修士学位論文

後続語予測可能文脈に基づく
大語彙連続音声認識に関する研究

東北大学大学院工学科電気・通信工学専攻
馬場 雅美

目 次

図目次

表目次

第1章

序論

1.1 研究の背景

“より安易に” そして “より自分が楽できるように” というのは、人間の飽くことなき欲求である。もっと簡単に様々な情報を得たい。もっと楽に機械を操作したい。それらの欲が現代の高度な技術に溢れた社会をもたらしたと言っても良いだろう。特に 20 世紀も終盤となったここ数年におけるコンピュータの高性能化や、通信技術の発展及び普及は瞠目に値する。10 年前の時点で、パーソナルコンピュータが一家に一台持たれるようになることを、誰が想像し得たであろう。そして今や、それらのシステムを “更に快適に” 使えるようにという欲求も出ている。

この欲求に合わせたように、近年電話やワープロなど音声認識を組み込んだシステムが非常に増えてきている。このような音声認識を使用するシステムの利点は、人間が用いる最も簡単かつ高速なコミュニケーションの手段である “言葉” を利用することで、特別な訓練をせずに、誰でも高速かつ容易に扱うことができるという点である。すなわち、声さえ出せればシステムを作動させることができるわけである。このように従来のマウスやキーボード、タッチパネルと違って、システムを利用しながら手では他の作業ができるというのは素晴しい魅力である。

しかし、このように様々な分野で利用されている音声認識システムも、まだ少なからず問題を抱えている。ここに主な問題点を 2 つ述べる。1 つめの問題は、認識精度を向上させるために、タスク依存性が高いシステムとなっているものが多いという点である。システムの認識精度が良いことは、もちろん喜ばしいことである。しかし反面、そのタスク依存性が高いということは、それだけ認識システムの融通が効かないことを意味している。つまり少しでも認識させたい文がシステム内のデータと違えば、極端な場合その文は全く認

識されなくなる恐れがある。これではもっと幅広い分野での同一システムの利用は困難である。

もう1つの問題は、我々人間が通常会話で発声している音声は、必ずしも全てが正文ではないところにある。正文とは、倒置や言い淀み、言い直し、余分な言葉の挿入等がない文のことを指す。また反対にこれらが含まれる文を非文という。人間は正文だけでなく非文を聞いても、かなり高精度にその文の意味を類推して認識することが可能だが、コンピュータではそうはいかない。従って、人間にとって自然な発話である連続音声を扱うシステムの研究は、未だ難航している。

そして、これらの問題点を改善するため、自然発話を扱う連続音声の認識では言語モデルが必ず用いられている。

1.1.1 認識に使用する言語モデル

現在の認識システムの認識性能はかなり向上しているとはいうものの、未だ人間によるものに比べると格段に悪い。それは前述した2つの問題点に起因している。一般に、人間の持つ認識能力はコンピュータの持つそれよりも、はるかに複雑かつ柔軟なものであると言われている。人間の持つそのような認識能力をコンピュータに持たせるには、コンピュータに自然言語を理解する能力を与える必要がある。しかし人間は、経験や学習を通して多様で膨大な情報を得て、それを類型的なパターンとして概念に対応づけて認識している [1] といわれ、これをコンピュータに加えるのは、大変困難なことである。

そこで近年では、コンピュータの認識能力を少しでも人間に近づけるために、音声認識時に自然言語の生成過程に対する近似を定式化した言語モデルを用いている。言語モデルは、認識時での曖昧性を削減すると同時に、単語仮説の探索空間を効率的に絞り込む役目を果たす。これは、認識結果をより確かなものとすることができ、連続音声認識では必要不可欠なものである。従って、認識精度の向上のためには、より高い性能を持つ言語モデルを使うことが重要となる。

連続音声認識で一般的に用いられている言語モデルには、有限オートマトンなどで表される文法規則に基づいた言語モデルと、多大なデータ (以下コーパスと呼ぶ) から求める統計的な言語モデルがある。特に近年では、コーパスに基づく統計的な言語モデルが盛んに用いられるようになっている。これは計算機の性能向上に加えて、言語コーパスの充実によるところが大きい。

1.1.2 大語彙連続音声認識の問題点

商用化されている音声認識システムの中には、単語を区切って発話する “孤立単語” を認識して扱っているものがある。しかし “孤立単語” は実際の人間の通常の活動においては不

自然であり、非常に恣意的である。それに対して連続音声認識、その中でも特に数万単語を越える語彙を扱う大語彙連続音声認識の分野は、日常で人間が発声する音声言語の大半を対象としている。よって、大語彙の連続音声を認識することで、人間にとって不自然な“孤立単語”でなく、自然かつ効率的な“連続発話”を用いることができる。つまり、大語彙の連続音声を認識できる技術が確立して初めて、音声認識技術の用途が大きく広がると予想される。このことから、音声認識技術の応用分野拡大において、大語彙連続音声認識は極めて重要な課題であるといえよう [2]。

現段階での大語彙連続音声認識における研究の流れは、ディクテーションと対話音声の認識を対象とした研究に分けられる。前者は正文や書き言葉に近い音声を発声し、読み上げられたすべての言葉を文字に変換するというものである。これは、入力される文は文法的にはほぼ正しいと想定できるが、認識しなければならない語彙や文体が非常に多様であるという点に難しさがある。近年では、新聞記事の読み上げによるディクテーションの研究が多く行なわれている [3]。また、後者の対話音声認識の場合には、タスクをある特定の分野に限定すれば、ある程度語彙を制限することができる。しかし話し言葉特有の曖昧な文章や、文法的に誤った文も認識の対象としなければならないため、困難なものとなっている [4]。

このような大語彙連続音声認識の、現在における主な問題として、単語列の最適な組み合わせをどのように決定するかということがある。そして、更にこのことは次のような問題点を含んでいる [5] [6]。

- 連続発声された単語の境界が明確でない
- 単語境界付近の音が、先行または後続の単語の影響で変形する調音 (word juncture) 結合が起こる
- 単語ひいては単語を構成する各音の継続時間がかなり短くなり、その発音も曖昧なものとなりやすい
- 単語仮説数が非常に多いため、網羅的なマッチング方法では処理量が膨大なものとなってしまう

これらの問題を解決するために、効率的に単語仮説の最適なマッチングを行なうためのDPマッチングというアルゴリズムが存在する。これは、処理量の増加を防いで、なおかつ最適な単語列を得ることができるというものである。

また更に、実際の認識システムで大語彙の連続音声を扱うためには、膨大な量の単語仮説の中から最適な文を現実的な処理量で探索することが必要である。つまり単語仮説の探索空間を縮小しなければならない。この問題の解決策として、A^* 探索法 [7] やビームサーチ [8] という方法がある。これらの手法は、なんらかの制限の下で認識経過の各単語仮説に

おいて次にくる単語を展開するか否かを定め、探索空間を縮小するというものである。探索空間を狭めるための制限はその経過までの単語仮説の尤度を用いており、次にくる単語についての条件等は考慮していない。

このような手法を用いて探索空間の縮小を行なうと、ビームエラーが生じる恐れがある。ビームエラーとは、正しい結果として出力され得る単語仮説を枝刈りしてしまったために、認識結果が正しく得られなくなる状況をいう。すなわち、処理量の大幅な削減を行なおうとすると認識精度が低下してしまう。このようなトレードオフがあるため、処理量を実用レベルまで減少させることができても、認識精度が非常に低下してしまう恐れがある。従って、より効率的な単語仮説の探索手法が必要とされる。

1.2 研究の目的

最近の音声認識で使用する言語モデルとしては、統計的な言語モデルが大半を占めている。しかし統計的な言語モデルを用いた大語彙の連続音声認識には、出現しうる仮説の数が極めて膨大なものとなるという欠点がある。従って、無駄な仮説を展開することなく効率的に、しかも最も確からしい仮説を見つけるための探索手法が重要であることは、1.1.2節で述べた通りである。またその対処法であるビームサーチを使用する場合、ビームエラーを減少させる必要もある。ここで、確からしい単語仮説をビームエラーで排除することなく求める手段として、認識途中で得られた単語仮説の後に展開する単語を限定するというものが考えられる。すなわち、単語仮説の次にくる単語を的確に限定して予測できるならば、確からしい単語仮説の累積尤度を高くすることができるというものである。今まで認識経過における単語仮説の後続単語について、その単語仮説毎に集合を特定して予測するような手法はなかった。

そこで本研究では、統計的手法による言語モデル n-gram を基盤とした高性能な単語予測モデルを提案し、そのモデルを用いて効果的に認識における処理量を削減することを目的とする。最終的にはそのモデルを実際に認識システムに実装して認識実験を行ない、認識における処理量の削減を確認する。単語予測モデルについて具体的には、特定の文脈の次にくる単語候補を限定、削減することにより、単語仮説を展開する処理量を減少させる。

1.3 本論文の構成

本論文の構成は次の通りである。

第1章 序論

研究の背景や目的、及び本論文の構成を述べる。

第2章 文脈の単語予測による認識の高性能化

音声認識処理の高精度化を実現させるために、次にくる単語を、文脈を用いて予測するモデルを考案し、その文脈を抽出する方法を提案する。また考案したモデルを用いて評価実験を行ない、どの程度高性能化が実現されうるか確認する。

第3章 単語予測モデルの認識システムへの実装

前章で考案した単語予測モデルを用いて実際の連続音声認識システムを実装する手法を提案する。また、その提案手法に基づいて実装したシステムにより認識実験を行なう。

第4章 大語彙連続音声認識における単語予測

第3章を踏まえて、形態素解析の最適化によるモデルの改変、及び単語辞書の語彙数が増えたときの単語予測モデルの処理方法について考案する。そして、それらを用いて連続音声認識システムを実装し認識実験を行なうことで、単語予測モデルの有効性を調べる。

第5章 結論

本研究の成果、及び今後の課題について述べる。

第2章

文脈の単語予測による認識の高性能化

2.1 はじめに

前章で述べたように、大語彙連続音声認識では、出現しうる単語の数が極めて膨大となる。従って認識時にそのまま全ての単語仮説を展開すると、莫大な処理量となり認識に時間がかかりすぎてしまう。そこで、その中から最適な文を効率的に探索する処理を行ない、実際に利用可能な範囲で行なえるように、探索空間を絞り込むことが要求される。本章では、従来の言語モデル及び連続音声認識の概略を述べ、その後文脈の単語予測可能性を考慮することで、探索空間を効率的に絞りこむ手法を提案する。

2.2 大語彙連続音声認識で用いる言語モデル

大語彙連続音声認識における利点は、前章で述べたように自然発話を扱うことができるところにある。そのため、特に大語彙連続音声認識についての研究が、多くの研究機関で進められている。現段階で、大語彙連続音声認識で用いられている言語モデルには、主に次に述べる2つがある。

2.2.1 文法に基づく言語モデル

1つめは、有限オートマトンや文脈自由文法といった文法に基づく言語モデルである。これは実際の文がどのような品詞から成っているか、各単語はどの品詞に含まれるか、更にはそれらがどのように連接しているか等を分析し、それを文法として記述した言語モデル

である。これを音声認識で使用する際には、認識で扱うタスクに応じて適切な文法を記述する。よって、このような文法に基づく言語モデルを用いた大語彙連続音声認識では、作成した文法規則に従わない単語仮説、すなわち文法的にあり得ない単語仮説が生成されることは初めからない。従って、認識時における音素のマッチング回数が少なくてすみ、処理量が小さいことが長所としてある。一方短所としては、文法規則がタスクに依存してしまい柔軟性に乏しいことや、規則を作るには文法や品詞等に関する専門的知識が必要なためにメンテナンスに多大な労力を要することが挙げられる。また、人間の発話自体が必ずしも定められた文法に則すとは限らないため、発話が正文であっても非受理となってしまうことがある。

従って実際の大語彙連続音声認識においては、文法に基づいた言語モデルは利用するには困難な点が多い。

2.2.2 n-gram

2つめは、近年主流となっているコーパスに基づく統計的な言語モデルの n-gram である。n-gram とは、単語 w の生成確率が直前の $n-1$ 個の単語にのみ依存していると統計的に考えるものである。

単語 w_i の直前の $n-1$ 単語から成る単語列 $w_{i-(n-1)}w_{i-(n-2)}\dots w_{i-2}w_{i-1}$ の出現頻度を $N(w_{i-(n-1)}w_{i-(n-2)}\dots w_{i-2}w_{i-1})$ と表すと、w_i の n-gram 確率は式 (2.1) のように推定することができる。

$$P(w_i|w_{i-(n-1)}w_{i-(n-2)}\dots w_{i-1}) = \frac{N(w_{i-(n-1)}w_{i-(n-2)}\dots w_{i-1}w_i)}{N(w_{i-(n-1)}w_{i-(n-2)}\dots w_{i-1})} \tag{2.1}$$

従って、単語列の生成確率である n-gram モデルは式 (2.2) のように与えられる [6]。ここで n-gram 確率の条件となる単語列 $w_{i-(n-1)}w_{i-(n-2)}\dots w_{i-1}w_i$ を文脈という。

$$P(w_1w_2\dots w_n) = \prod_{i=1}^{L} P(w_i|w_{i-(n-1)}w_{i-(n-2)}\dots w_{i-1}) \tag{2.2}$$

通常 $n=1$ の場合の n-gram を unigram、$n=2$ の場合を bigram、$n=3$ の場合を trigram という。n の値が大きいほど、限られたコーパスデータから信頼性の高い n-gram 確率の値を推定するのが難しくなる。例えば、認識語彙数を V とすると、$w_1w_2\dots w_{i-1}$ の異なる単語列数は V^{i-1} となり、これらの各々について確率を求めておくことは不可能である。従って、現在の段階で音声自然言語処理でよく使用されているのは、bigram あるいは trigram である。

大語彙連続音声認識では、逐一文法規則を生成するのは大変な作業となるため、文法に基づいた言語モデルを用いるのは困難である。それに対して n-gram は単語列の確率を求

めるものなので、大語彙においても比較的容易に作成することができる。そのような理由から、最近の大語彙連続音声認識では n-gram を用いるのが主流となっている。

n-gram の長所として、コーパスから統計的に導出するためタスクの依存性が高くなく様々なタスクに利用可能なことと、文法に基づく言語モデルに比べてメンテナンスが非常に容易であることが挙げられる。しかし一方で、実際の文として存在しうる単語列が学習用コーパス内に偶然存在しないと、n-gram 確率が零となってしまう。n-gram 確率が零であるということは、認識時にその単語列が出現しえないということである。そこでスムージング処理を行なう必要が出てくる。スムージングとは、学習用コーパス内に出現しない単語列が、認識時に生成されなくなるのを防ぐために、その単語列の n-gram 確率を正とする操作である。

2.3 n-gramのスムージング

スムージング (smoothing) とは、学習用コーパス内に出現しなかった単語列に関する n-gram 確率を正にする操作のことである。学習用コーパスの量が不十分な場合には推定が正しく行なわれるとは言い難いため、このような処理が必要不可欠なものとなる。次にスムージングの手法を次に 2 つ挙げる。

2.3.1 底上げ法

1 つは底上げ法 (flooring) という手法である。この手法は非常に簡単なもので、学習用コーパス内に出現しなかった単語列による n-gram 確率に ε なる正の定数を割り当てるというものである。これにより推定した確率は式 (2.3) で表される。

$$\bar{P}(w_i|w_{i-2}w_{i-1}) = \begin{cases} \varepsilon & \text{if } N(w_{i-2}w_{i-1}w_i) = 0 \\ \hat{P}(w_i|w_{i-2}w_{i-1}) & \text{otherwise} \end{cases} \qquad (2.3)$$

但し $0 < \varepsilon \ll 1$

ただこれにより推定された $\bar{P}(w_n|w_1w_2\ldots w_{n-1})$ は、明らかに確率の定義による以下の条件を満たすことはない。

$$\sum_{w_n} \bar{P}(w_n|w_1w_2\ldots w_{n-1}) = 1$$

2.3.2 back-off平滑化

スムージングのもう 1 つの手法として、back-off 平滑法 [9][10][11] がある。これの基本的な考え方は、学習用コーパス内に出現しない n-gram の確率を、$(n-1)$-gram 確率に基づい

て平滑化するというものである。これは、2.3.1節で述べた底上げ法よりも精度の高い確率推定が期待できる。

具体的な方法としては、まず学習用コーパス内で出現したn-gram確率を$\lambda(w_1w_2\ldots w_{n-1})$だけ小さくする。この作業をdiscountという。次にそのdiscountした分を、1回も出現しなかったn-gramに再配分(redistribution)してやることで、n-gram確率を全て正にする。この操作を式で表したものが、式(2.4)である[10]。

$$\bar{P}(w_n|w_1w_2\ldots w_{n-1}) = \begin{cases} K\lambda(w_1w_2\ldots w_{n-1})\bar{P}(w_n|w_2w_3\ldots w_{n-1}) & \text{if } N(w_1w_2\ldots w_n)=0 \\ \hat{P}(w_n|w_1w_2\ldots w_{n-1}) & \text{otherwise} \end{cases} \tag{2.4}$$

ここで、$N(w_1w_2\ldots w_n)$は単語列$w_1w_2\ldots w_n$の出現頻度である。またKは、推定された$\bar{P}(w_n|w_1w_2\ldots w_{n-1})$が以下の条件を確率の定義より満たすための定数である。

$$\sum_{w_n}\bar{P}(w_n|w_1w_2\ldots w_{n-1}) = 1$$

更に、$\hat{P}(w_n|w_1w_2\ldots w_{n-1})$と$\lambda(w_1w_2\ldots w_{n-1})$は、次のような条件下で与えられる。

$$0 \le \hat{P}(w_n|w_1w_2\ldots w_{n-1}) \le \bar{P}(w_n|w_1w_2\ldots w_{n-1})$$

$$\sum_{w_n}\hat{P}(w_n|w_1w_2\ldots w_{n-1}) = 1 - \lambda(w_1w_2\ldots w_{n-1})$$

2.4 連続音声認識における処理量の問題

連続音声認識における問題は、“人間が自然に発声した連続音声を与えたとき、単語の最適な組み合わせはどのように決定されるか”ということである[5]。この問題を解決するには、次のような問題を解く必要がある。

1.1.2節で述べたように、連続発声では調音結合のために単語境界を明確に指定することが難しいものが多い。これは言い換えるならば、発声のどこが単語の始まりでどこが終わりであるか正確には分からないということである。また通常、発話音声内に含まれる単語数Lは分からない。つまりこれらのことから、認識すべき音声のどの部分が何の単語に対応しているかを適切に定めることは難しいということがいえる。更に、V個の単語の集合と単語列中の単語数Lに対して、V^L通りの可能な組み合わせが存在する。すなわち、VとLが非常に小さな値である場合を除いて、マッチングパターンの合成で作成される指数的な単語列の組み合わせは膨大な数になることがわかる。

そこで、このような問題を解決するための効率的なアルゴリズムとして、単語仮説の最適なマッチングを行なうワンパスアルゴリズムについて述べる。このアルゴリズムは、実際の音声認識でしばしば利用されているものである。

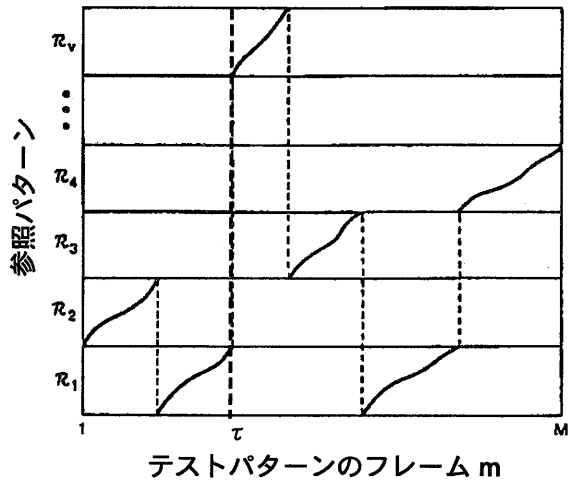


図 2.1: ワンパスアルゴリズム

2.4.1　ワンパスアルゴリズム (one-pass dynamic programing)

まず、参照パターン及びテストパターンを、次のようなベクトル系列で表わす。

$$\begin{aligned} &\text{参照パターン} \quad : \quad R_i = \{r_i(1), r_i(2), \cdots, r_i(N_i)\} \\ &\text{テストパターン} \quad : \quad T = \{t(1), t(2), \cdots, t(M)\} = \{t(m)\}_{m=1}^{M} \end{aligned}$$

