音声言語情報処理 34－18
(2000. 12. 21)

端点検出を行わない連続音声認識手法

瀬川 修[†*] 武田 一哉[†] 板倉 文忠[‡]

†名古屋大学大学院工学研究科
‡名古屋大学情報メディア教育センター
*中部電力株式会社電力技術研究所

〒464-8603 名古屋市千種区不老町1
segawa@itakura.nuee.nagoya-u.ac.jp

あらまし 入力音声の明示的な端点検出を必要としない新しい連続音声認識手法を提案する。本手法では数秒程度の一定時間長の処理ブロックを認識すると同時に終端で途切れた単語区間をバックトラックによって修復しながら連続的にデコードを続けるため、端点検出や発話単位の考慮無しに無限長の入力音声を認識することが可能である。基本的なアルゴリズムは次のとおりである。1) 一定時間長の処理ブロックの認識を行う。2) 処理ブロックの終端フレームに残った全ての単語終端ノードよりトレースバックによって一つの最尤パスにマージするフレームを探し、これを最適な単語境界フレームとする。3) 単語境界フレームまで戻ってサーチを再開する。本稿ではまずアルゴリズムの詳細を説明し、提案手法の有効性を検証するために行った約10分の連続した新聞読み上げ音声および男女各1名による約30分の車内音声対話の自動書き起こし実験の結果を示す。

キーワード 連続音声認識, 端点検出, サーチアルゴリズム, 音声モニタリング

Continuous Speech Recognition without End-point Detection

Osamu SEGAWA[†*], Kazuya TAKEDA[†] and Fumitada ITAKURA[‡]

†Graduate School of Engineering, Nagoya University
‡Center for Information Media Studies, Nagoya University
*Chubu Electric Power Co., Inc. Electric Power R&D Center

Furo-cho 1, Chikusa-ku, Nagoya 464-8603 JAPAN
segawa@itakura.nuee.nagoya-u.ac.jp

Abstract A new continuous speech recognition method that does not need the explicit speech end-point detection is proposed. In this method, the decoder proceeds to recognize a processing block of a predetermined length and then to fix a word section which is broken at the end of a processing block. Therefore, continuous speech recognition of infinite length can be executed without the explicit end-point detection and without considering an utterance unit. The basic algorithm is 1) decode a processing block of the predetermined length, 2) traceback and find the boundaries of the processing blocks where the word history in the preceding processing block is merged into one, and 3) restart decoding from the boundary frame with the merged word history. The effectiveness of the method is verified by the two dictating experiments.

Key words Continuous speech recognition, End-point detection, Search algorithm, Speech monitoring

1 はじめに

従来の音声認識の枠組では発話単位の始端と終端の同定が前提となっており、音声の端点検出の精度が全体性能に大きな影響を与えている。これまで端点検出の方式として様々な手法が報告されている [1][2]。端点検出の明示的なアプローチとしてはパワーとゼロ交差数を用いて音声区間を同定する手法が一般的であるが、これらの手法は背景音の無い静かな環境下では比較的良好に動作するが、背景雑音や様々な外乱の影響に対し音声区間の検出誤りを起こすことが多い。一方、非明示的な手法としては端点検出と認識処理を統合したアプローチもいくつか試みられている [3][4][5]。この手法では音声区間の前後で非音声区間のマージンも含めて認識を行い、無音の継続時間長と尤度を用いて認識と同時に音声区間の検出も行っている。非明示的な手法は入力音声レベルの変動や文中の無音との混同に起因する誤検出が低減できるなど、外乱要因に対してある程度頑健であることが報告されているが [6]、現在の音声認識の枠組が有限時間長の "発話単位" というものを前提としているため、デコーダーの観点からは端点検出の問題は不可避なものである。本稿では、これらの明示的・非明示的な端点検出を前提としたアプローチに対し音声の端点検出を全く行わないことによって音声区間同定に関わる問題を回避し、入力音声における発話単位を考慮せずに無限長の連続音声認識が可能な方式を提案する。

