音声言語情報処理 29－34
(1999. 12. 21)

音素グラフに基づく仮説制限法を用いた大語彙連続音声認識の検討

岡 直生　加藤正治　伊藤彰則　好田正紀

山形大学工学部
〒992-8510 米沢市城南 4 丁目 3-16
E-mail : naoki@ei5sun.yz.yamagata-u.ac.jp

あらまし 本報告では,大語彙連続音声認識のための高速な手法ー音素グラフに基づく仮説制限法ーについて検討する. 提案法は, 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, 効率的に探索手法を狭める手法であり,5000 語の語彙タスクにおいて誤り率を増加させることなく処理時間の約 70%が削減可能であると報告してきた.
本報告では, 音素グラフ生成の際に,Fast Match モデルを用いて 1 音素先読みを行うことで, 音素グラフ生成自体の処理量削減をはかる. さらに, 言語モデルにおける学習データの違いによる認識結果の違いについても検討する. 語彙サイズ 20000 の新聞記事読み上げ文音声で評価し, 音素グラフに基づく仮説制限法の導入は, 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 60%を削減できることを示す.
キーワード　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, 1 音素先読み

Study on Large Vocabulary Continuous Speech Recognition with a phoneme graph based hypothesis restriction

Naoki OKA,　Masaharu KATOH,　Akinori ITO,　Masaki KOHDA

Faculty of Engineering, Yamagata University
4-3-16 Johnan,Yonezawa-shi, 992-8510 Japan

Abstract　In this paper, we study about fast search strategies for large vocabulary continuous speech recognition (LVCSR). Many fast search strategies have been proposed until now. In [2], we proposed a new search strategy with a phoneme graph based hypothesis restriction, which efficiently reduces the search space. For 5000-word task, exprimental results showed that the method can reduce 70 % of the elapsed time without any error increasing. For further faster search, we incorporated 1-phoneme look-ahead technique into phoneme graph generation. We evaluate the proposed method with 20000-word Japanese newspaper task. Expremental results show that the method can reduce about 60 % of the elapsed time without error rate increasing.

key words　LVCSR, hidden Markov network, search strategy, phoneme graph, 1-phoneme look-ahead

1 はじめに

近年, 新聞記事やニュース音声を対象とする大語彙連続音声認識 (Large Vocabulary Continuous Speech Recognition : LVCSR) の研究が盛んになっている [1].
LVCSRは探索空間が非常に大きくなるために, 処理量削減のための認識アルゴリズムの設計は非常に重要な鍵となる. 文献 [2] では, 音素グラフに基づく仮説制限を提案した. そこでは, 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よ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がある. この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 システムに導入して, 文献 [2] では語彙サイズ 5000 の新聞記事読み上げ音声で評価し, 処理時間の約 70%を削減できることを示した.

本報告では, 音素グラフ生成の際に,Fast Match モデルを用いて 1 音素先読みを行うことで, 音素グラフ生成自体の処理量削減をはかる. そして今回は, 語彙サイズ 20000 の新聞記事読み上げ文音声で評価し, 音素グラフに基づく仮説制限法の導入は, 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 60%を削減できることを示す.

2 LVCSRシステム

2.1 システムの構成

現在の音声認識の多くは, 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. 入力音声Xが単語列Wである確率は, ベイズ則にしたがって

$$P(W|X) = \frac{P(X|W)P(W)}{P(X)} \tag{1}$$

のように与えられる. この確率 $P(W|X)$ が最大になる単語列Wを求めることが音声認識の目的である. 式 (1) は

$$P(\hat{W}|X) = \underset{W}{\operatorname{argmax}} P(X|W)P(W) \tag{2}$$

となる. ここで,$P(X|W)$ は音響モデル (Acoustic model),$P(W)$ は言語モデル (Language model) によって計算される確率である.$P(X)$ は W に関係ないので無視できる. 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. システム構成を図 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 パスサーチを採用している. 図 1 のシステムの音素グラフ生成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動するモジュールである.

