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†

本論文では，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が示された．

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†

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.

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さかんになり，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 の日本経済新聞の記事データを用いた研究[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，日本音響学会の新聞記事読み上げ音声データベースが整備され，情報処理振興事業協会（IPA）の大語彙連続音声認識プロジェクトがスタートするなど，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．

本論文では 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．したがって，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

† 山形大学工学部
Faculty of Engineering, Yamagata University

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，探索の過程でトレ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）もあわ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[10)~12)]．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したりするのに有効である[13)~16)]．

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[12),16)]．文献12)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いえず，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．

本論文では，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みをとっている．この枠組みに従えば，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

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

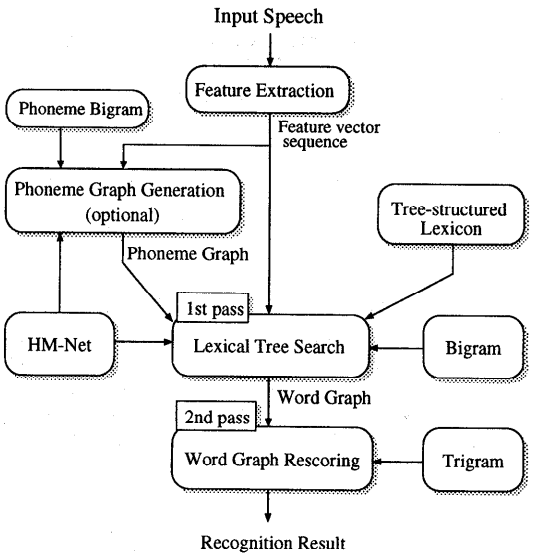


図1 LVCSRシステム
Fig. 1 The LVCSR system.

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みに基づいている．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章で述べる）．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中

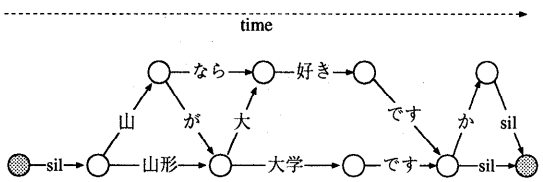


図2 単語グラフ
Fig. 2 A word graph.

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いえる[8]．単語グラフの例を図2に示す．

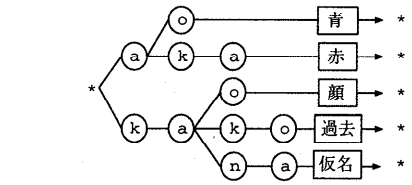
2.2.1 単語グラフ生成（第1パス）

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),8)]をベースに単語グラフの生成を行う．

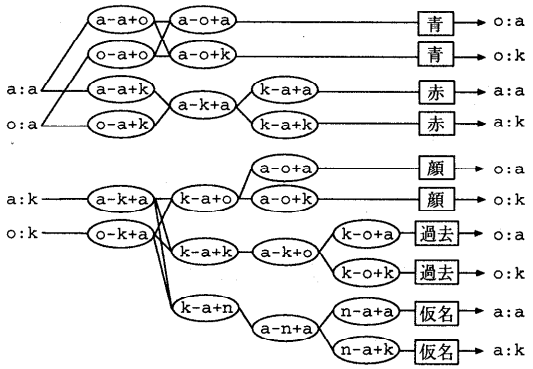
木構造辞書は単語の始端から発音が同じところまでの探索領域を共有する木構造ネットワークである．木構造辞書の例を図3(a)に示す．探索は木の根から葉に向かって行われ，葉で1つの単語を同定し，次の単語は根に戻って探索する．実際のシステムは，各ノードに triphone モデルを割り当て，かつ，単語間のつながりも考慮するので，図3(b)のような構造になっている．探索領域の共有は大語彙になればなるほどその効果を発揮する．しかし，単語の終端（木の葉）に到達するまではその単語を同定できないため，言語尤度の適用が1単語分遅れてしまう．これは，仮説数の増加を招くので，木の各節点でそれ以降に到達できる葉の言語尤度の最大値を予測値として利用する factorization[9)]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．ただし，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，単語履歴ごとに異なった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．したがって，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



(a) a monophone tree



(b) a cross-word triphone tree

図3 木構造辞書
Fig. 3 Tree-structured lexicon.

