社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－SLP－62 (6)
2006／7／7

フレーム単位の信頼度を用いた並列音声認識におけるデコーダ間枝刈りの検討

袴田智博† 南角吉彦†
李晃伸† 徳田恵一†

様々な環境や話者に対して高精度な不特定話者音声認識を実現するためのアプローチとして，対象の異なる複数のモデルによる結果を選択する方法が研究されている．この実現方法のひとつとしてモデルごとに独立した認識処理を統合する並列デコーディングがあるが，この場合デコーダの数に比例して計算量が増大する問題がある．本研究では，各デコーダの認識処理途中のフレームの情報に基づいてデコーダ間で動的に枝刈りを行うことを検討する．入力音声に対して相対的に適合しないと判断できるモデルの認識処理を中断することにより，最後まで処理を行うデコーダの数を絞り込み，トータルの計算量を削減する．枝刈りの基準として，フレームごとの現存単語仮説の最大累積尤度および，その現存仮説集合から得られる事後確率に基づく信頼度の上位の値を検討する．7 ～ 12 のモデルの組み合せを用いた評価実験において，認識処理終了後にモデル選択を行う従来の並列音声認識に比べてほぼ同じ精度を保ちつつ，計算量を全体の 1/3 程度に抑えることができた．

Inter-decoder Pruing on Parallel Decoding using Frame-wise Confidence Scores

Tomohiro Hakamata,† Yoshihiko Nankaku,† Akinobu Lee† and Keiichi Tokuda†

Parallel decoding based on multiple models has been studied on a speech recognition system to efficiently cover various conditions and speakers in real world. However, running many recognizers in parallel applying all models causes the total computational cost to grow. In this paper, an efficient way of finding and pruning unpromising decoding process while search based on frame-wise likelihoods of each model is proposed. By comparing temporal search statistics at each frame among all decoders, a decoder with relatively unmatched models can be pruned in halfway of recognition process. This method allows the model structures to be mutually independent. Experimental results on parallel recognition of various acoustic models showed that two thirds of the computational cost was reduced compared to full computation by using the both criteria without spoiling the recognition accuracy as compared with conventional post-selection.

1. はじめに

実環境を想定した音声認識システムでは，様々な音声の収録環境や話者の多様性を考慮する必要がある．これに対するアプローチのひとつとして，複数モデルを用いた音声認識の研究が行われている．様々な条件のデータ集合から単一のモデルを学習するマルチコンディションHMMや，異なる条件で個別に学習された音響モデルを統合するマルチミクスチャーHMMが提案されており，後者はより性能が改善することが報告されている[1]．しかしマルチミクスチャーHMMを作成する場合は音響モデルの構造が同一でなければならないため，汎用性に欠けるという問題点がある．

他の複数モデルを用いた認識の手法として並列デコーディング[2]が挙げられる．これは，異なるモデルやパラメータを与えたデコーダを並列に並べ，入力音声に対してそれら全てによる認識を行い，出力をまとめるという手法である．その際，複数のデコーダから出力された結果を用いて，ROVER[4]に代表される結果統合を行うのが一般的である．この手法の利点は，構造の異なる音響モデルや，語彙の異なる言語モデルを一斉に適用できることである．そのため，様々な環境やドメイン，発話方法や話題へ対応でき，実環境における発話に対してロバストな認識を実現できる．

しかし，並列デコーディングを行う場合，処理を行うデコーダの数に比例して計算量が増大する．そこで

† 名古屋工業大学大学院 工学研究科
Department of Computer Science and Engineering, Nagoya Institute of Technology

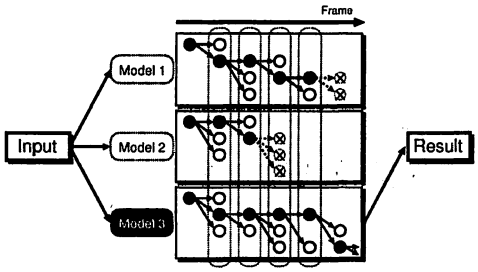

図 1　フレーム単位のデコーダ間枝刈り

本研究では，並列音声認識における計算量の削減を目的とした，フレーム単位の信頼度を用いたデコーダ間枝刈りの手法を検討する．認識処理中に入力音声に対するマッチ度合をデコーダ間で相対的に比較し，マッチしないと考えられるモデルの処理を中断させる．比較のための基準値として，各フレームの最大仮説尤度と，そのフレームに存在する単語終端の事後確率に基づく信頼度を用いる．並列デコーディングの枠組みにおいて，認識途中に計算を行うデコーダの数を動的に絞り込んでいくことで，計算量の削減を行う．