ここで、参照パターンについて、N_i は i 番目の単語の参照パターンのフレーム数で、i は V 単語の語彙に対して、$1 \leq i \leq V$ である。また、M はテストパターンのフレーム数である。さらに、可能な全ての単語連接のパターン数を l $(L_{\min} \leq l \leq L_{\max})$ とする。

基本的な考え方を、図 2.1 を用いて説明する。この図は、水平軸にテストパターン T を、垂直軸に参照パターンのセット $\{R_1, R_2, \cdots, R_V\}$ をとったものである。

参照パターン R_v $(1 \leq v \leq V)$ のフレームのインデックスを n $(1 \leq n \leq N_v)$ で表わすと、各テストフレームに対する累積距離は式 (2.5) のように計算される。ここで $d(m,n,v)$ は、$2 \leq n \leq N_v$ と $1 \leq v \leq V$ に対するテストパターンのフレーム $t(m)$ と参照パターン $r_v(n)$ の局所的な距離である。また、ここでは最大の経路の伸長を 2 倍までと仮定しているため、距離を接続する段階で参照パターンのフレームを 2 フレームだけ後ろに探索する。この再帰計算は、各参照パターンの内部の全てのフレームに対して実行される。

$$d^l(m,n,v) = d(m,n,v) + \min_{n-2 \leq j \leq n}(d^l(m-1,j,v)) \quad \text{if } n \geq 2 \tag{2.5}$$

参照パターンの境界 $n = 1$ では式 (2.6) のような再帰計算となる。

$$d^l(m,1,v) = d(m,1,v) + \min[\min_{1 \leq r \leq V}[(d^{l-1}(m-1,N_v,r), d^l(m-1,1,v)]] \tag{2.6}$$

従って、最適な単語列に対応する経路の距離は、式 (2.7) で求められる。

$$D^* = \min_{1 \le l \le L_{\max}} \min_{1 \le v \le V} [d^l(M, N_v, v)] \tag{2.7}$$

ここで言語モデルとしてスムージングした n-gram を用いると、図2.1のように、テストパターンのフレーム τ が参照パターン R_1 の最終フレームに対応している場合、フレーム τ の後に全ての参照パターンの初期フレームを展開していくこととなる。つまり、あらゆる参照パターンの連接による単語仮説を展開することになり、処理量は膨大なものとなる。

しかし、このワンパスアルゴリズムの利点は、テストパターンのフレーム m での計算を、フレーム同期で行なえることである。従って、式 (2.6) と式 (2.7) に必要な全ての計算を1つのフレーム時間内に行なえるような実時間実行に適している。そのため後述するビームサーチで用いるのに都合が良く、一般に連続音声認識に利用される。

2.4.2 探索空間の縮小

前節までで述べたように、スムージングした n-gram を用いて単語仮説を展開した結果、その時点で終端している単語の後にあらゆる単語が来る可能性を考慮しなければならなくなる。従って探索空間が非常に大きくなり、処理量が非常に多くなってしまうという問題がある。以下では、それを解決する手法として、認識システムでワンパスアルゴリズムと一緒によく用いられるビームサーチについて述べる。

2.4.3 ビームサーチ

認識時における各フレーム (時刻) について、次に続く可能性がある単語を全て連接していくと、単語仮説の総数が爆発的に増えてしまう。ビームサーチとは、そのような状態を避けるため各フレーム毎に、累積尤度の低い単語仮説をある制限のもとでその後の探索から除外するという手法である [8]。このように単語仮説を除外する処理を枝刈りという。枝刈りを行なうための制限には、次に挙げる2通りの方法がある。1つは、ある閾値以下の累積尤度を持つ単語仮説を全て削除するものである。もう1つは、最大累積尤度から累積尤度の高い順に定数個の単語仮説だけ残して、残りの単語仮説は全て削除するというものである。このように、フレーム毎に最尤なものからビーム幅で制限した単語列仮説に対してのみ計算を続けることによって、処理量及びメモリ量を大幅に減少させる。

しかし、制限をあまり厳しくすると処理量は削減できても、今度はビームエラーが生じてしまって認識率が低下してしまう。ビームエラーとは、正しい結果として出力され得る単語仮説を、早期段階で枝刈りしてしまったために、正しい認識結果が出力されなくなる状況をいう。これらのことが原因となって、処理量と認識率のトレードオフは難しい問題となっている。

2.5 状態間の遷移

2.5.1 文節構造

言語学的には、文節構造は有限状態オートマトンで良く近似されることが知られている。その概略の一例を図2.2に示す [12]。状態の総数が10個という非常に小さな有限状態オートマトンであるが、これを見ると、例えば活用語の後に名詞が続くことはなく、必ず活用語尾がくることがわかる。このように、有限状態オートマトンの中では、そこから出ている枝が限られた数となっている状態がほとんどである。従って、もし認識させようとしている発声が有限状態オートマトンで受理される文であるならば、認識時の処理量は少なくてすむ。

また有限状態オートマトンでは、そのオートマトンによって生成されうる文しか受理しないので、文として成り立たないものが結果として出ることはない。それらの点では、文節構造を近似した有限状態オートマトンは、高精度な言語モデルであるとされる。

状態から遷移する枝に着目すると、有限状態オートマトンは n-gram に比べて格段に少ないことが明らかである。これはすなわち、その状態の次にどのような単語がくるかを的確に予測しているといえ、本論文の目的に近いものがあるといえる。

しかし人間が自然発話するときには、必ずしも前もって作成した有限状態オートマトンで示されるような文体で話すわけではない。むしろそのように話すことの方が少ないと考えられる。つまり有限状態オートマトンは、実際の文としては正文であっても、そのオートマトンにおいて表すことができない文に関しては一切受け付けてはくれないという欠点も併せ持っている。更に言い淀みや言い直し、倒置などが含まれた、人間にとって自然な発話においても、それらを確実に受理できるような有限状態オートマトンを作成するのは手間がかかるため、必ずしも有効な手段ではない。

2.5.2 状態遷移数の削減

有限状態オートマトンでは、文節の先頭の状態から出る枝の数は数多くあるが、文節の先頭以外の状態では、そこから出ている枝の数は大変限られたものである。それに対して n-gram は、先に述べたスムージングの操作により各状態から出る枝の数が辞書に含まれる単語数分にもなってしまう。その結果として認識時における処理量が大きくなってしまうので、ビームサーチが必要となるのは前述した通りである。

しかし単語によっては、次にくる単語がある程度決まっているものもある。すなわち、n-gram の状態の中には枝の数を削減できるものも存在すると考えられる。具体的な例を図2.3に示す。もしこのような状態が実際に十分多いならば、その文脈に対して次にくる単語を制限してやることで、遷移する枝の数を削減することができる。このことから、次にく

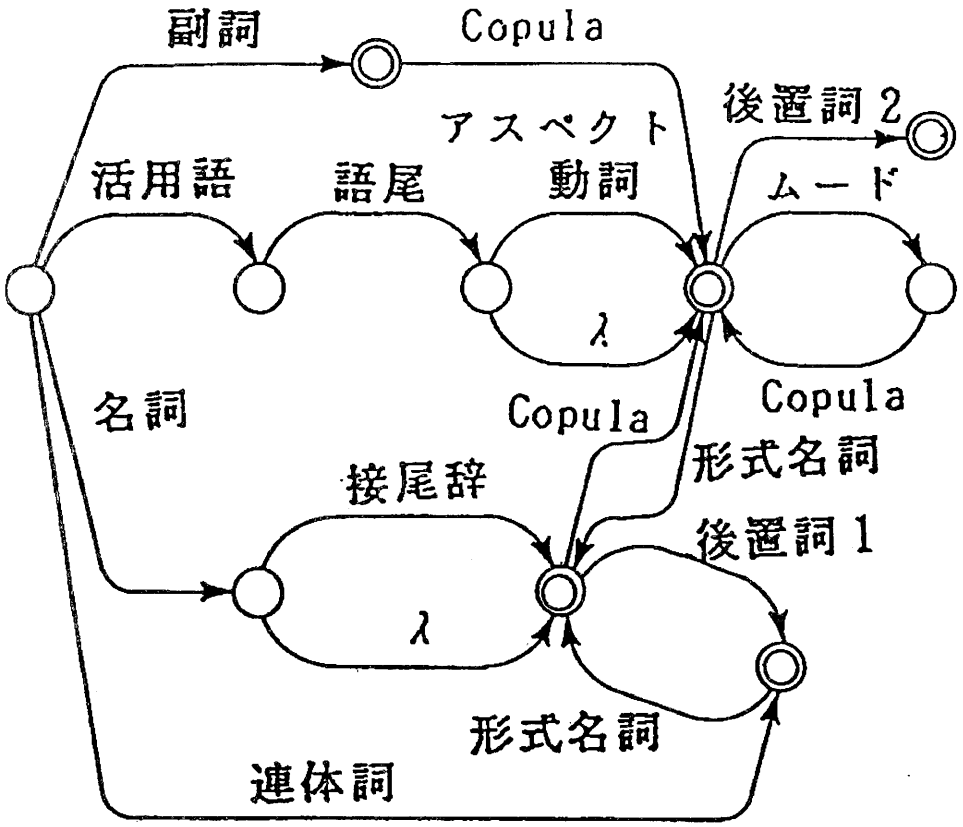


図 2.2: 文節構造を示す有限状態オートマトンの概略の一例 [12]

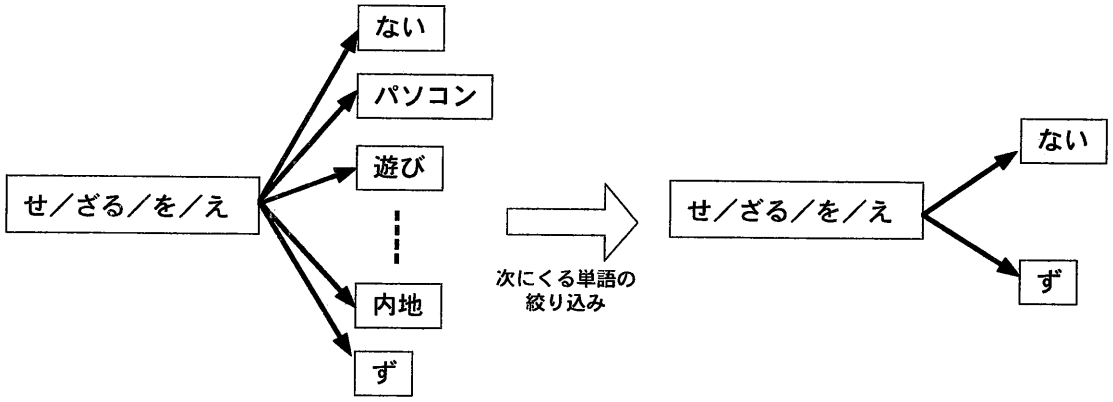


図 2.3: 単語候補数の削減

る単語が決まっている n-gram である場合に展開する単語を絞りこむことは、有限状態オートマトンを自動的に生成するのと同じ意味を成すと考えられる。

しかしこれには、本質的に次にくる単語の予測が不可能な状態の存在はないか、その単語の予測が統計的に信頼性があるか、また認識に悪影響を及ぼさないか等考慮すべき問題がある。それを解決するために、n-gram の各文脈において、それが次にくる単語を的確に予測することが可能か否かを選別する必要がある。

2.6　文脈による後続単語予測

2.6.1　文脈による後続単語予測の可能性

人間は会話をしているときだけに限らず、一方的な発話を聞いているときでも、ある文脈が出てきたならば、次にどのような言葉が続くかを経験的に推測することが可能である。例えば「次の授業は欠席せざるをえない」という発話を聞く場合を考える。この場合途中の「(欠席) せざるを」という文脈を聞けば、次にくる言葉は恐らく「えない」もしくは「えず」のどちらかであろうと想像することはたやすい。しかし「次の」という文脈の場合は、その後に続く単語が何かを推測することは非常に困難である。この理由は、文脈である「次の」の次に出現しうる可能性が高い (同程度のと言い換えてもよい) 単語が膨大な数存在するからである。

このようなことから、文脈によって次にくる単語を予測できないものと、予測できるものが存在することがわかる。これと 2.5 節から、次にくる単語を的確に予測できる文脈を求めれば、それを認識に用いることによって処理量を削減できると考えられる。

2.6.2 Successor-Predictable Context

前節までのことをふまえて、コーパスに基づく言語モデルである n-gram をベースとして、特定の文脈の次にくる単語を予測することによる連続音声認識の高性能化を図る。本論文では、このような次にくる単語が予測可能な文脈を Successor-Predictable Context(以下 SPC と呼ぶ) と名付け、これを用いて特定の文脈の次にくる単語の予測を行なうことにする。文脈を SPC として次にくる単語を予測する、このようなモデルを SPC モデルと呼ぶ。また、SPC の後に予測される単語を SPC の予測単語候補と呼ぶ。

SPC モデルによる効果として期待されることを次に挙げる。第1に、スムージング処理のために生じる、実際の文としてはあり得ない単語仮説が初めから排除される。それによって探索空間の絞りこみが行なわれ、同じビーム幅において従来の n-gram より多くの仮説を検証できる。これは言い換えれば、SPC モデルを使用すればビーム幅を削減しても従来の n-gram と同程度の精度を得られるということである。次に高次の n-gram の制約を効率良く用いることができる。これによって認識処理の最初の段階において、処理量の問題から、高精度な高次の n-gram ではなく、精度の劣る低次の n-gram を使わなければならないという問題を解決できる。最後の利点として、単語辞書に含まれない単語でも SPC の予測単語候補として存在させることができ、認識が可能となることがある。

手順としては、認識前の準備として SPC モデルを SPC 選別法に従って求めておき、認識時にその SPC モデルにより単語予測を行なうことで認識の高性能化を図る。各段階での手続きを次に示す。

- 認識の準備段階：SPC の選別
  - 学習用コーパスにおいて単語 n-gram を導出。その出現頻度、及び n-gram 確率を求める。
  - 上で求めた単語 n-gram から SPC 選別法に則して SPC モデルを求める。
- 認識の処理段階：文脈の次にくる単語の予測
  - 認識段階において、その時点までの単語仮説による文脈が SPC なら、SPC モデルによって決められた予測単語候補のみについて展開する。また SPC 以外の文脈ならば、従来通りスムージングを利用して全単語仮説を展開する。

つまり文中の文脈の多くが SPC であれば、認識時の処理量が小さくなり望ましいといえる。しかし、適当な文脈をむやみに SPC としては、単語予測を失敗してしまう可能性が高くなる。そうならないように SPC を選別する方法を考える必要がある。

2.6.3 SPCの選別手法

SPCモデルは、認識結果に悪影響が及ばないようにするため、単語予測を失敗する危険ができるだけ小さくなるように選別しなければならない。言い換えれば、認識精度を大幅に低下させることなく、処理量を削減することがSPCモデルの目的である。

そこで、単語列 $w_1w_2\ldots w_n$ の中で n-gram 確率が高くて出現頻度も多い単語列ならば、単語 w_n を単語列 $w_1w_2\ldots w_{n-1}$ の次にくる単語として信頼でき、その文脈をSPCとできるのではないかと考えた。この考えに基づいて、文脈がSPCか否かを選別する方法として次のようなものを考案した。

まず出現頻度 α 回以上の同一文脈 $w_1w_2\ldots w_{n-1}$ の単語 n-gram について、出現確率の上位から定数 β 個を取り出す。そして、それらの累積確率が定めた閾値 $\gamma\%$ 以上であれば、その単語 n-gram の文脈をSPCとして定め、単語予測に用いることとする。

SPCを定めたら、次に各SPCに対する予測単語候補をどのような条件で求めれば、単語の予測においてより高い信頼がおけるかを考える必要がある。SPC選別法における、n-gram の文脈に続く出現確率の上位 β 個の単語 w_n のみを単語予測候補にすると、SPCの次に展開する単語数が格段に絞られ大幅な処理量の削減が行なわれる。また、大抵の場合において n-gram 確率の高い単語 w_n を予測することができる。

しかし一方、このような定義で予測単語候補を抽出したならば、上位 β 個未満で累積確率が $\gamma\%$ となる場合には単語予測における信頼性の低い単語候補まで作ってしまうこととなる。また逆に、上位 β 個ではSPC選別の条件は満たさなかったが、もう少し多くの n-gram との累積尤度を求めれば条件を満たす場合も考えられる。このような点から、この手法で予測単語候補を抽出しては、単語の予測に関して信頼性の低下が危ぶまれる。

以上のことから、各SPCの予測単語候補は、そのSPCを文脈としてもつ、出現頻度が2以上の n-gram の単語 w_n のセットとすることを提案する。

上で述べたことをまとめて、SPC及びSPCの予測単語候補の選別手法を式で表すと以下のようになる。但しここで、V は語彙、$N(a)$ は a の出現回数、w_n^i は $P(w_n|w_1\cdots w_{n-1})$ の値でソートしたときの第 i 番目の w_n である。またこれを図で示したものが、図2.4である。

$$\text{SPC の集合} = \left\{ w_1\cdots w_{n-1} \;\middle|\; \begin{array}{l} \exists w_n \in V, N(w_1\cdots w_{n-1}w_n) \geq \alpha, \\ \sum_{i=1}^{\beta} P(w_n^i|w_1\cdots w_{n-1}) \geq \gamma \end{array} \right\},$$

$$\text{SPC } \psi \text{ の予測単語候補} = \{w_n \mid N(\psi w_n) > 1\}.$$

このように提案したSPC選別法を用いて作成したSPCモデルによる単語予測が、認識精度に与える影響について評価する必要がある。

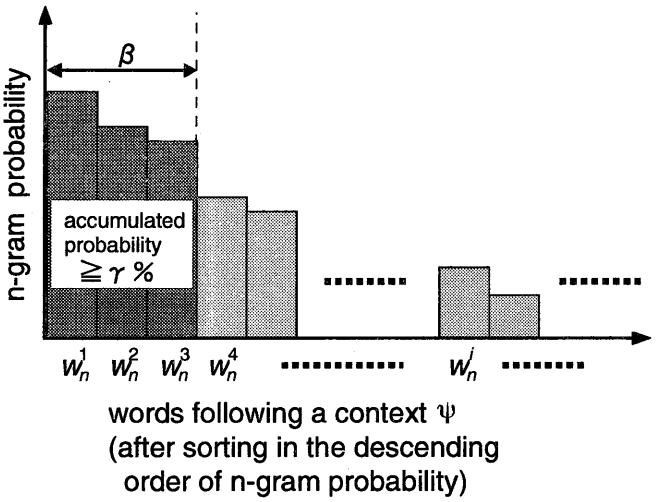


図 2.4: SPC 選別法

2.7　認識に与える影響についての評価実験 (1)

評価実験の結果において評価する際に、次のような言葉を用いる。

- 予測不能

  文脈が SPC でないため、次にくる単語は基本語彙に含まれる全ての単語となる。

- 予測可能

  文脈が SPC である場合で以下の 2 つに分けられる。

  – 予測失敗

    文脈は SPC だが、次にくる単語が SPC の予測単語候補の中に含まれていない。

  – 予測成功

    文脈が SPC で、次にくる単語が SPC の予測単語候補の中に含まれている。

2.7.1　実験条件

評価実験には毎日新聞記事 CD-ROM [13][14] を使用した。学習用データには、その 1991 年 1 月から 1994 年 9 月までの 45ヶ月分の記事データを用いた。また、評価用データとして、同 1994 年 10 月から 12 月の 3ヶ月分を用意した。

表 2.1: SPC モデルに対する評価実験の条件 (1)

SPC モデル n	$2, 3$
出現頻度の閾値 α 回	$3, 4, 5, 7, 10, 15, 30$
累積確率を求める際の上位 β 個	$3, 5, 7$
累積確率の閾値 $\gamma\%$	$90, 95$

ここでの毎日新聞記事 CD-ROM の形態素解析は、RWCP テキストデータベースを用いて行なった。また SPC モデルを求める際に使用した単語 n-gram の作成には、CMU-Cambridge SLM ツールキット [15] を利用した。

SPC モデルを求めるためのパラメータ、すなわち n-gram の出現頻度の閾値 α, 累積確率を求める際の上位からの個数 β 及び累積確率の閾値 $\gamma\%$ は表 2.1 の通りである。また共に SPC モデルの n についても示す。