以下、本稿では 2 節でアルゴリズムの詳細を説明する。3 節では新聞記事読み上げ音声を用いた基本性能の評価について述べ、4 節では高次言語モデルを用いたアルゴリズムの改良について述べる。5 節では実環境での評価として音声対話の自動書き起こしの実験について述べ、最後に 6 節でまとめを行う。

2 端点検出を行わない連続音声認識

本手法では、音声・非音声の区別なしに入力音声ストリームを数秒程度の一定時間長に機械的に分割した処理ブロック (セグメント) を逐次取り込みながら連続音声認識を行う。一つのセグメントの終端まで認識処理が終った時点で部分単語列を出力しながら認識を進めるため、一定間隔で逐次的に認識結果を出力しながら無限長の連続音声のデコードが可能である。発話の間に存在する無音、背景音などの非音声区間については音響モデルで用意された無音モデルによってデコードされるため、音声区間の同定を行わなくても連続音声認識を続けることができる。分割されるフレームの箇所は任意であるため、単語区間 (あるいは音素区間) の途中で途切れが起こる可能性が高い。このため、本手法ではセグメント単位の認識が終端まで達した時点で近傍の信頼性の高い単語境界のフレームを探索し、その時点から遡って再探索を行う。ここで単語境界推定の基本原理であるが、時間同期ビーム探索において終端までビーム内に残った全ての単語終端の状態からベストパスをトレースバックした部分文仮説を観察すると、最初は単語履歴が異なっているが途中から全ての仮説が一つの最尤パスと同じ履歴にマージしていることがわかる。そこで提案手法ではこれを利用して近傍の最も信頼性の高い単語境界を探索し、次のセグメントでは単語境界フレームまで戻ってサーチを再開する。

2.1 アルゴリズム

入力音声ストリームから一定時間長 T のセグメントを取り込み、始端から時間同期に One-Pass-Viterbi サーチを行う。サーチでは各フレームごとに下記の情報を保存する (単語トレリスインデックス [7])。

- ビーム内に終端が残った単語のインデックスと尤度
- 各単語終端に対応する始端フレーム

セグメントの終端フレームに達したところで以下の処理を行う。

1. 終端フレームでビーム内に残った全ての単語終端ノードより単語トレリス上でベストパスをバックトレースし、複数の部分文仮説を得る。

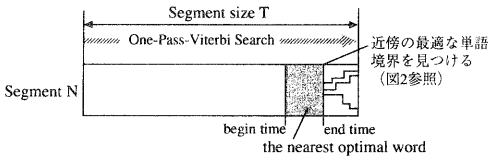

図 1: セグメント内での探索

2. 全ての部分文仮説が 1 つの最尤パスにマージするフレーム (図 2の $t_{boundary}$) を最適な単語境界フレームとする。ここで確定した最終単語 (図 2の W_{last}) の直前までの単語列を出力する。もし始端まで最適な単語境界フレームが見つからない場合は、終端の最尤状態からトレースバックした時の最尤パスの単語列を出力する。

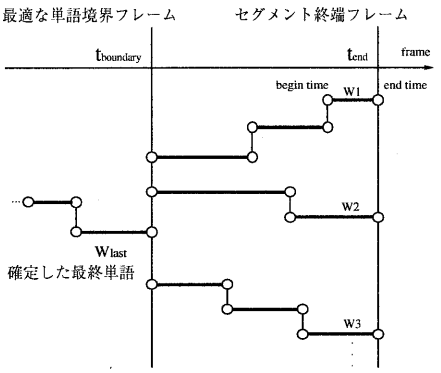

図 2: 単語トレリス上での最適な単語境界の探索

3. 次のセグメント $N+1$ では、直前セグメント N で確定した最終単語 (W_{last}) の始端を探索開始フレームとする。すなわち図3でセグメント N の終端から Δt だけ遡ったフレームより再探索を行う。ここでセグメント境界で bigram の言語制約を有効にするために直前セグメントの確定した最終単語の ID を伝搬させる。

