

大語彙連続音声認識のための音素グラフ に基づく仮説制限法の検討

堀 貴明 岡 直生 加藤正治 伊藤彰則 好田正紀

山形大学工学部

〒992 米沢市城南4丁目3-16

E-mail : hori@ei5sun.yz.yamagata-u.ac.jp

あらまし 本報告では、大語彙連続音声認識 (Large Vocabulary Continuous Speech Recognition : LVCSR) のための高速な探索手法について検討し、効果的に探索空間を狭める新しい手法—音素グラフに基づく仮説制限法—を提案する。本手法は、認識の前処理として音素グラフを生成し、認識段階ではその音素グラフの情報を利用して仮説の展開を制限しながら最良の単語列を探索する。音素グラフによる仮説制限は、音素境界制限と Forward-Backward Pruning からなり、これらは探索空間の大幅な削減を可能にする。語彙サイズ 5000 の新聞記事読み上げ音声を用いた認識実験において、本手法が誤り率を増加させることなく処理時間の約 70% を削減可能であることを示す。

キーワード 音声認識, 大語彙連続音声認識, 隠れマルコフ網, 探索手法, 音素グラフ

A Study on A Phoneme-Graph-based Hypothesis Restriction for Large Vocabulary Continuous Speech Recognition

Takaaki HORI, Naoki OKA, Masaharu KATOH, Akinori ITO, and Masaki KOHDA

Faculty of Engineering, Yamagata University
4-3-16 Johnan, Yonezawa-shi, 992 Japan

Abstract In this paper, we study about fast search strategies for Large Vocabulary Continuous Speech Recognition (LVCSR), and propose a new method — a phoneme-graph-based hypothesis restriction, which effectually prunes the search space. In the proposed method, a phoneme graph is generated at the pre-processing stage, and then the best word sequence is searched while restricting expansion of hypotheses using the information of the phoneme graph at the main recognition stage. The phoneme-graph-based restriction consists of the limitation of phoneme boundaries and the Forward-Backward Pruning, which enable to reduce the search space dramatically. The proposed method was tested on a 5,000-word Japanese newspaper reading task. The experimental results show that this method can reduce about 70% of the elapsed time without any error increasing.

key words speech recognition, LVCSR, Hidden Markov Network, search strategy, phoneme graph

1 はじめに

高精度かつ高速な大語彙連続音声認識 (Large Vocabulary Continuous Speech Recognition : LVCSR) を実現しようという試みは近年ますます盛んになり、新聞記事読み上げ音声や放送ニュース音声の認識を目的とした多くの検討がなされている [1]。ここ数年、日本でも NTT の日経新聞を用いた研究 [2] を皮切りに日本語 LVCSR システムに関する検討が始まっており、情報処理振興事業協会 (IPA) による日本語新聞記事読み上げ音声データベースも整備され、研究が本格化の兆しを見せている [3]。

我々は、高精度な日本語 LVCSR システムの構築を目指し、これまで音響モデルの性能改善について検討を行ってきた。そして、状態クラスタリングによる HM-Net の構造決定法 [4] を提案し、本手法によって生成した HM-Net が従来の逐次状態分割法 (SSS) [5] や Tree-based Clustering [6] によるモデルよりも高い認識性能を有することを示した。

本報告では LVCSR のデコーダに注目し、高速な LVCSR を実現する認識アルゴリズムについて検討する。デコーダは、音響モデルによる音響照合と言語モデルによる制約を統合し、入力音声に最も合致する解 (認識結果) を探索する音声認識システムの心臓部である。従って、デコーダにどのような探索アルゴリズムを採用するかはシステムのターンアラウンド時間に大きく影響する。特に LVCSR の場合は探索空間が非常に大きくなるため、探索アルゴリズムの設計は重要な鍵となる。

認識処理を高速化する手法は数多く提案されているが、探索の過程でトリス上の当該節点までの累積スコアだけでなく当該節点以降の累積スコア (の推定値) も併せて考慮する方法が広く用いられる。当該節点以降のスコアを用いる方法は、一般に先読み (look-ahead) と呼ばれる。時間同期ビームサーチにおいて先読みを行うと、将来的に枝刈りされる仮説を早い段階で枝刈りすることができる [8][9][10]。また、スタックデコーダの場合は、見込みの少ない仮説の展開を抑制したり、次に展開する単語を予備選択するのに有効である [11][12][13][14]。