2.7.2　実験に用いた評価尺度

評価基準としては、次に述べる 2 つの尺度を用いた。

- 文カバレッジ

  評価用データ内で SPC モデルによる単語の予測が一度も失敗しなかった文の割合

  文カバレッジが高いことは、SPC モデルによる単語予測が認識精度に与える悪影響が小さいことを意味する。しかしこの定義式では SPC が出現しなかった文も分母に入っているために、一概に文カバレッジが高ければ SPC モデルの効果があるとはいえない。

$$\text{Sentence Coverage}(\%) = (1 - \frac{\text{the number of miss-prediction sentences}}{\text{the number of sentences}}) \times 100$$

- SPC 出現率

  総単語数に対する SPC が評価用データ内で出現する回数の割合

  SPC 出現率が高いことは、それだけ探索空間を絞りこむ機会が多いということなので、認識における処理量を小さくできることを意味する。

$$\text{SPC rate}(\%) = (\frac{\text{the number of SPC}}{\text{the number of words}}) \times 100$$

2.7.3　実験結果及び考察

$n=2$ と $n=3$ のSPCモデルの、出現頻度の閾値 α と文カバレッジの関係を図2.5及び図2.6に、出現頻度の閾値 α とSPC出現率の関係を図2.7及び図2.8にそれぞれ示す。

これらの結果から次のような考察がなされる。

$n=2$ のSPCモデルの評価では、ほとんどの場合において文カバレッジについて90%以上を達成している。また半分以上の $n=3$ のSPCモデルでは、80%以上の文カバレッジを得ている。ここで、文カバレッジが90%であるということは、SPCモデルによって次にくる単語を予測する際に、予測を失敗する確率が10%だということである。すなわち、SPCモデルを使用しないで認識を行なったときの認識率が85%であれば、SPCモデルを使用することで最悪の場合76.5%以下まで認識率が低下することを意味する。このことから、SPCを用いることにより処理量を削減できる一方で、認識率が低下する恐れがあることがわかる。もっと文カバレッジを高くすることができれば、次にくる単語の予測失敗が少なくなり、SPCモデルは更に有効な手段になると考えられる。

また図2.5と図2.7、及び図2.6と図2.8を見ると、文カバレッジが高くなるとSPC出現率が低くなるというトレードオフの関係があることがわかる。また、SPC出現率は n-gram確率の累積個数である β を多くした方が高くなることがわかる。しかし、β を大きくとることは文脈をSPCと決定付けるための制約をそれだけ緩くするということなので、さほど単語予測に信頼のおけない文脈もSPCとなってしまう。そのため、認識時の単語予測失敗が増えてしまい、認識精度が下がってしまうことが推測される。従って、できるだけ小さな β で、高いSPC出現率を達成できることが望ましい。

また、同程度の文カバレッジを達成したときのSPC出現率の値は、ほとんどの条件下において $n=3$ のSPCモデルの方が高いことがわかった。これは文脈を長くすることで、次にくる単語の種類が少なくなりその文脈に続く確率が高くなるため、より信頼度の高い単語を予測できて予測失敗が少なくなるからである。

2.8　まとめ

次にくる単語候補を的確に予測できる文脈を求め、それを利用して探索空間を絞りこむ手法を提案した。そのような文脈をSPCとして定め、コーパスに基づく統計的な言語モデルである n-gramを基にして、選別法を考案した。また、SPC出現率と文カバレッジという2つの尺度を用いて、考案したSPC選別法に則して求めたSPCの単語予測に対する評価実験を行なった。この実験の結果から、文脈が長い $n=3$ のSPCモデルの方が $n=2$ のものよりも、次にくる単語の予測に信頼がおけることがわかった。

本章で行なった評価実験だけでは、実際の認識システムにおいてSPCモデルがどの程度

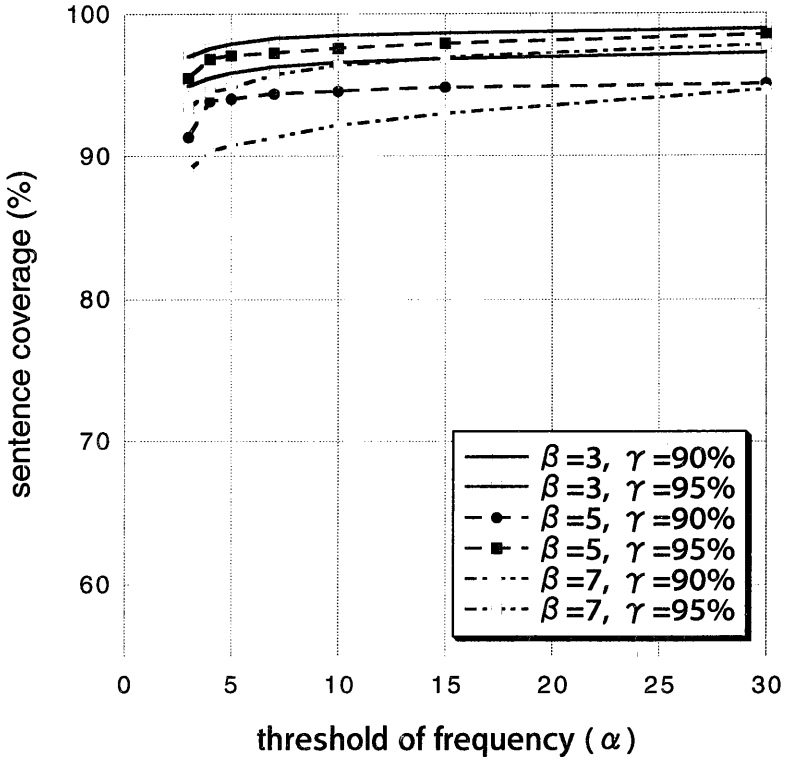


図 2.5: 出現頻度の閾値と文カバレッジの関係 ($n = 2$)

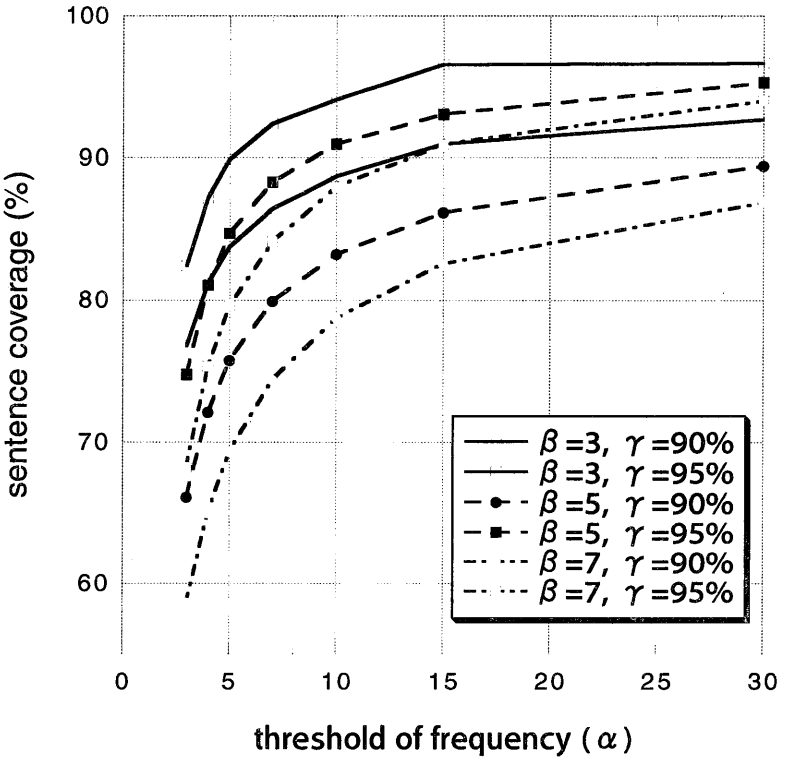


図 2.6: 出現頻度の閾値と文カバレッジの関係 ($n = 3$)

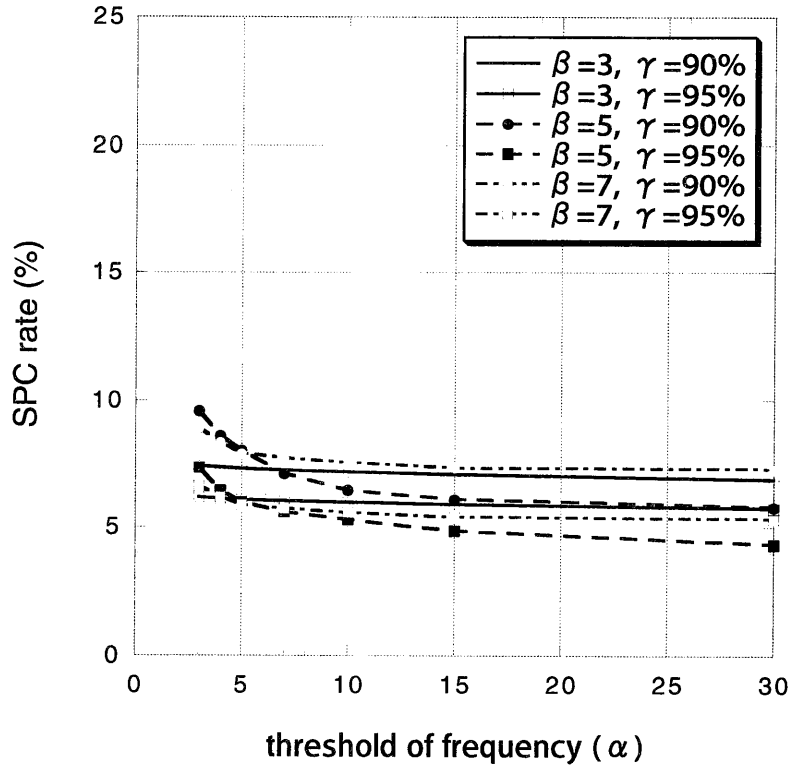


図 2.7: 出現頻度の閾値と SPC 出現率の関係 ($n = 2$)

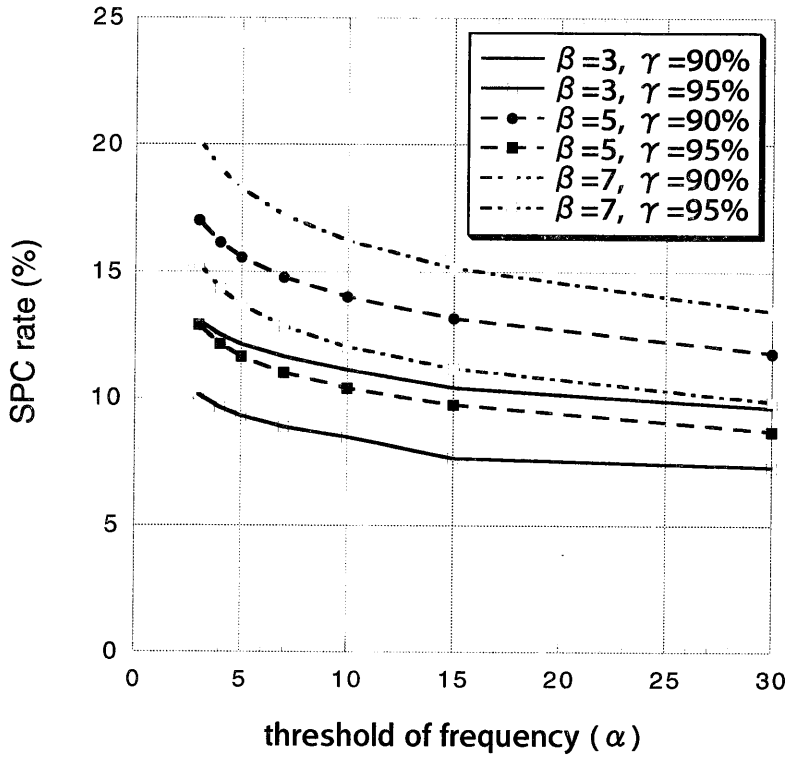


図 2.8: 出現頻度の閾値と SPC 出現率の関係 ($n = 3$)

有効性を示すかを測ることができない。従って、次の章ではSPCモデルを用いて実装した認識システムによる実験を行なう。

第3章

単語予測モデルを用いた認識システムの実装

3.1 はじめに

現在の大語彙連続音声認識システムの傾向として、認識処理を行なう際に第1パスで緩く探索空間を絞りこんだ後、第2パスでより高度な処理を行なう2パス処理が一般的となっている。このうち第1パスでの処理量が、認識における処理の大部分を占めていることは広く知られている。また、第1パスで枝刈りされた仮説が第2パスで結果として出力されることはありえない。従って、システムの性能向上を図るためには第1パスにおける処理の改善が不可欠である。

そこで本章では、前章で提案した単語予測モデルであるSPCモデルをそのような認識システムの第1パスに用いることを考えて認識システムを実装し、認識実験を行なう。

3.2 SPCモデルを用いた認識システムの実装

3.2.1 認識エンジン JULIUS

実際に音声認識を行なうために、SPCモデルを用いて連続音声認識システムを実装する。今回SPCモデルを用いる認識システムは、認識エンジン JULIUS [16][17] を基とする。

この認識エンジン JULIUSは、大語彙連続音声認識研究・開発の共通プラットフォームを開発するプロジェクトにより設計・作成されたものである。またこのプロジェクトは、情

表 3.1: 音素の一覧

a	i	u	e	o	a:	i:	u:	e:	o:	N	w	y
p	py	t	k	ky	b	by	d	dy	g	gy	ts	ch
m	my	n	ny	h	hy	f	s	sh	z	j	r	ry
q	sp	silB	silE									

報処理振興事業協会 (IPA) による「独創的情報技術育成事業」の支援を受けている。本章で使用する JULIUS は 1997 年度版のシステムで、基本語彙が 5,000 単語の連続音声認識を行なうものである。

以下に、JULIUS における各モデル及び各モジュールについて簡単に述べる [16]。

音響モデル

音響モデルは、混合連続分布 HMM(対角共分散) [6] に基づいている。また、音素環境独立 (monophone) モデルから音素環境依存 (triphone) モデルまでのさまざまな日本語の音響モデルを構築している。

この音響モデルで採用している日本語の 43 音素の一覧を表 3.1 に示す。この音素表記は、日本音響学会 (ASJ) の音声データベース委員会で策定されたものに基づいている。表中で、a: – o: は長母音を、q は促音を表わす。ポーズに関しては、silB、silE、sp の 3 種類のモデルを用いることとする。これらはそれぞれ、文頭、文末、文中 (単語間) のポーズに対応する。

音響モデルの学習には、日本音響学会の音素バランス文からなる研究用連続音声データベース (ASJ-PB) の全部と、毎日新聞記事読み上げ音声コーパス (ASJ-JNAS)[18] のうち 100 名分を利用した。これらは合計で、男女とも約 130 名の話者による 2 万文のデータである。

単語辞書

単語辞書は、音響モデルと言語モデルの両方と整合性がとれたものである。つまり、単語辞書に含まれる音素記号はすべて音響モデルに内包されており、かつ単語辞書に含まれる語彙は言語モデルに含まれる語彙と一致している。基本語彙は毎日新聞記事 CD-ROM [13][14] の、1991 年 1 月から 1994 年 9 月までの 45ヶ月分において、出現頻度上位の 5,000 単語 (ここでは単語とは形態素のことを指す) から構成される。

また日本語は英語等と異なり、文中で単語区切りがなされていないために、形態素解析が必要である。形態素解析を人手で行なうと時間やコストがかかる上に再現性が乏しくな

るため、形態素解析ツールによって処理されることが多い。認識エンジン JULIUS では、新情報処理開発機構の新聞記事タグデータ (RWCP テキストデータベース) に基づいている。

言語モデル

設定した基本語彙に基づいて、単語 n-gram モデルを構築する。今回のシステムにおける基本語彙は、学習データ中の出現頻度上位 5,000 単語である。言語モデルとして使用する単語 bigram と単語 trigram は、CMU-Cambridge SLM ツールキット [15] により作成した。従って、単語辞書に含まれない単語は未知語として扱われ、更には back-off 平滑化によりスムージング処理が行なわれている。単語 n-gram の学習用コーパスは、単語辞書と同様、毎日新聞記事 CD-ROM の 1991 年 1 月から 1994 年 9 月までの 45ヶ月分の記事データを使用した。

木構造辞書

単語を認識する際に用いる辞書として、root node から枝でつながっている複数の状態へ遷移し、単語を示す leaf node までをたどるような木構造を成す辞書を扱う。これを木構造辞書という。ここで、単語間の遷移数は語彙数の 2 乗となる。このとき、各枝における遷移確率は枝に対応する音素間の確率である。また、単語間の遷移には n-gram 確率を付与する。木構造辞書では、1 つの枝に 1 つの HMM が対応しており、単語の最初の部分で共通の音素を持つ単語は同じ枝を共有して遷移することで状態数を削減している。この効果は語彙が大きくなるにつれ顕著になり、大語彙連続音声認識では不可欠である。また、この木構造辞書では、単語の終端すなわち leaf node まで達しないと単語を同定できないため、n-gram の確率を静的に埋めこむことはできない。そこで、leaf node に達する毎に n-gram の確率を動的に加えるものとする。

デコーダ

認識エンジン JULIUS は、前述の音響モデルと言語モデルのインタフェースがとれるように開発された連続音声認識システムである。JULIUS は 2 パス探索を行ない、第 1 パスで単語 bigram、第 2 パスで単語 trigram を用いる。第 1 パスでは、木構造辞書に bigram 確率を動的に割り当てながらスコアを求め、フレーム同期ビームサーチを実行する。またここで、フレーム毎にスコアについてヒープソートを行なうことで、スコア上位の単語仮説を求めている。bigram 確率は、最尤の単語仮説と前の部分を共有する単語に応じて、木のすべてのノードに分配する。第 2 パスにおいては、言語モデルとして単語 trigram を用いると同時に、単語間の音素環境依存性 (Cotext-Dependent : CD) の処理を行なうことで、より高精度の認識を実現している。

以上の各モジュールを結合して、日本語ディクテーションシステムを設計・実装した。システムのブロック図を図3.1に示す。

デコーダの仕様に基づいて、音響モデルと言語モデルが統合されている。第1パスでは単語bigramを利用し、音素環境依存性(CD)の処理は単語内のみに限られている。より高精度で計算量の大きい単語trigramと単語間の音素環境依存性(CD)は、第1パスで絞られた候補を再探索・再評価する第2パスで適用される。

