音声言語情報処理 11－6
（1996．5．27）

確率モデルによる大語彙連続音声認識の評価

周 旻　堤真理子　中川聖一
{xiu, mariko, nakagawa}@slp.tutics.tut.ac.jp

豊橋技術科学大学 情報工学系
〒441 愛知県 豊橋市 天伯町 字雲雀ヶ丘1－1

概要 品詞 N-gram の確率モデルに基づく日本語のモデル化を検討し、テキストコーパスのパープレキシティ及び後続単語の予測能力について言語モデルの性能を評価した。次に、語彙数が4699単語の ATR 日本語対話データベース (ADD) について、大語彙日本語連続音声認識の評価実験を行なった。連続音声認識は話者独立と話者適応化の条件で、また認識対象音声の発声スピードは速い・普通・遅いの三通りを用いた。話者独立の場合には、音節の HMM モデルに音声の動的特徴量を導入する事によって、認識精度を改善した。

これらの認識実験を通して、品詞に基づく N-gram 言語モデル及びセグメント単位の音節 HMM の大語彙連続音声認識における有効性を明らかにした。また、助詞誤りと接頭語誤りが多いことに対処するため、助詞と接頭語の単語マルコフモデルを品詞 N-gram モデルと結合する事によって、認識精度の改善へのアプローチを試みた。また、疑似文節ごとにポーズを入れた音声と連続発声との比較認識実験を行なった。

和文キーワード： 音声認識、確率言語モデル、bigram 、trigram、大語彙、パープレキシティ

Evaluation of Large Vocabulary Continuous Speech Recognition Based on Stochastic Language Model

Min ZHOU, Mariko TSUTSUMI, Seiichi NAKAGAWA

Toyohashi University of Technology, Department of Information & Computer Sciences
Tempaku-cho, 441 Toyohashi, Aichi

Abstract In this study, the POS based bigram model is used to modeling of Japanese and evaluated the ability of the prediction of succeeding words and the perplexity of the task. The continuous speech recognition experiments are performed on a Japanese large vocabulary task including 4699 words which is a dialog database related with travel inquiry. The experiments are performed on both speaker independent/adaption mode on a set of 100 sentences which are uttered in three different speeds. In the case of speaker independent recognition, the performance is increased by using of dynamic feature of speech signal.

The experiment results illustrate that the POS-based bigram and syllable-based HMMs are valid on Japanese large vocabulary speech recognition. For the purpose of decreasing the errors due to postposition/prefix, a POS trigram and word bigram based on postposition/prefix are combined with POS bigram, which got some increase of recognition accuracy. As another approach, a try of recognizing utterances including pauses between pseud-phrases is also performed on the same task.

Key Words: speech recognition, stochastic language model, bigram, trigram, large vocabulary, perplexity

1 まえがき

確率言語モデルの連続音声認識における重要性と有効性が益々認められてきた。特に大語彙連続音声認識の場合に、言語モデルが後続単語を予測・制約する事によって、音声認識の探索空間を絞り込み、音声認識の性能を改善でき、音響的な特徴のみによる認識結果を補正し、認識精度を向上させることができる。

音声認識にとって重要な後続単語の予測に関しては、従来、文節文法や文脈自由文法を用いるものが代表的であった[1, 2]。しかし、多量のテキスト入力文や自然な発話文に対してはこのような文法の構築は難しく、これらに確率を導入したり、bigram、trigram を用いることが欧米を初め、日本でも盛んに研究されている[3, 4, 5, 6, 7, 8]。現実的には、確率文脈自由文法と trigram (bigram) の併用が有望だと考えられている[9]。

本報告では、確率モデルを学習するためのテキストコーパスが不十分な時に、品詞に基づく N-gram モデルでの近似方法を検討し、それに基づく後続品詞/単語の予測確率の計算法を検討する。また、日本語の ATR の対話データベース (ADD) を用いて単語レベルでその品詞モデルの有効性を比較して、言語モデルのエントロピーと後続品詞/単語の予測率を考察した。また、その品詞モデルを連続文認識アルゴリズムに組み込んで、話者独立/適応モードで大語彙の朗読発声の文認識の評価実験を行なった。更に、テスト文を異なる発声スピードで朗読した時の比較をし、音声信号の動的特徴量を用いて (4 フレームセグメントの統計量) 認識性能を改善した。

