TR-IT-0259

統計的手法による音声言語処理の研究

Research of Language Processing Using Statistical Method for Speech Recognition and Understanding

政瀧 浩和
Hirokazu Masataki

1998.3.31

本報告は、連続音声認識、および音声理解の性能向上を目的として、筆者がＡＴＲ音声翻訳通信研究所に在籍した３年間に行った統計的手法を用いた言語処理技術についての研究成果をまとめたものである。

もくじ

1 序論

1.1 本研究の背景

近年の連続音声認識技術の進歩には目覚しいものがある．これは，米国の ARPA(Advanced Research Projects Agency) プロジェクトが主催する Wall Street Journal 新聞を中心として，北米の種々の経済新聞 (NAB: North American Business News) の記事を読み上げに文章を認識するプロジェクトが強力に進められ，さまざまな研究機関が性能向上に向けて技術開発を競い合ったことが大きく貢献している．最近では，65,000 語彙で単語認識率 93% とかなり高い性能が報告されるている [1][2][3]．また，CPU の高速化，ディスク・メモリの大容量化等による計算機の性能の著しい向上と合間って，パソコンでも実時間に近い処理が可能になる等，ここ数年間でかなり実用レベルに近づいた感がある．

最近の ARPA プロジェクトでは，放送ニュースの認識や自然発話の認識等のより認識の難しい対象に興味が移っている一方，ATIS(Air Travel Information System) システムと呼ばれる音声による航空路線の案内システムのように，音声認識のアプリケーションとしての音声理解システムの研究も行われている．

また，日本においても，ディクテーションシステムを前提とした日本語大語彙音声認識で高い性能を得ており [5][6]，自然発話音声認識の研究 [7] や，音声認識結果を他国語に翻訳し異国語間の会話を可能とする音声翻訳システム [8] のような音声理解システムの研究も盛んに行われている．

音声認識技術においてその中心的役割を果たすのは，波形の特徴量から音素の認識を行う音響モデルである．音響モデルに隠れマルコフモデルを用いる手法が提案されて以来，音声の認識精度は大幅に向上した．しかし，連続音声認識における音響モデルの性能は，音素認識率で 60 ～ 70 パーセント程度であり，現状では，音響モデルだけで実用的な認識率を得ることはできない．このため，認識の単語を単語とし，単語間の接続関係を表現する「言語モデル」とを組み合わせることにより，連続音声認識の性能を格段に向上させている [4]．

従来，連続音声認識の言語モデルとしては，文法的ルールを用いる手法が盛んに研究されていたが [9]，近年連続音声認識において盛んに用いられている言語モデルは，N-gram と呼ばれるモデルである．この N-gram は，直前の (N-1) 単語から次の単語を確率的に予測するモデルであり，学習用に与えられたデータから予測確率を学習して自動的にモデルの構築を行う「統計的モデル」である．N-gram は極めて単純な言語モデルでありながら，構築の容易さ，統計的音響モデルとの相性の良さ，認識率向上や計算時間の短縮の効果が大きい等の理由で，連続音声認識には非常に有効であり，実際，最近の殆んどの連続音声認識システムにおいてこの統計的言語モデル N-gram が使用されていると言っても過言ではない．

一方，音声認識のアプリケーションとして盛んに研究されている音声理解システムにおいては，文法的ルールを用いた構文解析手法が言語理解技術として盛んに用いられているが，最近では，音声認識の言語モデルと同様，統計的手法を用いた言語理解方法も研究され始めている [10]．

1.2 本研究の目的

統計的言語モデル N-gram は，大語彙連続音声認識の言語制約として，その有効性が広く確認されている．しかし，実際に連続音声認識システムを構築する際には，次のような問題が発生する．

- 日本語の N-gram 学習用データを集めるのは困難
  英語等の文章は，単語がスペースで区切られているため，テキストデータから直接単語を単位とした N-gram の構築が可能であるが，日本語の文章では単語の区切りがなく，テキストデータから直接 N-gram を構築することはできない．このため，通常文章を品詞の単位で分割した形態素を単位とした N-gram

が使用される．しかし，形態素はテキストデータから単純に得られるものではないため，日本語の N-gram 学習用のデータを大量に収集するのは容易ではない．

- 少量の学習データから精度の良いモデルを得るのは困難
  N-gram はパラメータ数が多いため，学習には通常新聞記事等の大量のテキストデータが用いられる．しかし，音声理解システムに必要な話言葉のデータを大量に収集することは容易ではなく，日本語の場合はさらに困難が増す．学習データ量が少ない場合，N-gram の精度は低下し，システム構築上の大きな問題となる．

また，音声認識技術の応用である音声理解システムの言語理解技術においては，文法ルールを用いた構文解析による手法が盛んに用いられているが，この手法には次のような問題が考えられる．

- 音声認識システムとの相性が良くない
  連続音声認識では，N-gram が盛んに使用されているが，N-gram は局所的な制約しか持たず，認識誤りを生じると，非文法的な認識結果が得られる場合がある．この場合，文法ルールを用いた構文解析では，正しく言語理解を行うことはできない．

- 自由発話への適用が困難
  文法ルールは通常書き言葉を対象として作成されるが，音声を入力，すなわち話し言葉では一般的な文法では表現が困難であり構文解析するのは困難であると考えられる．

これらの問題点を解決するため，筆者は次のような研究を行った．

1. 品詞と可変長単語列の複合 N-gram
   少量のデータから精度の高い N-gram を構築するため，少ないパラメータ数で効率の良い表現が可能な N-gram モデルである「品詞と可変長単語列の複合 N-gram」を提案する．第 2 章では，まず N-gram について解説を行うとともに，提案する複合 N-gram の動作，生成方法について説明する．

2. 複合 N-gram を用いた自動形態素解析
   日本語の N-gram を構築するための大量の形態素解析データを収集するため，複合 N-gram を用いテキストデータに対し自動的に形態素解析を行う手法を 3 章で提案する．

3. 複合 N-gram を用いた連続音声認識
   少量のデータから精度の高いモデルが得られる複合 N-gram を連続音声認識に適用した．第 4 章にその実験結果を示し，連続音声認識における複合 N-gram の有効性を明らかにする．

4. N-gram 言語モデルタスク適応
   音声認識システムを構築する際には，タスクを限定し認識性能を向上させる方法が取られるが，対象とするタスクに関して大量のデータを収集することは容易ではない．このため，対象タスクの少量のデータを用いて，対象外タスクの大量データから構築される N-gram をそのタスクに適応するタスク適応手法を研究した．提案する手法は最大事後確率 (MAP) 推定を用いた手法で，第 5 章で MAP 推定による N-gram の定式化を行うと共に，タスク適応への応用方法について述べ，実験結果によりその有効性を示す．

5. クラスタ別 N-gram
   タスク適応の考えを拡張し，テキストデータ全体をクラスタリングし，そのクラスタ別の N-gram を構

築することによってN-gramの精度を向上させる手法を研究した．第6章にクラスタ別N-gramの構築方法を示し，実験結果によりその有効性を示す．

6. 統計的手法による音声理解

第7章では，音声認識誤りの理解，および自然発話の理解を目的として，統計的モデルを用いた言語理解の手法を提案する．

最後の8章は結論であり，本研究の成果を要約した．

2 統計的言語モデル

2.1 緒言

近年の連続音声認識技術は統計的手法が盛んに用いられている．統計的手法による連続音声認識では，入力波形の特徴パラメータ X から，確率的に最も高い単語列 W_L を探索し認識結果とする．これは，下式のように表される．

$$W_X = \arg max_W P(W|X) \tag{1}$$

本式は，Bayes' 則を用いることにより，次のように変形できる．

$$W_X = \arg max_W \frac{P(X|W)P(W)}{P(X)} \tag{2}$$

$P(X)$ は右辺の最大化には無関係な量であるから，次式と等しくなる．

$$W_X = \arg max_W P(X|W)P(W) \tag{3}$$

近年の連続音声認識技術は，本式を元にした処理を行っている．本式において，$\arg\max_W$ は考えられる全ての単語列の中から，確率 $P(X|W)P(W)$ の最大値を与えるものを探索することを意味し，サーチと呼ばれる処理である．確率 $P(X|W)$ は単語列 W から入力波形の特徴パラメータ X が得られる確率で，この確率を与えるモデルは音響モデルと呼ばれている．また，確率 $P(W)$ は単語列 W が出現する確率で，この確率を与えるモデルは言語モデルと呼ばれている．$P(W)$ すなわち，単語列 W が出現する確率を統計的な手法を用いて計算するのが本研究の主な対象となる統計的言語モデルである．本章の以下の節では，統計的言語モデルとして用いられる N-gram について述べ，その性能を改善するために提案されているクラス N-gram および可変長単語列 N-gram についても解説し，さらに我々が提案している品詞と可変長単語列の複合 N-gram について説明する．

2.2 N-gram の概念

統計的言語モデルは，単語列 W の生成確率を与えるモデルである．単語列 W は L 単語からなるとすると，生成確率は下式のように表される．

$$P(W) = P(w_1, w_2, \ldots, w_L) \tag{4}$$

しかし，この確率を求めるのは非常に困難である．そこで本式を，次のように変形してみる．

$$P(W) = \prod_{t=1}^{L} P(w_t|w_1^{t-1}) (= P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)\ldots P(w_L|w_1, w_2, \ldots, w_{L-1})) \tag{5}$$

ただし，w_x^y は単語列 W の x 番目から y 番目の単語列を意味する．$P(w_1)$ は単語 w_1 が文の最初に出現する確率である．また，$P(w_2|w_1)$ は，単語 w_1 の次に単語 w_2 が出現する確率である．これらの確率は適度な量のテキストデータから比較的容易に求めることができると考えられる．しかし，$P(w_L|w_1, w_2, \ldots, w_{L-1})$ は，$L-1$ 単語列の次に単語 w_L が出現する確率で，L が大きいと $L-1$ 単語の組み合わせは膨大な数になり，この確率を求めるのは不可能である．そこで，t 番目の単語 w_t は直前の $N-1$ 単語列のみに依存すると考えると，単語列 W の生成確率は次のように近似できる．

$$P(W) \approx \prod_{t=1}^{L} P(w_t|w_{t-N+1}^{t-1}) \tag{6}$$

N が比較的小さければ，$N-1$ の単語列の組み合わせはそれほど大きくなく，大量のテキストデータがあれば，求めるのは不可能ではない．本式で表されるモデル，すなわち直前の $N-1$ 単語から次の単語への遷移を確率に与えるモデルが N-gram 言語モデルである．通常，直前の 1 単語から次の単語の遷移確率を与える Bigram(2-gram)，および Trigram(3-gram) がよく用いられる．Bigram および Trigram は以下のように表される．

Bigram:

$$P(W) \approx \prod_{t=1}^{L} P(w_t|w_{t-1}) \tag{7}$$

Trigram:

$$P(W) \approx \prod_{t=1}^{L} P(w_t|w_{t-2}, w_{t-1}) \tag{8}$$

また，N-gram の遷移確率の推定には通常最尤推定が用いられる．以下に，最尤推定による単語 Bigram と Trigram の遷移確率の計算方法を示す．Bigram:

$$P(w_t|w_{t-1}) = \frac{n(w_{t-1}, w_t)}{n(w_{t-1})} \tag{9}$$

Trigram:

$$P(w_t|w_{t-2}, w_{t-1}) = \frac{n(w_{t-2}, w_{t-1}, w_t)}{n(w_{t-2}, w_{t-1})} \tag{10}$$

ただし，n(#) は単語または単語列 # のデータ中の出現頻度を表す．このように，単語 N-gram の遷移確率は，単語および単語列の出現頻度から容易に計算できる．

2.3 クラス N-gram・可変長単語列 N-gram

単語 N-gram のパラメータ数，すなわち単語遷移の組合せは V^N （V は語い）であり，N を大きくするとパラメータ数が格段に多くなるため，それぞれの値の推定が困難になるという大きな問題が存在する．例えば，語いが 10,000 語の時，Trigram のパラメータ数は $10,000^3 = 10^{12}(= 1\text{兆})$ となり，それぞれのパラメータを推定するためには，数兆語からなるテキストデータが必要となるが，これほどの大規模のデータを収集することは事実上不可能に近い．このため，補完（平滑化）と呼ばれる学習テキスト上に出現しない単語遷移に対しても 0 でない確率を与える手法が提案されているが [11][12][13]，データ量が少ない場合は信頼できる確率を与えることは困難と考えられる．

この問題点を解決する手段ために，パラメータ数を削減することを目的とした次の 2 種類のモデルが提案されている．

1. クラス N-gram[14][15][16]

   複数の単語をまとめてクラスとして扱い，クラス間の遷移を考えることによりパラメータ数を削減し，推定量の信頼性を高めるものである．L 単語からなる文の生成確率は一般に下式で表される．

   $$P(w_1^L) = \prod_{t=1}^{L} P(w_t|c_t) \cdot P(c_t|c_{t-N+1}^{t-1}) \tag{11}$$

   （c_t は w_t の属するクラスを，c_i^j は i 番目から j 番目の単語列に対応するクラス列を表す）

   上式で，$P(c_t|c_{t-N+1}^{t-1})$ は直前の *N-1* 単語列に対応するクラス列から次の単語の属するクラスへの遷移確率を表し，$P(w_t|c_t)$ は，次クラスから次単語が出現する確率を表す．クラス数が 100 の時，Trigram の全てのクラス間の遷移の組は $100^3 = 10^6(= 100\text{万})$ であるから，単語 N-gram に比べてパラメータ数は極めて少なく，比較的信頼できる遷移確率が求めることができる．しかし，単語間特有の連接関係を表現することができないため，全体としては．次単語の予測精度は悪くなると考えられる．

2. 可変長単語列 N-gram[17][18][19]

可変長単語列 N-gram は，単語列を結合させて N-gram の単位として扱うもので，固定長の N-gram と比較して，局所的に N を大きくさせる効果があり，パラメータ数の増大を抑えながら，より長い単語間の関係を表現するものである．L 単語からなる文の生成確率は一般に下式で表される．

$$P(w_1^L) = \prod_{t=1}^{K} P(ws_t | ws_{t-N+1}^{t-1}) \tag{12}$$

但し，ws_t は文章の t 番目の単語列（単独の単語も含める）を意味する．また，K は単語列の個数を表し，$K \leq L$ である．日本語の場合は，単語の境界が曖昧であり，また，「～ではないでしょうか」「～られませんでした」等，複数の単語の結合により一つのまとまった機能を有する単語列が多く存在するため，N-gram の単位の再決定という点からも非常に有効な手法であると考えられる．しかし，パラメータ数は，同次元の単語 N-gram より多くなり，少量の学習データから，信頼性の高いパラメータ推定を行うのは困難である．

2.4 品詞と可変長単語列の複合 N-gram

2.4.1 品詞と可変長単語列の複合 N-gram の概念

「品詞と可変長単語列の複合 N-gram」[20] は，クラス N-gram と可変長単語列 N-gram とのそれぞれの長所を生かしながら，それぞれの短所を補い合えるよう，出現頻度の低い単語はまとめて品詞クラスとして扱い，出現頻度の高い単語は品詞クラスから分離させ独立して扱い，さらに出現頻度の高い単語列を結合させたものである．従って，低出現単語への信頼性を高めると共に，高出現単語に関する次予測精度を高めることができる．Bigram を例にして，単語 N-gram，クラス N-gram，可変長単語列 N-gram，複合 N-gram との比較を図1に示す．

複合 N-gram は，品詞クラス，単語，単語列を同時に扱うため複雑なモデルとなるが，表現を簡単にするため，複合 N-gram を次の 3 種類のクラス間の N-gram として表現する．