一方、音素や単語の境界を限定することによって探索空間を狭めるアプローチがある [10][14]。文献 [10] では、音節デコーディングによって音節境界候補を求め、それらを単語境界の制限に利用している。しかし、音節デコーディングによって得られる境界は第 1 位候補の音節系列に対する境界なので精度が十分とは言えず、境界に許容誤差を設けて候補を増やしている。そのため境界制限による処理量の大きな改善は得られていない。

本研究では、音素境界制限と先読みによる枝刈りを統

合的に扱う高速化手法を提案する。本手法は認識の前処理として音素グラフを生成し、この音素グラフに含まれる情報を利用して、音素境界制限と先読みによる枝刈り (Forward-Backward Pruning) を行う。音素グラフは複数の音素列候補を含んでいるため精度良く境界を限定することが可能であり、かつ、音素グラフ上を後向きに累積したスコアを先読みスコアとすることによって 2 音素程度の先読み効果を期待できる。

本手法を状態クラスタリングによる HM-Net と単語 N-gram を用いた LVCSR システムに実装し、語彙サイズ 5000 の新聞記事読み上げ音声により評価する。実験結果より、本手法が単語誤り率をほとんど増加させることなく処理時間の約 70% を削減可能であり、かつ、従来の 1 音素先読みによる高速化手法と比べて処理量削減の効果が大きいことを示す。

2 LVCSR システム

2.1 システムの構成

現在用いられる連続音声認識の過程は、情報理論に基づく確率的音声認識の枠組をとっている。この枠組に従えば、入力音声 O が与えられたときに、それが単語列 W である確率はベイズ則によって

$$P(W|O) = \frac{P(O|W)P(W)}{P(O)} \quad (1)$$

のように与えられる。この確率が最大になる単語列 \hat{W} を求めることが音声認識のゴールである。ここで、 $P(O|W)$ は音響モデル (Acoustic Model)、 $P(W)$ は言語モデル (Language Model) によって与えられる。 $P(O)$ は \hat{W} に無関係なので無視できる。本研究で構築した LVCSR システムもこのような枠組に基づいている。構築したシステムの構成を図 1 に示す。

システムには、音響モデルとして HM-Net、言語モデルとして単語 N-gram (bigram, trigram)、デコーダとして単語グラフを中間表現とする 2 パスサーチが採用されている。音素グラフ生成 (Phoneme Graph Generation) の部分は音素グラフに基づく仮説制限を行う場合のみ稼働するモジュールである (詳しくは 3 節で述べる)。まずベースラインのデコーダについて説明する。

2.2 ベースラインデコーダ

大語彙連続音声認識は探索空間が非常に大きいため、始めから複雑な言語モデルを使用すると、探索処理も複雑になり、処理量が増大してしまう。そこで、まず簡単な言語モデルを使用して探索空間を絞り、その後で複雑

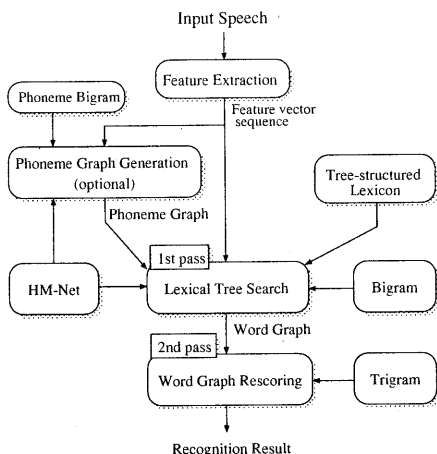


図 1: LVCSR システム.

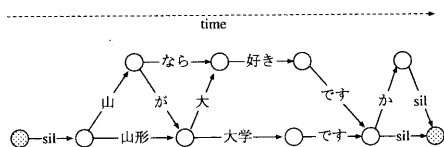


図 2: 単語グラフ.

な言語モデルを適用する手法が有効とされている。このような段階的探索をマルチパスサーチと呼ぶ。構築した大語彙連続音声認識システムのデコーダは、第1パスで HM-Net と bigram を用いて単語グラフを生成し、第2パスで trigram を用いて単語グラフから認識単語列を求める構成になっている。

単語グラフは認識単語列を段階的に絞り込む際の間表現として生成される。第1パスにおいて単語の音響尤度と始端・終端の集合を求め、グラフとして表現する。このグラフ中には多くの文候補が含まれるので、上位 N 個の文を求める N -best よりも効率的な表現と言える [7]。単語グラフの例を図 2 に示す。

2.2.1 単語グラフ生成 (第1パス)

