音声言語情報処理 34－19
(2000. 12. 21)

単語グラフ生成における パラメータ最適化の検討

加藤 正治　斎院 俊典　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 : hori@ei5sun.yz.yamagata-u.ac.jp

あらまし 本報告では，マルチパスサーチに基づく音声認識における，単語グラフ生成のためのパラメータ推定法について検討する．種々の条件を単語グラフを利用して，仮想的に表現することで，言語重み・挿入ナルティ，ビーム幅などのパラメータを効率よく推定することができる．ビーム幅を固定した場合，第1パスと同じ条件で単語グラフをリスコアすることで1，2回の繰り返し操作で最適な言語重み・挿入ペナルティを求められる．ビーム幅を含めた最適化では，あらかじめ，大きなサイズのグラフを用意することでパラメータを最適化する．提案法を新聞記事読み上げ音声に適応したところ，開発用650文で作成したパラメータは，評価用100文で作成したものよりも良好な良好が得られた．

キーワード 大語彙連続音声認識，隠れマルコフ網，単語グラフ，リスコアリング，パラメータ最適化

Optimization of the Parameter Set for Word Graph Generation

Masaharu KATOH, Toshinori SAIIN Akinori ITO, and Masaki KOHDA

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

Abstract Language model weight and insertion penalty greatly affects the recognition preformance of LVCSR system. In the multi-pass LCVSR system that uses word graphas an intermediate data structure, theses decorder parameter should be optimized in order to generate a good word graph. We proposed the rescoring based method that uses bigram LM insted of generating many word graphs for each parameter setting. As the rescoring is much faster than the re-generation of a word graph, the optimization time of the proposed method is much shorter. In this paper, we tested proposed method on Japanese News Article Sentences (ASJ-JNAS). When obtaied enough development data, the recognition performance is improved.

key words LVCSR, Hidden Markov Network, word graph, rescoring, optimize parameters

1 はじめに

大語彙連続音声認識 (LVCSR) システムには，システムを調整するパラメータがいくつか存在する．これらは，認識結果を大きく左右する要因となることも知られている．最も良い認識結果が得られるようにシステムのパラメータを調整することが望まれる．しかし，多くの場合は具体的な最適化手段が無く，全てのパラメータの組合せについて認識実験が必要となる．文献 [1] でのLVCSR システムは，マルチパスシステムに基づく探索を行っている．第 1 パスで，言語モデルに bigram を用いてビームサーチを行い単語グラフを生成する．第 2 パスでは，単語グラフを trigram でリスコアすることで最終的な認識結果を得る．このような，2 パスサーチの場合，第 1・第 2 パスそれぞれ言語重み・挿入ペナルティの組について認識を繰返し，実験的に最適なパラメータを得ることができる．しかし，第 1 パスの単語グラフ生成は処理用が非常に大きく，単語グラフ生成を繰返して最適なパラメタを求めることは効果的ではない．これまでの研究では，単語グラフを第 1 パス設定の条件に基づいてリスコアする手法について検討した [2]．ここでは，単語グラフを bigram を用いて，種々のパラメータでリスコアし最適なパラメータを求めている．また，音響スコアのリスコアを併用することでビーム幅に対する最適化の検討も行っている [3]．新聞記事読み上げ音声に対する実験について検討したところ，手法の有効性や収束性を示すことができた．

本報告では，評価用は別に同じ環境で収録された開発用の音声データを用いてパラメータの最適化をすることを検討する．具体的には，新聞記事読み上げ音声データのうち 650 文を開発用データとして，本手法によりパラメータの最適値を求め，別の 100 文を評価用とする場合について検討する．

2 実験条件

2.1 音声資料・音声分析

音声資料は ASJ-JNAS 日本音響学会新聞記事読み上げコーパスの中から，5000 語彙で閉じた 750 文を用いる．開発用として 650 文を用い，本手法によって最適パラメータを求める．評価用に残り 100 文を用いることで open 実験による性能を評価する．

音声データを 16kHz，16bit で量子化し，フレーム長さ 32msec，分析周期 8msec で対数パワーと 1〜12 次の 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．更に発話毎のケプストラム平均正規化を行う．

2.2 音響モデル・言語モデル

音響モデルには，状態クラスタリングによる HM-Net を用いる．音声資料として，新聞記事読み上げ音声コーパス (ASJ-JNAS) より男性 102 名，15732 文を用いる．音素カテゴリは，長母音，二重母音を含む 35 種類とする．1600 状態 4, 8, 16 混合，2000 状態 4, 8, 16，計 6 種類について実験する．

言語モデルには，毎日新聞記事 91 年 1 月〜94 年 9 月までの 45ヵ月より推定された語彙サイズ 5000 の bigram と trigram を用いる．