2. デコーダ間枝刈り

2.1 アイデア

複数モデルを用いた並列音声認識においては，いくつかの実現方法が考えられる．代表的なものとしては，単一のデコーダに全てのモデルを与え，従来の音声認識と同様のアルゴリズムによって全モデルから一つの最尤系列を得るものや，複数デコーダによって N-Best やラティス形式でそれぞれのモデルでの認識結果を出力し，それを事後的に統合して候補を確定していくものがある[5]．ただし後者においては，語彙がある程度共通であることを前提としている．

有効なモデルの組み合せは，認識対象の話者層やシステムの設置条件によって様々であると考えられるため，実環境において汎用性の高い並列音声認識を実現するために，本稿では，モデルごとにデコーダを用意し独立した認識処理を行いつつ，認識処理中にフレームごとの尤度等の解探索情報を相互に見比べることで，相対的にマッチ度の低いモデルの認識処理を計算途中で中断することを提案する．入力音声に対して最適なモデルが与えられたデコーダを動的に自動選択し，その出力を最終的な認識結果とする．並列に認識処理を行っている最中に不適切なデコーダを枝刈りすることができれば，最適なモデルによる出力が通常の並列認識と比較して少ない計算量で得られる．この提案手法を図示したものを図 1 に示す．

フレーム同期ビーム探索アルゴリズムによる認識デコーダを用いた音声認識では，フレーム毎に可能性のある仮説が展開され，それぞれの仮説についてのスコアが計算される．このとき，上位の仮説に与えられたスコアが他のモデルと比較して相対的に小さかった場合は，そのモデルは不適当だと判断されて処理が中断される．

2.2 枝刈り基準

提案手法であるデコーダ間枝刈りの性能は，複数モデルを用いて最後まで計算したときに得られる認識精度を損なわないまま，どれだけ早期に不適切なモデルの計算を中断できるかに依存する．

本稿では，フレーム単位の音声認識処理において得られる最大の単語仮説尤度と，事後確率による信頼度の上位の値の二つの基準値を用いた枝刈りを行う．以下に詳細を述べる．

2.2.1 最大仮説尤度

音声認識において入力音声に対して不適切なモデルは，適切なモデルと比較して仮説に与える尤度が低くなる．そのため，各々のフレームにおいて各モデルが出力する仮説尤度をモデルの比較基準とすることが考えられる．

入力始端フレームからフレーム t において終端が存在する単語仮説 w を $[w,t]$ とし，$W_{\mathrm{best}}(k)$ を仮説 k の最尤コンテキストと定義する．モデル m の仮説尤度 $g_m([w,t])$ は，以下のように定義される．

$$g_m([w,t]) = \log P(x_t|W_{\mathrm{best}}([w,t]))P(W_{\mathrm{best}}([w,t])) \quad (1)$$

なお，x_t はフレーム 0 から フレーム t の間の入力音声に相当する．ここで $W_m(t)$ をフレーム t においてモデル m を与えたデコーダで生き残っていた仮説の集合とすると，各デコーダの最大仮説尤度 $g_m(t)$ は以下の式で定義される．

$$g_m(t) = \max_{w \in W_m(t)} g_m([w,t]). \quad (2)$$

認識処理中，フレーム t においてデコーダ m は，以下の基準を満たした場合に処理を中断される．

$$g_m(t) + g_{\mathrm{off}} < \max_m g_m(t) \quad (3)$$

なお，閾値 g_{off} は固定値をあらかじめ与える．

さらにモデル比較の開始点としてパラメータ *start* を与える．これは例えば雑音成分を多く含む入力音声などで，入力始端付近の無音区間におけるスコアのぶれなどによる不適当な枝刈りを避けるためである．

