連続音声認識システム NINJA での音韻連鎖統計情報の利用　15－4

Phone N-gram model in Continuous Speech Recognition System : NINJA

伊藤克亘　速水 悟†　田中穂積
ITOU Katunobu　HAYAMIZU Satoru　TANAKA Hozumi
東京工業大学工学部
Tokyo Institute of Technology
†電子技術総合研究所
Electrotechnical Laboratory

It is difficult to recognize individual phones completely in continuous speech recognition system with large vocabulary. This paper reports a new attemp to improve recognition performance of continuous speech recognition system : NINJA using a statistical language model of phone n-gram approach. The phone n-gram model is used for both improvement of recognition accuracy and pruning in this system. In this paper, we added the phone n-gram model to the system using a dictionary and grammar. The probabilities of the phone n-gram model are estimated using text dababases. The databases consist of newspaper articles and dialogue about international conference. They contain 99530 bunsetu and 859311 phones. A bunsetu recognition error reduction rate of 12.1 % and a sentence recognition error reduction rate of 17.3 % were achived in continuous speech.

1　はじめに

音声による自然言語を用いたマンマシンインターフェイスを考えるとき、大語彙で連続音声が扱えることが必須条件となる。

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) の試作を行っており [1]、本稿ではそのシステムでの音韻連鎖統計情報の利用について述べる。

本システムでは、処理途中の候補の絞り込みに音響的な情報として、HMM のスコアを用いている。しかし、大語彙の連続音声を認識する場合、音韻認識の精度には限界があることから、言語情報を利用しなければならないと考えられる。候補の絞り込みに有効な言語情報は様々なものが考えられるが、そういった情報のひとつとして、音韻連鎖に関する統計情報が注目されている [3] [6]。

村上らは、音節入力のシミュレーションに対して、このモデルの有効性を示している [6]。また、川端らは、辞書を用いない場合の文節音声に対しての有効性を示している [3]。しかし、実際の連続音声認識システムにおいて、音韻連鎖に関する統計情報を辞書や文法と併用する場合の検討はまだなされていない。

本稿では、これらの観点をふまえて連続音声認識システムにおける音韻連鎖の統計モデルの有効な利用法について検討し、認識実験により、その有効性を実証する。

2　NINJA の構成

システムの構成の概略を 図 1 に示す。システムは大きく分けて構文解析と辞書引きのふたつの部分から構成される。処理はこのふたつの部分が相互に情報を受渡すことによって進められる。

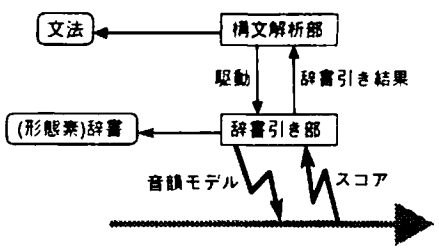

図 1 : システムの概略図

まず、あらかじめ用意された文法規則にしたがって、構文解析部が辞書引き処理を駆動する。辞書引きの部分は、あらかじめ用意された辞書を参照して音韻 HMM を駆動

連絡先 : 伊藤克亘　東京工業大学工学部　情報工学科　田中研究室　〒 152 東京都目黒区大岡山 2 - 12 - 1
電話　(03) 3726 - 1111 (内線) 4175　FAX　(03) 3720 - 4925　e-mail　itou@cs.titech.ac.jp

する。ここでは、音韻 HMM として音素文脈依存モデルを用いる [5] 。

そして、音韻ラティスなどを構成することなく、音声入力のパラメータ系列から、並列に複数の辞書引きをフレーム同期をとりながら行う。HMM の照合には、One Pass DP アルゴリズムを基本にした経路探索を、拡張 LR 構文解析法を用いた辞書引きアルゴリズムを用いて制御する。一般には、このような手法で N-Best の解を求めるアルゴリズムは、辞書に登録されている語数に依存する。しかし、本システムのように辞書引きの結果を組合せて、文節や文を構成していく手法では、辞書引きのアルゴリズムが登録されている語数に依存すると、候補が組合せ的に爆発してしまうおそれがある。そこで、本システムでは、辞書に現れる同じ音韻の照合をひとつにまとめて行う。このことにより、解の理論的な最適性は失われるが、音韻照合に必要な計算量は辞書に登録された語数に依存しないので、時間的・空間的に効率のよい探索ができる [1]。また、解析途中で候補の数が爆発的に増大しないように各フレームごとに枝刈りを行う。