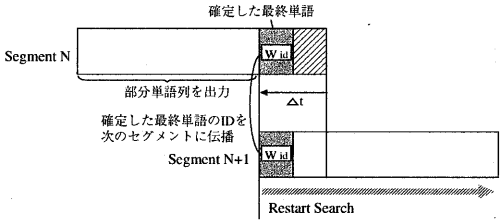

図 3: セグメント間での再探索

3 基本性能の評価

アルゴリズムの実装は大語彙ディクテーション Julius[7] の 1 パスサーチのモジュールをベースとして行った。ここでは提案手法の基本性能評価として新聞記事の連続読み上げ音声を用いた認識実験について述べる。実験条件を表 1に示す。

表 1: 基本性能評価の実験条件

音響分析	16kHz サンプリング、16bit 量子化、特徴量 (MFCC 12 次、ΔMFCC 12 次、Δ パワー) 分析フレーム長 25ms、フレームシフト 10ms Hamming 窓、プリエンファシス 0.97
音響モデル	IPA[8] の PTM HMM (男性話者モデル)
言語モデル	IPA[8] の新聞記事 20k 語彙 bigram

評価データは JNAS 新聞記事読み上げコーパスの中から男性話者 23 名発声の合計 100 文を連結した音声を用いた。時間長は約 583 秒 (58385 フレーム) である。

3.1 実験結果

実験では文節から 1 文程度の発声時間長を考慮して、セグメントサイズを 200,300,400,500 フレームの 4 通りに設定して認識を行った。提案手法 (Proposed method) の単語正解精度 (%accuracy) を図 4に示す。また比較のため、正確な端点で切り出した音声を与えた場合の Julius の 1 パスサーチのベースライン性能 (Baseline) と、オーバーラップなしで 200～500 フレームの各サイズに機械的に分割した音声を入力とした場合の Julius の 1 パスサーチの認識結果 (Not using proposed method) も併せて示す。

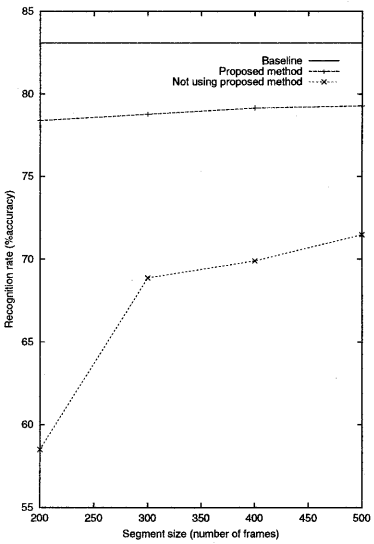

図 4: 基本性能 (認識率) の評価

3.2 考察

実験結果より、音声の明示的な端点検出処理を全く行わずに端点を正確に与えた従来法の 1 パスサーチの数%落ちの認識性能を達成できることを示した。

3.2.1 途切れた単語修復の効果

再探索処理を行わない場合 (図 4の Not using proposed method) と比較して、単語正解精度で 8%から最大 20%の認識率向上が確認された。提案手法のアルゴリズムによってセグメント分割により途切れた単語区間の復元が適切になされていることがわかる。セグメ

ントサイズについては200フレームまではほとんど認識性能の劣化は見られないが、100フレームの長さでは再探索で単語境界の探索が不安定になるため50.19%と認識性能の大幅な低下が見られた。認識誤りの内訳では置換と脱落の抑制に効果が見られた。特にセグメント始端では隣接するセグメント間でbigramによる言語モデル制約が働くため、単語の置換誤り(特に同音語の置換誤り)が大幅に減少した。

3.2.2 ベースライン性能との差

提案法とベースライン性能の差であるが、まず一つはセグメント内で最適な単語境界が見つからない場合のサーチエラーが考えられる。参考として図5に各セグメント内の探索において、終端から推定した単語境界までバックトラックしたフレーム数(セグメントサイズ400の場合)を示す。図5の結果では181回のセグメント単位の処理が行われた内、1回だけ単語境界が定まらなかったことを示している。