時間同期ビームサーチと木構造辞書を用いた One-pass アルゴリズム (Lexical Tree Search[7]) をベースに単語グラフの生成を行う。

木構造辞書は単語の始端から発音が同じところまでの探索領域を共有する木構造ネットワークである。木構造辞書の例を図 3(a) に示す。探索は木の根から葉に

向かって行われ、葉で一つの単語を同定し、次の単語は根に戻って探索する。実際のシステムは、各ノードに triphone モデルを割り当て、かつ、単語間の継りも考慮するので、図 3(b) のような構造になっている。探索領域の共有は大語彙になればなるほどその効果を発揮する。しかし、単語の終端 (木の葉) に到達するまではその単語を同定できないため、言語尤度の適用が 1 単語分遅れてしまう。これは、仮説数の増加を招くので、木の各節点でそれ以降に到達できる葉の言語尤度の最大値を予測値として利用する factorization [8] を併用する。

単語グラフは One-pass アルゴリズムを拡張することによって生成される。One-pass アルゴリズムは各単語境界で最良の仮説のみを選択して探索を進めるので、基本的には第1位候補しか求めることができない。しかし、このとき最良仮説だけでなく第2位以降の仮説の情報 (単語の始端と終端、その間の音響尤度) を記録しておけば、探索終了後にそれらの情報から単語グラフを生成することができる。但し、単語の尤度や境界はそれ以前の単語の影響を受けるので、単語履歴毎に異なる候補として扱うべきであるが、そのようにすると候補の数が膨大になるので直前の単語のみに依存させる単語対近似を仮定する。従って、先行単語ごとに異なる木構造辞書 (のコピー) を用意する。

単語 bigram を用いて単語グラフを生成するアルゴリズムを以下に示す [7]。説明のために、まず次の量を定義する。

$\alpha_v(t, s) :=$ 先行単語が v の辞書木の状態 s における時刻 t までの局所最適パスの累積尤度

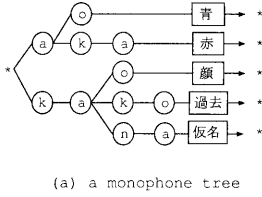
$B_v(t, s) :=$ 先行単語が v の辞書木の状態 s における時刻 t までの局所最適パスの開始時刻 (v の終了時刻)

$B_v(t, s)$ は、探索終了後に最尤単語列をトレースバックするためのバックポインタである。これら二つの量を用いて動的計画法の漸化式を表すと

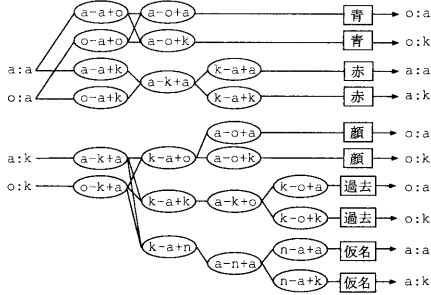
$$\begin{aligned} \alpha_v(0, s = 0) &= 1.0 \\ \alpha_v(t, s) &= \max_{\sigma} \{q(x_t, s|\sigma) \alpha_v(t-1, \sigma)\} \end{aligned} \quad (2)$$

$$\begin{aligned} B_v(0, s = 0) &= 0 \\ B_v(t, s) &= B_v(t-1, \sigma_v^{max}(t, s)) \end{aligned} \quad (3)$$

のようになる。ここで、 $s = 0$ は辞書木の根の状態、 $q(x_t, s|\sigma)$ は状態 s から s へ遷移し特徴ベクトル x_t を出力する音響尤度、 $\sigma_v^{max}(t, s)$ は (2) 式の右辺を最大にする σ を表す。



(a) a monophone tree



(b) a cross-word triphone tree

図 3: 木構造辞書.

単語 w の最終状態 S_w では最良の先行単語 v を

$$H(w; t) := \max_v \{p(w|v)\alpha_v(t, S_w)\} \quad (4)$$

のように選択することで単語 w の終端に到達した仮説はマージされる。ここで、 $p(w|v)$ は単語 v の後に単語 w が続く bigram 確率である。その後、

$$\alpha_w(t, s = 0) = H(w; t) \quad (5)$$

$$B_w(t, s = 0) = t \quad (6)$$

として、 t を 1 増やし先行単語を w とする辞書木の探索を進める。

仮説のマージにおいて、そこに到達した単語の情報

$$\text{単語境界} \quad \tau(t; v, w) := B_v(t, S_w) \quad (7)$$

$$\text{単語の音響尤度} \quad h(w; \tau, t) := \frac{\alpha_w(t, S_w)}{H(v; \tau)} \quad (8)$$