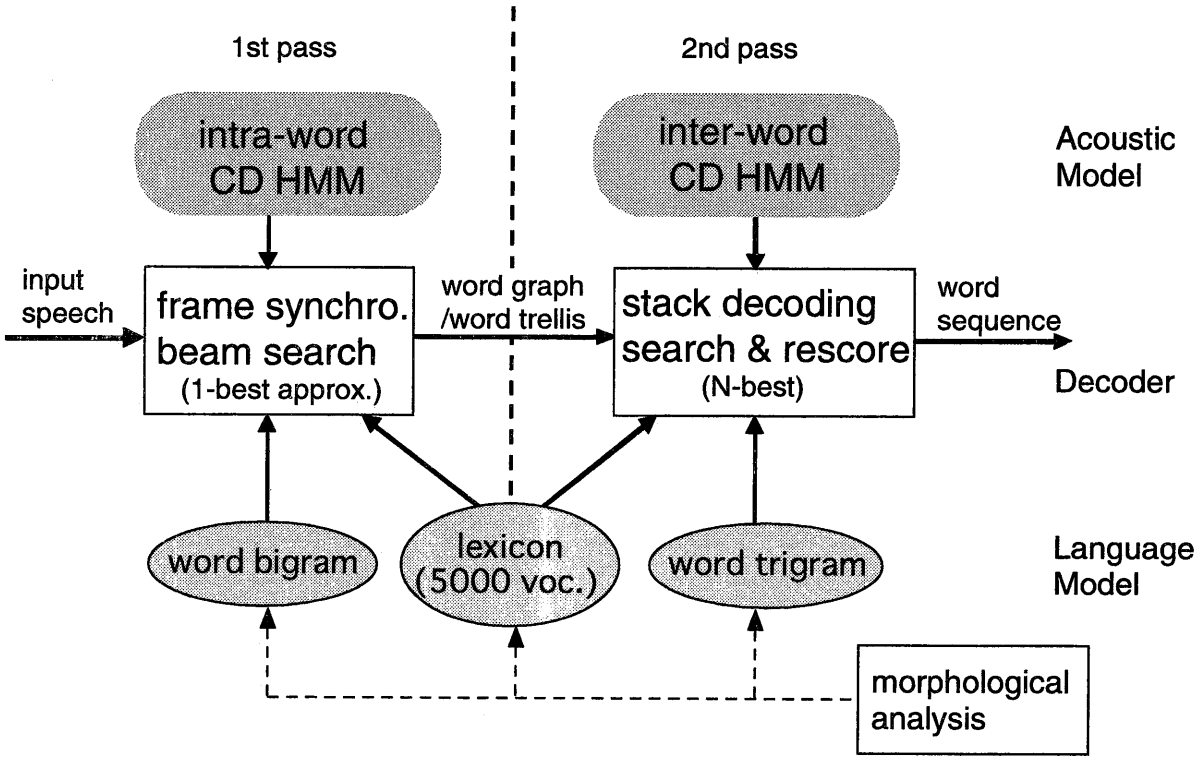


図 3.1: 認識エンジン JULIUS の構成

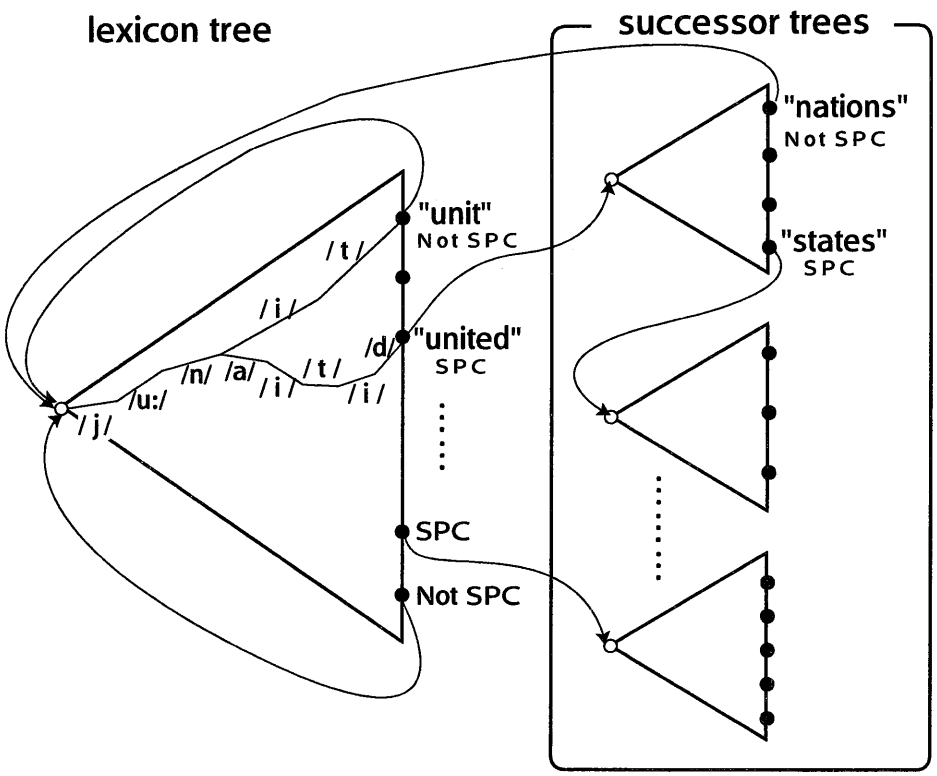


図 3.2: SPC モデルによる木構造辞書の拡張

3.2.2　木構造辞書の拡張

前節で述べたように、JULIUS で用いている n-gram はスムージングされたものである。従って、その単語 n-gram を扱うために、木構造辞書では木の root node から基本語彙の単語全てに対して遷移するように枝を配している。しかし n-gram の文脈が SPC であるならば、基本語彙の全単語に枝が出ている root node に遷移する必要はなく、次にくると予測される単語候補のみに遷移すればよい。これにより、全基本語彙との連接を考慮するよりも、格段に少ない処理量で認識を実行することが可能になると考えられる。本章では、基本語彙は JULIUS の単語辞書に含まれる 5,000 単語を示す。SPC モデルを連続音声認識システムに組み込む際の、木構造辞書の拡張法を図 3.2 に示す。

図 3.2 における三角形は全て木構造辞書を表している。また lexicon tree 及び successor tree [10] の各 tree は、1 つの枝が HMM の状態遷移に対応した木構造辞書である。ここに示す lexicon tree とは、基本語彙の 5,000 単語だけからなる木構造辞書を意味する。また successor tree は、SPC における予測単語候補を展開するための木構造辞書である。すなわち各 successor tree はそれぞれ 1 つの SPC のみに対応しており、その SPC の予測単語候補だけから構成されている。予測単語候補内に基本語彙に含まれない単語が存在した場合、

単語辞書にその単語を付加した上でsuccessor treeに内包する。どのように単語を付加するかについては、3.2.3節で説明する。

認識処理の段階で、単語仮説が完成した時点での文脈がSPCであれば、そのsuccessor treeに遷移することでSPCモデルの利用による単語の絞りこみが行なわれるところが、JULIUS等の従来のデコーダと異なる点である。またSPC以外の文脈の場合は、通常通りにスムージング処理されたn-gramにより確率を与え、基本語彙内の全ての単語との接続を仮定しながら処理を行なうものとする。従って、2.7節で求めたSPC出現率が高いSPCモデルほど、処理量を多く削減することが可能となる。

図3.2を例に挙げて、$n = 2$のSPCモデルの場合について詳しく説明する。この例では文脈“united”がSPCで、その予測単語候補として“states”や“nations”など複数の単語が存在している。従って、認識時に展開された単語仮説における文脈が“united”であれば、その文脈はSPCなので対応するsuccessor treeへと遷移する。それによって、lexicon treeに遷移したときよりも単語仮説の後に展開する枝数が削減される。もし単語仮説から“states”へと遷移したなら、文脈はまたSPCなので対応するsuccessor treeへと遷移する。一方、単語仮説における文脈が“nations”であるならば、SPCではないのでlexicon treeのroot nodeへと遷移し、通常のスムージングされたn-gramを利用して全単語とのつながりを考慮する。

このような木構造辞書の拡張により、認識システムでSPCモデルを用いることができ、処理量を削減することが可能となると考える。

3.2.3　SPCモデルの作成及び単語の付加

システムを実装する際に用いるSPCモデルの形態素解析は、JULIUSに標準添付されている言語モデル及び単語辞書におけるものと、可能な限り同じとなるようにした。SPCモデルの作成に関する一連の流れを図3.3に示す。

初めにRWCPテキストデータベースを用いて、毎日新聞記事CD-ROMの形態素解析を行なう。次に、形態素解析した文からCMU-Cambridge SLMツールキットによって、文中に出現する全てのn-gramを構築する。構築したn-gramから、SPC選別法に基づいてSPC及びその予測単語候補を選別し、SPCモデルを作成する。

次の段階として、作成したSPCモデルにおける基本語彙に含まれない単語に関しては、読み及び音素表記を付与する必要がある。今回は、音素表記を付与する前処理の段階でjumanとkakasiを共に使用した。jumanは日本語の形態素解析を行なうと共に読みも与えるツールであり、kakasiとは漢字かなまじり文をひらがな文やローマ字文に変換することを目的とする解析ツールである。

まず、基本語彙に含まれない単語に対してjumanを用いることで、単語の読みを得る。更に得られた仮名表記の読みを、kakasiによって処理することによって、単語をローマ字

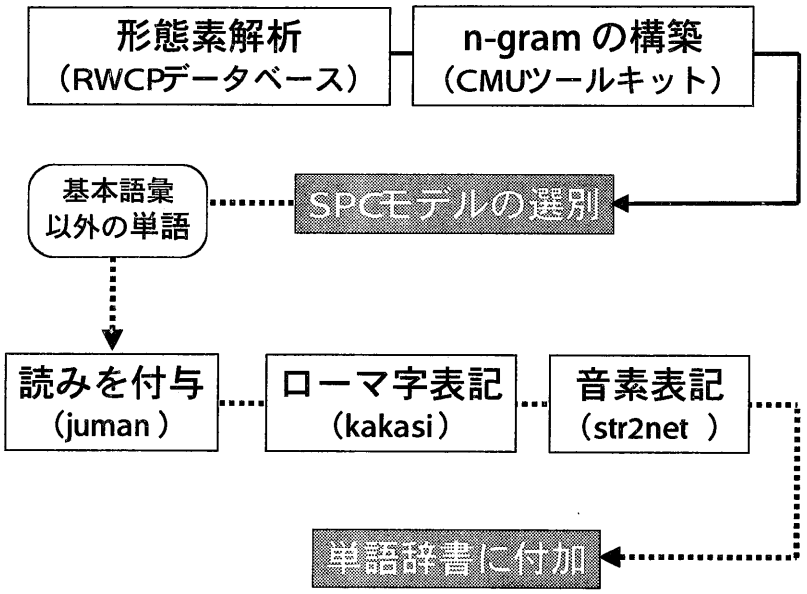


図 3.3: SPC モデル作成における一連の流れ

表記に変換する。最後に JULIUS に添付されているプログラムを用いて、ローマ字表記を音素表記に直すことによって、基本語彙に含まれない単語の音素表記を作成した。

但し、ここで juman の辞書に含まれない単語をもつものについては、SPC モデルから排除している。

3.3　SPC モデルを用いて実装した認識システムによる認識実験

3.3.1　実験条件

今回の認識実験では、前章で述べた SPC 選別法におけるパラメータの値を $\alpha = 7$、$\beta = 5$、$\gamma = 90\%$ として SPC モデルを求めた。そしてそのモデルを用いて連続音声認識システムを実装した。このパラメータは、2.7 節による評価実験の結果における文カバレッジと SPC 出現率の値から定めた。実際には、文カバレッジが高い値を出しながらグラフの傾斜が緩やかとなり、かつ SPC 出現率ができるだけ高くなるようにパラメータを選んだ。ここで計算量の問題から、予測には使用されないことが明らかな SPC モデルは除外している。SPC モデルについての詳細を表 3.2 に示す。

今回の実験で用いる認識システムは、認識システム JULIUS の第 1 パス処理部分に対して 3.2.2 節で提案した木構造辞書の拡張を行なって、SPC モデルを実装したものである。

実験条件を表 3.3 に示す。JULIUS の基本語彙の単語数は標準実装のままの 5,000 単語と

表 3.2: SPC モデルの情報 (基本語彙 5,000 単語)

	$n=2$	$n=3$
SPC の種類	763	47,663
単語辞書に加えた単語数	1,743	4,818
平均予測単語候補数	12.1	2.4

表 3.3: 認識実験の条件 (基本語彙 5,000 単語)

音響モデル	triphone (状態数 2000 混合数 16)
サンプリング周波数	16kHz
フレーム周期	10ms
ハミング窓長	25ms
フィルタバンク	24 チャネル
特徴パラメータ	MFCC(12 次)+Δ MFCC+Δ Pow (計 25 次)
話者	男性 10 名, 女性 10 名

している。

評価尺度としては Word Accuracy 及び Word Correct Rate を用いた。各評価尺度の定義式は以下の式の通りである。

$$\text{Word Accuracy } (\%) = \frac{N-S-I-D}{N}$$

$$\text{Word Correct Rate}(\%) = \frac{N-S-D}{N}$$

N : 単語総数
S : 置換された単語数
D : 削除された単語数
I : 挿入された単語数

今回の実験においては、単語の読みだけを用いて認識結果が正解であるか否かを判定した。すなわち、品詞や漢字表記が異なっていても、読みさえ正解文の単語と合致していればその単語は正しく認識されたものとする。

対象とする音声資料は、毎日新聞記事読み上げ音声コーパス JNAS の中の、出現頻度上位の 5000 語で閉じた文のセット MID、同 5000 語で未知語が 1 つの文のセット MID+、及び高頻度の 20,000 語で閉じた文のセット LARGE、同 20,000 語で未知語が 1 つの文のセット LARGE+、同 20,000 語で未知語が 2 つ以上の文のセット LARGE++とした。

表 3.4: 基本語彙 5,000 単語の認識実験の結果 (MID, MID+)

		Word Accuracy(%)		Word Correct Rate(%)	
speaker	the method	MID	MID+	MID	MID+
男性	bigram	72.5	63.2	73.3	65.7
	SPC モデル ($n = 2$)	73.9	63.3	74.7	65.9
	SPC モデル ($n = 3$)	74.0	63.8	74.8	66.7
女性	bigram	71.7	61.4	72.3	64.3
	SPC モデル ($n = 2$)	73.7	62.4	74.4	64.9
	SPC モデル ($n = 3$)	73.4	62.4	74.1	65.2

表 3.5: 基本語彙 5,000 単語の認識実験の結果 (LAR, LAR+, LAR++)

		Word Accuracy(%)			Word Correct Rate(%)		
speaker	the method	LAR	LAR+	LAR++	LAR	LAR+	LAR++
男性	bigram	54.8	47.1	45.1	60.0	55.5	54.0
	SPC モデル ($n = 2$)	54.9	49.7	46.6	60.3	57.3	55.2
	SPC モデル ($n = 3$)	57.2	49.8	47.4	62.0	57.0	55.4
女性	bigram	54.6	48.5	44.0	59.3	55.0	52.7
	SPC モデル ($n = 2$)	54.9	49.5	44.4	59.9	56.1	52.9
	SPC モデル ($n = 3$)	55.5	50.6	45.0	60.0	56.7	53.0

3.3.2 実験結果と考察

出現頻度上位 5,000 単語の MID 及び MID+における実験の結果を表 3.4 に、出現頻度上位 20,000 単語の LARGE、LARGE+及び LARGE++における実験の結果を表 3.5 にそれぞれ示す。表 3.5 中では、LARGE を略して LAR と記している。また各表中における bigram は、JULIUS が第 1 パスで使用している言語モデルである。

出現頻度上位 5,000 単語からなる MID 及び MID+について表 3.4 から、$n = 2$ の SPC モデルを導入することによって、Word Accuracy が男性話者においては最大 1.4%、女性話者では最大 2.0%向上していることがわかる。$n = 3$ の SPC モデルでは、男性話者においては最大 1.5%、女性話者では最大 1.7%向上している。また同様に表 3.5 から、出現頻度上位 20,000 単語からなる LARGE、LARGE+及び LARGE++は、$n = 2$ の SPC モデルを導入することによって Word Accuracy が男性話者において最大 2.6%、女性話者では最大 1.0%の向上が見られる。$n = 3$ の SPC モデルでは、男性話者においては最大 2.7%、女性話者では最大 2.1%の向上が見られる。この結果から、$n = 3$ の SPC モデルの方が $n = 2$ のモデルよ

表 3.6: 基本語彙に含まれない単語が予測された割合 (%)

	n	MID	MID+	LARGE	LARGE+	LARGE++
男性	2	0.5	0.5	0.8	0.8	1.4
	3	0.0	1.5	0.9	1.9	3.3
女性	2	0.5	0.0	1.6	1.2	0.1
	3	0.4	1.1	1.6	2.0	3.8

りも、次にくる単語の予測において信頼性が高いといえる。

また、Word Accuracy よりも Word Correct Rate の方が、bigram による結果と SPC モデルによる結果の差が小さくなっている。このことから、SPC モデルを用いた方が認識結果で挿入された単語が多いことがわかる。これは、bigram では誤って複数の単語として認識してしまうような単語を、SPC モデルを使用することで正確に予測できるために、挿入される単語が減少したからであろう。

更に、2.6.2 節で述べた基本語彙に含まれない単語を認識できるという効果が表れているか調べた。SPC によって次にくると予測された単語のうち、基本語彙に含まれなかった単語の割合を表 3.6 に示す。この割合は以下のような式を用いて算出した。

$$\text{基本語彙に含まれなかった単語の割合 (\%)} = \left(\frac{\text{SPC モデルで予測された基本語彙に含まれない単語数}}{\text{SPC モデルによって予測された単語総数}}\right) \times 100$$

表 3.6 から、若干ではあるが基本語彙に含まれない単語も予測されていることがわかる。

次に、SPC モデルを使用した実際の認識において、bigram を用いた場合と異なる結果についての例を図 3.4 に示す。図中の文において下線を付した部分が SPC である。

(ex.1) を見ると、SPC モデルを使用することによって、単なる bigram を用いた際に結果として出力されてしまった、正文としてはあり得ない単語列が排除されていることがわかる。また、(ex.2) で SPC である “私” によって “ども” が予測されているが、この “ども” は認識システムの基本語彙にはない単語である。この例と表 3.6 より、SPC モデルを使うことで、若干ではあるが確かに語彙を拡張することが可能であることが確認できた。

(ex.3) は改悪例を示したもので、SPC である “が／私” の予測単語候補に “以外” が含まれていなかったために正しい結果が得られなかった。これは SPC モデルの文カバレッジが 100%でないため、当然起こりうる結果である。このような結果を減らすためには、SPC モデルによる更なる文カバレッジの向上が必要である。

この実験によって、同じビーム幅を用いた場合には認識の精度は SPC モデルを使用することにより向上することがわかった。そこで、次にビーム幅を減少させて認識を行ない、実際に処理量をどの程度削減できるかについて調べる実験を行なった。

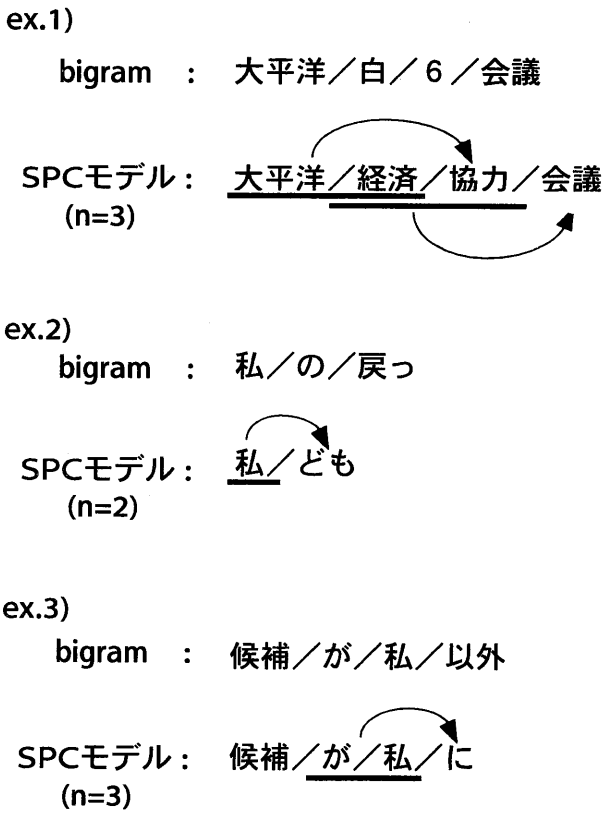


図 3.4: SPC による単語予測の実例

表 3.7: ビーム幅の値

ビーム幅	500, 600, 700

3.4 ビーム幅減少による処理量削減に関する実験 (1)

SPC モデルがもたらす効果として 2.6.2 節で述べた通り、ビーム幅を減少しても従来の n-gram と同程度の精度を得られることが期待できる。そこで、実際にビーム幅を減少させてその効果を調べた。

3.4.1 実験条件

ビーム幅を表 3.7 のように変化させて、認識実験を行なった。実験に使用した SPC モデルは、3.3 節の認識実験で用いたものと同様のモデルである。またその他の実験条件も表 3.3 と同様の条件を用いている。但し評価尺度は、今回は Word Accuracy のみを採用した。認識対象とする音声資料は、含まれる単語が出現頻度上位 5,000 単語の MID と 20,000 単語の LARGE を用いた。

3.4.2 実験結果と考察

男性話者と女性話者の MID における実験の結果を図 3.5 と図 3.6 に、また LARGE における実験の結果を図 3.7 及び図 3.8 にそれぞれ示す。

図 3.5 から図 3.8 のグラフより、SPC モデルを用いることによってビーム幅を上位 700 仮説から 500 仮説にしても、ビーム幅が上位 700 仮説の従来手法の bigram と同程度以上の Word Accuracy を実現することがわかった。この結果から、bigram と同程度の認識精度を得るときには、SPC モデルを利用することで認識時における探索空間を 70%以上に削減できるといえる。

また、$n = 2$ の SPC モデルを用いた場合の Word Accuracy の変化に比べ、$n = 3$ の SPC モデルを使用したときの方が小さいことが、これらのグラフから見てとれる。このことは、$n = 3$ の SPC モデルの方が、ビーム幅の削減による影響を受けにくいことを意味している。従って、SPC モデルは高次の n-gram の制約を有効に利用できていると考えられる。

3.5 まとめ

SPC モデルを用いた認識システムを実装するために、木構造辞書の拡張手法を提案した。その手法に基づいて、前章の SPC 選別法によって作成した SPC モデルを用いて大語彙連

