音声言語情報処理 30−11
(2000. 2. 5)

マルチパス探索における第2パス探索法

小川 厚徳　野田 喜昭　松永 昭一
NTTサイバースペース研究所
〒239−0847 神奈川県横須賀市光の丘1−1
{ogawa,noda,mat}@nttspch.hil.ntt.co.jp

あらまし 大語彙ディクテーションにおける探索法としては，膨大な数の文候補の中から高速かつ高精度に解(認識結果文)を得るために，段階的に探索の精度を上げて文候補を絞り込んでいくマルチパス探索が採用されることが多い．本稿では，第1パス探索の出力(中間表現)として単語ラティスの形式をとるマルチパス探索における第2パス探索法について検討した．検討した探索法は，単語ラティス上に記憶されている第1パス探索のスコアをヒューリスティックとして用いる時間非同期ビーム探索である．2万語彙の放送ニュース音声タスクによる評価実験の結果，今回検討した時間非同期ビーム探索では，従来の代表的な第2パス探索法であるN-bestリスコアリング，A*よりも高速かつ高精度に解を得ることができた．

キーワード マルチパス探索，第2パス探索，時間非同期ビーム探索，ヒューリスティック．

A Second-Pass Search Algorithm for Multi-Pass Speech Recognition Strategy

Atsunori Ogawa, Yoshiaki Noda and Shoichi Matsunaga
NTT Cyber Space Laboratories.
1-1 Hikarinooka Yokosuka-Shi Kanagawa 239-0847 Japan.
{ogawa,noda,mat}@nttspch.hil.ntt.co.jp

Abstract One of the most common search strategies in large vocabulary continuous speech recognition is the multi-pass search strategy which employs more accurate models in a later search stage for fast recognition. In this paper, we propose a time-asynchronous beam search algorithm for the second-pass search in the multi-pass search strategy. This algorithm works on a word lattice which the first-pass search generates and uses heuristics that are scores on word lattice nodes in the first-pass search. A Japanese broadcast news speech recognition experiment for a 20k vocabulary shows that the time-asynchronous beam search algorithm is more accurate and faster than other second-pass search algorithms.

Keywords multi-pass search, second-pass search, time-asynchronous beam search, heuristics.

1. はじめに

近年，音声ワープロや放送ニュース音声自動字幕作成システムなどの実現を目指した大語彙ディクテーションの研究がますます盛んになり，音響モデル，言語モデル，探索法のそれぞれの面において，様々な検討が進められている[1].

大語彙ディクテーションにおける探索法としては，膨大な数の文候補の中から高速かつ高精度に解(認識結果文)を得るため，段階的に探索の精度を上げて文候補を絞り込んでいくマルチパス探索が採用されることが多い．マルチパス探索には，第1パス探索法，第1パス探索の出力の中間表現，第2パス探索法の組み合わせにより様々な形式がある[2][3][4][5][6][7]．我々もマルチパス探索による大語彙ディクテーションの検討を行っている[8].

本稿では，中間表現として単語ラティスの形式をとるマルチパス探索における第2パス探索法について検討した.

従来の代表的な単語ラティス上での第2パス探索法としては，N-best リスコアリングや A*探索が挙げられる．文献[9]では，高次言語モデルによるN-bestリスコアリングを行っている．N-bestリスコアリングは実装が単純で確実に解を得ることができる．しかし，より精度を上げるための音響スコアの再計算は処理量が増えるため不向きである．A*探索では，単語ラティス上に記憶されている第1パス探索スコアをヒューリスティックとして利用することで効率的な探索を期待できるが，逆にヒューリスティックの精度が悪い場合には探索が困難になる等の問題がある.

これらに対し，今回検討した第2パス探索法は，A*探索と同様に，単語ラティス上に記憶されている第1パス探索のスコアをヒューリスティックとして用いる時間非同期ビーム探索である[10][11].