更に、大語彙日本語連続音声認識によく出る助詞と接頭語の誤認識に対して、助詞・接頭語の単語モデルを結合した方法を検討し、最後に、疑似的に文節ごとに区切って発声した場合との認識率の比較を行なった。

2 確率言語モデル

N-gram を用いて、欧米を初めとして、日本でも多数の大語彙連続音声認識システムが実現されている [3, 4, 5, 6, 7, 8]。

しかし、一つの主な問題点はモデルの学習に膨大なテキストデータベースの存在が前提となっていることである。例えば、5000 語の trigram を求めるために全ての三つ組の数が 1,250 億あるので、それらのパラメータを学習するために、パラメータ数の数倍の単語からなる膨大なるテキストデータが必要となる。これほどの膨大なテキストデータを収集する事は大変なことである。

学習データがそれほど十分でない時に、単語に基づく N-gram の実現は困難であって、モデルのパラメータの信頼度に問題がある。本報告では、日本語の ATR の大語彙対話データベース (ADD) を用いて品詞に基づく N-gram モデルを検討し、それらに基づくテストコーパスのパープレキシティに基づく比較をする。また、朗読発声の連続文音声認識実験によって、それらの確率モデルの比較をする。

2.1 品詞に基づく確率言語モデル

単語系列$\boldsymbol{w}_1^L = w_1 w_2 \cdots w_L$が与えられる時に、言語モデル$\boldsymbol{G}$ による部分文の生成確率 $P(\boldsymbol{w}_1^n)$ は次式のように表される。

$$P(\boldsymbol{w}_1^n) = \prod_{t=1}^{n} P(w_t | \boldsymbol{w}_1^{t-1}, \boldsymbol{G}) \tag{1}$$

上式を bi/tri-gram の確率言語モデルで近似すると、文の生起確率は次式で表される。

$$P_{\mathrm{trigram}}(\boldsymbol{w}) = \prod_{i=1}^{n} P(w_i | w_{i-2}, w_{i-1}) \tag{2}$$

$$P_{\mathrm{bigram}}(\boldsymbol{w}) = \prod_{i=1}^{n} P(w_i | w_{i-1}) \tag{3}$$

学習データが足りない時に、事後確率 $P(w_t | w_{i-2}, w_{w-1})$ を直接求めるのは困難であるため、我々は品詞レベルの言語モデルに基づいて、その単語列の確率及びテキストのエントロピーと後続単語の予測率を求める。本稿で用いる品詞に基づく N-gram モデルは以下のように定義された一次近似の確率モデルである [10]。

単語列$\boldsymbol{w}_1^L$ に対応する品詞列を$c_1^L = c_1 c_2 \cdots c_L$、品詞による条件付き予測確率を $P(w_t | c_{t-2}, c_{t-1})$ とすれば、部分文の生成確率は

$$P_{\mathrm{trigram1}}(\boldsymbol{w}_1^n) = \prod_{t=1}^{n} P(w_t | c_{t-2}, c_{t-1}) \tag{4}$$

$$P_{\mathrm{bigram1}}(\boldsymbol{w}_1^n) = \prod_{t=1}^{n} P(w_t | c_{t-1}) \tag{5}$$

で近似できる。

2.2 助詞・接頭語の単語単位のモデル

予備実験により、助詞や接頭語の認識が非常に難しいことが分かったので [10]、これらの予測精度を高めるために助詞と接頭語に対して単語 bigram を用いる方法についても検討を行なった。これは (1) 助詞・接頭語単語から後続単語を予測する方法、(2) 助詞・接頭語の品詞 trigram に基づく方法である。即ち、近似 bigram で $P($品詞$_{i-1}$→単語$_i)$ の確率を、$P($助詞・接頭語単語$_{i-1}$→単語$_i)$ と $P($品詞$_{i-2}$品詞$_{i-1}$→単語$_i)$ という精度の良いモデルに切替えるという方法である。