また，複数のデコーダが入力終端フレーム T まで生き残った場合は，その時点での最大仮説尤度 $g_m(T)$ が最も大きいモデルを最終結果として選択する．

2.2.2 上位の単語信頼度

事後確率による信頼度は，音声認識結果に対して付与する信頼度として一般的な手法である[7]．提案法では，各フレームにおいて終端が存在する単語仮説集合の累積尤度から計算された信頼度をデコーダ間の比較に用いる．事後確率は競合候補における任意の単語仮説尤度の相対的な分布を表し，合計値は常に 1 とな

るため，その中での上位の信頼度は，そのモデルの弁別度を反映しているものと考えられる．

各フレームにおいて，それぞれのモデルによって展開された仮説の信頼度は，その時点での累積尤度を元にした事後確率から計算することができる．入力のフレーム τ から t に単語仮説 w が存在するとき，その単語仮説 $[w;\tau,t]$ を含む全ての系列は $W_{[w;\tau,t]}$ で表されるものとする．このとき単語仮説 $[w;\tau,t]$ の入力音声 x_T に対する事後確率 $P([w;\tau,t]|x)$ は，その仮説をパス上に含む全文仮説の出現確率の和より求められ，以下の式で表される．

$$P([w;\tau,t]|x_T) = \sum_{W \in W_{[w;\tau,t]}} \frac{P(x_T|W)P(W)}{P(x_T)} \quad (4)$$

認識処理中に，全てのパスの合計値をその段階での最尤系列で近似することにより[7]，フレーム t $(0 < t \leq T)$ における事後確率は，

$$P([w;\tau,t]|x_t) = \sum_{W \in W_{[w;\tau,t]}} \frac{P(x_t|W)P(W)}{P(x_t)} \quad (5)$$

$$\approx \frac{P(x_t|W_{\mathrm{best}}([w,t]))P(W_{\mathrm{best}}([w,t]))}{P(x_t)} \quad (6)$$

となる．$P(x_t)$ は x_t に対して存在する全系列の合計と近似されるため，フレーム t におけるモデル m による単語仮説 w の事後確率は以下の式で表現され，

$$P_m([w,t]|x_t) \approx \frac{e^{g_m([w,t])}}{P(x_t)} \quad (7)$$

$$\approx \frac{e^{g_m([w,t])}}{\sum_{w' \in W_m(t)} e^{g_m([w',t])}} \quad (8)$$

さらにスムージング係数 α を加えることで，信頼度の式は以下のように表される．

$$C_m([w,t]) = \frac{e^{\alpha \cdot g_m([w,t])}}{\sum_{w' \in W_m(t)} e^{\alpha \cdot g_m([w',t])}} \quad (9)$$

大きい信頼度を付与された仮説を持つモデルは，その時点での競合仮説間において仮説を一意に識別できているということが期待される．そのため，上位の信頼度によってモデルの弁別度が判定可能だと考えられる．同程度の大きさの信頼度を持つ仮説を複数展開していることも考えられるため，実際は上位 N の信頼度をそのモデルの信頼度 $C_m(t)$ として用いる．これは以下の式で表され，

$$C_m(t) = \sum_{\text{best } N \text{ words at } t} C_m([w,t]) \quad (10)$$

認識時に以下の条件を満たしたデコーダ m は処理を中断される．

$$C_m(t) + c_{\mathrm{off}} < \max_m C_m(t) \quad (11)$$

信頼度は正規化された値であるため，尤度基準によるモデル比較と異なり，モデルを同じスケールで比較することが可能である．よってモデル規模が大幅に異なる場合でも，候補の比較が可能となる．

3. 評価実験

提案法の評価実験を行った．提案手法であるデコーダ間枝刈りは，オープンソース認識エンジン Julius Ver.3.5[8] に実装された．Julius の第 1 パスのフレーム単位ビームサーチにおいて，提案法であるデコーダ間の枝刈りが行われる．第 2 パスの計算は枝刈りされずに生き残ったデコーダのみで行われ，最終的な結果を得ることになる．

3.1 実験条件

連続音声認識コンソーシアム (CSRC)[9] により提供される言語モデル・音響モデルを使用した．学習用データベースには ATR-BLA と JNAS+ASJ が用いられており，ATR-BLA データベースは 3,769 話者，合計 162 時間の自発的な発話が収録されている．また，JNAS+ASJ データベースは 361 話者，合計 98 時間の読み上げ音声で構成されている．