を保持することで、単語グラフのエッジが登録される。それ以降の探索は 1 位のみについて行われる。終端に達した最良仮説から、各時点で保持した単語仮説をトレースバックすることによって単語グラフが構築される。

2.2.2 枝刈り (Pruning)

第 1 パスにはビームサーチが用いられる。ビームサーチにおいて見込みの少なくなった仮説を枝刈りするために、以下の 3 種類の枝刈りが行われる。

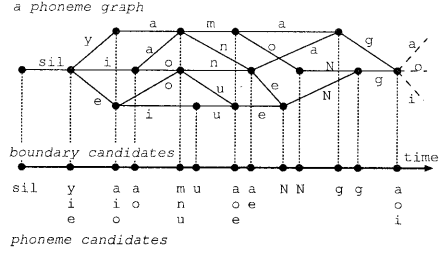


図 4: 音素グラフ内の境界候補と音素候補

(1) 単語内の枝刈り

$$\alpha_v(t, s) < f_{AC} \max_{v,s} \{\alpha_v(t, s)\} \quad (9)$$

であるとき仮説を枝刈りする。

(2) 単語間の枝刈り

$$\alpha_v(t, s = 0) < f_{LM} \max_v \{\alpha_v(t, s = 0)\} \quad (10)$$

であるとき仮説を枝刈りする。

(3) 仮説数による枝刈り

時刻 t に生き残っている仮説の上位 $MaxHyp$ 個を選択し、それ以外は枝刈りする。

枝刈りの閾値は一般に $0 < f_{AC} < f_{LM} < 1$ のように設定される。

2.2.3 リスコアリング (第 2 パス)

リスコアリングは、単語グラフ上の文候補の中から第 1 パスよりも複雑な言語モデルを用いて認識単語列を求める処理である。単語グラフのノードには単語境界の時刻 $\tau(t; v, w)$ が、エッジには単語の音響尤度 $h(w; \tau, t)$ が保持されているので、グラフ上を動的計画法によって時間方向に探索を進めることができる。エッジで音響尤度、単語境界で trigram 確率を累積していき、累積尤度最大となる単語列を認識結果として出力する。

3 音素グラフに基づく仮説制限法

音素グラフに基づく仮説制限法を提案する。音素グラフは音素の境界、各境界における音素候補およびその音響尤度を保持している (図 4)。これらの情報を適切に利用して第 1 パスの仮説数を削減することを検討する。音素グラフは単語グラフと同様のアルゴリズムによって生成することができる。このとき、音素 bigram と音素対近似を用いる。音素グラフの生成にかかる計算量は、大語彙の認識に比べればそれほど大きくないため、第 1 パスの前処理として音素グラフを生成することはさほ

ど問題にならない。

音素グラフに含まれる音素境界、音素候補、音響尤度を利用した仮説制限法として以下の手法を検討する。

3.1 音素境界制限

サーチの段階で、音素グラフ上の境界候補以外で音素間の遷移を許さない制限を加える。音素グラフは複数の音素列候補とその境界情報を保持しているため、音素グラフによる境界制限は、1位候補に基づく手法 [10] よりも精度良く真の境界を含むことが期待できる。但し、適切な音素境界は使用する音響モデルによって異なるので、本手法は音素グラフの生成と次のパスにおいて同じ音響モデルを用いることが前提となる。

3.2 Forward-Backward Pruning

音素グラフ中の音素候補以外の音素の展開を許さないという制限も考えられるが、音素グラフに正解音素列が100%含まれる訳ではないので、このような厳密な制限は認識誤りの増加を招く。そこで、本研究では厳密に音素候補を制限するのではなく、現時点までの仮説の尤度と音素グラフから求めた先読み尤度に基づいて展開を制限する。具体的には Forward-Backward Pruning [9] を導入する。Forward-Backward Pruning とは、始端から当該時刻までの Forward スコアと終端から当該時刻までの Backward スコア (の推定値) を併せて仮説を評価し、枝刈りを行う手法である。当該時刻以降のスコアを考慮することにより、見込みの少ない仮説を早い段階で枝刈りすることが可能となる。提案法では音素グラフを用いて Backward スコアを求める。