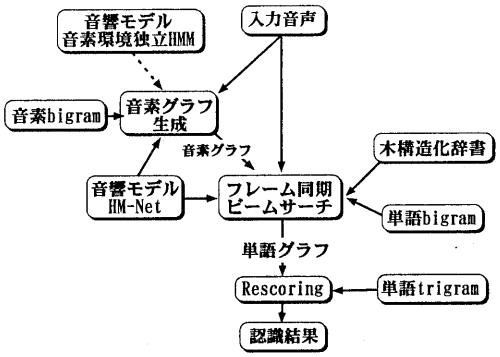


図 1: LVCSR システム

2.2 ベースラインデコーダ [2] [3]

LVCSR は探索空間が非常に大きいため, はじめから複雑なモデルを使用すると, 探索処理が複雑になり, 処理量が増加する. そこで, まず簡単なモデルを使用して探索空間を絞り, その後で複雑なモデルを適用するマルチパスサーチが有効とされている. システムのベースラインデコーダは, 第 1 パスで HM-Net と bigram を用いて, 木構造化辞書と単語対近似による one-pass アルゴリズムをフレーム同期, ビームサーチで実行し, 単語グラフを生成する. その際, 単語間の音素環境,bigram の factoring による言語モデルの先読みも考慮する. 第 2 パスで trigram を用いて単語グラフをリスコアリングして認識結果を得る. 単語グラフは, 認識単語列を段階的に絞り込む際の中間表現として生成される. 単語グラフは, 第 1 パスで求めた単語の音響尤度, 始端・終端の情報を持っている.

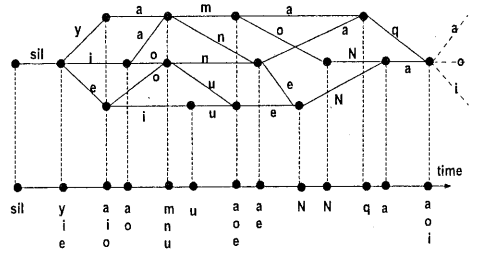


図 2: 音素グラフ内の境界候補と音素候補

2.3 音素グラフに基づく仮説制限法 [2]

(1) 音素グラフの生成

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 2). 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. 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. さらに,Fast Match モデルを用いて 1 音素先読み [4] を行うことで, 音素グラフ生成自体の処理量削減もはかる. ここでは, 当該フレームから 8 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利用する. 音素グラフの生成にかかる時間は, 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.

(2) 音素境界制限

単語グラフ生成の段階で, 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, 一位候補に基づく手法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, この手法では音素グラフの生成と単語グラフの生成において同じ音響モデルを用いることが前提となる.

(3)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, 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 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. ここでは音素グラフを用いて Backward スコアを求める.

この手法は,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ので, 1 音素先読み手法よりも仮説削減の効果が大きい. また, 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.

3 大語彙連続音声認識実験

3.1 音声資料・分析条件

音響モデル学習用の音声資料は,, 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読み上げ文+音素バランス文, 計 15732 文である.

評価は男性 23 名, 計 100 文である. 評価用の 100 文はそれぞれ毎日新聞の記事 1994 年 10 月～12 月から選ばれた未知語を含まない 20000 語で閉じた内容である.

音声データを 16kHz,16bit でディジタル化し, フレーム長 32msec, 分析周期 8msec で対数パワーと 12 次元の LPC メルケプストラムおよびそれらの 1 次と 2 次の回帰係数 (計 39 次元) を抽出する. さらに発話毎のケプストラム平均正規化を行う.

3.2 音響モデル

HM-Net の音素カテゴリーは 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.

1 音素先読みの Fast Match モデルとして音素環境独立,3 状態,4 混合分布の HMM を使用する.

3.3 言語モデル

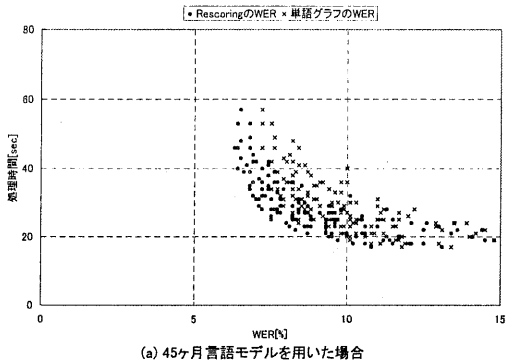
言語モデルとして,JNAS 付属の毎日新聞 91 年 1 月～94 年 9 月 (45ヶ月) と 91 年 1 月～94 年 9 月・95 年 1 月～97 年 6 月 (75ヶ月) の記事より推定された N-gram(bigram,trigram) を用いる. また, 音素グラフを生成する場合に使用する音素 bigram は, 音響モデルの学習に用いた音声データ 15732 文の発音表記から推定する.