辞書引きが完了すると、辞書引き部は形態素の構文的なカテゴリを構文解析部に渡す。この部分には、拡張 LR 構文解析法を用いる。結果を渡された構文解析部は、その形態素を含む経路の適切さを構文などの制約によって判別し、文法規則にしたがって次の辞書引き処理を駆動する。このような処理を繰返して、解析結果を次々に出力していく。

このようにして、得られた結果のうち最もスコアの高いものを認識結果とする。しかし、大語彙のシステムでは、HMM のスコアだけでは得られる精度に限界がある。そこで、認識精度を上げるための新たな情報として、本稿では、3 章で説明する音韻連鎖に関する統計情報を導入する。

また、この情報は局所的ではあるが、あらゆる日本語の系列に対して計算できる。したがって、本システムのようなフレーム同期での枝刈りの評価基準としても適している。枝刈りの評価基準の精度が上がれば、途中で正解の候補が枝刈りされる可能性が減り、認識精度の向上が期待できる。

3 音韻連鎖の統計モデル

3.1 N-gram モデル

ある発話を音韻の系列とみなす。すると、音韻系列 $P(=p_1, p_2 \cdots, p_n)$ が発話される確率 $P(\mathrm{P})$ は式 (1) で表される。

$$P(\mathrm{P}) = \prod_{i=1}^{n} P(p_i|p_1, p_2, \cdots, p_{i-1}) \qquad (1)$$

ここで、$P(p_i|p_1, p_2, \cdots, p_{i-1})$ は音韻系列 $p_1, p_2, \cdots, p_{i-1}$ が発話され、次の音韻として p_i が発話される確率である。理論上は n が無限大になりうるため、$P(p_i|p_1, p_2, \cdots, p_{i-1})$ の値をあらかじめ用意するわけにはいかない。そこで、なんらかの方法で $P(p_i|p_1, p_2, \cdots, p_{i-1})$ を近似的に求めなければならない。

本稿では、音韻連鎖の統計的な言語モデルとして N-gram モデルを用いる。N-gram モデルでは、式 (1) の $P(\mathrm{P})$ を式 (2) で示すように近似する。

$$P(\mathrm{P}) = \prod_{i=1}^{n} P(p_i|p_{i-N+1}, p_{i-N+2}, \cdots, p_{i-1}) \qquad (2)$$

このような近似方法を用いるため、N の値が大きければ大きいほど利用できる情報が増え、推定精度がよくなる。しかし、一般には、この統計値を実際のデータでの出現確率から得るので、N が大きくなると組合せの数は急速に大きくなる。したがって、データの数が限られる場合には、あまり N が大きくなると、確率値の推定精度が悪化する。

また、音韻連鎖の統計モデルは N が大きくなるとタスクによる違いが生じることが知られている [2]。

3.2 N-gram モデルの生成

表 1 に示すデータベースを用いて N-gram モデルを生成した。

表 1 : テキストデータベース

出典	文数	文節数	音韻数
新聞	8147	77237	648753
キーボード会話	1376	7717	76781
電話会話	1684	14576	133777
合計	11207	99530	859311

「新聞」は日本経済新聞 1982 年の新聞記事 19 日分、「キーボード会話」「電話会話」は、それぞれ国際会議に関する会話を含んでいる。

これらのデータベースは図 2 に示す形式のデータで構成されている。

```
k o t o g a
d e k i m a s u
$
```

図 2 : テキストデータの例

ここで、$ は文の区切りを示す。データはこのように文節ごとに区切られた文から構成されている。実際に音韻連鎖の統計を調べるときには、文節の境界を示す記号を用意して、文内では文節の境界をこえる音韻連鎖に関しても考慮している。