A) 品詞クラス

B) 独立した 1 単語のみで構成されるクラス

C) 1 連接単語列のみで構成されるクラス

このクラス分類を用いると，複合 N-gram による文の生成確率は，下式のクラス N-gram の形で与えることができる．

$$P(w_1^L) = \prod_{t=1}^{K} P(ws_t | c_t) \cdot P(c_t | c_{t-N+1}^{t-1}) \tag{13}$$

但し，ws_t は文章を上記のクラス分類を用いた場合の，t 番目の単語列（単独の単語も含める）を意味する．また，K は文章の単語列の個数を表し，$K \leq L$ である．例として，次の文章（7 単語）を考える．

「わたくし - 橋本 - と - 言 - い - ま - す」

「橋本」は出現頻度が高くないため．固有名詞クラスとして扱う方が適切であると考えられる．「わたくし」および「と」は日本語の文章で頻繁に出現する単語であるため．品詞クラスより分離して単独で扱う．また，「言 - い - ま - す」は日本語で頻繁に用いられるフレーズであるため．結合させて一単位として扱う方が効果的であると考えられる．従って，この文章の生成確率は，次の式で与えられる．

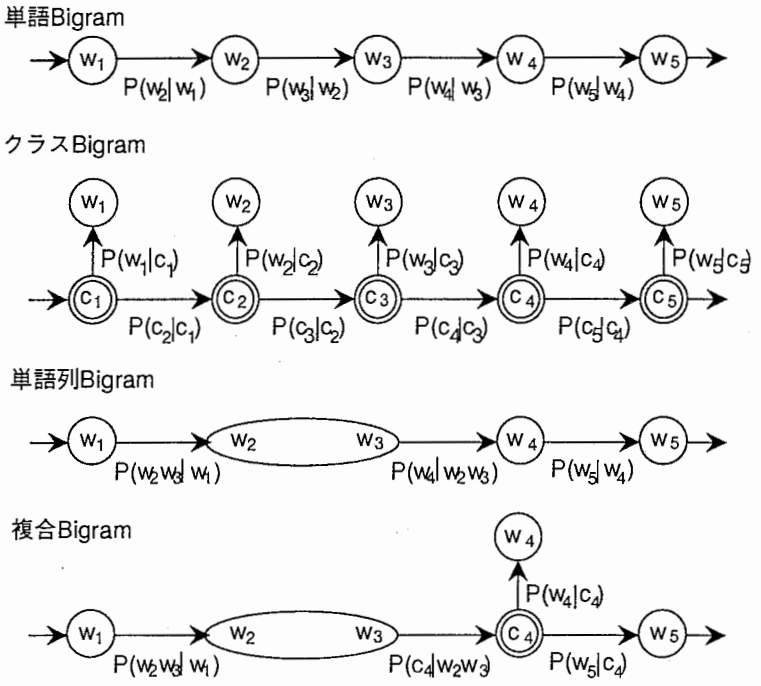

図 1: 各種 Bigram の確率計算方法の比較

$$
\begin{aligned}
P(w_1^L) &= P(\text{わたくし}|\{\text{わたくし}\}) \cdot P(\{\text{わたくし}\}) \\
&\cdot \quad P(\text{橋本}|\langle\text{固有名詞}\rangle) \cdot P(\langle\text{固有名詞}\rangle|\{\text{わたくし}\}) \\
&\cdot \quad P(\text{と}|\{\text{と}\}) \cdot P(\{\text{と}\}|\langle\text{固有名詞}\rangle) \\
&\cdot \quad P(\text{言います}|[\text{言います}]) \cdot P([\text{言います}]|\{\text{と}\})
\end{aligned}
$$

但し，$\langle\rangle\{\}[]$ はそれぞれ，クラス A) B) C) に属していることを表す．B) および C) のクラスは，単語（列）とクラスの出現頻度は等しいため ($P(w_t) = P(c_t)$)，上式は次のように変形することができ，複合 N-gram と等価であることがわかる．

$$
\begin{aligned}
P(w_1^L) &= P(\text{わたくし}) \\
&\cdot \quad P(\text{橋本}|\langle\text{固有名詞}\rangle) \cdot P(\langle\text{固有名詞}\rangle|\text{わたくし}) \\
&\cdot \quad P(\text{と}|\langle\text{固有名詞}\rangle) \\
&\cdot \quad P(\text{言います}|\text{と})
\end{aligned}
$$

2.4.2 エントロピー最小化基準に基づく複合 N-gram の自動生成

より少ないパラメータで次単語予測精度の高い効率的な複合 N-gram を得るためには，初期クラスから独立させる単語，および結合させる単語列を適切に選択する必要がある．このため，品詞クラスを初期クラスとし，

初期クラスからの単語独立によるクラス分離，および単語列結合によるクラス分離の2種類のクラス分離を逐次的に行うことによって，複合N-gramのためのクラス分類を決定する方法を提案する．単語独立，および単語列結合候補の決定は，式14により求められるエントロピーを最小にさせる候補を1つのみ選択する．

$$H\quad(\{c_i\}) = -\sum_i P(c_i)$$
$$\sum_k P(ws_k|c_j)\cdot P(c_j|c_i)\log_2\{P(ws_k|c_j)\cdot P(c_j|c_i)\}$$
$$where\ ws_k \in c_j \tag{14}$$

エントロピーはあいまいさを表す尺度であり，また，エントロピーをHとしたときパープレキシティは2^Hで与えられる．すなわち，エントロピーが小さいことはあいまいさが小さく，また，次単語予測の分岐も少なく，言語モデルの精度が高いことを意味する．従って，クラス分離を行う際に，常にエントロピーを最小にする候補を選択する本手法は，より少ないパラメータで精度の高い複合N-gramを生成するための最も適した手法であると考えられる．なお，本手法において，エントロピーの減少は常に正になることが保証されており，クラス分離によって，学習データに関してエントロピーは単調に減少する．

次に，生成アルゴリズムの詳細を示す（図2参照）．

1. 初期設定

   対象とする全単語を品詞クラスに分類する．

$$\forall x(1 \le x \le V)\ \ w_x \in c_\xi$$

(Vは語い)

2. クラス分離

   次の (a) ～ (c) の手順でクラス分離を行なう

   (a) 分離クラス候補のリストアップ

   i.ii. の2種類のクラス分離を考える

   i. 単語の品詞クラスからの分離

   品詞クラスに属している任意の単語に対して，その単語の独立したクラスと，残りの単語からなる品詞クラスとに分離する．

$$c_\xi \rightarrow \{w_x\} \oplus \{c_\xi \backslash w_x\}$$

c_ξは単語w_xの属するクラス

$a\backslash b$ はaの要素からbを除くこと意味する

   ii. 連接単語の結合によるクラス分離

   既に初期クラスより分離されている単語クラス，および単語列クラス間の任意の2クラスに対して，結合して生じる新たな単語列クラス，その単語列に連接しない単語クラスとに分離する．

$$\{w_x\} \oplus \{w_y\} \rightarrow \{w_x w_y\} \oplus \{w_x \widetilde{w_y}\} \oplus \{\widetilde{w_x} w_y\}$$

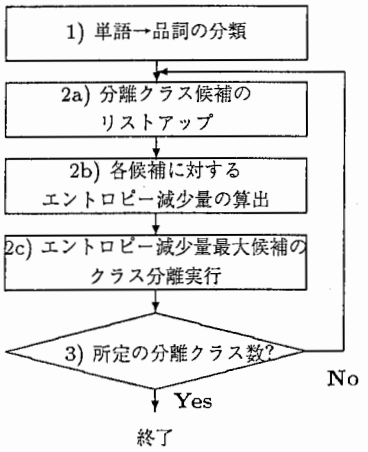

図 2: 複合 N-gram の自動生成アルゴリズム

$\{w_x w_y\}$は連接単語列$w_x w_y$のクラス

$\{w_x \widetilde{w_y}\}$は単語w_yが後続しない単語w_xのクラス

$\{\widetilde{w_x} w_y\}$は単語w_xより接続しない単語w_yのクラス

本式は，単語の結合に関する式であるが，単語列と単語の結合，および単語列と単語列との結合も同様に候補とする．

(b) エントロピー減少量の計算

(a) の各分離候補に対して，エントロピー減少量を計算する．

i. 単語の品詞クラスからの分離

$$\Delta H = H(\{c_i\}) - H(\{c_i \backslash c_\xi\} + \{w_x\} + \{c_\xi \backslash w_x\})$$

ii. 連接単語の結合によるクラス分離

$$\Delta H = H(\{c_i\}) - H(\{c_i \backslash (\{w_x\} \oplus \{w_y\})\} + \{w_x w_y\} + \{w_x \widetilde{w_y}\} + \{\widetilde{w_x} w_y\})$$

(c) 分離クラスの決定

(a) の全候補内で，(b) で計算したエントロピー減少量を最大にするものを 1 候補のみ選択し，実際にクラス分離を行う．

3. 生成終了

分離クラス数が所定の個数に達したら生成終了．そうでない場合は，（2）のクラス分離を繰り返す．

3 品詞と可変長単語列の複合 N-gram を用いた形態素解析

3.1 緒言

N-gram は当初，英語の連続音声認識に対して適用され，その有効性が示された．英語の文章は，単語がスペースで区切られており，テキストデータから単語を単位とした N-gram が容易に構築できる．しかし，日本語の文章は文字が連続しており単語の境界が明らかではなく，テキストデータのみでは単語 N-gram を構築することはできない．この問題を解決するため，文字を単位とした N-gram も提案されている [21][22][19]．しかし我々は 3.2 節で述べるように，文字を単位とする場合の問題を回避するため，形態素を単位とした N-gram の利用を考える．

形態素を単位とした N-gram を構築する場合，テキストデータに形態素を付与する，いわゆる形態素解析を行う必要がある．しかし，N-gram を構築するのに必要な，大量のテキストデータを全て人手で形態素解析を行うには多大な労力と時間が必要であり，また，かなりの経験がある人が作業を行わなければ，付与された形態素の揺れも大きくなると考えられる．従って，大量のデータをより正確に形態素解析を行うためには，自動的に形態素解析する手法が望ましい．自動形態素解析は，従来人手で作成したルールにより解析を行う方法が主流であったが，ルールの作成の作業は相当の知識・経験が必要であり，また，話し言葉等のより自然な文を全てカバーできかつ矛盾のないルールを作成するのは困難であると考えられる．これに対し，本章では N-gram 統計に基づく形態素解析手法を考える．N-gram は統計的言語モデルであり，データから容易に構築可能であるため，ルールの作成等の複雑な作業の必要がなく，また，より自然な文に対しても，データさえ収集できれば容易に適用可能である．3.3 節では，N-gram を用いた形態素解析のしくみを説明する．

本研究では，形態素解析のための N-gram 言語モデルとして，より少ないデータから精度の高い予測精度の言語モデルを得るため，品詞と可変長形態素列の複合 N-gram[20] を用いることを提案する．複合 N-gram は，基本的には品詞を単位とした N-gram であるが，言語モデルとしての精度を高めるため，特定の形態素は品詞クラスから分離させ独立して扱い，さらに特定の形態素列を結合させて新たな単位として扱うモデルである．このため，品詞という単位では表現できない形態素独自の特徴を表現でき，かつ長い範囲の形態素間の連接関係を効率良く表現することができるモデルである．

通常連続音声認識では，辞書に登録されている語いを対象とした認識が行われている．しかし形態素解析では，大量のテキストデータをまとめて処理するため，辞書に登録されていない未知語が含まれている場合も多く存在する．このため，形態素解析においては，未知語を含む文に対しても正確に処理が行えることが重要であると考える．本研究では，品詞から未知語が出現確率する確率を考えることにより，未知語の形態素解析も行えるよう，3.4 節で定式化を行った．本研究で使用した複合 N-gram は，品詞を基本単位とした N-gram であるため，このような未知語処理が容易である．

3.5 節では，形態素解析実験により，形態素 N-gram や品詞 N-gram に対する複合 N-gram の有効性を示す．

3.2 形態素を単位とした連続音声認識

日本語の文は連続した文字列から構成されており，単語の区切りがないため，英語の場合とは異なり，単にテキストデータのみを収集しても，連続音声認識用の N-gram を構築することはできない．この問題点を解決するため，文字あるいは文字列を単位とした手法が提案されているが [21][22][19]，これらは以下のような問題がある．

- 読みの付与が困難

  多くの漢字は複数の読みが可能であり，また，助詞の「は」は'w a'，「へ」は'e' と発音されるなど，文

字だけでは，正確な読みを与えることができない．文字に可能な全て読みを与えてしまうと，本来の読みとは異なる読みをも許すことになり，例えば文献 [21] に示されているように，「大切 (t a i s e ts u)」という語に対して「t a i g r i」のような誤った読みをも許してしまうため，認識率が低下すると考えられる．

- ポーズ (無音) の位置が不明
  連続音声といっても，発声の中には通常ポーズが挿入される．文字を認識の単位とした場合，全ての文字の前後にポーズの挿入が可能であるとすると，連続音声認識の探索空間が大きくなってしまい，また，認識率の低下にもつながると考えられる．文字列を単位とした手法も提案されているが [19]，この場合は，文字列の途中でポーズが挿入された場合，その文字列は認識できないという問題が発生する．

- 言語理解システムとの相性が悪い
  音声認識のアプリケーションを考えた場合，文字入力を音声で行うディクテーションシステムの他に，機械翻訳等の言語理解システムとを組み合わせた音声理解システムが考えられる [8]．言語理解の手法としては，形態素解析・構文解析を行う手法が盛んに用いられているが，文字を単位とした音声認識システムでは，認識誤りが生じた場合，誤りの前後では形態素解析の精度が著しく低下すると考えられる．

以上の理由から，文字を単位とした音声認識では，良好な結果を得ることは困難である．また，言語理解と組み合わせた音声理解システムを構築する場合も，良い結果は得られないと考えられる．

これに対して，形態素を単位としたN-gramを考えた場合，事前に形態素に読みを与えておく必要はあるものの，各文字毎に読みを与える方法よりも正確な読みの付与が可能であると考えられる．また，通常ポーズは形態素の前後に挿入されると考えられ，各文字の前後にポーズが挿入される可能性があるとする方法よりも，音声認識の探索空間を狭めることができ，また認識率も向上すると考えられる．さらに，音声理解システムを考えた場合，音声認識の単位を形態素とすることにより，認識結果に対し言語理解部で改めて形態素解析を行う必要がない．以上より，連続音声認識では，形態素を認識単位とする方が望ましいと考えられる．

3.3 N-gramによる形態素解析

日本語の形態素解析は，文の文字列 L から，それに対応する形態素列 W_L を獲得することである．統計的手法では，L に対して最も高い確率を与える形態素列 W_L を探索することにより形態素解析を実現する．これは，以下の式で与えられる．

$$W_L = \arg\max_W P(W|L) \tag{15}$$

ベイズ則により本式は下式のように変形される．

$$W_L = \arg\max_W \frac{P(L|W)P(W)}{P(L)} \tag{16}$$

本式において，$P(L)$ は右式の最大値を与えるためには無関係な量である．従って，式16は下式と等価となる．

$$W_L = \arg\max_W P(L|W)P(W) \tag{17}$$

右辺の確率 $P(L|W)$ は形態素から文字列を与える確率であるが，これは，形態素の表記と文字列が一致する場合は必ず1であり，一致しない場合は0である．また，確率 $P(W)$ は，形態素列 W の生成確率である．従って，統計的手法による形態素解析は，与えられた文字列と一致する全ての形態素列の中から，生成確率が最も高くなる形態素列を探索することによって実現できる．