リズムを以下に示す[8)]．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2つの量を用いて動的計画法の漸化式を表すと

$$\alpha_v(0,s_0) = 1.0$$
$$\alpha_v(t,s) = \max_{\sigma}\{q(x_t,s|\sigma)\alpha_v(t-1,\sigma)\} \quad (2)$$
$$B_v(0,s_0) = 0$$
$$B_v(t,s) = B_v(t-1,\sigma_v^{max}(t,s)) \quad (3)$$

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 σ から s へ遷移し特徴ベクトル x_t を出力する音響尤度，$\sigma_v^{max}(t,s)$ は式(2)の右辺を最大にする σ を表す．

そして，時刻 t に単語 w の終端に到達する最良仮説の尤度 $H(w;t)$ は

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

のように計算され，このとき右辺を最大にする先行単語 v を選択することによって w の終端に到達した仮説はマージされる．ここで，S_w は単語 w の最終状態，$p(w|v)$ は単語 v の後に単語 w が続く bigram 確率を表す．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

単語境界 $\tau(t;v,w) := B_v(t,S_w)$ (7)

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

を保持することで，単語グラフのアークが登録される．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．

(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，アーク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．アークで音響尤度，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．

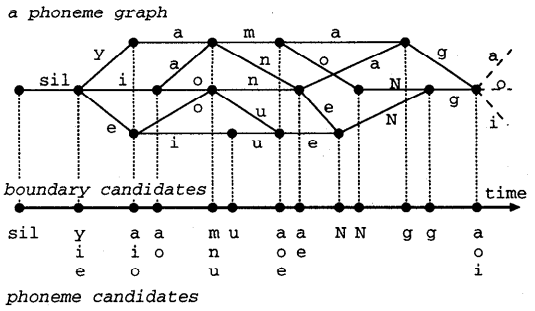


図 4 音素グラフ内の境界候補と音素候補

Fig. 4 Boundary and phoneme candidates in a phoneme graph.

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[12]よりも精度良く真の境界を含むことが期待できる．ただし，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%含まれるわけではないので，このような厳密な制限は認識誤りの増加を招く．そこで，本論文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[11]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）をあわせて仮説を評価し，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 から後向きにたどって，時刻 t に音素 q に至る Backward スコアを $\beta(t,q)$，音素グラフ内の音素 q の時刻 t から τ のアークにおける音響尤度を $h(q;t,\tau)$ とする．このとき漸化式，

$$\beta(T,q)=1.0$$

$$\beta(t,q)=\max_{\tau>t}\left\{h(q;t,\tau)\max_{r}\beta(\tau,r)\right\} \quad (11)$$

$$t,\tau\in\text{音素グラフの境界候補}$$

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，

$$\hat{\beta}(t,q)=\eta\min_{x}\beta(t,x) \quad (12)$$

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．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．

各音素が triphone の場合は，Backward スコア $\hat{\beta}(t,\mathrm{p-q+r})$ は以下のように求める（ただし，p−q+r は先行音素が p，中心音素が q，後続音素が r の triphone を表す）．

if 時刻 t に始まる p − q + r のアークが存在する

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

elseif 中心音素が一致するアークが存在する

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

else

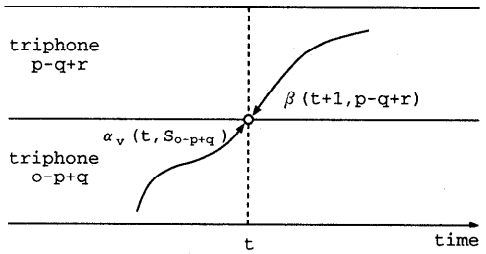
$$\hat{\beta}(t,\mathrm{p-q+r})=\eta_2\min_{x,y,z}\beta(t,x-y+z) \quad (15)$$

ただし，η_1，η_2 は $0<\eta_2<\eta_1<1$ の定数とする．

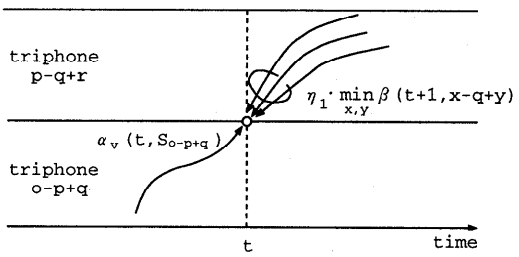
時刻 t に triphone o − p + q から p − q + r へ遷移する Forward-Backward スコアは

$$\gamma_v(t,\mathrm{o-p+q},\mathrm{p-q+r})=\alpha_v(t,S_{\mathrm{o-p+q}})\hat{\beta}(t+1,\mathrm{p-q+r}) \quad (16)$$

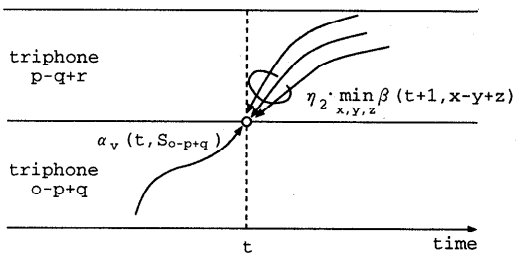
となる．ただし，$S_{\mathrm{o-p+q}}$ は triphone o − p + q の HMM の最終状態を表す．本手法における Forward-Backward スコアの概念を図 5 に示す．音素グラフに



(a) Forward-Backward Score.



