電子情報通信学会　第二種研究会（1991. 7）
講演討論会
「音声認識の新潮流をめざして」

音素文脈依存モデルと高速な探索手法を用いた連続音声認識システム niNja

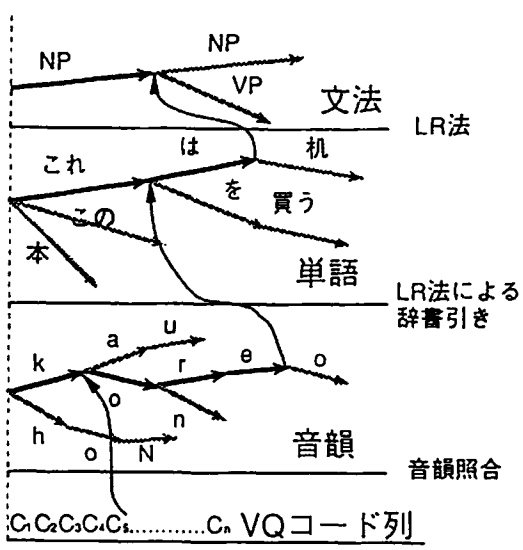
伊藤克亘　速水　悟†　田中穂積
東京工業大学工学部
†電子技術総合研究所

1　概要

われわれは、音声を含めた日本語による自然言語インターフェイスを目指すシステム niNja (Natural language INterface in JApanese) の試作をおこなっている [1]。本稿では、そのシステムの精度をあげるために音素文脈依存音韻モデルを導入する手法と、効率化のために音韻レベルでの処理を統合する手法について述べる。

2　システムの構成

本システムでは、各階層の履歴を木構造で保持する。このようにすると、それぞれの階層の情報に対する制約の評価は、同じ節点に対して 1 度おこなえば、その他の仮説では、過去の結果を利用できる。したがって、無駄な再計算を防げる。



システムの動作を簡単に説明する。まず、VQ コードを音韻モデルと照合する。音韻モデルの最終状態に達した経路は、辞書の LR 表を参照して、次に探索する音韻モデルを決める。辞書の LR 表で単語 (形態素) が完成したら、その構文的なカテゴリを (文法レベルの) LR 表にかえす。LR 表は、辞書の LR 表からうけとった構文カテゴリに応じて、次の辞書引き処理を駆動する。これらの動作を各フレームごとに、全ての階層でおこなう。

このように処理をすすめると、例えば、o N s e i という発話を処理する場合に、t_i フレームで照合を終了した o N と t_{i+1} フレームで照合を終了した o N と t_{i+2} フレームで照合を終了した o N では、全て音韻照合のスコアが異なる。本システムでは、これらの仮説を全て別々に記憶する。この単位を認識候補とよぶ。認識候補は、それぞれがスコアを保持する。このスコアは、その仮説の音韻の生起確率の対数をとったものとする。また、その認識候補が各階層でどのような履歴を持つかを、各階層の木構造での該当する節点へのリンクという形で保持する。

このように、音韻ラティスや単語ラティスなどの中間的な構造を介さず、VQ コード列から直接認識結果をつくる。したがって、情報の欠落が防げ、高精度な処理をおこなうと期待できる。しかし、このような処理方法では、処理をすすめるとともに認識候補が爆発的につくられるという問題が生じる。

解析途中で認識候補の数が爆発的に増大しないように、本システムではビームサーチを導入する。しかし、単にビームサーチを導入するだけでは精度が悪くなるので、なんらかの対策が必要である。本システムでは以下で述べるように、音韻モデルの精度の向上のために音素文脈依存モデルを導入し、同じ認識候補数での処理量の削減のために音韻レベルでの処理の統合をおこなう。

LR 構文解析法は、基本的に次の終端記号を先読みして動作を決定することで、無駄な処理をおこなわない特徴をもつ。しかし、例えば、

名詞 → h o N

のように、辞書を音韻を終端記号にした規則とみなすと、ふつうの LR 構文解析法では、この h o N の辞書引き処理を完了するためには、N の次の終端記号を先読みしなければならない。しかし、語 (形態素) の境界では、未知語や自由な語順をゆるしたばあい、あらかじめ動作をきめて LR 表をつくろうとすると、先読みとしてすべての音韻 (終端記号) をかんがえることになる。したがって、先読みをおこなう利点がない。

このような理由から、辞書の LR 表については、レデュースのときに先読みをおこなわないようにする。つまり、先の例では、終端記号 N を処理したところでレデュースする。