続音声認識システムを実装した。その後、実際に実装した認識システムを用いて認識実験を行なった。実験の結果から、あり得ない単語列をSPCモデルによって排除でき、かつ認識できる単語を増やすことができるという有効性が示された。またビーム幅を変化させた認識実験の結果から、SPCモデルによって処理量の削減を実現できることがわかった。

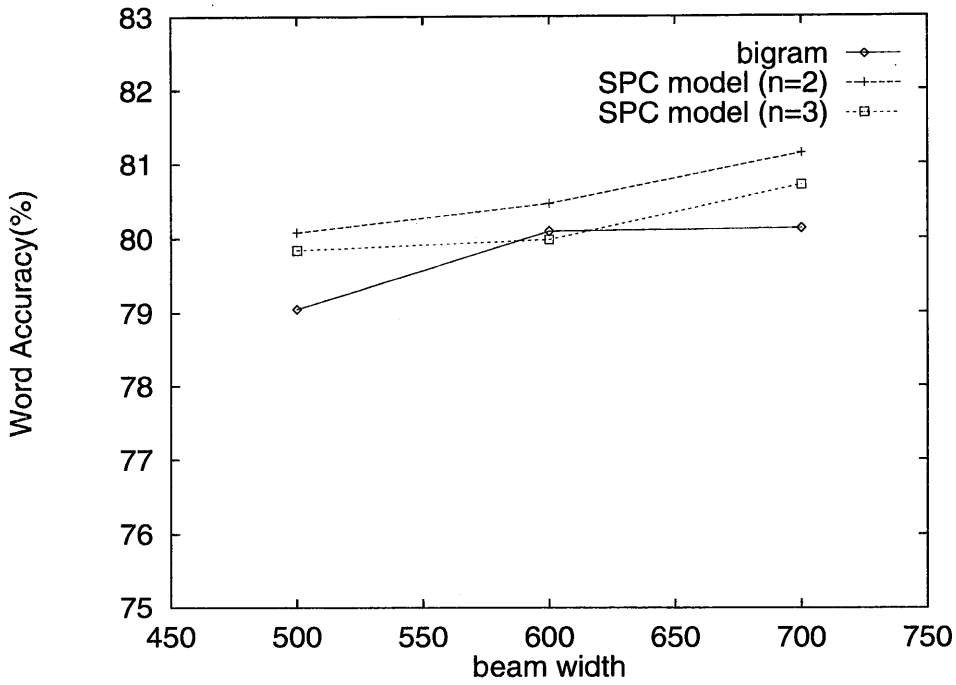


図 3.5: ビーム幅の変化による基本語彙 5,000 単語の認識精度の推移 (男性, MID)

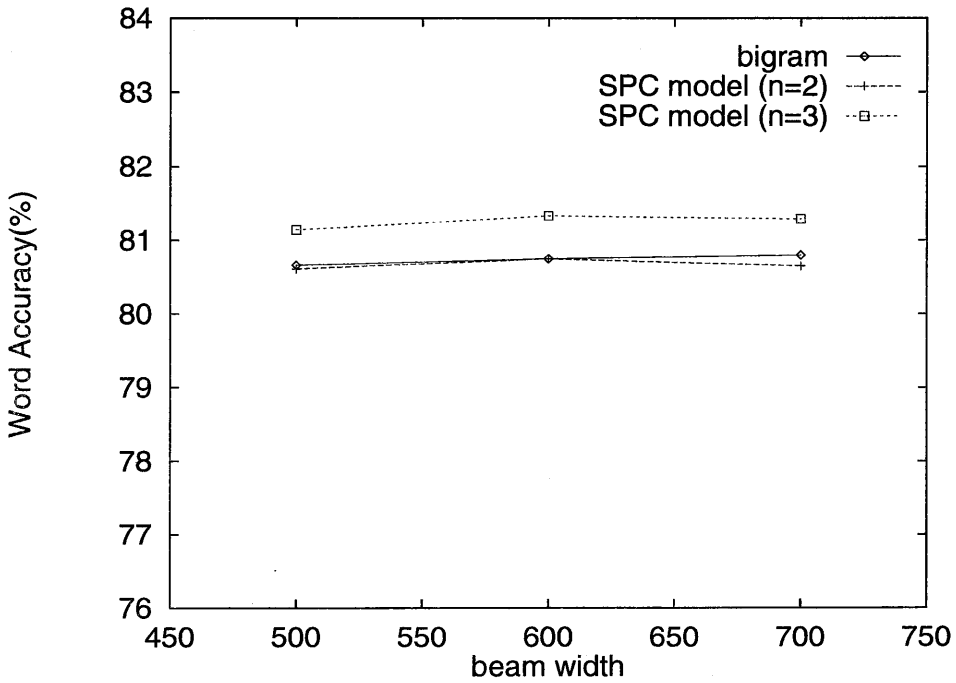


図 3.6: ビーム幅の変化による基本語彙 5,000 単語の認識精度の推移 (女性, MID)

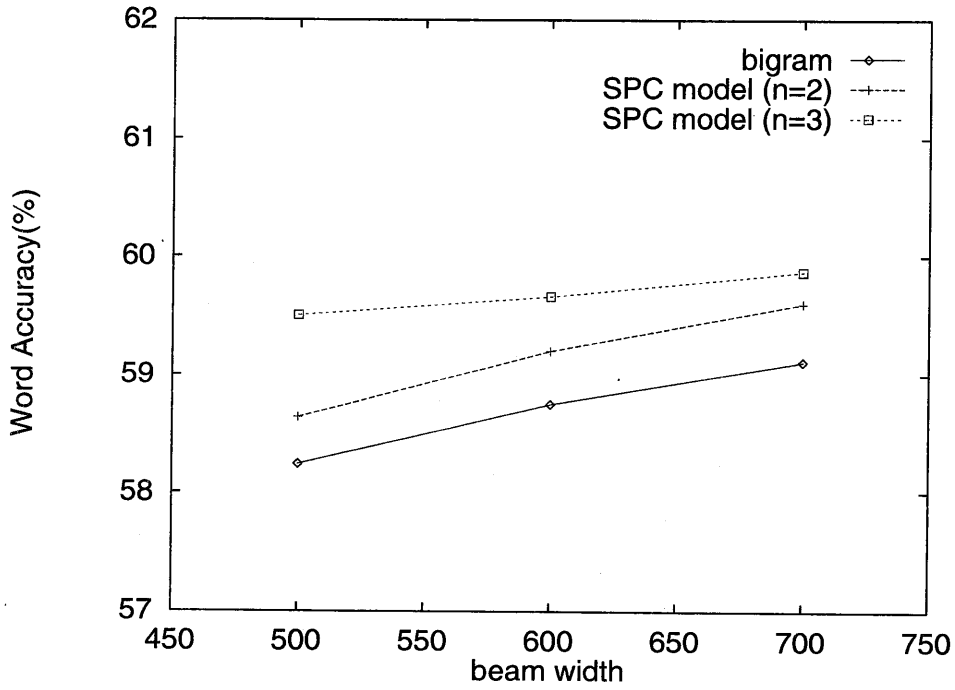


図 3.7: ビーム幅の変化による基本語彙 5,000 単語の認識精度の推移 (男性, LARGE)

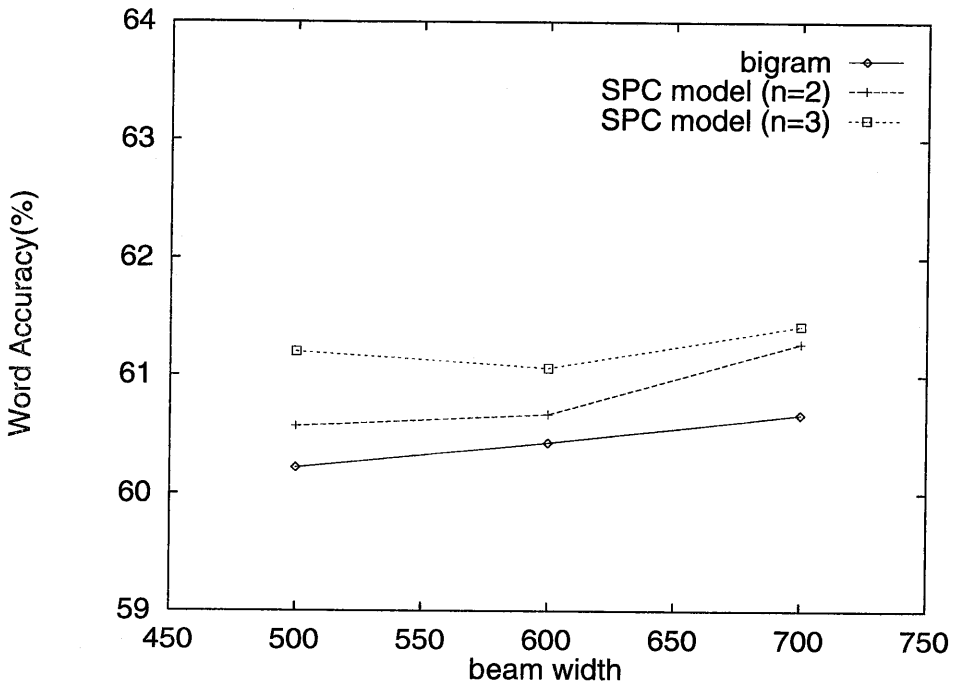


図 3.8: ビーム幅の変化による基本語彙 5,000 単語の認識精度の推移 (女性, LARGE)

第4章

大語彙連続音声認識における単語予測

4.1 はじめに

前章では、考案したSPCモデルを用いて基本語彙5,000単語の連続音声認識システムを実装し、その効果を調べた。結果として、SPCモデルを用いることによって認識精度を向上させることができ、また同程度の認識精度を保ちながら処理量をおよそ71%にまで削減することが可能であった。

しかしこれまで行なってきた形態素解析は、音声認識に適したものではなかった。そこで本章では、音声認識に対して最適化された形態素解析を用いてSPCモデルを作成する。また、SPCモデルを更に大きな語彙を扱う連続音声認識に用いた場合に生じうる問題点を分析し、その解決法について述べる。そしてそれらを踏まえて、新たなSPCモデルを用いた連続音声認識システムを実装し、認識実験を行なう。

4.2 大語彙連続音声認識におけるSPCモデルの問題点

3.2.2節で、successor treeを利用する木構造辞書の拡張法を提案したが、この手法では同じ単語をlexicon tree及び各successor treeにおいて何回も異なる木構造辞書の中に含まなければならない。すなわち辞書木が1つの場合に比べて、重複している分だけ乗算的にノード数を必要とすることとなる。従って基本語彙数が増えることで各辞書木での単語の重複も増加し、結果としてノード総数が膨大な数となってしまう。このことは、実際に大語彙の認識システムでSPCモデルを利用する際には大きな問題となる。

また、2.6.2節でSPCモデルが認識に与える効果として、“基本語彙に含まれない単語で

もSPCの予測単語候補として存在させることができ、認識が可能となる" 点を挙げた。これについては、基本単語を大語彙にすることによって、未知語が少なくなるためにあまり効果が出なくなると推測される。

更に3.2.1節で述べたように、日本語を対象として音声認識を行なうためには形態素解析が必要不可欠である。特に大語彙連続音声認識では、形態素による認識への影響が大きくなるので、形態素解析の最適化を行なう必要がある。

4.3 認識システムの実装における手法の改善

本章ではSPCモデルを用いて実装する新たな大語彙認識システムとして、IPAによって作成された1998年度版認識エンジンJULIUS[19]を基に扱う。以下、前章のJULIUSと混同を避けるため、大語彙認識システムのJULIUSは98年度版JULIUSと記す。98年度版JULIUSは基本語彙20,000語の大語彙連続音声認識エンジンで、システムの構成については基本語彙数の他は図3.1と変わりはない。語彙数の増加及びシステム細部の変更に伴う木構造辞書の構築、及び形態素解析の最適化におけるSPCモデルの作成の点において改善・変更を要する。以下にそれぞれの変更手法を述べる。

4.3.1 拡張lexicon tree

前節で述べた通り、3.2.2節で提案した木構造辞書の拡張法では、同じ単語を何回も異なる木構造辞書の中に含まなければならない。実際、3.3節で使用した$n = 3$のSPCモデルでは、lexicon treeのみのときは約85,000なのに比べて、全てのsuccessor treeとlexicon treeのノードを足した総数が、およそ1,500,000と非常に増大している。

これでは単語辞書が数万のものを扱うとき、複数の辞書木で重複して含まれる単語の数が増えてしまい、ノード総数が更に増加することが予想される。従ってこの手法を用いる場合、語彙数が少ないシステムならば問題は特にないが、大語彙を扱う際には要するメモリの量が膨大になり不向きであると考えられる。

そこで、SPCモデルを大語彙の認識システムに組み込むために、木構造辞書の構成を変更した。その木構造辞書を拡張lexicon treeと名付け、図4.1に示す。

これは、各SPCに対応してその予測単語候補のみからなるsuccessor treeを個々に作成するのではなく、全ての単語を包含する1つの大きな木構造辞書を作成するというものである。すなわち、lexicon treeにSPCモデルにおける基本語彙に含まれない単語も全て付加して、新たなlexicon treeを作成する。このlexicon treeを以下では拡張lexicon treeと呼ぶ。拡張lexicon treeを用いてSPCモデルによる遷移数の削減を行なうために、認識時において文脈がSPCとなったならば、そのSPCの予測単語候補にのびている枝のみを辿るように処理を行なう。

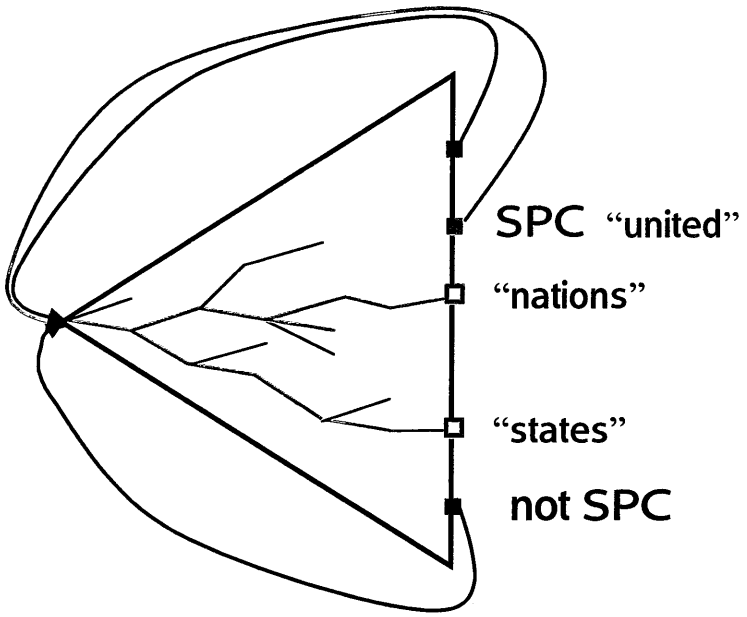


図 4.1: 拡張 lexicon tree

図 4.1 に示す $n = 2$ の SPC モデルを例に挙げて、拡張 lexicon tree による SPC モデルの処理方法を説明する。図 4.1 では SPC として "united"、その予測単語候補として "nations" 及び "states" が存在する。認識時に展開された単語仮説での文脈が SPC である "united" となったならば、拡張 lexicon tree の root node に遷移し、その後 "united" の予測単語候補を leaf node にもつ赤い枝に対してのみ展開していく。文脈が SPC でなければ、root node に遷移した後、従来のスムージングされた bigram として基本語彙に含まれる全ての単語につながる枝を展開する。ここで文脈が SPC でないならば基本語彙に含まれない単語には遷移させない理由は、スムージングした bigram が基本語彙を基盤に作成されたものであるためである。

4.3.2 形態素解析の最適化

日本語において、形態素解析が必要不可欠であることはすでに述べた。特に大語彙連続音声認識では形態素による認識への影響が大きくなるので、形態素解析の最適化を行なう必要がある [20][21][22]。また読みにおいても、「言う (いう・ゆう)」や「体育 (たいいく・たいく)」のように文法的な読みと人間が発音する音とは異なることがある。このような場合、文法的な読みを与えただけでは実際の音声認識では正しい結果を得ることはできない。更に、「京都」に対して「きょうと」と「きょーと」のどちらの読みを使用するかによっても、認識結果は変わることとなる。

これらの問題を解消するよう、98 年度版 JULIUS では "形態素解析サブプログラム ChaSen [23][24]"

並びに “読み付与サブプログラム ChaWan [25] 及び PostProcess” が添付されている。本章における単語の形態素解析と読みの付与には、それらを利用した。各プログラムにおける処理は次の通りである。

ChaSen で使用する辞書は、3.2.3 節で用いた解析プログラム juman の辞書に修正を加えて IPA 品詞体系辞書としたものである。従ってこれらの解析ツールを使用することによって、前章までで用いた形態素解析よりも音声認識に適した形態素解析が行なわれる。但し、ChaSen の辞書に未登録のために形態素解析及び読みの付与が成されなかった単語に関しては、SPC モデルから排除した。

おおよその読みは ChaSen だけでも付与できるが、数詞及び助数詞などの読みの変化に対しては正しい読みを出力しない。すなわち、「1 個」という単語列の正しい読みは「いっこ」であるが、ChaSen の解析結果では「いちこ」、同様に「1999」に対しては「いちきゅうきゅうきゅう」となってしまう。ChaWan は、そのような ChaSen の読みに対して読み変化を与える処理をするプログラムである。更に PostProcess は、単語に付与する読みを統一すると共に、音便に対して読みの変化を与える処理を行なう。これらの読み変化を考慮するための処理は 3.2.3 節の作成時には行なっていないが、認識においては非常に重要な前処理である。

また SPC モデルを求める際に使用した単語 n-gram の作成には、3.2.3 節と同様に CMU-Cambridge SLM ツールキットを使用した。更に単語辞書に必要な単語の音素表記は、98 年度版 JULIUS に添付されている音声認識用辞書を作成するプログラムを用いて作成した。

4.4 認識に与える影響についての評価実験 (2)

形態素解析及び作成手法の違いによって SPC モデルが変わったので、SPC モデルによる単語予測が認識に与える影響を評価する実験を改めて行なった。

4.4.1 実験条件

評価実験には毎日新聞記事 CD-ROM [13][14] を使用した。学習用データには、その 1991 年 1 月から 1994 年 9 月までの 45ヶ月分の記事データを用いた。また、評価用データとして、同 1994 年 10 月から 12 月の 3ヶ月分を用意した。

SPC モデルを求めるためのパラメータ、すなわち n-gram の出現頻度の閾値 α、累積確率を求める際の上位からの個数 β 及び累積確率の閾値 $\gamma\%$ は表 4.1 の通りである。

評価尺度は 2.7 節と同じく、文カバレッジと SPC 出現率を用いる。

表 4.1: SPCモデルに対する評価実験の条件 (2)

出現頻度の閾値 α 回	3, 4, 5, 7, 10, 15, 30
累積確率を求める際の上位 β 個	3, 5, 7, 10
累積確率の閾値 γ%	90, 95

4.4.2 実験結果及び考察

$n = 2$ と $n = 3$ のSPCモデルの、出現頻度の閾値 α と文カバレッジの関係を図4.2及び図4.3に、出現頻度の閾値 α とSPC出現率の関係を図4.4と図4.5にそれぞれ示す。

各グラフの傾向は2.7節の結果と同じく、SPCの選別条件が厳しいほど文カバレッジは向上する一方、SPC出現率は低下している。このことは、単語予測の信頼度と処理量削減の度合のトレードオフ関係に帰着する。

同じ条件下における2.7節の結果と比較すると、文カバレッジは平均1%強向上している。またSPC出現率も、ほとんどの条件下で2.7節よりも高い値を結果として出している。このことから、今回行なった形態素解析の最適化は、SPCモデルの単語予測に対しても有効であると考えられる。