(b) An estimate of Forward-Backward Score when arcs with the same center phoneme exist.



(c) An estimate of Forward-Backward Score when no corresponding arc exists.

図 5 前向き後向きスコア
Fig. 5 Forward-Backward score.

対応する triphone が存在する場合は (a)，存在しない場合は (b) または (c) のように求められる．Forward-Backward スコアが

$$\gamma_v(t,\mathrm{o-p+q},\mathrm{p-q+r}) < f_{FBP}\max_{v,x,y,z,z'}\gamma_v(t,x-y+z,y-z+z') \quad (17)$$

であるとき，triphone o − p + q から p − q + r 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 $\hat{\beta}(t,\mathrm{p-q+r})$ が音素 p の後に q，r 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．したがって，従来の 1 音素先読み手法[10]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15)~17)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．発音表記どお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うる．文献16)では全探索で音素グラフを求めているが，音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．したがって，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．

表1 音声分析条件
Table 1 The condition of speech analysis.

標本化周波数	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	発話ごとのケプストラム平均正規化

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．したがって，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，各10文，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

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

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．したがって，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．ただし，形態素の区切りが異なっている場合は誤りとする．たとえば，“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（クロック周波数200 MHz，SPECint95 = 14.3，SPECfp95 = 21.4）であり，処理量は1文あたりのCPUタイム，および，1フレームあたりの平均仮説数（モデル数）で評価する．評価文100文の平均発声時間は3.9秒である．

4.5 実験結果と考察

(1) 実験結果

ビーム幅（$-\log f'_{AC}$）を130~160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．ただし，f'_{AC}は音素グラフ生成時のビーム閾値で，第1パスにおける式(9)のf_{AC}に相当する．

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（正解1音素あたりの音素グラフ内の音素候補数），音素グラフ誤り率（音素グラフ中で最も正解に近い音素列に対する音素誤り率），1文あ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

表2 音素グラフによる仮説制限を用いた実験結果
Table 2 Experimental results using phoneme-graph-based hypothesis restriction.

		音素グラフ生成のビーム幅 ($-\log f'_{AC}$)			
		160	150	140	130
音素グラフの性能	境界候補 [%]	33.1	28.1	23.7	19.8
	ヒット率 [%]	99.0	98.6	98.2	97.5
	PGD	24.2	15.8	10.3	6.8
	GER [%]	0.7	0.9	1.2	1.6
	処理時間 [sec]	26 (1003)	23 (827)	19 (678)	17 (552)
単語誤り率 [%] (漢字仮名評価)	baseline	5.7 (8.1)			
	境界制限	5.7 (8.2)	5.8 (8.5)	6.3 (8.9)	6.7 (9.1)
	境界制限+FBP	5.8 (8.4)	5.8 (8.7)	6.0 (9.0)	6.3 (9.2)
単語誤り率 [%] (読み評価)	baseline	3.8 (6.4)			
	境界制限	3.8 (6.4)	3.9 (6.6)	4.1 (7.0)	4.5 (7.1)
	境界制限+FBP	3.9 (6.6)	3.9 (6.7)	4.0 (7.1)	4.3 (7.2)
1文あたりの処理時間 [sec] (音素グラフ生成を含む)	baseline	146 (6834)			
	境界制限	111 (3670)	94 (3079)	84 (2548)	80 (2086)
	境界制限+FBP	50 (1565)	44 (1347)	40 (1151)	35 (979)

イメントを行った場合の音素境界を意味する．また，処理時間の欄の括弧内は，音素グラフ生成時の1フレームあたりの平均仮説数（モデル数）を示している．

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文あたりの処理時間 [sec] を示している．誤り率の欄の括弧内は第1パスの単語誤り率，処理時間の欄の括弧内は第1パスにおける1フレームあたりの平均仮説数（モデル数）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．ただし，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おり，平均仮説数にも音素グラフ生成時の仮説数が加えてある．