3 音素文脈依存モデルの導入

3.1 音素文脈依存モデル

本稿では、文献 [2] の手法を用いて推定した音素文脈依存音韻 HMM を用いる。この手法では、先行音素または後続音素の調音特徴にもとづいて決定木を用いて音素文脈を分類し、音韻モデルをあらかじめ決めたモデル数に応じて最適に分類することができる。

しかし、音素文脈依存モデルを連続音声認識に用いる場合、形態素 (単語) の境界にある音素では、あらかじめ音素文脈が決まらないので、その扱いについてかんがえなければならない。本論文では、音素文脈があらかじめ決まっている形態素内の音素については、はじめから音素文脈依存音韻モデルをわりあてておき、形態素の境界の音素については、動的に音素文脈依存音韻モデルをわりあてるという手法を用いる。以下、その手法について説明する。

3.2 音素文脈依存モデルの導入

本システムでは、辞書と文法のふたつの LR 表を用いる。辞書の LR 表の終端記号は、(音素文脈独立な) 音韻であり、文法中には音韻があらわれないようになっているので、本システムでは、実行前に決定している音素文脈の音韻については、あらかじめ文脈に応じて書き換えてしまい、実行前に決定しない部分については、パーザを拡張して動的に対応する。

3.2.1 辞書の変換

音素文脈独立な音韻表記の辞書を用意する。

名詞 → h o N
名詞 → z a q sh i

次に、音素文脈ごとの音素と音素文脈依存モデルの対応表を用意する。

音素文脈	モデル
h(i,o)	h3
o(h,N)	o10
a(z,q)	a4
q(a,sh)	q0
sh(q,i)	sh5

この表にしたがって、辞書を変換する。

名詞 → h(*,o) o10 N(o,*)
名詞 → z(*,a) a4 q0 sh5 i(sh,*)

このとき、辞書の項目の最初の音素と最後の音素については、辞書のレベルでは音素文脈が決まらないので、モデルには変換せず、決定している音素文脈だけ記録しておく。h(*,o) は、後続の音素が o であり、先行の音素が決定していない音韻 h を表し、N(o,*) は先行の音素が o であり、後続の音素が決まっていない音韻 N を表す。

このように辞書を変換したら、あとは、通常と同じ方法で辞書の LR 表をつくる。(一部を示す。)

状態	h(*,o)	z(*,a)	o10	N(o,*)	a4
0	sh1	sh3			
1			sh2		
2				re0	
3					sh4

終端記号の種類がふえるので LR 表は大きくなるが、どれだけモデルの数を多くしても、LR 表のエントリ数はたかだか辞書項目の音素数の総和になる程度である。

3.2.2 LR パーザの拡張

認識候補が音素文脈を保持できるようにする。そして、以下のように辞書引きの処理を変更する。ここでは、「ここに本があります。」という発話の「本」の部分を例にとって説明する。

(1) 辞書引き処理を駆動するとき

辞書引き処理を駆動する認識候補は、それぞれ次にくる音素を決定している。k o k o n i (ここに) の i の場合であれば、モデル数が 128 個の場合には、i(n,*) にたいして、次の 5 つのモデルがわりあてられる。

モデル	後続音素
i5	a, a-, e, e-, i, i-, o, o-, u, u-
i7	ch, f, k, ky, p, py, ry, s, sh, t, ts
i13	b, by, d, dy, j, r, w, y, z
i15	N, g, gy, m, my, n, ny
i17	#, h, hy, q, >

これらのモデルはそれぞれ、表に示された可能な複数の後続音素をもつ。つまり、k o k o n i には、最後の i に対するモデルがことなる 5 つの認識候補があり、それぞれ、次にくる音素を決定している。

次に駆動する辞書の LR 表の開始状態 (状態番号が 0 の状態) をみる。このとき、i17 をもつ認識候補の場合は、後続音素として h を許しているので、LR 表で示される音素文脈 h(*,o) を h(i,o) とみなして、この節の初めに示した音素文脈と音素文脈依存モデルの対応表をみてモデル h3 を駆動する。そして、そのモデルの照合が終了したら、LR 表にしたがって、状態 1 にすすむ。同様に、i13 をもつ認識候補は、後続音素として z を許しているので、状態 3 にすすむことができるが、その他の認識候補は、ここですてられる。

(2) 開始状態以外の状態でのシフトのとき

先読み記号が音素文脈依存モデルなので、そのままそのモデルを駆動して次の状態にすすむ。例えば、(1) で h3 を駆動した認識候補では、状態 1 にすすむと、o2 を駆動して状態 2 にすすむ。