3.3 NINJA での N-gram モデルの利用

NINJA では 式 (3) で示したスコアをそれぞれの認識結果について求める。

$$P_{total} = \mathrm{P}_{hmm} + w_{gram}\mathrm{P}_{gram} \qquad (3)$$

ただし、

$$\mathrm{P}_{hmm} = \frac{\sum_{t=1}^{N_{frame}} \log P_{hmmt}}{N_{frame}}$$

$$\mathrm{P}_{gram} = \frac{\sum_{p=1}^{N_{phone}} \log P_{gram_p}}{N_{phone}}$$

ここで、w_{gram} は N-gram モデルの重みである。本稿では、この値を P_{hmm} と P_{gram} のダイナミックレンジを考慮して適当に決めている。

また、本システムでは、各フレームごとに、辞書引き部で音韻を組合せるときに枝刈りを行なう。そのときも各候補ごとに式 (3) のスコアを求めて評価に用いる。

4 文認識実験

4.1 実験条件

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) [4] 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。これらの収録は防音室でおこなった。HMM は、[5] に示した方法で訓練した音素文脈依存モデルであり、モデル数は 512 個である。

言語モデルは、表 1 のデータを全て用いて生成した。

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。このテキストは、言語モデルを作成したデータには含まれていない。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いない。

辞書は、音韻バランス単語集合 WD-I (492 単語) [4] に実験用のテキスト 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いくつかつながったものを文とみなす、非常に制限のゆるいものを用いる。

4.2 評価基準

連続音声認識システムでの言語モデルの効果を評価する尺度として、文認識率と文節の挿入・脱落を考慮した文節認識率を用いる。

文節認識率は、次のようにして求める。まず、認識結果の文節系列と正しい文節系列との DP マッチングを行い、文節の置換・挿入・脱落誤りの個数を求める。これらの個数から、式 (4) を用いて文節認識率を計算する。

$$\text{文節認識率} = \frac{\text{全文節数} - \text{置換} - \text{挿入} - \text{脱落}}{\text{全文節数}} \qquad (4)$$

4.3 認識実験

言語モデルのスコアが音声認識システムにおいて、どのような効果を及ぼすかを調べるために、1 名の話者で式 (3) の重み (w_{gram}) をいろいろ変えて実験を行った。言語モデルとしては、式 (2) の n を 3 としたもの (トライグラム) を用いた。認識率を 表 2 に、認識結果の例を 図 3 に示す。

表 2： 言語モデルの重みと認識率の関係

重み	0.0	0.1	0.2	0.3	0.4
文節認識率	66.7	72.7	78.8	72.7	78.8
文認識率	45.5	54.5	54.5	54.5	63.6

0.5	1.0
60.6	51.5
45.5	18.2

言語モデルなし ：< n e d a N w a # i k u r a #
i m i k o t o b a >
(値段はいくら忌み言葉)
重み = 0.4 ：< n e d a N w a # i k u r a
d e s u k a >
(値段はいくらですか) (正解)
重み = 1.0 ：< d a r e # d a r e w a #
i k u r a d e s u k a >
(誰誰はいくらですか)

図 3： 認識結果の例

ここで、表 2 の重みが 0.0 の場合とは、言語モデルを使わない場合である。図 3 の # は文節の区切り、< . > はそれぞれ文頭、文末をあらわす。

この結果から、言語モデルを適当に用いることによって認識率が向上することがわかる。ただし、言語モデルに対するスコアの重みを大きくしすぎると、逆にデータによく出現する系列 (単語) を含む候補ばかりが残り、正解が途中で枝刈りされ、認識率が悪くなる。このことは、解析途中でどの候補を枝刈りするかにも、言語モデルが強く影響することを示している。