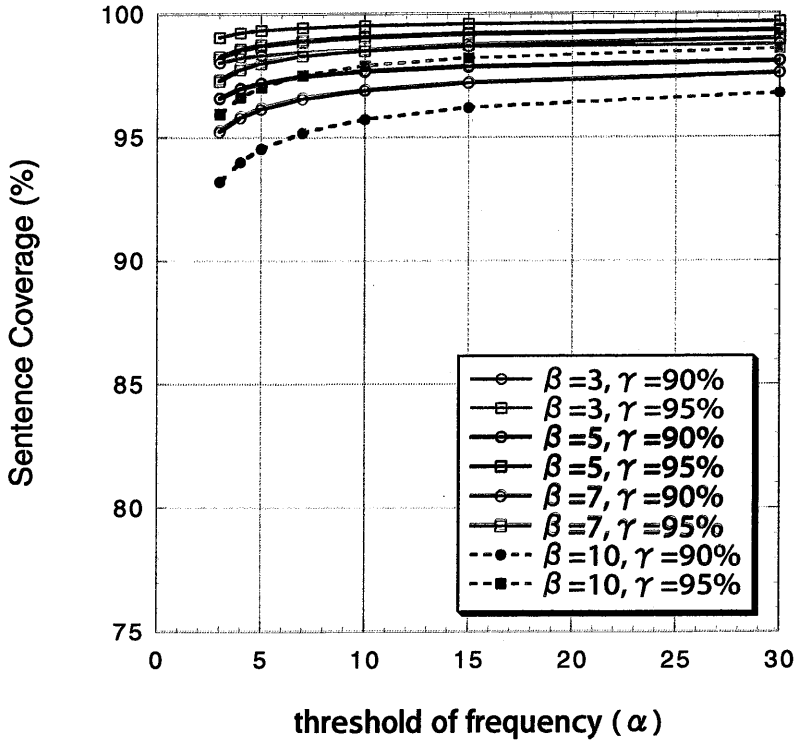


図 4.2: 出現頻度の閾値と文カバレッジの関係 ($n = 2$)

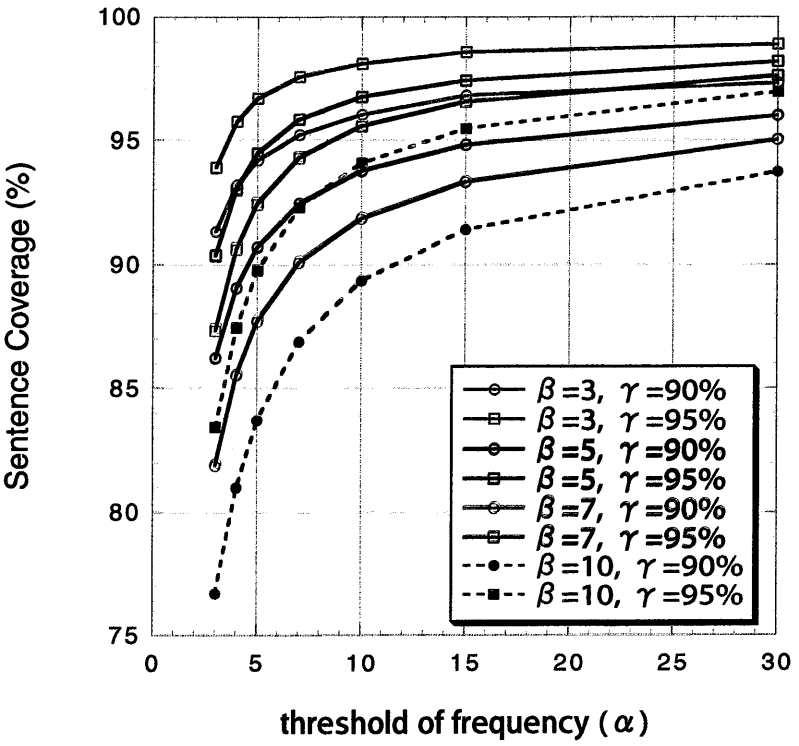


図 4.3: 出現頻度の閾値と文カバレッジの関係 ($n = 3$)

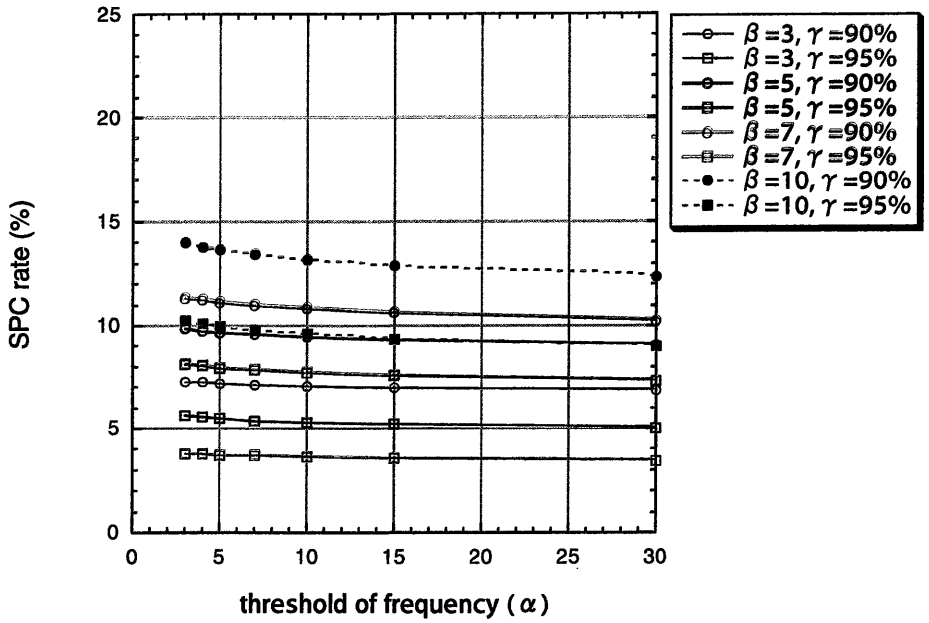


図 4.4: 出現頻度の閾値と SPC 出現率の関係 ($n = 2$)

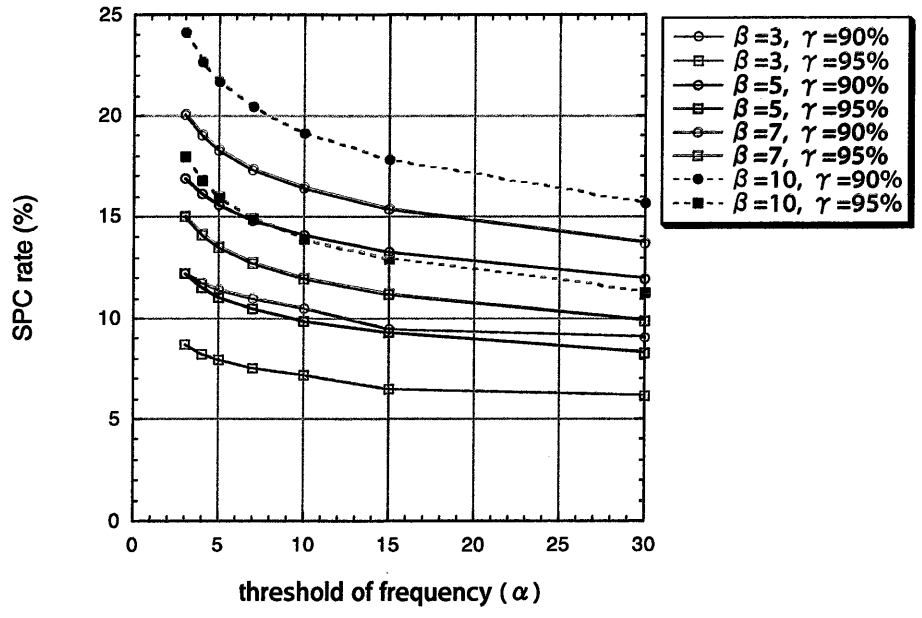


図 4.5: 出現頻度の閾値と SPC 出現率の関係 ($n = 3$)

表 4.2: 認識実験の条件 (基本語彙 20,000 単語)

音響モデル	triphone (状態数 2000 混合数 16)
サンプリング周波数	16kHz
フレーム周期	10ms
ハミング窓長	25ms
フィルタバンク	24 チャネル
特徴パラメータ	MFCC(12 次)+Δ MFCC+Δ Pow
	(計 25 次)
話者	男性 20 名, 女性 20 名 (MID, MID+)
	男性 23 名, 女性 23 名 (LARGE, LARGE+, LARGE++)

表 4.3: SPC モデルの情報 (基本語彙 20,000 単語)

	$n=2$	$n=3$
SPC の種類	16,794	46,718
単語辞書に加えた単語数	16,045	9,477
平均予測単語候補数	3.9	2.2

4.5 大語彙認識システムによる認識実験

4.5.1 実験条件

次に、4.3.2 節のように最適化した形態素解析に基づいて作成した SPC モデルを用いて大語彙認識システムを実装し、認識実験を行なった。認識実験の条件を表 4.2 に示す。

今回の認識実験では、基本語彙は 20,000 語として認識システムを実装した。またシステムに組みこむ SPC モデルには、2.6.3 節で述べた SPC 選別法におけるパラメータの値を $n=2$ では $\alpha=5$、$\beta=7$、$\gamma=90\%$、$n=3$ では $\alpha=3$、$\beta=15$、$\gamma=90\%$ として求めたモデルを用いた。作成した SPC モデルについての情報を表 4.3 に示す。

評価尺度として、3.3 節と同様に Word Accuracy と Word Correct Rate を用いる。但し、今回の実験では 2 つの評価尺度の求め方は 3.3 節とは異なり、品詞及び漢字表記も一致する単語を正解としている。

認識対象とする音声資料は、毎日新聞記事読み上げ音声コーパス JNAS の中の、出現頻度上位の 5,000 語で閉じた文のセット MID、同 5,000 語で未知語が 1 つの文のセット MID+、及び出現頻度上位 20,000 語で閉じた文のセット LARGE、同 20,000 語で未知語が 1 つの文のセット LARGE+、同 20,000 語で未知語が 2 つ以上の文のセット LARGE++を用いた。

表 4.4: 基本語彙20,000単語での認識実験の結果 (MID, MID+)

		Word Accuracy(%)		Word Correct Rate(%)	
speaker	the method	MID	MID+	MID	MID+
男性	bigram	72.7	72.2	76.7	74.2
	SPC モデル ($n=2$)	72.7	72.6	76.9	74.7
	SPC モデル ($n=3$)	73.1	72.9	77.0	75.0
女性	bigram	76.1	73.5	77.2	75.6
	SPC モデル ($n=2$)	76.1	73.5	77.3	75.6
	SPC モデル ($n=3$)	76.3	73.8	77.4	75.8

表 4.5: 基本語彙20,000単語での認識実験の結果 (LAR, LAR+, LAR++)

		Word Accuracy(%)			Word Correct Rate(%)		
speaker	the method	LAR	LAR+	LAR++	LAR	LAR+	LAR++
男性	bigram	72.6	64.5	54.2	74.6	67.9	60.0
	SPC モデル ($n=2$)	72.6	64.1	53.7	74.7	67.7	59.7
	SPC モデル ($n=3$)	72.8	64.5	54.3	74.7	68.0	60.2
女性	bigram	74.7	65.5	59.3	76.5	68.8	64.3
	SPC モデル ($n=2$)	74.5	65.4	59.0	76.4	68.6	64.1
	SPC モデル ($n=3$)	74.8	66.0	59.2	76.5	69.2	64.2

4.5.2 実験結果と考察

SPC モデルによる単語予測が認識に与える影響を、Word Accuracy 及び Word Correct Rate を求めることによって調べた。MID 及び MID++におけるこれらの値を表 4.4 に、更に LARGE、LARGE+及び LARGE++におけるこれらの値を表 4.5 に示す。表 4.5 中で LARGE を略して LAR と記す。

これらの結果より、出現頻度上位 5,000 単語に基づく MID 及び MID+に対しては、SPC モデルを用いたことによる認識精度の向上が見られた。一方、出現頻度上位 20,000 単語の LARGE、LARGE+及び LARGE++に関しては、SPC モデルを用いても結果に良い影響は出ているとは言い難い。このことから、実際の音声認識における SPC モデルによる単語の的確な予測は、出現頻度上位 5,000 単語について行なわれることが多いといえる。この現象は、単語予測に多く用いられる定型的な言い回しや動詞の接尾語等に関する SPC モデルが、出現頻度上位 5,000 単語を多く含むことに起因すると考えられる。

表 4.4 と表 4.5 をみると、$n=3$ の SPC モデルを用いた方が $n=2$ の SPC モデルを用い

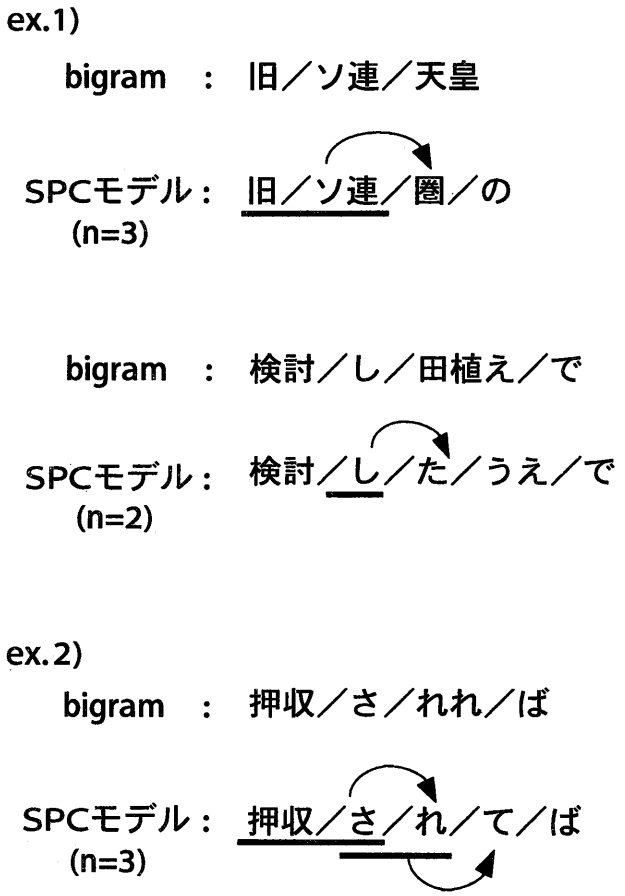


図 4.6: SPC による単語予測の実例 (基本語彙 20,000 単語)

た場合よりも、認識精度が良い。このことから、文脈を長くすることで次にくる単語をより絞りこむことができ、SPC モデルが単語の予測に信頼がおけるものとなっていることがいえる。

次に、SPC モデルを使用した実際の認識において、bigram を用いた場合と異なる結果となった例を図 4.6 に示す。図中の文において下線を付した箇所が SPC、矢印で指されている先が SPC によって予測された単語である。

(ex.1) を見ると、従来の bigram を用いて出力された、正文として成りたたない単語列が、SPC モデルを使用することによって排除されていることがわかる。最初の実例は $n = 3$ の SPC モデルを用いた結果で、SPC である “旧／ソ連” によって正解単語の “圏” が的確に予測されている。またもう一方の例では、bigram を用いた結果では、読みとしては合っているが品詞及び漢字表記では誤っている。このように単語列が同音であるために陥りやすい誤認識に関しても、SPC モデルによって後続単語の予測を行なうことで回避できた。

(ex.2) は改悪例を示したものである。この例では、SPC である “押収／さ” の予測単語候補に正解単語である “れれ” が含まれていない。そのため単語予測候補の中で正解に近い結

果を出してはいるが、正しい結果は得られなかった。これはSPCモデルの文カバレッジが100%でないために起こった結果である。

また表4.3より、$n = 2$のSPCモデルの予測単語候補の平均数が、3.3節で用いたSPCモデルに比べてかなり減少していることがわかる。この現象の1つの理由として、形態素解析ツールChaSenを使用することにより、3.3節では除外されていたn-gramがSPCモデルとして採用されたためであると考えられる。すなわち、文脈となる単語列が難解なために次にくる単語が非常に少数となっているSPCモデルが増え、予測単語候補の平均数が減少した。その例として“櫟／画廊”、“毀誉／褒貶”などがある。

更に、ChaSenを用いて形態素解析を行なったことにより、当然3.2.3節の手法で作成した形態素と異なる形態素が結果として得られる。例として“全欧安保協力会議”を挙げる。この単語列に対して3.3節で処理した手法で形態素解析を行なうと、“全／欧／安保／協力／会議”という5つの形態素に分けられる。一方、4.3.2節で述べた手法で解析処理をすると、たった1単語で“全欧安保協力会議”という形態素として解析される。つまり、今回の手法では頻繁に出現する単語列については1つの形態素とみなすために、4.3.2節に比べてSPCモデルの予測単語候補の平均数が減少している。

もう1つの理由として読みの細分化が挙げられる。本章ではChaWanやPostProcessを利用することで、数詞及び助数詞、音便などの読みの変化を考慮している。その結果として、例えば数詞“1”に対して3.3節では“イチ／イッ／ツイ／ツイタ／ヒト／ワン”という読みが1つとして与えられていたのに比べ、ChaWan等を用いることにより“イチ／イッ／ヒト”、“イチ／ヒト”、“ヒト”の三種類の読み方が付与される。従って2.6.3節で提案したSPC選別法を用いると、読みが異なって与えられた単語“1”の個々に対してSPCモデルを考慮することになり、単語候補数が減る場合がある。

ここで、先に述べた2番目の現象から、頻出する単語列についてはSPCモデルによる予測を行なうことはできない場合が生じる。また読みの細分化のために、更に後続単語の予測について信頼度の低い文脈もSPCとしてしまうことがいえる。これらの事象が原因で、SPCモデルを用いた際に認識精度の向上が見られなかったものと考えられる。

次に、基本語彙に含まれない単語をどの程度認識結果として出力しているかについて調べた。SPCによって次にくると予測された単語のうち、基本語彙に含まれなかった単語の割合を表4.6に示す。

表4.6から、基本語彙20,000単語の大語彙音声認識においては、基本語彙に含まれない単語をSPCモデルに内包することによる認識に対する影響はほとんどない。従って、SPCモデルを使用して得られる認識結果への効果は、SPCの後に展開する探索空間を削減することで、尤度が下位だった単語仮説を浮上させたことによるものであると結論づけられる。

表 4.6: 基本語彙に含まれない単語が予測された割合 (%)

	n	MID	MID+	LARGE	LARGE+	LARGE++
男性	2	0.2	0.0	0.2	0.0	0.1
	3	0.0	0.0	0.0	0.0	0.0
女性	2	0.0	0.2	0.2	0.0	0.3
	3	0.0	0.0	0.0	0.0	0.2

表 4.7: ビーム幅削減の実験条件

話者	男性 20 名, 女性 23 名
ビーム幅	300, 700, 1000, 1500 (MID)
	500, 1000, 1500 (LARGE)

4.6　ビーム幅減少による処理量削減に関する実験 (2)

ビーム幅を変化させて認識実験を行ない、実際に SPC モデルが処理量に対してどのような影響を及ぼすか調査する。

4.6.1　実験条件

実験に使用した SPC モデルは、4.5 節の認識実験で用いたものと同じである。実験条件は表 4.7 に示す通りである。その他の実験条件については、表 4.2 と同様とする。今回の評価尺度は Word Accuracy のみを採用する。

認識対象とする音声資料は、含まれる単語が 5,000 単語の MID と 20,000 単語の LARGE を用いる。

4.6.2　実験結果と考察

男性話者と女性話者の MID における実験の結果を図 4.7 と図 4.8 に、また LARGE における実験の結果を図 4.9 及び図 4.10 にそれぞれ示す。

図 4.8 と図 4.10 より、SPC モデルを用いるとビーム幅の減少に基づく認識精度の低下が緩いことがわかる。この結果から、SPC モデルを使用することによって、bigram よりもビーム幅を削減したことによる空間量の縮小の影響を受けにくくなると考えられる。特に $n = 3$ の SPC モデルにおいて、この現象は顕著である。

また図 4.7 から図 4.10 のグラフより、ビーム幅を変化させた場合でも、$n = 2$ の SPC モ

デルに比べて $n = 3$ のSPCモデルの方が、認識結果に良い影響を与えていることが見てとれる。このことからも、文脈を長くすることによって、SPCモデルの単語予測における信頼性の向上がなされているといえる。従って基本語彙20,000単語の連続音声認識においても、SPCモデルによって高次の n-gram の制約が有効に利用されていると考えられる。

4.7 まとめ

基本語彙数万単語の大語彙連続音声認識におけるSPCモデルの問題点について考え、その解決策を提案した。形態素解析の最適化をすることによるSPCモデルの単語予測の変化を調べた。これにより、形態素を最適化することで文カバレッジの向上が得られることがわかった。また改善した実装手法を用いて、SPCモデルを用いた基本語彙20,000単語の大語彙認識システムを実装した。実験の結果から、さほど認識精度を下げずにSPCモデルを使用できることがわかった。更にビーム幅を変化させて同様の認識実験を行なった。この結果から、SPCモデルを使用することによって、bigramよりも処理量削減の影響を受けにくくなることがいえる。

