語が抽出されており、「X,Y」:と訳されるものは、(ホテル、部屋),...) などがある。これら学習パターンにおける変数に相当する単語と入力文の変数に相当する単語との意味的な類似性を調べると、入力文の(一万円、部屋)と学習パターンの(円、部屋)、及び、入力文の(京都ホテル、部屋)と

学習パターンの（ホテル、部屋）は非常に類似している。従って上記二通りの構造の中では、（京都ホテルの（一万円の部屋））の構造の方が学習パターンと類似しており最適表現であると決定される。この際の意味的類似性は意味シソーラス辞書に従って算出される。この最適表現パターンの探索は、left-to-right のボトムアップ探索処理にて行なわれる。従って、もし文全体に相当する表現パターンが探索できなくても、部分文に相当する表現パターンは探索可能である(Furuse 1996)。

（II）正解部分特定

上記の自由発話文解析手法は、その結果として、入力文を構成している部分的な表現パターンを出力する。しかも、各表現パターン毎に学習パターンとの類似性が明らかにされている。これらの結果は、以下の理由から、音声認識誤りを判断できる可能性を持っていると考えられる。

- ・ 文脈自由文法に基づいた文法規則では、話し言葉も音声認識誤りも同様に解析が困難であるため、両者の区別がつかない場合が多かったが、上記解析法では実際の話し言葉の言い回しを学習することにより、解析できない部分を音声認識誤りと判断できる可能性がある。
- ・ 上記解析法で扱う表現パターンは、N-gram のN個以上の単語からなるものも多くある。この表現パターン単位で文の自然性を判断することで、N-gram よりさらに大局的な部分を評価できる枠組である。

これらの解析結果を用いて音声認識誤りを判断する際には、音声認識の際に考慮されていない、長い単語連鎖からなる大局的な部分からの意味的な不自然さを判断する必要がある。そこで、「入力文と学習文との単語間の意味的類似性」と「1つの表現パターンに含まれる単語数」に注目して各表現パターンを評価することにより正解部分を特定することを試みた。正解部分の特定条件は次の二つである。

- （1） ある入力文中の表現パターンが音声認識誤りを含んでいる場合には、学習パターン内の単語との意味距離値は大きくなると予想される。意味距離値が一定値より小さい表現パターンを正解部分とする。
- （2） 単語 N-gram により制御された認識結果は、N 個以下の単語からなる部分については既に制御されており、結果が正解であるか誤りであるかに無関係に「自

然」な系列である場合が多い。単語数が従って、 $N+1$ 個以上からなる範囲から判断して「自然」である場合のみ、正解部分とする。

(Ⅲ) 正解部分特定手順

(Ⅱ) 節で示した条件に基づき、次の手順で正解部分の特定を行う。

- (1) 長い語句の範囲の表現パターンから順にその意味的距離を判断し、意味的距離が閾値より小さい場合は、その範囲に含まれる全ての部分を正解部分とする。
- (2) もし意味的距離が閾値より大きい場合には、その部分のどこかに誤りが含まれているとみなして、より小さい部分の表現パターンについて(手順 1)の処理を繰り返す。
- (3) この処理が繰り返され、非常に短い語句の範囲での局所的な語句からなる表現パターンを扱うに至る場合は、対象となる短い部分は、他の部分と依存関係がなく文全体から見て不自然な部分であると判断する。

そこで、解析範囲に含まれている単語数に下限閾値を設け、解析範囲が細分化されてその形態素数が閾値に達した場合には、意味的距離がたとえ小さくても、その部分を誤り部分と判断する。

図 3-1 に正解部分特定例を示す。これは単語 bi-gram を用いた音声認識結果を解析した例である。3.1.2 節の (Ⅰ) 節にて述べた単語間依存関係を基本とした自由発話文解析法では、まず音声認識結果に対しボトムアップな解析を行うことで、いくつかの部分文に対し表現パターンが適応され、結果として、図 3-1 に示したような構造と意味的距離値を得る。この解析結果から正解部分の特定を行う。単語数の下限閾値を bi-gram の N より 1 つ大きい 3、意味的距離の上限閾値を 0.2 とすると、この結果は次のように処理される。この結果では、「料金」が「料決ま」と認識誤りを起こしているため、文全体としての解析は失敗している。そこで、解析できた部分で最も大きい表現パターンである「エエ それぞれおいくらなんですか」の意味的距離 0.4 を閾値と比較する。この場合は閾値を上回っているので、この範囲のどこかに誤りがあるとみなし、一段階小さな表現パターン「それぞれおいくらなんですか」を処理する。この部分の意味的距離 0.013 は閾値より小さいので、この部分は全て正解部分と特定する。

まだ、判断していない残りの「料決ま」の部分の意味的距離は閾値より小さいものの、この範囲に含まれる形態素数2が、形態素数閾値の3未満であるので、この部分は誤った部分と判断する。

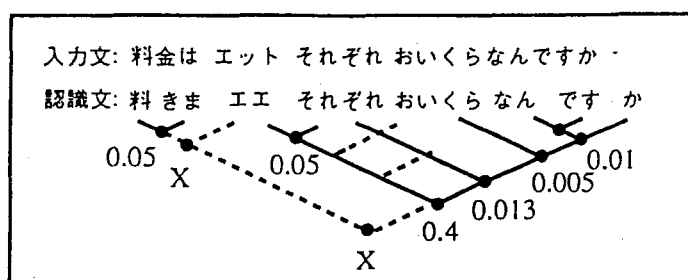


図 3-1 正解部分特定例

(Ⅳ) 正解部分特定法の文理解度に及ぼす効果

音声認識結果に対して、3.1.2 節の (Ⅲ) 節で述べた手順で正解部分を特定し、特定された部分のみから、どれだけ入力文の意図が理解できるかを評価した。音声認識結果はまず原言語解析部に入力され、解析部は文全体、または部分文の構造と意味距離値を出力する。これらを用いて正解部分を特定し、特定された部分文のみを出力する。出力された部分文を以下の観点から評価した。

(1) 正解部分特定率

(2) 特定された部分のみを提示した場合の全体文に対する文理解率

音声認識手法として、音素HMMと単語 bi-gram 言語モデルを使ったマルチパス探索でワードグラフを出力する連続音声認識方式(Shimizu1996)を用いた。データベースには旅行会話データベース(Morimoto1994)の中の9会話分(9話者119文)を用いた。本方法は正解部分をトップダウンに特定するため、正しい文が入力され文全体が従来の解析にて成功した場合には従来と同様文全体を出力し、文全体の解析が成功しなかった場合のみ正解部分を特定する。特定された部分のみを出力することの効果进行评估するため、ここでは、この119文を用いて音声認識実験を行い、認識誤りを起こした70文のみを評価に使用した。

(IV－I)正解部分特定率

本方法の正解部分の特定性能を確認するために、正解であると特定した部分に含まれる正解単語の再現率及び適合率を調べ、正解部分の特定を行う前の音声認識結果文の再現率及び適合率と比較した。

$$\text{再現率} = \frac{\text{結果文に含まれる正解単語数}}{\text{正解文の総単語数}} \quad (3-1)$$

$$\text{適合率} = \frac{\text{結果文に含まれる正解単語数}}{\text{結果文の総単語数}} \quad (3-2)$$

正解部分の特定に使用した意味的距離の閾値と一つの表現パターンに含まれる形態素数の下限閾値とを様々に変えた場合の性能の違いも合わせて評価した。図 3-2、図 3-3 に適合率と再現率を示す。次のことがわかった。

- ・どのような閾値の条件下でも、正解単語適合率は 92%を上回っており、特定する前の認識結果と比較しても約 15%前後向上している。特定した部分が正解である信頼性は高い。
- ・正解単語再現率は 20%前後低下しており、本方法が特定しきれない正解部分も多い。
- ・形態素数の閾値については、閾値が大きくなるほど適合率は向上する。特に閾値が 3 以上の場合は、閾値が 2 の場合と比べて適合率は大きく向上している。本実験では音声認識時に単語 bi-gram を用いており、たとえ誤った単語であっても 2 単語連鎖間だけを取り上げると自然なものが多い。閾値が 2 の場合は誤った 2 単語連鎖を正解部分と特定してしまったことが、3 以上の場合に比べて適合率が低かった原因である。
- ・形態素数の閾値が 4 以上になると再現率が極端に低下する。
- ・意味的距離の閾値については、その値が小さくなるほど再現率は高く適合率が低くなる傾向がある。しかし、閾値の違いによる性能の違いは、形態素数の閾値を変化させた時に比べて少ない。

- ・ 意味的距離の閾値が 0.2 以下になると適合率は徐々に低下する。

本方法は、学習された表現パターンに基づく解析結果から正解部分を特定するため、たとえ正しく認識された単語であっても、表現パターンと適合しなければ正解部分と特定されず再現率は低くなっている。しかし、本方法による適合率は高く、これは特定された部分が実際に正解単語である信頼性は高いことを示している。従って、単語 N-gram により制約された認識結果から正解部分を特定するために、学習パターン及び入力パターンに出現する単語間の意味的類似性を用いる本方法は、高い信頼度で正解部分を特定するのに有効であると考えられる。

また、正解部分特定性能は単語数閾値の変動に対しては敏感であり、たとえば、音声認識時の言語モデルの制約範囲より小さい値に閾値を設定すると適合率は低下する。言語モデルの制約範囲を越えた単語数閾値を用いることが必要である。

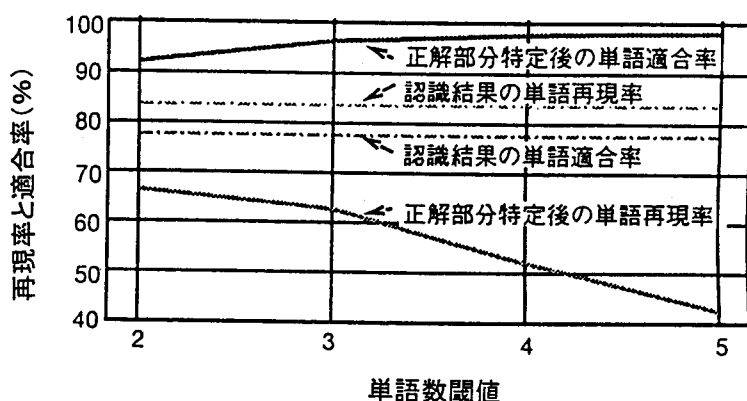


図 3-2 単語数の閾値と正解部分特定率との関係

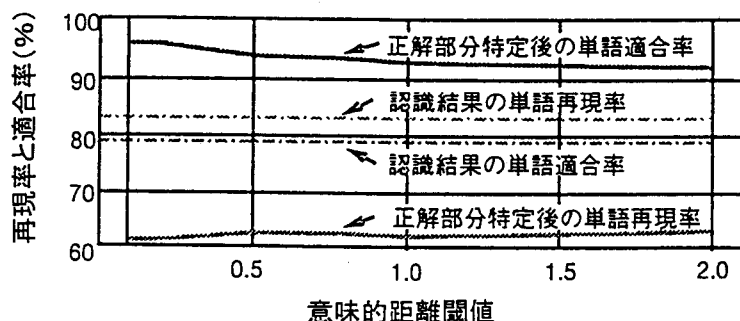


図 3-3 意味的距離の閾値と正解部分特定率との関係

(Ⅳ－Ⅱ) 特定部分からの文理解度

本正解部分特定法が、文の意味を理解する上で有効かどうかを確認する。ここでは、音声認識誤り部分を含んだ文全体を提示した場合（特定前）と、本正解部分特定法により特定された部分のみを提示した場合（特定後）との、各々の文理解度を評価した。形態素数および意味距離値の閾値は、前章の実験での最適値(3 と 0.2)を設定した。評価方法は、特定前及び特定後の各々の結果文とそれに相当する正解文を比較して、次の4段階の評価レベルを結果文に与えるものである。評価は日本人5名で行った。4段階の評価レベル内容を以下に示す。