今回検討した時間非同期ビーム探索を，NTTで開発した音声認識エンジン VoiceRex[12]上で実装し，2万語彙の放送ニュース音声タスクで評価した結果，N-best リスコアリング，A*探索よりも高速かつ高精度に解を得ることができた.

2. マルチパス探索

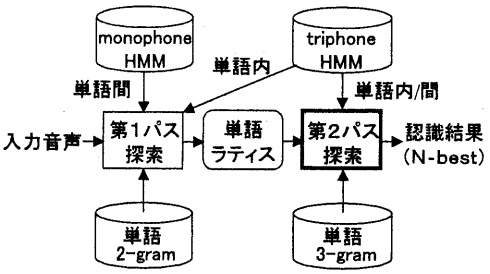

図1．マルチパス探索

本稿では，図1に示すようなマルチパス探索を考える.

第1パス探索では，粗いモデル(単語間 monophone HMM，単語内 triphone HMM，単語 2-gram)を用いて高速に文候補の絞り込みを行い，単語ラティスを出力する．単語ラティス上には，第1パス探索のスコアと単語境界時刻が記憶される.

第2パス探索では，第1パス探索の出力として得られた単語ラティス上を，高精度のモデル(単語内/間 triphone HMM，単語 3-gram)を用いて再計算を行う形で探索し，最終的に N-best 解を得る．本稿では，単語ラティス上での第2パス探索法を検討した.

3. 従来の第2パス探索法

従来の代表的な単語ラティス上での第2パス探索法としては，N-best リスコアリングと A*探索が挙げられる.

3.1. N-best リスコアリング

N-best リスコアリングでは，粗いモデルによる探索で得られた N-best 文候補のスコアを高精度のモデルによるスコアで置き換えて，候補の順序を入れ換える.

N-best リスコアリングを第2パス探索に用いる場合は，第1パス探索のスコアを基に単語ラティスから N-best 文候補を作成し，単語 2-gram スコアを単語 3-gram スコアで置き換えて，候補の順序を入れ換える．N-best リスコアリングは，実装が単純であり，確実に解を得ることができる.

しかし，単語ラティスから N-best 文候補を作成する際に，1単語のみ異なるような類似文候補が多数出現するため，十分な認識精度を得るには

非常に多くの文候補を対象にリスコアリングを行う必要がある．また，より精度を上げるための音響スコアの再計算も行えるが，類似文候補の重複した単語についても計算の共有ができず，それぞれに計算を行わなければならないため，効率的ではない．

3.2. A*探索

A*探索[13]では，(1)式で定義される時刻（フレーム番号）t における全区間推定スコア $\hat{f}_n(t)$ が最も高い仮説 n から優先的に展開を行う（best-first 探索）．

$$\hat{f}_n(t) = g_n(t) + \hat{h}_n(t) \qquad (1)$$

ここで，$g_n(t)$ は既に探索を終えた区間のスコア，$\hat{h}_n(t)$ は未探索区間の推定スコア（ヒューリスティック）である．A*探索で第1位に（最も早く探索が終わる解として）最適解を得るためには，$\hat{h}_n(t)$ の値がその真値 $h_n(t)$ よりも大きくなければならない（A*実行可能性）．さらに，$\hat{h}_n(t)$ が $h_n(t)$ に近いほど効率の高い探索が可能である．

A*探索を第2パス探索に用いる場合は，単語ラティス上を第1パス探索とは逆向きに文末から文頭に向かって単語単位の仮説展開を行う（図2）．$g_n(t)$ は単語内/間 triphone HMM と単語 3-gram を用いて再計算する第2パス探索スコアで，$\hat{h}_n(t)$ には単語ラティス上に記憶されている第1パス探索スコアを用いる．

