TR-IT-0249

HMMの状態系列で表された誤認識特性を用いた正解音素、正解語彙探索手法について

脇田 由実　匂坂 芳典

Yumi WAKITA　Yoshinori SAGISAKA

1997 年 12 月

概要

従来の音響モデルと認識手段では認識が困難な語彙を獲得するために、下記の特徴を持つ誤認識特性を用いて、従来認識できなかった正解音素または正解語彙を探索する手法を提案し、その探索性能を評価した。誤認識特性は次の特徴を持つ。

- 誤認識特性がHMMの状態系列として表現されている。
- 認識結果の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誤認識特性を抽出する。

本報告書では、まず、上記の誤認識特性の抽出法および誤認識特性を利用した次の２つの探索方法について説明する。(a) 認識結果から正解音素探索行なう方法、(b) 発音辞書作成して正解語彙を探索する方法。

次に探索性能の評価結果を報告する。評価結果は、上記 (a),(b) ２つのどちらの探索方法においても、上記に示した本誤認識特性の特徴が、従来認識できなかった音素系列や語彙を探索するのに優位であることを示している。さらに、誤認識特性の品詞依存性を考慮しながらマルチ発音辞書を作成することが、より探索性能を向上させることも示している。

エイ・ティ・アール音声翻訳通信研究所

ATR Interpreting Telecommunications Research Laboratories

目次

1 はじめに 1
1.1 本研究所での研究の位置付け 1
1.2 音声認識における研究の背景 1

2 現状の音響モデルによる誤認識の分析 3

3 誤認識特性を用いた正解音素、正解語彙の探索 6
3.1 誤認識特性の抽出について 6
3.1.1 誤認識特性抽出法 6
3.1.2 誤認識特性抽出法の評価 7
状態表現と音素表現との誤認識特性記述力の比較 8
シンボルＤＰによる抽出法との比較 9
3.2 誤認識結果からの正解音素系列探索 10
3.2.1 探索手法 10
3.2.2 本探索手法の認識率改善に対する効果 12
3.2.3 学習データ量と探索性能の関係 12
3.2.4 従来の誤認識特性抽出法との探索性能比較 13
状態表現と音素表現との誤認識特性記述力の比較 13
シンボルＤＰによる抽出法との比較 13
3.3 マルチ発音辞書化による正解語彙探索 14
3.3.1 誤認識特性の品詞依存性について 14
3.3.2 誤認識特性の品詞依存性を考慮したマルチ発音辞書作成 16
3.3.3 本マルチ発音辞書による正解語彙探索の評価 16
状態表現による誤認識特性を用いることの効果 17
品詞依存性を考慮した誤認識特性の効果 17
3.3.4 誤認識特性の学習用データ量と単語認識率との関係 17

4 おわりに 19

5 著者の関連発表一覧 21

6 ツール関連 22

第 1 章

はじめに

1.1 本研究所での研究の位置付け

高性能な音声翻訳システムを構築するためには、その基本システムである音声認識部、言語翻訳部、音声合成部などが高性能であることが必要条件である。本研究所では、従来、まずは各基本システムの性能を向上させることに重みを置き、研究開発を進めてきた。たとえば音声認識については、HMMを基本とした音響モデルの構築、話し言葉を意識した統計的言語モデルの構築を行ない、それらの性能を音素認識率や単語認識率で評価してきた。言語翻訳については、文法的に記述が困難な話し言葉の解析と翻訳を可能にするために、用例から多くの話し言葉の言い回しなどを学習し、それを用いた変換主導型の翻訳システムを構築し、翻訳文の正確さや自然性を評価してきた。次のステップとして、音声翻訳システムを実現するために、さらに各システムが互いの規則、知識を活かしながら、またはお互いの欠点を補いながら処理をするしくみを導入し、各システムを統合した完成度の高い音声翻訳システムを構築する必要があると考える。この観点から、私は、以下のしくみを新たに音声翻訳システムに導入することが必要だと考えている。

音声認識部については、これを言語翻訳部への入力文作成部と考えると、必ずしも認識結果の音素認識率や単語認識率が100%である必要はなく、むしろ次の条件を満たすことが重要である。

(1) 言語翻訳に必要な言葉が認識されていること

(2) 誤訳を招くような単語に誤認識しないこと

また言語翻訳部は、音声認識が度々認識誤りを起こす現実を考慮し、誤りを含んだ文の処理が必要になる。具体的には、

(3) 誤認識単語を含んだ認識結果文の解析が可能であること

(4) 誤認識結果に対応した誤訳をしないこと

が重要である。

上記4つの中の(1)(2)を満たす音声認識手法を確立するための取り組みとして、音響的な誤認識特性を用いて、従来の音響モデルでは認識が困難な音素または語彙を探索する手法を研究してきた。本論ではこの研究成果を報告する。

1.2 音声認識における研究の背景

従来音声認識では、なるべく完璧な音声認識率をめざして、音響モデルの改良に向けての研究がさかんに行なわれている。たとえば、誤認識の大きな原因となる音素環境コンテキスト特有の発声変形を吸収するために、コンテキスト依存型のHMMが提案され、有効性も確認されている[1; 2; 3]。また、話者特有の発声変形を吸収するために、様々な話者適応法も提案されてきた[4; 5; 6]。しかし、音素環境コンテキスト依存型HMMや話者適応後の音響モデルを用いた認識でも、なおコンテキストに依存した誤認

識、話者に依存した誤認識は多い。また、このような音響モデルの不完全をカバーするために、多くの音声認識システムでは、認識中の探索ビーム幅を増やしたり、最終的に上位N-best文を候補として出力することで、なるべく多くの正解音素または正解単語を出力するしくみを導入している。しかし、探索のビーム幅をいくら増やしても、正解が結果候補に入ってこない場合がある。この原因は次のように考えられる。

第1に考えられる原因として、コンテキスト依存型のHMMが考慮しているコンテキストは音素環境だけであり、話者適応法が考慮しているのは話者の違いだけである。発声速度、発声レベルといったその他の要因が考慮されていない。誤認識を軽減するためには、発声変形の要因を分析し、各要因ごとに様々な環境のデータを準備し、これを用いてHMMを作成する必要があるが、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、発話条件に応じてHMMの条件を最適に設計することの困難さが考えられる。通常の連続認識系では、予め、発声中の小さな単位（たとえば音素、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。発声変形の状況が変化しても、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と、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1つのHMMに対応するような不自然なマッチングが生じ得る。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[7]が必要であるが、全学習データ量、学習に要する時間等を考えると、さらに工夫が必要である。

音素または単語認識率を向上させるために、従来音響モデル上で解決しようとしていたコンテキストや発話条件の違いに起因する問題を、全て、音響モデル単独で解決をはかるのは得策ではないと考える。しかも、言語翻訳のために一部の必要な言葉を優先的に正しく認識させることを考えると、音響モデル作成時ではなく、既に与えられた音響モデルの不備を認識中または認識後に積極的に救済する手法が是非必要であると考える。そこで私は、音響的な誤認識特性を用いて、単独のHMMラベルとしては正解候補に入ってこない音素系列または単語系列を新たに音素または単語候補系列として追加し、従来認識できなかった音素または単語系列を認識結果として出力することを可能にするメカニズムを確立した。