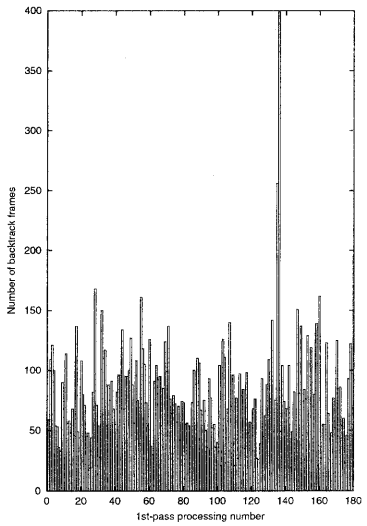

図5: 単語境界までバックトラックしたフレーム数(セグメントサイズ400)

他の要因としては言語モデルが1文単位で構築されていることが挙げられる。このため文と文の間にサーチがまたがる場合は言語制約が有効に機能していないものと推測される。検証のため、評価に用いた新聞読み上げ音声の中で30文を連結したデータ(男性7名、時間長16957フレーム)を分割なしで全体をJuliusの1パスサーチで認識させたところ、単語正解精度は76.31%であった。これに対し正確な端点で分割した場合は84.97%と約8%程度の差が生じた。このことから提案手法のサーチにおいてもセグメントの中に2つの文の境界が入った場合は言語モデルに起因する認識誤りが生じているものと考えられる。

4 高次言語モデルを用いた性能改善

これまで述べたbigramによる1パスサーチのベース性能が十分でないことから、改善すべき問題としてサーチアルゴリズムの改良が挙げられる。以下では高次言語モデルとしてtrigramを用いたアルゴリズムの改良について検討を行う。高次言語モデルを用いたサーチを実現するには、具体的に次の2つの方法が考えられる。

時間同期1パスtrigramサーチ: 第1パスからtrigramによるフレーム同期サーチを行う。セグメント間にまたがる処理は2節の基本アルゴリズムとほぼ同様に行う。

2パスtrigramサーチ: 第1パスではbigramによるフレーム同期サーチを行い、第2パスではセグメント内で確定した最終単語からtrigramによる単語同期の後向きサーチを行う。これをセグメント単位で逐次的に実行する。

前者の手法では探索の処理コストが高く実装も複雑になることから、以下では後者の手法について検討を行った。

4.1 2パスtrigramサーチの検討

まず数文節程度の1~3秒の短い時間長の音声に対してtrigramによる2パスサーチが有効であるかどうかを調べる予備実験を行った。実験では、前述のJNAS新聞読み上げコーパスの男性話者23名の100文を一定区間に機械的に分割した評価データ(150~300フレーム、オーバラップなし)に対し、Juliusの1パスサーチ(bigram)と2パスサーチ(bigram+trigram)で性能比較を行った。その結果、2パスサーチを実行した場合は各セグメントサイズについて単語正解精度で約3%から5%程度の認識率向上が見られた。よって1~3秒程度の短い音声区間に対しても高精度の言語モデルを用いた2パスサーチが有効であることがわかる。さらに、提案手法のセグメント間の単語修復処理を2パスサーチに合わせて適切に行えば、より性能向上が期待できる。

4.2 アルゴリズム

一つのセグメント内で2節で述べた手順によって最適な単語境界フレームを決定し、部分単語列を確定する時に以下の処理を行う(図6参照)。

1. 確定した最終単語 W_{last} の始端フレームの直前より、第1パスのサーチ結果をヒューリスティクスとした単語単位の best-first な後向きスタックデコーディング探索 [7] を実行し該当区間の部分文仮説の再評価を行う。