次に、N の数が異なるときに N-gram モデルの音声認識システムにおける効果を調べるために 1 名の話者で、さまざまな言語モデルを用いて実験を行った (重みは 0.4)。結果を 表 3 に示す。

表 3：音韻連鎖の長さと認識率の関係

N	0	1	2	3	4	5
文節認識率	66.7	60.6	69.7	78.8	66.7	51.5
文認識率	45.5	45.5	54.5	63.6	54.5	45.5
音韻連鎖出現率	—	100	100	100	98	94

N が 1, 2, 3 の場合は学習データ中に、実験データに出現する音韻連鎖が全て含まれている。このことから、これらの N の場合には今回利用したデータ量でも、N-gram モデルの信頼性はかなり高いと考えられる。このうち、N が 1 の場合は、言語モデルを使わない場合よりも文節認識率が悪くなっており、連続音声認識に利用するには、不十分であると考えられる。N が 4, 5 のように連鎖が長くなると、音韻連鎖出現率が 100 % ではなくなる。また、認識率も悪くなる。これらのことから、今回利用したデータ量では、4-gram や 5-gram モデルを生成するには不十分であるか、もしくは、データベースと実験データのタスクの類似度が低いと考えられる。

最後に、言語モデルが実際の不特定話者を対象とした連続音声認識システムでどの程度有効であるかを調べるために、10 人の話者に対して 11 文ずつ (計 110 文) で実験を行った (用いた言語モデルはトライグラムで、重みは 0.3)。

表 4：不特定話者認識実験の結果

言語モデル	文節認識率	文認識率
なし	70.3	52.7
あり	73.9	60.9

表 4 に示すように、文節認識率で 12.1 %, 文認識率で 17.1 % の誤りが減少した。この実験では、収録条件に差があるため、話者ごとの文節認識率には大きなばらつきがある。話者ごとの認識率の変化をみると、言語モデルなしで認識したときに、60 % 程度の認識率になる話者については、文節認識率が向上するが、75 % を越えるようなものや、60 % 以下の認識率のものについては、言語モデルの効果はそれほど大きくない。

認識率が低い場合に言語モデルの効果が小さい理由としては、N-gram モデルが局所的な音韻連鎖しか考慮しないため、正解からかけ離れた候補を含む場合には、全く正解とは異なった系列に対して高いスコアを与えてしまう可能性があることがあげられる。

5 おわりに

連続音声認識システムにおける音韻連鎖統計情報の利用法を示し、実際のシステムで不特定話者の連続音声認識実験をおこない、その効果を示した。

今後の課題としては、テキストデータベースのデータ量とタスクの種類を考慮に入れて、より推定精度の高い音韻連鎖の統計モデルを生成すること、今回よりもさらに難しいタスクを用いた定性的・定量的な評価などが上げられる。

謝辞

本稿で使用したテキストデータベースのうち、日本経済新聞の記事については、NTT 情報通信処理研究所メッセージシステム研究部から、国際会議に関するものについては、ATR 自動翻訳電話研究所から提供していただきました。貴重なデータを使用させていただいたことを感謝いたします。

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. 信学会技術報告, SP90-74:49-56, 12 1990.

[2] 伊藤克亘. 日本語音韻の統計的なふるまいを利用した連続音声認識. 修士論文, 東京工業大学, 2 1990.

[3] 川端豪, 花沢利行, 伊藤克亘, 鹿野清宏. HMM 音韻認識における音節連鎖統計情報の利用. 信学会技術報告, SP89-110:7-12, 1 1990.

[4] 速水悟, 田中和世, 横山晶一, 太田耕三. 研究用音声データベースのための VCV/CVC バランス単語セットの作成. 電総研彙報, 49(10):803-834, 1985.

[5] 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. 信学会技術報告, SP90-64:55-62, 12 1990.

[6] 村上仁一, 荒木哲郎, 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. 信学会技術報告, SP88-29:63-70, 1988.