3.4 評価方法

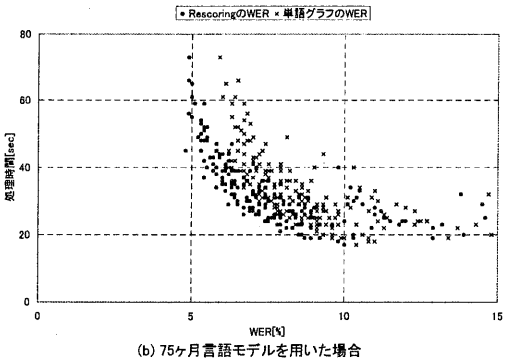
各単語は形態素単位 (品詞番号付き) とし, 漢字仮名まじり表記と読み表記, 品詞番号で表されている. 同じ漢字仮名表記であれば同じ単語として扱い, 品詞番号は考慮しない. 評価は 100 文の認識結果に対する単語誤り率

$$WER[\%] = \frac{S+I+D}{N} \times 100 \qquad (3)$$

で評価する. ここで S, I, D は置換, 挿入, 脱落誤りの数であり ,N は正解単語列の単語数を表す. 誤り率を求める際



(a) 45ヶ月言語モデルを用いた場合



(b) 75ヶ月言語モデルを用いた場合

図 3: WER と処理時間の関係の分布

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. 使用計算機は VT-Alpha 600 である.

3.5 実験結果と考察

(1) 実験結果

音素グラフ生成のビーム幅を 120〜180(対数値) としたときの音素グラフの性能と, その音素グラフに基づく仮説制限で生成された単語グラフとリスコアリングの性能を, 音素グラフ 1 音素先読みの有無, 及び,45ヶ月言語モデルと 75ヶ月言語モデルについて表 1 に示す.

表 1 の音素グラフの性能欄の境界候補, ヒット率,PGD,GER, 生成時間はそれぞれ, 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 音素当りの音素グラフ内の音素候補数), 音素グラフ誤り率 (音素グラフ内で最も正解に近い音素列に対する音素誤り率),1 文当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, 単語グラフの性能欄の WGD,GER はそれぞれ, 単語グラフ密度 (正解 1 単語当りの単語グラフ内の単語候補数), 単語グラフ誤り率 (単語グラフ内で最も正解に近い単語列に対する単語誤り率) を表している. なお,WER・GER の評価基準は漢字仮名まじりの一致で評価した単語誤り率である.

音素グラフ・単語グラフ生成のビーム幅, 言語尤度の重み, 挿入ペナルティ, 及び, リスコアリングの言語尤度の重み, 挿入ペナルティを様々に変えて,WER と処理時間の関係を求めると, 図 3 のような分布が得られる.

表 1 は, 図 3 の中から, 音素グラフ生成の各ビーム幅についてほぼ最適な条件の結果をまとめたものである. なお, 最右欄 (baseline) は, 音素グラフに基づく仮説制限を行わない, ベースラインデコーダの性能である.

表 1 の WER と処理時間の関係を, 音素グラフ生成における 1 音素先読みの有無について比較したものを図 4 に,45ヶ月言語モデルと 75ヶ月言語モデルについて比較したものを図 5 に示す.

(2) 音素グラフに基づく仮説制限の効果

ベースラインデコーダと音素グラフに基づく仮説制限 (ビーム幅 180) の性能比較を表 2 に示す.

音素グラフ生成のビーム幅が 180 の場合,WER は若干 (0.2%) 増加するが, ベースラインデコーダに対して処理時間を 58〜60%削減できる. 音素グラフに基づく仮説制限の効果は大きいといえる.

なお, 語彙サイズ 5000 の実験では, 音素グラフ生成の際に 1 音素先読みなしで, 処理時間を約 70%削減できていたが, 語彙サイズ 20000 の今回の実験では,1 音素先読みを行うことによって前回に近い削減が達成できたことになる.

(3) 音素グラフ生成における 1 音素先読みの効果

音素グラフ生成における 1 音素先読みは,PGD を小さくし, 音素グラフ生成時間を 58〜69%に短縮する効果がある. しかしながら,Fast Match モデルと HM-Net の音響尤度の不整合によりヒット率が若干 (0.5〜2.8%) 低下する. リスコアリング後のシステムの性能で見ると, 図 4 に見られるように,WER と処理時間の関係を改善する. その効果は,45ヶ月言語モデルを用いる場合のほうが若干大きい.