(3) レデュースのとき

レデュースのときは、次にくる音素が決まらないので、先行音素だけから決まる可能な音素文脈依存モデルを全て駆動する。例えば、辞書の LR 表の状態 2 では、N(o,*) となっているので、この音素文脈に対してゆるされるモデルの数に応じて認識候補をつくる ((1) 参照)。

このような手法では、辞書に登録されている項目数が少ないときには、辞書のレデュースのところで無駄に駆動される音韻モデル数が多くなってしまう。しかし、辞書に登録される項目数がふえれば、その差は小さくなるので、大語彙をかんがえる場合にはそれほど問題ない。

4 音韻レベルの処理の統合

One Pass DP 法は、有限状態オートマトンによる比較的単純な制御構造を持ち、入力音声のフレームに同期して処理が進むという優れた特徴を持つ。

本システムでは、2 節で述べたように辞書やそれまでのスコアにもとづいて音韻レベルの木構造を作っている。認識単位が音韻なので、この木構造を有限状態オートマトンとみなして、One Pass DP 法の制御をおこなう。

つまり、認識候補は、そのフレームでの有限状態オートマトンのある状態に対応し、ひとつずつが異なった音韻モデルの系列をあらわしている。この音韻系列には、次の特徴がある。(1) 語彙数や規則数がふえても、音韻系列にあらわれる音韻モデル数はある一定数である。(2) 異なる音韻系列に同じ音韻モデルが何度もあらわれる。

本論文で提案する手法では、このふたつの特徴に着目して、同じ音韻モデルに対する照合は、ある認識候補だけがおこない、他の認識候補はその結果を使う。このようにすると、理論的な最適性は失われることもある [1] が、照合のための処理量は語彙数や規則数に依存しない量で、(最適性は失われているので、近似的にではあるが) N-Best な解を求めることができる。したがって、従来の N-Best な解を求める手法よりも処理量が少ない。

アルゴリズム

各フレームで、以下の手順で処理をおこなう。

(1) 音韻モデルの最終状態に達している経路については、その音韻での経路の初期状態をみる。各経路の初期状態には、認識候補の集合が記憶されている。このそれぞれの認識候補について、スコアの更新と次に連結する音韻モデル (複数のこともある) の決定をおこなう。

経路のスコアは、(3) でのべるように、初期状態での認識候補の集合の要素のうち、最もスコアのよい認識候補に対するスコアである。初期状態での認識候補は、それぞれ最大のスコアをもつ認識候補との差を保持している。したがって、その分を経路のスコアから引いたものを現在のスコアとする。

次に連結する音韻モデルは以下のようにきめる。認識候補の LR 表での動作がシフトのときは、表がしめす次の状態に遷移する。動作がレデュースのときは、辞書引き結果の構文カテゴリにしたがって、文法の LR 表が動作して次の辞書引き処理を始めて、辞書の LR 表の開始状態に遷移する。このようにして、それぞれ遷移した状態で先読みとなっている音韻モデル (複数のこともある) を次に連結する。

(2) 音韻モデルごとに、(1) の処理をおえた認識候補の集合をつくる。

(3) 各音韻モデルについて、認識候補の集合の要素のうち、最大のスコアをもつもののスコアを、そのフレームでの初期状態でのスコアとする。他の認識候補は、最大のものとのスコアの差を保持しておく。

(4) 音韻モデルの最終状態以外の状態については、ビタビサーチをおこなう。ただし、フレームごとに、初期状態に設定される過去の経路が違う ((3) 参照)。したがって、最適でない探索がおこなわれることもある [1]。

(3) のように各音韻モデルは一度しか照合しないので、フレームごとの音韻照合の処理量は、認識候補の数にかかわらず、音韻モデル数を n とすると $O(n)$ である。しかし、(1) でのスコアの更新のところの処理は、認識候補数を m とすると $O(m)$ なので、候補数がふえると処理量が多くなる。

本システムでは、そういった問題を解決するため、ビームサーチを導入する。具体的には、各フレームごとに、そのフレームの認識候補の最大のスコア (正規化したもの) から λ の対数をとった値を正規化した値を引いた値以下の認識候補については、枝刈りする。このような手法をとると、認識候補の数が一定以下になる保証はないが、認識候補をソートする必要はない。

5 認識実験

5.1 音声資料

実験に用いた HMM の訓練用音声資料は単語音声と連続音声からなる。単語音声資料の話者は成人男性 5 名で、発声用テキストは音韻バランス単語集合 WD-II (1542 語) である。連続音声資料の話者は成人男性 2 名で、発声用

