学習データ量に対するマルコフ連鎖確率値の収束性と単語のHMMとBigramを用いた文節音声認識システムについて

A Study of Convergence Properties of N-gram Language Models for Japanese and A Phrase Recognition System using the Word Bigram Model applied to HMM Word Models

村上仁一　坪井俊明

Jin'ichi Murakami　Toshiaki Tuboi

ATR音声翻訳通信研究所　NTTヒューマンインターフェース研究所

ATR Interpreting Telephony Research Labs. NTT Human Interface Labs.

概 要

This paper describes the frequency of occurrence of n-grams of syllables, kanji-kana, part-of-speech, and word in Japanese newspaper text and X-ray CT scanning reports. An algorithm for continuous speech sentence recognition using word HMMs and a word bigram model is also discussed.

It is well known that the word bigram or word trigram models is effective tools for a speech recognition system. However, the convergence properties for n-gram probabilities have not been reported for Japanese.

In this paper, we firstly report the convergence of unigram, bigram, trigram and 4-gram language models for syllable, kanji-kana, word and part-of-speech units extracted from newspaper text and X-ray CT scanning reports.

Secondly, we report a phrase recognition algorithm using the word bigram model and word HMMs for X-ray CT scanning reports. The usual high amount of training data for word HMMs could be reduced to a single utterance using a technique known as fuzzy-vector-quantization.

A sentence recognition experiment using this algorithm was carried out to test the efficiency of the word bigram model with a vocabulary of about 3000 words. The following phrase recognition rates were obtained in this experiment: 96.8% for text-closed data and normal findings, 78.1% for text-closed data and abnormal findings, 86,5% for text-open data and normal findings, 72.1% for text-open data and abnormal findings. (The term "abnormal findings" refers to the content of the X-ray CT scanning reports.)

1　まえがき

人が発声した自然言語をコンピュータに認識させる文(文節)音声認識システムでは、認識性能を向上させるために言語情報を使用している [1]。この言語モデルとしてbigram・trigramなどのマルコフモデルが有効であることが知られている [2],[3],[4],[5],[6]。しかし信頼性のあるマルコフ連鎖確率値を得るには大量のテキストが必要である。したがって学習データ量に対するマルコフ連鎖確率値の変化について調査する必要がある [7]。

この調査を行なうためには大量のテキストデータが必要である。英語では、古くからデータベースの必要性が認識されていて、ブラウンコーパス [8] やAPコーパスなどが作成されている。しかし日本語で

は自由に扱える大規模なテキストデータベースは見当たらない。現在ＡＴＲにおいて対話文を中心とするテキストデータベースの作成が進められているが [9]、規模はまだ 100 万文字程度である。最近、新聞記事が CD-ROM で販売されているため、漢字仮名文字のデータは多く入手できるが、形態素解析はされていない。しかし現在自然言語処理の研究が進んできており、任意の漢字仮名交じり文に対して、ある程度高い精度で形態素解析ができるようになってきている [10]。

そこで本論文では、一般的な日本語として新聞記事を選び、学習データ量に対する音節および漢字仮名および品詞の unigram・bigram・trigram・4-gram のマルコフ連鎖確率値の収束性について調べる。次に、日本語の専門的な文章として X 線 CT 所見の文章を選び、新聞記事と同様に音節および漢字仮名および単語の unigram・bigram・trigram・4-gram のマルコフ連鎖確率の収束性について調査する [11]。

最後に X 線 CT 所見作成の文章の入力を目的とした特定話者の文節音声認識システムについて報告する。多くの音声認識システムでは認識単位を音素にしているが、このシステムでは認識単位として単語を選択した。この場合、調音結合などの問題が軽減されるため文節認識率は向上するが、HMM のための学習データは大量に必要であると考えられてきた。しかし実験から Fuzzy-VQ を利用することによって HMM の学習データが 1 つでも高い認識性能が得られることが示された。そして、言語情報として単語の bigram と組み合わせることにより、text-open data の正常所見で 86.5%、異常所見で 72.1%の文節認識率が得られた。

2 学習データ量とマルコフ連鎖確率値の収束性について

2.1 エンロトピーと頻度別出現率

学習データ量の変化に対するマルコフ連鎖確率の値の変化を調べるために、まず学習データ量に対するエントロピーの収束率を調査した。unigram・bigram・trigram・4-gram のエントロピーは次の式によって計算できる。

$$unigram \quad \Sigma_i p(w_i) log[p(w_i)]$$
$$bigram \quad \Sigma_{i,j} p(w_i, w_j) log[p(w_j|w_i)]$$
$$trigram \quad \Sigma_{i,j,k} p(w_i, w_j, w_k) log[p(w_l|w_i, w_j)]$$
$$4-gram \quad \Sigma_{i,j,k,l} p(w_i, w_j, w_k, w_l) log[p(w_l|w_i, w_j, w_k)]$$