(4)45ヶ月言語モデルと 75ヶ月言語モデルの比較

75ヶ月言語モデルは 45ヶ月言語モデルと比較して, 図 5 にみられるように,WER と処理時間の関係を確実に改善する. 表 2 の性能比較によると,75ヶ月言語モデルは, 処理時間が若干 (7%〜11%) 増加するが,WER が 45ヶ月言語モデルを用いるよりも 1.1%(誤り改善率で 19%) 良

表 1: 音素グラフに基づく仮説制限法を用いた実験結果

(a) 音素グラフ生成における 1 音素先読みなしの場合

音素グラフ生成のビーム幅			120	130	140	150	160	170	180	baseline
	音素グラフの性能	境界候補 [%]	17.8	21.0	25.1	29.5	34.8	40.6	46.8	-
		ヒット率 [%]	96.9	97.6	98.3	98.6	99.0	99.3	99.5	-
		PGD	4.8	7.1	10.7	16.4	25.3	39.4	61.2	-
		GER[%]	2.4	1.8	1.3	1.0	0.8	0.6	0.5	-
		生成時間 [sec]	11	13	15	18	20	24	27	-
45ヶ月	単語グラフの性能	WGD	4.3	4.5	4.5	4.2	5.0	6.9	4.2	4.8
		GER[%]	4.9	4.0	3.5	3.0	3.0	2.2	2.3	2.2
		WER[%]	9.8	9.1	8.4	7.5	8.2	7.2	7.1	6.8
	Rescoring	WER[%]	8.7	7.7	7.5	6.5	6.5	6.1	5.8	5.7
	処理時間 [sec](音素グラフ生成時間を含む)		25	30	31	38	47	47	59	101
75ヶ月	単語グラフの性能	WGD	3.7	3.7	3.7	3.8	5.7	5.5	6.8	8.1
		GER[%]	5.3	4.3	3.5	3.0	2.6	2.2	2.0	1.8
		WER[%]	9.2	8.2	7.7	6.8	6.6	6.2	6.2	5.9
	Rescoring	WER[%]	8.0	7.4	7.1	6.1	5.6	5.2	4.9	4.6
	処理時間 [sec](音素グラフ生成時間を含む)		24	27	30	34	40	46	59	112

(a) 音素グラフ生成における 1 音素先読みありの場合

音素グラフ生成のビーム幅			120	130	140	150	160	170	180	baseline
	音素グラフの性能	境界候補 [%]	17.5	20.6	24.3	29.0	33.0	33.9	44.9	-
		ヒット率 [%]	94.1	95.3	96.3	97.8	98.3	98.7	99.0	-
		PGD	4.6	6.6	9.8	15.4	23.3	35.1	52.1	-
		GER[%]	3.1	2.4	1.9	1.3	1.0	0.9	0.7	-
		生成時間 [sec]	7	8	9	12	13	14	16	-
45ヶ月	単語グラフの性能	WGD	4.3	4.5	4.6	4.2	4.3	4.2	4.1	4.8
		GER[%]	5.3	4.4	3.6	3.2	3.0	2.8	2.5	2.2
		WER[%]	10.3	9.4	8.4	7.8	8.0	7.4	7.0	6.8
	Rescoring	WER[%]	9.3	8.1	7.5	6.9	6.8	6.3	5.9	5.7
	処理時間 [sec](音素グラフ生成時間を含む)		21	23	25	32	35	38	42	101
75ヶ月	単語グラフの性能	WGD	3.5	6.1	5.2	5.6	5.7	5.6	5.5	8.1
		GER[%]	6.0	3.8	3.2	2.9	2.5	2.4	1.8	1.8
		WER[%]	9.6	8.3	7.2	6.7	6.7	6.3	5.8	5.9
	Rescoring	WER[%]	8.9	7.2	6.5	5.8	5.4	5.4	4.8	4.6
	処理時間 [sec](音素グラフ生成時間を含む)		19	26	30	34	37	42	45	112

表 2: 音素グラフに基づく仮説制限とベースラインデコーダの性能比較

		45ヶ月言語モデル		75ヶ月言語モデル				
						45ヶ月モデルとの比較		
		WER [%]	処理時間 [sec]	WER [%]	処理時間 [sec]	WER 減少分	WER 改善率	処理時間 増加分
ベースラインデコーダ		5.7 (4.9)	101	4.6 (4.1)	112	1.1	19%	11%
音素グラフに基づく仮説制限 (ビーム幅 180)		5.9 (5.1)	42	4.8 (4.3)	45	1.1	19%	7%
ベースラインとの比較	WER 増加分	0.2		0.2				
	処理時間減少分		58%		60%			