しかし，第1パス探索と第2パス探索では用いるモデルが異なるため，$\hat{h}_n(t)$ が A*実行可能性を満たすとは限らない．このため，第1位に最適解が得られるとは限らず，現実的には，第N位までの解を探索結果スコア $g_n(0)$ でソートしながら待ち，N-best 解を得る必要がある．また，精度の良い $\hat{h}_n(t)$ が得られるとも限らない．このため，入力音声によっては，膨大な数の仮説を展開することになり，探索が前に進まなくなる場合がある．これを回避するため，現実的には，何らかの枝刈りを導入する必要がある[14]．

4. 時間非同期ビーム探索

今回は第2パス探索法として，A*探索と同様に，単語ラティス上に記憶されている第1パス探索スコアをヒューリスティックとして利用する時間非同期ビーム探索を検討した．

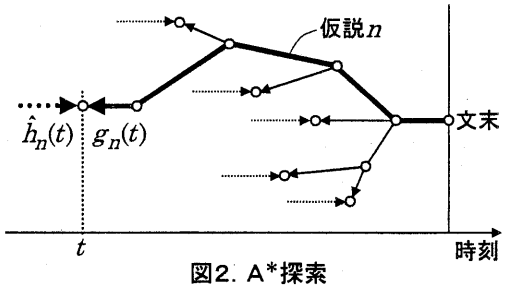

図2. A*探索

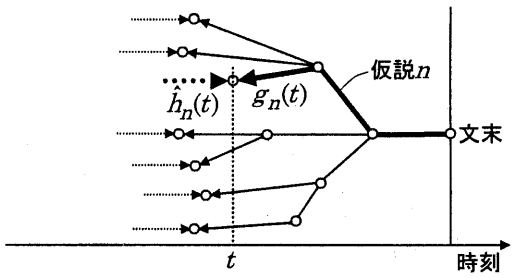

図3. 時間非同期ビーム探索

4.1. 基本アルゴリズム

今回検討した時間非同期ビーム探索では，A*探索と同様に(1)式を仮説のスコアとして用いながら，単語単位の breadth-first な仮説展開を行う（図3）．このとき，仮説数が一定となるように，(1)式の全区間推定スコア $\hat{f}_n(t)$ を基にして，仮説数の制限による枝刈りを行う．

時間非同期ビーム探索においても，A*探索と同様に，$\hat{h}_n(t)$ がその真値 $h_n(t)$ に近いほど効率の高い探索が可能である．その一方で，A*探索のように最適解を得られる保証はないが，必ずしも $\hat{h}_n(t)$ を $h_n(t)$ よりも大きく設定する必要がないため，(2)式のように $\hat{h}_n(t)$ に適当な重み α をかけて実験的に調整することも可能である．

$$\hat{f}_n(t) = g_n(t) + \alpha \cdot \hat{h}_n(t) \qquad (2)$$

時間非同期ビーム探索では，探索は展開すべき仮説がなくなるまで行う．このとき，探索結果スコア $g_n(0)$ を基に上位N個の解をソートしながら保持し，探索終了後，これらを N-best 解とする．

4.2. スコアに基づく枝刈り

第2パス探索では，単語内/間 triphone HMM と単語 3-gram を用いてスコアの再計算を行う．こ

のときのViterbi計算のコストが高いため，時間同期ビーム探索と同様に，時刻が同一の状態仮説に対してスコアに基づく枝刈りを行う．

具体的には，図4に示すように，時刻(フレーム番号) t ごとに第2パス探索スコア $g_n(t)$ の最高値を逐次更新しながら記憶しておき，その最高値の包絡から一定のスコア幅内に入らない HMM 状態の Viterbi 計算は行わない．

精度の高い枝刈りを行うには，$g_n(t)$ の最高値の包絡をより多くの仮説から求める必要がある．このため，時間非同期ビーム探索における仮説展開を，探索が最も遅れている仮説から優先的に行い(shortest-first 探索)，各仮説の時間的な長さがなるべく揃うようにする．

4.3. 単語境界時刻の幅の考慮

第1パス探索では，単語間で monophone HMMを用いるため，単語間音素環境を考慮しないのに対し，第2パス探索では，単語間でも単語内と同様にtriphone HMMを用いて単語間音素環境を考慮する．