(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 3.8%という低い誤り率が得られているものの，1文あたりの処理に 146 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 150 や 160 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 36%，24%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 28.1%，33.1%に制限しつつ真の境界を 98.6%，99.0%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 150 や 160 の場合で処理量は baseline に対しそれぞれ 70%，66%も削減される．しかも，誤り率の増加はほとんど見られない．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ほとんど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．

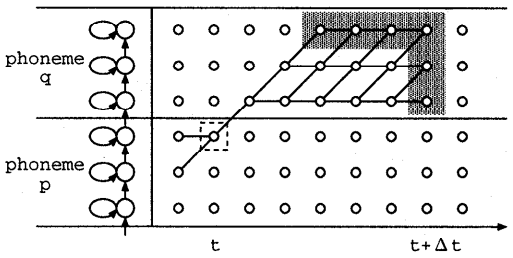


図6 1音素先読み
Fig. 6 1-phoneme look-ahead.

(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[10]と比較する．文献 10) 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あわせて仮説の枝刈り判定に利用する．1音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

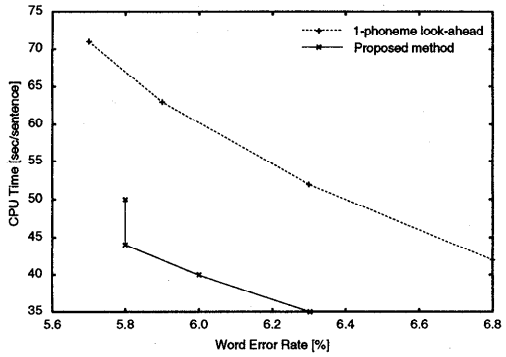


図7 提案法と1音素先読みの比較

Fig. 7 The comparison between the proposed method and 1-phoneme look-ahead.

と比べて誤りを増加させずに50%程度の処理量削減の効果があり，これは文献10)の結果とも一致する．提案法と1音素先読み手法を比べると，同程度の誤り率で提案法と1音素先読みの処理時間は約2:3の割合となっている．この差は，1音素先読みの場合は，先読みに音素環境独立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．

しかし，オンラインの音声認識の場合，提案法は発声開始とともに前向き処理で音素グラフを生成するが，発声が終了するまで第1パスを開始できないことを考慮する必要がある．表2より，音素グラフの生成とそれ以降の計算に要する処理時間が約1:1であることから，提案法が実時間の2倍以上の処理時間を要する状況では，本論文で述べたような計算量の削減効果を期待できる．それよりも高速な場合には音素グラフ生成時にCPUを100%使用できなくなるためその効果は徐々に小さくなる．提案法が実時間の1倍で動作する状況では，発声開始から結果が出力されるまでの応答時間は実時間の1.5倍を要し，これは，図7の関係から，1音素先読みの場合の応答時間とほぼ同等となる．さらに高速な場合は1音素先読みの方が応答時間が短くなる．以上のことから，オンライン認識の場合，1音素先読みに対する提案法の優位性は，提案法が実時間の1倍以上の処理時間を要する場合に現れ，2倍以上の処理時間を要する場合には本論文で述べた削減効果がそのまま現れる．

このような問題を解決するために，入力音声を数十フレーム幅のブロックに分割して，ブロックごとに処理を進める方法がある[16]．提案法に導入する場合は，ブロックごとに音素グラフを生成して仮説制限を用いたビームサーチ（第1パス）を行う手順になる．ただし，各ブロックの終端部分でBackwardスコアの精度が落ちるのを防ぐために，ブロックをオーバラップさせて処理する必要がある．この手法によって提案法の第1パスを開始するまでの処理ディレイはブロック幅程度に抑えられるので，オンラインの音声認識にも対応可能になることが期待できる．提案法に対するブロック処理の導入は今後の検討課題である．

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，境界制限だけでも35%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ほとんど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．本論文において構築したLVCSRシステムの現時点の性能は，誤り率5.8%で処理量が実時間の11倍，誤り率6.3%で処理量が実時間の9倍である．今後は，認識精度，処理時間をさらに改善していくとともに，オンライン認識への対応や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

