音声言語情報処理 29－37
(1999. 12. 21)

逐次2パスデコーダを用いたニュース音声認識システム

今井 亨、小林彰夫、佐藤庄衛、安藤彰男

ＮＨＫ放送技術研究所

〒157-8510 東京都世田谷区砧 1-10-11

{imai, akio, shoe, ando}@strl.nhk.or.jp

あらまし 放送ニュースの自動字幕化のための音声認識システムでは、少ない遅れ時間で認識結果を確定することが望ましいが、従来のマルチパスデコーダは文末でしか結果を確定できないという問題点がある。そこで本稿では、発話中に結果を早期確定する逐次2パスデコーダを提案する。逐次2パスデコーダは第1パス実行中に第2パスを繰り返し実行し、安定な最尤単語列を検出して早期確定する。この手法は原理的に最適性を保証するものではないが、単語認識誤りを有意に増やすことなく、単語の早期確定を実現する。放送ニュースの音声認識実験において、逐次2パスデコーダは単語正解精度の劣化を0.22%に抑えて単語確定平均遅れ時間を554msecに短縮し、その有効性が確かめられた。

キーワード 音声認識、放送ニュース、探索、早期確定

Broadcast News Transcription System with a Progressive 2-Pass Decoder

Toru Imai, Akio Kobayashi, Shoei Sato, and Akio Ando

NHK (Japan Broadcasting Corp.) Sci. & Tech. Res. Labs.

1-10-11 Kinuta, Setagaya, Tokyo 157-8510 JAPAN

{imai, akio, shoe, ando}@strl.nhk.or.jp

Abstract This paper describes a broadcast news transcription system with a 2-pass decoder which progressively outputs the latest available results. The decoder practically eliminates the disadvantage of multiple-pass decoders which delay a decision until the end of a sentence. During the first pass of search the proposed decoder periodically executes the second pass and detects most likely and stable words. This method is not theoretically optimal but makes a quick response with a negligible increase in word errors. In a recognition experiment on Japanese broadcast news, the decoder worked with an average decision delay of 554msec for each word and degraded word accuracy only by 0.22%.

key words speech recognition, broadcast news, search, early decision

1. はじめに

聴覚障害者や高齢者向けの字幕放送番組は、ドラマや情報番組などを中心に、着実に増加しつつある[1][2]。しかし、日本ではニュース番組の字幕化はまだ実現していない。英語の場合は速記用キーボードによって生放送番組にも字幕付与が可能だが[3]、日本語は仮名漢字変換や同音異義語の選択が必要なため、キーボードでのリアルタイム字幕作成は困難である。ニュースの場合、その約8割は元になる電子原稿が存在するものの[4]、アナウンサーが読み上げる原稿は放送直前まで手書きで加筆・修正されるため、電子原稿をそのまま字幕にすることはできない。そこで我々は、ニュース番組中のアナウンサーらの声を音声認識によって自動的に文字に変換し、最終的には人による確認・修正を経てリアルタイムに字幕を作成することを検討している。

ニュース音声の認識に関しては、欧米ではDARPAを中心としたプロジェクトが、実時間の10倍以内の処理量での検討を行っている [5]。しかし、生放送のニュース番組に字幕を付与する場合、発声から字幕表示までの遅れ時間はできる限り小さい方が望ましく、ほぼリアルタイムでの処理が求められる。したがって、音声認識を行うデコーダは、音声にほぼ同期して認識結果を出力できるように設計する必要がある。

音声認識手法の1つに、入力音声を2回以上のパスで探索するマルチパスデコーダがある[6]。マルチパスデコーダは一般に、第1パスで簡易なモデルを使って時間同期サーチを行い、単語候補を絞り込む。そして、発話終了後に詳細なモデルを使って第2パス以降の処理を行い、認識結果を確定することが多い。マルチパスデコーダは計算量を抑えつつ、比較的高精度な認識を行うことができるという利点があるが、認識結果の確定を発話終了時に行うという欠点を持つ。1パスのデコーダでも、過去の探索経路が唯一に絞られない限り[7]、やはり文末まで待たなければ最適な認識結果を確定することはできない。従来の我々のデコーダ[8]も文末で第2パスの処理を行っていたため、発話終了後にしか認識結果を確定できなかった。

そこで本稿では、音声入力中に逐次的に第2パスの処理を行うデコーダ「逐次2パスデコーダ」を提案し、認識結果の早期確定を試みる。逐次2パスデコーダによるニュース音声の認識実験を行い、認識率と単語確定までの遅れ時間について考察する。