このため，図5Aのように，第2パス探索の最適パスが，単語ラティス上に記憶されている第1パス探索の最適パスの単語境界時刻(t_1, t_2)と同時刻に単語境界を通るように拘束すると，第2パス探索スコアとして本来のスコアよりも低いスコアが得られる可能性がある．

そこで，図5Bのように，第2パス探索の最適パスは，第1パス探索の最適パスの単語境界時刻に前後数フレーム分の幅を持たせた範囲のどの時刻においても単語境界を通ってよいことにした．これにより，より高精度の第2パス探索を行えると期待できる．

5. 評価実験

今回検討した時間非同期ビーム探索と比較対象であるN-bestリスコアリング，A*探索を音声認識エンジン VoiceRex 上で実装し，2万語彙の放送ニュース音声タスクで比較評価した．

5.1. 実験条件

音響モデルは，ニュース番組1ヶ月分から選んだ6700文を学習データとする総状態数120，混

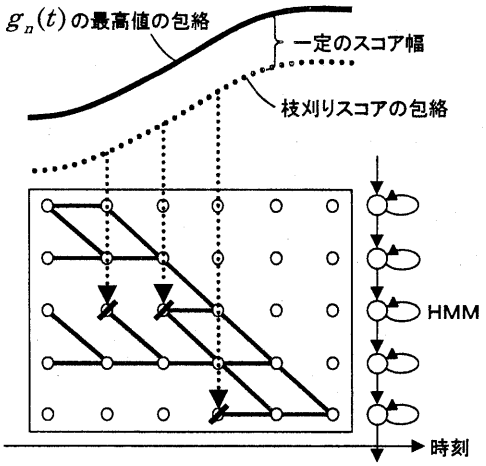

図4. スコアに基づく枝刈り

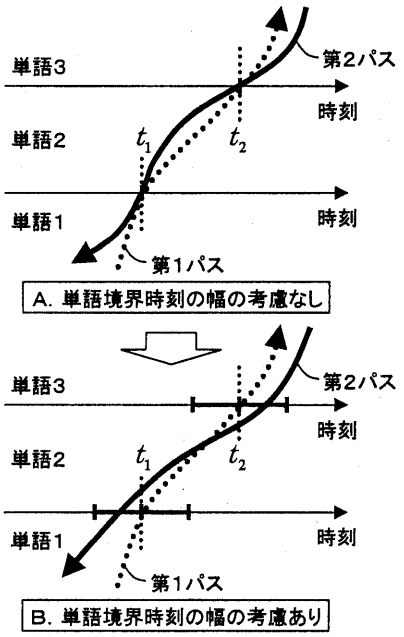

図5. 単語境界時刻の幅の考慮

合数16の monophone HMMと総状態数2000，混合数8の triphone HMM である[15]．サンプリング周波数16kHz，フレーム長30msec，フレーム周期10msec，特徴量は，MFCC12次元とその1次，2次回帰係数，対数パワーとその1次，2次回帰係数の計39次元である．

言語モデルは，ニュース番組原稿4年分の50万文と，1ヶ月分のニュース番組音声の書き起こし

で学習された単語 2-gram と単語 3-gram である.

評価セットはニュース番組5日分から50文(総単語数1800, 平均発声長12. 05秒)を選択した. 言語モデルの学習データの1ヶ月分は, 評価セットを収集した時期から直前の1ヶ月である. 評価セットの perplexity は, 単語 2-gram で105. 8, 単語 3-gram で57. 0であった.

実験に使用した計算機は, Sun Ultra Enterprise 450 (UltraSPARC-II 296MHz) である.

なお, 第1パス探索の最適解の単語正解精度は90. 49%であった.

5.2. 実験結果

5.2.1. 枝刈りの効果

まず, 時間非同期ビーム探索における仮説数の制限による枝刈りとスコアに基づく枝刈りの効果を調査した(ここでは, 4. 3節で述べた単語境界時刻の幅は考慮していない). 2つの枝刈りの条件をいくつか設定し, 1文あたりの平均第2パス探索時間(sec)と単語正解精度(%)を求めた.