- 4：正解と比べて同じ意味であると理解できる。
- 3：少し不自然だが、意味は理解できる。
- 2：全体の意味はわからないが、部分的には理解できる。
- 1：間違えた意味にとってしまう。

各レベルにおける5名の回答頻度を表 3-1 に示す。表の「正解部分なし」は正解部分を特定した結果、文全体が誤り部分であり正解部分がないと判断された場合であり、本評価の対象外となる文の頻度を示している。上記次のことがわかった。

- ・本方法にて特定された部分のみ提示することで、誤った意味に解釈される文が半分以上減少した。(評価レベル1：35.5%⇒15.2%)
- ・本方法にて特定された部分のみ提示することで、正しく意味が伝えられた文は増加したが、その割合は僅かであった。(評価レベル4と3の和：41.6%⇒42.9%)

本正解部分特定法により特定された部分のみを提示することは、認識誤り文がそのまま誤った意味に解釈されてしまう問題を軽減する効果があった。これは、特定した部分が確かに正しいという本方法の適合率の高さを裏付けている。しかし、より正しい意味を伝えるための効果は僅かであった。

表 3-1 正解部分特定前後の文理解率の違い [%]

レベル	4(完全理解)	3(理解可能)	2(部分理解可能)	1(理解不可能)	正解部分なし
特定前	19.6%	22.0%	23.0%	35.5%	0.0%
特定後	20.3%	22.6%	36.8%	15.2%	5.4%

3.1.3 正解部分のみ翻訳する部分翻訳

本正解部分特定法が、翻訳結果に及ぼす影響を確認する。3.1.2節で述べたように、解析された表現パターンと変数に相当する単語間の意味的類似性とを用いて、認識結果の中の正解部分を特定し、特定された部分のみの部分翻訳が可能な音声翻訳システムを構築した。図3-4にシステムの概要を示す。

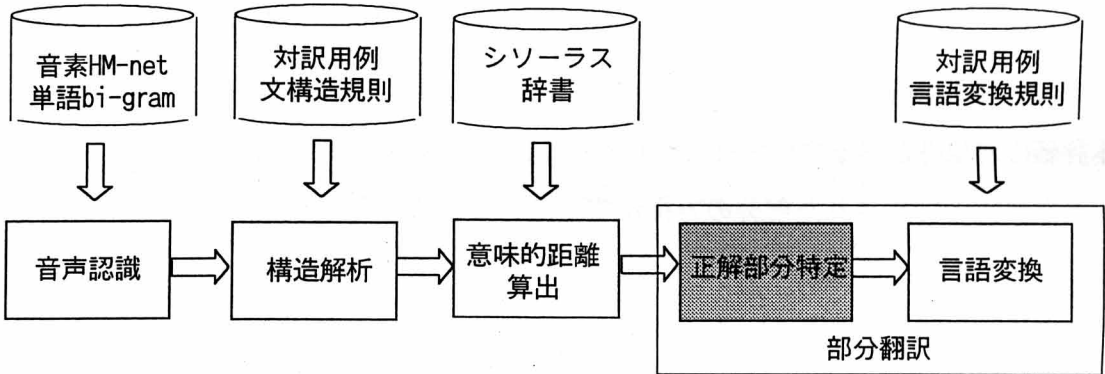


図3-4 部分翻訳を用いた音声翻訳システム構成

3.1.4 部分翻訳の翻訳性能に及ぼす効果

ここでは、音声認識誤り部分を含んだ文全体を翻訳した場合（特定前）と、本正解部分特定法により特定された部分のみの翻訳した場合（特定後）との、各々の翻訳文理解度を評価した。閾値の条件は前章と同様に設定した。評価方法は、特定前及び特定後の各々の原言語文に対する翻訳結果文とそれに相当する正解原言語文に対する翻

訳結果文とを比較して、次の4段階の評価レベルを訳結果文に与えるものである。

- 4：正解と比べて同じ意味であると理解できる。
- 3：少し不自然だが、意味は理解できる。
- 2：全体の意味はわからないが、部分的には理解できる。
- 1：間違えた意味にとってしまう。

評価は英会話能力の高い日本人3名で行なった。各評価レベルにおける回答数の3名の平均値を表3-2に示す。評価レベル4から1までは、前章の文理解率を評価する評価基準と同じである。特定された部分がない、または訳過程で処理が失敗するという理由で、訳結果が出なかったときに相当する場合を「訳結果なし」とした。次の結果が得られた。

- ・従来11.9%であった訳正解率が、本特定法の導入後は約2倍の25.7%に向上している。
- ・本正解部分特定法の導入前は、85.2%の文が訳できなかったが、本特定法の導入後は、69%（評価レベル4と3と2の和：25.7%+16.7%+26.6%）の文に対し、正しいかもしくは部分的にも理解できる訳文を出力できた。
- ・本正解部分特定法の導入前は、誤った意味に訳されることはほとんどなかったが、本特定法の導入後は、誤訳文が21%に増加した。

表 3-2 正解部分特定前後の訳率の違い [%]

レベル	4(完全理解)	3(理解可能)	2(部分理解可能)	1(理解不可能)	訳結果なし
特定前	11.9%	0.0%	0.0%	2.4%	85.7%
特定後	25.7%	16.7%	26.6%	21.0%	10.0%

以上の結果は、本正解部分特定法が訳結果に及ぼす効果は大きく、従来ほとんど訳できなかった誤認識文の69%に対し、部分的にも意味を理解できる訳文を出力可能であることを示している。また、文全体を訳できなくても、内容の理解に必要な語が認識されていれば、その部分のみを訳することで、ほぼ正しく理解できる訳結果が出力できていた。この結果は、特定されなかった部分の修復を検討する際に

は、特定されない部分を全て修復する必要がなく、必要な単語のみの修復で十分であり、場合によっては、修復の必要がないものも多いことを示唆している。しかしまた、翻訳誤り文も増加している。これは、特定された部分はほとんどが実際に正解部分であったが、部分翻訳結果は、特定された原言語部分の順序で相当する部分翻訳結果を並べて提示しているため、各部分文は適切に翻訳されていたとしても、部分文系列から判断すると誤った意味に翻訳される場合があることが原因であると考えられる。このような誤訳に関しては、今後解決していく必要がある。次節にて、上記の翻訳結果の分析を通して、本正解部分特定法の音声翻訳に及ぼす効果と悪影響について考察する。

3.1.5 考察

従来の翻訳でも、認識誤り文の 11.9% に対しては正しく翻訳することができている。これは使用した翻訳システム TDMT が、話し言葉に対応するために助詞や一部の文末表現が欠落した言い回しを学習しており、認識誤りがこれらの脱落誤りであった場合にも正しく解析できたためである。しかし、実際の話し言葉には存在しないような言い回しに認識誤りした場合には、翻訳することができなかった。本正解部分特定法の導入により、効果が見られた認識誤り文の特徴は、主に以下のものである。

- (a) 表現パターン間の挿入誤り
- (b) 文末の言い回しの誤り
- (c) N 個以上の単語からなる不自然な表現への誤り
- (d) 学習データに稀な表現への誤り

また、認識誤りではないが、従来翻訳できなかった文が、本方法にて翻訳可能になったものに、次のものがある。

- (e) 複数文からなる発話

また逆に、本正解部分特定法を導入することで誤訳してしまった文の主な特徴は、次のものである。

- (f) 文末の、人称、立場、肯定文か否定文か、などを決定する表現の誤り

(a)表現パターン間の挿入誤りについて

自由発話文では、「ええ」や「あ」「ん」などの冗長語が多く話されるため、音声認識では、これらの単語を認識単語として登録している場合が多い。ところが、日本語の冗長語は、短音素で構成されているものが多く、認識時には、他の単語の一部が、度々、冗長語と誤って認識されてしまう傾向がある。さらに、冗長語は文全体のどこで話されるかが予測しにくく、単語 N-gram の枠組では制御しにくい性質がある。これらの理由で、音声認識結果では、この冗長語の挿入誤りが頻繁に起こる。表 3-3 は冗長語の挿入誤りのために、従来の翻訳処理では翻訳できなかった認識誤り文の例である。正解部分を特定して翻訳することで、「電話番号は」や「五二七」などの翻訳結果を出力することができている。ただし、「九です」は、冗長語「あっ」の挿入のために、他の部分とは独立している 2 形態素の局所的な部分であるとみなされ、正解部分と特定することができなかった。

表 3-3 挿入誤りに対する部分翻訳結果例

入力文	電話番号は五二七九です。
認識結果	電話番号は <u>っお</u> 五二七 <u>あっ</u> 九です。
正解部分特定結果	電話番号は ... 五二七
部分翻訳結果	The telephone number ... five two seven

(b) 文末表現の誤り

間接的な表現や丁寧な表現が文末で話されることが頻繁にあり、文末の発声が不明瞭な場合は、この部分で認識誤りが起こることも多い。しかし、これらの表現が聞き取れなくても問題なく文の内容が理解できることも多く、そのような場合には、文末に至るまでの部分のみを翻訳することで、ほぼ意味が理解できる翻訳文を出力することができる。表 3-4 は文末での丁寧な表現の部分で音声認識誤りを起こした例である。前半部のみを特定し翻訳することで、意味の理解できる翻訳結果を出力することができている。ただしこの例の場合は、文末表現が人称を決定しているため、主語の「I」を翻訳することはできなかった。

表 3-4 文末での置換誤りに対する部分翻訳結果例

入力文	部屋の予約をお願いしたいんですけれども。
認識結果	部屋の予約をお願いしたい <u>ねすとも</u>
正解部分特定結果	部屋の予約をお願いしたい
部分翻訳結果	 would like to reserve the room.