2. この時、後向きサーチの初期仮説として単語 W_{last} を言語制約として与える。

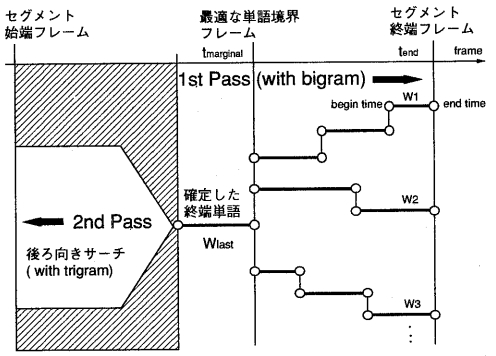

図 6: セグメント内での 2 パス trigram サーチ

4.3 実験

2 パス trigram サーチの効果を検証する実験を行った。実験条件および評価データは基本性能評価の時と同一である。ただし、言語モデルとして 20K 語彙の後向き trigram を追加している。実験結果を図 7に示す (Proposed method-2pass)。図には比較のため 2節の基本アルゴリズムによる結果 (Proposed method-1pass) 及び、端点を正確に与えた場合の Julius の 1 パスサーチ (Baseline-1pass) と 2 パスサーチ (Baseline-2pass) の各ベースライン性能を示す。

4.4 考察

実験結果より、セグメントサイズが 500 フレームにおいては端点を正確に与えた場合のベースラインの認識結果の約 3%落ちの性能が得られた。またセグメントサイズが 300 フレーム以上であれば、端点を正確に与えた場合の 1 パスサーチのベースライン性能を上回る結果を示している。trigram を用いた探索では bigram の場合と比較してセグメントサイズが大きくなる (単語数が多くなる) ほど効果が高いことがわかる。300 フレーム以下の短い間隔であっても言語モデルによる精度向上の効果が見られた。以上により、基本アルゴリズム

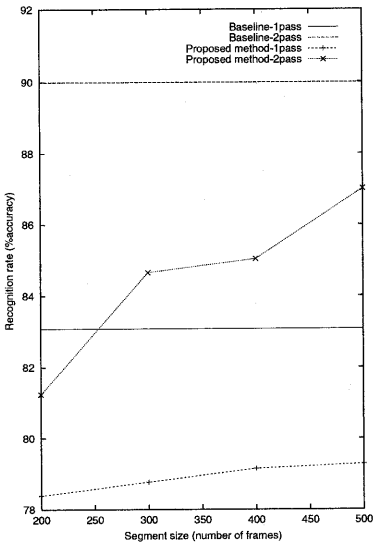

図 7: 2 パス trigram サーチによる認識率改善

に対して trigram を用いた 2 パスサーチの改良手法は性能改善に非常に効果が大きいことがわかった。

5 車内音声対話の書き起こし実験

提案手法の実環境での有効性を検証するために、現在名古屋大学 CIAIR で収集を行っている車内音声データベース [9] の自動書き起こしの試行実験を行った。CIAIR では音声対話による情報検索システムを想定してドライバーと助手席のシステム役の対話を収集している。タスクはレストランや店舗などの情報検索、予約、道案内である。通常音声対話の書き起こしとタグ付けを行うには人手による多大なコストを要するため、提案手法を応用した自動書き起こしシステムが実現できれば有用性は非常に高いと考えられる。

5.1 評価データと実験条件

評価データとして、1999 年 12 月に事前収集した店舗情報検索の音声対話コーパスより音響モデル、言語モデルの学習に用いていない 230 発話、約 30 分 (181939 フレーム) の対話音声を用いた。発話の内訳はドライバー 79 発話 (男性 1 名)、システム役 151 発話 (女性 1 名) である。評価音声の収録には接話マイクを使用している。

実験条件を表 2に示す。クラス bigram の学習には CIAIR 音声対話コーパスの書き起こし 3424 文を用いた。テストセットパープレキシティは 29.46、未知語率は 2.49%である。