- (品詞 bigram + 助詞・接頭語の品詞 trigram) モデル

$$P(w_i | w_1^{i-1}) = \begin{cases} P(w_i | c_{i-2} c_{i-1}) & \text{if } (w_i \in \boldsymbol{C}) \\ P(w_i | c_{i-1}) & \text{otherwise} \end{cases}$$

- (品詞 bigram + 助詞・接頭語の単語 bigram) モデル

$$P(w_i | w_1^{i-1}) = \begin{cases} P(w_i | w_{i-1}) & \text{if } (w_i \in \boldsymbol{C}) \\ P(w_i | c_{i-1}) & \text{otherwise} \end{cases}$$

ただし$\boldsymbol{C}$={w | w is 助詞 | 接頭語 }。

2.3 エントロピーとパープレキシティ

以上の言語モデルの能力を比較するために、品詞レベルと単語レベルで種々の確率言語モデルによる後続品詞と後続単語の予測及びモデルパラメータ数とエントロピーの関係などの比較、評価実験を行なった。

(a) 定義 言語モデル $\boldsymbol{G}$ において、文 (単語列) $\boldsymbol{w}_i = w_1^{T_i}$ の出現確率を $P(\boldsymbol{w}_i)$ とすれば、文集合$\{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_N\}$ のエントロピーは次式で求められる [11]。

$$H(L) = -\sum_{i=1}^{N} P(\boldsymbol{w}_i) \log_2 P(\boldsymbol{w}_i) \tag{6}$$

全テキスト文の連接を $\boldsymbol{W} = \boldsymbol{w}_1 \boldsymbol{w}_2 \cdots \boldsymbol{w}_T$ とすれば

$$H(L) = -\log_2 P(\boldsymbol{W})$$

一単語当たりのエントロピーは

$$H_0(L) = -\frac{\log_2 P(\boldsymbol{W})}{\sum_i T_i} \tag{7}$$

また言語の複雑さ・パープレキシティは

$$P(L) = 2^{H_0(L)} \tag{8}$$

と定義される。言語の複雑さは、全ての生成可能な文に対する平均値だが、実際の学習・認識実験は有限の文集合であるため、テスト文集合に対するパープレキシティを使う方が現実的である。これを特に、テストセットパープレキシティと呼び、(8) 式で求める。

表 1: 言語モデルの比較 (perplexity)

(a) 品詞の一次近似モデル

モデル	Bigram	Trigram	Bigram-HMM
train set	59.6	43.5	49.7
test set	62.7	52.7	54.9

(b) 助詞・接頭語モデルの高精度化

モデル	データ	$P(w_i\|c_{i-2}c_{i-1})$	$P(w_i\|w_{i-1})$
助詞	train set	48.0	33.1
	test set	45.1	36.0
助詞+接頭語	train set	47.7	31.4
	test set	44.8	33.3

(b) 言語モデルの比較 言語モデルの有効性を調べるために、ATR で作成された日本語の旅行案内に関する問い合わせの対話データベース (ADD) を用いて言語のモデル化の評価実験を行なった。このデータベースは学習セットとテストセットにわけられ、それぞれ 10495 文と 1073 文を含む。学習とテストセットを併せて、冠頭詞を取り除いて、全セットに出現する単語の数は 4699 種類である。全てのデータが人手で単語ごとに区切られ、24 種類の品詞にラベリングが付けられている。表 1中のテストセットは実際の連続音声認識に使われる 100 文を指し、1073 テストセットから適当に選択したものである。