2. 従来の2パスデコーダ

従来の我々の2パスデコーダ[8]は、第1パスでバイグラムとトライフォンHMMを使って時間同期Viterbiビームサーチを行い、単語ラティスを生成していく。そして、発話終了後に単語ラティスを再帰的にトレースバック[9]してN-best文を得る。最後に第2パスの処理として、N-best文をトライグラムでリスコアリングし、最適な単語列を確定する。

第1パスでは、単語依存N-best探索[9]を行い、HMMの各状態で直前単語が異なる最大n個（$n<<N$）のパスを残しながら探索を進める。直前単語の同じ複数のパスが同一状態に達した時は、最大スコアのパスを残す。単語終端では、現単語のIDと終端時刻、n個の各パスに対するスコアと直前単語終端へのバックポインタを保存し、単語ラティスを生成していく。次単語へは、現単語の最良スコアのパスだけを進める。ビームサーチの枝刈りには、対数尤度差に基づく大域的ビーム幅と、より狭い単語終端ビーム幅を併用する。

第1パスの探索は、語頭部分の音素を複数単語で共有する木構造[10]の音素ネットワーク上で行う。ただし、木構造音素ネットワークは動的にコピーするのではなく、1つの静的なものを使う[8]（ただし探索エラーを避けるためにN=1でもnは大きい方が望ましい）。木構造の音素ネットワークは、語頭部分のノード（トライフォン）を複数単語で共有するので、語頭でのアクティブなノード数を削減できるという利点がある。しかし、現単語を特定できるノードに処理が進むまで、スコアにバイグラムを適用することができない。そこで枝刈りには、ノードを共有する単語のうちで最大のバイグラムを使うことが広く行われている[11]。そのためには、次のノードに遷移した時、共有単語のリストが変化したなら、直前単語に依存した最大バイグラムを求めなければならず、実装時には処理量を削減する工夫が必要になる。語彙サイズが小さい場合には、あらかじめ全ノードで直前単語

ごとに最大バイグラムを計算しておき、テーブル化しておくことが可能だが、語彙サイズが大きくなるにつれて、必要とするメモリー量は増加する。そこで、語彙サイズに応じて、アクティブになる割合が高い語頭の音素Lレベル目までのノードのみ、事前に最大バイグラムを計算しておくことにする[8]。Lより大きいレベルのノードでは、デコード中で必要になった時に最大バイグラムを計算し、キャッシュに格納して再利用する。

従来のこの2パスデコーダでは、認識結果を文末で確定するので、ニュース番組にリアルタイムで字幕をつけるためには、認識結果確定までの遅れ時間が問題であった。

3. 逐次2パスデコーダ

第1パス実行中に最尤単語列を観察してみると、現フレームの近傍の最尤単語は比較的頻繁に入れ替わるものの、少し前の時点の最尤単語は変化が少ない。そこで、認識結果を早期確定するために、音声入力中に逐次的に第2パスの処理を行うことを試みる[12]。基本的な方針は、「音声入力中に適当な間隔で第2パスを実行し、安定したと考えられる単語列を認識結果として確定していく」というものである。逐次的に確定された認識結果の発話全体に対する最適性は保証されないが、実用上、認識率の劣化はわずかであると期待できる。

図1に従来の2パスデコーダと逐次2パスデコーダの動作の比較を示す。従来の2パスデコーダは第2パスを文末で一度だけ実行して認識結果を確定するのに対して、逐次2パスデコーダは発話中の第1パス実行中に、逐次的に第2パスを実行して、部分的に認識結果を確定していく。

提案する逐次2パスデコーダのアルゴリズムを以下で説明する。

(1) 第1パスのΔtフレームおきに、最高スコアの音素を出発点として単語ラティスをトレースバックし、N通りの単語列候補（N-best単語列）を得る。ただし、すでに確定された最終単語の終端フレーム以前での単語ラティスの分岐は考慮しなくてよい。

(2) N-best単語列をトライグラムでリスコアリングし、最良の単語列（1-best単語列）を得る。

(3) Δtフレーム前の1-best単語列と現1-best単語列を比較し、すでに確定された単語以降で一致する単語列を認識結果として確定する。一致する単語列がなければ、何も確定しない。ただし、現フレームの近傍の単語は不安定なので、文末以外では、現1-best単語列の終端M単語は確定対象としない。すでに単語が確定されている区間で現1-best単語列に変化が生じていても、それは無視する。