本論では、従来の音響モデルと認識手段では解決が困難な誤認識の特性を分析し、この分析結果に基づいた誤認識特性抽出法について提案し、さらに抽出した誤認識特性を用いて従来認識できなかった音素や語彙を探索する方法について提案し、その評価結果を報告する。まず2章で、音響的な誤認識特性についての分析結果を報告し、3章で、誤認識特性の抽出法、正解音素または正解語彙探索方法、評価結果を順次報告する。最後に4章で、まとめと今後の課題について述べる。

第 2 章

現状の音響モデルによる誤認識の分析

音響モデルを改良したり、ビーム幅を大きくしても、なお残る誤認識に対応できる音素または単語候補追加モデルを提案するために、現在の音響モデルが抱える誤認識の特徴を調べ、正解音素または正解語彙探索モデルが解決するべき問題点を明らかにした。前章に述べたように、現行の音響モデルの検討では、コンテキストの違いや話者の違いによる認識性能の変化を吸収するための検討がなされているのでコンテキスト依存型の音響モデル用い、さらに音響モデルを改良する手段として話者適応を取り上げ、話者の違いによる誤認識に限って調べることにした。

まず、現行の音響モデルがビーム幅を増やすことにより救済できる誤認識の範囲を見定めるため、ビーム幅と音素認識率との関係を調べた。さらに話者適応の効果も併せて調べるために、不特定話者用ＨＭＭを話者適応した後の誤認識特徴を、話者適応前の誤認識特徴と比較した。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）[6; 8]を採用し、これにコンテキスト環境を考慮した平滑化を加え、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 2.1に示す。音響モデルに直接起因する誤認識特性を調べるために、極力、言語的な制約を用いない認識系で比較した。ここでは、日本語の音素の隣接についての制約（たとえば、子音どうしの隣接を許さない など）のみの言語制約下で one-pass DP 法で音素認識を行なう音素タイプライター型の認識システムを用いた[3; 9]。分析データとして、各話者のデータから、話者適応に用いた文と内容が同じで発声が異なる文（以下「既知文」と呼ぶ）33 文を選択した。

図 2.1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（適応前で 2000、適応後で 1000）以上になると、認識率は飽和している。

次に、話者適応前後の誤認識特性を比較した。まず誤認識特性を正確に抽出するため、 Viterbi-alignment を用いて調べた正解音素系列の各音素のマッチング区間と認識結果の各音素のマッチング区間とを比較することで、誤認識部分を決定した。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

表 2.1: 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 HM-net、延べ状態数 201、混合分布数 10 継続時間制御なし
認識法	One-pass DP 法による N-best 候補探索
構造決定	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
分析用データ	33 文 (637 音素) ／ 話者
話者	3 男声、3 女声

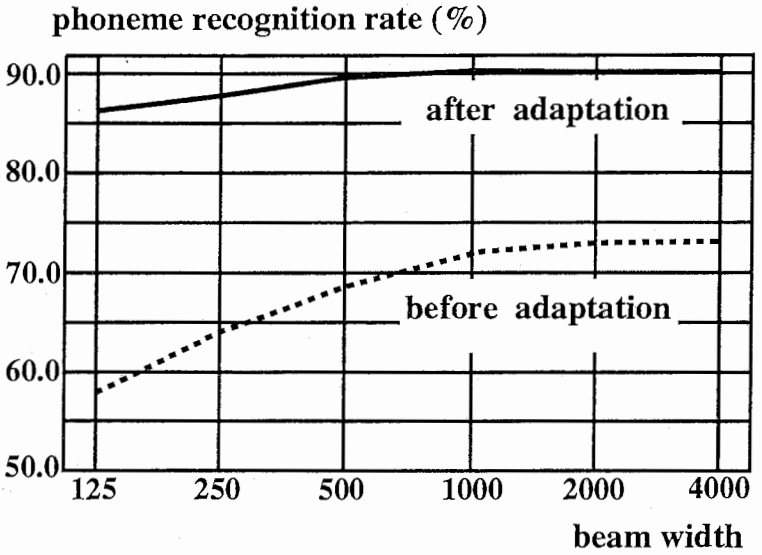


図 2.1: ビーム幅と音素認識率との関係

付け、印に挟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の音素系列の種類が異なっている場合、その区間全体を誤認識部分とみなし、正解と認識結果との音素系列ペアを抽出した。出現数が多いペアの上位 6 位まで（括弧内は、総出現数）を表 2.2に示す。話者適応前には、/u/ → /e/ や /o/ → /u/ や /t/ → /k/ のように、単音素間の誤認識特性の頻度が上位をしめているが、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誤認識特性の頻度が相対的に高くなることがわかる。

さらに、全誤認識の中の、複数音素にまたがる誤認識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 2.3が示すように、話者適応により複数音素にまたがる誤認識特性の絶対数（ 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誤認識特性の割合は増加している。

これらの複数音素にまたがる誤認識系列は，表 2.2の例からもわかるように、誤認識系列の音素数と正解系列の音素数とが 1 対多になっているものが多い。従って、これらの誤認識における各音素の継続時間が不自然であり、継続時間の制御を導入することによりこれらの誤認識が解決されることも考えられる。継続時間を制御することで複数音素にまたがる誤認識がどれだけ改善されるかを把握するために，以下に示す方法で、複数音素にまたがる誤認識特性の各音素のマッチング区間長が不自然でないかどうかを調べた。予め学習データを用いて、各音素または状態毎に、継続長の最大値と最小値を調べておき、表 2.2に示した複数音素にまたがる誤認識特性における各音素及び各状態のマッチング区間長が、同じ種類の音素または状態の継続長の最大値と最小値の間に収まっているかどうかを確認した。その結果、誤認識特性（/o g o/ → /o/）の /o/ について 2 例、（/i g i/ → /i/）の /i/ について 1 例が、それぞれ学習データの /o/,/i/ の最大長を上回っていたが、残りの音素及び状態のマッチング区間長は、学習データにおける最大値と最小値との範囲内に収まっていた。この結果が示すように，表 2.2に示す誤認識音素系列の各音素のマッチング区間長は、その音素の種類から判断した範囲では不自然なものではなく，各音素または状態毎にマッチング区間に制限を与えながら認識を行ったとしても，表 2.2、表 2.3における誤認識特性の傾向は大きく変わらないと思われる。

以上の分析結果が示すように、話者適応により誤認識は減少するが、適応に用いた文と同じ内容の文を認識した場合でも、誤認識の修復には限界がある。特に、各音素が 1 対 1 対応した誤認識に対する有効性に比べ、複数音素にまたがる誤認識に対してはさらに対策が必要であることがわかる。

表 2.2: 出現頻度が多い誤認識（上位 6 位）
(/ 正解系列 / → / 誤認識系列 / : 括弧内は、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 2.3: 複数音素にまたがる誤認識系列の全誤認識に対する割合

(括弧内は、複数音素にわたる誤認識系列数／全誤認識系列数)

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

第 3 章

誤認識特性を用いた正解音素、正解語彙の探索

前章で調べた誤認識の救済方法として、予め誤認識の特性を学習しておき、それを用いて認識中または認識後に認識結果候補から結果候補には含まれない新たな音素または語彙候補を探索する方法を提案する。

この方法は、学習データを用いて誤認識特性を抽出する部分と、学習時に抽出された誤認識特性を用いて正解音素または正解語彙を探索する部分とから成り立つ。

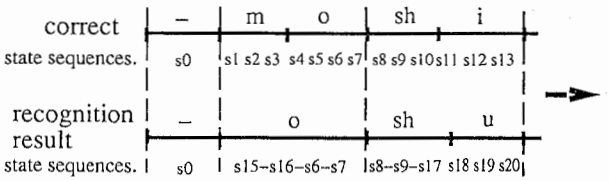
また、正解音素、正解語彙を探索する手段には少なくとも2通りあると考えられる。1つは、認識後に認識結果のラベル系列上で正解音素または語彙の探索を行なう方法である。もう1つは、誤認識特性を用いてマルチ発音辞書を作成し、それを用いて認識することで、認識時に正解語彙の探索を行なう方法である。