3 LVCSR システム

LVCSR システムは，探索空間が非常に大きいため，始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，簡単なモデルを利用して探索空間をしぼり，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 [6]．構築したシステムのデコーダは，第 1 パスで HM-Net と bigram を用いて，木構造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．その際に，単語間の音素環境や factoring による言語スコアの先読み [5] も考慮している．第 2 パスでは trigram を用いて単語グラフをリスコアし認識結果を得る．

4 言語重み・挿入ペナルティの最適化

4.1 言語重み・挿入ペナルティ

音響モデルの尤度は多次元正規分布の混合で表現する確率密度値であり，n-gram を用いる言語尤度は離散確率分布の値である．さらに，音響尤度は音声フレーム毎に計算され，言語尤度は単語毎に適用される．これらの値は大幅にダイナミックレンジが異なる．音響モデルの尤度と言語モデルの尤度のバランスを考慮して，言語重み λ を導入すると次式で表す．

$$\hat{W} = \underset{W}{\operatorname{argmax}} \{\log P(O|W) + \lambda P(W)\} \quad (1)$$

また，挿入誤りや脱落誤りを制御するためにペナルティ δ を導入した次式が良く用いられる．

$$\hat{W} = \underset{W}{\operatorname{argmax}} \{\log P(O|W) + \lambda P(W) + n\delta\} \quad (2)$$

ここで，n は W を構成する言語モデルの単位数である．

言語重みと挿入ペナルティは種々に変えて，最適な認識率を導く値が使わので，原理的にはこれらのパラメータの全ての組合せについて認識実験を行う必要がある．

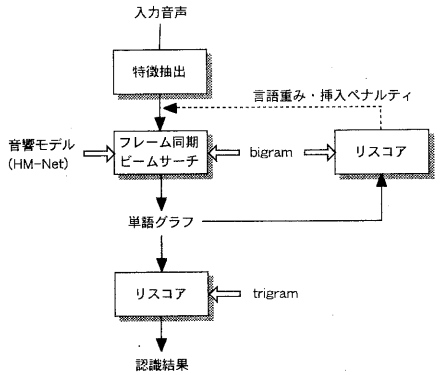

図 1: 言語重み・挿入ペナルティの最適化

4.2 bigram を使ったリスコア

第 2 パスのリスコアにかかる処理はわずかであることから，幾通りかの言語重み・挿入ペナルティの組について認識実験を繰り返し，実験的に最適なパラメタを求めても問題は無い．しかし第 1 パスの単語グラフ生成は，リスコアに比べて処理量が非常に大きく，単語グラフ生成を繰り返して実験的に最適なパラメータを求めることは効率的ではない．

単語グラフ生成の言語重み・挿入ペナルティを最適化するために，生成した単語グラフに対して第 1 パスで用いた言語モデルでリスコアする手法を提案する．本研究では，第 1 パスで bigram を第 2 パスで trigram を用いているので bigram を使ったリスコアがこれにあたる．

単語グラフを bigram でリスコアする際の言語重みと挿入ペナルティを種々に変えることで，これらのパラメータに対応した認識結果を得ることができる．この結果を参考にして，最も良い WER(単語誤り率) を導く言語重み・挿入ペナルティの組を第 1 パスの最適パラメータとする．

リスコアにかかる処理量は単語グラフ生成の 1%に満たないため，数百通りのパラメータの組合せについてのリスコアは，単語グラフを数回生成する処理量に相当する．

4.3 アルゴリズム

単語グラフ生成の言語重み・挿入ペナルティ最適化の処理を追加した LVCSR システムを図 1 に示す．言語重み・挿入ペナルティ最適化のアルゴリズムは次のようになる．

(1) 第 1 パスの言語重み・挿入ペナルティを初期値に設定し，単語グラフを生成する．
(2) 単語グラフを bigram でリスコアする．リスコアは言語重み・挿入ペナルティのあらかじめ定めた範囲内で行い，WER が最小になる条件を求める．
(3) (2) で求まる条件に，言語重み・挿入ペナルティを更新し，再び単語グラフを生成する．
(4) 言語重み挿入ペナルティが収束するまで (2)，(3) を繰り返す．

単語グラフの精度は WER，GER(グラフ誤り率) 等で評価されるが，このアルゴリズムの評価基準は，第 1 パスの WER 最小化である．第 2 パスの WER 改善や単語グラフの GER 改善は保証されていない．なお，このアルゴリズムを用いる際，ビーム幅は固定する．ビーム幅が途中で変わると，言語重み・挿入ペナルティの最適値も変化して，アルゴリズムが収束しないことがある．