音素グラフを発声終了時刻 T から後向きに辿って、時刻 t に音素 p に至る Backward スコアを $\beta(t, p)$ 、音素グラフ内の音素 p の時刻 t から τ のエッジにおける音響尤度を $h(p; t, \tau)$ とする。このとき漸化式、

$$\begin{aligned} \beta(T, p) &= 1.0 \\ \beta(t, p) &= \max_{\tau > t} \left\{ h(p; t, \tau) \max_q \beta(\tau, q) \right\} \quad (11) \\ t, \tau &\in \text{音素グラフの境界候補} \end{aligned}$$

によって Backward スコアを求めることができる。しかし、音素グラフに含まれない音素に対しては Backward スコアが計算できないので、

$$\hat{\beta}(t, p) = \eta \min_x \beta(t, x) \quad (12)$$

をその推定値とする。ここで、 η はペナルティであり、 $0 < \eta < 1$ の定数とする。

各音素が triphone の場合は、Backward スコア $\beta(t, o-p+q)$ は以下のように求める (但し、 $o-p+q$ は先行音素が o 、中心音素が p 、後続音素が q の triphone を表す)。

if 時刻 t に始まる $o-p+q$ のエッジが存在する

$$\hat{\beta}(t, o-p+q) = \beta(t, o-p+q) \quad (13)$$

elseif 中心音素が一致するエッジが存在する

$$\hat{\beta}(t, o-p+q) = \eta_1 \min_{x,y} \beta(t, x-p+y) \quad (14)$$

else

$$\hat{\beta}(t, o-p+q) = \eta_2 \min_x \beta(t, x) \quad (15)$$

但し、 η_1, η_2 は $0 < \eta_2 < \eta_1 < 1$ の定数とする。Backward スコアの計算は音素単位で進むので、その処理量は無視できるほど小さい。

時刻 t に音素 p から q へ遷移する Forward-Backward スコアは

$$\gamma_v(t, p, q) = \alpha_v(t, S_p) \hat{\beta}(t+1, q) \quad (16)$$

となる。但し、 S_p は音素 p の HMM の最終状態を表す。本手法における Forward-Backward スコアの概念を図 5 に示す。音素グラフに対応する音素が存在する場合は (a)、存在しない場合は (b) のように求められる。Forward-Backward スコアが

$$\gamma_v(t, p, q) < f_{FBP} \max_{v,x,y} \gamma_v(t, x, y) \quad (17)$$

であるとき、音素 p から q への遷移は制限される。ここで、 f_{FBP} は枝刈り閾値で $0 < f_{FBP} < 1$ の定数である。

本手法は、triphone モデルを用いて音素グラフを生成することにより 2 音素程度の先読み効果が期待できる。それは、Backward スコア $\hat{\beta}(t, o-p+q)$ が音素 o の後に p, q が続いて終端に達するスコアであるため、そのスコアは未来の 2 音素に依存する。従って従来の 1 音素先読み手法 [8] よりも仮説削減の効果が大きいことが期待できる。

また、本手法は文献 [13][14][15] にあるような音素 (音節) グラフまたは音素 (音節) トレリスから直接単語列を探索しようとするアプローチとは異なっている。ビームサーチによって音素グラフを生成した場合、その中に正解音素列が 100% 含まれるとは限らない。発音表記通りに明瞭に単語が発声される場合はよいが、そうでない場合はしばしば音素グラフから正解音素が脱落することがある。1 音素脱落するだけでも、音素グラフから単語列を探索するときには重大な誤りの原因になり得る。文献 [14] では全探索で音素グラフを求めているが、音

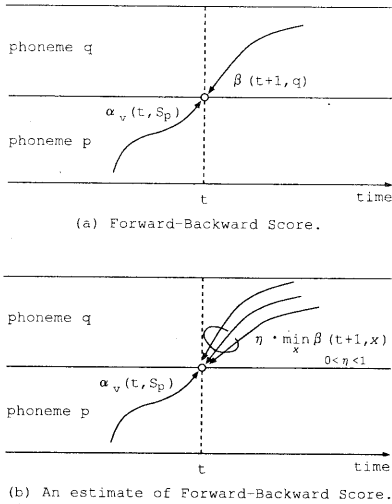


図 5: Forward-Backward Pruning

表 1: 音声分析条件

標準化周波数	16 kHz
量子化	16 bit
フレーム長	32 ms
フレーム周期	8 ms
分析窓	ハミング窓
高域強調	$1 - z^{-1}$
特徴ベクトル	1~12 次の LPC メルケプストラム係数と対数パワー, およびその一次と二次の回帰係数 (計 39 次元)
正規化	発話毎のケプストラム平均正規化