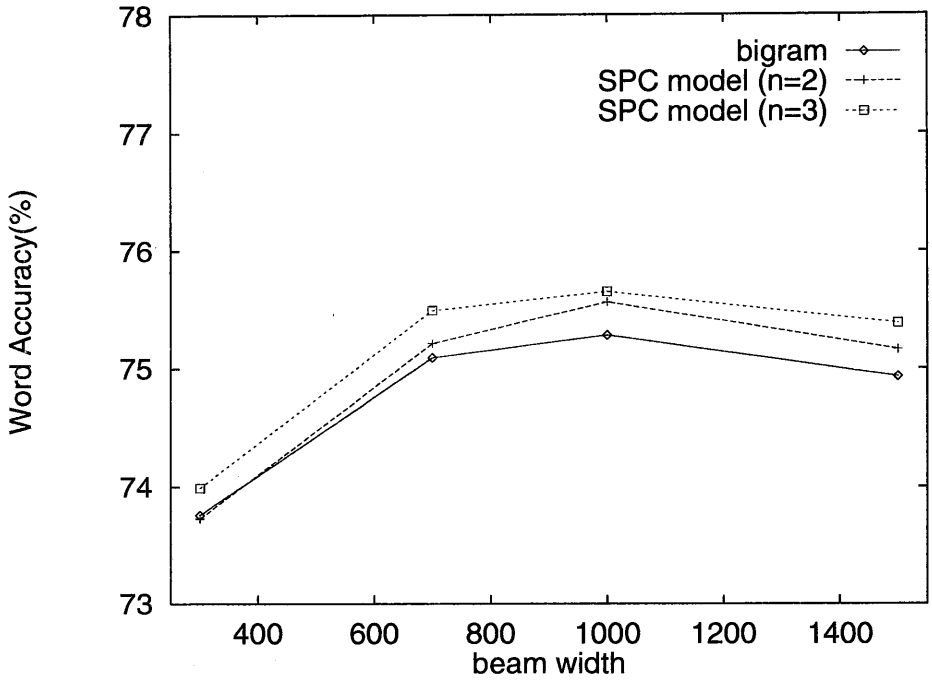


図 4.7: ビーム幅の変化による基本語彙 20,000 単語の認識精度の推移 (男性, MID)

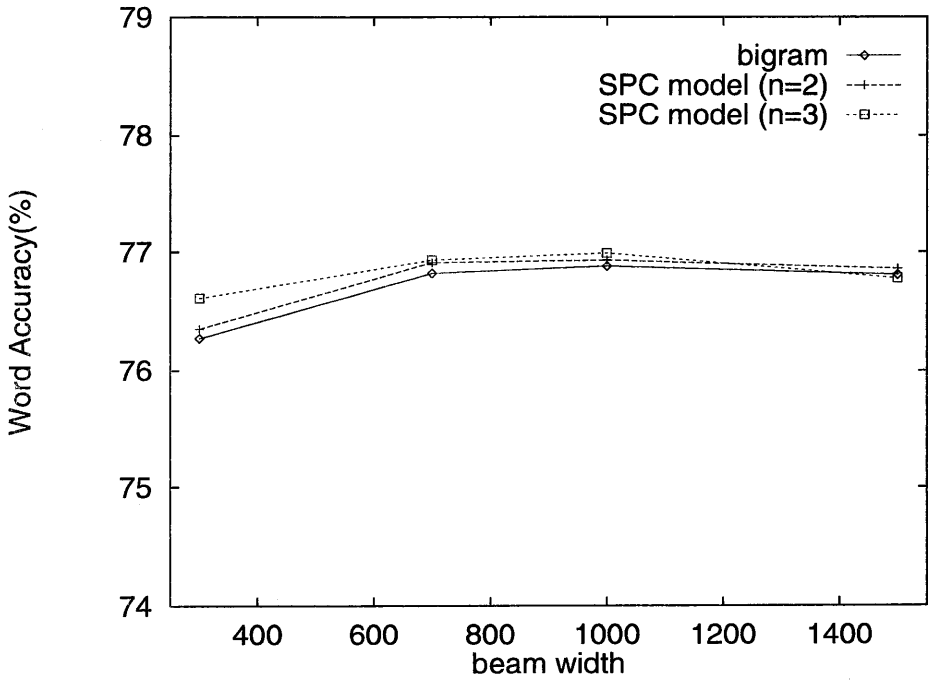


図 4.8: ビーム幅の変化による基本語彙 20,000 単語の認識精度の推移 (女性, MID)

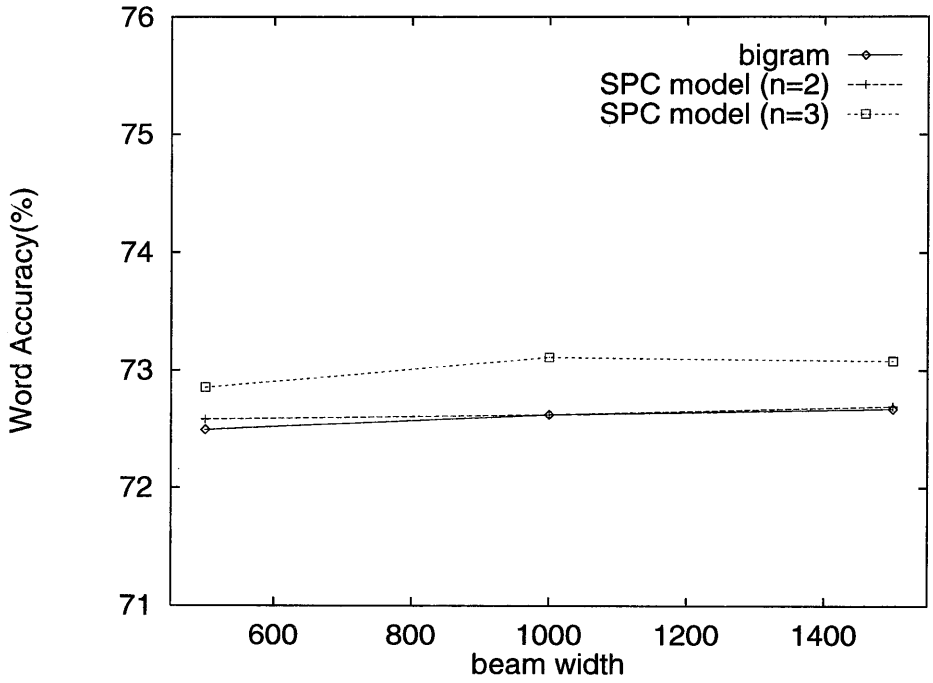


図 4.9: ビーム幅の変化による基本語彙 20,000 単語の認識精度の推移 (男性, LARGE)

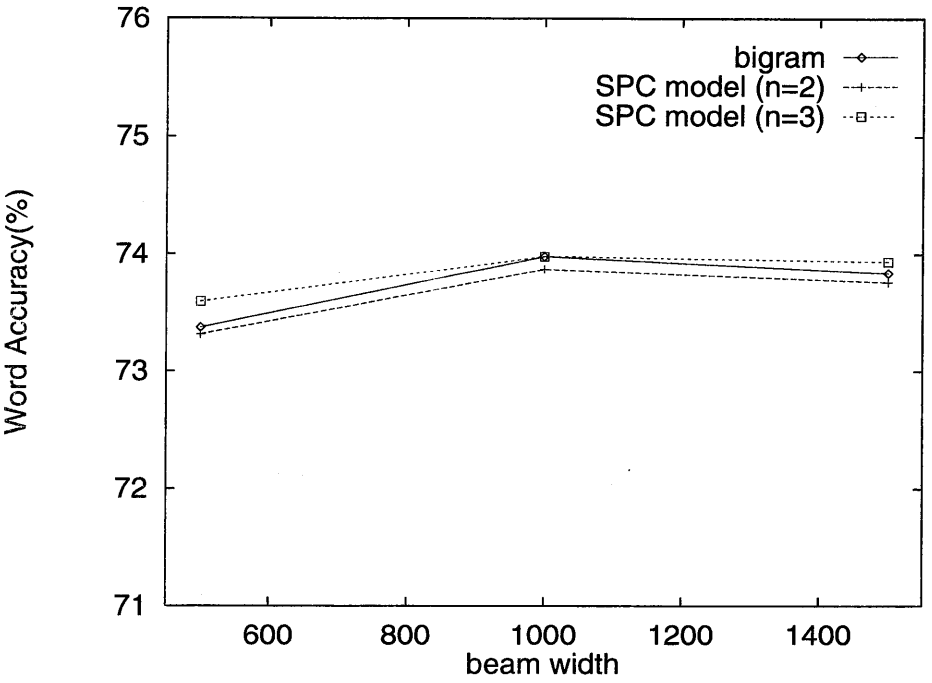


図 4.10: ビーム幅の変化による基本語彙 20,000 単語の認識精度の推移 (女性, LARGE)

第5章

結論

5.1 本研究の成果

本研究の成果について述べる。

第2章では、従来のスムージングしたn-gramの問題点である膨大な処理量について明示し、その解決手法を提案した。具体的には、特定の文脈に対して次にくる単語を的確に予測して、認識結果に悪影響をなるべく及ぼさずに処理量を削減するモデルを提案した。その特定の文脈をSuccessor-Predictable Context(SPC)と名付けた。また、認識に及ぼす悪影響をできるだけ防ぐようなSPCを選別する手法を提案した。そして、その選別法に基づいて作成したSPCモデルが認識にどの程度影響を及ぼすかについて、評価実験を行なった。これにより、文脈が長い$n=3$のSPCモデルの方が$n=2$のものよりも、次にくる単語の予測に対して信頼がおけることがわかった。

第3章においては、実際にSPCモデルを用いた認識システムを実装するために、木構造辞書の拡張手法を提案した。その手法に基づいて、前章のSPC選別法に基づいて作成したSPCモデルを用いて、基本語彙5,000単語の連続音声認識システムを実装した。そして、実装した認識システムを実際に用いて認識実験を行なった。実験の結果から、あり得ない単語列をSPCモデルによって排除できるという有効性が示された。またビーム幅を変化させた認識実験の結果から、SPCモデルによる処理量の削減の検証も行なった。これらの結果より、同じビーム幅では認識精度が向上し、ビーム幅を減少させた場合には空間量に関して約30%の削減が実現することがわかった。

第4章では、第3章でSPCモデルを用いて実装した基本語彙5,000単語の音声認識システムによる実験結果を踏まえて、SPCモデルを用いて更に基本語彙の大きな音声認識システムを実装する際の問題点を考え、その解決策を提案した。更に、音声認識において有効

な形態素解析の最適化を行ない、それに基づいたSPCモデルを作成した。評価実験から、形態素解析の最適化はSPCモデルの単語予測の信頼性向上に有効であると考えられる。また、SPCモデルを処理するための改善手法を用いて、新しいSPCモデルを用いた大語彙認識システムを実装し認識実験を行なった。その結果SPCモデルが認識にどのような影響を及ぼすか、例を挙げて述べた。次にビーム幅を変化させて認識実験を行ない、SPCモデルを用いることによる影響を調べた。この結果から、SPCモデルを使用することによって、bigramよりも処理量削減の影響を受けにくくなることがいえた。

5.2 今後の課題

今回第2章で提案したSPC選別法を用いてSPCモデルを作成しそれを認識に用いることで、先に述べたような結果を得た。しかしこの選別法にも問題点がある。

本論文で提案したSPC選別法は、n-gram確率上位β個のn-gram確率を累積した確率を基にSPCを選別するものである。しかし、出現頻度が非常に多い文脈で、ほぼ均等に後続単語が存在するためにn-gram確率にあまり差がない場合を考えると、作成したSPCモデルによる単語予測の信頼度が低くなり、文カバレッジが低下してしまうと考えられる。このことから、ある程度の出現頻度の差を持つn-gramに着目し、差が開くまでのn-gram確率を累積してSPCを選別すれば、より単語予測精度を高くできると考えられる。

このようなことを考慮に入れ、単語予測に更に信頼がおけるSPCモデルを選別する手法を考案する必要がある。

謝辞

本研究を進めるにあたり、全般的な御指導とともにこの研究の機会を与えて下さった東北大学工学部 阿曽弘具教授に心から感謝致します。

東北大学工学部 阿部健一教授におきましては、本論文をまとめるにあたり至らない点を御指摘いただきましたことを深く感謝致します。

音声認識の分野におきましては、さまざまな御指導、御助言を頂いた東北大学大型計算機センター 牧野正三教授に深く感謝致します。

日々の研究においては、東北大学工学部 大町真一郎助教授に感謝致します。また距離を問わず、細かく御指導、御教授頂いた宇都宮大学工学部助手 森大毅氏に感謝致します。音声ゼミにおきまして貴重な御意見、御助言を頂いた東北大学大型計算機センター助手 鈴木基之氏に感謝致します。研究室ゼミにおいて様々な御指摘をしていただいた東北大学工学研究科 佐藤俊治氏、並びに毎日の研究室生活において活気づけて下さった阿曽研究室の皆様に感謝致します。

つらいときに心の支えとなって下さった家族及び友人たちに心より感謝致します。

参考文献

[1] 小川 英光, “パターン認識・理解の新たな展開”, 電子情報通信学会, 1994.

[2] 武田 一哉, 伊藤 克亘, 松岡 達雄, 竹沢 寿幸, 鹿野 清宏, “大語彙連続音声認識研究のためのテキストデータ整備”, 情報処理学会研究報告, SLP-96-55, pp.49–54, 1996.

[3] 亀山 誠裕, 加藤 正治, 伊藤 彰則, 好田 正紀, “新聞記事コーパスからの言語モデル作成におけるカットオフ及び時期差の検討”, 日本音響学会平成10年度春期研究発表会講演論文集, 1-6-23, 1998.

[4] 古井 貞熙, “大語彙連続音声認識の現状と展望”, 日本音響学会平成10年度春期研究発表会講演論文集, 1-6-10, 1998.

[5] L. Rabiner, B. Juang, 古井 貞熙監訳, “音声認識の基礎(上)(下)”,NTTアドバンステクノロジ株式会社, 1994.

[6] 中川 聖一, “確率モデルによる音声認識”, 電子情報通信学会, 1988.

[7] 李 晃伸, 河原 達也, 堂下 修司, “A^* 探索に基づく大語彙連続音声認識”, 情報処理学会研究報告, SLP-96-55, pp.19–24, 1996.

[8] 村上 仁一, “フレーム同期型フルサーチアルゴリズムを用いた連続音声認識と自由発話への応用”, 信学技報, SP-95-32, 1996.

[9] 川端 豪, 田本 真詞, “二項事後分布に基づくN-gram言語モデルのBack-off平滑化”, 信学技報, NLC95-58, SP95-93, 1995.

[10] M. Federico, M. Cettlo, F. Brugnara and G. Antoniol, “Language modelling for efficient beam-search”, Computer Speech and Language, vol.9, pp.353–379, 1995.

[11] Katz, S. M., “Estimation of Probabilities form Sparse Data for the Language Model Component of a Specch Recognizer”、IEEE Trans., ASSP-35, 3, pp.400–401, 1987.

[12] 伊藤 彰則, “タスクに依存しない日本語文音声の認識に関する研究”, 東北大学博士学位論文, 1991.

[13] K. Itou, K. Takeda, T. Takezawa, T. Matsuoka, K. Shikano, T. Kobayashi, and S. Itahashi, “Design and development of Japanese speech corpus for large vocabulary continuous speech recognition assessment”, In Proc. Oriental COCOSDA, 1998.

[14] 伊藤 克亘, 伊藤 彰則, 宇津呂 武仁, 河原 達也, 小林 哲則, 清水 徹, 田本 真詞, 荒井 和博, 峯松 信明, 山本 幹雄, 竹沢 寿幸, 武田 一哉, 松岡 達雄, 鹿野 清宏, “大語彙日本語連続音声認識研究基盤の整備–学習・評価用テキストコーパスの作成–”, 情報処理学会研究報告, 97-SLP-18-2, 1997.

[15] Cambridge University, “The CMU-Cambridge Statistical Language Modeling Toolkit v2”, http://svr-www.eng.cam.ac.uk/ prc14/toolkit.html, 1997.

[16] 河原 達也, 李 晃伸, 伊藤 克亘, 伊藤 彰則, 宇津呂 武仁, 小林 哲則, 清水 徹, 田本 真詞, 荒井 和博, 峯松 信明, 山本 幹雄, 竹沢 寿幸, 武田 一哉, 松岡 達雄, 鹿野 清宏, “大語彙日本語連続音声認識研究基盤の整備–評価用連続音声認識プログラムの開発–”, 情報処理学会研究報告, 97-SLP-18-1, 1997.

[17] 情報処理振興事業協会 (IPA), “日本語ディクテーション基本ソフトウェア–1997 年度版–”, 1998.

[18] 日本音響学会, “新聞記事読み上げ音声コーパス (JNAS)”, 1997.

[19] 情報処理振興事業協会 (IPA), “日本語ディクテーション基本ソフトウェア–1998 年度版–”, 1999.

[20] 高木 一幸, 小黒 玲, 橋本 顕示, 尾関 和彦, “ニュース音声認識における言語モデルの検討”, 信学技報, NLC-97-46, SP-97-79, 1997.

[21] 高木 一幸, 桜井 直之, 岩崎 淳, 古井 貞熙, “ニュース音声を対象とした言語モデルと話題抽出の検討”, 信学技報, SP-98-33, 1998.

[22] 伊藤 克亘, 田中 和世, 山本 俊一郎, 踊堂 憲道, 鹿野 清宏, “音声認識用統計言語モデルのための形態素解析済みテキストの後処理”, 日本音響学会講演論文集, 1998-3

[23] 奈良先端科学技術大学, 日本語形態素解析システム『茶筅』version 2.0, http://cl.aist-nara.ac.jp/lab/nlt/chasen/, 1999.

[24] 山本 俊一郎, 踊堂 憲道, 中村 哲, 鹿野 清宏, 伊藤 克亘, “形態素解析結果の後処理により改良された統計的言語モデルの評価”, 日本音響学会平成10年度秋期研究発表会講演論文集, 2-1-16, 1998.

[25] 山本 俊一郎, 伊藤 克亘, 山田 篤, 宇津呂 武仁, 鹿野 清宏, “ディクテーションにおける形態素辞書エントリと読みの整備の効果”, 日本音響学会平成11年度春期研究発表会講演論文集, 2-1-1, 1999.

研究業績一覧

1. “頻出文脈に着目した単語予測法”,
馬場雅美, 森大毅, 牧野正三, 阿曽弘具,
東北大学通信工学研究所 第297回音響工学研究会, (1998.5).

2. “文脈に基づく単語予測を用いた音声認識”,
馬場雅美, 森大毅, 牧野正三, 阿曽弘具,
日本音響学会平成11年度春季発表会. 2-1-9 (1999.3).

3. “Proposition of Successor-Predictable Context and its Application to Large-Vocabulary Continuous Sppech Recognition”,
Masami Baba, Hiroki Mori, Shozo Makino, Hirotomo Aso,
International Conference on Speech Processing, vol. 2, pp.371–376(1999.8).