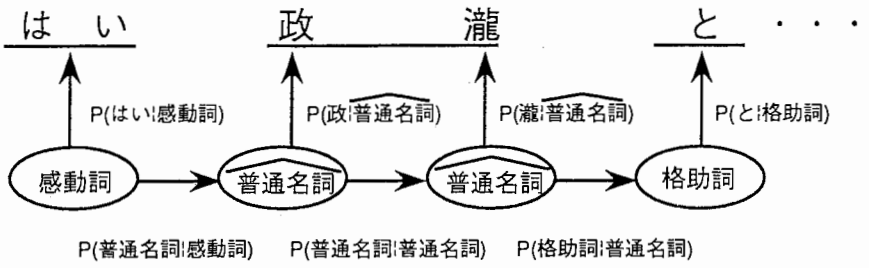

図 3: 未知語の形態素解析処理

形態素列 W を $w_1, w_2, \ldots w_m$ とすると，その生成確率 $P(W)$ は次のように表される．

$$P(W) = \prod_{t=1}^{m} P(w_t|w_1^{t-1}) \tag{18}$$

本式において，w_x^y は x 番目 y から番目までの形態素列を表す．右辺の確率を直接求めるのは困難であるから，各形態素は，直前の N-1 形態素から確率的に予測できると近似する．これが，形態素 N-gram である． N-gram を用いると，上式の形態素列 W の遷移確率は次のように近似される．

$$P(W) = \prod_{i=1}^{m} P(w_i|w_{i-N+1}^{i-1}) \tag{19}$$

3.4 未知語を含んだ文の形態素解析

未知語の形態素解析を行うために，品詞クラス c_ξ に対して，同一品詞の未知語のためのクラス $\hat{c_\xi}$ を導入する．クラス $\hat{c_\xi}$ は，任意の文字を 1 次を出力するクラスであり，同一未知語クラス $\hat{c_\xi}$ が連続した場合は，それらをまとめて一つの未知語とみなす．図 3に，「政瀧」という未知語を含んだ文の品詞 Bigram を使用した形態素解析の処理例を示す．以下に，$\hat{c_\xi}$ に関する確率の導出を行う．

Turing 推定によると，データ上に r 回出現する形態素は，次式の r^* 回と推定される．

$$r^* = (r+1)\frac{n_{r+1}}{n_r} \tag{20}$$

ただし，n_r はデータ上に r 回出現した形態素の種類数を表す．従って，r 回出現する形態素 w の品詞からの出現確率 $P(w|c_\xi)$ は，

$$P(w|c_\xi) = \frac{r^*}{N(c_\xi)} \tag{21}$$

となる．これを，クラス c_ξ に属する全ての形態素について計算し，1 から引いた残りが品詞 c_ξ から未知語出現する確率 $P(\hat{c_\xi})$ である．

$$P(\hat{c_\xi}) = 1 - \sum_{w \in c_\xi} P(w|c_\xi) \tag{22}$$

品詞 c_ξ の未知語の文字 l の出現する確率 $P(l|\hat{c_\xi})$ は，全ての文字が等しい確率で出現すると仮定し，未知語出現確率 $P(\hat{c_\xi})$ から均等に割り当てる．

$$P(l|\hat{c_\xi}) = \frac{P(\hat{c_\xi})}{V} \tag{23}$$

ただし，V は文字の種類数とする．

表 1: 各種言語モデルの形態素解析性能比較 (品詞のみの評価)

	形態素 N-gram	品詞 N-gram	複合 N-gram （分離クラス数）			
			500	1000	1500	2000
Bigram	98.90	98.56	<u>99.13</u>	99.07	99.02	99.01
Trigram	98.95	98.94	<u>99.17</u>	99.08	99.01	99.03

また，$P(\hat{c_\xi}|\hat{c_\xi})$ は，未知語が連続する確率であるが，未知語の長さが二項分布に従うと仮定すると，その品詞に属する語 w の文字列長 $len(w)$ より下のように求められる．

$$P(c_\xi|c_\xi) = \sum_{w(\in c_\xi)} \frac{len(w)-1}{len(w)} \tag{24}$$

3.5 評価実験

3.5.1 各種 N-gram モデルの形態素解析性能評価

自然発話旅行会話データベース [23] を用いて形態素解析の評価実験を行った．本データベースには，間投詞や感動詞のほか，ら抜き表現，助詞落ち等の自然発話特有の言語現象が頻出する．データベースは，1,334 対話，44,091 文，559,711 形態素から成り，語いは 7,724 語である．．このうち，約 4 分の 1 (334 対話, 11,321 文，137,691 形態素) を評価用データとし，残り (1, 000 対話，32,770 文，402,020 形態素) を言語モデル学習に使用した．

形態素解析精度の比較対象として，複合 N-gram と形態素 N-gram，および品詞 N-gram を構築した．複合 N-gram は，活用形，および活用型を含めた 234 品詞を初期状態とし，最大 2,000 クラスまで分離を行い，500 分離おきにデータを採取した．また，形態素 N-gram，品詞 N-gram，複合 N-gram ともに，形態素および品詞クラスの遷移確率を back-off Smoothing [12] により学習データに出現しない形態素および品詞クラス遷移に対して 0 でない確率を与えた．また，本節の実験では，辞書には学習データ，評価データに出現する全ての形態素が登録されており，未知語は存在しない．ただし，学習データに出現しない形態素（未学習語）に対する遷移確率は，全てのモデルにおいて 1/(100 * 語い) という確率を与えた．

形態素の正解率の評価には，音声認識で広く用いられている単語正解率 (Accuracy) にならい形態素 Accuracy を用いた．形態素 Accuracy(%) は下式で表される．

$$100 \times \frac{W-S-D-I}{W} \tag{25}$$

ただし，W: 正解の形態素数，S: 置換誤り形態素数，D: 削除誤り形態素数，I: 挿入誤り形態素数を表す．

通常，形態素解析では，形態素の分割が正しく，かつ付与された品詞が正しければ，正解とみなされれる．この場合の形態素 Accuracy(%) を表 1に示す．また，形態素解析結果を音声認識に用いることを考えると，同一品詞の形態素でも読みが異なるものは，別単位として扱うことが好ましい．形態素に読みまで考慮した場合のた場合の形態素 Accuracy(%) を表 2に示す．

表中で下線を付した値が，その次数の複合 N-gram の最高の形態素正解率を示す．表 1および 2より，いずれの評価の場合も，複合 N-gram の最も高い形態素正解率は，同次数の形態素 N-gram および品詞 N-gram よりも高い正解率を得ることができ，複合 N-gram の他のモデルに対する優位性が実験的に示された．形態素正解率最高の値を与える分離クラス数は，品詞のみの評価の場合は分離クラス数 500，読みも含めた評価の場合は

表 2: 各種言語モデルの形態素解析性能比較 (品詞と読みを含めた評価)

	形態素 N-gram	品詞 N-gram	複合 N-gram （分離クラス数）			
			500	1000	1500	2000
Bigram	98.54	96.78	98.62	98.64	98.62	98.63
Trigram	98.64	97.12	98.68	98.68	98.61	98.66

表 3: 複合 N-gram の学習データ量と形態素解析性能の関係

	データ量						
全学習データに対する割合	1/64	1/32	1/16	1/8	1/4	1/2	1/1
学習形態素数	6,306	14,293	25,931	50,794	101,227	200,105	402,020
複合 Bigram(500)	94.45	95.87	96.97	97.53	98.02	98.45	98.62
複合 Bigram(1000)	94.27	95.66	96.85	97.50	97.98	98.41	98.64

分離クラス 1000 であり，それ以上増やしても逆に形態素正解率は低下する傾向にある．これは，クラス数が増加すると共に，パラメータ数も増加するため，各パラメータの確率推定が正しく行われないことに起因すると考えられる．

3 種類のモデルを比較すると，品詞 N-gram は読みを含めた評価の場合に他のモデルと比較して形態素正解率が著しく低下している．これは，ある形態素の読みはその前後の形態素の読みに影響されると考えられるが，品詞という枠組みでは，前後の読みの関係が表現できないためと考えられる．形態素 N-gram と複合 N-gram では，読みまで含めた形態素を単位として扱うことができるため，このような大きな低下は見られない．また，複合 N-gram と形態素 N-gram との正解率の差は大きくはないが，3.4 節で示した未知語処理の容易さとを考えると，複合 N-gram が有利である．

3.5.2 学習データ量と形態素解析率との関係

前節の実験で，約 40 万語のデータより構築した複合 N-gram モデルは，読みまで考慮した形態素解析率が 98% 以上の，高い解析率が得られることが分かった．しかし，40 万語の形態素データを集めることは容易ではなく，連続音声認識に使用する N-gram を学習するための，大量の形態素データを容易に集めるという本研究の目的と矛盾する．従って，データ量が少ない時にどの程度の形態素解析率が得られるかは，本研究の趣旨において重要なことである．これを調査するため，前節の実験で用いたデータを量を 1/2,1/4 から最小 1/64 とした時の形態素正解率を調べた．言語モデルには，複合 Bigram の分離クラス数 500 と 1000 を用い，形態素正解率は読みも含めた場合の形態素 Accuracy で評価した．実験結果を 3に示す．

表 3より，データ量が減少するに比例して，形態素正解率は低下することが分かる．しかし，データ量が全体の 1/64 の場合は，形態素数がわずか 6,306 であるが，このような非常に少ない量の学習データから構築したモデルでも，94% 程度の比較的高い正解率が得られる．94% の形態素正解率は自動で形態素解析を行うには高い精度とは言えないが，自動形態素解析の結果を見て，人手で誤り個所を修正するような，半自動の形態素解析としては，使用に耐える性能であると考える．

全学習データを使用した場合は，複合 Bigram の分離クラス数 1000 の場合が分離クラス数 500 の場合より

表 4: 未知語を含む文の形態素解析性能

品詞 Bigram	複合 Bigram（分離クラス数）			
	500	1000	1500	2000
97.66	98.31	98.26	98.24	98.26

も正解率が高いが，データ量が減少するにつれて，正解率は逆転している．これは，パラメータ数の多い分離クラス 1000 のモデルでは，データ量が少ない場合では，正確なパラメータ値を推定することが困難になることが原因であると考えられる．

以上より，大量の形態素データを得るためには，まず，一万形態素程度のデータを人手で作成し，クラス数の少ない複合 N-gram を構築して半自動の形態素解析を行い，数十万形態素程度のデータが集まった段階で，クラス数の大きい複合 N-gram を構築し，その後は自動で形態素解析を行う，というのが効果的な手段であると考えられる．

3.5.3 未知語を含む文の形態素解析結果

次に，未知語を含む文の形態素解析実験を行った．学習・評価には，6.1 節の実験と同一データを使用した．ただし，辞書には学習データに出現した形態素しか登録しておらず，評価データのみにしか出現しない形態素が未知語となる．このような未知語は 632 語い存在した．未知語処理は，3.4 節に示した方法で行った．ただし，形態素 N-gram は，この処理は行えないため，品詞 N-gram と複合 N-gram のみで比較実験を行った．ただし，処理時間の都合上，両モデル共に Bigram のみを用いた．また，形態素解析の評価は，本アルゴリズムでは，未知語に対して読みを付与することはできないため，品詞のみの評価による形態素 Accuracy(%) で評価した．表 4に結果を示す．

未知語処理を行った場合でも，複合 Bigram が品詞 Bigram よりも高い正解率を得た．辞書に全語いが登録されている 6.1 節の実験では正解率が 99.13% であったから，0.8% 程度低下はしているものの，98% 以上の比較的高い正解率が得られた．また，未知語の形態素解析誤りを分析したところ，「防音」が「防」と「音」のように，1 形態素が複数の形態素に分割された例が多数見られた．これは，「音」という形態素が辞書に登録されているため「防」という文字のみが未知語として解析された結果生じた現象である．「防」も「音」も両方普通名詞であるから，これらの語を結合させることにより，誤りを低減することが可能であると考えられる．

3.6 結言

本章では，連続音声認識用の N-gram 言語モデルを構築するのに必要な形態素データを大量に収集することを目的として，品詞と可変長形態素列の複合 N-gram を用い，テキストデータから自動で形態素解析を行う方法を示した．形態素解析実験の結果，最大 99.17% の精度であり，読みまで考慮した結果でも最大 98.68% の精度を得ることができた．これは，従来の品詞 N-gram および形態素 N-gram よりも高い精度であり，提案手法の有効性が示された．また，実験により，一万語程度の少ない学習データから学習したモデルでも 94% 程度の比較的高い精度が得られることを確認した．さらに，品詞から未知語の出現確率を考えることにより未知語を含む文の形態素解析が行えるよう改良を行い，実験の結果，未知語が登録されている場合と比較して形態素解析精度の低下は 0.8% 程度であることを確認した．

今後の課題としては，形態素解析に最適な複合 N-gram の分離クラス数を自動決定することが重要と考える．また，未知語に関しては，同一品詞の未知語と既知語とを結合させ，新たな未知語と考えること等により形態素解析率を向上させ，さらに，音声認識に直接活用できるよう，未知語に対して読みを自動的に付与する手法の開発も行いたい．

表 5: 単語 Bigram と複合 Bigram との性能比較

	単語 Bigram	複合 Bigram （分離クラス数）				
		0	500	1000	1500	2000
パープレキシティ	20.02	40.27	19.34	17.64	16.98	16.75
パラメータ数	51,018	8,109	28,046	42,700	52,586	60,748

表 6: 単語 Trigram と複合 Trigram との性能比較

	単語 Trigram	複合 trigram （分離クラス数）				
		0	500	1000	1500	2000
パープレキシティ	15.13	38.76	16.65	15.16	14.61	14.48
パラメータ数	148,711	19,086	104,043	140,131	159,781	173,129

4 複合 N-gram による連続音声認識

4.1 緒言

近年，大語い連続音声認識の言語モデルとして N-gram が盛んに用いられているが，N-gram を構築するためには，通常新聞記事等の数千万語の大量のテキストデータが用いられる．しかし，音声認識のアプリケーションとしては，ATR のホテル予約タスク等，使用するタスクを限定したシステムが盛んに研究されている．こういったシステムでは，新聞記事等の通常のテキストデータとは異なる話言葉を扱わなければならず，また，発話の内容も新聞記事とはかなり異なったものであるため，新聞記事で構築した N-gram を用いても高い認識率を得ることはできないと考えられる．このため，使用するタスクに合わせてデータを取ることになるが，その場合，数千万語といった大量のテキストデータを収集するのは困難である．このため，筆者は，通常の単語 N-gram に変わり，少ない量のテキストデータからも高い精度が得られるよう，第 2 章で提案した品詞と可変長単語列の複合 N-gram を連続音声認識の言語モデルに使用し，実験によりその有効性を示す．

4.1 節では，まず，パープレキシティにより言語モデルの評価を行い，4.2 節で連続音声認識に適用し，その効果を示す．

4.2 言語モデルの評価実験

提案言語モデルの性能を確認するため，パープレキシティ，およびパラメータ数について従来の単語 N-gram との比較実験を行った．実験に用いたデータはＡＴＲの自然発話旅行会話データベース [23] で，現在，27,054 文，475,009 単語（6,396 異なり語）から構成される．このうち，約 4 分の 1（6,763 文，118,732 語）をランダムに選択して評価用テストセットデータとし，残り（20,291 文，356,277 語）を言語モデルの学習用として使用した．学習データ量は新聞記事 [5] 等と比較すると極めて少量なため，Bigram で比較実験を行った．

複合 Bigram は，活用形，および活用型を含めた 156 品詞を初期状態とし，最大 2000 クラスまで分離を行い，100 分離おきにデータを採取した．また，複合 Bigram・単語 Bigram ともに，クラス，および単語の遷移確率を削除補完法 [11] により学習データに出現しない単語遷移に対して確率を与えた．