結果を表1に示す. 表1より, 2つの枝刈りがそれぞれ効果的に導入されていることが分かる.

5.2.2. 単語境界時刻の幅を考慮する効果

次に, 時間非同期ビーム探索において, 4. 3節で述べた単語境界時刻の幅を考慮する効果を調査した. 前節の結果より, 枝刈りの条件を, 仮説数250, スコアに基づく枝刈りありと設定し, 考慮する単語境界時刻の幅を前後0～5フレーム(0～50msec)と変化させた.

結果を表2に示す. 表2より, 単語境界時刻の幅を考慮することで, 認識精度が改善されることが分かる. 本実験では, 2フレームの幅を考慮することで最も精度が改善された. また, 幅を考慮すれば, 探索の処理量が増加し, 平均第2パス探索時間も増加すると予想したが, 結果は幅を考慮しない場合とほとんど変わらなかった. これは, 幅を考慮することで, より精度の高い探索が可能となったためであると考えられる.

5.2.3. 比較評価

表1. 枝刈りの効果

仮説数	スコア	第2パス探索時間(sec)	単語正解精度(%)
1000	なし	19. 10	93. 29
1000	あり	5. 00	93. 34
250	なし	4. 02	93. 06
250	あり	1. 78	93. 34

表2. 単語境界時刻の幅を考慮する効果

考慮する幅(フレーム)	第2パス探索時間(sec)	単語正解精度(%)
0	1. 78	93. 34
1	1. 84	93. 90
2	1. 77	93. 96
3	1. 79	93. 85
4	1. 95	93. 85
5	1. 89	93. 57

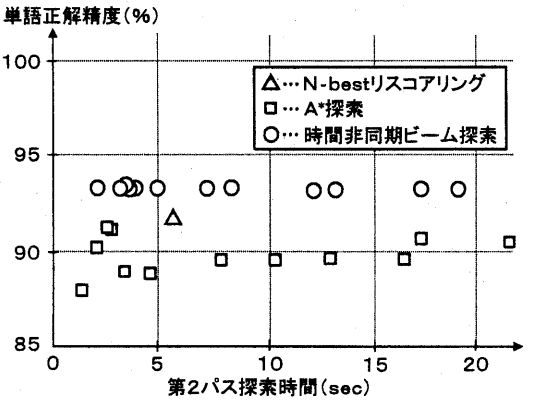

図6. 比較評価

表3. 最高性能の比較

第2パス探索法	第2パス探索時間(sec)	単語正解精度(%)
N−best	5. 37	92. 45
A*探索	2. 22	92. 11
ビーム	1. 77	93. 96

最後に, 時間非同期ビーム探索と N-best リスコアリング, A*探索の比較実験を行った. N-best リスコアリングでは, 予備実験の結果より, 300文候補のリスコアリングを行った. また, 3. 2節で述べたように, 純粋な A*探索では, 入力音声によっては大幅に探索時間がかかる場合が生じたので, 時間非同期ビーム探索と同様の2つの枝刈りを導入した. さらにヒューリスティック $\hat{h}_n(t)$ が A*実行可能性を満たさない場合を考慮し, 第10位まで

の解を待ち，最高スコアの解を求めた．A*探索，時間非同期ビーム探索ともに，5．2．2節の結果から2フレームの単語境界時刻の幅を考慮した．また，時間非同期ビーム探索においては，予備実験より，(2)式のヒューリスティック $\hat{h}_n(t)$ の重み α を0．99に設定した．枝刈りの条件は，表1に示した他にもいくつかの条件を設定した．

横軸に1文あたりの平均第2パス探索時間(sec)，縦軸に単語正解精度(%)をとった結果を図6に示す．図6より，時間非同期ビーム探索ではN-best リスコアリング，A*探索よりも高速かつ高精度に解を得られることが分かる．表3に，3つの第2パス探索法の最高性能の比較を示す．

