論文 / 著書情報
Article / Book Information

論題(和文)	連続音声認識のためのネットワーク構造を用いた効率的探索手法
Title(English)	An efficient search method for continuous speech recognition using network structure
著者(和文)	花沢健, 南泰浩, 古井貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	日本音響学会 1997年春季講演論文集, Vol. , No. 2-6-4, pp. 51-52
Citation(English)	, Vol. , No. 2-6-4, pp. 51-52
発行日 / Pub. date	1997, 3

連続音声認識のためのネットワーク構造を用いた効率的探索手法*

○花沢健† 南泰浩‡ 古井貞熙†‡
(†東京工業大学 ‡NTTヒューマンインタフェース研究所)

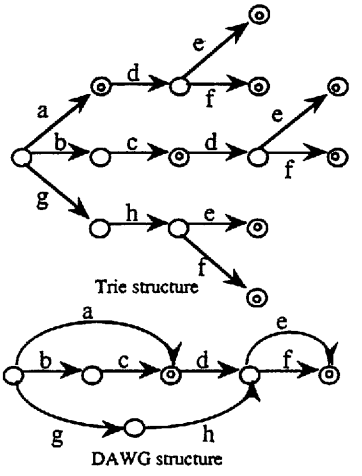
1. はじめに

従来我々の提案していたHMM-LR [1] では、LR文法を用いていたために冗長な探索を多く行っていた。HMM-LRや、木構造を単語辞書の記述方法として用いる手法では、辞書の構造において接頭辞は併合しているが接尾辞は併合していなかった。このため、共通の接尾辞を持つ単語(人名など)を非常に多く含む辞書を構成する場合には効率的でない。そこで本論文では、大語彙連続音声認識のためのコンパクトなデータ構造と効率の良い探索手法を提案する。提案手法では、探索空間を削減するために辞書の構造としてDAWGというコンパクトな構造を用いる。また、この構造の利点を生かし、N-bestの仮説を得るための効率の良い探索アルゴリズムを提案することにより、認識速度、認識性能の双方を向上させることを試みる。

2. ネットワーク構造

情報検索のために用いられているデータ構造に、Trie構造 [2] がある。Trie構造は、対象となるキーワード集合の共通の接頭辞を併合した探索木であり、文字単位の検索を行うことから高速な検索が可能となる。また、あるキーワードの探索中にそのすべての接頭辞が検索できる。この Trie 構造の共通接尾辞を併合すると、よりコンパクトなDAWG(directed acyclic word-graph)構造 [3] が得られる。図1は、あるキーワード集合Kを表現するTrie構造とDAWG構造の例である。我々は、青江らによるDAWG構造を動的に自動生成する手法 [4] を用いて、単語辞書を作成することにした。この手法では、辞書への単語の追加や削除を一単語単位で高速に行うことができるため、音声認識の単語辞書の構成法に適している。

我々の対象とするタスクは電話番号案内システムである。このタスクは認識対象の住所氏名として5つのキーワードクラスを持つ。認識対象となる文章は、これらのキーワード、及びその間に入るノンキーワードの組合せから成る。ノンキーワードは、不要語や冗長語、慣用表現などを受け入れるためのクラスである。本システムの文法は、図2のようにメインとサブの2つの部分より成る。メインは、サブ間の関係を「意味」に基づいて制御する。ここで「意味」とはキーワードの組合せのことで、例えば {市、氏名} {町、番地、氏名} {町、氏名} などである。今回、メインの文法はTrie構造によって記述されている。サブの文法は、キーワードとノンキーワードをそれぞれ扱う。これらは、DAWG構造によって記述されている。



K={a, ade, adf, bc, bcde, bcdf, ghe, ghf}
◎ terminal node

図1. Trie構造とDAWG構造の例

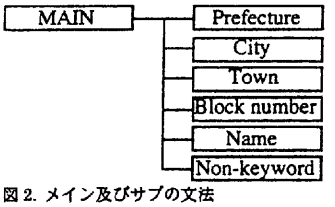


図2. メイン及びサブの文法

3. 音声認識システム

DAWG構造を使ってN-bestの仮説を得るためには、図2の枝c、d、またはfの後の節のように、ネットワークが数多くの節において併合されており、探索時に単語を一意に決定できないために、新しい探索アルゴリズムが必要となる。