以下に、まず誤認識特性の抽出について述べた後、2つの探索方法を順次述べる。

3.1 誤認識特性の抽出について

3.1.1 誤認識特性抽出法

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



error state sequences table

pair no.	correct state seq. : error state seq.	apperance rate
1.	s1–s2–s3– s4–s5–s6–s7 : s15–s16–s6–s7	60%
2.	s1–s2–s3– s4–s5–s6–s7 : s1–s2–s3– s4–s5–s6–s7	40%
3.	s8–s9–s10– s11–s12–s13 : s8–s9–s17–s18–s19–s20	30%
4.	s8–s9–s10– s11–s12–s13 : s8–s9–s10– s11–s12–s13	70%

図 3.1: 誤認識特性抽出例 (/moshi/ → /oshu/)

表 3.1: 本誤認識特性抽出法で抽出された出現確率の高い誤認識特性

(/ 正解系列 / → / 誤認識系列 / : 括弧内は、出現確率)

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%)

特有の発声変形は、同じ誤認識音素系列を頻繁に出力する可能性がある。従って、この誤認識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。

誤認識の傾向を把握するために、誤認識した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、誤認識特性を用いて、認識誤りを訂正する研究がなされており[10; 11]、これらの研究における誤認識部分の抽出法は、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1対1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、誤認識系列と正解系列とのペアを抽出している。しかし、2章で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、正解の音素と誤認識ラベルとが1対1対応しているとは限らない。複数HMMで成り立つ誤認識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誤認識特性を抽出するために、1つの誤認識部分含まれるHMMラベル数に制限を与えず、複数の誤認識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誤認識に反映させるには、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。そこで誤認識を、音素ではなくHMMの状態系列として表現した。

誤認識特性は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誤認識部分を決める。この方法は、前章で誤認識部分を抽出した場合と同じであるが、違いは前章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/ と誤認識した場合の抽出例を図3.1に示す。図3.1では、正解状態系列 $\{s_0, s_1, \cdots, s_{13}\}$ と認識結果状態系列 $\{s_0, s_{15}, \cdots, s_{20}\}$ とを比較し、両者の音素系列の境界点が同じになるところ（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、誤認識状態系列と正解状態系列のペアを抽出した後、各誤認識状態系列の信頼度を表すものとして、誤認識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。ある誤認識状態系列 $\{E_i = e_1, e_2, e_3, \cdots, e_n\}$ と正解状態系列 $\{C_i = c_1, c_2, c_3, \cdots, c_m\}$ (e,c は各状態）のペア (E_i, C_i) の出現確率 $\mathrm{P}(E_i, C_i)$ は、次式で定義する。

$$P(E_i, C_i) = \frac{\text{ペア}(E_i, C_i)\text{の出現数}}{\text{認識結果上の系列}\{E_i\}\text{の出現数}} \tag{3.1}$$

認識結果の状態系列 $\{E_i\}$ が正解であった場合は、ペア (E_i, E_i) として上記ペア (E_i, C_i) 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3.1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のペアの誤認識状態系列が正解であった場合として登録されている。

2章で話者適応に用いた3 3文を用いて、実際に話者6名分の誤認識特性を抽出した。抽出された誤認識特性の傾向は2章の表2.2に示した話者適応後の音素表現による誤認識特性の傾向と似ており，出現確率の高い誤認識特性は，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，各話者の誤認識特性の約半分は各話者特有の誤認識特性であった．たとえば表3.1は，6話者のうちの2名（男女1名ずつ）について、出現確率の高かった上位5位の誤認識特性例を示している．実際には誤認識は状態系列で表現されているが，傾向をわかりやすくするために表3.1では音素系列で表現する．この2名については，誤認識特性　/o g o/ → /o/ は両者に共通しているが，他の誤認識特性は各話者特有のものである。

3.1.2 誤認識特性抽出法の評価

上記の誤認識特性抽出法が、正解音素または正解語彙の探索において有効であることを予測するために、音素認識結果と抽出された誤認識特性とを比較して、誤認識系列と同系列が結果に含まれていた場合に、その部分を正解系列に置き換えたもの結果候補に追加するしくみを構築し、追加された音素系列に含まれる正解音素の割合を調べた。さらに、従来の抽出法を用いた場合や N-best 法による複数候補

表 3.2: 評価実験データ

誤認識特性抽出用	32 文 (754 音素) ／ 話者
評価用（既知文)	抽出用データと同文章、異発声 33 文 (637 音素) ／話者
評価用（未知文）	抽出用データと異文章、異発声 18 文 (629 音素) ／話者
話者	3 男声、3 女声

における正解音素の割合と比較し、誤認識特性を用いることの優位性も確認した。実験条件は 2.1と同じであるが評価データは，前章で用いた既知文（学習データと，文内容は同じ発声は異なる）と、文内容も発声も抽出用データとは異なる文（以降未知文という） 18 文とを用い，既知文と未知文とを別々に評価した。予備実験により、N-best 法では、正解音素含有率は N=13 で限界に達していたので、以下の N-best 法の実験では N=13 に設定した。

本抽出法の特徴は次の 2 つである。

1. 認識結果の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誤認識特性を抽出する。

2. 誤認識特性がHMMの状態系列として表現されている。

従来の誤認識特性抽出法では、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ない、なるべく音素と音素、音節と音節を 1 対 1 対応させながら、誤認識を抽出していた。本モデルの各々の特徴の有効性を調べるため、特徴毎に従来の抽出法との比較を行なった。

状態表現と音素表現との誤認識特性記述力の比較

本モデルで、誤認識特性を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比較した。ここでは次の 3 つの表現を比較した。

(a) 誤認識系列を状態系列として表現。（状態表現）

(b-1) 誤認識系列を音素系列として表現。（音素表現）

(b-2) 誤認識系列をその前後の 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[3]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、誤認識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。音素系列を追加する際には、(b-1) による音素表現の場合が、最も多くの音素系列が追加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。

結果を表 3.3に示す。表 3.3 には、下記により定義された正解音素含有率と、全候補の増加率を示している。

$$\text{正解音素含有率} = \frac{\text{全候補に含まれる正解音素数}}{\text{正解文に含まれる全音素数}}$$

$$\text{全候補増加率} = \frac{\text{全候補に含まれる音素数}}{\text{第 1 候補に含まれる音素数}}$$

もし、異なる候補文の同じ位置に同じ音素がある場合は、カウントされる音素数は 2 回ではなく、1 回とする。カウント例を下に示す。

表 3.3: 誤認識特性の表現の違いが正解音素含有率および候補増加率に及ぼす影響

		N-best (N=13)	状態表現 (a)	音素表現 (b-1)	拡張音素表現 (b-2)
正解音素含有率	既知文	77.9%	82.3%	84.9%	77.5%
	未知文	71.7%	72.7%	76.3%	69.4%
候補増加率	既知文	1.76%	1.50%	3.67%	1.47%
	未知文	1.52%	1.38%	3.04%	1.23%

```
correct       h a i
------------------          正解音素含有率 = 3 / 3
1st cand:     h a u         候補増加率 = 4 / 3
2nd cand:       a i
```