音響モデルは，表 1 に示した 7 種類を使用した．この中で “JNAS-PTM” と “JNAS-tri” は JNAS のデータベース，“ATR-PTM” と “ATR-tri” は ATR のデータベースによって学習されている．“Senior-PTM” は高齢話者による JNAS のデータベースと同数の発話データによって学習されており，“Child-PTM” は 400 人の子供話者の単語発話によるモデルである．また，“JNAS-Tel-tri” は JNAS データベースから学習されたものだが，認識タスクを電話音声の帯域に限定したモデルとなっている．なお，モデル名の末尾の “-PTM” と “-tri” はモデルの構造を表しており，前者は音素内タイドミクスチャー (PTM) モデルを，後者は状態共有のトライフォンモデルを表す．

言語モデルは毎日新聞の記事から学習した語彙 20,000 語の単語 3-gram を用いた．

テストセットは 2 種類のデータベースから集めたデータを用いた．男性話者 23 名，女性話者 23 名による，それぞれ 50 文章の JNAS データベースの文章を読み上げたものと，男性話者 4 名による ATR のデータベースの読み上げ 50 文章である．なお，それらの文章は学習データには含まれていない．

全ての音声はサンプリング周波数 16kHz，25ms のハミング窓によって特徴抽出を行った．特徴ベクトルは 12 次元の MFCC 及びそのデルタとパワーである．

全ての実験においてデコーディングにおける第 1 パスのビーム幅は 2000 とし，信頼度のスケーリング係数 α は 0.05 に設定した．

3.2 評価基準

本実験においては，7 個の音響モデルを並列に 7 個のデコーダに与え，認識を行う．提案法による計算量

表 1 各テストセットに対する単一モデルでの認識精度

モデル	WER		
	JNAS	ATR	Total
JNAS-PTM	7.28%	21.37%	10.11%
ATR-PTM	7.70%	16.16%	**9.40%**
Senior-PTM	11.81%	27.22%	14.90%
Child-PTM	39.81%	36.97%	51.16%
JNAS-Tel-tri	96.67%	92.12%	95.77%
JNAS-tri	6.56%	21.49%	9.55%
ATR-tri	7.27%	20.16%	9.85%

表 2 事後的なモデル選択を行う複数モデル音声認識の性能

手法	WER	Cost
最適な単一モデル (ATR-PTM)	9.40%	14.29
人手によるモデル選択	5.36%	100.00
第 2 パスの尤度による自動選択	8.94%	100.00
第 1 パスの尤度による自動選択	9.33%	100.00

削減の評価のために，単語誤り率 (WER) に加えて，全てのデコーダで最後まで認識を行った場合のフレーム数と実際に計算を行ったフレーム数の比率を $Cost$ という値として定義した．例えば全デコーダにおいて第 1 パスの最終フレームまで処理を行った場合は $Cost = 100$ となり，単一のデコーダで認識を行った場合は，その他のデコーダでの処理を行ったフレーム数は全て 0 となるので $Cost = 14.29$ と計算される．なお，第 2 パスの計算量はここでは無視している．

3.3 単独モデルおよび事後モデル選択の評価

全ての音響モデルをそれぞれ単一に適用した場合の単語誤り率を表 1 に示す．この結果，ATR-PTM モデルが最良の単語誤り率である 9.40% を示した．なおこの場合の計算コストは $Cost = 14.29$ となる．

続いて，比較のために，全てのモデルの計算終了後に最終フレームでの値で事後的にモデル選択を行う，静的な並列音声認識の実験を行った．全てのモデルによって第 1 パス，もしくは 第 2 パスの最終フレームまでデコーディングが行われ，その時点での最大尤度をもつモデルが最終的な結果を出力するものとして選択される．この結果を表 2 に示す．第 1 パスの尤度を用いた場合単語誤り率が 9.33%，第 2 パスの最終尤度を用いた場合 8.94% となり，並列音声認識による精度改善の効果がみられた．

さらに比較のため，テストセットの発話それぞれに対して最適なモデルを人手によって選択した場合を調べたところ，単語誤り率 5.36% を得た．これが自動モデル選択で得られる単語誤り率の上限値となる．