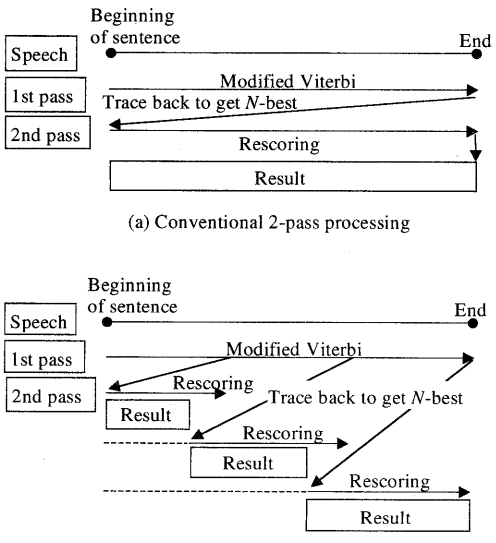

図1　2パスデコーダの動作比較

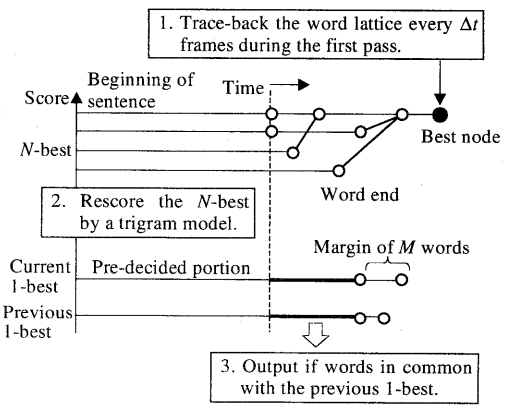

図2　逐次2パスデコーダ

逐次２パスデコーダでは、第１パスは従来通り実行され、第２パスや確定単語列からのフィードバックは一切受けない。第２パスが、部分的なN-best単語列に対して定期的に繰り返し実行される点が従来と異なる。認識結果を早期確定することによる認識率の劣化と確定までの遅れ時間はトレードオフの関係にあり、第２パスの実行間隔Δtと確定遅れ単語数Mで制御できる。Δtが長くなる、あるいはMが大きくなるほど認識率の劣化は少なくなるが、確定までの遅れ時間は長くなる。

逐次２パスの実現方法は他にも、トレースバックのタイミングを非一定間隔にしたり、現フレームの複数のアクティブな音素を考慮したり、既確定単語との接続を考慮するなど、より詳細なものも考えられる。また、単語の確定については、1-best単語列の継続性やN-best単語列の共通性を全フレームで観測して判断するという方法もある[13]。しかし、処理量の点から、本稿では上記のような簡易な方法による性能（認識率の劣化、確定までの遅れ時間）を検討することにする。

4. 実験

4.1. 条件

提案する逐次２パスデコーダを用いて、ニュース音声の認識実験を行う。評価音声はNHKのニュース（'98.9.30昼・夜７時放送）から収録された122文（男性話者２名65文、女性話者１名57文、計4,942単語、文平均長12.5秒）である。これらはスタジオのアナウンサーによって発声されており、背景雑音や音楽等は含まれていない。評価音声の音響分析条件は、標本化周波数16kHz、ハミング窓（窓幅25msec、シフト幅10msec）、39次パラメータ（12次MFCC、対数パワー、それぞれの１次および２次の回帰係数）とした。

音響モデルは、男女別状態共有化トライフォンHMM（８混合ガウス分布）で、NHKのニュース音声およびアナウンサーの発声したバランス文（トータルで男性話者48.3時間分、女性話者38.6時間分、評価話者を含む）で学習された。音素数は42で、ニュース用の２万単語の語彙に対して、論理HMMの数は5,873、物理HMMの数は男性話者用が3,093（状態数は1,950）、女性話者用が2,894（状態数は1,656）である。

言語モデルは、第１パス用のバイグラムと第２パス用のトライグラムを、NHKのニュース原稿約７年半分（'91.4.1～'98.9.29）から学習した。それぞれのカットオフは１と２で、Good-Turingのディスカウントを用いたバックオフスムージングを行った。

デコーダでは、HMMの各状態での最大保存パス数n=4、単語列候補数N=200、大域的ビーム幅160、単語終端ビーム幅110、言語スコア重み14、挿入ペナルティなし、最大バイグラムのテーブル化音素レベル数L=2とした。評価音声に対するテストセットパープレキシティは、バイグラムが69.2、トライグラムが28.0で、未知語率は0.67%である。