5 言語重み・挿入ペナルティ最適化の実験

5.1 実験条件

単語グラフ生成の単語内，単語間ビーム幅は 300, 200. 初期単語グラフ生成の言語重み・挿入ペナルティは 30, 10 とする．また，bigram によるリスコアは言語重み 2〜50(2 刻 : 25 通り)，挿入ペナルティ −50〜50(5 刻 : 21 通り) の 525 通りとする．最終的な認識結果は，trigram によるリスコア言語重み 2〜50(2 刻 : 25 通り)，挿入ペナルティ −50〜 − 2(2 刻 : 25 通り) の 625 通りの探索範囲で評価する．開発用 650 文，及び，評価用 100 文について実験する．

5.2 結果と考察

1600 状態の結果を表 1(a)〜(c) に，2000 状態の結果を表 2(a)〜(c) に示す．表には，第 1 パスの WER(単語誤り率) のほかに，第 2 パスでの認識性能．GER(グラフ誤り率)，WGD(単語グラフ密度) を示した．結果から，いずれの状態数の場合にも，1，2 回で収束している．比較として，100 文で直接最適値を求めた場合と，グラフ誤り率を 1.0% 以下となるように設定した，baseline(従来法) についても示している．bigram リスコアリングにおけるパラメータと認識率の関係を図 2，3 に示す．

本手法を 650 文で行った場合でも，文献??の 100 文の場合と同様に，パラメータが 1，2 回で第 1 パス WER の下限に収束した．bigram によるリスコアの WER の

表 1: 最適パラメータの性能評価：1600 状態

(a) 1600 状態 4 混合

評価文/単語グラフ	w/p	WER (%) 1st	WER (%) 2nd	GER (%)	WGD
650/初期	30, 10	12.1	8.6	3.1	10.2
/再生成	20, 0	10.2	7.2	1.1	37.7
100/baseline	24, −5	7.7	6.0	0.4	75.4
/初期	30, 10	9.6	6.4	2.5	10.1
/opt.100	18, −40	**7.1**	5.9	1.0	24.2
/opt.650	20, 0	8.6	**5.8**	0.5	31.8

(b) 1600 状態 8 混合

評価文/単語グラフ	w/p	WER (%) 1st	WER (%) 2nd	GER (%)	WGD
650/初期	30, 10	10.5	7.0	2.1	10.7
/再生成	20, −20	8.9	6.2	0.7	29.3
/再生成	20, 0	8.7	6.2	0.6	45.8
100/baseline	26, −5	7.8	6.4	0.2	69.3
/初期	30, 10	8.6	6.0	1.1	10.2
/opt.100	16, −35	**7.5**	5.8	0.3	49.4
/opt.650	20, 0	8.5	**5.7**	0.1	36.8

(c) 1600 状態 16 混合

評価文/単語グラフ	w/p	WER (%) 1st	WER (%) 2nd	GER (%)	WGD
650/初期	30, 10	9.4	6.5	1.7	10.2
/再生成	22, 0	8.1	5.6	0.6	28.0
/再生成	20, −10	8.0	5.8	0.5	33.6
100/baseline	22, −5	7.6	6.0	0.2	94.9
/初期	30, 10	8.1	5.4	0.6	9.1
/opt.100	14, −30	**6.5**	5.7	0.1	90.7
/opt.650	20, −10	7.5	bf5.6	0.2	25.4

opt.100 は 100 文から求めた最適値,
opt.650 は 650 文から求めた最適値.

3 次元グラフにの結果を見ると，100 文の場合に比べて 650 文の場合は等高線がほぼ同心円上で極小値がひとつになっている．サンプル数が増えたことによって収束が安定したといえる．さらに，第 2 パスの WER については，どの音響モデルの場合でも初期単語グラフを上回る性能を示した．(100 文の場合では，初期単語グラフのほうが性能が高い場合もある．)

開発用 650 文で求めたパラメータを評価用 100 文に適用すると，100 文から求めたパラメータと比べ，第 1 パスでの性能は劣る．しかし，第 2 パスの性能は，同等かそれ以上を示した．単語グラフを複数個作成し，GER が 1.0% 以下のものを採用した baseline の性能を考慮すると，本手法が第 1 パスのパラメータを決める上で有効であるといえる．

6 ビーム幅を考慮した最適化

6.1 音響スコアの再評価

単語グラフでは，節点には時刻と連接する弧の情報が記録され，弧には，単語とスコアが記録されている．ビーム幅を考慮する場合は，グラフに対して音響スコア・言語スコアを再計算することになる．このとき，再計算はグラフ内に限られるため，全ての単語を対象とする通常の探索をするよりも高速に探索を行うことができる．また，グラフの構造 (単語のセグメント位置) を変えない事とする．

6.2 最適化の手順

ビームサーチより作成される単語グラフは，元の単語グラフを越えることは無い．そこで，あらかじめ大きなサイズの単語グラフを作成しておき，それを再評価する．言語重み・挿入ペナルティの初期値を適当に与えて，単語グラフを作成する．言語重み・挿入ペナルティの最適化とビーム幅の最適化を交互に行うことでパラメータの最適化を実現する (図 4).