N-best リスコアリングでは，リスコアリングの前に単語ラティスから300ベストの文候補を作成するのに時間がかかり，単語間の音素環境を考慮した音響スコアの再計算を行わないために，認識精度が低くなったと考えられる．

A*探索で認識精度が低いのは，ヒューリスティック $\hat{h}_n(t)$ が A*実行可能性を満たしていない場合に，最適解が枝刈りされた，または，第10位までに最適解が入らなかったためと考えられる．

6. まとめ

本稿では，単語ラティスを中間表現とするマルチパス探索における第2パス探索法として，単語ラティス上に記憶された第1パス探索のスコアをヒューリスティックとして用いる時間非同期ビーム探索を検討した．

時間非同期ビーム探索では，仮説数の制限による枝刈りとスコアに基づく枝刈りを効果的に導入することができる．特にスコアに基づく枝刈りを効果的に導入するため，shortest-first な仮説展開を行った．また，第1パス探索と第2パス探索で用いる音響モデルの違いを考慮し，単語境界時刻の幅を考慮したより精度の高い第2パス探索を検討した．

2万語彙の放送ニュース音声タスクによる評価実験の結果，時間非同期ビーム探索では，従来の代表的な第2パス探索法であるN-bestリスコアリング，A*探索よりも高速かつ高精度に解を得ることができた．

謝辞

ニュース原稿とニュース音声を提供して頂いた日本放送協会に感謝します．

音響モデル，言語モデルを提供して頂いた NTT サイバースペース研究所メディア処理プロジェクト音声認識グループ，山口義和氏，大附克年氏，堀貴明氏，中川聡氏に感謝します．

参考文献

[1]古井：“大語彙連続音声認識の現状と展望”，音学講論，1-6-10，1998．3．

[2]R.Schwartz, et al.：“Multiple-pass Search Strategies”，in Automatic Speech and Speaker Recognition Advanced Topics, Kluwer Academic Publishers, 1996.

[3]S.Ortmanns, et al.：“A word graph algorithm for large vocabulary continuous speech recognition”, Computer Speech and Language, Vol.11, No.1, 1997.

[4]S.Austin, et al.：“The forward-backward search algorithm”，Proc. ICASSP, Vol.1. 1991.

[5]E.-F.Huang, et al.：“The use of tree-trellis search for large-vocabulary Mandarin polysyllabic word speech recognition”，Computer Speech and Language, Vol.8, No.1, 1994.

[6]Z.Li, et al.：“Bi-directional graph search strategies for speech recognition”，Computer Speech and Language, Vol.10, No.4, 1996.

[7]T.-H.Ho, et al.：“Improved search strategy for large vocabulary continuous Mandarin speech recognition”, Proc. ICASSP, Vol.2, 1998.

[8]野田 他：“単語グラフを用いた大語彙連続音声認識における近似演算手法の検討”，信学技報，SP96-12，1997．1．

[9]今井 他：“ニュース音声認識用デコーダーの開発”，音学講論，3-1-12，1998．9．

[10]野田 他：“前向きヒューリスティック関数を用いたビーム探索による HMM-LR 連続音声認識”，信学論(D-II)，Vol．J79-D-II，No．8，1996．

[11]小川 他：“マルチパス探索における第2パス探索法の検討”，音学講論，2-1-3，1999．9．

[12]野田 他：“音声認識エンジン VoiceRex の開発”，音学講論，2-1-19，1999．9．

[13]N.J.Nilsson 著，合田 他訳：人工知能－問題解決のシステム論，コロナ社，1973．

[14]李 他：“大語彙連続音声認識エンジンJuliusにおける A*探索法の改善”，情処研報，SLP27-5，1999．7．

[15]山口 他：“音声認識エンジン VoiceRex によるニュース放送音声認識”，音学講論，2-1-20，1999．9．