ここではエントロピーの他に “頻度別出現率” も調査した。 “頻度別出現率” とは次のように定義する。

“頻度別出現率 98%” が示す値は、学習データの中で 98%をカバーするのに必要な最小のマルコフ連鎖確率の種類の数である。また “頻度別出現率 100%” が示す値は、学習データ量全てをカバーするのに必要なマルコフ連鎖の種類の数である。

調査は頻度別出現率 96%、頻度別出現率 98%、頻度別出現率 100%、およびエントロピーの合計 4 つの値で行なった。

なお以後の図 1 から 6 までの横軸は学習データ量で、縦軸は出現したマルコフ連鎖確率の種類の数およびエントロピーの値である。また図中における太い実線は頻度別出現率 96%、太い断線は頻度別出現率 98%、細い実線は頻度別出現率 100%、細い断線はエントロピーを示している。また “Entropy” の横に示した値は、全学習データを利用したときのエントロピーの値である。

2.2 新聞記事

標準的な日本語として 1982 年 1 月 4 日から 3 月 30 日までの 64 日分の日経新聞の新聞記事を選んだ。この記事を日本語形態素解析プログラムで形態素解析を行ない、音節と品詞を自動的に付与した。そして、文節に区切り、このデータから音節および漢字仮名および品詞のマルコフモデルの収束性を調べた。ただし、過度の複雑さを避けるため、記号・外国語読み・数詞の文字が存在する文は文全体を削除した。データ量は漢字仮名の文字数にして約 170 万文字である。また、使用した日本語形態素解析プログラムの形態素解析の精度は単語認定率で約 95%である [10]。表 1 に新聞記事の一部を載せる。

表 1: 新聞記事の例

大蔵省はことし四月から新銀行法が施行されるのに伴い、在日外銀の営業活動を日本の銀行同様に扱うとの基本方針を決め、これを盛り込んだ政令を二月中にも公布する。おもな内容は（１）企業向け貸し出しに対する大口融資規制を在日外銀にも適用し、五年間の猶予期間を設けるなどの配慮をする（２）利益準備金の積み立てを義務づけ、外銀に対する信頼を高める（３）邦銀の支店を買収することや現地法人化を認める――など。大蔵省はこれによって在日外銀に関する法的根拠が明確になるほか、在日外銀の国内活動がしやすくなり、欧米諸国の間に出始めているわが国の金融制度に対する不満を和らげるのに役立つとみている。（在日外国銀行は「きょうのことば」参照）

2.2.1 新聞記事における音節のマルコフ連鎖確率の収束率

音節の unigram・bigram・trigram・4-gram の学習データ量に対するエントロピーおよび頻度別出現率のグラフを図 1 に示す。音節の種類の数は、外来語を除き鼻音化したガ行を加え長音を 1 音節として 111 種類である。これらから以下のことがわかる。

1. エントロピーは比較的少ないデータで収束する。
2. 頻度別出現率 98%や 96%が収束するのに必要な学習データの量は、エントロピーを収束させるのに必要な学習データの量よりも多くのデータが必要である。
3. 頻度別出現率 100%は学習データを増やしても収束する傾向がみられない。これは、学習データを増加させるにともない、全体に占める割合は少ないが、新しいマルコフモデルの組み合わせがたえず出現することを意味している。
4. エントロピーは unigram・bigram・trigram・4-gram になるにしたがい低下する。