( ) 内：読み評価の WER

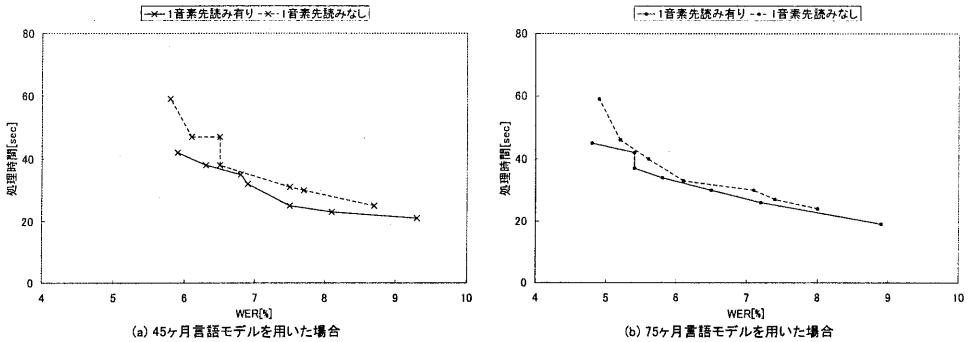


図 4: WER と処理時間の関係 (音素グラフ生成における 1 音素先読みの有無の比較)

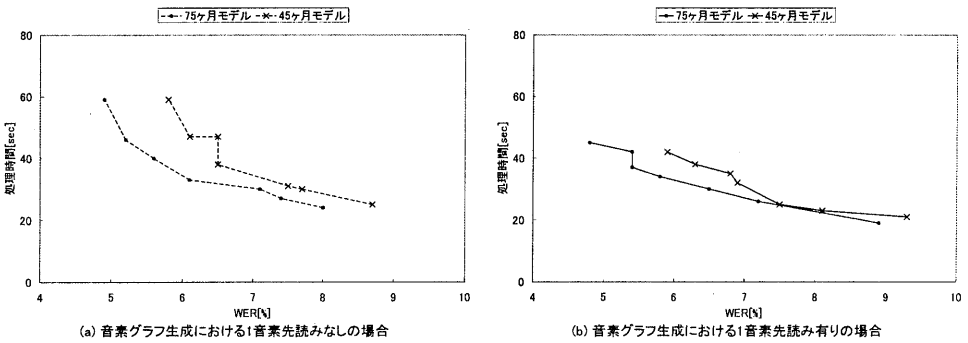


図 5: WER と処理時間の関係 (45ヶ月言語モデルと 75ヶ月言語モデルの比較)

くなる.

4 おわりに

音素グラフに基づく仮説制限法を用いた大語彙連続音声認識システムにおいて, 音素グラフ生成の際に FastMatch モデルを用いて 1 音素先読みを行うことで, 音素グラフ生成自体の処理量削減をはかった. 語彙サイズ 20000 の新聞記事読み上げ音声で評価し, ベースラインデコーダと比較して, 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 60%を削減できることを示した. この結果より音素グラフに基づく仮説制限は大語彙連続音声認識に有効な手法であり, 音素グラフ生成の際に 1 音素先読みを用いることでさらに処理時間を削減することが可能である.

今後は自然発話文の認識等を行っていく予定である.

参考文献

[1] 河原 達也, 李 晃伸, 小林 哲則, 武田 一哉, 峯松 信明, 伊藤 克亘, 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 (98 年度版) の性能評価, 情報処理学会研究報告,99-SLP-26-6(1999-05).

[2] 堀 貴明, 岡 直生, 加藤 正治, 伊藤 彰則, 好田 正紀:大語彙連続音声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, 情報処理学会論文誌,vol.40,No.4,pp.1365-1373(1999-04).

[3] S.Ortmanns,H.Ney,X.Aubert:A word graph algorithm for large vocabulary continuous speech recognition,Computer Speech and Language,Vol.11,No.1,pp.43-72(1997).

[4] S.Ortmanns,A.Eiden,H.Ney,N.Coenen:Look-ahead techniques for fast beam search,ICASSP97,pp.1783-1786(1997).