(c) N個以上の単語からなる不自然な表現への誤り

単語 N-gram における N 個の単語を越えた大局的な範囲で判断した時に、不自然であると思われる認識誤りは、認識時には発見されない。本方法で正解部分を特定することにより、このような誤りを発見できる利点がある。表 3-5 は、入力文の「お部屋」を「親子」と誤って認識した例である。しかしこの例では、単語 bi-gram で制御されている単語の 2 連鎖間の関係には不自然さはなく、文頭に「親子」と話すことも、「親子」+「の」の連鎖で話すことも不自然ではない。しかし、「親子」+「の」+「ご希望」の 3 連鎖でみると、ホテル予約タスクとしてはその内容は少し不自然である。本正解部分特定法を用いると、「親子のご希望」の意味的距離値が閾値を越えることから、「親子の」が除去され「ご希望はございますか」のみ翻訳することができる。

表 3-5 大局的にみて不自然な認識誤りに対する部分翻訳結果例

入力文	お部屋のご希望はございますか？
認識結果	<u>親子</u> のご希望はございますか？
正解部分特定結果	ご希望はございますか？
部分翻訳結果	 do you have any preference ?

(d) 学習データに稀な表現への誤り

本翻訳システムは用例主導型のアプローチを採用しており、一般に構造的または意

味的に曖昧性を含む文が入力されても、既に学習された表現パターン採用することで、その曖昧性を解消することができる。そのために、認識誤りの結果、学習データには存在しないが一般には不自然ではない文に誤ってしまった場合でも、誤ったまま翻訳は行なわず、正解部分を特定して部分的にのみ翻訳することができる。表 3-7 は、入力文中の「言い」が「い」と認識誤りを起こした例である。しかし、認識結果の「鈴木直子といいます」は、一般的には自然な文であり、誤りを判断することは難しい。しかし、本旅行案内タスクの学習データには稀な文であるため、意味距離値が閾値を越えてしまい、「鈴木直子と」の部分のみを翻訳することができた。

表 3-6 データに稀な言い回しに対する部分翻訳結果例

入力文	鈴木直子と言います。
認識結果	鈴木直子と <u>い</u> ます。
正解部分特定結果	鈴木直子と ...
部分翻訳結果	Naoko Suzuki ...

(e) 複数文からなる発話文の翻訳

音声認識結果に誤りがなくとも、一発声に複数の文が含まれている場合には、従来の翻訳システムでは文の境界を解析できないため、翻訳結果を出力できなかった。本方法を用いることで、各文を独立に翻訳することが可能となるため、結果的には、文境界を検出して各文を翻訳した場合と同じ結果を出力することが可能となる。たとえば一発話として「わかりました、ありがとうございました」と発声した場合、従来の翻訳システムでは、結果を出力できなかったが、本部分翻訳を行うことで、「I see, thank you very much」と翻訳することができた。

(f) 正解部分のみを翻訳することによる誤訳

文末での、肯定文か否定文か、あるいは、平叙文か疑問文か、を決定する表現が認識誤りを起こした場合、本方法で特定された部分文のみ翻訳することで、違った意味

に解釈される翻訳文を出力してしまうことがある。表 3-7 の例は、入力文中の「なく」という否定を表す助動詞が、認識されず欠落してしまった例である。結果的に、「都合で泊まれ」のみが特定され、正しくは、「can not」と否定文として翻訳されるべきものが、「can」と肯定文となり、逆の意味に翻訳されてしまった。

表 3-7 文末表現の認識誤りに対する部分翻訳の誤訳例

入力文	都合で泊まれなくなった。
認識結果	都合で泊まれ _____ になった。
正解部分特定結果	都合で泊まれ
部分翻訳結果	 can stay for some reason

3.1.6 まとめ

自由発話文における音声認識誤り文を解析するために、予め学習された話し言葉の表現パターンと入力文における表現パターンとの意味的類似性を用いて、認識結果から正しく認識された部分を特定する手法を提案した。さらに、本正解部分特定機能を音声翻訳システムに導入し、音声認識結果の正解部分のみを部分翻訳するシステムを構築した。本正解部分特定率の評価は、本方法が特定した部分の 96% が実際に正解部分であり、高い信頼度で正解部分を特定できることを示していた。また、本正解部分特定法が文の意味の理解に及ぼす効果を評価した。認識結果から特定された部分のみを提示することは、誤りを含んだ認識結果をそのまま提示するよりも誤った意味に理解してしまう割合を半分以上軽減する効果があった。さらに、特定した正解部分のみを部分翻訳することで、従来翻訳できなかった音声認識誤り文の 69% に対して、正しいか部分的にも正しい意味が理解できる翻訳文を出力できる効果が示された。このことは、音声認識誤り文に対して誤り部分を全て修復しなくても、必要な部分のみの修復、場合によっては修復を行わなくても、意味の通じる音声翻訳が可能であることを示唆している。しかし、音声翻訳の実用化を考えると、本部分翻訳手法による誤訳の増加は必ず解決しなければならない問題であると考えられる。次節では、実用化をめ

ざしてより確実に翻訳結果を保証する手法をめざす。

3.2 単語間の依存関係を用いた類似用例選択

3.2.1 はじめに

3.1節では、用例主導型の音声翻訳手法（隅田1994、古瀬1994）に基づいて、用例を用いて解析不可能な部分木を音声認識誤り部分として除去し、解析可能な部分のみを翻訳する部分翻訳手法を構築した。その結果、従来翻訳できなかった音声認識誤り文に対して、正しいか部分的にも正しい意味が理解できる翻訳結果を出力できることを確認した。しかし部分翻訳手法では、音声認識誤り部分は翻訳結果からも除去されてしまうため、発声文の意味を完全に伝えるには限度があり、翻訳できなかったものが、逆に誤った翻訳結果を出力してしまう場合も増加していた。従来、発声文の意味を完全に理解できる翻訳結果を出力するために、入力文と用例との文字の類似性、構文的類似性、意味的類似性などの情報を用いて認識誤り部分を訂正する方法が提案され、訂正による翻訳性能の向上効果を確認している（Ishikawa1998）。しかし、音声認識誤りが多数の部分で起こった場合は訂正効果にも限度があり、訂正後の結果を保証することは難しい。

音声翻訳の最も重要な機能は、発声文の意味を異なる言語で正しく伝えることであると考えられる。従って、認識結果を直接翻訳しなくても、既に対訳が保証されている用例データベースから同じ意味を持つ用例を選択し、その対訳を翻訳結果の代わりに出力することで、上記機能は達成すると考えられる。発声文の意図が限られるような限定したドメインを扱う場合を想定すると、訂正効果の限界を超えるような音声認識誤りが発生する現状を解決するためには、上記に述べた類似用例選択による翻訳手法は、有効なアプローチの一つであると考えられる。

入力文と類似した用例の対訳を模倣する考え方は、従来、事例に基づく翻訳手法として既に提案されている（Nagao1984、佐藤1991）。しかし、これらの方法は、音声認識誤り文が入力されることは想定されていない。類似用例選択法を用いた音声翻訳手法において、認識誤り文に対して、なるべく正しい意味を伝えるかまたは誤った意味に理解される翻訳結果の出力を避けるための対策が別途必要であると考えられる。

ここでは、旅行会話ドメインに限定された音声翻訳システムにおいて、意味を理解する上で必要な単語（以後キーワードと呼ぶ）を認識結果から抽出し、同じキーワー

ドを含む用例を類似した意味を持つ用例とみなして選択し、その対訳を翻訳結果とする音声翻訳手法について述べる。特に、音声認識誤りを起こしても用例選択をなるべく誤りなく行うために二つの対策を行った。一つは、キーワードが音声認識誤りを起こした場合でも誤った用例を選択することを避けるために、認識誤りを起こしたキーワードを推定し、残りの正しく認識されたとみなされたキーワードのみから用例選択を行うことである。認識結果に含まれるキーワード間の依存関係の自然性を調べ、他のキーワードとの依存関係が不自然なキーワードを認識誤りであると推定した。もう一つは、少ない用例でなるべく各キーワードが発声されるであろう言い回しを漏れなくカバーするために、類似した言い回しをするキーワードが同じクラスとなるように、用例データベースを用いてキーワードを自動クラス化したことである。

本節では、3.2.2節で音声認識誤りを考慮した類似用例選択による音声翻訳手法について述べ、本手法の認識誤りに対する有効性を評価して報告する。また3.2.3節では、キーワードの自動クラス化法について説明し、類似用例選択性能への効果を報告する。最後に3.2.4節でまとめる。

3.2.2 音声翻訳のための類似用例選択

(I) 認識誤り推定を用いた類似用例選択法

限定されたドメインのみを扱う音声翻訳システムにおいて、音声認識誤りに頑健なシステムを構築するために、音声認識結果に含まれる内容語をキーワードとみなし、キーワードのみを用いて類似した用例の対訳文を出力するシステムを構築している。いくつかのキーワードが音声認識誤りを起こした場合に意味の異なる用例を多く選択してしまうことを避けるために、認識誤りを起こしたキーワードを推定し除去するしくみを導入している。

類似用例選択による音声翻訳システムを図3-5に示す。この方法では、認識結果を直接翻訳する代わりに、認識結果に含まれる正しく認識されたと推定されるキーワードのみを最も多く含んでいる用例を意味が類似している用例であるとみなして選択する。選択された用例が一つに絞り込めなかった場合には、選択された全ての用例がユーザ

に提示され、ユーザにより最適な用例が一つ決定される。そして決定された用例の対訳文が翻訳結果として出力される。認識結果に認識誤りを起こしたキーワードが含まれていると、意味が異なる用例が多数選択されてしまいユーザが最適な用例を決定する際の作業が煩雑になってしまう。場合によっては、正しく類似性の判断ができず、意味が異なる用例ばかりが選択されることも考えられる。この問題を避けるために、認識誤りを起こしたキーワードを推定し、推定されたキーワードを除いた残りのキーワードのみを用いて用例の選択を行うことで、全く意味の違う用例がなるべくユーザに提示されないようにした。認識誤りを起こしたキーワードの推定は、次のように行う。予め学習用例を用いて依存関係にあるキーワード対を抽出しておく。音声翻訳時には、音声認識結果に存在する複数のキーワードと既に抽出された依存関係にあるキーワード対とを比較し、同じ発声内の他のどのキーワードとも依存関係が見出せないキーワードがあれば、そのキーワードを認識誤りキーワードと推定する。従って、用例の選択は、お互いに依存関係にあって自然であると推定されるキーワードが同じ発声内に見つかったキーワードのみを用いて行われる。

以下に動作例を示す。たとえば、「地元のビールを飲んでみたいです」という発声に対し、「地元のビールを吸ってみたいです」と認識誤りを起こした場合を考える。内容をキーワードとした場合には、「地元」「ビール」「吸っ」がキーワードとして抽出される。一方、用例を学習した結果、依存関係を持つキーワード対として「ビール」と「飲む」、「タバコ」と「吸う」、「地元」と「ビール」、などが予め抽出されている。これらと認識結果に含まれるキーワードを比較した結果、「地元」と「ビール」は依存関係が認められるが、「吸っ」は、「地元」とも「ビール」とも依存関係は認められない。そこでこの例では、「吸っ」を音声認識誤りとみなして除去し、残りの依存関係が認められたキーワード対「地元」と「ビール」のみを用いて、それらが含まれる用例を用例データベースから選択する。その結果、たとえば「このビールは地元のですか?」「地元のビールが欲しいです」の二つの用例が選択された場合、これら二つの用例をユーザに提示し、最適な用例がユーザにより選択される。選択された「地元のビールが欲しいです」の対訳である「Any local beer, please」が翻訳結果として出力される。