複合 Bigram のテストセットパープレキシティの値の変化の様子を単語 Bigram の値と共に図 4に示す．

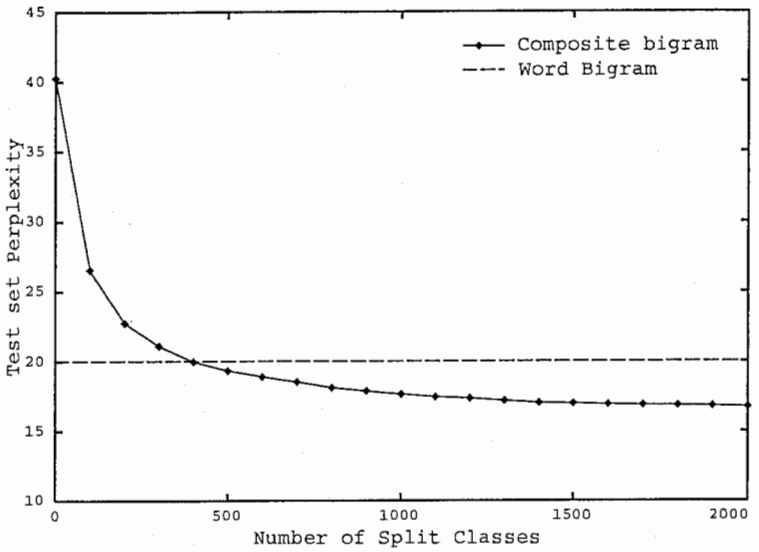

図 4: テストセットパープレキシティの変化

図 4より，複合 Bigram のテストセットパープレキシティは，分離クラス数が増加するに従って単調に減少している．分離クラス数が 400 の時単語 Bigram と同程度の値となり，分離クラスがそれ以上の場合は従来の単語 Bigram よりもパープレキシティが低くなる．従って，分離クラスが特定の個数以上で，複合 Bigram は単語 Bigram よりも高い精度を持つことができる．但し，複合 Bigram は，分離クラス数が 1,500 を越えると，パープレキシティの減少が頭打ち状態となるため，この辺りが本モデルの性能の限界であると考えられる．

表 5には，パープレキシティ，およびパラメータ数について，単語 Bigram と分離クラス数 500 おきに採取した複合 Bigram との比較を示す．但し．パラメータ数は．学習テキスト上で実際に観測した単語遷移（複合 Bigram の場合は．品詞・単語列の遷移含む）の数を表す．

表 5 より，複合 Bigram(500) （括弧内の数字は分離クラス数を表す．以下同様）は単語 Bigram よりもパープレキシティ値が 3% 低いが，パラメータ数は単語 Bigram に比べて 45% も少ない．また，複合 Bigram(1000) は単語 Bigram よりもパープレキシティは 12% 低いが，パラメータ数は 14% 少ない．従って，複合 Bigram は，与えられたパラメータで極めて効果的な表現が可能なモデルであることが確認できた．

また，各モデルの頑健性を確認するため，学習データ量を変化させた時のテストセットパープレキシティ値の比較を図 5に示す．但し，図 5において，横軸の値は学習に用いた単語数を表す．

図 5より，全ての学習データを用いた時は，単語 Bigram と複合 Bigram(500) とのパープレキシティの差は僅か 0.7 である．しかし，学習単語数を減少させた時，複合 Bigram(500)，は単語 Bigram よりもパープレキシティの増加は比較的小さい．これらの事実が示すように，複合 Bigram はパラメータ数が少ないため，頑健性が高く，少ない学習データでも従来の単語 Bigram と比較して，精度の高い言語モデルが構築できる．

N を増加した場合の様子を知るため， Trigram の比較を表 6に示す．複合 Trigram の作成にあたっては，単純に複合 Bigram と同じクラス分類を用いた．表 6より，複合 Trigram は，クラス数 1000 で単語 Trigram と同程度のパープレキシティを示し，それ以上のクラス数では単語 Trigram よりも低くなっている．式 (14) は Bigram のエントロピー計算であるため，これを基準としたクラス分類は Bigram に対して最適なクラス分類となるが，これを， Trigram でのエントロピー計算を行うことにより，複合 Trigram の性能はさらに向上すると考えられる．

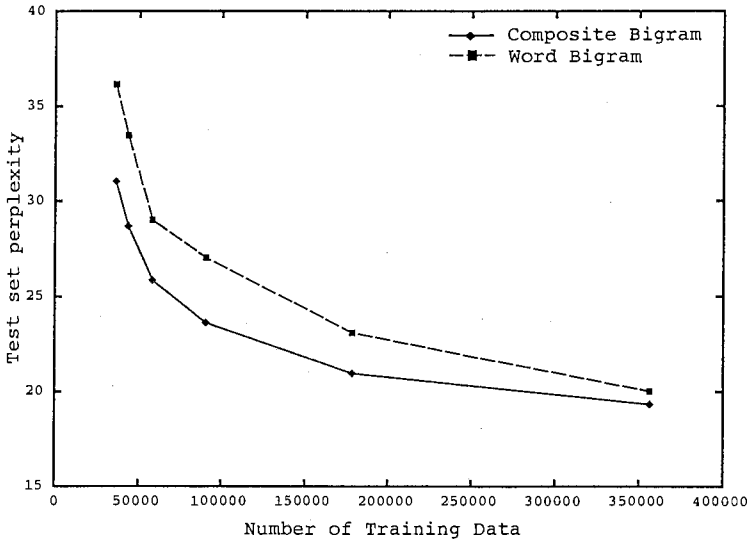

図 5: 学習データ量とパープレキシティとの関係

4.3 連続音声認識の評価実験

連続音声認識に適用し，言語モデルの効果を確認した．表 7に示す条件に基づいて音響モデルを作成し，単語グラフによる連続音声認識法 [7] により認識候補を探索し，単語 Bigram，複合 Bigram（分離クラス 0,500,1000）との性能比較を行った．認識対象文は，データベース中のホテル予約タスクより選択した 16 対話であり，これらの会話については，言語モデルの学習の対象外である．また，認識対象語は次の 2 通りで実験した．

- 辞書 Fullset:

  データベースに出現する全ての単語（6,396 語）

- 辞書 Subset:

  ホテル予約タスクに出現する単語（1,321 語）

各言語モデルで尤度 1 位の文認識候補の単語 Accuracy を表 8に示す．なお，メモリ容量と計算時間の都合上で，Fullset の辞書では，単語 Bigram の認識実験を行うことができなかった．

表 8より，辞書 Subset の場合の比較では，複合 Bigram で分離クラス数 500，および 1000 の場合で，単語 Bigram よりも認識率がおよそ 4 ～ 5% 程度向上しており，連続音声認識結果においても，複合 Bigram が優れていることが確認できた．複合 Bigram(500) と単語 Bigram とでは，パープレキシティの差は小さいが，複合 Bigram では，語尾・格助詞等，比較的出現頻度の高い短い単語が結合されており，それらの単語の湧きだし誤りの発生が抑えらることが認識率の向上に寄与していると考えられる．

4.4 結言

本章では N-gram 言語モデルの次単語予測精度，およびパラメータ推定の信頼性向上の両立を図るため，品詞および可変長単語列を単位とする複合 N-gram 言語モデルを連続音声認識に適用し，その効果を検討した．言語モデルの評価実験の結果，品詞および可変長単語列を単位とする複合 N-gram は，，パープレキシティ値の比較により，従来の単語 N-gram よりも次単語予測精度が高いことがわかり，また頑健性が高く，少ないデー

表 7: 音響分析条件

分析条件
サンプリング周波数 12KHz 20ms ハミング窓 , フレーム周期 10ms
使用パラメータ
16 次 LPC ケプストラム +16 次 Δ ケプストラム + log パワー + Δlog パワー
音響モデル
HMnet(ML-SSS) [24][25]
男女別不特定話者モデル [26]
800 状態 , 5 Mixture

表 8: 単語認識率の比較

	単語 Bigram	複合 Bigram （分離クラス数）		
		0	500	1000
辞書 Subset	54.62	48.46	59.74	58.13
辞書 Fullset	-	47.28	58.23	60.41

タ量でも比較的精度の高いモデルが構築できることを確認した．また，連続音声認識に適用した結果，通常の単語 Bigram よりも高い認識性能を得ることが確認できた．また，インプリメントの観点からも，パラメータ数の少ない複合 Bigram は単語 Bigram よりも有利であり，大容量のメモリを必要とする大語いの連続音声認識システムの構築も容易である．

今後の課題としては，初期クラスの細分化，クラスまでも含めた可変長統計等により，言語モデルの精度をさらに向上させる予定である．

5 最大事後確率推定による N-gram 言語モデルのタスク適応

5.1 緒言

N-gram は言語モデルはパラメータ数が多く，それぞれの値を正確に求めるためには，莫大な量のテキストデータが必要とされる．この問題を解決する方法として，学習テキストに出現しない単語遷移に対しても遷移確率を与える平滑化の手法や [11][12][13]，クラス分類，可変長 N-gram 等パラメータの数を減少させる手法 [14][27][20] 等が数多く提案されている．しかし，これらの手法を用いても，精度の良い言語モデルを構築するためには，相当量のデータを用いる必要があると考えられる．

現在，実用化に向けて研究が行われている連続音声認識システムは，タスクを限定しシステムの性能を向上させている場合が多い．しかし，タスク毎に大量の言語データを集めるのは困難である．特に，日本語の場合は，英語等のように単語の区切りが明確ではなく，通常人間が手作業で単語の切り出し・形態素解析の作業を行うため，大量のデータを集めるのはさらに困難である．しかし，データ量を増やすために他のタスクのデータを用いる場合，言語的特徴はタスク毎に異なるため，単純にデータを混合しても目的のタスク特有の言語特徴を効果的に表現することはできないと考えられる．

これらの問題を解決する手段として，言語モデルのタスク適応を考える．すなわち，目的のタスク以外のデータも含めた大量のデータを学習することによりデータ量の問題を解決し，得られたモデルの言語特徴を目的のタスクに適応させる方法である．タスク適応の手法として，複数の N-gram の遷移確率を線形結合する方法が提案されているが [28][29][30]，本研究では，最大事後確率推定（以後 MAP 推定と略：Maximum A-posteriori Probability Estimation）を用いた手法を提案する．提案する手法は，複数のタスク毎の遷移確率の分布を事前分布とし，目的のタスクのデータを観測データとみなすことにより MAP 推定を適用する手法であり，データ量に応じて安定性の高いパラメータ推定を行うことが可能である．

5.2 MAP 推定による N-gram 遷移確率

5.2.1 MAP 推定の概念

通常，N-gram の遷移確率は，最ゆう推定 (以後 ML 推定と略（Maximum Likelihood Estimation）) を用いて推定される．ML 推定では，観測したサンプル値 x に対して，ゆう度関数 $f(x|p)$ を最大にさせる値として確率値 p_{ML} が定められる．

$$p_{ML} = \arg\max_{p} f(x|p) \tag{26}$$

N-gram の場合，推定対象の確率 p は，直前の N-1 単語列 h から次の単語 w への遷移確率 $p(w|h)$ である．観測サンプル，すなわちテキストデータ x において単語列 h が $c(h)$ 回観測され，その内単語 w が後続する場合（確率 p）が $c(h,w)$ 回，w 以外の単語が後続する場合（確率 $1-p$）が $c(h)-c(w)$ 回であるから，ゆう度関数 $f(x|p)$ は

$$f(x|p) = p^{c(h,w)}(1-p)^{c(h)-c(h,w)} \tag{27}$$

となる．$f(x|p)$ の最大化条件 $d\log f(x|p)/dp = 0$ を解くことにより，N-gram の遷移確率は次のように計算される．

$$p_{ML}(w|h) = c(h,w)/c(h) \tag{28}$$

従って，もし単語列 h,w が観測データ上で出現しない場合，$c(h,w) = 0$ であるから，遷移確率は 0 と推定されてしまう．

これに対して，MAP 推定では，観測したサンプル値 x に対して，事後関数 $l(p|x)$ を最大化する値として，確率を推定する．

$$p_{MAP} = \arg\max_p l(p|x) \tag{29}$$

Bayes' 則を用いると，本式は次のように変形される．

$$p_{MAP} = \arg\max_p f(x|p)g(p) \tag{30}$$

ここで，$g(p)$ は，確率 p に関して何らかの情報より先見的に得られる事前分布である．すなわち，MAP 推定を用いると，N-gram の遷移確率はある事前知識より得られる分布 $g(p)$ に従う変数とし，この事前分布と実際に観測されたサンプル値とを用いて，実際の遷移確率が推定される．このため，観測データで出現しない単語遷移に対しても，事前知識により 0 でない遷移確率を与えることができる可能性がある．

5.2.2 MAP 推定による N-gram の遷移確率の導出

本節では，MAP 推定による N-gram の遷移確率を求める方法を示す．但し，変数の定義は前節と同じものを用いる．

まず，遷移確率 p の事前分布 $g(p)$ としてベータ分布（$ap^{\alpha-1}(1-p)^{\beta-1}$，$a$ は正規化のための定数）を用いる．ベータ分布を用いる理由は次の 2 点である．

- 2 項分布の共役分布で，MAP 推定によるパラメータの解が直接計算可能である．
- パラメータ α, β を変化させることにより，様々な形状の分布を表すことができる．

式 (30) の MAP 推定の定義に従うと，遷移確率 p は，ゆう度関数 $f(x|p)$ と事前分布 $g(p)$ とを用いて次のように求められる．

$$\begin{aligned} &p_{MAP}(w|h) = \\ &\arg\max_p p^{c(h,w)}(1-p)^{c(h)-c(h,w)} a p^{\alpha-1}(1-p)^{\beta-1} \\ &\qquad \equiv \arg\max L(p) \end{aligned} \tag{31}$$

$L(p)$ が最大となるための条件 $d\log L(p)/dp = 0$ を p について解くと，単語 Bigram の遷移確率は次のように求まる．

$$p_{MAP}(w|h) = \frac{c(h,w)+\alpha-1}{c(h)+\alpha+\beta-2} \tag{32}$$

α，および，β は，事前分布であるベータ分布のパラメータであるが，これらは，次のように求めることができる．

ベータ分布の平均 μ および 分散σ^2 は以下の式となることが知られている [13].

$$\mu = \frac{\alpha}{\alpha+\beta} \quad , \quad \sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{33}$$

これらの式を $\alpha, \alpha+\beta$ について解くと，

$$\alpha = \frac{\mu^2(1-\mu)}{\sigma^2} - \mu \quad , \quad \alpha+\beta = \frac{\mu(1-\mu)}{\sigma^2} - 1 \tag{34}$$

が得られる．

以上より，事前分布の平均 μ・分散 σ^2 を式 (34) に代入することにより α および $\alpha+\beta$ が得らる．これらの値と，テキストデータから単語列 h および h, w の頻度を求めることにより得られる $c(h)$, $c(h,w)$ とを式 (32) に代入することにより，MAP 推定による単語 N-gram の遷移確率 $P(w|h)$ を求めることができる．

5.3 MAP 推定によるタスク適応

5.3.1 タスク適応における MAP 推定の事前・事後知識

MAP 推定を行うためには，事前分布および観測サンプルを定義する必要がある．本研究では，MAP 推定をタスク適応に応用するため，事前分布をタスク毎の遷移確率の分布とし，また，観測サンプルを目的のタスクのテキストデータと定める．この定義の下で，MAP 推定によるタスク適応 N-gram の遷移確率を求める手順を以下に示す．

複数のタスク毎の遷移確率の分布を MAP 推定に用いる事前分布とした場合，これをベータ分布に従うと仮定した場合，式 (34) より，この分布の平均および分散から α および $\alpha+\beta$ を求めることができる．但し，出現頻度を考慮するため，平均および分散は加重平均および加重分散を用いる．これらの値は，次式により計算される．

$$\mu = \sum_i c_i(h) p_i(w|h) / \sum_i c_i(h) \tag{35}$$

$$\sigma^2 = \sum_i c_i(h) p_i(w|h)^2 / \sum_i c_i(h) - \mu^2 \tag{36}$$

と表される．本式において，$c_i(h)$ はタスク i のテキスト内の (N-1) 単語列 h の出現頻度，$p_i(w|h)$ はタスク i における単語列 h から w への遷移確率である．但し，各タスク語との遷移確率 $p_i(w|h)$ は最ゆう推定によって求める．