結果より次のことがわかる。

1. 状態表現を用いた場合、N-best 法に比べて、正解音素含有率は高く候補増加率は低い。本方法は N-best 法にに比べて、効率よく正解音素を含んでいる。

2. 音素表現を用いた場合には、N-best 法に比べ、正解音素含有率は高いが、候補増加率は約 2 倍になり、効率が悪い。

3. 拡張音素表現を用いた場合には、正解音素含有率が N-best より低い。

この結果より、状態表現を用いた誤認識特性は、N-best 法に比べ正解音素を追加するのに有効であることがわかった。状態表現の場合のみが、顕著に有効性を示したことから、誤認識特性の前後のコンテキストの音響的特徴を考慮した上で、正解音素候補の追加を行なうことが重要であると思われる。N-best 法においては、話者特有の発声がHMMで吸収されていない場合には、N-best 数を増やしても、その部分の正解音素系列は出力されない。上記の結果は、HMMの性能に限界がある場合に、本モデルを用いることで、N-Best 候補に入らなかった正解音素が出力できることを示している。

シンボルＤＰによる抽出法との比較

本モデルが音響的なマッチング区間の考慮して誤認識部分を特定していること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１対１に対応させながら誤認識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なうために、誤認識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いた。

音素と音素をなるべく１対１に対応させて誤認識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いた（たとえば文献[12]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を行なう。正解音素系列のｉ番目までの音素系列 $\{P_1^c, P_2^c, \cdots, P_i^c\}$ と認識結果音素系列のｊ番目までの音素系列 $\{P_1^r, P_2^r, \cdots, P_j^r\}$ 間のＤＰ距離 $D_{i,j}$ は次式のように算出される。

$$D_{i,j} = min \left\{ \begin{array}{c} D_{i-1,j} + 1.0 \times f_{i,j} \\ D_{i-1,j-1} + 1.1 \times f_{i,j} \\ D_{i,j-1} + 1.0 \times f_{i,j} \end{array} \right\} \tag{3.2}$$

$$f_{i,j} = \left\{ \begin{array}{ll} 0, & \text{if } P_i^c = P_j^r \\ 1, & \text{if } P_i^c \neq P_j^r \end{array} \right. \tag{3.3}$$

表 3.4: 音響的なマッチング区間を考慮した誤認識特性抽出法が正解含有率と候補数増加率に及ぼす影響

		N-best 法	シンボルＤＰ	音響マッチング
正解音素	既知文	77.9%	83.5%	84.9%
含有率 (%)	未知文	71.7%	72.5%	76.3%
候補数	既知文	1.76	2.95	3.67
候補数 (%)	未知文	1.52	2.66	3.04

このようなＤＰを用いた場合、認識結果の各音素は系列に従ってなるべく同音素とマッチングし、同種の音素がない場合には、異種の音素となるべく１対１に対応する。この場合、音素は置換誤りであるとして処理される。両者の音素系列に含まれる音素数が異なっており１対１対応できない場合には、余ったり不足した音素は、挿入または脱落誤りとして処理される。

誤認識を抽出する際には、上記で置換誤りと処理された音素は、正解音素と誤認識音素とが１対１に対応している誤認識特性として、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誤認識及び正解系列に含むように誤認識部分を決定する。たとえば、正解音素系列が /aiueo/ 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/、誤認識系列 /ie/ のペアがテーブルに登録される。同じ正解系列に対し誤認識系列が /aio/ の場合は、正解系列 /iueo/、誤認識系列 /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/ に対し誤認識系列が /ao/ であった場合には、正解系列 /iu/、誤認識系列 /o/ のペア誤認識特性に対応する。上記の処理を、以降、シンボルＤＰによる誤認識特性抽出法と呼ぶ。

結果を 3.4に示す。以下のことがわかる。

- 従来の N-best 法に比べて誤認識特性を用いた方が正解音素含有率は高く、従来認識できない音素を候補とすることができる。
- 既知文については、シンボルＤＢより音響的なマッチング区間を考慮した本方法が正解音素含有率は僅かに良いが、両者の差は少ない（ 83.5% と 84.9%）。むしろ、シンボルＤＢを用いた方が、より少ない候補数の増加で効率良く正解を含有している。
- 未知文については、音響マッチング区間を考慮した場合の正解音素含有率に対する効果は高く、シンボルＤＢとの正解音素含有率の差は大きかった。

3.2 誤認識結果からの正解音素系列探索

ここではまず、認識後に認識結果から正解音素系列を探索する手法について説明する。この方法は、認識結果の状態系列と誤認識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、その部分を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のである。この際に、誤認識を状態で表現することにより、追加登録には次の制約が与えられる。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いても、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、誤認識を状態系列で表現している場合は、誤認識と同じ音素系列が、認識結果の音素系列または言語モデルに記述されている語彙の中に存在しても、これらの全てが正解音素系列に置き換えられる訳ではなく、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。

3.2.1 探索手法

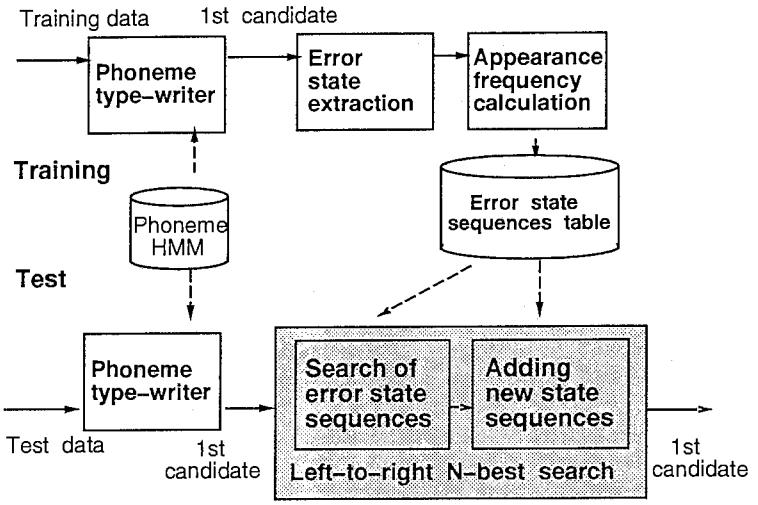


図 3.2: 音素タイプライターを用いた left-to-right の N-best 探索による正解音素系列探索

正解音素の探索をここでは、音素タイプライター型の認識システムを用いて実現している。音素タイプライターでは、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、認識処理を行なっている。誤認識特性と認識結果を用いて、新たな音素系列を結果候補として追加することは、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追加できる可能性があり、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。

前章の実験で、ビームの幅を大きくしてもなおビームから落ちてしまう正解音素のうち、既知文で約 20%、未知文で約 4% の音素が、本モデルで追加した新たな音素系列に含まれることを確認している。しかし、追加した音素系列が認識結果の第 1 候補として出力される保証はない。そこで、本モデルが認識率を向上させるのに有効であるかを評価するために、認識結果の第 1 候補の音素系列のみを対象とした正解率の向上をねらう。実際には、2 章で用いた音素タイプライター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しする。探索後に誤認識傾向の強い音素系列を優先的に書き換えた探索結果候補を第 1 候補として出力するために、認識時に算出された音響ゆう度を使用せず、誤認識特性の出現確率のみに基づいた N-best 探索を行なった。システム構成を図 3.2に示し、探索手順を次に説明する。