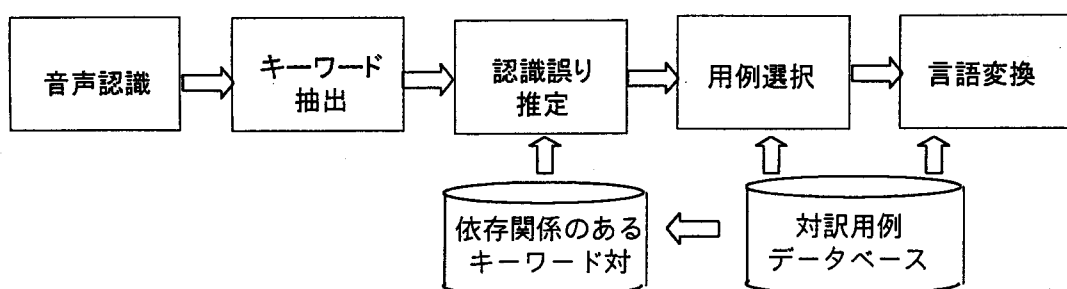


図 3-5 類似用例選択による音声翻訳システム

(II) 類似用例選択が音声認識誤りの校正に及ぼす影響

本用例選択手法が、音声認識誤りに頑健な音声翻訳手法として有効かどうかを調べるために、直接認識結果を翻訳する代わりに類似用例を選択することの有効性と、類似用例選択において認識誤り推定を行うことの有効性について調べた。発声文と同じ意味を持つ用例を選択できれば、既に品質が保証されている対訳文は意味の正しく伝わる翻訳結果であると考えられる。そこで、有効性の評価は類似用例選択性能を用いて行い、正しい用例が選択されているかどうかで翻訳性能を評価する。

(II - I) 音声認識結果と選択された類似用例との意味理解度の比較

認識誤りを起こした場合、類似用例の選択を行うことが直接認識誤り文を翻訳するよりも有効であるかどうかを推定するために、音声認識結果と選択された類似用例からユーザにより決定された最適用例とのどちらが、発声文の意味を正しく伝えられるかを比較する。そこで、元の発声文と音声認識結果またはユーザにより決定された用例とを比較し、発声文と同じ意味と理解できるかどうかを以下の 5 段階の評価レベルで評価した。評価は、日本人 2 人で行った。評価レベルを以下に示す。また、評価実験データ仕様を表 3-8 に示す。

1. 異なる意味である。
2. 部分的には同じ意味であるが、入力文の意味は理解できない。
3. 部分的に同じ意味であり、入力文の意味を推定することができる。
4. ほぼ同じ意味である。
5. 完全に同じ意味である。

表3-8 意味理解度比較実験データベース仕様

ドメイン	ATR言語データベース②日英会話表現集旅行会話
用例DB構築用文	1051文
抽出されたキーワード対	1607単語
評価用入力文	上記1051文中の誤認識302文(closed文)
誤認識文におけるキーワード認識率	58.43%

旅行会話 1051 文を用いて用例データベースを作成し、同じ 1051 文を入力した場合の認識誤り結果の一部 302 文を評価の対象とした。本データベースは会話文を想定しているため、場面を想定しないと意味が曖昧である場合がある。そこで、評価の際には、使用される場面を評価者に提示した。評価結果を表 3-9 に示す。次の結果が得られた。

- ・「意味が理解できる」（評価レベル 3～5）と評価された割合を比較すると、音声認識誤り結果では、評価レベル 3～5 の総和が 28%であるのに対し、用例選択結果では、51%とほぼ倍増している。
- ・「完全に同じ意味である」（評価レベル 5）と回答した割合は約 25%であり、音声認識が誤った場合においても 4 分の 1 は、完全に回復していることがわかる。
- ・「異なる意味に理解される、または理解できない」（評価レベル 1～2）は、音声認識誤り結果では 71.3%であるが、用例選択結果では 48.7%と 31.6%削減されている。

表 3-9 音声認識誤り文と用例選択文との意味理解度の比較 [%]

レベル	5(完全理解)	4(理解可能)	3(推定可能)	2(理解不可能)	1(異なる理解)
認識結果	5.63	9.44	13.58	24.67	46.69
用例選択	25.33	8.11	17.88	14.74	33.94

(II - II) 認識誤りキーワード推定が用例選択性能に及ぼす効果

本節では、キーワードの認識誤りを推定することの用例選択性能に及ぼす効果を評価し報告する。認識誤りを起こしたキーワードを推定し、推定されたキーワードを除き残りのキーワードのみを用いて類似用例選択を行うことの効果を確認するために、この方法で用例選択を行った場合の用例選択性能を、認識誤りキーワードの推定を行わず認識結果に含まれている全てのキーワードを用いて類似用例選択を行った場合と比較した。ここでは、用例文と同じ文を入力した場合の認識誤り結果に対して、入力文と同じ用例文が選択されるどうかを、以下式3-3、式3-4に示した再現率及び適合率を用いて評価した。用例データ及び評価実験に用いたデータ仕様は表3-8に示したものと同一である。

$$\text{再現率} = \frac{\text{用例選択に成功した文数}}{\text{用例数}} \times 100 \text{ [%]} \quad (3-3)$$

$$\text{適合率} = \frac{\text{用例選択に成功した文数}}{\text{出力文数}} \times 100 \text{ [%]} \quad (3-4)$$

表 3-10 に用例選択結果を示す。認識誤りを起こしたキーワードの推定をしなかった場合には、用例候補が多数出力されることが多かったが、認識誤りの推定を行い正しく認識されたと推定されたキーワードのみを用いて用例を選択することにより、再現率が 8.13% 下がり、適合率が 40.5% 向上している。再現率の低下は、依存関係にあるキーワード対の一方が認識誤りを起こしてしまったために、依存関係にあるキーワード対が見つけれず、全キーワードが認識誤りとみなされリジェクトしたことが主な原因であった。しかし、適合率が 5.6 倍と大幅に改善されている。認識誤りを起こしたキーワードを推定することにより、類似用例選択を行う際の用例候補数を効率よく絞り込むことが可能であることを示

唆している。

表 3-10 認識誤りキーワード推定が類似用例選択性能に及ぼす効果 [%]

- 用例と同文入力の場合の認識誤り文に対する結果 -

認識誤り推定	再現率	適合率
あり	74.89	49.26
なし	82.96	8.72

3.2.3 類似用例選択のためのキーワードクラス化法

(I) キーワードクラス化の必要性

限定したドメインを扱うことで、発声される意図を表す用例及び使用されるキーワードをコーパスとして準備できたとしても、各キーワードが使われるであろう言い回しを全て用例として学習するのは困難である。なるべく様々な言い回しをカバーするためには、各キーワードが用例中で用いられた言い回し以外の言い回しで入力された場合であっても適切な用例を選択可能にする必要がある。そこで、同じ言い回しで利用できるキーワードが同じクラスとなるように、シソーラスと用例とを用いてキーワードを自動クラス化し、同じクラスに属するキーワードであれば各々のキーワードが含まれる言い回しを共有して使用できるしくみを考えた。たとえば「コーヒーが飲みたい」という用例については、コーヒーも含めてその他の飲み物を表すキーワードをクラス化し、「コーヒー」というキーワードの代わりに「(飲み物クラス) が飲みたい」という飲み物を表すクラスに置き換えて扱う。「紅茶」や「ジュース」など、用例に含まれないかまたは用例に含まれていても他の言い回しの中でのみ使用されている「飲み物」を表すキーワードに対して、この用例を選択することが可能となる。従来、このようなクラスの概念としてシソーラス辞書が頻繁に使用されている (Furuse1994、隅田 1994)。しかし一般的なシソーラス辞書を用いる場合には次のような問題がある。

- ・一般的なシソーラス辞書では多くの単語が二つ以上の意味クラスに含まれている。

そのため、使用されているドメインに適切な意味クラスに絞ってクラス化することが困難である。

- ・シソーラスで同じクラスであっても、ドメインに依存して他の単語とは異なった特別な使われ方をする単語を区別できない。たとえば、「酒が強い」の「酒」の部分に「コーヒー」や「ミルク」が置き換えられると不自然であるが、シソーラスでは、これらのものはすべて「飲み物」として同じクラスになってしまう。

これらの問題を解決するために、シソーラス辞書に基づいたクラスを初期クラスとして、依存関係にあるキーワード対の情報を用いて、上記の問題を回避するように初期クラスを調整しながら、最終的にドメインに適切なクラスを決定する方法を提案する。

(II) キーワードクラス化法

本クラス化の目的は、各キーワードが用例にはない言い回しで入力されても、正しい意味を表す用例を選択できることであり、そのためには、依存関係にあるキーワード対を適切に推定することで適切な用例文を選択可能にすることが必要である。従って、同じキーワードに依存するまたは同じキーワードに依存されるキーワードどうしが同じクラスに属するようにクラス化を行うべきであると思われる。そこで、同じクラスに属するキーワードを決定するための「キーワード間の類似度」を、以下のように定義した。

依存関係にある二つの単語を係り元単語と係り先単語と呼ぶことにする。たとえば、「コーヒーを飲む」という文については、係り元単語は「コーヒー」であり、係り先単語は「飲む」である。係り元であるキーワードをクラス化する場合には、式 3-5 にて算出される係り元単語間の類似性を用いる。2つの係り元単語 W_j と W_i 間の類似度 $sim(W_j, W_i)$ は、同じ係り先単語に頻繁に依存している場合値が大きくなり、類似性が高いと判断される。

$$Sim(W_j, W_i) = \sum_{m=1}^M R(W_j, Z_m) \times \sum_{m=1}^M R(W_i, Z_m) \quad (3-5)$$

$$R(W_j, Z_m) = \frac{FreqPair(W_j, Z_m)}{Freq(W_j)} \quad (3-6)$$

if $R(W_j, Z_m) = 0$, then $R(W_i, Z_m) = 0$

if $R(W_i, Z_m) = 0$, then $R(W_j, Z_m) = 0$

$Sim(W_j, W_i)$: 係り元単語 W_j と W_i の類似度

Z_m : m 番目の係り先単語

$FreqPair(W_j, Z_m)$: 係り元単語 W_j が係り先単語 Z_m に依存する数

$Freq(W_j)$: 単語 W_j の出現数

M: 全係り先単語数

同じように、係り先単語 Z_m と Z_n の類似度 $sim(Z_m, Z_n)$ は、上記の式 3-5、式 3-6 の係り元単語と係り先単語とを入れ替えた式 3-7、式 3-8 で求められる。

$$Sim(Z_m, Z_n) = \sum_{j=1}^J R(W_j, Z_m) \times \sum_{j=1}^J R(W_j, Z_n) \quad (3-7)$$

$$R(W_j, Z_m) = \frac{FreqPair(W_j, Z_m)}{Freq(Z_m)} \quad (3-8)$$

if $R(W_i, Z_m) = 0$, then $R(W_i, Z_n) = 0$

if $R(W_i, Z_n) = 0$, then $R(W_i, Z_m) = 0$

$Sim(Z_m, Z_n)$: 係り先単語 Z_m と Z_n の類似度

W_j : j 番目の係り元単語

$FreqPair(W_j, Z_m)$: 係り元単語 W_j が係り先単語 Z_m に依存する数

$Freq(Z_m)$: 単語 Z_m の出現数

J: 全係り元単語数

これらの式を用いて求められたキーワード間の類似度を用いて、シソーラスのみから決定された初期クラスを再度クラス化する。手順を図 3-6 に示す。ドメインに最適なクラス化を行うために、以下の二つの条件を用いてクラス化した。