この探索アルゴリズムは、前向き探索とトレースバックから成る。前向き探索は、基本的には通常のN-bestにおける単語グラフを生成する方法と同じであるが、我々の手法では単語グラフのかわりに音素グラフを生成する。トレースバックは、この音素グラフを探索してN-bestの仮説を生成する。

3.1. 前向き探索

認識システムは、時間同期の Viterbi アルゴリズムを用いてDAWG ネットワークの中を探索し、音素グラフを生成する。また、探索中には枝刈りも行われている(beam-search)。

探索プロセスにおいて、複数の候補が同時刻に同じ節に到達し、しかもそれらの持っている意味が同じだった時、最も高いスコアを持つ候補のみが以降の処理を続け、その節にはその候補が到達したという履歴情報が残される。他の候補については、履歴情報を残すだけでそれ以降の処理は行わない。この処理をマージと呼ぶ。

このことにより、探索コストを削減できる。各候補はそこまでのスコアだけを持ち、音素系列を持たないので、Viterbi デコードにおけるメモリ資源も削減される。各節に残された履歴情報は、その候補が到達した時の時刻、スコア、その直前の節、及び持っていた意味である。これら履歴情報は、スコアの降順にソートされて残る。すなわち、このアルゴリズムはN-best アルゴリズムを単語の内部まで適用したものである。

---

* An efficient search method for continuous speech recognition using network structure.
By Ken Hanazawa, Yasuhiro Minami, and Sadaoki Furui
(Tokyo Institute of Technology, NTT Human Interface Laboratories)

マージは、メインにおいても行われる。同時刻に同じ節に到達した複数の候補はマージされる。これらの候補は、キーワードのクラスさえ同じであれば、そのキーワードが何であるかに関わらずにマージされる。

3.2. トレースバック

前向き探索終了後は、最大のスコアのみが得られる。トレースバックは、音素グラフを逆向き(後向き)にトレースして、すべての可能な仮説を生成する。そして上位 N 個の仮説を N-best の仮説として出力する。

トレースバックに必要な情報は、いつ、どこから、どの音味を持ってある候補がその節に到達したかである。これらの情報はすべて前向き探索時に音素グラフの各節に履歴情報として残され、そのスコアによってソートされている。ここで、すべての仮説を無駄に生成しないために次の手法を提案する。まず、N 個の仮説のためのバッファを用意する。音素グラフは入力文章の終端からトレースされ、降順に残された履歴情報をもとにして各節で候補が作られていく。生成された仮説はスコアで降順にバッファに加えられることになる。もしある仮説のスコアがバッファの N 番目の仮説のスコアより低かった場合は、その仮説及びそこから生成され得るすべての仮説は棄却される。なぜなら降順に履歴情報が残っているため、そこから生成される仮説のスコアがバッファの N 番目の仮説のものより高くなることはないからである。この処理は、すべての節について再帰的に行われる。

同じキーワード系列を持つ複数の仮説が生成された場合は、最もスコアの高いもののみがバッファに残される。

4. 音声認識実験

まず実験条件について説明する。単語辞書は 4776 単語のキーワードを持つ。メインの文法において、各キーワードはその順序に従って最大一回の出現しか許されていない。また、氏名無しでは電話番号の検索ができないため、氏名は一文章に必ず一回だけ出現を許すようにする。これらの文法に沿わない文章を、ここでは Out-Of-Syntax と考える。

評価尺度としては、すべてのキーワードが正しく認識されたときに、その仮説を正解とする。

4.1. 学習及び評価データ

音素 HMM としては、4 状態 3 ループ 4 混合のガウス分布型で、文脈依存モデルを用いる。特徴量は、それぞれ 16 次元のケプストラム、△ケプストラム、及び 1 次元の△パワーである。音素 HMM は、音素バランス文を男性 34 名女性 30 名の話者により発声した ASJ の不特定話者データベース 9600文章を用い、連結学習によって学習を行っている。