素環境に依存する大規模・高精度な音響モデルを用いる場合には非現実的である。提案法は、音素グラフの中から解を探索するのではなく、その情報を利用して単語列の探索における仮説数を削減する。従って、音素グラフから脱落した音素を含む単語が認識不能になることはない。

4 大語彙連続音声認識実験

4.1 音声資料・分析条件

音声資料として、日本音響学会の新聞記事読み上げ音声コーパス (JNAS), 男性 102 名が発声した新聞記事と音素バランス文, 計 15732 文を用いる。音声分析条件を表 1 に示す。

4.2 HM-Net の生成

音素カテゴリーは 34 音素 + 無音とする。各音素 3 状態の音素環境独立 HMM を並列に接続した 102 状態の HM-Net を初期モデルとして、状態クラスタリングによる構造決定法 [4] を用いて 2000 状態まで分割を行う。無音モデルは 3 状態の音素環境独立 HMM とする。構造決定後、各状態の出力確率分布を 16 混合分布に再構成してパラメータを再学習する。

4.3 N-gram の生成

言語モデルとして、JNAS 付属の毎日新聞 91 年 1 月 ~ 94 年 9 月の記事より推定された N-gram (bigram, trigram) を用いる。また、音素グラフを生成する場合に使用する音素 bigram は、音響モデルの学習に用いた音声データ 15732 文の発音表記から推定する。

4.4 評価方法

認識タスクは語彙サイズ 5000, 未知語なしの新聞記事読み上げ音声のディクテーションである。各単語は形態素単位とし、漢字仮名まじり表記と品詞番号の対で表される。従って、同じ漢字仮名表記であっても品詞が異なる場合は別の単語として扱われる。

評価データとして JNAS の学習データとは異なる男性 10 名 (M006, M014, M017, M021, M026, M089, M102, M109, M115, M122) が各 10 文発声した計 100 文を用いる。評価は 100 文の認識結果に対する単語誤り率

$$WER[\%] = \frac{S + I + D}{N} \times 100 \quad (18)$$

で行う。ここで、 S, I, D は置換, 挿入, 脱落誤りの数, N は正解単語列の単語数を表す。誤り率を求める際の単語の一致に関する基準は、漢字仮名まじりの一致と読みだけの一致の 2 通りを考える。品詞の違いは考慮しない。従って、漢字仮名まじり評価の場合は漢数字とアラビア数字の違いも誤りとなる。読みだけの評価の場合は同音異義語も正解となる。但し、形態素の区切りが異なっている場合は誤りとする。例えば、“20 日” に対して “20 / 日” のように、“20” と “日” が離れて認識された場合は、置換誤り 1, 挿入誤り 1 としてカウントされる。

使用計算機は汎用ワークステーション HP-C200 であり、処理時間は 1 文当たりの CPU タイムで評価する。評価文 100 文の平均発声時間は 4.1 秒である。

4.5 実験結果と考察

(1) 実験結果

ビーム幅 ($-\log(\text{ビーム閾値})$) を 120~150 に変えて生成した音素グラフの性能と、その音素グラフを仮説制限に適用した認識実験の結果を表 2 に示す。表の上段の、境界候補、ヒット率、PGD、GER、処理時間は、それぞれ、境界候補数の全フレーム数に対する割合、境界候補に真の境界が含まれる割合、音素グラフ密度、音素グラフ誤り率、1 文当たりの音素グラフ生成に要する処理時間を表している。ここで、真の境界とは同じ音響モデルで正解文の Viterbi アライメントを行った場合の音素境界を意味する。

一方、表の下段はこれらの音素グラフを仮説制限に適用した場合の単語誤り率 (漢字仮名評価, 読み評価) と 1 文当たりの処理時間 [sec] を示している。括弧内は第 1 パスの単語誤り率である。単語誤り率は、システムのパラメータ (言語尤度の重み, 挿入ペナルティ, ビーム幅 f_{AC} , f_{LM} , $MaxHyp$, Forward-Backward Pruning のペナルティ η_1, η_2) を様々に変化させた場合の最適値である。但し、Forward-Backward Pruning の閾値は $f_{FBP} = f_{AC}$ としている。処理時間には音素グラフ生成の処理時間も含まれている。