観測サンプルを目的のタスクのテキストデータとすると，前節の $c(h)$ および $c(h,w)$ は次のように表される，

- $c(h)$: 目的のタスクのデータ中の単語列 h) の出現頻度
- $c(h,w)$: 目的のタスクのデータ中の単語列 h, w の出現頻度

これら μ , σ^2 , $c(h)$, $c(h,w)$ を前節の式 (32),(34) に代入することにより，N-gram 遷移確率 $p(w_n|w_1^{n-1})$ が得られる．6図に，これらの一連の手順を図示した．

この遷移確率の計算を，全ての N 単語列の組み合わせ h, w について行うことにより MAP 推定によるタスク適応 N-gram を生成することができる．

5.3.2 Back-off Smoothing による遷移確率の平滑化

前節で，MAP 推定によるタスク適応の基本原理を述べたが，実際に言語モデルとして使用するには，2つの問題がある．1つは，平滑化の問題である．不特定タスクの大量のデータを用いても，出現しない単語列が存在し，MAP 推定の事前分布の平均・分散を求めることができない．従って，平滑化によりテキストに出現しない単語組に対して，遷移確率を与える必要がある．もう1つの問題は，本研究で提案するタスク適応手法は，全ての遷移確率を独立に求める手法であるため，遷移確率の和が1になるとは限らない．連続音声認識等に適用する際は特に支障はないが，パープレキシティ等の計算により他の言語モデルとの比較を行う場合は，正しい評価ができない．本節では，近年盛んに用いられている Back-Off 平滑化の手法 [12] を応用して，これらの問題を解決する方法を示す．

遷移確率を求めようとする N 単語列 $w_1^n (= h, w)$ が不特定タスクデータに含まれる場合は，既に示した MAP 推定によるタスク適応手法により遷移確率 $p_{MAP}(w_n|w_1^{n-1})$ を求め，Turing 推定により，確率 $p_{MAP}(w_n|w_1^{n-1})$ $(= p_{MAP}(w|h))$ を小さく (Discounting) し，遷移確率 $P_s(w_n|w_1^{n-1})$ とする．但し，Discounting の係数は全タスクのデータにおける単語列 w_1^n の出現頻度 $c_I(w_1^n)$ を用いて計算する．Discounting により生じた確率の余剰分を w_1^n が不特定タスクデータに含まれない単語連鎖に対して，(n-1)-gram の遷移確率 $P_s(w_n|w_2^{n-1})$ に比

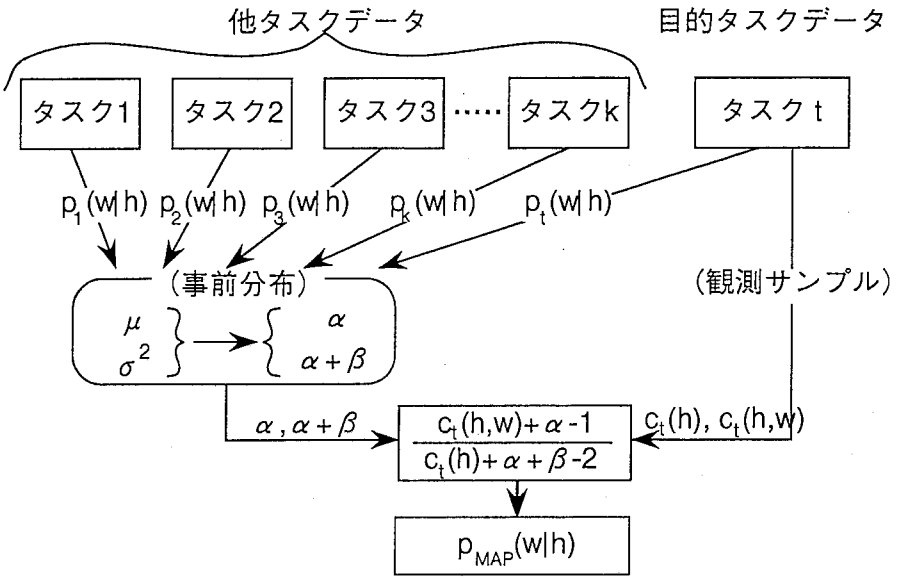

図 6: MAP 推定によるタスク適応 N-gram の遷移確率算出

例して配分する．こうして，より低次の N-gram の遷移確率を再帰的に割り当てることにより高次 N-gram の遷移確率を求める．

以上をまとめると，タスク適応 N-gram の平滑化後の遷移確率 $P_s(w_n|w_1^{n-1})$ は次式で表される．

$$P_s(w_n|w_1^{n-1}) =$$

$$\begin{cases} \cdot\, c_I(w_1^{n-1}) > 0 \text{ の場合} \\ \quad \tilde{P}(w_n|w_1^{n-1}) \\ \cdot\, c_I(w_1^{n-1}) = 0, c_I(w_2^{n-1}) > 0 \text{ の場合} \\ \quad \alpha(w_1^{n-1}) P_s(w_n|w_2^{n-1})) \\ \cdot\, c_I(w_1^{n-1}) = 0, c_I(w_2^{n-1}) = 0 \text{ の場合} \\ \quad P_s(w_n|w_2^{n-1}) \end{cases} \tag{37}$$

上式で，$\tilde{P}$ はタスク適応により得られる確率に Discount 係数をかけたものである．

$$\begin{aligned} &\tilde{P}(w_n|w_1^{n-1}) \\ &= \frac{c_I(w_1^n)+1}{c_I(w_1^n)} \cdot \frac{n_{c_I(w_1^n)+1}}{n_{c_I(w_1^n)}} \cdot p_{MAP}(w_n|w_1^{n-1}) \end{aligned} \tag{38}$$

但し，n_k は不特定タスクテキスト中に k 回出現する単語列の種類数である．また同式で，$\alpha(w_1^{n-1})$ は正規化のための係数であり，次のように求められる．

$$\begin{aligned} &\alpha(w_1^{n-1}) \\ &\quad = \frac{1 - \sum_{w_n : c_I(w_1^n)>0} \tilde{P}(w_n|w_1^{n-1})}{1 - \sum_{w_n : c_I(w_1^n)>0} \tilde{P}(w_n|w_2^{n-1})} \end{aligned} \tag{39}$$

表 9: タスク一覧

タスク番号	会話数	内容
1)	491	ホテルのサービス
2)	351	ホテルの部屋の予約
3)	50	旅行パックの問い合わせ
4)	36	ホテルの会議室の相談・予約
5)	28	交通手段の問い合わせ
6)	24	ホテルの部屋の相談
7)	22	飛行機のフライトの予約
8)	22	バス・列車の切符の問合せ
9)	20	レンタカーの問い合わせ
10)	14	コンサートのチケットの予約
11)	12	レストランの予約
12)	8	トラブル・忘れ物
13)	8	料理の注文
14)	8	道案内
15)	4	ショッピング

以上の Back-off 平滑化を応用した手法を用いることにより，求める N-gram 遷移確率の N 単語遷移がデータ中に出現しない場合は，(N-1)-gram 以下の低次の遷移確率によって確率値を与えることができる．また，式 (39) において α を求める際に正規化を行うため，遷移確率の和は自動的に 1 に正規化される．

5.4 評価実験・考察

5.4.1 パープレキシティを評価基準としたタスク適応の効果

提案したタスク適応の有効性を確認するため，評価実験を行った．実験用いたデータは，ＡＴＲ自然発話データベース [23] で，1,098 会話，449,070 単語 (のべ)，6,797 (異なり) 単語からなる．また，このデータベースは表 9に示すように，15 タスクから構成されている．これらのデータの内，約 4 分の 1 の会話をランダムに選んでテストセットとして，残りの会話を学習セットとして使用した．但し，各タスクから最低でも 1 会話はテストセットとして選択している．

言語モデルとしては，次の 3 種類のモデルを考える．

- 不特定タスクモデル：
  全タスクのテキストで作成した N-gram
- 特定タスクモデル：
  各タスクのテキストのみで作成した N-gram
- タスク適応モデル：
  MAP 推定により各タスクに適応させた N-gram

これらのモデルをタスク毎に，単語 Bigram, Trigram で作成した．モデル・タスク毎のテストセットパープレキシティ値を表 10に示す．

表 10: 各モデルのタスク別パープレキシティ

番号	単語数		不特定タスクモデル		特定タスクモデル		タスク適応モデル	
	Training	Test	Bigram	Trigram	Bigram	Trigram	Bigram	Trigram
1)	136,175	42,698	23.177	17.954	22.922	18.261	22.159	17.391
2)	118,124	38,697	14.844	10.080	13.843	9.942	13.404	9.553
3)	19,471	6,610	26.539	17.398	23.934	17.201	20.042	14.364
4)	15,302	5,075	31.285	24.706	38.158	32.851	27.994	23.041
5)	10,791	2,983	24.191	16.563	21.772	16.577	17.471	13.187
6)	8,802	2,999	17.136	11.199	14.666	11.402	11.867	8.712
7)	8,617	2,722	21.114	14.186	18.391	14.649	14.644	11.060
8)	8,537	2,193	21.148	14.296	14.225	11.307	12.769	10.208
9)	8,567	2,528	25.171	18.167	26.007	20.822	19.141	14.922
10)	5,036	1,608	16.592	10.832	14.060	10.929	10.915	7.815
11)	5,326	1,439	12.982	8.887	12.179	9.631	9.350	6.908
12)	3,578	1,165	32.918	19.395	25.369	18.372	18.034	12.756
13)	2,378	1,075	30.288	22.405	34.246	32.202	18.185	16.735
14)	2,572	908	35.545	27.109	46.611	42.088	25.396	21.947
15)	1,750	509	44.156	34.232	47.543	44.545	25.948	23.186
(平均)			25.139	17.827	24.928	20.719	17.821	14.119

タスク適応モデルのパープレキシティは，不特定タスクモデルと比較して，タスク全体の平均で約 29% (Bigram)，21% (Trigram) 低くなっている．特定タスクモデルと比較しても，約 29% (Bigram) ，31% (Trigram) 低く，提案したタスク適応手法により言語モデルの精度が向上することが確認できた．また，タスク適応モデルのパープレキシティは，全てのタスクで Bigram，Trigram の両方において，不特定タスクモデル・特定モデルのいずれよりも低く，安定してタスク適応の効果が得られることが分かった．

不特定タスクモデルと特定タスクモデルのパープレキシティを比較すると，Bigram では，特定モデルのパープレキシティの方が不特定モデルよりも低い値を示す場合が多いが，Trigram では，不特定タスクモデルの方が特定タスクモデルよりも低い場合が多い．これは，単語 Bigram では，学習データのスパース性が低いため，特徴を表すことのできる特定タスクモデルの方が有利であるが，Trigram では，学習データがよりスパースであるため，特定タスクの少ない量のデータでは，信頼できるパラメータ推定が行われていないことが原因と考えられる．従って，タスク適応を行うと，大量のデータを用いたことにより，学習データのスパース性が解決でき，さらに，適応を行うことにより，そのタスクの言語特徴を表現できたものと考えられる．

データ量が少ない 12),15) 等のタスクでは，タスク適応によるパープレキシティの減少の割合が大きくなる傾向にある．特にタスク 15) では，不特定タスクモデルと比較して 41%(Bigram) および 32%(Trigram)，特定タスクモデルと比較して 45%(Bigram) および 48%(Trigram) パープレキシティの減少が非常に大きい．すなわち，目的のタスクのデータが少量しか集まらない場合に，タスク適応を使用する効果が大きいと言える．

表 11: 各モデルの単語認識率の比較

不特定タスク Bigram	特定タスク Bigram	タスク適応 Bigram
66.65	73.62	74.82

5.4.2 連続音声認識におけるのタスク適応の効果

連続音声認識におけるタスク適応の効果を調べた．認識の対象は，前節のタスク 8)，バス・列車の切符の問合せタスクとした．音響パラメータ・音響モデルは第 4 章で行った実験と同一のものを使用した．認識解の探索には，単語グラフサーチ [7] を用いた．特定タスクモデルはそのタスクに出現しない単語への遷移確率が 0 になり，音声認識においては，不特定タスクモデル・タスク適応モデルと比較して有利になると考えられるため，認識対象語いを 8) タスクに出現する 713 語に限定した．言語モデルの学習には，表 10と同じく 8,537 単語を用い，評価データの 2,193 単語について認識を行った．不特定タスクモデル・特定タスクモデル・タスク適応モデル単語 Bigram モデルについて，単語 Accuracy(%) による性能比較を行った結果を表 11に示す．

表 11より，タスク適応 Bigram は不特定タスク Bigram よりも 8.17% 高く (改善率 24.9%)，また，特定タスク Bigram よりも 1.2% が高い (改善率 4.5%) 認識率が得られた．タスク適応 Bigram は，不特定タスク Bigram より，認識率が大幅に向上したことより，タスク適応の効果は示せた．しかし，特定タスク Bigram と比較すると，認識率の向上は小さかった．すなわち，特定タスクモデルは学習データがわずか 8,537 単語しかないのにかかわらず，比較的高い認識率が得られている．これは，データベースを収録する際，チェックシートを用いて会話の制御を行ったため，テストセットと評価セットで，固有名詞や，日付，電話番号等で同一単語がよく用いられ，評価セットがかなり学習セットに近い内容であったと考えられる．実際のフィールドテスト等で評価を行うと，タスク適応モデルが特定タスクモデルよりも認識率が高くなると考えられる．

5.5 結言

本研究では，最大事後確率（MAP）推定を用いることにより，大量のデータから作成される N-gram をデータ量に応じて目的のタスクに効果的に適応を行う手法を示した．実験の結果，タスク適応によるパープレキシティの減少効果が確認され，数千語程度の少量のテキストを用いるだけで，適応前のモデルよりも大幅に精度の良い N-gram が構築できることがわかった．また，連続音声認識に適用した結果，単語認識率 1.2% の向上 (改善率 4.5%) が得られ，連続音声認識においても有効であることが示せた．

今後は，連続音声認識への適用，あるいは，新聞記事等の大規模なテキストデータで作成した N-gram からのタスク適応について実験を行いたい．

6 最大事後確率推定を用いた適応によるクラスタ別 N-gram 言語モデル

6.1 緒言

前章では，特定のタスクの N-gram 言語モデルの精度を向上させるため，MAP 推定（最大事後確率推定）によるタスク適応の手法を提案している [31]．しかし，この手法には次の 3 つの問題点が考えられる．

- 複数のタスクからなるコーパスが必要

単語列毎にタスク間の N-gram 遷移確率の平均，および分散を用いてパラメータ推定を行うため，テキストコーパス全体があらかじめ複数のタスクに分割されている必要があり，コーパスが単一タスクより構成される場合，また複数タスクが存在するものの，あらかじめタスクの分類がされていないコーパスに対しては適用できないという問題点が存在した．

- タスクという分類が適切ではない

同一タスクと言っても，実際にはその中にもさまざまな内容の文が存在する．「よろしくお願いします」「はい，分かりました」のような会話に頻出するフレーズ等，似た内容の文が複数のタスクで出現すると考えられる．このため，タスクという尺度よりも，文の内容による分類の方が言語的特徴がより明確になる考えられる．

- パラメータ数が多い

MAP 推定による N-gram の遷移確率を求めるためには，単語列毎に 2 つのパラメータが存在する．このため，通常の単語 N-gram ではパラメータの数が多く，それぞれのパラメータについて信頼できる推定値を得ることができない．