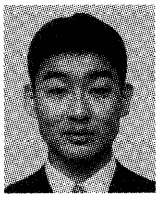
1) Young, S.J.: A review of large-vocabulary continuous-speech recognition, *IEEE Signal Processing magazine*, Vol.13, No.5, pp.45–57 (1996).
2) 松岡達雄，大附克年，森 岳至，古井貞熙，白井克彦：新聞記事データベースを用いた大語い連続音声認識，信学論，J79-D-II, 12, pp.2125–2131 (1996).
3) 河原達也，李 晃伸，小林哲則，武田一哉，峯松信明，伊藤克亘，伊藤彰則，山本幹雄，山田 篤，宇津呂武仁，鹿野清宏：日本語ディクテー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，信学技報SP97-75, pp.47–52 (1997).
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，J76-D-II, 10, pp.2155–2164 (1993).
6) Young, S.J., Odell, J.J. and Woodland, P.C.:

Tree-based state tying for high accuracy acoustics modeling, *Proc. ARPA Human Language Technology Workshop*, pp.307–312 (1994).
7) Odell, J.J., Valtchev, V., Woodland, P.C. and Young, S.J.: A one pass decoder design for large vocabulary recognition, *Proc. ARPA Human Language Technology Workshop*, pp.405–410 (1994).
8) Ortmanns, S. and Ney, H.: A word graph algorithm for large vocabulary continuous speeh recognition, *Computer Speech and Language*, Vol.11, No.1, pp.43–72 (1997).
9) Ortmanns, S., Ney, H. and Eiden, A.: Language-model look-ahead for large vocabulary speech recognition, *Proc. ICSLP'96*, Vol.4, pp.2095–2098 (1996).
10) Ortmanns, S., Eiden, A., Ney, H. and Coenen, N.: Look-ahead techniques for fast beam search, *Proc. ICASSP'97*, Vol.4, pp.1783–1786 (1997).
11) Austin, S., Schwartz, R. and Placeway, P.: The forward-backward search algorithm, *Proc. ICASSP'91*, Vol.1, pp.697–700 (1991).
12) 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).
13) Paul, D.B.: New developments in the Lincoln stack-decoder based large-vocabulary CSR system, *Proc. ICASSP'95*, Vol.1, pp.45–48 (1995).
14) Gopalakrishnan, P.S., Bahl, L.R. and Mercer, R.L.: A tree search strategy for large-vocabulary continuous speech recognition, *Proc. ICASSP'95*, Vol.1, pp.572–575 (1995).
15) Huang, E.-F., Soong, F.K. and Wang, H.-C.: 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).
16) Li, Z., Boulianne, G. Labute, P., Barszcz, M., Garudadri, H. and Kenny, P.: Bi-directional graph search strategies for speech recognition, *Computer Speech and Language*, Vol.10, No.4, pp.295–321 (1996).
17) Ho, T.-H., Yang, K.-C., Huang, K.-H. and Lee, L.-S.: Improved search strategy for large vocabulary continuous Mandarin speech recognition, *Proc. ICASSP'98*, Vol.2, pp.825–828 (1998).

(平成 9 年 10 月 13 日受付)
(平成 11 年 2 月 8 日採録)

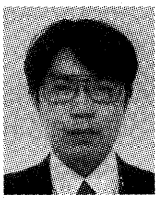
堀　貴明

平成 6 年山形大学工学部電子情報工学科卒業．平成 8 年同大学大学院博士前期課程修了．現在，同大学院博士後期課程在学中．音声認識の研究に従事．電子情報通信学会，日本音響学会各会員．



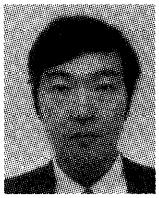
岡　直生

平成 10 年山形大学工学部電子情報工学科卒業．現在，同大学大学院博士前期課程在学中．音声認識の研究に従事．



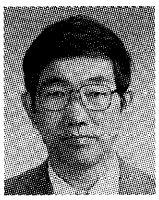
加藤　正治（正会員）

平成 3 年山形大学工学部情報工学科卒業．平成 5 年同大学大学院修士課程修了．現在，山形大学工学部電子情報工学科助手．音声認識の研究に従事．電子情報通信学会，日本音響学会各会員．



伊藤　彰則（正会員）

昭和 61 年東北大学工学部通信工学科卒業．平成 3 年同大学大学院博士課程修了．同年同大学情報処理教育センター助手．現在，山形大学工学部電子情報工学科講師．工学博士．音声認識における言語処理の研究に従事．電子情報通信学会，日本音響学会各会員．



好田　正紀（正会員）

昭和 40 年名古屋大学工学部電子工学科卒業．昭和 42 年同大学大学院修士課程修了．同年日本電信電話公社電気通信研究所入社．昭和 62 年山形大学工学部情報工学科教授．音声認識を主とする音声情報処理の研究に従事．工学博士．電子情報通信学会，日本音響学会，人工知能学会，言語処理学会，信号処理研究会各会員．