評価セットは 362 文章で、男性 12 名女性 8 名の話者の発声である。すべての発声は、マルチモーダル電話番号案内システム [5] によって収録された。

評価セットの発声文章は、システムに受理されるべきいくつかのノンキーワードを含むことがあるが、未知語や繰り返し、言い直しは含まれていない。

4.2. 認識実験結果

4.2.1. 状態数の比較

表 1 に、木構造の一種である Trie 構造と DAWG 構造の状態数を示す。各キーワードクラスはそれぞれ単一のネットワークによって表されている。時間同期の認識システムでは、ネットワークの状態数が探索空間の大きさを表していると言えるので、表 1 より DAWG 構造を採用することで探索空間が大幅に削減できることがわかる。

今回ノンキーワードの単語辞書を作成するにあたり、Trie 構造では状態数が爆発してしまったのに対し、DAWG 構造では 169 状態しか必要とならなかった。この理由は、日本語の文章は文末に様々な表現が可能で、その部分において木構造の枝分かれが非常に多くなってしまうためと考えられる。この問題は、Viterbi デコーディングのプロセスでスタックを用いて探索を行うことによっても回避できるが、そのためには余分なメモリが必要となってしまう。

表 1. 状態数の比較

ネットワーク	Trie 構造	DAWG 構造
都	18	7
市	38	17
町	665	115
番地	36,987	602
氏名	82,848	7,824
ノンキーワード	968,639	169

4.2.2. 従来のシステムとの比較

本手法と我々の従来手法 [1] を比較する。従来手法では、音素同期のトレリスサーチアルゴリズムを用いている。このアルゴリズムでは、CFG を解析する一般化 LR 文法が音素予測の言語モデルとして使われる。この LR パーザは複数の候補が同じ文法的意味を持っていたときにマージを行っている。

表 2 に、従来手法と提案手法の認識性能の比較を示す。単語正解率は、キーワードのみを考慮している。認識時間は、一文あたりにかかる平均認識処理時間である。計算には HP9000を使用している。提案手法で、候補のマージ及びソートにかかる時間は CPU 時間の 1%程度であった。表 2 より、提案手法を用いることで約 20~60%程度の誤り率削減が実現し、認識時間も約 6 倍向上したことが確認できる。

表 2. 認識性能の比較

	文正解率		単語正解率	認識時間
手法	top	top 5		(秒)
従来手法	83.2	91.5	92.8	334
提案手法	86.2	96.7	94.4	56

Trie 構造を用いた認識システムは、その状態数が爆発してしまうために実現できなかった。しかし、時間同期の音声認識システムでは状態数の比較がおおよそ探索空間の大きさの比較になるため、状態数の比較から、DAWG 構造を用いた場合に比べ Trie 構造を用いた場合は認識時間が大幅に増加してしまうことが予想される。

5. むすび

本論文では、大語彙連続音声認識のためのコンパクトなデータ構造(DAWG)を導入した。また、このデータ構造の利点を生かした、N-best の仮説を出力する効率的な探索手法を提案した。実験結果から、本手法が従来手法と比較して認識性能、認識速度ともに優れていることが確認できた。

本手法は、大語彙連続音声認識システムの一般的なタスクに対しても適用可能であると考える。

謝辞

NTT ヒューマンインタフェース研究所古井特別研究室の皆様、東工大古井研究室の皆様に感謝します。

参考文献

[1] Y. Minami, et al, "Large vocabulary continuous speech recognition algorithm applied to a multi-modal telephone directory assistance system", *Trans. Speech Communication* 15, 1995,pp. 301-310.

[2] E. Fredkin, "Trie memory", *Commun. ACM*, 3, 9, Sept. 1960,pp. 490-550.

[3] A. W. Appel, and G. J. Jacobson, "The world's fastest scrabble program", *Commun. ACM*. 31, 5, May 1988, pp. 572-578.

[4] 青江順一, 森本勝士, 長谷美紀, "トライ構造における共通接尾辞の圧縮アルゴリズム", 電子情報通信学会論文誌, D-II, Vol. J75-D-II, No. 4, Apr. 1992,pp. 770-779.

[5] O. Yoshioka, Y. Minami, and K. Shikano, "A multi-modal dialogue system for telephone directory assistance", *Proc. ICSLP94*, Sept. 1994,pp. 887-890.