言語のモデル化実験として (1) 品詞の bigram + (助詞+接頭語) の品詞 trigram、(2) 品詞の bigram + (助詞+接頭語) 単語の bigram モデルなどの言語モデルについてパープレキシティによる比較を行なった。表 1(b) にそれらの結合モデルの場合のテストセットのパープレキシティを示す [12]。ここでの助詞は 84 個の助詞を含む助詞リストで、接頭語は 29 個の接頭語リストである。表 1(a) は比較として、bigram、trigram、bigram-HMM によるパープレキシティを示す [10]。

表 1(a) と比べれば、(b) の助詞・接頭語の単語 bigram を結合した場合のモデルのテストセットのパープレキシティが他より小さいので、言語制約における効果があると思われ、このモデルを連続認識実験に適用し、評価した。

3 連続文認識アルゴリズム

3.1 One-Pass 文認識アルゴリズム

従来の連続音声認識のアルゴリズムとして知られている Viterbi サーチ (One-Pass DP) は各フレーム毎に各単語の境界と仮定して、言語モデルによる確率の対数値と音響累積尤度を足すことを繰り返すことによって、次の式を満たす最尤の単語列の候補を計算することができる。

$$w^* = \arg\max_{\{w_1^N\}\{t_1^N\}} \left(\sum_{n=1}^{N} \log(P_a(w_n|y_{t_{n-1}+1}^{t_n})) + weight \cdot \sum_{n=1}^{N} \log(P(w_n|c_{n-1})) \right) \quad (9)$$

ここで $P_a(w_t|y_{t_{n-1}+1}^{t_n})$ は観測パターン系列 $y_{t_{n-1}+1}\ y_{t_{n-1}+2} \cdots y_{t_n}$ に対して単語 w_n の音響的な出現確率、$P(w_n|c_{n-1})$ は品詞が c_{n-1} の単語 w_{n-1} の次に単語 w_t が接続する確率である。即ち、(HMM の音響尤度 + 単語 bigram の尤度 + Viterbi ベストスコア + ビームサーチ) を基本的な認識の枠組みとして、各時刻・各状態において累積尤度を計算する。

3.2 計算量の削減

ビームサーチ 各時刻 (フレーム) について、全て予測される単語を全部連接し、保存していくと全部の単語列の数が爆発的に増える。従って、各フレーム毎において累積尤度の低い単語列は以後の探索から除外する。そのため、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候補に対してのみ計算を続けることにより、計算量及びメモリ量を大幅に減らしている。

枝がり 枝がりは計算量を減少するためのもう一つの制御方法である。上述のビームサーチは一旦、残された候補は入力終端までずっとマッチングを続行する (HMM では自己ループを許しているいるためで、我々のように継続時間制御を用いても似たふるまいをする)。これを制限するために枝刈りを行なう。各フレームの最尤累積尤度と比較して、パスの累積尤度をフレームごとにチェックして、一定値以下であれば、そのパスの計算を打ち切る方法である。これによって、フレーム当たりの計算量が単調に増加することを防ぎ、処理時間とメモリの節約を行なうことができる [15]。

後続単語予測リスト N-gram はフロアリングされている ($P(w_{t+1}|w_t) > 0$) から、任意の単語間の接続関係が許される。即ち、探索認識の時に後続単語として辞書中の全ての単語がマッチングの対象になっている。後続予測率の予備実験により、品詞 bigram モデルの時に 512 位までの予測的中率が約 92%になっている。従って、品詞 bigram モデル ($P(w_t|c_{t-1})$) に対して、各品詞の後続単語予測リストを作り、計算量を大部に減少した。今回の大語彙認識実験では、この予測リストの長さ (予測リストの単語数) を 512 にした。

以上の方法は固定長後続単語予測リストの方法であるが、実際は各品詞に対して、同じ予測的中率になるまでの単語リストの長さは異なる。従って、計算効率をもっとよくするために可変長後続単語予測リストの方法を試した。即ち、各品詞に対して同じ後続単語予測率になる単語リストを作る。今回の大語彙認識実験の時、固定長の場合と同じ認識率の条件では、最短リスト長は 13、最長リスト長は 1680、平均長は 562 である。これによって全体的に約 18%の計算量が減少された。