逐次２パスの実行間隔Δtを1～50フレーム（１フレームシフト幅は10msec）、確定遅れ単語数Mを0～2に変化させ、単語正解精度と確定までの遅れ時間（確定を決断したフレームと単語終端フレームの差）を調べた。実行マシンのCPUは、Alpha21264-500MHz（メモリ768MB）である。

4.2. 結果

従来の２パスデコーダによる音声認識では、単語正解精度は96.76%で、単語確定平均遅れ時間は7.2秒（最大37秒）と大きかった。一方、逐次２パスを実行したところ、図３～４に示すように、例えば逐次２パス実行間隔Δt=30フレーム、確定遅れ単語数M=1の時、単語正解精度は96.54%で従来法からの劣化は0.22%にしかすぎず、単語確定平均遅れ時間も554ms（最大2.7秒）と大幅に短縮された。単語を早期確定しながらも、有意な単語正解精度の劣化は見られなかったので、提案手法の有効性が確かめられたと言える。

逐次２パスデコーダにおける認識率と確定遅れ時間の関係を、パラメータの違いで見てみると、M=2の時は余裕をもって単語を確定するので、Δtによらず単語正解精度の劣化が小さいが、確定遅れ時間は長い。一方、M=0の時はΔtが小さいほど確定遅れ時間は短いが、単語正解精度の劣化が大きくなる。

全評価データに対するオフライン（A/D後のディスクからの音声入力）での処理時間は、従

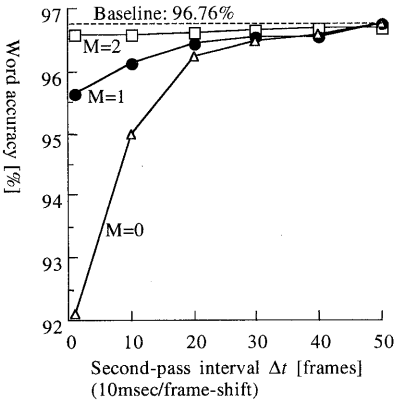

図3　単語正解精度

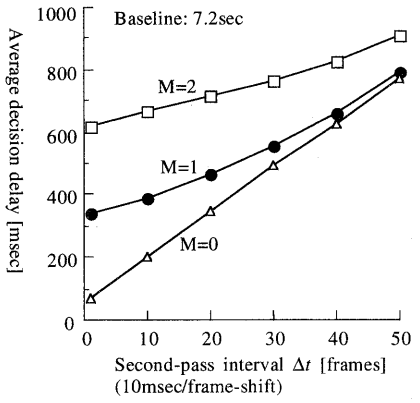

図4　確定遅れ時間

来の2パスデコーダは実時間の0.8倍であったが、逐次2パスデコーダはこれよりも約1割短かった。発話全体の単語ラティスに対して*N*-best文を得るよりも、Δtフレームおきに短い区間で*N*-best（実際はN=200よりもずっと小さい）単語列を得る方が、処理量は少ないことになる。認識精度の点からも、区間が短いほど、同じ*N*でより詳細な候補単語列が得られるという利点もある。

オンライン（A/Dしながらの音声入力）で逐次2パスデコーダ（Δt=30フレーム、M=1）を実行したところ、発話開始からほぼ1～2秒以内の遅れで次々と単語を確定することができ、ほぼリアルタイムでの動作が確認できた。

4.3. 保証された早期確定法との比較

動的計画法によって時間同期探索を進めると、直前最適ノードの選択や枝刈りによって、生き残ったすべてのパスが文頭の同一区間を共有することがある（図5）。この区間の単語は、文全体に対する最適なパスの一部でもあるので、文末を待たずに確定しても認識率は劣化しない。この原理は、アクティブな後続単語の数を各単語終端で管理することで、トレースバックよりも少ない処理量で実現でき、連続数字音声認識においてその有効性が報告されている[7]。