本章では，これらの問題を解決し精度の高い N-gram 言語モデルを得るため，テキストコーパ全体を自動的にスクラスタリングし，MAP 推定を用いてそれぞれのクラスタ毎に N-gram 言語モデルを構築する方法を考える．また，従来の単語 N-gram に代り，精度を向上させるため，品詞と可変長単語列の複合 N-gram[20] を用いることを同時に提案する．6.2 節では，まず，提案する言語モデルを音声認識に適用するための手法について述べる．続く 6.3 節では，クラスタリングの手法を説明する．6.4 節では，MAP 推定を複合 N-gram に適用するため，クラス N-gram に対する MAP 推定の定式化を行う．6.5 節では，パープレキシティ，および連続音声認識結果により，提案手法の有効性を示す．

6.2 連続音声認識システム概要

図 7に本稿で提案する連続音声認識システムの概要を示す．本システムの特徴は，言語モデルの学習に用いるテキストコーパスをクラスタリングし，クラスタ毎の言語特徴を明確にさせ，言語モデルの精度を向上させるものである．しかし，入力された音声の発話文が属するクラスタをあらかじめ知ることは不可能である．このため，まず最初に，入力音声をコーパス全体で作成した言語モデルで認識を行い，次に，認識結果からクラスタ別の言語モデルを 1 つのみ選択し再度認識を行う，という 2 段階の認識を考える．

中間認識結果 W からクラスタ別の言語モデル LM_C の選択は，次式のように確率が最も高いものを選択することにより行う．

$$LM_C = \arg\max_{LM_k} P(LM_k|W) \tag{40}$$

本式は Bayes' 則を用いると，次のように表される．

$$LM_C = \arg\max_{LM_k} P(W|LM_k)P(LM_k) \tag{41}$$

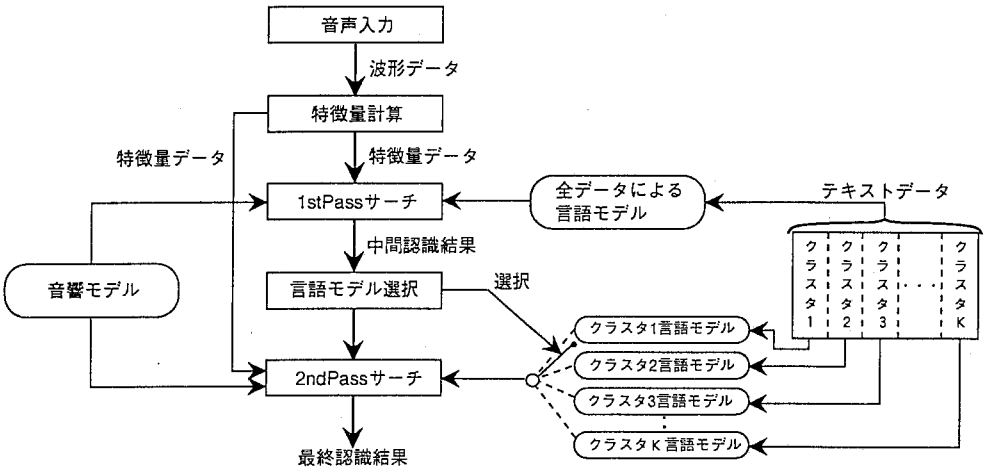

図 7: 音声認識システム概要

$P(LM_k)$ は，言語モデル LM_k の事前確率である．この確率は，対話システムでは前発話の内容等から推定できると考えられるが，一般的には，発話される内容を事前に求めることはでないため，現在のところこの確率は考慮しない．従って，クラスタ言語モデルは次式により選択される．

$$LM_C = \arg\max_{LM_k} P(W|LM_k) \tag{42}$$

すなわち，K 個のクラスタのそれぞれのモデル LM_k で認識結果文 W に対する生成確率 $P(W|LM_k)$ $(1 \leq k \leq K)$ を求め，確率の最も高いクラスタモデル LM_C を選択する．

6.3 コーパスのクラスタリング

コーパスを自動クラスタリングするために，K-means 法に類似した方法を用いた．K-means 法は，サンプルを距離が最も近いクラスタ中心に所属させる形でクラスタリングを行う手法である．この手法を文のクラスタリングに適用するため，次の 2 点で通常の方法と異なる．

- クラスタ中心をそのクラスタに属する文で生成される言語モデルとする．
- 距離尺度に文の生成確率 $P(W|LM_k)$ を用いる．

これらは，前節で述べた認識結果からクラスタモデルの選択で用いる手法と同一であり，妥当な方法であると考える．以下に，クラスタリングの手順を示す．

1. クラスタモデルの初期化：
   クラスタ数を K とし，適当な手法によりコーパスから K 文を選択して全クラスタに 1 文ずつ配置し，クラスタ毎の言語モデル $LM_1, LM_2, \ldots, LM_K$ を作成する．

2. クラスタの選択：
   コーパスの全文について，各クラスタにおける言語モデルの文生成確率を求め，最も確率の高いクラスタを選択し，その文を所属させる．

3. クラスタモデルの変更：
   各クラスタ毎に，3. で選択した文を用いて言語モデル $LM_1, LM_2, \ldots, LM_K$ を更新する．

4. 終了条件：

文の属するクラスタが1文も変化しない場合クラスタリングを終了する．それ以外の場合は，2.～4.の処理を繰り返す．ただし，ある程度の回数を繰り返してもクラスタリングが収束しない場合は強制終了させる．

6.4 MAP 推定による N-gram の適応

クラスタリングを行うことにより，クラスタ毎の言語的特徴は明確になるものの，クラスタ毎のデータ量は減少するため，N-gram のパラメータ推定の精度が低下することが考えられる．このため，MAP 推定を用いた適応 ([31]) を用い，パラメータ推定の精度を向上させる．

文献 [31] によると，MAP 推定による単語列 h から次単語 w への単語 N-gram の遷移確率 $P(w|h)$ は次式により与えられる．

$$P(w|h) = \frac{N(h,w)+\alpha-1}{N(h)+\alpha+\beta-2} \tag{43}$$

ただし，$N(\#)$ はそのクラスタでの単語 (列)# の出現頻度である．また，α および β は事前分布として用いるベータ分布（$ap^{\alpha-1}(1-p)^{\beta-1}$, a）のパラメータであり，次式により求められる．

$$\alpha = \frac{\mu^2(1-\mu)}{\sigma^2}-\mu \ , \quad \alpha+\beta = \frac{\mu(1-\mu)}{\sigma^2}-1 \text{と同様に与えることができる．} \tag{44}$$

上式の μ および σ^2 は，クラスタ毎の遷移確率 $P(w|h)$ の分布の平均および分散である．

文献 [20] で提案されている複合 N-gram(可変長 N-gram) は，クラス N-gram を基本としたモデルであり，遷移確率は $P(ws|c(ws))\cdot$
$P(c(ws)|c(h))$ として与えられる．ただし，ws は可変長の単語列で，$c(\#)$ は単語（列）# の属するクラスである．
$P(c(ws)|c(h))$ はクラス間の遷移確率であり，式 44と同様に与えることができる．また，$P(ws|c(ws))$ は単語列 ws の属するクラス $c(ws)$ から単語列 ws が出現する確率であり，MAP 推定により次式で与えられる．

$$P(ws|c(ws)) = \frac{N(ws)+\alpha-1}{N(c(ws))+\alpha+\beta-2} \tag{45}$$

また，Back-off smoothing[12] を用い，コーパス上に出現しなかった単語遷移に対して確率を与えるとともに，遷移確率の和が 1 になるよう確率の正規化を行う．

6.5 評価実験・考察

自然発話旅行会話データベース [23] を用いて評価実験を行った．本データベースのサイズは，1,332 対話，32,074 文，
597,626 単語で，語いは 7,221 語である．このうち評価用として「ホテルの部屋の予約」タスクから 40 対話，1166 文，
18,381 単語を選択し，残りのデータを言語モデルの学習に使用した．

最初にテストセットパープレキシティにより評価を行った．複合 N-gram は活用形・活用型を含む 158 品詞による初期クラスから，500 クラス分離を行ったモデルを使用した．クラスタ数 4, 8, 16, 32, 64 の時のクラスタモデルと，データベース全体で作成したモデル (クラスタ数 1) とのパープレキシティの比較を表 12に示す．本表より，クラスタ数に比例してパープレキシティが減少しており，クラスタ毎の言語的特徴がより明確になっていると考えられる．クラスタ数が 64 の時は，全体モデルよりもパープレキシティが約 32% 減少した．また，評価に用いた「ホテルの部屋の予約」タスクのデータは，データ量が多いため文献 [31] ではタスク

表 12: パープレキシティによる比較

全体モデル	クラスタモデル（クラスタ数）				
	4	8	16	32	64
14.21	13.00	12.33	11.44	10.44	9.72

表 13: 連続音声認識における性能比較

	全体モデル	クラスタモデル (クラスタ数)		
		4	16	64
単語認識率	77.66	78.69	79.06	78.54
文認識率	33.43	35.82	36.12	37.31

適応の効果は，単語 Bigram で 5% 程度と小さかったが，本稿で提案した手法では，文の内容毎に適応モデルを作成するため，大きな精度向上が得られたと考えられる．計算量の都合上クラスタ数は最大 64 としたが，さらにクラスタ数を増加させることにより，パープレキシティは減少すると考えられる．ただし，クラスタ数を多くしすぎると各クラスタのデータ量が少なくなりすぎ，パラメータ推定が困難になるため，限界はあると考えられる．

次に，連続音声認識に適用した際の認識率によって評価を行った．音響モデルには ML-SSS[24][25] による隠れマルコフ網
(801 状態 5 混合分布) 不特定話者モデルを用い，単語グラフサーチ [7] により認識解の探索を行った．言語モデルは，学習データ全体で作成したモデルとクラスタ数 4, 16, 64 のクラスタモデルとを比較した．表 13に単語認識率 (Accuracy)(%) 及び文認識率 (%) を示す．本表より，単語認識率はクラスタ数 16 の時に全体モデルより約 1.4% 向上 (改善率約 6%) し，文認識率はクラスタ数 64 の時に最大約 3.9% 向上 (改善率約 6%) し，連続音声認識における有効性を確認した．クラスタ数 64 の時の単語認識率はクラスタ数 4,16 の時よりも低下しているが，これは，誤認識が生じた際にクラスタモデルの選択が正しく行われないことが原因と考えられる．

6.6 結言

本章では，コーパスの各文を自動クラスタリングし，それぞれのクラスタ毎に MAP 推定による N-gram 型の言語モデルを作成することにより言語特徴をより効果的に表現できるモデルを提案した．実験の結果パープレキシティは最大約 32 %減少し，また，連続音声認識に適用した際，単語認識率及び文認識率共に最大約 6% 改善し，本手法の有効性を確認した．今後は，クラスタリング手法の改善，発話履歴よりクラスタ言語モデル事前確率 $P(LM_k)$ を用いた認識結果からクラスタモデルの選択精度の改善等により，さらなる認識精度の向上を行いたい．

7 統計的手法による音声言語理解

7.1 緒言

近年，隠れマルコフモデルによる音響モデル，およびN-gramによる言語モデルを用いた連続音声認識が盛んに研究されており，数万語いの認識で，単語認識率が90%程度とかなり実用レベルに近くなっている [4][2]. 音声認識技術を用いたアプリケーションとしては，読み上げ文をそのまま出力するディクテーションシステムに加え，旅客機案内システム [32]，電話番号案内システム [33]，音声翻訳システム [8] 等，音声認識結果を理解し，ユーザーに情報を提供するいわゆる「音声理解システム」も盛んに研究されている.

現在，音声理解システムのための言語理解の手法は，システムが扱うことのできる文型を限定したもの [32][33]，キーワードを用いた手法 [34]，文法ルールを用いて構文解析を行う手法 [35] 等が提案されている．発話内容の文型を限定する手法は，理解のための処理が容易であり理解率の向上が期待できるが，限定された文型以外の入力に対処できず，柔軟なシステム構成をとりにくい．一方，キーワードを用いた手法は，より自由な発話を扱える利点はあるが，キーワードのみでは正確な理解を得るのは困難であると考えられ，文献 [34] ではユーザーインターフェースによりキーワード間の単語の補完を必要としている．また，文法ルールを用いた構文解析による手法は，文型を限定する手法よりは，発話の自由度が高いが，文法的に正しい文章でないと理解できず，自然発話や認識誤りを含む文等の非文法的な文の理解は困難であると考えられる.

我々は，より自然な発話を扱うことができ，多少の認識誤りを含む文に対しても頑健な理解も可能であり，かつ正確な理解が可能な音声理解システムを目標とする．この目標を実現するため，本研究では，中間表現の各要素から文の生成確率を与える隠れマルコフモデル，および中間表現の各要素間の共起確率を用いた，統計処理による言語理解手法を提案する．統計処理による手法は，文章を解析して言語理解に必要な情報を自動的に学習できるため，専門家による文法ルールの作成を必要としない上に，文法として規定しにくい自由発話に対する理解も期待できる．また，隠れマルコフモデルは，文の構造を状態の遷移として表現することができるためキーワードによる理解系よりも正確な理解ができる可能性があり，また，文の局所的な構造をモデル化するため，局所的な音声認識誤りに対しても頑健な理解が行える可能性がある.

7.2節では，研究開発の対象とした音声理解システムの概要，および言語理解部の動作例を述べる．7.3節では，本システム中の統計的手法による言語理解手法について説明し，7.4節の実験により，本手法の有効性を示す.

7.2 音声理解システム概要

7.2.1 システム概要

今回開発した音声理解プロトタイプシステムは，スキー場の案内を対象としている．これは，入力された音声を理解し，スキー場のデータが入力されたデータベースへアクセスし，ユーザーの要求する情報を表示するシステムで，次の3種類の動作を行うことができる.

- ユーザーが要求する条件を満たすスキー場の検索 (SHOWLIST)
- 各スキー場のデータ（県・標高差・リフト数等12項目）の表示 (SHOWVALUE)
- スキー場の地図の表示 (SHOWIMAGE)

システム全体の構成を図8に示す．本システムは，主に「音声認識部」と「言語理解部」からなる.

