木構造を用いた音韻連鎖統計モデル

1E−3

田本 真詞　伊藤 克亘　田中 穂積
(東京工業大学)

1 はじめに

計算機における連続音声認識では、処理の効率化のために種々の言語情報を用いている。これらの言語情報としてシンボル連鎖に関する統計情報があり、認識の誤り訂正や曖昧性の解消に有効であることが知られている[1]。

連鎖の統計情報には音韻などのシンボルの生成をマルコフ過程とみなし、シンボル列の生成確率を近似するN-gramモデルがあり、統計的言語モデルとして注目されている[2]。N-gramによる言語モデルは与えられた観測データの量が限られている場合やマルコフ過程の次数を上げ、コンテキストの弁別性を高めた時にコンテキストの組合せ数の増加で生じるデータの減少・欠落によって統計的信頼性を損ねることがあった。

私達は、コンテキストに応じて参照するシンボルの連鎖長を変化させ、統計モデルの信頼性の低下を避けながら弁別性の高いモデルを生成する手法、「木構造を用いた音韻連鎖統計モデル(Vari-gram)」を提案した。Vari-gramは、コンテキスト生成と後続シンボル生成の結合確率のエントロピーを最大化するコンテキストに注目し、そのコンテキストの連鎖を後方に伸長して新たなモデルを生成する[3]。さらに、このモデルがN-gramより優れた特性を持つことを示した[4]。本論文ではVari-gramの設計方針と統計的連鎖モデルの良さの指標のひとつとされる条件付きエントロピーの関係について考察するとともに、Vari-gramモデルを連続音声認識システムの言語モデルに適用し、実際の音声認識における有効性について検証する。

2 音韻連鎖統計モデルの設計

マルコフ過程の次数を一意に定めるN-gramに対し、次数を固定せずに、観測データの持つ統計的性質やシンボルのクラス分けに応じて連鎖の長さを動的に変化させる可変長モデルが近年注目されている。可変長モデルとして、任意の音韻系列を分割して個々のモデルとする分割モデルと木構造を用いて段階的に連鎖長を増加し、徐々に次数の高いマルコフ過程を生成する手法がある。木構造統計モデルには、確率分布のコンテキスト依存性を示す条件付きエントロピーを最小化するモデル化[5]、相互情報量の最大化[6]などの生成方針がある。これらの方針はいずれもモデルのサイズに比較して十分大きな観測データを想定しており、先に述べた問題を個々のモデルの選択時には考慮せず、モデル生成のstopping criteriaや生成後のsmoothingなどで解消を行なっている。

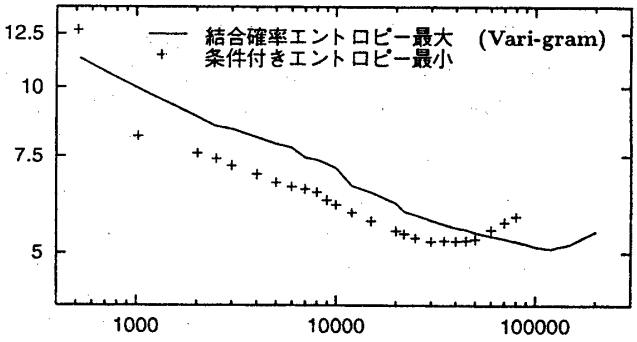
3 比較実験

Vari-gramと同様に新たなコンテキストが生成された時に条件付きエントロピーの増加が最小となるモデルを生成し、Vari-gramと比較する。実験に用いたテキストデータベースは、1982年の日本経済新聞の37日分の記事、文節数は145,718、音韻数で1,385,082である。また、生成されたモデルの評価用として、同新聞の1日文の記事、6260文節、54733音韻を別に用意した。

3.1 結果

両モデルを比較するとVari-gramでは、テスト文におけるモデルのカバー率を示すcoverageがモデル数の多い(50,000〜)ときに低下が少なく、モデルの平均連鎖長が全般的に大きい。また、モデルの条件付きエントロピーは、両モデルにほとんど差がない。モデル数の増加にしたがって認識処理の複雑さを示すパープレキシティーが変化する様子をグラフに表す。

図1: パープレキシティーの変化



Vari-gramモデルはパープレキシティーの増加がモデル数が増加に対して低く抑えられ、その最小値も若干小さい。一方、条件付きエントロピーの最小化をモデル生成の基準とした場合、コンテキスト依存性の高い確率分布がモデル化され、コンテキストの標本数が考慮されない。このためコンテキストの増加に伴い信頼性の低いモデルが生成されてモデル全体の信頼性が損なわれたと考えられる。

Vari-gramモデルの結合確率エントロピー最大化を基準としたモデル生成が、どのように信頼性の低下を避けているか考察する。Vari-gramモデルのすべてのコンテキストをL、後続する音韻をpとすれば、木構造全体の結合確率エントロピーは、H(L·p)、条件付きエントロピーは、H(p|L)となる。一般に両者の間にはH(L·p) = H(p|L) + H(L)が成り立つことから、Vari-gramでは結合確率エントロピーを増加させ、一方、条件付きエントロピーは減少するためにコンテキストLのエントロピーが増大する。この作用で各コンテキストの出現確率が均一になるように分割・生成が行なわれ、標本数の多いコンテキストのみがモデル化される。

A Tree-based Stochastic Phone Sequence Modeling.
Masafumi Tamoto, Katunobu Itou and Hozumi Tanaka
Tokyo Institute of Technology