まず認識を行なう前の学習作業として、誤認識状態系列と正解状態系列とのペアを抽出し、ペアの出現確率とともにテーブルに登録しておく。これは 3 章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なった後、認識結果の第 1 候補の状態系列と誤認識特性とを比較し、誤認識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、その部分を誤認識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 に行なう。1 つの誤認識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誤認識状態系列長はまちまちで、お互いに一部が重なっていたり包含関係にある。そのため、同じ入力音素系列から、様々な誤認識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、誤認識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 探索を行なうことで、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誤認識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 $\{s_1, s_2, \cdots, s_n, E_i, \cdots\}$ 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 $\{C_i\}$ である確率が $\mathrm{P}(E_i, C_i)$ であるとする。この場合、状態系列 $\{s_1, s_2, \cdots, E_i\}$ 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 $\{s_1, s_2, \cdots, C_i\}$ に与えられるスコア $\mathrm{S}(s_1, s_2, \cdots, s_n, C_i)$

表 3.5: 正解音素系列探索後の音素誤認識率の減少

	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%

は、

$$\begin{aligned} & \mathrm{S}(s_1, s_2, \cdots, s_n, C_i) \\ = \quad & \mathrm{S}(s_1, s_2, \cdots, s_n) \times \mathrm{P}(E_i, C_i) \end{aligned} \tag{3.4}$$

となる。 $\mathrm{S}(s_1, s_2, \cdots, s_n)$ は、部分的な状態系列 $\{s_1, s_2, \cdots, s_n\}$ に与えられているスコアであり、状態系列 $\{s_1, s_2, \cdots, s_n\}$ の中の全ての状態が誤認識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、状態系列の長さに無関係に $\mathrm{S}(s_1, s_2, \cdots, s_n) = 1$ となる。一部が置き換えられている場合は、 $\mathrm{S}(s_1, s_2, \cdots, s_n)$ は式 (2) 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。

3.2.2 本探索手法の認識率改善に対する効果

提案している正解音素探索法の有効性を確認するために、上記探索結果の第 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 2.1と同じである。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誤認識が、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ＶＦＳにて話者適応を行なった後のモデルを用い、ビーム幅はこれは 2 章の実験結果図 2.1で認識率が既に飽和している際のビーム幅 4000 を用いた。評価データは，表 3.2を用い、既知文と未知文とを別々に評価した.

探索を行なう前後の話者 6 名の音素誤認識率を比較した。結果を表 3.5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%) 誤認識が減少した。本正解音素系列探索法が、評価データの種類に関係なく誤認識を減少させ、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、誤認識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 ）がおり、効果の度合いは話者により異なっていた。

3.2.3 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てみた。未知文データを表 3.2の条件に固定し、学習データに用いた文章数を徐々に増やして、正解音素系列の探索を行なった場合の音素認識率を図 3.3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 3.3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 3.5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でることが期待される。図 3.3に

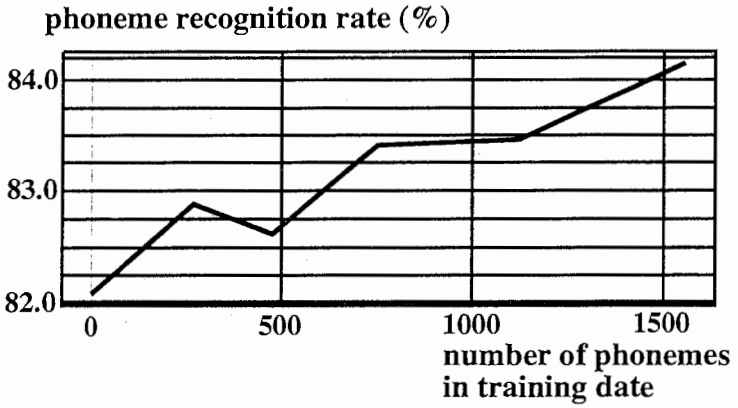


図 3.3: 誤認識特性の学習データ量と効果との関係

表 3.6: 誤認識系列の状態表現と音素表現との音素誤認識率の比較

	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%

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、誤認識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。

3.2.4 従来の誤認識特性抽出法との探索性能比較

状態表現と音素表現との誤認識特性記述力の比較

誤認識特性をHMMの状態で表現したことの優位性を確認するために、3.1 章で比較した場合と同様、状態表現 (a)、音素表現 (b-1)、拡張音素表現 (b-2) の各々における探索性能を比較した。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。

実験結果を表 3.6に示す。表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の状態の共有化が誤認識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。

シンボルＤＰによる抽出法との比較

音響的なマッチング区間を考慮して誤認識特性を抽出したこと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 1 対 1 に対応させながら誤認識特性を抽出する方法場合との探索性能を比較した。

表 3.7に、探索前後の音素誤認識率を示す。" マッチング区間考慮" の結果は、4.4.1. の実験結果の表 3.6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

表 3.7: 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ＤＰとの音素誤認識率の比較

	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%

表 3.8: 実験条件

タスク	旅行に関する会話
音響モデル	不特定話者 HM-net, 延べ状態数 401, 混合分布数 10, 継続時間制御なし
認識法	One-pass DP 法による N-best 候補探索
構造決定	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
評価用データ	109 文 (4854 音素) ／話者,
話者	3 女声

誤認識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１対１に対応させていることにより、実際の誤認識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、誤った探索を行ないながらも、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。

以上の結果をまとめると、次のようになる。

1. 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。

2. 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なうこと、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誤認識の抽出を行なうことは、特に誤認識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得る上で、必要条件である。

3.3 マルチ発音辞書化による正解語彙探索

ここでは、誤認識特性を用いて予めマルチ発音辞書を作成することでしておき、認識中に、従来認識できなかった語彙を探索する手法を説明する。これは、従来の各単語の状態系列とテーブルに保管されている誤認識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単語があった場合、その部分を対応する誤認識状態系列で表現したものを、新たな発音として辞書に追加登録する。認識時には、追加登録されたマルチ発音辞書を用いて認識を行なう方法である。さらに、本誤認識特性抽出法にて抽出された誤認識を分析した結果、一部の誤認識特性は、単語や品詞の種類に依存していることがわかった。ここでは、誤認識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を提案し、本辞書の有効性を確認したので報告する。

3.3.1 誤認識特性の品詞依存性について

表 3.8に示した評価用データを用いて誤認識特性を抽出し、誤認識された音素が属する単語の性質（品詞）に注目して、誤認識傾向を整理した。分析は話者毎に行ない、各音素毎の誤りに対してどのような品詞で誤りが起きているかを調べた。認識条件、出現数の多い誤り例を表 3.9 に示す。

表 3.9の例に見られるように、一部の誤認識は特定の品詞に偏って出現していることがわかった。たと

表 3.9: 出現数の多い誤認識例における品詞別出現数
correct : 正解文における正解状態系列の出現数
confusion : 認識結果文における誤り状態系列の出現数

誤認識特性 * (/ 正解 / → / 誤り /)	品詞別出現数 品詞 (出現数)
/z/ → /d/	*correct* : 名詞 (23), 代名詞 (21), 本動詞 (24), 感動詞 (24), 副詞 (2)
	confusion : 本動詞 (6), 感動詞 (3)
/ga/ → /e/	*correct* : 代名詞 (14), 感動詞 (20)
	confusion : 感動詞 (8)
/o/ → /e/	*correct* : 名詞 (7), 代名詞 (9), 感動詞 (19), 助詞 (11)
	confusion : 代名詞 (4)