- ・初期クラスにおいてクラス内の全キーワード間の類似度が一定以下である場合は、

使用するドメインには不適切なクラスであると判断し、相当するクラスに属していたキーワードは1キーワード1クラスとして扱われる。

- ・初期クラスにおいて、クラス内にあるキーワードが同じクラスにある他のどのキーワードとも類似度が一定以下である場合、そのキーワードのみをクラスから外し、1キーワード1クラスとして扱う。

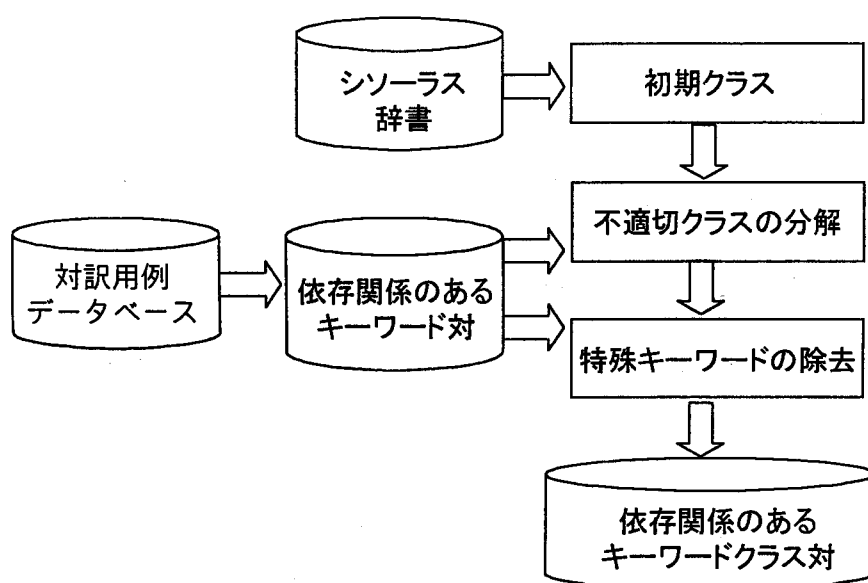


図3-6 キーワードクラス化手順

(Ⅲ) キーワードクラス化法の類似用例選択性能に及ぼす効果

本節では、先に述べたキーワードクラス化法が用例選択性能に及ぼす効果を評価し報告する。上記のキーワードクラス化が、用例選択性能に効果があることを確認するために、シソーラス辞書のみを用いてキーワードクラスを決定した場合、つまり、本方法の初期クラスをそのまま採用した場合と本方法によるクラス化を行った場合との、用例選択性能を比較した。

(Ⅲ - I) 用例選択実験条件

用例選択実験の仕様を表3-11に示す。シソーラス辞書には国立国語研究所による分類語彙表（中野1996）を用い、意味コードの上から4桁が同じ場合に同クラスとみなした。ドメインは「ATR言語データベース②日英会話表現集旅行会話」の中のレストランにてお客側が使用する発話とし、対象となる930文うち770文を用例および依存関係を持つキーワード対学習用に、残り160文を評価用に用いた。ここでは一般的な内容語をキーワードとして扱い、具体的には、名詞、動詞、形容詞、副詞、感動詞に相当する単語をキーワードとした。本実験で扱ったレストランのお客側発話においては、食事の「メニュー内容」に相当するキーワードが多数発声されることが想像され、これらのキーワードを含む言い回しも多種想像できる。しかし、各「メニュー内容」に対して全ての単語を用例中にカバーすることは困難であり、「メニュー内容」は、本クラス化の有効性が最も顕著に図れるクラスであると思われる。従って本実験では、「メニュー内容」に相当するキーワードのほとんどが係り元単語であることから、まず、係り元単語であるキーワードのみをクラス化の対象としクラス化の有効性を評価した。本学習用例に含まれるクラス化の対象となるキーワードの異なり語数は675語であった。また、単語間類似度の閾値が0、つまり1つでも同じ係り先単語に依存している単語対は同じクラスになるようにクラス化を行った場合、提案法では、675語中392のキーワードが61クラスに分類され、その他のキーワードは1キーワード1クラスとして扱われた。シソーラス辞書のみを用いる従来法のクラス化では675キーワードが367クラスに分類された。

上記の条件下で、発声文の認識結果から類似用例選択法にて用例を選択し、選択された用例文と元の発声文との意味の一致度を評価した。選択された用例が複数個ある場合は、選択された用例の中から、最適なものを一つ選択して用いた。また本用例選択手法は、用例中に含まれる依存関係にあるキーワード対がどれだけ認識結果に含まれているかを調べることで行われる。従って入力音声中に依存関係にあるキーワード対が認められない場合や、もともと入力音声中に一つしかキーワードが含まれない場合には、用例が選択できない。本実験ではこの場合をリジェクトとし、該当用例のないことを発声者に知らせるメッセージを翻訳文のかわりに出力することにした。用例が選択された場合には、正しい入力文に対する選択された用例文の正しさを、次の「1」

から「5」の五つのレベルで評価を行った。評価者は3名の日本人である。

- 1：選択された用例は全て異なる意味であり、最適用例を選択できない。
- 2：部分的には同じ意味であるが、全体としては違う意味である。
- 3：何とか同じ意味であると判断できる。
- 4：ほぼ同じ意味である。
- 5：完全に同じ意味である。

表3-11 用例選択実験仕様

ドメイン	ATR言語データベース②日英会話表現集旅行会話(レストラン)
キーワード内容	名詞、動詞、形容詞、副詞、感動詞
キーワードクラス構築用DB	レストランドメイン:770文
キーワードクラス構築条件	係り先側単語675語のみを対象
認識方式	One-passDPによるN-best候補探索
評価用DB	正解文:レストランドメイン160文(オープン文) 認識誤り文:上記文の認識誤り結果88文

(Ⅲ - Ⅱ) キーワードクラスが用例選択性能に及ぼす効果

(a) 用例とは異なる発声文に対する効果

まず、用例とは異なる文セット160文を正しい認識結果文とみなし、これらを入力した際の用例選択性能を評価した。図3-7に結果示す。この結果は選択された用例に与えられた各評価レベルの出現頻度を表している。

- ・理解できると評価された(評価レベル3以上5以下)頻度は提案法の方が多い(提案法64.8%、従来法57.8%)。提案法により、正しく理解できる用例を出力する場合が増加したことを示しており、提案法が類似用例文の選択に有効であることを示唆している。
- ・提案法により理解できない(評価レベル1, 2)頻度が減り(従来法30.1%、提案法17.2%)、代わりにリジェクトされる割合が増えている。用例にはない意味を持つ文に対し、従来法では誤った用例を選択していたが、提案法では正しくリジェクト

されていることがわかった。

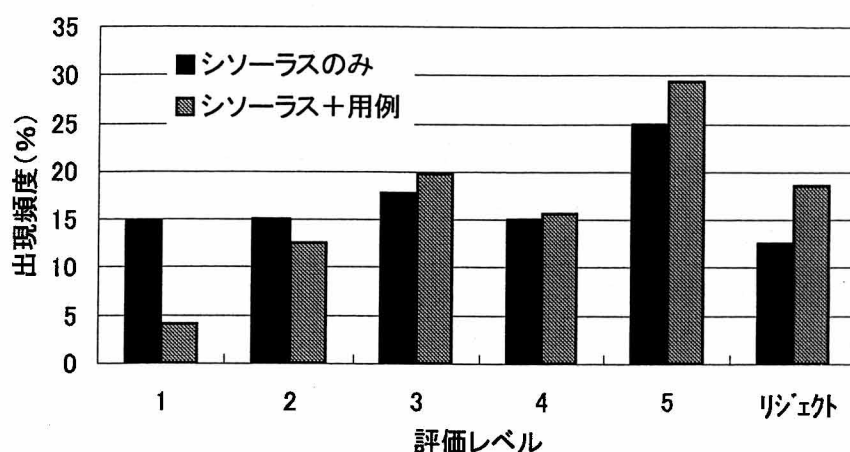


図 3-7 本クラス化手法が類似用例文選択性能に及ぼす効果
- 用例文とは異なる文に対する結果 -

(b) 音声認識誤り文に対する効果

次に、認識誤りを起こした認識結果 88 文に対する類似用例選択性能を、上記と同様の方法により 5 段階レベルで評価した。結果を図 3-8 に示す。

- ・ 正しく理解できると評価された（評価レベル 3 以上 5 以下）文数に両者の差が殆ど無い（従来法 22.6%、提案法 27.2%）。正しい用例の選択を可能にする提案法の効果は、みられなかった。
- ・ リジェクト率は 53.4% から 60.2% に増加しており、一方、理解できない用例を出力する頻度（評価レベル 1 または 2）は減少している（従来法 25%、提案法 12.6%）。このことから、提案法には、先のオープン文に対する評価実験の傾向と同様に、従来法で理解できない用例の選択率を減少させる効果があることがわかった。

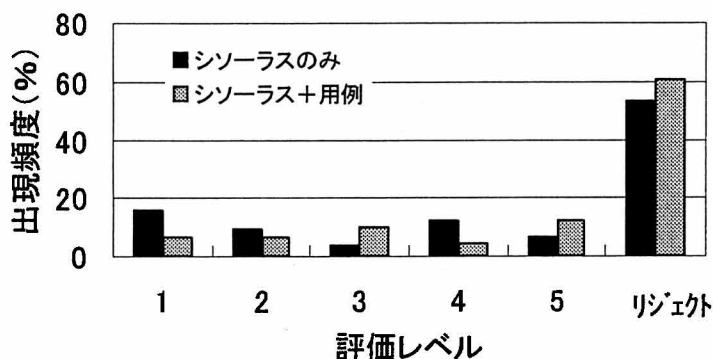


図 3-8 本クラス化手法が類似用例選択性能に及ぼす効果
- 認識誤り結果に対する結果 -

従来法では正しい用例が選択されなかったが本クラス化法ではリジェクトされた例に以下のものがある。「ホットドッグには何を付けますか?」という入力文を「ホットドッグには何をしますか?」と認識誤りした結果に対して用例選択を行った。従来法では「デザートはありますか?」という用例文が選択され、「デザート」が「ホットドッグ」が置き換えられ「ホットドッグはありますか?」という意味の異なった用例がユーザに提示された。従来のシソーラスでは「デザート」と「ホットドッグ」とは同クラスであったが、本クラス化では、「デザート」を特別な単語としてクラスから除去していたため、誤った用例を選択することはなくリジェクトされた。