(2) 音素境界制限の効果

音素グラフによる仮説制限を適用しない場合 (baseline) は、漢字仮名評価で 5.7%, 読み評価で 3.7% という低い誤り率が得られているものの、1 文当たりの処理に 138 秒を費している。これに対し音素境界制限を適用すると、音素グラフ生成のビーム幅が 140 や 150 の場合で誤り率をほとんど増加させることなくそれぞれ 53%, 49% の処理量が削減される。これは、音素グラフによる境界推定精度が高いことを示しており、表 2 上段に見られるように境界候補を 22.9%, 27.3% に制限しつつ真の境界を 98.0%, 98.5% の割合で含むことが分かる。

(3) Forward-Backward Pruning の効果

音素境界制限に Forward-Backward Pruning (FBP) を加えると、音素グラフ生成のビーム幅が 140 や 150 の場合で処理量は baseline に対しそれぞれ 71%, 69% も削減される。しかも、誤り率の増加は全く見られない。これらの結果より、提案した音素グラフに基づく仮説制限法が誤りを増加させることなく処理量の 70% を削減できる有効な手法であることが示された。

(4) 1 音素先読みとの比較

提案法の有効性を示すために、従来の 1 音素先読みによる高速化手法 [8] と比較する。文献 [8] と同様に、Fast Match モデルを作成し、当該フレームから Δt フレーム先までの音声に対する照合を行い、これを当該フレームまでの尤度と併せて仮説の枝刈りに判定に利用する。1 音

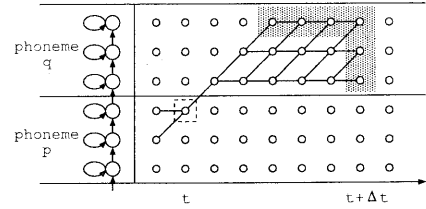


図 6: 1 音素先読み。

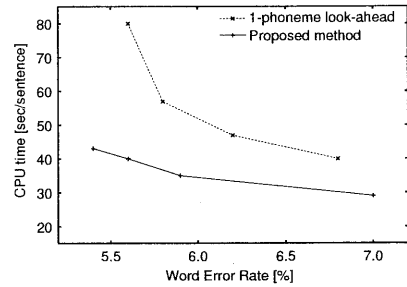


図 7: 提案法と 1 音素先読みの比較。

素先読みの概念を図 6 に示す。Fast Match モデルとして JNAS の 15732 文から作成した音素環境独立、混合数 4 の HMM を利用する。 Δt は音素の平均継続時間から 8 フレームとする。

1 音素先読みにおいて枝刈り閾値を変化させた場合の認識結果と表 2 の音素グラフによる仮説制限 (境界制限 + FBP) の結果を図 7 にまとめる。図は単語誤り率 (漢字仮名評価) と処理時間の関係を示している。図 7 より、1 音素先読みによる枝刈りは baseline と比べて誤り率 5.6% の場合で 42%, 誤り率 5.8% の場合で 59% の処理量削減の効果がある。

提案法と 1 音素先読み手法を比べると、誤り率 5.6% の場合で 1 : 2, 誤り率 5.8~5.9% で約 7 : 11 の処理時間となっている。この差は、1 音素先読みの場合は、音素環境独立 HMM を用いているために仮説を十分に絞り込めないこと、境界の曖昧性が残されているために無駄な仮説の展開が行われていること等によるものと考えられる。提案法は、音素グラフの生成に要するオーバーヘッドはやや大きいものの、境界制限と 2 音素の先読み効果によって従来の 1 音素先読み手法よりも無駄の少ない探索を実現している。

5 おわりに

大語彙連続音声認識システムのための音素グラフに基づく仮説制限法を提案し、新聞記事読み上げ音声を用

表 2: 音素グラフによる仮説制限を用いた実験結果

		音素グラフ生成のビーム幅 (尤度幅)			
		150	140	130	120
音素グラフ の性能	境界候補 (%)	27.3	22.9	19.0	15.9
	ヒット率 (%)	98.5	98.0	97.3	96.2
	PGD	14.8	9.8	6.4	4.3
	GER (%)	1.7	2.0	2.4	3.3
	処理時間 [sec]	23	19	17	14
単語誤り率 [%] (漢字仮名評価)	baseline	5.7 (6.5)			
	境界制限	5.4 (7.0)	5.8 (7.4)	6.0 (7.9)	7.1 (8.5)
	境界制限+FBP	5.4 (7.1)	5.6 (7.5)	5.9 (7.9)	7.0 (8.5)
単語誤り率 [%] (読み評価)	baseline	3.7 (4.1)			
	境界制限	3.5 (4.7)	3.9 (5.0)	4.2 (5.5)	5.2 (6.1)
	境界制限+FBP	3.5 (4.8)	3.7 (5.1)	4.1 (5.5)	5.2 (6.1)
1文当たりの 処理時間 [sec] (音素グラフ生成を含む)	baseline	138			
	境界制限	71	65	57	49
	境界制限+FBP	43	40	35	29