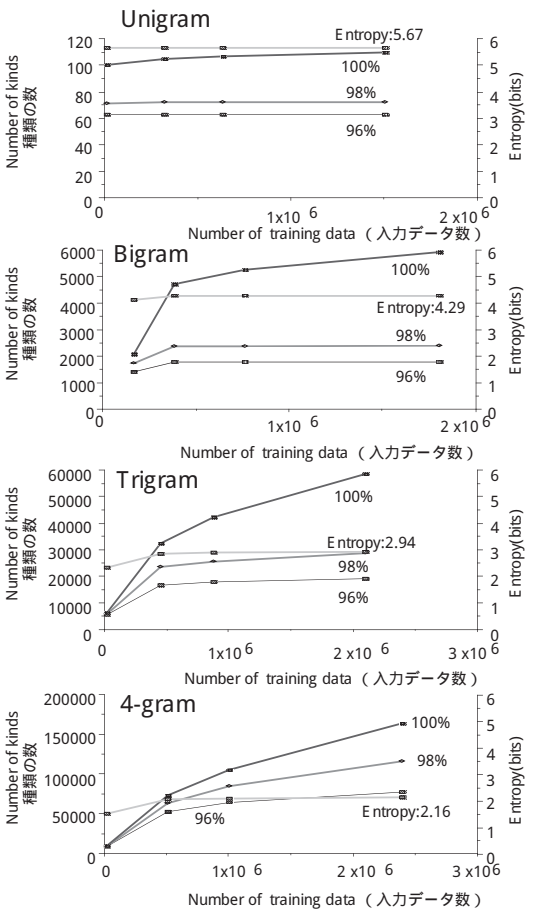
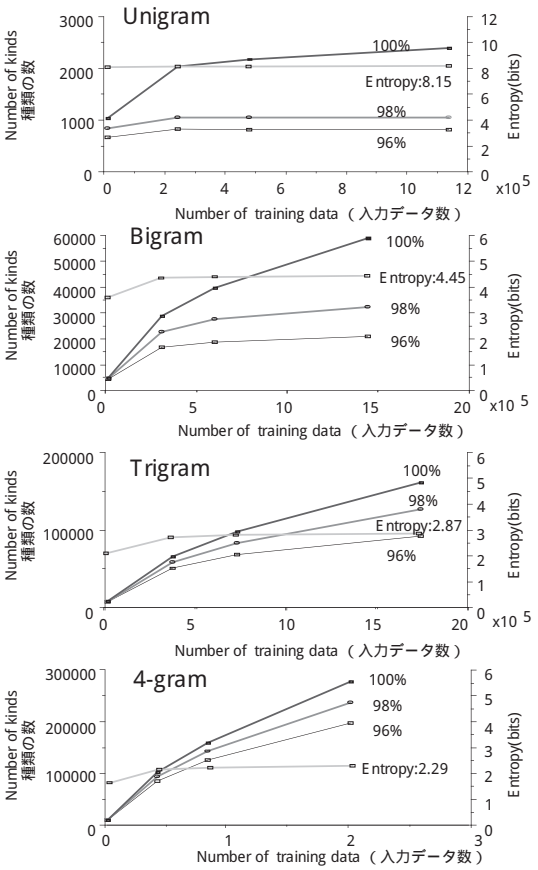


図 1: 新聞記事における学習データ数に対する音節のマルコフ連鎖確率値の収束率

2.2.2 新聞記事における漢字仮名文字のマルコフ連鎖確率の収束率

新聞記事における漢字仮名文字の学習文字数に対するエントロピーおよび頻度別出現率のグラフを図 2 に示す。なお、使用した漢字仮名の種類は JIS 1 級、約 3000 種類に限定した。これらから以下のことがわかる。



新聞記事における入力データ数に対する
漢字仮名のマルコフ連鎖確率値の変化

図 2: 新聞記事における学習データ数に対する漢字仮名のマルコフ連鎖確率値の収束率

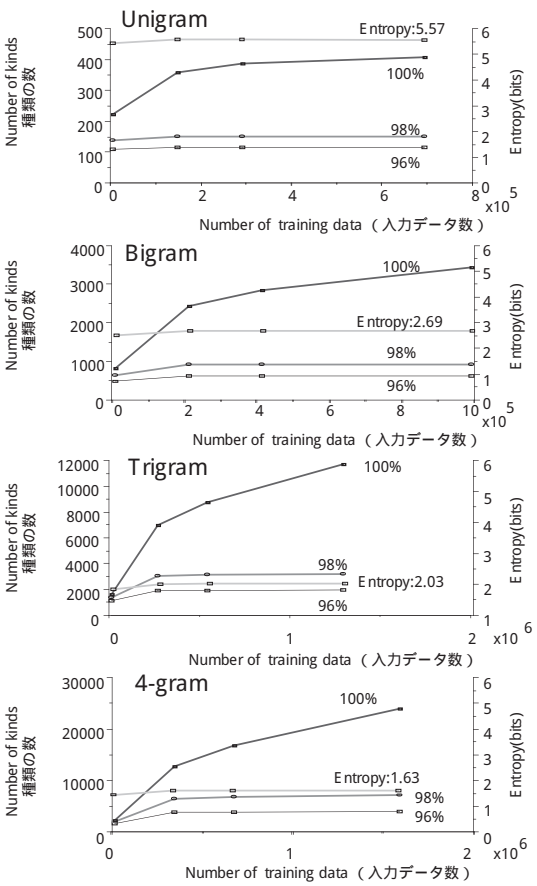
1. 漢字仮名文字の場合、連鎖確率の値を収束させるためには音節の場合よりも大量のデータが必要である。

2. 頻度別出現率 98%,96%の収束に必要な学習データの量は、音節と同様にエントロピーの場合よりも多く必要である。

3. 漢字仮名と音節のエントロピーの値を比較すると、unigram と bigram においては、音節のエントロピーの方が低いが、trigram では漢字仮名文字のエントロピーの方が低い。漢字仮名の種類の数は音節の種類の数の約 30 倍もあることを考えると、漢字仮名文字の trigram の持つ情報量は、音節と比較すると、かなり多いと思われる。