表 1: 認識システム及び認識用試料

認識システム	
標本化周波数	15kHz
フレームシフト	10ms
音響分析	14 次メルケプストラム 及び 時間変化分 パワーの時間方向の変化分 29 パラメータ コードブックサイズ 1024 に量子化
音韻数	43 種類
音韻認識	4 状態 3 ループ 離散型 HMM
音韻モデル	音韻バランス単語 WD-II (成人男性 5 名) ATR 音韻バランス文 150 文 (成人男性 2 名)
発声試料	
実験資料	日本経済新聞記事 20 文 111 文節、951 音韻
発声様式	文節発声

表 2: 音声認識結果

	結合確率エントロピー最大 (Vari-gram)	条件付きエントロピー最小
w_{lang}	2.5×10^{-1}	2.0×10^{-1}
$w_{duration}$	2.0×10^{-2}	1.0×10^{-2}
音韻認識率	55.4%	48.2%
音韻仮説数	24649.4	41575.8

4 連続音声認識システムでの認識実験

連続音声認識システム niNja[7] の音韻タイプライタを用いた認識実験を通して Vari-gram の有効性を検証する。音韻タイプライタは、辞書の語彙に制約されることなく任意の音韻を仮説として生成する。

モデル生成に用いたテキストデータベースは、先の日本経済新聞の記事に対話データなどを加え、音韻数で 1,869,868 である。このデータベースから Vari-gram モデルを生成し、認識用のオートマトンをテストセットパープレキシティー最小のモデル (100,000 モデル) から作成した。また、同じデータから パープレキシティーが最小となった 4-gram モデル (39,815 モデル) を生成し、比較用とした。実験では音韻モデルのスコアに継続時間長モデルのスコアと言語モデルのスコアを重みをつけて加え最終的なスコアとする。継続時間長モデルのスコアの重み ($w_{duration}$) と言語モデルのスコアの重み (w_{lang}) を調整して認識率が最大になる点で比較を行なう。

認識率の評価は、認識結果の音韻列と正解とのDP マッチングをとり、文節内の音韻の置換・挿入・脱落誤りを減点して音韻認識率とした。

$$\text{音韻認識率} = \frac{\text{文節数} - \text{置換} - \text{挿入} - \text{脱落}}{\text{文節数}}$$

音韻認識率は、Vari-gram の方が若干優れており、音韻仮説の量も少ない。これは、Vari-gram のタスク抽出能力が高く、音韻仮説からの正解の選択性に優れていることを示している。

また、一般に音韻仮説の量が少なければ処理が高速に進むことから認識時間も短くなると考えられる。

5 結論・考察

Vari-gram は、コンテキストと後続音韻の結合確率エントロピーを最大化する方針でモデルを生成するが連鎖統計モデルの良さを示す条件付きエントロピーは、このエントロピーを最小化させるモデルとほとんど変わらない。

Vari-gram では、結合確率エントロピーと条件エントロピーの差分であるコンテキストのエントロピーが増加し、コンテキストの出現確率を分散させ、結果的に信頼性の高いモデルを生成している。

Vari-gram を言語モデルに用いた音声認識では、N-gram と比較して文節認識の誤り率が 14%減少し、音韻仮説の量も減少する。

統計連鎖モデルでは、タスクの曖昧さを示す条件付きエントロピーの抑制とコンテキストの示す確率の統計的信頼性を維持することが重要である。統計的信頼性を維持する基準としてコンテキストのエントロピーを最大化する手法が適切かどうか、また、コンテキストごとの標本分布から推定される確率の確からしさを評価する尺度をどう設定し、これを設計方針とすることでさらに信頼性の高い連鎖統計モデルを生成する手法の検討が今後の課題となる。

謝辞

本稿で使用した日本経済新聞の記事に関するテキストデータベースは、NTT 情報通信処理研究所メッセージシステム研究部から提供して頂きました。貴重なデータを使用させて頂いたことを感謝いたします。また、音韻モデルを作成するのに日本音響学会の研究用データベースの一部を使わせて頂きました。関係各位の御尽力に感謝いたします。実験に協力して頂いた電子技術総合研究所の速水悟氏に深く感謝致します。最後に、日頃から研究について御助言、御討論頂く田中研究室の皆さんに感謝いたします。

参考文献

[1] 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, Vol SP89-110, pp. 7-12, 1 1990.

[2] 中川聖一, 山本幹雄, 周旻. 連続音声認識のための確率文脈自由文法による言語のモデル化.「マルコフモデル・ニューラルネットワークを包含する新しい音声認識手法の総合的研究」研究成果報告書, pp. 339-348, 2 1992.

[3] 田本真詞, 伊藤克亘, 田中穂積. 木構造を用いた音韻連鎖統計モデル. 情報処理学会第 44 回全国大会講演論文集, 第 2 巻, pp. 167-168, 3 1992.

[4] 田本真詞, 伊藤克亘, 田中穂積. 木構造を用いた音韻連鎖統計モデル. 情報処理学会第 45 回全国大会講演論文集, 第 2 巻, pp. 9-10, 10 1992.

[5] 村上仁一, 嵯峨山茂樹. 単語連鎖可変長統計の自動学習に基づく連続音声認識. 日本音響学会講演論文集, pp. 95-96 10 1992.

[6] Lalit Bahl, Peter F. Brown, and Peter V. de Souza. A tree-based statistical language model for natural langae speech recognition. *IEEE, ASSP*, Vol. 37, No. 7, pp 1001-1008, 1989.

[7] 伊藤克亘. 連続音声認識システムに関する研究. Master's thesis, 東京工業大学, 3 1993.