* (注) 上記の誤認識特性は状態系列で表されるものであるが、表が煩雑になることを避けるため、表では対応する音素系列で記述している。

表 3.10: 正解系列の品詞別分布と誤認識特性の品詞別分布との相関で表現された誤認識特性の品詞依存性

誤認識特性 (/ 正解系列 / → / 誤認識系列 /)	相関値
/eo/ → /o/	0.999
/go/ → /ko/	0.998
/s/ → /sh/	0.992
/m/ → /g/	0.972
/ch/ → /z/	0.873
...	
/m/ → /n/	0.545
/do/ → /ro/	0.541
/d/ → /t/	0.411

えば誤認識特性 /z/ → /d/ では、正解系列 /z/ が含まれる単語の品詞の種類はいろいろあるが、この誤認識は感動詞と本動詞でしかみられない。正解系列 /o/ も多くの品詞に出現しているが、誤認識特性 /o/ → /e/ は代名詞にでしかみられない。

この性質をより明確にするために、以下の 2 つ分布の相関関係を調べた。1 つは入力文におけ正解系列の品詞別出現分布であり、もう 1 つは認識結果文における御認識特性の品詞別出現分布である。この 2 つの分布の相関値が高ければ、多くの正解系列を含む品詞で誤認識も多く出現していることになり、その誤認識特性の品詞依存性は小さいことになるが、相関値が低い場合には、誤認識が特定の品詞に偏って出現していることになり、誤認識の品詞依存性は高いといえる。表 3.10は、各々の誤認識特性とその相関値を示している。ここでは信頼性の高い相関値のみを評価するため、25 回以上出現した誤認識特性のみを扱った。上 5 つの誤認識特性に対する相関値は高く、これらは品詞依存性が低いと考えられるが、下 3 つの誤認識特性に対する相関値は低く、一部の品詞に偏って誤認識が起こったことがわかる。

この結果より、一部の誤認識特性は品詞に依存していることがわかった。これらの誤認識特性については、限られた品詞に絞って正解音素または正解語彙の探索を行なうことで、無駄な探索を軽減でき、結果的に探索性能の向上が期待できる。

表 3.11: 認識実験条件

音響モデル	不特定話者用 HM-net 型 [3] [9] 延べ状態数 401 、 混合分布数 10
言語モデル	認識対象単語数 2343 単語 bi-gram、 smoothing なし
認識	One-pass DP 法での N-best 候補探索 ビーム幅 1000 、
データベース 言語モデル学習用 誤認識特性学習用 認識実験用	旅行会話データベース 3476 文 のべ 44147 単語 話者 3 名文、平均 109 文 / 話者 話者 3 名、平均 80 文 (907 単語)/ 話者 (誤認識特性学習用とは異文章、異発声)

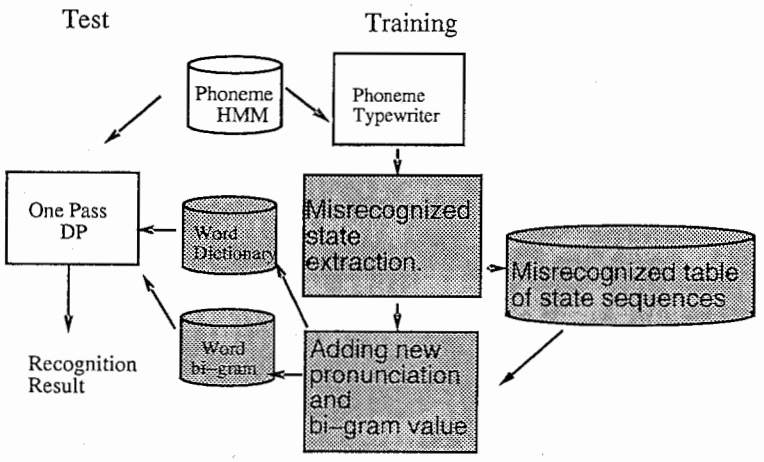


図 3.4: 誤認識特性から作成されたマルチ発音辞書を用いた音声認識システム

3.3.2 誤認識特性の品詞依存性を考慮したマルチ発音辞書作成

認識中に正解語彙を探索するために、HMMの状態系列で表現された誤認識特性から決定された新たな発音を、従来の発音辞書に付加することでマルチ発音辞書を作成した。作成の手順は以下のとおりである。

(1) 音素タイプライタを用いて、複数音素にわたる状態系列として誤認識特性を抽出。

(2) 誤認識毎に出現数、出現率を算出。但し一定数以下の品詞にしか出現しない誤認識系列は，誤認識系列の品詞依存性が強いとみなし、その品詞毎の出現数，出現率を算出。

(3) 単語辞書の全単語を対象に、抽出された誤認識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。抽出した誤認識特性の正解系列を誤認識系列に置き換えた発音を各単語に追加。但し、品詞に依存している誤認識特性については、その品詞に限って発音を追加。

3.3.3 本マルチ発音辞書による正解語彙探索の評価

本マルチ発音辞書の有効性を評価するために、図 3.3.3の連続音声認識システムにて認識実験を行なった。認識実験条件は表 3.3.3に示す。誤認識特性が話者に依存していることを考慮して、音響モデルは話者適応後のモデルを用い、誤認識特性テーブルおよびマルチ発音辞書も話者毎に作成した。高精度なマ

表 3.12: 品詞依存性を考慮したHMM状態の誤認識特性を用いた場合の単語誤認識率の減少

	話者 A	話者 B	話者 C	average
without re-entry	10.38%	5.79%	22.21%	12.79%
PHONE-withoutPOS	11.40%	5.33%	23.40%	13.38%
STATE-withoutPOS	9.18%	5.20%	19.96%	11.45%
STATE-POS	8.85%	3.81%	17.31%	9.99%

ルチ発音辞書を作成するため、2回以下しか出現しなかった誤認識特性は発音辞書作成には用いなかった。また、発音辞書サイズが大きくなり過ぎないように、1つの単語につき登録する発音数を3以下と限定し、3以上の発音候補がある場合には、誤認識特性の出現率が高い上位3つの発音に絞って登録した。

本提案のマルチ発音辞書の有効性を確認するため、本辞書の特徴である次の2つについてその効果を評価した。

(A) HMMの状態表現を用いることの有効性

(B) 誤認識特性の品詞依存性を考慮することの有効性

状態表現による誤認識特性を用いることの効果

誤認識特性を表現する際にHMMの状態表現を用いることによる、正解語彙探索への効果を確認するために、マルチ発音辞書を作成する際に音素表現による誤認識特性を用いた場合 (PHONE-withoutPOS) と状態表現による誤認識特性を用いた場合 (STATE-withoutPOS) との単語誤認識率を比較した。各々の単語誤認識率を表 3.12に示す。全話者において、状態表現による誤認識特性を用いた方が、単語誤認識率が低い (平均 11.45%)。話者Aや話者Cでは、音素表現による誤認識特性を用いると、従来の単発音辞書を用いた際と比較して、逆に誤認識率が増加してしまう。この結果は、HMMの状態系列を用いること、つまり、誤認識部分の前後のコンテキストを考慮しながら誤認識特性を抽出することが、単語認識率を高い信頼度で向上させるのに有効であることを示唆している。誤認識部分の前後のコンテキストを考慮しない場合は、不必要な発音を無駄に登録してしまうことで、逆に認識率が低下してしまう場合がある。