2.2.3 新聞記事における品詞のマルコフ連鎖確率の収束率

品詞は、名詞・助詞などの機能的な分類の他に地名・人名・色の種類など意味的にも分類されていて、約 450 種類ある。学習データの量の変化に対する品詞のエントロピーおよび頻度別出現率のグラフを図 3 に示す。これらから以下のことが示される。



新聞記事における入力データ数に対する
品詞のマルコフ連鎖確率値の変化

図 3: 新聞記事における学習データ数に対する品詞のマルコフ連鎖確率値の収束率

1. 品詞は、音節や漢字仮名と比較すると小量のデータで収束する。

2. 音節や漢字仮名では、unigram,bigram,trigram になるにしたがいエントロピーは半減している。しかし、品詞の場合、unigram のエントロピーの値に対して bigram のエントロピーの値は約半減するが、bigram のエントロピーの値に対して trigram のエントロピーの値は、あまり減少しない。したがって品詞の trigram の情報量は、少ないと思われる。

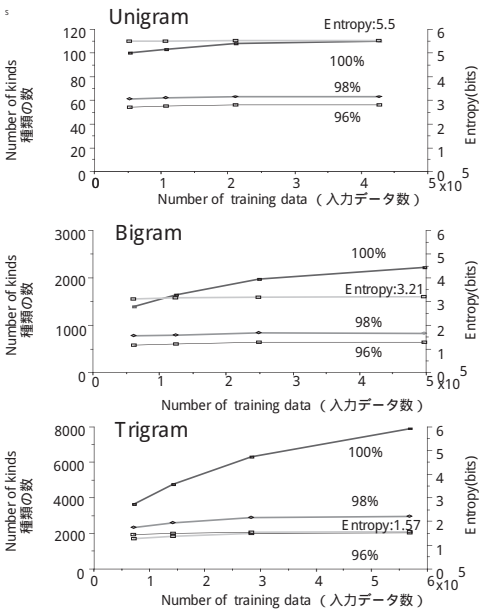
2.3 X 線 CT 所見作成

次に日本語の専門的な文章の例として X 線 CT 所見作成の文章を調べた。表 2 に文の一部を載せる。

表 2: X 線 CT 所見作成の例

頭部ＣＴ 単純および造影
１、３月１３日のＣＴと比較した。
２、スライスのレベルが若干異なっているので正確な比較はできないが、鞍上槽の正中からやや右上方へ向かって進展している増強効果を示す腫瘤の大きさは本質的に変わっていない。ただし前回のＣＴでこの結節性腫瘤の右前方に見られた嚢胞性の成分については今回は描出されていない。
３、側脳室の大きさ形も前回と同様である。

impression.....
鞍上槽の頭蓋咽頭腫の残存については明らかな変化はないが、右後方に見られた嚢胞性成分が消失しているかもしれない。



X 線CT所見作成の文章における学習データ数に対する音節のマルコフ連鎖確率値の変化

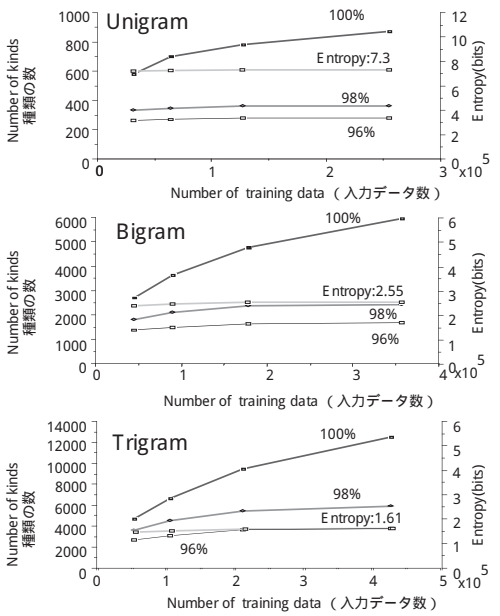
図 4: X 線 CT 所見における学習データ数に対する音節のマルコフ連鎖確率値の収束率

2.3.1 X 線 CT 所見作成における音節のマルコフ連鎖確率の収束率

X 線 CT 所見作成の文章は”mass effect”,”large magna”, などの外来語が数多く出現する。そのため音節の種類の数は 118 種類になった。図 4 に学習データ量に対する音節の unigram・bigram・trigram およびエントロピーの値の変化を示す。

2.3.2 X 線 CT 所見作成における漢字仮名のマルコフ連鎖確率の収束率

X 線 CT 所見作成の文章には外来語が多く出現する。ここでは、これらの外来語を全て 1 文字の全角文字として (例えば “mass effect” は MASS EFFECT) 漢字仮名のマルコフ連鎖確率の収束性を調べた。この結果を図 5 に示す。