3.4 デコーダ間枝刈りの評価

モデル評価基準として，最大仮説尤度と上位信頼度を用いた場合のフレームごとの値の変遷をそれぞれ図 2 および図 3 に示す．なお，事前実験によって，信頼度を用いる場合は上位 $N = 4$ 個を用いた場合に最良の結果が得られたため，図 3 では $N = 4$ とした．

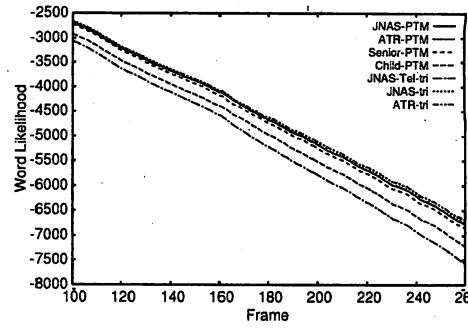

図 2 最大仮説尤度の変遷

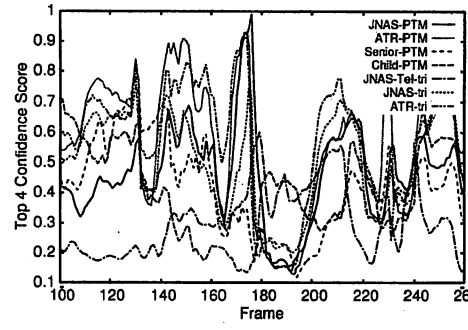

図 3 上位 4 個の信頼度の変遷

表 3 枝刈り基準の比較

基準	WER	Cost
仮説尤度	9.17%	39.42
信頼度	8.19%	51.71
仮説尤度 + 信頼度	8.53%	36.53

最大仮説尤度を用いた場合は，子供モデルと電話音声モデルのスコアが大幅に小さくなり，その他のモデルによるスコアはそれほど大きな差は出なかった．このことから仮説尤度が，モデルを学習したデータの音響空間を反映していることが分かる．それに対して上位信頼度を用いた場合は，子供モデルと電話音声モデルだけでなく，高齢者モデルの値も平均して小さくなっている．これらの変遷において，最も良い値から一定値よりも小さくなったスコアを持つモデルは，その時点でデコーディングが中断される．

次に，これらのモデル評価基準を用いて，提案法であるデコーダ間の動的枝刈りの評価実験を行った．それぞれのモデル評価基準を単独で用いた場合，および併せて用いた場合の単語誤り率と計算コストの関係を図 4 に示す．それぞれ，パラメータ g_{off}, $start$, C_{off} を変化させたときの値である．また，単語誤り率をほとんど下げない条件で最も計算量を抑えたときの値を表 3 に示す．

尤度基準の枝刈りによって，従来の事後的なモデル

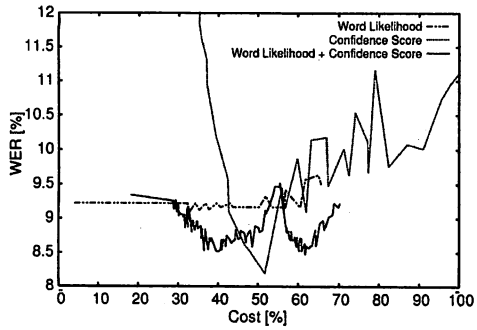

図 4 デコーダ間枝刈り (提案手法) による単語誤り率と計算量

選択とほぼ同じ精度を得ながら，計算コストを 39.42% まで削減できた．信頼度基準を用いた枝刈りでは計算コストは若干大きいもののより良い精度が得られた．

仮説尤度を用いた場合の枝刈りでは，単語誤り率と計算コストに明瞭な相互関係は見い出せない．また，モデル間での最大仮説尤度の順位が変わることは稀であり，差分の変化もそれほど大きくない．そのため，閾値 g_{off} を大きく設定した場合は，途中で枝刈りされるモデルが少なくなる．

仮説尤度と信頼度を併せて基準として用いた場合に最良の結果が得られた．認識の処理の初期段階で，入力音声に対して大幅にドメインの異なるモデルである子供や電話音声モデルが尤度基準によって除去され，後段のモデル比較が信頼度を用いて上手く行われた場合に，最も良い認識結果が低い計算コストで得られた．最終的には，双方の基準を併せて使用した枝刈りによって，最適な計算コスト 36.53% が得られた．