3.2.4 まとめ

音声翻訳システムの実用化をめざして、音声認識誤りを起こしても誤った意味に理解される翻訳結果の出力を避けるために、類似用例選択による音声翻訳手法を提案した。本手法は、認識結果を直接翻訳するのではなく、認識結果に含まれるキーワードを用いて、予め学習された用例から意味が類似した用例を選択し、その対訳を翻訳結果として出力するものである。選択された用例が一つに絞り込めなかった場合には、

選択された複数用例をユーザに提示し、ユーザが選択した用例の対訳文を翻訳結果として出力する。本手法において二つの対策を行った。一つは、音声認識誤りに頑健なシステムを構築するために、用例選択する前に予め学習されたキーワード間の依存関係を用いて音声認識誤りを起こしたキーワードを推定し、これらを除去した残りの依存関係を確認できたキーワードのみを用いて用例を選択するしくみにしたことである。本用例選択手法の効果及び用例選択の際に音声認識誤りを起こしたキーワードの推定を行うことの効果調べた。その結果、音声認識誤りを起こした場合には、認識結果を直接翻訳するよりも、類似した意味を持つ用例を選択してその対訳を翻訳結果にした方が、意味の正しく伝わる翻訳結果の出力頻度が約 2 倍に向上することを確認した。また、用例選択の際に音声認識誤りを起こしたキーワード推定を行うことで、用例選択再現率は約 8 % 低下するが、用例選択適合率が 5.6 倍と大幅に改善され、ユーザが複数用例から最適用例を決定する際の作業の煩雑さを改善できる可能性を確認した。もう一つの対策は、用例にはない言い回しでキーワードが入力された場合でも類似した用例の選択を可能とするために、類似した言い回しの中で同じように用いられるキーワードが同じクラスとなるようにキーワードを自動クラス化したことである。一般的なシソーラスを用いて作成された初期クラスを、用例から抽出されたキーワード間の依存関係を用いてドメインに最適に自動調整するクラス化法を提案し有効性を評価した。その結果、正しく認識された文に対する用例選択率を 1.2 倍に向上させる効果と、認識誤り文または用例にはない文に対する誤選択率を 57.1% に削減する効果を確認した。

第4章 まとめ

本論文では、連続音声認識技術の実用化をめざす立場から、現状の連続音声認識が抱える問題点を整理し、認識技術を音声翻訳システムに応用する研究開発を通して、先の問題点の解決を図ったものである。

第1章は序論であり、従来の連続音声認識の現状と実用化する上での問題点を明らかにするとともに、本研究の目的を述べた。

第2章では、連続音声認識性能の向上をめざした研究成果を報告した。特に、現状の統計的手法による連続音声認識技術では導入が困難な、音声単位にまたがる音響的特徴を用いた単語候補探索手法を提案し、認識性能に対する本手法の効果について述べた。

一つの音響的特徴は音声単位の継続時間である。音声単位の継続時間は、通常のHMMでは、状態間の遷移確率の形で間接的に反映されているが、認識箇所に対してダイナミックな形で反映されていない。そこで、各音声単位毎に継続時間を独立に制御するのではなく、それらの相互関係を考慮した継続時間制御を施すことで認識性能の向上を図った。音節継続時間は、当該音節の種類だけでなく、その近傍の音節の種類、特に2音節以内の音節の種類に大きく左右されることを明らかにし、この知見を踏まえた音節継続時間予測法を提案した。本予測法を用いて音節マッチング区間を制御して認識性能を確認した結果、継続時間を予測しない場合や音節間の相互関係を用いないでマッチング区間を制御した場合と比べて、認識率の向上が確認できた。また、本音節継続時間予測法を用いて逐次的に状態継続時間を制御する手法も合わせて提案した。Left-to-right型の認識処理において、認識の最小単位である状態の継続時間を制御することは、先の音節継続時間を制御する効果に加えて、認識処理と同期して逐次的に継続時間を制御できる効果がある。認識実験から、単語または文の認識率の向上と、処理時間の短縮を確認できた。

他の一つの音響的特徴は音声単位にまたがる音声認識誤り特性である。話者適応技術などによる従来の音響モデルの改良方法では、音声単位にまたがった部分での音声認識誤り対策が不十分のままになることが考えられる。そこで、複数の音声単位にまたがった音声認識誤り特性を分析し、この認識誤りを解決することにより認識性能の

向上をめざした。本音声認識誤り特性を用いた音素系列候補追加モデルにより、従来では候補として出力できなかった正解音素を正解候補に含むことができることが確認できた。さらに、音声認識誤り特性が品詞に依存していることを明らかにし、品詞の違いを考慮した音声認識誤り特性をマルチ発音辞書の構築に応用することを提案した。本認識誤り特性を用いて構築された発音辞書が、品詞を考慮しない場合に比べてさらに認識性能を向上させることが確認できた。

第3章では、音声認識誤りが生じた場合に全く意味の理解できないまたは意味の異なる認識結果が出力されてしまう従来の問題の解決策として、認識結果全体の意味的な整合性を評価することで、従来の統計的に求めた単語連接制約のみでは特定できない認識誤り部分特定し、これを校正した結果を提供できる処理を考案した。

まずは、特定された音声認識誤り部分を除去し、正解部分のみを翻訳することで、意図を正しく伝達できるしくみを考案した。音声認識誤り部分は、予め学習された話し言葉の表現パターンと入力文における表現パターンとの意味的類似性を用いて特定される。その結果、本手法が高い信頼度で音声認識誤りを特定できることを確認した。さらに、正解部分を部分翻訳した結果、従来翻訳できなかった音声認識誤り文に対して、正しいか部分的にも正しい意味が理解できる翻訳文を出力できる効果を確認できた。しかし、音声認識誤り部分を修復せずに除去する部分翻訳手法は、除去した部分が発声文の意味を理解する必要な部分であった場合、逆に誤った翻訳結果が出力されてしまう場合がある、という課題を残した。

次に、上記課題を踏まえて、音声認識誤りが発声文の意味を理解する上で必要な単語で起こった場合にでも、意味が理解できない翻訳結果または異なった意味に理解されてしまう翻訳結果を出力しないために、認識結果を直接翻訳するのではなく、認識結果に含まれるキーワードのみを用いて発声文と意味が類似した用例を自動選択し、その対訳を翻訳結果として出力する類似用例選択による音声翻訳手法を構築した。音声認識誤りに対応するために、キーワード間の依存関係を用いて音声認識誤りを起こしたキーワードを推定し、それらを除去した残りのキーワードつまり依存関係にあるキーワード対のみから用例を選択した。さらに、キーワードが用例には存在しない言い回しで発声された場合にも適切な用例を選択できるように、類似した言い回しの中で使用されるキーワードが同じクラスとなるようにキーワードの自動クラス化を行っ

た。評価した結果、本用例選択手法は音声認識誤りを起こした場合に、意味の理解できる翻訳結果を出力するのに有効であること、また認識誤りを起こしたキーワードを推定することが、用例選択性能を向上させるのに有効であること、さらに、従来の音声翻訳に比べて、音声認識誤り文に対して理解できないまたは全く意味が異なる翻訳結果を出力してしまうという従来の問題点を軽減する効果があることを確認した。

今後は、実用化が可能な音声認識技術、音声翻訳技術を確立するために、さらに、実用の際の使用目的、ドメインの種類、ハード仕様などのイメージを具体化していく中で、音声認識、音声翻訳処理のより具体的な性能目標設定し、実用化の可能性評価とそれを踏まえた新たな研究課題設定を行っていく必要があると考えている。

- 参考文献 -

第 1 章

Schwartz1985 :

R.Schwartz, Y.Chow, O.Kimball, S.Roucos, M.Krasner, and J.Makhoul.

Context-dependent modeling for acoustic-phonetic recognition of continuous speech.

Proc. ICASSP85, pp.1205-1208, 1985.

Lee1990 :

K.Lee, S.Hayamizu, H.Hon, C.Huang, J.Swartz, and R.Weide.

Allophone Clustering for Continuous Speech Recognition

Proc. ICASSP90, pp.749-752, 1990.

Takami1992 :

J.Takami and S.Sagayama.

A successive state splitting algorithm for efficient allophone modeling.

Proc. ICASSP92, Vol.1, pp.573--576, 1992.

Takami1993 :

鷹見淳一、嵯峨山茂樹

逐次状態分割法による隠れマルコフモデル網の自動生成

電子通信学会論文誌(D-II), Vol. J76-D-II, No.10, pp.2125-2164, 1993

Hori 1997 :

堀貴明、加藤正治、伊藤彰則、好田政紀

音素決定木に基づく逐次状態分割法による HM-Net の検討

電子通信学会論文誌(D-II), Vol.J80-D-II, No.10, pp.2645-2154, 1997

Singer1996 :

H.Singer and M.Ostendorf

Maximum likelihood successive splitting

Proc. ICASSP96, pp.601-604, 1996

Ostendorf1989 :

M.Ostendorf and S.Roukos.

A stochastic segment model for phoneme-based continuous speech recognition

IEEE Trans, Acoust.,Speech & Signal Process, Vol.37, No.12, pp.1857-1869, 1989

Franzini1990 :

M. Franzini and K-F Lee.

Connectionist viterbi training a new hybrid for continuous speech recognition.

Proc. ICASSP, Vol.1, pp.425-428, 1990

Hanna1999 :

P.Hanna, J.Ming, and F.J.Smith

Inter-frame dependence arising from preceding and succeeding frames-Application to speech recognition.

Speech communication, Vol.31, No.4, pages 301-312, 1999

Sato1991 :

S.Sato.

Example-based Machine Translation.

Doctorial Thesis, Kyoto University

Sumita1992 :

E.Sumita and H.Iida.

Example-Based Transfer of Japanese Adnominal Particles into English.

IRICE TRANS. Inf. & SYST., Vol.E75-D, No4, pp.585-594, 1992

Maruyama1992 :

H.Maruyama and H.Watanabe:

Tree Cover Search Algorithm for Example-based Translation.

Proc. TMI-92 pp.173-184, 1992

Grishman1992 :

R.Grishman and M.Kosaka:

Combining Rationalist and Empiricist Approaches to Machine Translation.

Proc. TMI-92, pp.263-274, 1992

Furuse1994 :

O.Furuse and H.Iida

Constituent Boundary Parsing for Example-based Machine Translation

Proc. Coring94, pp.105-111, 1994

Bahl1983 :

L.R.Bahl, F.Jelinek, and R.L.Mercer

A Maximum likelihood to continuous speech recognition.

IEEE Trans. on Pattern Analysis and Machine Intelligence, pp.179-199, 1983

第 2 章

2 - 1 節

Ferguson1980 :

J.D.Ferguson.

Variable duration models for speech.