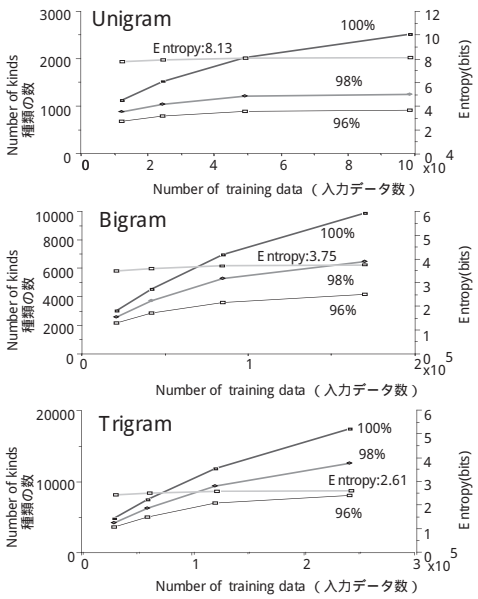


X 線CT所見作成の文章における入力データ数に対する漢字仮名のマルコフ連鎖確率値の変化

図 5: X 線 CT 所見における学習データ数に対する漢字仮名のマルコフ連鎖確率値の収束率

2.3.3 X 線 CT 所見作成における単語のマルコフ連鎖確率の収束率

X 線 CT 所見作成の文章の語彙数は約 3000 語である。ただし、全体の認識性能を向上させるため文節出現率が高いものから上位 100 文節は単語として登録してあるため、通常、文節と考えられるものまで単語と見なしている (例えば" 脳実質を" は 1 単語)。X 線 CT 所見作成における単語のマルコフ連鎖確率の収束性を図 6 に示す。



X 線CT所見作成の文章における入力データ数に対する単語のマルコフ連鎖確率値の変化

図 6: X 線 CT 所見における学習データ数に対する単語のマルコフ連鎖確率値の収束率

2.4 入力データ量に対するマルコフ連鎖確率値の収束率の問題

ここでは、新聞記事および X 線 CT 所見作成において、学習データ量に対する音声・漢字仮名・品詞・単語のマルコフ連鎖確率値の収束率について調べた。これらの結果から、以下のことが示される。

2.4.1 エントロピーと頻度別出現率

エントロピーと頻度別出現率の収束性を比較すると、全てのデータにおいてエントロピーは頻度別出現率よりも少ない学習データ量で収束することが示された。これは学習データ量に対するマルコフ連鎖確率値の変化について調査する場合、エントロピーだけでなく、頻度別出現率も調査する必要があることを意味していると思われる。

2.4.2 頻度別出現率 100%と 98%

頻度別出現率 100%のデータと頻度別出現率 98%のデータの比較から、学習データが増加した場合、全体に占める割合は少ないが、たえず新しい種類のマルコフモデルの組み合わせが出現することが予想される。これは、言語モデルとしてマルコフモデルを選択したときの妥当性に関して、 2 つの異なった解釈を出す。1 つは、滅多に出現しない言語現象は、あえてモデルに適合させる必要がないと判断するもので、もう 1 つは、所詮、言語モデルとしてマルコフモデルは使用できないと判断するものである。この判断は、人によって異なると思われる。

2.4.3 新聞記事と X 線 CT 所見作成の比較

X 線 CT 所見作成の文章と新聞記事を比較すると、音節・漢字仮名、いずれの場合もエントロピーが低く、かつ少ない学習データ量で収束している。これらから X 線 CT の所見作成の文章は新聞記事と比較して文章が単純であると言える。

2.4.4 形態素解析プログラムの精度

新聞記事におけるマルコフ連鎖確率の収束性を調べるために使用した形態素解析プログラムは単語認定率で約 95%の精度しかないため、人手によって文節単位に区切られた場合のマルコフ連鎖確率の値と、ここで得られた値に差がある可能性がある。特に品詞に関しては、品詞の trigram に有意性が見られなかった。これは、品詞の定義が人によって異なる (例えば形容動詞) などの問題点もあるが、形態素解析の精度の問題と関連している可能性があり、今後検討が必要である。

3 単語の HMM と bigram を利用した文節音声認識

ここでは、X 線 CT の所見作成入力用の音声ワードプロセッサーを目指して、認識単位として単語、言語情報として単語の bigram を使用した文節音声認識

システムを作成した。語彙数は約 3000 である。このシステムの概要と実験結果について述べる。

3.1 認識単位を単語とした文節音声認識

従来の多くの文（文節）音声認識システムでは認識単位として音節や音素を選択している [12][3]。しかし、現実の音声データでは音素境界が曖昧な音素が多い。したがって、高い認識性能を目指す場合、長い認識単位が有利であると考えられる。したがって、ここでは認識単位として単語を選択した。しかし、単語を認識単位とした場合、単語の HMM の学習の時に、大量の単語発声の音声データが必要であること、また認識のときに、HMM のパラメータの記憶のために多くのメモリー空間が必要であることなどから、従来はあまり多く行なわれてきていない [6]。そこで、本論文では学習データを減らすため、 1 つの単語の HMM の学習に 1 つの単語発声の音声データのみ使用することにした。つまり X 線 CT の所見作成入力用の音声ワードプロセッサーを使用する人に、事前に 3000 単語を 1 回発声してもらい、このデータで単語の HMM を学習した。そして、少ない音声データで精度の高い HMM のパラメータを推定するために Fuzzy-VQ HMM を用いた。また認識時において HMM のパラメータの記憶のためのメモリー空間を減らすために、単語の HMM のモデルは全て 4 状態 3 ループとした。