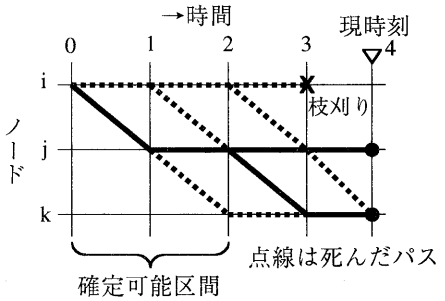

図5　認識率を劣化させない早期確定

認識率を劣化させずに単語を確定できるこの手法は魅力的であるが、現実には必ずしも常に過去の探索経路が唯一に絞られるわけではない。実際、この手法を我々の2パスデコーダに実装して同じニュース音声の認識実験を行ったところ、この手法によって早期確定できたのは全体の3.6%の単語（文頭の1～2単語）にしかすぎず、残りの単語は文末での通常のトレースバックによって得られた（もちろん認識率は劣化していないが）。この認識タスクでは、文全体に対する最適性を確保しようとすると、単語確定までの遅れ時間をかなり犠牲にしなければならないことがわかる。

そもそもマルチパス探索における第1パスは、後段での絞り込みを前提にして多くの単語候補を生成するのが目的であるから、早期確定可能な唯一の探索経路は存在しにくい。大きな語彙や広いビーム幅、多くの履歴の保存など、探索空間が広がるほど、原理的に正確に早期確定できる区間が存在する可能性は低くなる。また、この区間は候補が1通りであるから、第2パス以降でも結果は変わりようがなく、マルチパス

探索の意味がない。

ニュース音声認識のような大語彙でリアルタイム性が要求される応用においては、提案する逐次2パスデコーダのように認識率の劣化をできるだけ抑えつつ、早期に結果を確定する方が実用上は好ましいと考える。

5. まとめ

認識結果を早期確定するために、音声入力中に逐次的に第2パスの処理を行う逐次2パスデコーダを提案した。これは、従来のマルチパスデコーダが文末でしか結果を確定できなかったのに対して、音声入力中に逐次単語を確定していくという特徴を持つ。認識結果の文全体に対する最適性は保証されないが、逐次2パスの実行間隔と確定遅れ単語数によって、認識率の劣化と確定遅れ時間を制御することができる。逐次2パスデコーダによるニュース音声の認識実験を行ったところ、単語正解精度に有意な劣化はなく、ほぼリアルタイムで認識結果を確定することができ、その有効性が確認された。

参考文献

[1] Nippon Hoso Kyokai Annual Report 1999, NHK広報局 (1999.6).

[2] "視聴覚障害者向け放送の現状"、郵政省放送行政局長定例記者会見資料、http://www.mpt.go.jp/pressrelease/japanese/housou/991012j701.html (1999.10).

[3] "Frequently Asked Questions About Closed Captioning: Online (Live) Captioning," http://www.robson.org/capfaq/online.html

[4] 尾上和穂、今井亨、安藤彰男、"記者原稿を用いたニュース音声認識結果の修正法"、音講論集、1-6-6 (1998.3).

[5] D.S. Pallett, J.G. Fiscus, J.S. Garofolo, A. Martin, and M. Przybocki, "1998 Broadcast News Benchmark Test Results: English and Non-English Word Error Rate Performance Measures," Proc. DARPA Broadcast News Workshop, pp. 5-10 (1999.3).

[6] R. Schwartz, L. Nguyen, and J. Makhoul, "Multiple-Pass Search Strategies," Automatic Speech and Speaker Recognition Advanced Topics, Kluwer Academic Publishers, MA, pp. 429-456 (1996).

[7] P.F. Brown, J.C. Spohrer, and P.H. Hochschild, "Partial Traceback and Dynamic Programming," Proc. ICASSP-82, pp. 1629-1632 (1982.5).

[8] 今井亨、尾上和穂、小林彰夫、安藤彰男"ニュース音声認識用デコーダーの開発"、音講論集、3-1-12 (1998.9).

[9] R. Schwartz and S. Austin, "A Comparison of Several Approximate Algorithms for Finding Multiple (N-Best) Sentence Hypotheses," Proc. ICASSP-91, pp. 701-704 (1991.5).

[10] H. Ney, R. Haeb-Umbach, B.-H. Tran, M. Oerder, "Improvements in Beam Search for 10000-Word Continuous Speech Recognition," Proc. ICASSP-92, pp. 9-12 (1992.3).

[11] J.J. Odell, V. Valtchev, P.C. Woodland, and S.J. Young, "A One Pass Decoder Design for Large Vocabulary Recognition," *Proc. Human Language Technology Workshop*, pp. 405-410 (1994.3).

[12] 今井亨、小林彰夫、安藤彰男、"認識結果早期確定のための逐次2パスデコーダ"、音講論集、2-1-4 (1999.9).

[13] 中野裕一郎、小林哲則、"連続音声認識における部分単語列の早期検出"、音講論集、3-1-8 (1998.9).