4. 追加実験

前説の実験に加え，奈良先端大で運用されている実環境音声情報案内システム「たけまるくん」[10] の収集データについて同様の実験を行った．

4.1 実験条件

公共施設に設置された「たけまるくん」によって収集されたデータにより作成された，年齢層別の 2 種の言語モデルおよび 6 種の音響モデルの組み合わせについて提案手法を適応し，その効果を検証した．

音響モデルは adult (成人話者)，junior (小学校高学年話者)，element (小学校低学年話者)，infant (小児話者) と，更に全ての子供話者モデルを統合したものとして allchild，その内 infant を除いた child モデルを使用した．音響モデルは全て PTM であり，2,000 状態 64 混合からなる．

言語モデルは adult (成人モデル)，child (子供モデル) の 2 種類であり，共に語彙数約 40,000 語の単語 3-gram である．

たけまるシステムにおいて収集されたデータは，情報案内システムの性質から発話内容がほぼ同一ものが多数存在するため，書き起こしの頻度を元にそれぞれ異なる発話内容のデータをテストセットとして選択した．2002 年 12 月から 2004 年 10 月の収録データから抽出した 5,000 文章のうち，本実験では，成人話者による 200 発話，小学校高学年話者による 100 発話，小学校低学年話者による 100 発話，小児話者による 100 発話の計 500 文章をテストセットとして用いた．なお，音響モデルの学習に用いられたデータも同様の期間に収録されたものであるが，その中に評価データは含まれていない．

デコーディング時の第 1 パスのビーム幅は 1500，信頼度のスケーリング係数は 0.05 とした．

4.2 単独モデルおよび事後モデル選択の評価

表 4 に，各テストセットに対する各モデルでの認識性能を示す．単独モデルでは，allchild 音響モデルと adult 言語モデルの組み合わせによって最良の単語誤り率 24.78% が得られた．

続いて静的並列音声認識の実験を行った．結果を表 5 に示す．自動選択では，第 1 パスの尤度基準で 20.12%，第 2 パスの尤度基準で 19.34% の単語誤り率を得た．また，人手によるモデル選択では 11.56% の単語誤り率が得られた．これらの結果より，複数モデルを用いた並列認識による誤り低減が確認された．

表 5 事後的なモデル選択を行う複数モデル音声認識の性能

手法	WER	Cost
最適な単一モデル	24.78%	8.33
人手によるモデル選択	11.56%	100.00
第 2 パスの尤度による自動選択	19.34%	100.00
第 1 パスの尤度による自動選択	20.12%	100.00

4.3 デコーダ間枝刈りの評価

それぞれのモデル評価基準を単独で用いた場合及び，併せて用いた場合に得られた単語誤り率と計算コストの関係を図 5 に示す．前節の実験と同様，それぞれパラメータ g_{off}, $start$, C_{off} を変化させた値であり，信頼度の上位は $N = 4$ に固定している．また，単語誤り率を保ちながら計算コストを下げる場合の最良の結果を表 6 に示す．提案手法により，静的選択に近い単語誤り率を保ちつつ，計算量を全体の 29.00% まで削減することができた．

表 6 枝刈り基準の比較 (たけまる)

基準	WER	Cost
仮説尤度	20.48%	35.04
信頼度	20.28%	50.76
仮説尤度 + 信頼度	20.44%	29.00

表 4 各テストセットに対する単一モデルでの認識精度 (たけまる)

音響モデル	言語モデル	WER				
		adult	junior	element	infant	total
adult	adult	11.01%	22.13%	27.82%	67.63%	27.92%
allchild	adult	17.74%	25.72%	23.91%	39.27%	**24.68%**
child	adult	18.74%	24.92%	22.36%	42.73%	25.30%
junior	adult	16.76%	22.54%	24.50%	56.23%	27.36%
element	adult	28.50%	27.25%	24.52%	40.82%	29.92%
infant	adult	54.33%	46.94%	34.44%	40.93%	46.20%
adult	child	13.30%	21.86%	26.40%	72.43%	29.56%
allchild	child	22.85%	24.44%	21.04%	42.03%	26.64%
child	child	21.41%	22.36%	20.50%	45.35%	26.21%
junior	child	20.53%	18.13%	19.66%	62.08%	28.19%
element	child	29.84%	25.86%	22.50%	44.25%	30.46%
infant	child	58.71%	43.97%	30.00%	40.93%	46.46%