3.2 言語モデル

言語モデルには単語の bigram のみをもちいた。bigram の連鎖確率値の計算には、今まで入手できた X 線 CT の所見作成の全文章、71198 単語から計算した。また、連鎖確率値が 0.0 である場合は $\exp(-1000.0)$ に置き換えた。ただし、deleted-interpolation[2] などのスムージングはおこなっていない。

3.3 単語の bigram を用いた文節音声認識アルゴリズム

実験に用いた認識アルゴリズムの基本は、単語の HMM に Viterbi サーチ (one-pass DP) に単語の bigram とした。これの認識アルゴリズムを表 3 に示す。

ただし、実験では HMM の状態 $G(l, w, i)$ を複数 (N 個) 持たせることによって複数の候補を出力する N-best サーチを行なった。また、計算量およびメモリー量を減らすために次のような方法を使用した。

1. Viterbi サーチの経路計算

   one-pass DP では最尤度のワード列を知るために、計算の途中において選択した経路を残す必要がある。このために [語彙数 ×HMM の状態数 × 最大認識時間のフレーム長] のメモリー空間が必要である。しかし、どの経路を選択したかを最大シンボル出力確率 $G_t(l, w_0, i)$ とともに、次の状態に渡すことによって [語彙数 ×HMM の状態数 × 文を構成する最大単語数] のメモリー空間で計算可能である。後者を選択したとき計算時間は少し増加するが、一般的には文を構成する最大単語数は最大認識時間のフレーム長より小さいため、メモリー量を削減できる。

表 3: 単語の bigram を利用した Viterbi サーチアルゴリズム (one-pass DP)

[定義]
sn_w : 単語 w の HMM の状態数
$a^w_{i,j}$: 単語 w の HMM における状態 s_i から状態 s_j への遷移確率
$b^w_{i,j}(v)$: 単語 w の状態 s_i から状態 s_j における ベクトル v の出力確率
$Bi(w_1, w_0)$: 単語 w_1 が出現したときに w_0 に遷移する確率 $P(w_0\|w_1)$
Q : 語彙数
T : 入力フレーム数
L:最大連続単語認識数
$O(t)$: フレーム t における観測ベクトル
$G_t(l, w_0, i)$: レベル l, 単語 w_0, 状態 s_i での先頭から フレーム t までの経路の HMM の最大シンボル出力確率
[初期化]
1) $w_0 = 0, ..., Q-1$ において step2 を実行
2) $G_0(0, w_0, i) = Bi(start, w_0)$ $start$ は文頭を意味
[Viterbi サーチ]
3) $t = 0, 1.., T-1$ において step4 を実行
4) $l = 0, 1, .., L-1$ において step5,step8 を実行
5) $w_0 = 0, 1...Q-1$ において step6 を実行
6) $i = 0, 1, ..., sn_{w_0} - 2$ において step7 を実行
7) $G_t(l, w_0, i) = \max(G_{t-1}(l, w_0, i) \times a^{w_0}_{i,i} \times b^{w_0}_{i,i}(O_t),$ $G_{t-1}(l, w_0, i-1) \times a^{w_0}_{i-1,i} \times b^{w_0}_{i-1,i}(O_t))$
[単語境界の計算]
8) $w_0 = 0, 1, ..., Q-1$ において step9 を実行
9) $\Delta = \max_{0 \le w_1 \le Q-1}(G_{t-1}(l-1, w_1, sn_{w_1} - 2)$ $\times a^{w_1}_{sn_{w_1}-2, sn_{w_1}-1} \times b^{w_1}_{sn_{w_1}-2, sn_{w_1}-1}(O_t) \times Bi(w_1, w_0))$ もし $\Delta \ge G_t(l, w_0, 0)$ ならば $G_t(l, w_0, 0) = \Delta$

2. 単語 bigram の値の記憶

   単語の bigram の値を記憶しておくには [語彙数 ^{2}] のメモリー空間が必要である。しかし、テキストデータ中に存在する bigram の組み合わせの値をリスト形式で記憶することにより、メモリーを節約できる。