表 2: 車内音声対話書き起こしの実験条件

音響分析	基本性能評価と同じ
音響モデル	IPA の PTM HMM (男女共用モデル)
言語モデル	品詞をクラスとした 2K 語彙のクラス bigram

試行実験ではコーパス書き起こしの学習文が十分に用意できなかったことからクラス trigram の学習は信頼できるパラメータ推定が困難であり、言語モデルの構築はクラス trigram までとした。したがってデコードの手法も 2節の基本アルゴリズムを用いて認識を行った。

5.2 実験結果

実験はセグメントサイズ 400 フレームで行った。認識結果 (%correct と%accuracy) を表 3に示す。

表 3: 実験結果 (bigram を用いた基本アルゴリズム)

Speaker	% Correct	% Accuracy
男性 (driver)	69.72	66.42
女性 (system)	78.64	70.60
Total	75.44	69.10

5.3 考察

実験結果より車内音声データの自動書き起こしの実現性を示すことができた。特に複数話者の音声や環境音が混在したデータに対して全く前処理を必要としない本手法は大量の音声の書き起こしに適している。単語正解精度はまだ十分とはいえないが、書き起こしの補助ツールとしては、ある程度実用レベルであると考えられる。今後の性能改善のためには対話タスクの trigram 言語モデルの構築と車内音響モデルの最適化を行っていく必要がある。

6 おわりに

入力音声の明示的な端点検出を必要としない新しい連続音声認識手法を提案した。本手法では数秒程度の一定時間長の処理ブロックを認識すると同時に終端で途切れた単語区間をバックトラックによって修復しながら連続的にデコードを続けるため、端点検出や発話単位の考慮無しに無限長の入力音声を認識することが可能である。提案手法の有効性を検証するために、連続した約 10 分の新聞読み上げ音声のディクテーション実験を行った。その結果、基本性能評価では端点を正確に与えた場合のベースライン性能の約 3%落ちの性能が得られた。また実環境での性能評価のため、男女各 1 名による約 30 分の車内音声対話を用いた評価実験を行い音声対話の自動書き起こしの実現性を示した。実環境下での認識性能については今後十分な評価を行っていく必要があるが、長時間の音声モニタリングなど様々な応用が考えられる。

謝辞 本研究の一部は文部省科学研究費補助金 COE 形成基礎研究費 (課題番号 11CE2005) の補助を受けて行われた。

参考文献

[1] L.R.Rabiner and M.R.Sambour. "An algorithm for determining the endpoints of isolated utterances", The Bell System Technical Journal, vol.54, no.2, pp.297-315, Feb 1975.

[2] J.G.Wilpon, L.R.Rabiner and T.Martin. "An improved word-detection algorithm for telephone-quality speech incorporating both syntactic and semantic constraints", AT&T Bell Laboratories Technical Journal, vol.63, no.3, pp.479-498, Mar 1984.

[3] J.G.Wilpon and L.R.Rabiner. "Application of hidden Markov models to automatic speech endpoint detection", Computer Speech and Language, vol.2 pp.321-341, 1987.

[4] A.Acero. "Robust HMM-based endpoint detector", In Proc. European Conference on Speech Communication Technology, pp.1551-1554, 1993.

[5] J.C.Junqua, B.Mak and B.Reaves. "A robust algorithm for word boundary detection in the presense of noise", IEEE Trans. on Speech and Audio Processing, 2(3) pp.406-412, 1994.

[6] 内藤, 黒岩, 山本, 武田: "部分文仮説のゆう度を用いた連続音声認識のための音声区間検出法", 信学論 D-II Vol.J80-D-II No.11 pp.2895-2903 1997.

[7] 李, 河原, 堂下: "単語トレリスインデックスを用いた段階的探索による大語彙連続音声認識", 信学論 D-II Vol.J82-D-II No.1 pp.1-9 1999.

[8] 河原他: "日本語ディクテーション基本ソフトウェア (99 年度版) の性能評価", 情処研報 SLP31 pp.9-16 2000.

[9] 河口, 梶田, 岩, 松原, 武田, 板倉: "実走行環境下における車内音声データベースの構築", 音講論 2000 春季 pp.191-192 2000.