音声認識部では，入力された波形データに対し特徴量計算を行った後，隠れマルコフ網による音響モデル [24][25]，および複合N-gramによる言語モデル [20] を用いて，単語グラフサーチ [7] により解の探索を行い，認識結果

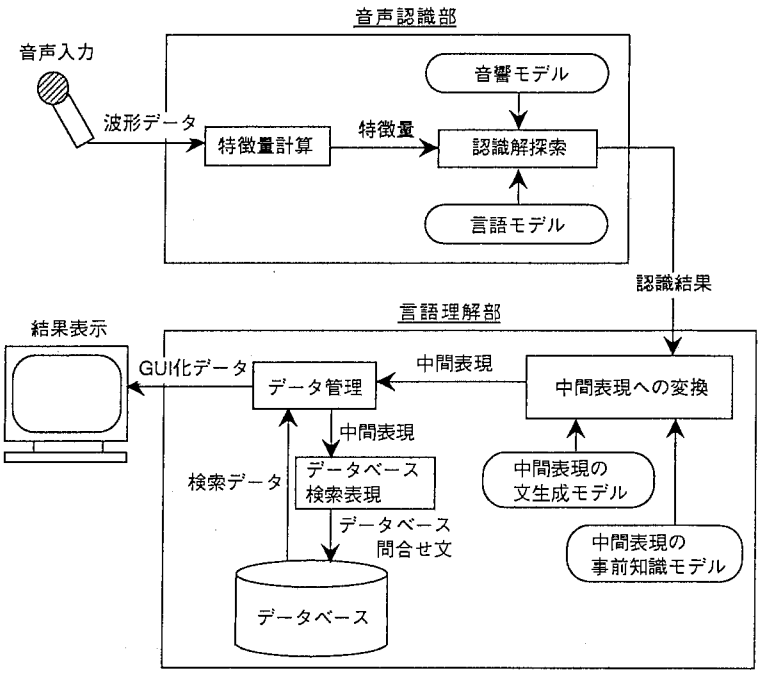

図 8: 音声理解システム概要

を出力する．言語理解部では，音声認識部で得られた認識結果の単語列を中間表現に変換する．さらにデータベース検索用表現（SQL 問合せ文）を生成することによりデータベースにアクセスし，検索されたデータをユーザーの要求に応じて表示する．

7.2.2 言語理解部概要

本システムにおける言語理解部の目的は，認識結果をデータベース検索用表現に変換することである．しかし，データベース検索用表現は，文脈に直接関係のないデータベース言語特有のキーワード等が含まれるため，認識結果文から検索用表現に直接変換するのは効率的でない．このため，本システムでは，データベース言語特有のキーワードを除いた中間表現を用い，「入力文から中間表現への変換」，「中間表現からデータベース検索用表現への変換」と二段階の変換を行うことを考える．中間表現は，データベース検索用表現へ正確かつ容易に変換できるように設計されている．また，中間表現には，データベース検索の条件と共に，ユーザーの要求動作のタイプも含まれている．

本システムで用いた中間表現は次の要素から構成される．

- R_(コマンド名):

  要求動作の指定 (Request)

- O_(対象物名):

  動作の対象 (Object)

- D_(ドメイン名):

  データベースの検索項目 (Domain)

- C_(比較方法):

  データベース検索の比較条件 (Comparison)

- V_(値):

  データベース検索の値 (Value)

中間表現は，これらの要素の列として表現され，次のフォーマットで与えられる．

R_(コマンド名) O_(対象物名)
  D_(ドメイン名) C_(比較方法) V_(値)

以下に，自然言語から中間表現への変換例を挙げる．

例 1)

入力文：

標高差が 1000m 以上のスキー場を教えて下さい．

中間表現：

$$R\_SHOWLIST\ O\_\text{スキー場名}$$
$$D\_\text{標高差}\ C\_>=\ V\_1000$$

例 2)

入力文：

八方尾根スキー場の標高差は何メートルですか．

中間表現：

$$R\_SHOWVALUE\ O\_\text{標高差}$$
$$D\_\text{スキー場名}\ C\_=\ V\_\text{八方尾根}$$

例 3)

入力文：

八方尾根のゲレンデマップを見せて下さい．

中間表現：

$$R\_SHOWIMAGE\ O\_\text{ゲレンデ地図}$$
$$D\_\text{スキー場名}\ C\_=\ V\_\text{八方尾根}$$

データベース検索用表現は SQL 言語のサブセットを用いている．中間表現からデータベース検索用表現への変換例を以下に示す．

中間表現：

$$R\_SHOWLIST\ O\_\underline{\text{スキー場名}}$$
$$D\_\underline{\text{標高差}}\ C\_\underline{>=}\ V\_\underline{1000}$$

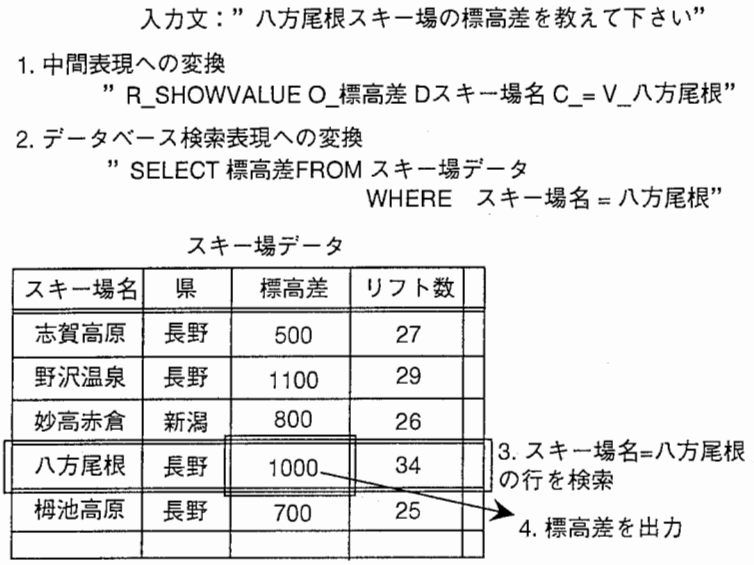

図 9: 言語理解部の動作概要

データベース検索表現：

$$SELECT\ \underline{\text{スキー場名}}\ FROM\ \text{スキー場データ}$$
$$WHERE\ \underline{\text{標高差}}\ \underline{>=}\ \underline{1000}$$

中間表現の $O\_$, $D\_$, $C\_$, $V\_$ の各要素をデータベース表現の下線の部分にはめこむだけで機械的に変換が可能である．

言語理解部の一連の動作例を図 9に示す．言語理解部は，認識結果が入力されると，次の順序で処理を行う．

1. 認識結果から中間表現への変換
2. 中間表現からデータベース検索用表現を生成
3. 条件に適合するデータをデータベースから検索し，動作の対象情報を獲得
4. 対象物名に対して中間表現のコマンド名で規定された動作を実行

7.3 自然言語から中間表現への変換

前節で述べたように，本システムでの言語理解は，「入力文から中間表現への変換」，「中間表現からデータベース検索用表現への変換」と二段階の変換を行う．中間表現からデータベース検索用表現への変換は機械的に可能であるため，入力文から中間表現への変換が，本システムの言語理解における重要な役割を果たす．本研究の主眼点は，この「入力文から中間表現への変換」を統計的手法により実現する点にある．

入力文から中間表現への統計的手法による変換の原則は，単語系列 W が与えられたとき，確率的に最ももっともらしい中間表現列 S_W を得ることである．これは，次式で表される．

$$S_W = \arg\max_S P(S|W) \tag{46}$$

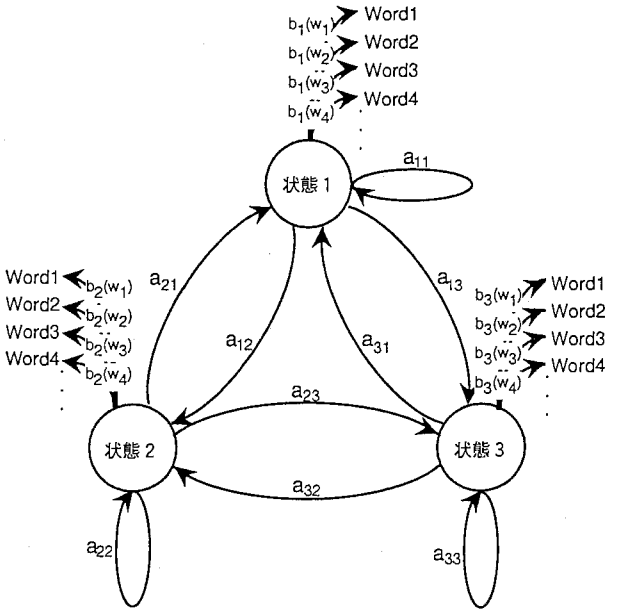

図 10: 隠れマルコフモデル

上式の右辺に Bayes' 則を用いると，次のように変形できる．

$$S_W = \frac{\arg\max_S P(W|S)P(S)}{P(W)} \tag{47}$$

右辺の $P(W)$ は最大値を求める S には無関係の量であるため，次式と等価である．

$$S_W = \arg\max_S P(W|S)P(S) \tag{48}$$

本式において，$P(W|S)$ は，ある中間表現 S から入力文の単語列 W を生成する確率である．この確率を求めるために，隠れマルコフモデルを用いる．一方，$P(S)$ はある中間表現 S が出現する確率で，入力文とは独立に求められる事前確率である．この確率を求めるために，中間表現の各要素間の共起確率を用いる．それぞれのモデルに関しては，続く 3.1 および 3.2 節で説明する．また，これらのモデルを用いて，入力文から中間表現を得る方法を 3.3 節に述べる．

7.3.1 HMM による中間表現の文生成確率

前節の確率 $P(W|S)$，すなわち中間表現から入力単語列の生成確率を与えるモデルとして，隠れマルコフを用いる．隠れマルコフモデルは，複数の状態から構成され，単語が入力される毎に，状態 i から状態 j へ確率 a_{ij} で遷移し，遷移後の状態 j から確率 $b_j(w_k)$ で単語 w_k を出力する．図 10には 3 状態の隠れマルコフモデルを示す．隠れマルコフを用いることにより，出力確率 $b_j(w_k)$ により，中間表現の各要素と単語との共起関係を表すことができ，また，状態遷移確率 a_{ij} により，単語の並びとして表される文の構造をも統計的に反映することができるため，単なるキーワードに基づく理解系に比べて，統計的により精度の高い理解が可能であると考えられる．

隠れマルコフモデルは，中間表現の各要素毎に作成し，文が入力されると，全てのモデル毎に独立に動作し，全てのモデルそれぞれが入力文の単語全体に対して生成確率を計算する．中間表現 S から入力文 W の生成確

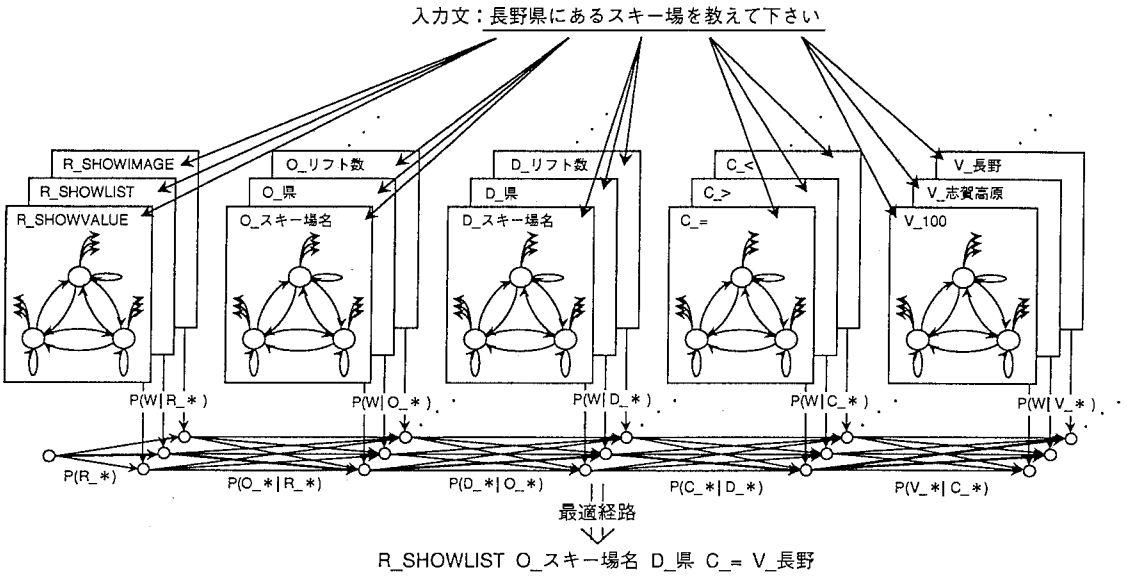

図 11: 入力文から中間表現への変換

率は，次式のように 2.2 節で示した 5 要素それぞれの文生成確率の積として近似する．

$$P(W|S) = \prod_{t=1}^{5} P(W|s_t) \tag{49}$$

但し，s_t は中間表現 S の t 番目の要素を表す．

音声認識では通常 left-to-right 型，すなわち一方通行型のモデルが盛んに使用されている．しかし，言語理解のための隠れマルコフモデルとして，図 10のように，全ての状態間の遷移が可能なエルゴディック隠れマルコフモデルモデルを用いた．これは，自然な発話多彩な言い回しに対応可能な理解モデルを構築するため，構造をあらかじめ決定することは避け，理解に必要な言語構造の特徴の獲得を，モデルのパラメータ推定により自動的に獲得することを狙ったためである．

隠れマルコフモデルのパラメータの推定には，通常最ゆう推定が用いられるが，中間表現間の各要素の特徴をより明確に表現し，識別の精度を向上させるため，本研究では識別誤り最小化 (MCE: Minimum Classification Error) 学習 [36][37] を用いた．MCE 学習は，正解と誤りとの距離を表す識別誤り関数を最小化するように行われる学習である．MCE 学習を本モデルに適用する際，中間表現の要素のグループ ($R\_*, D\_*\ldots$) 毎に行われ，正解の中間表現列に含まれる要素を正解，同一グループに属するの他の要素を誤りとして学習を行った．ただし，MCE 学習を行う際の HMM の初期パラメータは，最ゆう推定による値を用いた．

7.3.2 要素の共起確率による中間表現の事前確率の利用

中間表現は，各要素がランダムに出現する訳ではなく，例えば，ユーザーの要求がスキー場のリストを表示する (R_SHOWLIST) 場合その対象は必ずスキー場名 (O_スキー場名) になる等，中間表現の要素の共起関係が存在する．この共起関係を確率的に表現するため，中間表現の各要素間の Bigram として表される共起確率を用いた．この時，中間表現の事前確率 $P(S)$ を，次式によって求められる．

$$P(S) = P(s_1)\prod_{t=2}^{5} P(s_t|s_{t-1}) \tag{50}$$

それぞれの確率は最ゆう推定により次式で容易に求められる．

$$P(s_1) = N(s_1)/L \tag{51}$$

$$P(s_t|s_{t-1}) = N(s_t, s_{t-1})/N(s_{t-1}) \tag{52}$$

但し，$N(\#)$ は中間表現データ中の要素'#' の出現回数を表し，L は学習データの文数を表す．

7.3.3 入力文から中間表現への変換

入力文から中間表現への変換行うためには，文生成確率 $P(W|S)$，および中間表現の事前確率 $P(S)$ の積 $P(W|S)P(S)$ の最大値を与える中間表現列を求めればよい．

図 11に中間表現列を求める手法の概念図を示す．まず，文が入力されると，中間表現の各要素に対応する全ての HMM において，$P(W|s_i)$ を計算する．

次に，$P(W|S)P(S)$ を計算する．これは，式 (49)，および (50) より，下式のように求められる．

$$P(W|S)P(S) = \{\prod_{t=1}^{5} P(W|s_t)\}\{P(s_1)\prod_{t=2}^{5} P(s_t|s_{t-1})\} \tag{53}$$

右辺を変形すると，次のようになる．

$$P(W|s_1)P(s_1)\prod_{t=2}^{5}\{P(W|s_t)P(s_t|s_{t-1})\} \tag{54}$$

本式を用いると，中間表現の各要素の順番 (R_* , O_* , D_* , C_* , V_*) に HMM と共起確率との確率との積を計算し，それらの最大値を与える中間表現の要素列 $s_t(1 \leq t \leq 5)$ を求めることにより，入力文から中間表現への変換結果を得ることができる．なお，Viterbi アルゴリズムを用いることにより，この条件を満たす中間表現の要素列を容易に求めることができる．

7.4 評価実験・考察

7.4.1 言語理解部の評価実験

提案した手法による言語理解の性能を実験により評価した．言語理解部を単独に評価するため，まず，正解文からの言語理解率を評価した．実験に用いたデータは，スキー場案内システムのために収集している会話で，現在，2,700 文（34,098 単語）あり，全ての文章に，それに対応する中間表現を人手で付与している．この内，2,618 文（33,017 単語）を言語理解モデルの学習に使用し，残りの 82 文（1,081 単語）を評価用のデータとした．なお，中間表現の要素は 126 種類ある．