のテキストは ATR 音韻バランス文 150 文である。これらの収録は簡易防音室でおこなった。

認識実験に用いたモデルは、文献 [2] の方法でおこなった。モデルの数は、43 個 (音素文脈独立)、128 個、256 個、512 個、1024 個である。

実験に用いた音声資料の発声用テキストは 11 文 (文節数は 33) の疑問文などである。このテキストを成人男性の 2 名分を防音室で、 8 名分を計算機室で収録した。これらの話者・テキストは HMM を訓練した資料には含まれておらず、不特定話者・語彙独立な実験条件とした。

5.2 認識実験

以下のふたつの実験をおこなった。

(1) 音素文脈依存モデルの効果

音素文脈依存モデルのモデル数と認識率の関係をしらべるために、辞書・文法・枝刈りのためのパラメータ λ は一定で、モデル数だけを変化させて実験をおこなった。

この実験に用いた総単語数は 113 である。文法は、単語をアークとするネットワーク表現と等価なものを文脈自由文法で記述した。その平均分岐数は 4.1 である。なお、文節間には、任意の無音区間を許す。

結果を次にしめす。

モデル数	43	128	256	512	1024
文認識率 (%)	81.8	82.7	85.5	90.9	92.7
文節認識率 (%)	88.2	89.4	92.7	95.5	96.1

この表からあきらかなように、モデル数が多くなるのにつれて認識率も向上した。モデル数が 1024 個の場合には、音素文脈独立モデルを用いた場合と比較すると、文節認識率で誤りが 67 % 低減し、文認識率で 60 % 誤りが低減した。

(2) 音韻レベルの処理を統合した効果

音韻レベルの処理の統合によってどれだけの処理がはぶけるかをしらべるために、実験 (1) に使った文のうち、ひとつの文を用いて、各フレームごとの認識候補の数と実際に照合をおこなった音韻数を枝刈りのためのパラメータ λ をかえて測定した。この実験では、辞書として音韻バランス単語集合 WD-I (492 単語) に実験 (1) で用いた 11 文に含まれる単語 22 語 と付属語 12 語を加えたものを用いる。文法は、辞書に含まれる単語から作られた文節が任意につながったものを文とみなすものを用いる。この文法は、非常に制限がゆるい。

認識実験の結果を次にしめす。

$\lambda = 1.0 \times 10^{-3}$ のとき

< s o n o # h o N o # k a e u t a # i- k i o #
u o i ch i b a # j i p u sh i- w a >

(その本を替え歌好い気を魚市場ジプシーは)

$\lambda = 1.0 \times 10^{-20}$ のとき

< s o n o # h o N o # k a i t a i t o #
o m o i m a s u k a >

(その本を買いたいと思いますか) (正解)

この例の場合、枝刈りの条件をゆるくする (λ を小さくする) ことで、認識結果として正解がえられた。このとき、認識中に生成された認識候補の総数は、$\lambda = 1.0 \times 10^{-3}$ のときは 881363、$\lambda = 1.0 \times 10^{-20}$ のときは 1764036 と枝刈りの条件をゆるくすると、2 倍程度になる。しかし、途中で照合された音韻モデルの総数は、$\lambda = 1.0 \times 10^{-3}$ のときは 239285、$\lambda = 1.0 \times 10^{-20}$ のときは 249376 と、ほとんど変わらない。このように、本論文で提案した音韻レベルの処理を統合する手法をビームサーチと組合せて用いると、音韻照合の処理量をほとんどかえずに、ビーム幅を大きくして認識精度をあげられることがわかる。

6 まとめ

本稿では、連続音声の VQ コード列からフレーム同期で、音韻系列・単語 (形態素) 系列を自動的に構成するシステムでの精度をあげる手法と効率をあげる手法について報告した。精度をあげるため、音韻モデルに音素文脈依存モデルを用いて、連続音声認識にも有効であることをしめした。音素文脈依存モデルを用いた場合、モデル数が 1024 のときに文認識率 92.7 % がえられた。効率をあげるためには、VQ コード列を音韻モデルと照合して音韻系列を構成する部分での処理を統合する手法について提案した。この手法では、照合の処理量が途中に保持される候補の数に依存しない。

なお、音韻・単語・文法やそれ以上のレベルでの言語情報の導入や、今回実験をおこなったものより、さらに複雑なタスクにおいても、本論文で提案した手法が有効であるかどうかを検討することなどが、今後の課題としてあげられる。

参考文献

[1] 伊藤克亘, 速水悟, 田中穂積. 拡張 LR 構文解析法を用いた連続音声認識. 信学技報, SP90-74,pp. 49-56, (1990-12).

[2] 速水悟, 田中和世. 木構造音韻モデルによる未知音素文脈中の音響的変動の予測と評価. 信学技報, SP90-64,pp. 55-62, (1990-12).