3.3 音節数に依存した結合重みの決定

言語モデルとして N-gram 確率モデルを用いた場合に、助詞や接頭語といった音節数の少ない単語の脱落を防止するために、以下の方法を検討した。

Bigram 確率は単語レベルで求められ、その単語の長さ (音節数) は考慮していない。そのため、文認識を行なう際に音節数の長い単語の方が bigram 確率を乗じる回数が少なくなり、有利になってしまう。

そこで音節数の短い単語を優先的に認識するために、bigramの重み係数を接続する単語の音節数に依存するよう、重み係数として音節数を確率にかける(連乗する)ようにし、音節数の短い単語を優先する。これによって、予備実験では認識率が大幅に改善された[13]。

今回の大語彙連続音声認識実験に於いても、言語モデルの結合尤度を単語の音節数に依存するようにセットした。すなわち、式(9)の定数の結合重み係数を次のような音節数に依存するような変数に変更する。

$$weight = \alpha + \beta \cdot \gamma(w_n) \tag{10}$$

ここで$\gamma(w_n)$は単語w_nに含む音節数に依存するような関数である。以下の認識実験の結果は$\gamma(w_n) = w_n$の音節数、$\alpha = 10$、$\beta = 0.5$とほぼ最適な値にした時の結果である。

4 大語彙連続音声認識

音声認識における言語モデルの役割は文認識性能を向上させるためである。その有効性を評価するために上述した言語モデルを大語彙認識システムに組み込んで、日本語の連続音声認識実験を行なった[6]。

4.1 実験条件

データベース 今回の大語彙連続音声認識に用いるテスト文はATR観光案内対話データのテストセットからランダムに抽出した100文である。文の長さが3〜20単語に制限され、テストセットの平均長さは約9単語(18.7音節)/文である。

音声データと分析条件 ATR観光案内対話データのテストデータから抽出した100文を男性話者三人が3通りの発声速度(速い、普通、遅い)で発声したものを評価用データとし、各話者30文で以下に述べる音響モデルの話者適応化を行なった。表2にその分析と認識条件を示す。表3にそれぞれの発声速度を示す。

4.2 音響モデル(HMM)

大語彙の連続音声認識において、一般的に単語より小さい単位のものを扱うことが多い。ここで実際に使うのは113個の文脈独立の日本語の音節モデルである。それらの音節モデルは4状態5遷移連続出力型(全共分散行列使用)離散継続時間長分布制御のHMMである。

話者独立の基本HMM ATR研究用日本語音声データベース中の連続発声文データ(503文)と216単語を男性6名が発話した音声から切り出した音節データを用いて、最尤推定で学習した基本HMMを更に日本語音響学会研究用連続音声データベースのVol.1〜3(30名の男性が発話した4500文)でMAP推定法によって学習されたものである。HMMの出力分布の特徴パラメータベクトルは(10cep+10Δcep)のものである。

話者適応化のHMM 静かな録音室で録音した30文の発話を用いて、話者独立の基本HMMをMAP推定法によって話者毎に適応化したものである。

表2: 連続音声認識の実験条件

語彙	4699単語
言語モデル	品詞-単語のバイグラム
学習データ	10495文(117027単語)
実験文数	100文×3人
発声様式	朗読発話
発声内容	旅行案内に関する問い合わせ
音節モデル	5状態4ループHMM
音節カテゴリ数	113音節
話者適応化文数	30文 / 話者
平均文長	9単語 / 文
sampling周波数	12kHz
窓関数	21.33msハミング窓
フレーム周期	8ms
分析	14次元のLPC分析
特徴パラメータ	10次LPCメルケプストラム + 10次回帰係数

表3: 評価データの発声スピード(音節/秒)

Speaker	KO	MH	TI	平均
fast	10.2	9.0	10.1	9.8
normal	7.2	7.8	6.3	7.1
slow	5.7	6.3	7.0	6.3