言語理解のためのモデルは，最ゆう推定 (ML) 学習による隠れマルコフモデル，および，識別誤り最小化 (MCE) 学習を行った 2 種類用のモデルを用意した．隠れマルコフモデルの状態数は，全てのモデルで同一数とし，1 から１０まで変化させて実験を行った．評価には言語理解率を用いた．但し，言語理解率は，入力文章から中間表現へ正確に変換できた割合であり，中間表現の全ての要素が正しく変換できた場合のみ正解とする．表 14にこれらの条件での言語理解率 (%) を示す．

隠れマルコフモデルの学習に最ゆう推定 (ML) を用いた状態数が 1 の時で理解率は 82.9% であり，状態数を大きくしても理解率は向上せずむしろ逆に低下する傾向にある．これに対して，MCE 学習を用いた場合は，状態数の増加に従って理解率が向上しており，状態数 6 の時に言語理解率は最大 91.5% まで向上している．これは，本来状態数が多くなるほど，モデルの自由度が増し，より精度の高い表現が可能であるが，MCE 学習を行うことにより，最ゆう推定では困難であった中間表現の要素の識別が効果的に行われた結果と考えられる．

表 14: 各モデルの言語理解率

	隠れマルコフモデル状態数									
	1	2	3	4	5	6	7	8	9	10
ML	82.9	80.5	80.5	78.0	81.7	79.2	79.2	79.2	78.0	80.5
ML+MCE	82.9	87.8	87.8	90.2	90.2	91.5	90.2	89.0	89.0	90.2

表 15: 音声認識実験条件

音響分析条件	
標本周波数	12 kHz
窓関数	20 ms ハミング窓
フレーム周期	10 ms
パラメータ	16 次元 LPC ケプストラム + Log パワー 16 次元 Δ ケプストラム + ΔLog パワー
音響モデル	
隠れマルコフモデル網 [24][25] 803 状態, 5 混合分布	
言語モデル	
複合 N-gram [20] 54 品詞クラス + 300 分離クラス	

状態数 1 の隠れマルコフモデルは語順を全く考慮することができず，状態数を複数にすることにより語順を考慮できるようになる．従って，状態数の増加に従って理解率が向上するという実験結果より，正確な理解のためには，出現単語の種類だけでなく，語順すなわち文構造の把握が重要であり，提案モデルは隠れマルコフモデルの状態遷移という形で理解に必要な文構造の表現が自動獲得できたと考えられる．

7.4.2 音声理解システム全体の評価実験

言語理解部を音声認識部と接続し，音声理解システムとしての性能を評価した．連続音声認識システム部の条件を表 15に示す．なお，音声認識部の言語モデルの学習には，前節の言語理解モデルの学習に用いたデータと同一データを使用した．言語理解部へは，音声認識結果のゆう度最大候補のみを用いて処理を行った．また，言語理解部の隠れマルコフモデルは，前節の実験で文理解率の最も良かった 6 状態の MCE 学習を行ったモデルを使用した．音声認識率，および音声理解率を表 16に示す．

表 16より，音声認識部の文認識率は約 60% であるのに対して，音声理解率は約 73% と文認識よりも約 13% 程度高い値を示している．すなわち，認識誤りを含む文から正しく理解された例が全体の約 13% あり，多少の認識誤りが発生しても正しい理解が得られることが確認できた．

誤認識文のうち，正しく理解できた文の代表例を以下に示す．

表 16: 音声理解率

単語認識率	文認識率	音声理解率
91.4	59.8	73.2

謝辞

研究の機会を与えて頂いた，ＡＴＲ音声翻訳通信研究所の山崎泰弘前社長（現国際電信電話株式会社研究所），および山本誠一社長に深く感謝致します．また，研究を進めるにあたって，多大な指導を頂きました第一研究室室長の匂坂芳典博士，および元主任研究員の松永昭一博士（現ＮＴＴヒューマンインターフェース研究所）に感謝いたします．

連続音声認識の実験にあたっては，清水徹研究主任（現ＫＤＤ研究所），および山本博史研究技術員にはプログラムの提供，実験の補助等多大なご協力を頂きました．プログラムの作成にあたっては岩澤亮祐氏，および高島浩司氏に補助して頂きました．また，音声理解システムの実験においては，内藤正樹研究員をはじめとした多数の方に音声データの収集に協力して頂きました．また，研究員の方々には，言語モデルに関する議論，音響モデルの提供等，研究を進める上で多いに有益な情報・データ・ツール等を提供して頂きました．また，データ整備を行って頂いたラベラーの方々，研究環境の整備を行って頂いた Technical Supporting Group の方々，事務手続きを行って頂いたＡＴＲの職員の方々には日頃大変お世話になりました．

以上記して感謝致します．

付録 A）関連発表論文

〔主論文〕

1. 政瀧 浩和, 松永 昭一, 匂坂 芳典: “品詞と可変長単語列の複合 N-gram の自動生成, ” 信学論 D-II （印刷中）.

2. 政瀧 浩和, 匂坂 芳典, 久木 和也, 河原 達也：“最大事後確率推定による N-gram 言語モデルのタスク適応,” 信学論（投稿中）.

3. 政瀧 浩和, 谷垣 宏一, 匂坂 芳典：“統計的モデルによる音声言語理解, ” 信学論 D-II （投稿中）.

4. 政瀧 浩和, 匂坂 芳典: “品詞と可変長単形態素列の複合 N-gram を用いた日本語形態素解析,” 自然言語処理 (投稿中).

〔副論文〕

1. H. Masataki and Y, Sagisaka: “Variable-Order N-gram Generation by Word-Class Splitting and Consecutive Word Grouping,” Proc. of ICASSP 96, pp. 188-191 (1996.5)

2. T. Shimizu, H. Yamamoto, H. Masataki, S. Matsunaga and Y. Sagisaka: “Spontaneous Dialogue Speech Recognition Using Cross-Word Context Constrained Word Graphs,” Proc. of ICASSP 96, pp. 145-148 (1996.5)

3. 清水徹, 山本博史, 政瀧浩和, 松永昭一, 匂坂芳典: “大語い連続音声認識のための単語仮説数削減,” 信学論 Vol.J79-D-II, No.12, pp.2117-2124 (1996.12)

4. H. Masataki, Y. Sagisaka, K. Hisaki and T. Kawahara: “Task Adaptation Using MAP Estimation in N-gram Language Modeling” Proc. of ICASSP 97 (1997.4)

〔研究会資料〕

1. 政瀧 浩和, 松永 昭一, 匂坂 芳典: “連続音声のための可変長連鎖統計言語モデル,” 電子情報通信学会技術報告, SP95-73, pp. 1-6 (1995.11).

2. 政瀧 浩和, 匂坂 芳典, 久木 和也, 河原 達也: “MAP 推定を用いた N-gram 言語モデルのタスク適応,” 電子情報通信学会技術報告, SP96-103, pp. 59-64 (1997.1).

3. 政瀧 浩和, 谷垣 宏一, 匂坂 芳典: “統計的処理による音声・言語理解モデル,” 電子情報通信学会技術報告, SP97-98, pp. 25-32 (1998.1).

〔講演報告〕

1. 政瀧 浩和, 松永 昭一, 匂坂 芳典: “連続音声認識のための品詞・単語可変長 N-gram” 日本音響学会平成 8 年度春季研究発表会講演論文集, 1-P-17, pp. 195-196 (1996.3).

2. 政瀧 浩和, 匂坂 芳典, 久木 和也, 河原 達也: “MAP 推定による N-gram 言語モデルの適応,” 日本音響学会平成 9 年度春季研究発表会講演論文集, 1-6-3, pp. 5-6 (1997.3).

3. 政瀧浩和，谷垣宏一，匂坂芳典: “統計的手法による認識結果から中間表現への変換を用いた音声理解システム,” 日本音響学会平成１０年度春季研究発表会講演論文集, 1-6-7, pp. 13-14 (1998.3).

4. 政瀧浩和 “MAP 推定に基づく N-gram 言語モデルの自動分類されたコーパスへの適応,” 日本音響学会平成１０年度春季研究発表会講演論文集, 1-6-19, pp. 41-42 (1998.3).

5. 清水 徹，山本 博史，政瀧 浩和，松永 昭一，匂坂 芳典: “単語グラフと可変長 N-gram を用いた大語彙自然発話音声認識,” 日本音響学会平成 8 年度春季研究発表会講演論文集, 1-P-18, pp. 197-198 (1996.3).

6. 谷垣宏一，政瀧浩和，匂坂芳典: “決定木を用いた発話の意味タグ推定,” 日本音響学会平成１０年度春季研究発表会講演論文集, 1-6-2, pp. 3-4 (1998.3).

7. 内藤正樹，政瀧浩和， Harald Singer， 塚田元， 匂坂芳典: “日英音声翻訳システム ATR-MATRIX における音声認識用音響・言語モデル,” 日本音響学会平成１０年度春季研究発表会講演論文集, 2-Q-20, pp. 159-160 (1998.3).

参考文献

[1] Proc. ARPT Speech Recognition Workshp, Morgan Kaufmann Publishers, 1996.

[2] P. C. Woodland et.al, "THE 1994 HTK Large Vocabulary Speech Recognition System," Proc. ICASSP95', Vol.1, pp73-76, 1995.

[3] L. R. Bahl et.al: "Performance of the IBM Large Vocabulary Continuous Speech Recognition System on the ARPA Wall Street Journal Task," Proc. ICASSP95', Vol.1, pp.41-44, May 1995.

[4] L. R. Bahl, F. Jelinek, and R. L. Mercer: "A Maximum Likelihood Approach to Continuous Speech Recognition," IEEE Trans. on Pattern Analysis and Machine Intelligence, pp 179-190, 1983.

[5] 松岡達雄, 大附克年, 森岳至, 古井貞熙, 白井克彦: "新聞記事データベースを用いた大語い連続音声認識," 信学論 Vol.J79-D-II No.12, pp2125-2131, December 1996.

[6] 西村雅史, 伊東伸泰: "単語を単位とした日本語ディクテーションシステム," 信学論 Vol.J81-D-II No.1, pp10-17, January 1998.

[7] 清水 徹, 山本 博史, 政瀧 浩和, 松永 昭一, 匂坂 芳典: "大語い連続音声認識のための単語仮説数削減," 信学論 Vol.J79-D-II, No.12, pp.2117-2124, December 1996.

[8] 竹沢 寿幸, 森元 逞, 匂坂 芳典, ニック キャンベル, 飯田 仁: "日英音声翻訳システム ATR MATRIX," 第５６回情処全大, 6Q-07, March 1998.

[9] 永井, 鷹見, 嵯峨山: "逐次状態分割法 (SSS) と音素コンテキスト依存 LR パーザを統合した SSS-LR 連続音声認識システム," 信学技報, SP92-33, 1992.

[10] R. Pieraccini and E. Levin: "Stochastic representation of semantic structure for speech understanding," Proc. EUROSPEECH-91, Vol2, pp.383-386, 1991.

[11] F. Jelinek and R. L. Mercer: "Interpolated estimation of Markov Source Parameters from Sparse Data," Proc.Workshop Pattern Recognition in Practice, pp.381-37, 1980.

[12] S. M. Katz: "Estimation of Probabilities from Sparse Data for the Language model Component of a Speech Recognizer," IEEE Trans. on Acoustics, Speech, and Signal Processing, 400-401, 1987.

[13] 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 SP95-93, pp1-6, December 1995.

[14] P. F. Brown et al.: "Class-Based n-gram models of natural language," Computational Linguistics, Vol.18, No.4, pp467-479, 1992.

[15] Masaaki Nagata: "A stochastic Japanese morphological analyzer using a forward-DP backward A^* N-best search algorithm," COLING-94, pp.201-207, 1994.

[16] 田本真詞, 川端豪: "連接共起に注目した単語のクラスタリング," 信学技報 SP93-125, pp55-62, January 1994.

[17] Giachin, E. P.: "Phrase Bigrams for Continuous Speech Recognition," Proc. ICASSP-95, Vol.1, pp225-227, May 1995.

[18] Deligne, S. & Bimbot, F.: "Language Modeling by Variable Length Sequences," Theoretical Formulation and Evaluation of Multigrams, May 1995. Proc. ICASSP-95, Vol.1, May pp169-172.

[19] 伊藤彰則, 好田正紀: "かな・漢字文字列の連鎖統計による言語モデル". 信学論 Vol.J79-D-II No.12, pp2062-2069, December 1996.

[20] 政瀧浩和, 松永昭一, 匂坂芳典: "連続音声認識のための可変長連鎖統計言語モデル," 信学技報, SP95-73, pp.1-6, 平成7年

[21] 山田 智一, 川端 豪, 松永 昭一, 鹿野 清宏: "かな・漢字の文字列連鎖情報を利用した統計的言語モデル," 信学技報 SP91-26, pp.65-72, June, 1991.

[22] 山田 智一, 川端 豪, 松永 昭一, 鹿野 清宏: "音声認識におけるカナ・漢字連鎖確率に基づく統計的言語モデル," 信学論 (A), vol.J77-A, no.2, pp.198-205, 1994.

[23] T. Morimoto et al.: "A Speech and Language Database for Speech Translation Research," ICSLP, pp1791-1794, September 1994.

[24] 鷹見淳一, 嵯峨山茂樹, "逐次状態分割法による隠れマルコフモデル網の自動生成," 信学論 Vol.J76-D-II, No.10, pp.2155-2164, October 1993.

[25] M. Ostendorf and H. Singer: "HMM Topology Design Using Maximum Likelihood Successive State Splitting," Computer Speech and language, ,11, pp. 17-41, 1997.

[26] 小坂哲夫, 鷹見淳一, 嵯峨山茂樹: "話者混合SSSによる不特定話者音声認識," 日本音響学会講演論文集 2-5-9, pp135-136, October 1992.

[27] T. R. Niesler and P. C. Woodland, "A Variable-Length Category-Based N-gram Language Model," Proc. ICASSP'96, Vol.1, pp.164-167, 1996.

[28] R. Kneser, & V. Steinbiss: "On the Dynamic Adaaptation of Stochastic Language Models," Proc. ICASSP'93, Vol.2, pp.585-588, 1993.

[29] P. Clarkson & A. Robinson: "Language Model Adaptation using Mixtures and an Exponentially Decaying Cache," Proc. ICASSP'97, Vol.2, pp.799-802, 1997.

[30] S. Matsunaga, T. Yamada & K. Shikano: "Task Adaptation in Stochastic Language Models for Continous Speech Recognition," Proc. ICASSP'92, Vol.1, pp.165-168, 1992.

[31] 政瀧 浩和, 匂坂 芳典, 久木 和也, 河原 達也: "MAP推定を用いたN-gram言語モデルのタスク適応," 電子情報通信学会技術報告, SP96-103, pp. 59-64 (1997.1).

[32] 坂井 信輔, 畑崎 香一郎, 水野 正典, 渡辺 隆夫: "音声入力を用いたパソコンネットワーク旅客機空席案内システムの試作," 信学技報, SP94-89, pp.29-36, January 1995.

[33] 内藤 正樹, 黒岩 眞吾, 武田 一哉, 山本 誠一, 谷戸 文廣: "大規模内線電話受付システムの試作," 信学技報, SP94-90, pp.37-42, January 1995.

[34] 遠藤 充, 伊藤 達朗, 星見 昌克, 二矢田 勝行: "音声による文例検索方法の検討," 音響講論 2-Q-12, pp.163-164, March 1997.

[35] S.Seneff: "TINA: A Natural Language System for Spoken Language Applications," Computational Linguistics, Vol.18, No.1, March 1992.

[36] W. Shou, C. H. Lee and B. H. Juang: "Minimum error rate training based on N-best string models," ICASSP93, pp. 652-655, 1993.

[37] 村上 哲範, 武田 一哉, 河井 恒, 山本 誠一: "行列によるトレリス計算を用いたHMMの文レベルでの識別学習," 信学技報, SP95-25, pp.1-6, July 1995.