3. one-pass DP における単語境界での計算

Viterbi サーチでは各フレームごとに、単語境界の計算をするために、全ての前の単語の最終状態の値に前の単語から現在の単語に遷移する bigram の値を加算してから最大値を選択する (表 3 step9)。しかし bigram の値はリスト形式で記憶されているため、bigram の値を得るために計算コストがかかる。そこで、このコストを減らすために、全ての前の単語の最終状態の値で最大値を選択してから bigram の値を加算して最大値を選択した。このアルゴリズムを次に示す。

9) $\Delta = \max_{0 \leq w_1 \leq Q-1}(G_{t-1}(l-1, w_1, sn_{w_1}-2) \times a^{w_1}_{sn_{w_1}-2, sn_{w_1}-1} \times b^{w_1}_{sn_{w_1}-2, sn_{w_1}-1}(O_t))$

もし $\Delta \times Bi(w_1, w_0)) \geq G_t(l, w_0, 0)$ ならば
$G_t(l, w_0, 0) = \Delta \times Bi(w_1, w_0))$

このアルゴリズムを選択した場合 bigram の値をアクセスする回数が減らせるため、全体の計算コストが減少する。ただし、得られる尤度は近似解になる。

3.4 実験条件

認識実験では duration control と N-best のサーチ幅を変化させて行なった。また、単語の HMM の学習のデータを増加させた場合の実験も行なった。これらの実験の条件を表 4 に示す。その他の実験条件は表 5 にまとめた。なお duration control は同一話者の単語発声の 3 回分のデータの平均発声時間と分散を測定し、この値からガウス分布を計算し、duration control に使用した。

表 4: 実験条件

実験番号	duration control	N-best	学習データの個数
実験 1	なし	2	1
実験 2	あり	2	1
実験 3	あり	8	1
実験 4	あり	2	3

3.5 テストデータ

X 線 CT 所見作成の文章は大きくわけて正常所見と異常所見に分類される。そして異常所見は正常所見と比較すると文章が複雑なため、認識率が低くなることが知られている [11]。そこで実験は、bigram の連鎖確率を計算するのに使用したテキストを発声した音声データ (text-closed data) と bigram の連鎖確率を計算するのに使用しなかったテキストを発声した音声データ (text-open data) について、各々異常所見と正常所見について合計 4 つの条件で行なった。実験は平均 100 文節行なった。

1. text-closed data の正常所見
2. text-closed data の異常所見
3. text-open data の正常所見
4. text-open data の異常所見

表 5: 文節音声認識の実験条件

使用アルゴリズム	word HMM + Viterbi search + word bigram 特定話者認識
認識単位	word
語彙数	約 3000
学習データ	単語発声
言語情報	単語 bigram
音響パラメータ	log power + 16 次 LPCcepstrum + Δlog power
距離尺度	簡易マハラノビス
VQ コード数	256
単語モデル	4-state 3-loop Fuzzy-VQ HMM
フレーム窓長	18ms
フレーム周期	9ms
ファジネス	1.5
近傍数	5
サンプリング周波数	12kHz
HMM と bigram の結合値	32

3.6 実験結果

実験結果を表 6 に示す。この結果からわかることを以下に示す。

1. 実験 1 から test-closed の正常所見で 96.8%、異常所見では 78.1%、text-open の正常所見でも 86.5%、異常所見では 72.1%の高い文節認識率が得られた。したがって HMM の学習データが 1 つでも Fuzzy-VQ を使用することにより高い文節認識性能が得られることがわかった。

2. 実験 1 と実験 2 の比較から、duration control を行なうと認識性能が低下した。この原因として duration control に使用した平均・分散の値の不正確さが考えられる。これらの値は同一話者が発声した 3 つの単語発声の音声データから計算したため値の信頼度はかなり低い。

3. 実験結果 2 と実験結果 3 の比較から、N-best の幅を広げた方が高い認識率を出すことが示された。

4. 実験結果 2 と実験結果 4 の比較から、音声データを増加させることによって認識性能が向上することが示された。これは HMM のパラメータを推定するための学習データが 1 つでは、不十分であることを示している。しかし不特定話者認識の場合、一人の発話データが 1 つしかなくても、複数の話者が発話することによって、多くの音声データが利用できるため、認識単位が単語でも問題はないと思われる。

表 6: 実験結果

実験番号	1	2	3	4
duration control	なし	あり	あり	あり
N-best	2	2	8	2
学習データ数	1	1	1	3
text-closed data の正常所見	96.8%	82.6%	100.0%	100.0%
text-closed data の異常所見	78.1%	76.3%	78.9%	84.2%
text-open data の正常所見	86.5%	86.5%	89.2%	94.6%
text-open data の異常所見	72.1%	68.9%	72.1%	77.0%