4フレームセグメントのHMM [14] 基本HMMと似ているような、24次元の出力分布(14cep+10Δcep)を持つHMMである。このうち14次元のパラメータは音声特徴量の10次元のメルケプストラムの4フレームを1つのセグメントにして得られた40次元をKL展開で14次元に変換したものである。これを1フレームづつシフトしたものを用いる。

ΔΔケプストラムを含むHMM [14] 上述の(14cep+10Δcep)にさらに10ΔΔcepを追加した34次元の音響特徴量をHMMの出力分布のパラメータにするものである。

4.3 実験結果

表4(a)〜(c)は以上で述べた条件で行なった連続文音声の認識実験の結果である。表4(a)はビーム幅が512、話者適応モードで品詞-単語ペアレベルのbigramによる連続文音声の(単語)認識結果である。表4(b)はビーム幅が512、話者独立モードで4フレームセグメントのHMMを用いた時の連続文音声の(単語)認識率の結果である。表4(c)は話者独立モードで表(b)と同じ条件にΔΔcepを追加して得られた認識結果である。表より、セグメントの統計量の有効性が分かる(表4(a)と同じ音響HMM(10cep+10Δcep)の話者独立モードの場合では単語認識率は(fast:60.8%、normal: 63.7%、slow:67.8%))。3.3節で述べたweightに対しては音節数に依存させない通常の方法($\alpha = 10, \beta = 0$)と比べて、認識率にあまり差はなかった。

表 4: Bigram による連続音声認識結果 (%)

(a) 話者適応化 (10cep+10Δcep)

発声	置換	挿入	脱落	正解	文正解
fast	15.8	3.1	15.4	68.7	15.3
normal	21.6	4.4	7.1	71.3	20.3
slow	23.1	6.7	5.0	73.5	16.3

(b) 話者独立 (14cep+10Δcep)

発声	置換	挿入	脱落	正解	文正解
fast	27.3	5.9	7.3	65.5	11.6
normal	26.2	4.9	7.4	66.4	15.3
slow	26.1	8.7	4.0	69.9	12.3

(c) 話者独立 (14cep+10Δcep + 10ΔΔcep)

発声	置換	挿入	脱落	正解	文正解
fast	25.9	6.6	5.8	68.2	9.3
normal	26.7	9.0	3.8	69.5	15.3
slow	24.9	12.8	2.2	72.9	9.3

表 5: 助詞・接頭語モデルによる認識結果 (%)

(a) 品詞 bigram + 助詞・接頭語の品詞 trigram

話者	置換	挿入	脱落	正解	文正解
KO	23.8	6.4	6.8	69.3	15
MH	20.0	8.5	3.6	76.4	20
TI	24.5	5.5	6.7	68.8	17
平均	22.8	6.8	5.7	71.5	17.7

(b) 品詞 bigram + 助詞・接頭語の単語 bigram

話者	置換	挿入	脱落	正解	文正解
KO	22.9	6.6	6.7	70.4	15
MH	20.4	8.3	3.5	76.1	21
TI	24.6	5.8	6.3	69.1	18
平均	22.6	6.9	5.5	71.9	18.0

表 5は表 4(a) の条件 (但し、発声スピード: normal のみ) に対して、助詞・接頭語を特別に考慮した言語モデルを使用した場合の認識率である。助詞・接頭語モデルの併用の効果はそれほど顕著ではなかったが、助詞・接頭語の単語 bigram の方で若干正解率が改善された。

4.4 認識結果による後続予測率

音声認識における言語モデルの一つの重要な役割はある時点まで認識できた単語列から、次にどんな単語がどんな確率で生じるかを求めることである。従って、よい言語モデルとは次の単語が正しく予測できる確率が高いモデルのことである。本稿では、単語列 $w_1 w_2 \cdots w_i$ と後続可能な単語 w_x を結合した $w_1 w_2 \cdots w_i w_x$ の生起確率を言語モデルで求め、確率の高い順の w_x を予測順位とする。