■ 修士学位論文

後続語予測可能文脈に基づいた大語彙連続音声認識に関する研究

2000年2月17日
東北大学大学院工学研究科電気・通信工学専攻
馬場雅美

第1章 序論

■ 背景

現在音声認識は様々な分野で実利用
- 孤立単語認識(例:カーナビ・電話)
- 連続音声認識(例:ワープロ用ディクテーションシステム)

言語モデル
認識結果をより正しい文に近づけるために
計算機に与える「文法知識」

例
浅い花
⇩
赤い花

近年の大語彙認識システム:
統計的言語モデル n-gram を使用する傾向

■ 目的

連続音声認識における問題：処理量が膨大

次にくる単語の的確な予測に基づく
高精度な単語予測法を用いた
探索空間の絞り込みによる処理量の削減

次にくる単語を限定できる特定の文脈に
着目して後続語の数を削減

■ 本論文の構成

第2章　文脈の単語予測による認識の高性能化

n-gramとそのスムージング(1)

n-gram

単語の生成確率が直前のn-1個の単語のみに依存

$$P(w_N \mid w_{N-n+1} w_{N-n+2} \cdots w_{N-1}) \approx P(w_N \mid w_1 w_2 \cdots w_{N-1})$$

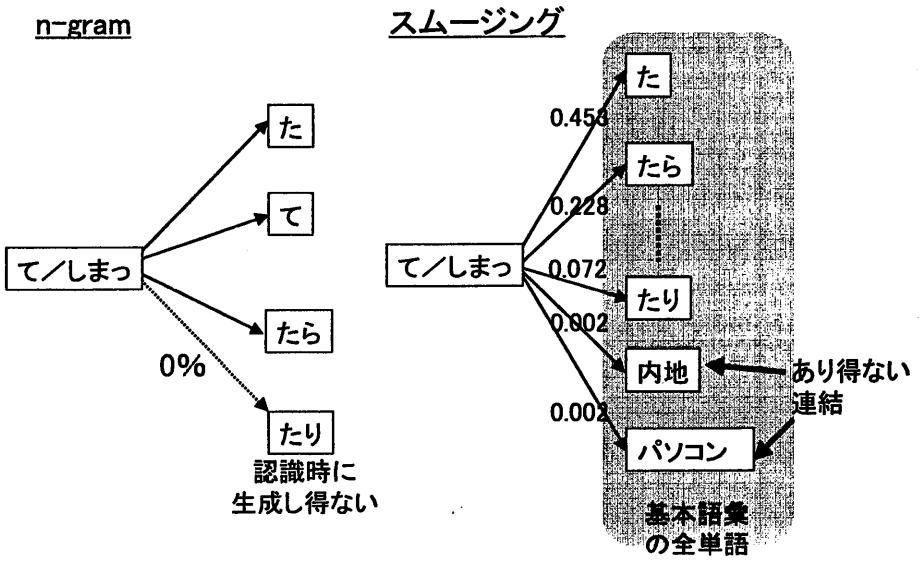
n-gram確率: $P(w_n \mid w_1 w_2 \cdots w_{n-1}) = \dfrac{N(w_1 w_2 \cdots w_n)}{N(w_1 w_2 \cdots w_{n-1})}$

$N(w_1 w_2 \cdots w_n)$: 単語列 $w_1 w_2 \cdots w_n$ の出現回数

n-gramモデル:

$$P(w_1 w_2 \cdots w_{N-1} w_N) = \prod_{i=1}^{N} P(w_i \mid w_{i-(n-1)} w_{i-(n-2)} \cdots w_{i-1})$$

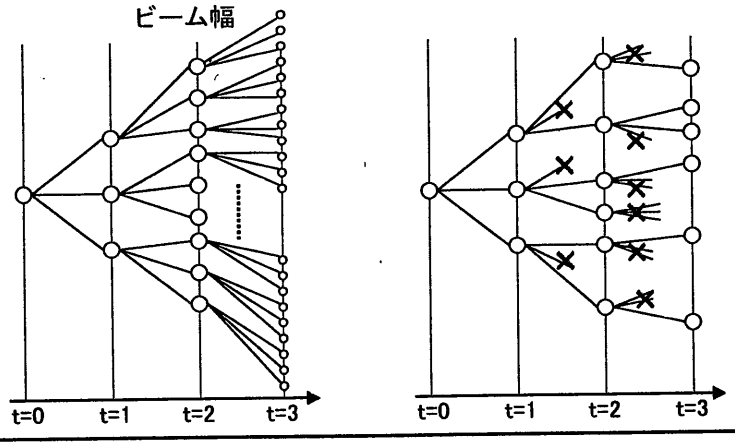
n-gramとそのスムージング(2)



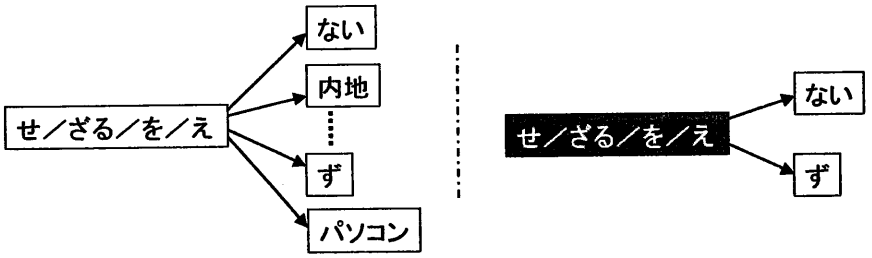
ビームサーチ

尤度の低い単語仮説を制限に従って
その後の探索から除外することで処理量削減

最大尤度の単語仮説から尤度の高い順に
ある定数個の単語仮説を残して他は除外



Successor-Predictable Context

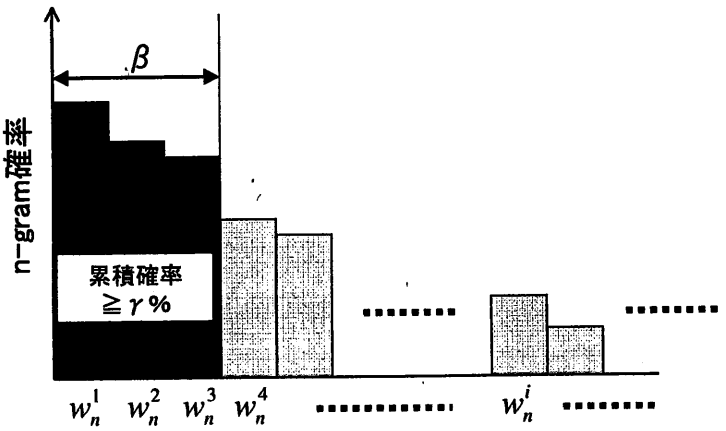


統計的言語モデルを用いて次にくる単語を
的確に同定できる文脈を自動的に獲得

Successor-Predictable Context (SPC)

SPCでない文脈：従来通りスムージングを用いて展開

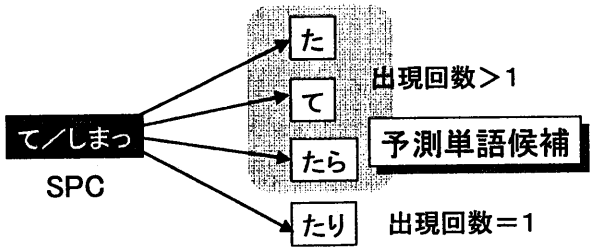
SPCの選別手法(1)



文脈Ψの次にくる単語(出現回数α回以上)

SPCの選別手法(2)

SPCの予測単語候補の選出



認識に与える影響についての評価実験

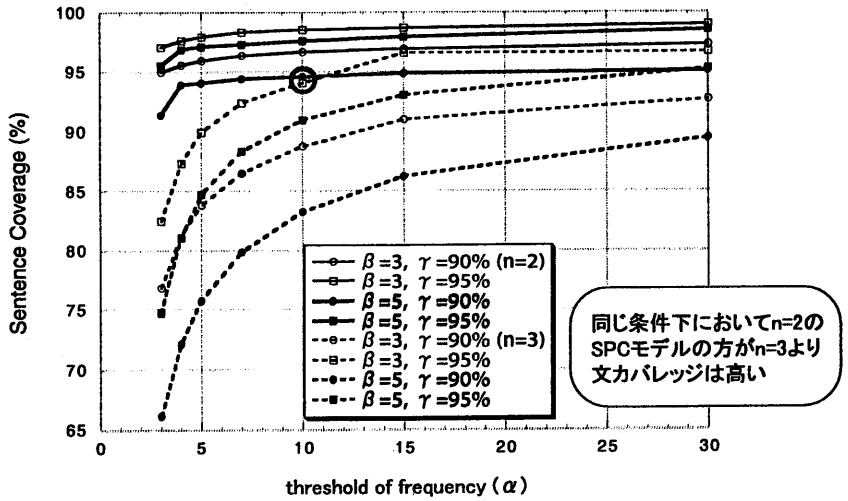
実験条件

SPCモデル	n = 2, 3
学習用データ	毎日新聞記事CD-ROM45ヵ月分 (Jan.1991 - Sep.1994)
評価用データ	同3ヵ月分(Oct.1994 - Dec.1994)
出現頻度の閾値 α	3, 4, 5, 7, 10, 15, 30
上位 β 個	3, 5
累積確率の閾値 γ %	90, 95

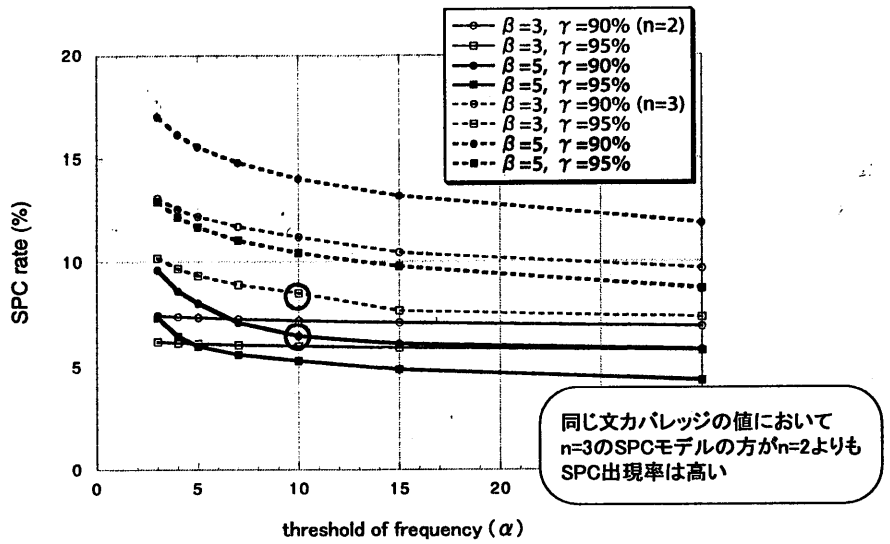
$$\text{文カバレッジ(\%)} = \left(1 - \frac{\text{SPCモデルによる予測が失敗した文数}}{\text{文の総数}}\right) \times 100$$

$$\text{SPC出現率(\%)} = \frac{\text{SPCの出現回数}}{\text{文脈総数}} \times 100$$

実験結果-文カバレッジ-



実験結果-SPC出現率-



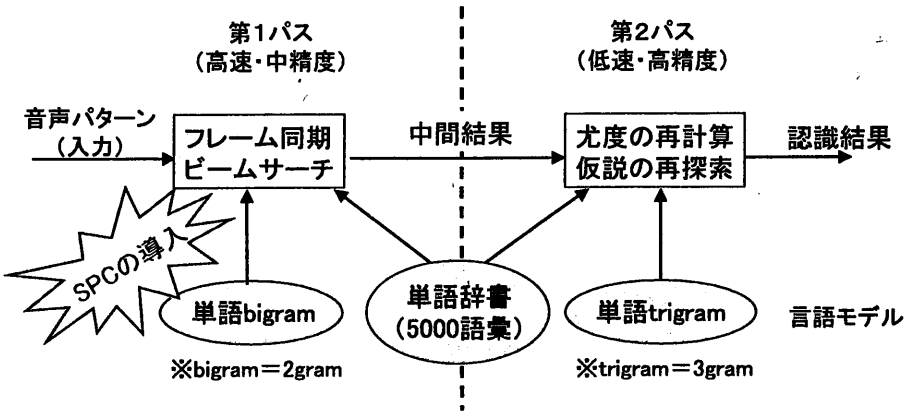
第2章まとめ

- 次にくる単語を的確に予測できる文脈を用いて探索空間を絞り込む単語予測モデル(SPCモデル)を提案した
- SPCモデルが認識にどの程度影響を及ぼすかの評価実験を行なった
- 文脈が長いn=3のSPCモデルによる単語予測の方がn=2の場合よりも信頼がおけるといえる

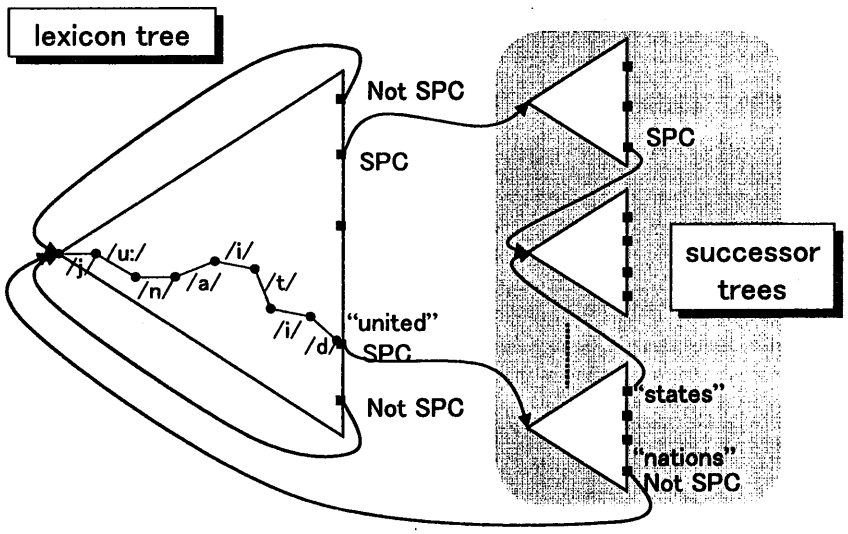
第3章　単語予測モデルの認識システムへの実装

■ 2パス処理システム

-大語彙連続音声認識システムで主流



■ SPCモデルを用いたシステムの実装



SPCモデルを実装したシステムによる認識実験

・既存のシステムを基礎としてSPCモデルを用いたシステムを実装

-システムの基本語彙数：5,000単語

実験条件

音響モデル	triphone (状態数：2000　混合数：16)
サンプリング周波数	16 kHz
フレーム周期	10 ms
フレーム長	25 ms
フィルタバンク	24 channel
特徴パラメータ	MFCC(12)＋ΔMFCC＋ΔPOW (total：25)
認識対象音声	ASJ-JNAS 音声データベース male：10名, female：10名
ビーム幅	500, 600, 700

認識実験に用いた評価尺度

結果に対して正解文とDPマッチングを行なう

$$\text{Word Accuracy (\%)} = \frac{N - S - D - I}{N} \times 100$$

N ：正解文の単語総数

S ：単語の置換総数

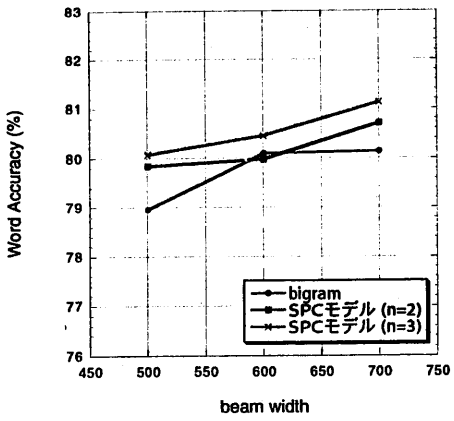
D ：単語の脱落総数

I ：単語の挿入総数

認識結果の単語の読みを用いて算出
品詞や表記が異なっても正解

実験結果

認識対象音声:
- 高頻出5,000単語で閉じた文セット
- 男声



- ビーム幅を削減しても同程度の精度を達成
  - 上位700仮説から500仮説へ
    空間量約30%削減

認識結果の具体例(1)

改善例

Ex.1) ありえない単語列の排除

bigram : アジア／太平洋／白／6／会議

SPCモデル (n=3) : アジア／太平洋／経済／協力／会議

Ex.2) 基本語彙に含まれない単語の認識

bigram : 私／の／戻っ

SPCモデル (n=2) : 私／ども

認識結果の具体例(2)

改悪例

bigram　：候補／が／私／以外

SPCモデル：候補／が／私／に／意外
(n=3)

正解単語である“以外”が SPCである“が／私”の単語候補に含まれていなかった

第3章まとめ

- システムに実装するための木構造辞書の拡張法を提案した
- SPCモデルを用いて連続音声認識システムを実装し認識実験を行なった
- 実験結果より同じビーム幅ならば認識精度が向上
  空間量を約30%削減して同程度の精度を実現した

第4章　大語彙連続音声認識における単語予測

■ 基本語彙数の大規模化における処理量問題

処理量の膨大化

・語彙の大規模化に伴う単語仮説数の増加

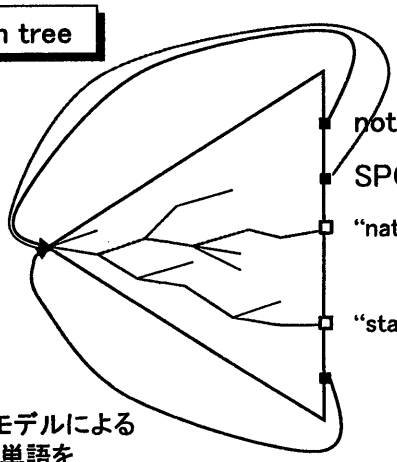
lexicon tree と successor tree を個別に作成するため
同じ単語が複数の tree 内で重複して存在

-計算に要するメモリの見積もり概略：102.8Mb

SPCモデルをシステム内で処理する木構造辞書の
拡張法の改善が必要

■ 大語彙における木構造辞書の拡張

拡張 lexicon tree



"united"の
予測単語候補

基本語彙とSPCモデルによる
基本語彙以外の単語を
一つの辞書木で表す
（元の辞書木よりも大きい）

計算に要するメモリの見積もり概略
103Mb ⇒ 10Mb

SPCモデルを用いた大語彙認識システムの認識実験

-システムの基本語彙数：20,000単語

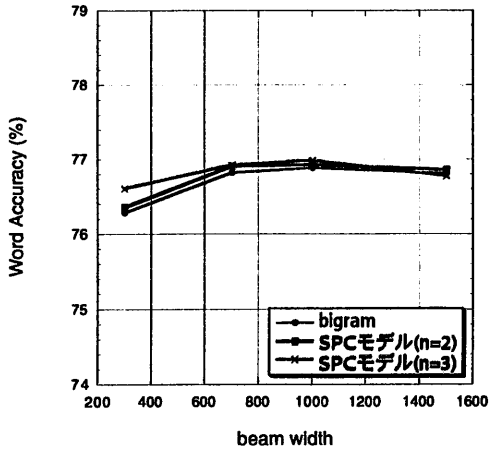
実験条件

音響モデル	triphone (状態数：2000　混合数：16)
サンプリング周波数	16 kHz
フレーム周期	10 ms
フレーム長	25 ms
フィルタバンク	24 channel
特徴パラメータ	MFCC(12)＋ΔMFCC＋ΔPOW (total：25)
認識対象音声	ASJ-JNAS 音声データベース male：20名, female：20名
ビーム幅	300, 700, 1000, 1500

評価尺度：Word Accuracy (%)

実験結果(female) -基本語彙20,000単語-

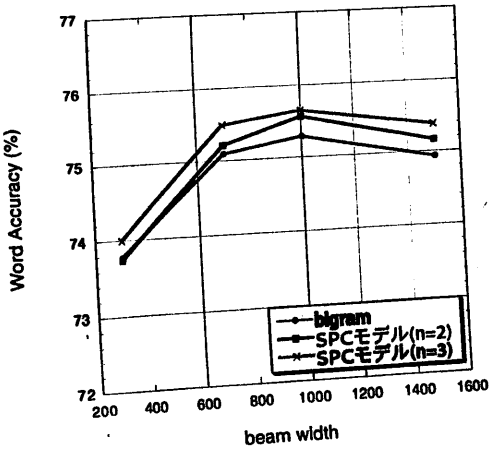
認識対象音声：
- 高頻出5,000単語で閉じた文セット
- 女声



グラフの傾きからビーム幅の縮小による影響が軽減されていることがわかる

実験結果(male) -基本語彙20,000単語-

認識対象音声:
- 高頻出5,000単語で閉じた文セット
- 男声



文脈を長くすることで単語予測の信頼性向上がみられた
-大語彙認識での高次のn-gramの制約の活用が可能

第4章まとめ

- 基本語彙が更に大きくなったときの問題点を挙げそれを解決する手法を提案した
- 提案した手法に基づき新たな連続音声認識システムを実装して認識実験を行なった
- ビーム幅を削減による認識結果への影響がSPCモデルを用いることで軽減できることがわかった
- 大語彙の認識で高次のn-gramを効率よく利用することが可能となる

第5章　結論

本研究の成果

- 特定の文脈に対して次にくる単語を的確に予測して処理量を削減するモデル(SPCモデル)を提案した
- 提案したSPCモデルを認識システムに実装して実際に認識実験を行なった
  その結果SPCモデルを用いることで性能の向上が見られた
- 文脈が長いSPCモデルを用いることによって高次のn-gramの制約を大語彙認識でも利用できるといえる

今後の課題

・SPC選別手法の洗練