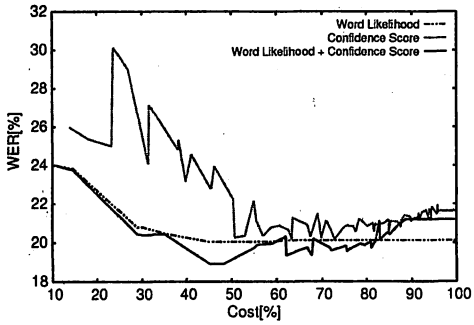

図 5 提案法による単語誤り率と計算量 (たけまる)

5. おわりに

並列音声認識における計算量の削減のため，デコーダ間の枝刈り手法を提案，検討した．この手法はフレーム単位のビーム探索において，認識処理中に得られる最大の仮説尤度及び上位の信頼度を用いることで，フレームごとにデコーダ間の入力に対するマッチ度を動的に比較するものである．7 ～ 12 のモデルの組み合せを並列処理する評価実験において，認識処理終了後に静的にモデル選択を行う従来の並列音声認識の精度を保ちつつ，計算量を全体のおよそ 1/3 に抑えることができた．

6. 謝　辞

この研究の一部は，文部科学省のリーディングプロジェクト「e-Society 基盤ソフトウェアの総合開発」によって行われた．

本稿の実験を行うにあたって，音声情報案内システム「たけまるくん」のテストセットを貸与していただいた，鹿野清宏教授をはじめとする奈良先端大の皆様に感謝します．

参考文献

1) 鈴木基之，加藤裕介，伊藤彰則，牧野正三．SN 比に頑健なマルチミクスチャー HMM の性能評価．電子情報通信学会研究会技術報告，SP2005-31, pp.25-30, (2005).

2) 松田繁樹，中村哲．パラレルデコーディングを基礎とした ATR 音声認識システム．2006 年日本音響学会春季研究発表会，1-P-29, pp.195–196, (2006).

3) Masahiko Matsushita, Hiromitsu Nishizaki, Yasuhiro Kodama, Takehito Utsuro, and Seiichi Nakagawa, "Evaluating multiple LVCSR model combination in NTCIR-3 speech-driven web retrieval task," in *Proc. Eurospeech*, pp.1205–1208 (2003).

4) Jonathan G. Fiscus, "A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER)," in *Proc. IEEE ASRU*, pp.347–352 (1997).

5) Yonghong Yan, Chengyi Zheng, Jianping Zhang, Jielin Pan, Jiang Han, and Jian Liu, "A dynamic cross-reference pruning strategy for multiple feature fusion at decoder run time," in *Proc. Eurospeech*, pp.1177–1180 (2003).

6) Simo O. Kamppari and Timothy J. Hazen, "Word and phone level acoustic confidence scoring," in *Proc. IEEE ICASSP*, pp.1894–1897 (2000).

7) 李晃伸，鹿野清宏，河原達也．2パス探索アルゴリズムにおける高速な単語事後確率に基づく信頼度算出法．電子情報通信学会技術研究報告，SP2003-160, pp35–40, (2003).

8) 李晃伸．〔特別講演〕大語彙連続音声認識エンジン Julius の開発の進展．情報処理学会研究報告，2005-SLP-59-22, (2005).

9) Akinobu Lee, Tatsuya Kawahara, Kazuya Takeda, Masato Mimura, Atsushi Yamada, Akinori Ito, Katsunobu Itou, and Kiyohiro Shikano, "Continuous speech recognition consortium — an open repository for CSR tools and models —," in *Proc. IEEE LREC*, pp.1438–1441 (2002).

10) 李 晃伸，西村 竜一，山田 真士，鹿野 清宏．公共音声情報案内システム「たけまるくん」の運用および収集発話の分析，情報処理学会研究報告，2004-SLP-53-9, pp49-54 (2004).