実際の認識システムに言語モデルを組み込んだ場合には、単語認識誤りが生じるため誤りの混在した単語列から次単語を予測しなければならない。そこで、認識結果による単語列を用いて、後続単語の予測を行ない、単語の認識精度と予測率の関係を調べた。入力文 (単語認識率 100%) による後続単語の予測率と比べるために、両方とも同時に図 1に示した。

図1に示してある各線の word は単語の認識率、pos は品詞の認識率である。Ins.、Del. と Seg. はそれぞれ品詞の挿入、脱落とセグメンテーション率 (Seg.=1.0−Ins.−Del.) である。本実験の言語モデルは品詞情報を用いた単語予測を行なっているから、後者の値に意味がある。実験結果から分かるように、認識率が良ければ予測精度が良くなる傾向が見える。後続予測率の順序関係は認識率とほぼ対応した関係があることがわかる。全体的に認識率が上昇するに従って言語モデルの後続単語予測率も上がる。高精度な認識システムを構築するためには、後続単語を予測するために使う直前の認識率 (本実験では品詞認識率) が 80%程度、セグメンテーション率は 90%以上は必要であると思われる。

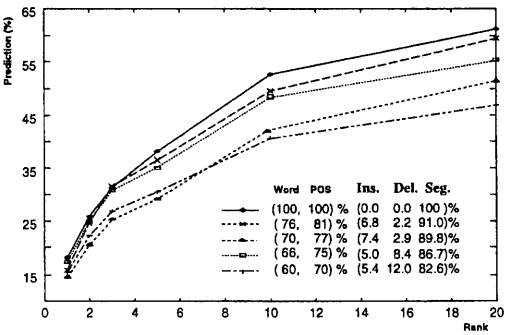

図 1: 認識結果に基づく後続単語の予測.

4.5 疑似文節単位の発声による認識実験

人間は一つの文を一気に連続して喋るのではなくて、その途中にポーズを入れる場合が多い。更に、そのポーズを入れる場所は文節の間が多いと思われる。そこで、ユーザの協力的な発声が期待できるディクテーションシステムとして、適当な場所でポーズを入れて発声してもらい、このことが認識率の改善に寄与するか調べた。

同じ 100 文を 3 人の被験者に自分の意思で文節に相当すると思われるところに適当にポーズを入れて発声してもらった。表 6は疑似文節単位の発声の場合の発声状況を示している。文毎の平均ポーズ数は 2.3 個、平均発声スピードは特に、連続発声の評価データの slow の場合よりも少し遅い。表中の文節境界の一致率は、第三者が決めた文節境界位置で正しくポーズを入れている割合である。単語の境界でポーズを入れた割合は只の一回だけで

表 6: 疑似文節単位の発声データ

Speaker	KO	MH	TI	平均
発声速度(音節/秒)	5.2	6.2	6.2	5.9
ポーズ数/文	2.5	2.4	2.1	2.3
文節境界の一致率	0.94	0.92	0.83	0.90

表 7: 疑似文節単位の発声による認識実験 (%)

話者	置換	挿入	脱落	正解	文正解
KO	25.6	13.7	2.8	71.7	8
MH	19.1	7.4	2.5	78.4	20
TI	21.8	10.6	3.2	75.0	13
平均	22.2	10.6	2.8	75.0	13.7

ある。

実験は表 4(a) の条件と同様で、ポーズ区間はパワーの閾値で検出し、無音区間を全て切り詰めた音声を用いる。認識時にはポーズの検出時に推定された文節境界のフレーム位置の情報を利用し、単語境界位置の制約を設けるようにした。

表 7にその認識実験の結果を示す。評価データは異なるが、表 4(a) で発声速度が slow の場合と比較すると認識精度は高くなっている。更に認識時にポーズ位置の制約を用いることも現在検討を行なっている。

5 結び

本報告では日本語の旅行案内対話データを用いて品詞 bigram モデルによる日本語の言語モデル化について検討し、大語彙連続音声認識へ適用し、約 5000 単語の ADD データベースについて認識実験を行なった。