4 考察

4.1 HMM の種類について

この実験では、HMM の学習に使用する音声データを 1 つとしたため Fuzzy-VQ HMM を使用した。しかし不特定話者認識のためには continuous mixture HMM が良いと考えられる。しかし、continuous mixture HMM を使用した場合、mixture 数にも依存するが学習に大量の音声データが必要である。そこで、単語を認識単位とするばあいは、学習データがある程度少なくてすむ semi-continuous HMM[1] が有望ではないかと考えている。

4.2 認識単位・単語

認識単位として音素を選択したとき、HMM の学習のために、音素ラベルが付与された音声データが必要になる。ラベリング作業は自動化がある程度可能であるが、最終的には人手に頼らざるを得ないため、音声データベースの作成のコストはかなり高い。一方認識単位を単語にしたばあい、ラベリング作業は不用になる。そのかわり、数個の単語発声が必要があるため、発話者の負荷が大きくなる。したがって認識システムの仕様や目的にも依存するが、認識単位を単語としたときのほうが、音声データベースの作成に必要なコストは低くなる可能性があると考えている。

4.3 リアルタイムにむけて

音声認識のリアルタイム化には 2 つの方法がある。1 つにはアルゴリズムによる計算量の削減であり、もう 1 つはハードウエアによる計算コストの分散化である。フレーム同期型の認識アルゴリズムにおいて計算量を削減する方法としてビームサーチが知られている [13]。しかし、この方法を用いるにはフレームごとに $G_t(l, w_0, i)$(表 3 step7) のソーティングが必要になってくる。この計算コストはかなり大きい。このため、超並列コンピュータなどを考えた場合、ビームサーチを採用しないほうが早くなる可能性がある。今後、リアルタイム化はハードウエアも考慮して最適なアルゴリズムを考えていく必要があると思われる。

5 まとめ

本報告では、学習データ量に対するマルコフモデルの収束率について調査した。この結果マルコフモデルの連鎖確率の信頼性を調べるためにはエントロピーだけでなく、頻度別出現率も調査する必要があると思われる。また、特定話者の文節認識実験を行なった結果、認識単位を単語とした場合、HMM の学習用の音声データが 1 つでも、かなり高い認識率が得られること、そして単語の bigram の情報と組み合わせることにより、text-open の正常所見でも 86.5%、異常所見では 72.1%の文節認識率が得られることが

示された。

参考文献

[1] X.D.HUNAG, Y.Ariki, M.A.Jack, “Hidden Markov Models for Speech Recognition” EDINBURGH UNIVERSITY PRESS (1990)

[2] A.Averbuch, “An IBM PC BASED LARGE-VOCABULARY ISOLATED-UTTERANCE SPEECH RECOGNIZER,” Vol.1, 2.4.1 PP.53-56 ICASSP (1986).

[3] Kai-Fu Lee, “Large-Vovabulary Speaker Independent Continuous Speech Recognitton: The SPHINX System,” 15213 CMU-CS-88-148 (April 18, 1988).

[4] 鹿野 清宏, “Trigram Model による単語音声認識結果の改善,” , 電子情報通信学会技術報告,SP87-23, pp.9-16 (1987).

[5] 村上 仁一, 荒木 哲朗, 池原 悟, “2 重マルコフ連鎖確率モデルを使用した単音節音声入力の改善,” 信学技報, SP88-29, pp.63-70 (June 1988).

[6] 南 泰浩、中川 正雄, “trigram モデルを用いた複数候補を求めるフレーム同期型 HMM 連続音声認識,” 電子情報通信学会論文誌, D-2,Vol.j73-D-2,No9 PP.1383-1392 (1990).

[7] 村上仁一, 坪井俊明, “BIGRAM をもちいた音節 HMM による文節音声認識,” 音響学会講演論文集, 3-5-15, pp. 117-118 (1991-04).

[8] Kucera.H.,Francis, W.N., “Computational Analysis of Present-day American English,” Brown University Express. Providence, Rhode Island. (1967).

[9] 江原 暉将, 小倉 健太郎, 森元 逞, “電話対話データベースの構築,” 第 40 回情報処理全国大会, pp.486-491 (1990).

[10] 宮崎正弘、大山 芳史 “日本音声出力のための言語処理方式,” 情処論文誌,Vol.27, No.11, pp.1053-1061 (1986).

[11] 坪井 俊明, 菅村 昇 “文章作成支援装置の評価,” 電子情報通信学会技術報告,SP90-36, PP.17-22 (1990.8).

[12] 永井, 他, “HMM-LR 連続音声認識装置の開発と性能評価,” 日本音響学会平成 3 年度秋季研究発表会, 1-5-23 pp.45-46, (Oct. 1991).

[13] 迫江博昭, 藤井浩美, 吉田和永, 亘理誠夫, “フレーム同期化、ビームサーチ、ベクトル量子化の統合による DP マッチングの高速化,” 信学論 D,Vol.J71-D, No.9, pp.1650-1659(Sep. 1988).