品詞依存性を考慮した誤認識特性の効果

誤認識特性の品詞依存性を考慮したことによる正解語彙探索の効果を確認するために、品詞依存性を考慮しない場合 (STATE-withoutPOS) と考慮した場合 (STATE-POS) との単語誤認識率を比較した。前セクションでの評価では、全ての誤認識特性が品詞に依存している訳ではなかった。そこで本実験では、品詞依存性を考慮した発音辞書を作成する場合でも、誤認識が2種類以下の品詞に限って出現する場合にのみ、品詞依存性を導入した。この結果、話者3名分の誤認識中 0.73% が品詞に依存していた。

単語誤認識率を表 3.12に示す。品詞依存性を考慮した場合には、考慮しない場合より単語誤認識率が低い。(考慮しない：11.45%、 考慮する：9.99%)。考慮した場合には、誤認識減少率は 21.9% になる。この結果から、誤認識特性の品詞依存性を考慮しながらマルチ発音辞書を作成することが、認識性能の向上に効果がることがわかる。

3.3.4 誤認識特性の学習用データ量と単語認識率との関係

これまでの実験では、話者毎に誤認識特性が大きく異なることを考慮して、話者適応後の音響モデルを用いて、各話者毎に誤認識特性を抽出してきた。話者に依存した誤認識特性を用いる場合には、次の2つの条件が重要であると考えられる。

- なるべく少ない学習データから、信頼性の高い誤認識特性を抽出できる
- 学習データ量が増えるとともに、安定して認識率が向上する

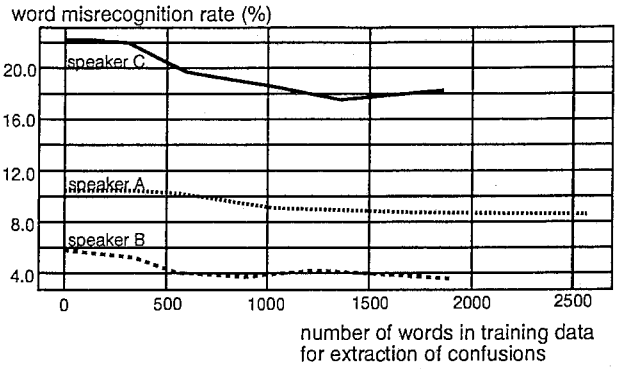


図 3.5: Relation between the decrease of word misrecognition rates and number of words in training data

本正解語彙探索手法が、上記の2条件を満たしているかどうかを確認するために、各話者毎に、学習データ量と誤認識減少率との関係を調べた。結果を 3.3.4に示す。単語誤認識率は話者により異なっているが、誤認識率が安定して減少しているのは、全話者に共通した傾向である。3話者中2話者で、500単語の学習データで誤認識の減少率が限界になり、残りの1話者も、1000単語の学習で限界となっている。

この結果から、話者毎の学習データが使用できるアプリケーションでは、この正解語彙探索法は効果があり、安定して認識性能を向上させることができると考える。話者毎の学習データが収集できないアプリケーションについては、今後、誤認識特性の話者依存性を明確にし、同じような誤認識特性を持つ話者クラスを作成するなどの方法で正解語彙を探索できる手法を確立していく必要があると考える。

第 4 章

おわりに

従来の音響モデルと認識手段では誤認識を起こしてしまう語彙を獲得するために、下記の特徴を持つ誤認識特性抽出法を提案し、この誤認識特性を利用して、認識結果から正解音素探索行なう方法、また、発音辞書作成して正解語彙を探索する方法を提案した。本誤認識特性の特徴は次の点である。

- 認識結果の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誤認識特性を抽出する。
- 誤認識特性がHMMの状態系列として表現されている。

各探索手法を評価した結果、認識結果から正解音素を探索する場合およびマルチ発音辞書を用いて正解語彙を探索する場合ともに、上記に示した本誤認識特性の特徴が、従来認識できなかった音素系列や語彙を確実に探索するのに有効であることを確認した。さらに、誤認識特性の品詞依存性を考慮しながらマルチ発音辞書を作成することが、ことが、より探索性能を向上させることも確認できた。

本研究では、言語翻訳部と組み合わせて音声翻訳システムとしての評価するには至らなかった。しかし、本論文で提案した語彙探索手法は、たとえば特定の語彙のみのマルチ発音辞書化したり音素系列探索処理を行なうことで、翻訳するために落せない語彙を優先的に探索することも可能な枠組になっており、翻訳性能の向上に貢献できる技術として期待できるものであると考える。

謝辞

最後になりましたが、本研究にあたり音素 HM-net 及び音素認識システムを提供して頂き、また有益な御意見および御協力を頂いた第一研究室の Harald Singer 滞在研究員、単語 n-gram を用いた音声認識システム及びマルチ発音単語ネットワークの作成に御協力頂いた（株）東洋情報システムの高橋誠氏、研究方針について度々有益な御助言を頂いた飯田仁第三研究室室長、熱心に討論頂いた音声翻訳研究所第一研究室と第三研究室の皆様に深謝致します。どうもありがとうございました。

1997 年 12 月 25 日

参考文献

[1] R. Schwartz, Y. Chow, O. Kimball, S. Roucos, M. Krasner, and J. Makhoul. "Context-dependent modeling for acoustic-phonetic recognition of continuous speech," ICASSP'85, pp. 1205-1208, 1985.

[2] K. Lee, S. Hayamizu, H. Hon, C. Huang, J. Swartz, and R. Weide. "Allophone Clustering for Continuous Speech Recognition," ICASSP90, pp.749-752, 1990.

[3] J. Takami and S. Sagayama. "A successive state splitting algorithm for efficient allophone modeling," ICASSP92, **vol.1**, pp.573-576, San Francisco, 1992.

[4] C.H. Lee, C.H. Lin, and B.H. Juang. "An study on speaker adaptation of thr parameters of continuous density hidden markov models," IEEE Trans. on Signal Processing, **vol.39**, No.1, 1991.

[5] C.J. Leggetter, P.C. Woodland. "Speaker adaptation of continuous density HMMs using multivariate linear regression" ICSLP94, pp.451-454, 1994.

[6] K. Ohkura, M. Sugiyama, and S. Sagayama. "Speaker adaptation based on transfer vector field smoothing with continuous mixture density HMMs," ICSLP92, pp.369-372, 1992.

[7] H. Singer, M. Ostendorf, "Maximum Likelihood Successive State Splitting, " ICCASP96, 1996.

[8] J. Takami, A. Nagai, and S. Sagayama. "Speaker adaptation of the SSS (successive state splitting)-based Hidden Markov network for continuous speech recognition," Fourth Ausalian SST, pp.437-442, 1992.

[9] H. Singer and J. Takami, "Speech Recognition without Grammar or Vocabulary Constraints, " ICSLP94, pp2207-2210, 1994.

[10] 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" 信学論 (D) 62-D,1, pp.24-31, 1979.

[11] T. Araki, S. Ikehara, and H. Yokokawa, "Using accent information to correctly select Japanese phrases made of strings of syllables," ICSLP94, pp.2155-2158, 1994.

[12] S. Young, J, Jansen, J. Odell, D. Ollason and P. Woodland, The HTK book manual 1996.