ADD 大語彙データベースの連続音声認識実験について、学習データが不十分の時に品詞による一次近似モデルは本タスクに対して予想通りの認識結果が得られた。普通の速度での発声の時に、話者独立/適応の場合にそれぞれ約 70/71%の単語認識率が得られた。やや遅い発声スピードの場合には、単語認識率がそれぞれ約 73/74% であった。

助詞・接頭語誤りに対して、助詞・接頭語の助詞モデルを品詞 bigram に結合し、大語彙連続音声認識へ改善を試みた。助詞/接頭語の品詞 trigram の場合は認識率の改善はあまり見られなかったが、単語助詞/接頭語の bigram モデルにより、認識率が少し良くなった。疑似的な文節単位で区切って発声された文データに対しては、話者適応化モードで、75%の単語認識率が得られた。

以上の大語彙連続音声認識へのアプローチによって、品詞に基づく確率モデルが学習コーパスが小さい時にも有効であることが分かった。セグメントの統計量を用いた音響モデルは話者独立の認識実験で非常に効果があった。今後、更なる認識精度の改善と処理時間の短縮が重要な課題である。

参考文献

[1] 中川, 伊藤 : "音節標準パターンと逆時間向き係り受け解析法を用いた日本語文音声の認識", 信学論, Vol. 70-D, No.2, pp. 2469-2478 (1987).

[2] 中川, 大黒, 橋本 : "構文解析駆動型日本語連続音声認識システムー SPOJUS-SYNO", 信学論, Vol.72-DII, No.8 pp.1726-1280 (1989).

[3] V.Zue, J.Glass, D.Goddeau, D. Goodine, L.Hirschman, M.Phillips, J.Polifroni, and S.Seneff: "The MIT ATIS System: February 1992 Progress Report", *Proc. ARPA Human Language Technology Workshop*, pp. 84-88 (1993).

[4] Douglas B.Paul: "The Lincoln Large-Vocabulary Stack-Decoder Based HMM CSR", *Proc. ARPA Human Language Technology Workshop*, pp. 399-404 (1994).

[5] J.L.Gauvain, L.Lamel, and M.Adda-Decker: "Developments in Continuous Speech Dictation using the ARPA WSJ Task", *Proc. of the IEEE*, pp. 65-68 (1995).

[6] M.Zhou: "A Study on Stochastic Models for Spoken Language", *Ph.D. thesis, Toyohashi University of Technology* (1996.1).

[7] 政瀧, 松永, 匂坂: "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 (1995-11).

[8] 大附, 森岳, 松岡, 古井, 白井 : "新聞記事を用いた大語彙連続音声認識の検討", 信学技報, NLC95-55, SP95-90, pp.63-68 (1995-8)

[9] J.H.Wright, G.J.F.Jones and E.N.Wrigley, "Hybrid grammar-bigram speech recognition system with first-order dependent model", Proc. ICASSP pp.I-169-172 (199).

[10] 周, 中川 : "確率モデルによる後続単語予測と大語彙日本語連続音声認識", 情報処理学会, 音声言語情報処理研究会, 95-SLP-6-1, pp. 1-8 (1995.5).

[11] 中川 : "確率モデルによる音声認識", 電子情報通信学会 (1988).

[12] 堤真理子: "確率モデルによる音声認識・言語モデルの研究", 豊橋技術科学大学, 実務訓練報告書 (1996.3).

[13] 堤, 周, 甲斐, 中川: "対話音声認識の言語制約としての文脈自由文法と統計的モデルの比較", 日本音響学会, 平成 8 年度春季研究発表会, 講演論文集, pp. 175-176 (1996.3).

[14] 山本, 中川 : "セグメント単位入力 HMM とその評価", 信学技報, SP95-104, pp.77-84 (1995-12).

[15] 中川, 甲斐 : "文脈自由文法制御による One Pass 型 HMM 連続音声認識法", 電子情報通信学会論文誌, Vol.J76-D-II, No.7, pp.1337-1345 (1993).