Proc. of Symposium on the Application on Hidden Markov Models to Text and Speech, pp. 143-179, 1980

中川 1989 :

中川聖一、橋本泰秀

HMM 法とベイズ確率を用いた連続音声のセグメンテーション

電子通信学会論文誌(D-II), Vol.J72-D-II, No.1, pp.1-10, 1989

Rabiner1989 :

L.R.Rabiner

A Tutorial on Hidden Markov Model and selected application in speech recognition.

Proc. IEEE, Vol.77, No.2, pp.257-286, 1989

Levinson1986 :

S.E.Levinson

Continuously variable duration Hidden Markov Models for automatic speech recognition.

Computer Speech and Language, Vol.1, No.1, pp.29-45, 1986

Russel1985 :

M.J.Russel and R.K.Moore

Explicit modeling of state accuracy in Hidden Markov Models for automatic speech recognition.

Proc. ICASSP85, pp.5-8, 1985

高田 1991 :

高田義広、脇田壽、坪香英一

継続時間分布モデルの比較検討ーポアソン分布、ガンマ分布、対数正規分布

音響学会講演論文集, 2-5-14, pp.75-76, 1991

Lee1989 :

Automatic speech recognition

Kluwer Academic Publishers.

Rabiner1989 :

L.R.Rabiner, J.G.Wilpon and F.K.Soong

High performance connected digit recognition using Hidden Markov Models.

Proc. ICCASP89, pp.1214-1225, 1989

松永 1989 :

松永昭一、嵯峨山茂樹

双方向網探索による文節音声認識と時間軸離散化の効果

電子通信学会技術報告書, SP89-16 (1989)

Singer1992 :

Singer Harald、嵯峨山茂樹

行列パーザとその HMM 音声認識への応用

電子通信学会技術報告書, SP92-76, 1992

比企 1967 :

比企静雄

連続音声中の音韻区分の持続時間の性質

電子通信学会論文誌, Vol.50, No.5, pp.849-856, 1967

佐藤 1997 :

佐藤大和

単語における音韻継続時間と発話のタイミング”

日本音響学会音声研究会資料, S77-31, 1977

匂坂 1980 :

匂坂芳典、東倉洋一

連続音声における音韻継続時間の性質

日本音響学会音声研究会資料, S80-34, 1980

海木 1990 :

海木延佳、武田一哉、匂坂芳典

文音声における音韻継続時間長の設定

電子通信学会技術報告書, SP90-2, 1990

Takeda 1990 :

K.Takeda, Y.Sagisaka and H.Kuwabara

On sentence-level factors governing segmental duration in Japanese

J.Acoust.Soc.Am., Vol.86, No.6, 1990

大黒 1988 :

大黒慶久、橋本康秀、中川聖一

構文解析駆動型日本語音声理解システム

電子通信学会技術報告書, SP88 - 87, 1988

大坂 1993 :

大坂幸弘、牧野正三、曾根敏夫

発声速度と音素環境を考慮した持続時間情報を用いた単語音声認識

音響学会講演論文集, 2-4-1, pp.49-50, 1991

松尾 1989 :

松尾広、牧野正三、城戸健一

音素の持続時間モデルを利用した単語音声認識に関する研究
電子通信学会技術報告書, SP89-90, pp.33-40, 1989.

2 - 2 節

Schwartz1985 :

R.Schwartz, Y.Chow, O.Kimball, S.Roucos, M.Krasner and J.Makhoul
Context-dependent modeling for acoustic-phonetic recognition of continuous speech.
Proc. ICASSP85, pp.1205-1208, 1985

Lee1990 :

K.F.Lee, S.Hayamizu, H.Hon, C.Huang, J.Swartz and R.Weide
Allophone clustering for continuous speech recognition.
Proc. ICASSP90, pp.749-752, 1990

Takami 1992 :

J.Takami and S.Sagayama
A successive state splitting algorithm for efficient allophone modeling.
Proc. ICASSP92, Vol.1, pp.573-576, 1992

Singer1996 :

H.Singer and M.Ostendorf
Maximum likelihood successive splitting
Proc. ICCASP96, pp.601-604, 1996

Ohkura1992 :

K.Ohkura, M.Sugiyama and S.Sagisaka
Speaker adaptation based on transfer vector field smoothing with continuous mixture density
HMMs

Proc. ICSLP92, pp.369-372, 1992

Lee1991 :

C.H.Lee, C.H.Lin, and B.H.Juang

An study on speaker adaptation of the parameters of continuous density HiddenMarkov Models.

IEEE Trans on Signal Processing, Vol.39, no.1, 1991

Leggeth1994 :

C.J.Leggetter and P.C.Woodland

Speaker adaptation of continuous densityHMMs using multivariate linear regression.

Proc. ICSLP94, pp.451-454, 1994

新津 1979 :

新津善弘, 三輪譲二, 牧野正三, 城戸健一

単語音声自動認識における言語情報の一利用法

電子通信学会論文誌 (D) , Vol.62-D, No.1, pp.24-31, 1979

Araki1994 :

T.Araki, S.Ikehara and H.Yokokawa

Using accent information to correctly select Japanese phrases made of strings of syllables.

Proc. ICSLP94, pp.2155-2158, 1994

Ito1992 :

A.Ito and S.Makino

Word pre-selection using a redundant hash addressing method for continuous speech recognition.

Proc. ICASSP92, pp.309-312, 1992

Ringer1996:

E.K.Ringer and J.F.Allen

Error correction via a post-processor for continuous speech recognition.

Proc. ICASSP96, pp.427-430, 1996

Young1996 :

S.Young, J.Jansen, J.Odell, D.Ollason and P.Woodland

The HTK book manual , 1996

Mori1990 :

T.Mori and T.Yoshida

Data analysis technical book, 北王子書房, pp.181-183

Morimoto1994 :

T.Morimoto

A speech and language database for speech translation research.

Proc. ICSLP94, pp.1791-1794, 1994

Wooters1994 :

C.Wooters and A.Stolcke

Multiple-pronunciation lexical modeling in a speaker independent speech understanding system.

Proc. ICSLP94, pp.1363-1366, 1994

Imai1995 :

T.Imai, A.Ando and E.Miyasaka

A new method for automatic generation of speaker-dependent phonological rules.

Proc. ICASSP95, pp.864-867, 1995

Sloboda1995 :

T.Sloboda

Dictionary learning: performance through consistency.

Proc. ICASSP95, pages453-456, 1995

Humphries1996 :

J.Humphries, P.Woodland, and D.Pearce

Using accent-specific pronunciation modeling for robust speech recognition.

Proc. ICSLP96, pp.2324-2327, 1996

Neumeyer1996 :

L.Neumeyer, H.Franco, H.Weintraub and P.Price

Automatic text-independent pronunciation scoring of foreign language student speech.

Proc. ICSLP96, pp.1457-1460, 1996

Fukada1997 :

T.Fukuda and S.Sagisaka

Automatic generation of a pronunciation dictionary based on pronunciation networks.

Proc. EuroSpeech97, pp.2471-2474, 1997

Ravishanker1997 :

M.Ravishanker and M.Eskenazi

Automatic generation of context-dependent pronunciation.

Proc. EuroSpeech97, pp.2467-2470, 1997

第3章

3 - 1 節

Bahl1983 :

L.R.Bahl, F.Jelinek, and R.L.Mercer

A Maximum likelihood to continuous speech recognition.

IEEE Trans. on Pattern Analysis and Machine Intelligence, pp.179-199, 1983

Saitou1988 :

H.Saitou and M.Tomita

Parsing noisy sentences

Proc. CORING88, pp.561-566, 1988

Mellish1989 :

C.S.Mellish

Some chart-based techniques for parsing ill-formed input.

Proc. the Annual Meeting of the ACL, pp.102-109, 1989

Lavie1996 :

A.Lavie, D.Gates, M.Gavalda, L.Mayfield, A.Waibel and L.Levin

Multilingual translation of spontaneously spoken language in a limited domain.

Proc. 16th ICCL, pp. 442-447, 1996

竹沢 1996 :

竹沢寿幸, 森元逞

部分木に基づく構文規則と前終端記号バイグラムを併用する対話音声認識手法

電子通信学会論文誌 (D-II) vol.J79-D-II, No.12, pp.2078-2085, 1996

古瀬 1994 :

古瀬蔵 隅田英一郎, 飯田仁

経験的知識を活用する変換主導型機械翻訳

情報処理学会論文誌(D-II) , Vol.35, no.3, pp.412-417, 1994

隅田 1994 :

隅田英一郎, 古瀬蔵, 飯田仁

英語前置詞句係り先の用例主導あいまい性解消

電子通信学会論文誌 (D-II) Vol.J77-D-II, No.3, pp.557-565, 1994

Furuse1996 :

O.Furuse and H.Iida

Incremental translation utilizing constituent boundary patterns.

Proc. COLING96, pp.412-417, 1996

Shimizu1996 :

T.Shimizu, H. Yamamoto, H.Masataki, S.Matsunaga and Y.Sagisaka

Spontaneous dialogue speech recognition using cross-word context constrained word graphs.

Proc. ICASSP96, pp.145-148, 1996

3 - 2 節

隅田 1994 :

隅田英一郎, 古瀬蔵, 飯田仁

英語前置詞句係り先の用例主導あいまい性解消

電子通信学会論文誌 (D-II) Vol.J77-D-II, No.3, pp.557-565, 1994

古瀬 1994 :

古瀬蔵 隅田英一郎, 飯田仁

経験的知識を活用する変換主導型機械翻訳

情報処理学会論文誌(D-II) , Vol.35, No.3, pp.412-417, 1994

Ishikawa1998 :

K.Ishikawa, E.Sumita, and H.Iida :

Example-based error recovery method for speech translation Repairing sub-trees according to the semantic distance.