第 5 章

著者の関連発表一覧

○ 脇田由実、Harald Singer、匂坂芳典 “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” 日本音響学会春季研究発表会 , 1994-10

○ 脇田由実、Harald Singer、匂坂芳典 “話者適応が誤認識特性に及ぼす影響について” 日本音響学会春季研究発表会 , 1995-3

○ 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” 電子情報通信学会技術研究報告 SP95-30, PP.41-48 1995-6

○ 脇田由実、Harald Singer、匂坂芳典 “Phoneme Candidate Re-entry Modeling using Recognition Error Characteristics over Multiple HMM States.” ESCA Workshop on Spoken Dialogue System, pp.73-76 1995-6

○ 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系列候補探索モデル” 電子通信学会論文誌 J79-D-II, No12, pp.2086-2095, 1996-12

○ 脇田由実、Harald Singer、匂坂芳典” 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の利用” 日本音響学会春季研究発表会 1-6-6, pp.11-12, 1997-3

○ 脇田由実、Harald Singer、匂坂芳典” Speech recognition using HMM-state confusion characteristics” Eurospeech97 M4A.2, Vol1, pp.36-39, 1997-9

○ 脇田由実、Harald Singer、匂坂芳典” Multiple Pronunciation Dictionary using HMM-state Confusion Characteristics” Computer Speech & Language に論文投稿中

第 6 章

ツール関連

上記に記載した実験に使用したツールは、以下のディレクトリにある。

○ 誤認識抽出関連： /home/as41/yumi/PCR/ERROR_EXT/

○ 認識結果からの正解音素系列探索関連： /home/as41/yumi/PCR/PCR_from_RESULT/

○ マルチ発音辞書作成関連： /home/as41/yumi/PCR/PCR_using_WORDDIC/

さらに、誤認識特性を用いて各単語毎のマルチ発音ネットワークを作成するツールを別途作成した。本ツールの仕様などについて以下に示す。これは、開発をお願いした（株）東洋情報システム 高橋氏の報告資料から抜粋したものである。ツールは、 /home/as41/yumi/PCR/PCR_using_WORDDIC/MK_MultiPron にある。

[マルチ発音単語ネットワーク作成ツールについて]

○ 概要

本作業は、平成 9 年前期に実施した「統合処理実験プログラム 認識誤り特性を利用した認識システムの構築および改良」の延長作業である。前期の作業において開発したマルチ発音単語ネットワークの作成部を独立したツールとして切り離す作業と、それに付随する機能の整理を実施した。

このツールは、標準入力した単語 ID に対して、誤り特性テーブルに基づいてマルチ発音ネットワークを作成し表示する。前期の作業では、

（1） 1 単語中に発生する HMnet 状態の挿入／脱落／置換誤り

（2） 2 単語間の渡りの部分に発生する HMnet 状態の挿入／脱落／置換誤りに対応できる

マルチ発音ネットワークを作成する処理を実装したが、本作業では、（1）のみを対象としている。なお、マルチ発音単語ネットワークの作成処理については、平成 9 年度 8 月末の作業報告書を参照のこと。

○ 入力仕様

入力ファイルは、音響モデル（HMnet）、単語辞書、誤り特性テーブルである。以下に、それぞれの形式とサンプルを述べる。

● 単語辞書

(1) 形式

単語 ID 表記形 '|' ヨミ '|' 正規形 '|' 品詞 '|' 活用型 '|' 活用形 '|' 音便 ':' 音素表記

(2) サンプル

```
10005   もしもし | モシモシ | もしもし | 感動詞 ||||:m,o,sh,i,m,o,sh,i
10006   わたし | ワタシ | わたし | 代名詞 ||||:w,a,t,a,sh,i
10007   田中 | タナカ | 田中 | 普通名詞 ||||:t,a,n,a,k,a
```

● 音響モデル

(1) 形式：HMnet の形式

● 誤り特性テーブル

(1) 形式

```
<レコード> :=  <正解状態パス> : <誤り状態パス> '|' <発生率> <品詞条件>
<正解状態パス> := <音素ラベル> '[' #state ... ']'
<誤り状態パス> := <音素ラベル> '[' #state ... ']'
<発生率> := 実数
<品詞条件> := '<' <品詞大分類> ... '>'
```

・正解状態パス側および、誤り状態パスは1音素までしか指定できない。
・品詞条件には、通常に使用している品詞に対する上位カテゴリが指定できる。
　例．名詞 ← 普通名詞、固有名詞、人名、地名

(2) サンプル

```
g [ 230 358 243 ] : k [ 285 305 90 ]   | 4 < $ >
ts [ 207 341 17 111 ] : ch [ 55 13 235 157 ]   | 2 < 接尾辞 >
t [ 343 22 250 ] : p [ 217 25 41 ]   | 3 < 間投詞 >
```

○ 出力仕様

●ネットワーク表示

単語ネットワークの形状をそのまま表示する。ノードを表す番号は、HMnet の状態番号であるため、一意にノードを特定する数字としては使用できないので注意が必要である。ネットワークの形状を見るために使用する。

(1) 形式

数字　HMnet の状態番号　-- ノード接続　$ ノード終端　< > 一度通ったノード

(2) サンプル

```
-1--39--7--394--226--34--5--333--89--215--259--38--140--47-- -1--$
                                                          +--113--< -1>
                                                          +--118--211--< -1>
                                                          +--191--< -1>
                                                  +--269--118--211--< -1>
                                                  +--387--118--211--< -1>
                                                          +--259--< -1>
  +-390--7--394--<226>
```

●接続情報表示

単語ネットワークの各ノードの接続情報を表示する。ノードを表す番号としてノード ID と HMnet 状態番号の組み合わせ（※）を使用しているため、一意にノードを特定することが出来る。他のツールの入力にする場合に使用する。

※ノードIDのみで一意に特定できるが、他のツールでの利用の利便性を考慮してHMnet状態番号を付加している。

(1) 形式
遷移元ノードID'_'遷移元状態番号'　'遷移先ノードID'_'遷移先状態番号

(2) サンプル
```
0_-1   1_39
1_39   2_7
2_7     3_394
3_394  4_226
4_226  5_34
5_34   6_5
6_5    7_333
7_333  8_89
8_89   9_215
```

○ 使用法

【ツール名】
MK_MultiPron

【環境設定】
単語辞書、HMNET、誤り特性テーブルを環境変数に設定する。

●単語辞書
```
setenv WORD_DIC_FILE_NAME  /home/TDMT/LM6.tdmt/MASTER_LEXICON_DEPT3.phone
```
● HMNET
```
setenv PDIC  /home/TDMT/AM/HMnet.F.201.retrain.150.3.state.10.mix
```
●誤り特性テーブル
```
setenv PRONOUNCE_ERROR_TABLE /home/TDMT/markov_cat/etc/pron_error.tab.for_test2.wids
```

環境変数で設定されていない場合は、実行ディレクトリにある、
".HMnet-NGRAMrc"の設定が使用される。

【オプション】
[-net] : マルチ発音単語ネットワークをネットワーク表示する
[-trans] : マルチ発音単語ネットワークの接続情報を表示する
[-org] : もとの単語ネットワークを表示する
[-h] : 使用法を表示する。

【使用例】
```
% echo "10054" | MK_MultiPron -net
##### After Added Word Net Work .... ######################
  -1--108--22--330--212--106--254--6--63--127--162--368--319--89--215--259-- -1--$
                                                          +--281--47--< -1>
                                                               +--113--< -1>
                                                               +--191--< -1>
                                                          +--279--103--< -1>
                                                          +--297--97--103--< -1>
```

```
                                                        +--279--103--< -1>
                                                        +--350--103--< -1>

                    +--331--<106>
   +--343--22--250--<106>
   +-- 69--10--82--<106>
   +--217--25--41--<106>
##################################################################
```