括弧内は第1パスの誤り率 [%] を表す

いて評価した。音素グラフによる仮説制限は、境界制限だけで50%程度の処理量削減が可能であり、Forward-Backward Pruning と併用することによって誤りを増加させることなく全体の70%もの処理量を削減可能であることが示された。本研究において構築したLVCSRシステムの現時点の性能は、誤り率5.4%で実時間の11倍、誤り率5.9%で実時間の9倍、誤り率7%で実時間の7倍の処理時間を要する。今後は、認識精度、処理時間を更に改善していくとともに、より大語彙のタスクに対する評価も行っていく予定である。

参考文献

- [1] S.J.Young : A review of large-vocabulary continuous-speech recognition, *IEEE Signal Processing magazine*, Vol.13, No.5, pp.45-57 (1996).
- [2] 松岡 達雄, 大附 克年, 森 岳至, 古井 貞熙, 白井 克彦 : 新聞記事データベースを用いた大語彙連続音声認識, 信学論 (D-II), J79-D-II, 12, pp.2125-2131 (1996).
- [3] 河原 達也, 李 兪伸, 小林 哲則, 武田 一哉, 峯松 信明, 伊藤 克亘, 伊藤 彰則, 山本 幹雄, 山田 篤, 宇津呂 武仁, 鹿野 清宏 : 日本語ディクテーション基本ソフトウェア (97年度版) の性能評価, 情報処理学会研究報告, 98-SLP-21-10 (1998).
- [4] 堀 貴明, 加藤 正治, 伊藤 彰則, 好田 正紀 : 状態クラスタリングによる HM-Net の構造決定法の検討, 信学論 (D-II), J81-D-II, 10, pp.2239-2248 (1998).
- [5] 鷹見 淳一, 嵯峨山 茂樹 : 逐次状態分割法による隠れマルコフ網の自動生成, 信学論 (D-II), J76-D-II, 10, pp.2155-2164 (1993).
- [6] S.J.Young, J.J.Odell and P.C.Woodland : Tree-based state tying for high accuracy acoustics modeling, *Proc. ARPA Human Language Technology Workshop*, pp.307-312 (1994).
- [7] S.Ortmanns and H.Ney : A word graph algorithm for large vocabulary continuous speech recognition, *Computer Speech and Language*, Vol.11, No.1, pp.43-72 (1997).
- [8] S. Ortmanns, A. Eiden, H. Ney, and N. Coenen : Look-ahead techniques for fast beam search, *Proc. ICASSP'97*, Vol.4, pp.1783-1786 (1997).
- [9] S. Austin, R. Schwartz, and P. Placeway : The forward-backward search algorithm, *Proc. ICASSP'91*, Vol.1, pp.697-700 (1991).
- [10] 甲斐 充彦, 廣瀬 良文, 中川 聖一 : N-gram 言語モデルと効率的探索手法を用いた大語彙連続音声認識システムの検討, 信学技報 SP97-99, pp.31-38 (1998).
- [11] D.B.Paul : New developments in the Lincoln stack-decoder based large-vocabulary CSR system, *Proc. ICASSP'95*, Vol.1, pp.45-48 (1995).
- [12] P. S. Gopalakrishnan, L. R. Bahl, and R. L. Mercer : A tree search strategy for large-vocabulary continuous speech recognition, *Proc. ICASSP'95*, Vol.1, pp.572-575 (1995).
- [13] E. -F. Huang, F. K. Soong, and H. -C. Wang : The use of tree-trellis search for large-vocabulary Mandarin polysyllabic word speech recognition, *Computer Speech and Language*, Vol.8, No.1, pp.39-50 (1994).
- [14] Z. Li, G. Boulianne, P. Labute, M. Barszcz, H. Garudadri, and P. Kenny : Bi-directional graph search strategies for speech recognition, *Computer Speech and Language*, Vol.10, No.4, pp.295-321 (1996).
- [15] T. -H. Ho, K. -C. Yang, K. -H. Huang, L. -S. Lee : Improved search strategy for large vocabulary continuous Mandarin speech recognition, *Proc. ICASSP'98*, Vol.2, pp.825-828 (1998).