Proc. ICSLP98, pp.1147-1150, 1998

Nagao1984 :

M.Nagao

A Framework of a Mechanical Translation between Japanese and English by Analogy Principle.
Artificial and Human Intelligence Elithorn & Banerji, Eds., Elsevier Science Publishers,
pp.173-180, 1984

佐藤 1991 :

佐藤理史

実例に基づく訳語選択

人工知能学会誌、Vol.6, No.4, pp.592-600, 1991

Furuse1994 :

O.Furuse and H.Iida

Constituent Boundary Parsing for Example-based Machine Translation,
Proc. Coring94, pp.105-111, 1994

中野 1996 :

中野洋

「分類語彙表」形式による語彙分類表（増強版）

一謝辞一

本研究を遂行しまとめるにあたり、多くの方々の御理解、御指導、御鞭撻、御協力を頂きました。ここに感謝の意を表します。

本研究をまとめるにあたり、懇切な御指導と御鞭撻を賜った神戸大学大学院自然科学研究科山本誠一教授に心より感謝致します。山本誠一教授は著者が(株)ATR音声翻訳通信研究所に在籍中より、同研究所の社長の立場から研究遂行の機会を与えて頂き、御支援頂きました。重ねて心より感謝致します。

また、本研究テーマの設定、遂行時から本論文をまとめるに至るまで、懇切な御指導と御鞭撻を賜った早稲田大学国際情報通信研究科匂坂芳典教授（神戸大学大学院自然科学研究科元教授）に心より感謝致します。匂坂教授は、著者が(株)ATR音声翻訳通信研究所に在籍中、同研究所第一研究室の室長を兼務されており、研究遂行にあたり終始御指導を頂きました。重ねて心より感謝致します。

本論文の 2.2 節、3.1 節は、著者が(株)ATR音声翻訳通信研究所に在籍中に行った研究成果をまとめたものです。その間、多くの方々にお世話になりました。著者が所属していた第三研究室の飯田仁室長（現(株)ソニーコンピュータサイエンス研究所）には、言語翻訳処理について丁寧に御指導頂き、数々の有益なご教示を賜りました。心より深謝致します。また、研究の機会を与えて頂き、御支援頂いた(株)ATR音声翻訳通信研究所当時社長山崎泰弘博士（現東海大学工学部電気工学科教授）に深謝致します。

(株)ATR音声翻訳通信研究所第一研究室に当時在籍されていた Harald Singer 博士（現 Speech Works International Inc.）には、2.2 節に関する研究の共同研究者として御協力頂き、評価実験に用いた音素タイプライターに関するソフトウェアを提供して頂きました。また、(株)ATR音声翻訳通信研究所第三研究室に当時在籍されていた河井淳研究員（現(株)東洋情報システム）には、3.1 節に関する研究の共同研究者として御協力頂き、さらに研究成果を導入した音声翻訳システムを開発頂きました。両氏に心より感謝致します。また ATR 人間情報科学研究所加藤宏明博士は神戸大学大学院自然科学研究科を卒業された先輩であり、本論文をまとめる過程で御指導頂きました。深謝致します。その他、(株)ATR音声翻訳通信研究所第三研究室隅田英一

郎博士（現 ATR 音声言語通信研究所）、古瀬蔵氏（現 N T T サイバーソリューション研究所）を始めとする A T R の研究員の方々には、研究内容について数多くの御指導を頂き、また熱心に御討論を頂きました。ここに深く感謝致します。

松下電器産業（株）中央研究所（現先端技術研究所）の当時所長新田恒治博士、同研究所平井誠博士（現大阪市立大学助教授）には本研究の機会と(株)A T R 音声翻訳通信研究所への出向の機会を与えて頂きました。深謝いたします。また、松下電器産業（株）先端技術研究所の当時所長山崎攻博士及び同研究所情報科学グループ当時グループマネージャー二矢田勝行博士には、本研究の機会を与えて頂き、本論文をまとめるにあたり御便宜を賜りました。また、松下電器産業（株）先端技術研究所安立正明所長、同研究所ヒューマンウェア研究所丸野進所長には、本研究に御理解を頂き、本論文をまとめる機会を与えて頂きました。深謝いたします。中央研究所に当時在籍されていた坪香英一博士には、著者の上司としてまた 2.1 節に関する共同研究者として、有益な御指導と御鞭撻を頂きました。先端技術研究所ヒューマンウェア研究所ヒューマンウェアグループ松井謙二グループマネージャーには、現在の著者の上司としてまた 3.2 節の共同研究者として、研究の方向性について有益な御指導を頂き、本論文をまとめるにあたり御協力と御鞭撻を頂きました。両氏に対しここに深く感謝致します。また、中央研究所に当時在籍中の中橋順一氏には、2.1 節の評価実験で用いた音声認識ソフトウェアを著者ととも開発頂き、研究遂行に御協力頂きました。先端技術研究所廣瀬良文氏には、3.2 節の共同研究者として御協力を頂き、本研究成果を導入したシステムを著者ととも開発頂きその評価実験実施に協力頂きました。先端技術研究所遠藤充氏及び小沼知浩氏には、3.2 節における評価実験のための音声認識結果を提供頂きました。ここに深謝致します。その他、先端技術研究所音声処理担当の研究者の方々には、様々な御協力と御支援を頂きました。ここに深く感謝致します。

－ 発表文献 －

論文

滝沢(現在脇田)由実、坪香英一：“音節継続時間予測法を用いた不特定話者連続音声認識” 電子情報通信学会論文誌 A, Vol. J77-A, No.2, pp.173-181 (1994.2)

脇田由実、Harald Singer、匂坂芳典：“複数音素にまたがる誤認識特性を用いた音素候補系列追加モデル” 電子情報通信学会論文誌 D, J79-D, No.12, pp.2086-2095 (1996-12)

脇田由実、河井 淳、飯田 仁：“意味的類似性を用いた音声認識正解部分の特定法と正解部分のみ翻訳する音声翻訳手法” 言語処理学会自然言語処理論文誌, Vol.5, No.4, pp.111-125 (1998.10)

Yumi Wakita, Harald Singer, Yoshinori Sagisaka：“Multiple Pronunciation Dictionary using HMM-state Confusion Characteristics” Computer Speech and Language, Vol.13, No.2, pp.143-153(1999.4)

脇田由実、廣瀬良文、匂坂芳典、山本誠一：“キーワード主導型類似用例検索を用いた認識誤りに頑健な音声翻訳手法”（電子通信学会論文誌に投稿中）

国際学会（査読あり）

Yumi Wakita, Eiichi Tsuboka : “ Syllable Duration Prediction for Speech Recognition ” International Conference of Speech Language Processing '92, pp.1371-1374 (1992.10)

Yumi Wakita, Eiichi Tsuboka : “State Duration Constraint using Syllable Duration for Speech Recognition” International Conference of Speech Language Processing '94, pp.195-198 (1994.9)

Yumi Wakita, Harald Singer, Yoshinori Sagisaka : “Phoneme Candidate Re-entry Modeling using Recognition Error Characteristics over Multiple HMM States” ESCA Workshop on Spoken Dialogue Systems, pp. 73-76 (1995.6)

Yumi Wakita, Jun Kawai, Hitoshi Iida : “Correct Parts Extraction from Speech Recognition Results using Semantic Distance Calculation, and its Application to Speech Translation” Spoken Language Translation Workshop in Conjunction with ACL '97/ESCA '97, pp.24-31 (1997.7)

Yumi Wakita, Harald Singer, Yoshinori Sagisaka : “Speech Recognition using HMM-State Confusion Characteristics” 5th European Conference of Speech Communication and Technology, Vol.1 pp.7-10 (1997.9)

Hideki Mima, Osamu Furuse, Yumi Wakita, Hitoshi Iida : “Multi-ligual Spoken Dialog Translation System using Transfer-Driven Machine Translation” Machine Translation Summit VI, pp.148-155 (1997)

Yumi Wakita, Kenji Matsui, Yoshinori Sagisaka : “Robust speech translation using fine keyword clustering” Workshop on Multi-Lingual Speech Communication, (2000.10)

Yumi Wakita, Kenji Matsui, Yoshinori Sagisaka : “Fine keyword clustering using a thesaurus and example sentences for speech translation.” International Conference of Speech Language Processing 2000, pp.390-393 (2000.10)

Chengqing Zong, Yumi Wakita, Bo Xu, Kenji Matsui, Zheebiao Chen : “Japanese – to – Chinese spoken language translation based on the simple expression.” International Conference of Speech Language Processing 2000, pp. 418-421 (2000.10)

Kenji Matsui, Yumi Wakita, Tomohiro Konuma, Mitsuru Endo, Masashi Murata : “An experimental multilingual speech translation system.” International Workshops on Perceptual User Interfaces 2001 (2001.11)

Tomohiro Konuma, Kenji Matsui, Yumi Wakita, Kenji Mizutani, Mitsuru Endo, Masashi Murata : “An experimental multilingual bi-directional speech translation system” TMI2002: The 9th conference on Theoretical and Methodological Issues in Machine Translation (2002.3)

参考国際学会（査読あり）

Masahiro Hamada, Yumi Takizawa, Takeshi Norimatsu : “A Noise Robust Speech Recognition System” International Conference of Speech Language Processing '90, pp.893-896 (1990.11)

Yumi Wakita, Masahiro Hamada : “Lombard Speech Recognition by Formant-Frequency-Shifted LPC Cepstrum” International Conference of Speech Language Processing '90, pp.293–296 (1990.11)

Jun Kawai, Yumi Wakita, Hitoshi Iida : “Stochastic Language Model using Semantic Recognition Error Characteristics over Multiple HMM States” Natural Language Processing Pacific Rim Symposium '95, pp.107 - 110 (1995.12)

Yumi Wakita, Jun Kawai, Hitoshi Iida : “An Evaluation of Statistical Language Modeling for Speech Recognition using Mixed Category of Both Words and Parts-of-speech” International Conference of Speech Language Processing '96, pp.293–296 (1996.10)

研究会（査読なし）

脇田由実、Harald Singer、匂坂芳典：“複数音素にわたるHMMの誤認識特性を用いた語彙候補の追加” 電子情報通信学会音声研究会 SP95-30 pp.41-47 (1995.06)

脇田由実、河井淳、飯田仁：“意味的類似性を用いた後処理的な音声認識正解部分特定法と音声翻訳手法への導入” 情報処理学会研究報告資料 IPSJ SIG Notes, 97-SLP-17 vol.97, No.66, pp.19-26 (1997)

廣瀬良文、脇田由実、小沼知浩、西崎誠、遠藤充：“音声翻訳のための認識誤りにロバストな類似用例選択手法の評価” 情報処理学会研究報告資料 2001-SLP-36 Vol.2001, No.55, pp.43-48 (2001.06)

口頭発表（査読なし）

滝沢（現在脇田）、中橋、堀岡、坪香：“音節継続時間予測の連続音声認識における有効性” 音講論集 3-1-11 pp.91,92(1992.3)

滝沢（現在脇田）、坪香：“連続音声における近傍音節を考慮した音節継続時間予測法” 音講論集 2-Q-22 pp.215-216(1992.10)

滝沢（現在脇田）、坪香：“音節を認識単位とした単語認識における音節継続時間制御法” 音講論集 2-4-2 pp.21-22 (1993-3)

滝沢（現在脇田）、坪香：“音節継続時間を考慮した状態継続時間制御 HMM” 音講論集 2-P-20 pp.197-198 (1994-3)

脇田、Singer、匂坂：“HMM のモデル間にまたがる状態系列を利用した誤り訂正法の検討” 音講論集 3-8-2 pp.111-112 (1994-10)

脇田、Harald、匂坂：“話者適応が誤認識特性に及ぼす影響について” 音講論集 2-5-7 pp.43-44 (1995-3)

脇田、Harald、匂坂：“単語 bi-gram を用いた連続音声認識への状態系列の誤認識特性” 音講論集 1-6-6 pp.11-12 (1997-3)

工藤、脇田：“構文情報と意味情報を用いた統計的言語モデルの検討” 音講論集 1-1-25 pp.49-50 (1999-3)

脇田、松井、匂坂：“単語の依存関係を用いた単語クラスの自動決定法” 音講論集 2-8-13 pp.53-54 (2000-3)

脇田、松井：“要約型音声翻訳における翻訳文の品質について” 情報処理全国大会論文集 2L-5 pp.139-140 (2000.3)