(1) 第 1 パスの言語重み・挿入ペナルティを初期値に設定し，大きい単語グラフを生成する.
(2) 単語グラフを bigram でリスコアする．リスコアは言語重み・挿入ペナルティのあらかじめ定めた範囲内で行い，WER が最小になる条件を求める.
(3) 言語重み・挿入ペナルティを最適値に設定し，(1) で求めた単語グラフをリスコアして，各ビーム幅について単語グラフを作成する.
(4) (3) で得られた複数の単語グラフに対して，それぞれ，言語重み・挿入ペナルティの最適値を求める.
(5) 言語重み・挿入ペナルティ，ビーム幅の最適値が収束するまで (3),(4) を繰返す.

前節の場合はビーム幅を固定することで，安定して最適値が得られることが分かっている．ここでは，元となる単語グラフを固定することで，安定して最適値を得ることを期待している.

表 2: 最適パラメータの性能評価:2000 状態

(a) 2000 状態 4 混合

評価文/		WER (%)		GER	WGD
単語グラフ	w/p	1st	2nd	(%)	
650/初期	30, 10	11.5	8.1	2.8	9.4
/再生成	20, 0	9.3	6.9	1.0	35.1
100/baseline	24, −5	8.0	6.4	0.4	84.0
/初期	30, 10	9.3	6.8	2.3	9.4
/opt.100	24, −10	**8.1**	6.4	1.6	14.3
/opt.650	20, 0	8.6	**6.0**	0.9	29.8

(b) 2000 状態 8 混合

評価文/		WER (%)		GER	WGD
単語グラフ	w/p	1st	2nd	(%)	
650/初期	30, 10	10.0	6.9	1.8	9.8
/再生成	18, −10	8.6	6.1	0.6	50.2
/再生成	20, −15	8.5	6.1	0.7	28.7
100/baseline	26, −5	8.3	6.1	0.2	83.1
/初期	30, 10	9.1	5.9	1.5	9.1
/opt.100	18, −40	**7.7**	5.6	0.4	22.8
/opt.650	20, −15	8.2	**5.6**	0.4	21.1

(c) 2000 状態 16 混合

評価文/		WER (%)		GER	WGD
単語グラフ	w/p	1st	2nd	(%)	
650/初期	30, 10	11.5	8.1	2.8	9.3
/再生成	20, −15	9.3	6.9	1.0	27.1
100/baseline	22, −5	6.9	5.6	0.3	81.9
/初期	30, 10	7.6	**5.4**	1.2	8.0
/opt.100	26, −15	**6.4**	5.5	0.4	20.4
/opt.650	20, −15	6.6	5.5	0.4	18.7

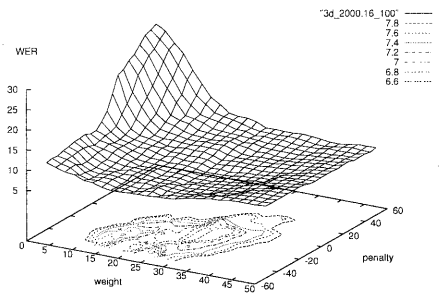

図 2: bigram を使ったリスコアの WER
2000 状態，16 混合，100 文

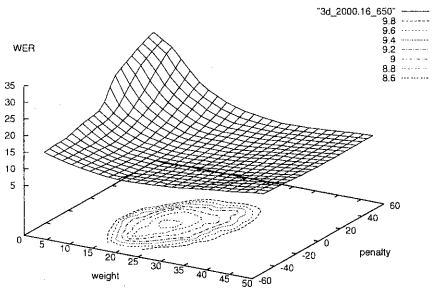

図 3: bigram を使ったリスコアの WER
2000 状態，16 混合，650 文

7 ビーム幅を考慮した最適化の実験

7.1 実験条件

パラメータの最適値を求めるための大きな単語グラフは，言語重み 30，挿入ペナルティ10，単語内ビーム幅 500，単語間ビーム幅 400 に設定して作成する．リスコアリングのビーム幅は 200〜400(20 刻：11 通り)，単語間のビーム幅は一律に「単語内 − 100」で設定した．初期単語グラフ生成の言語重み・挿入ペナルティは 30, 10 とする．また，bigram によるリスコアは言語重み 2〜50(2 刻み：25 通り)，挿入ペナルティ −50〜50(5 刻：21 通り) の 525 通りとする．最終的な認識結果は，trigram によるリスコア言語重み 2〜50(2 刻：25 通り)，挿入ペナルティ −50〜 − 2(2 刻：25 通り) の 625 通りの探索範囲で評価する．開発用 650 文，及び，評用 100 文を認識対象として実験する．

7.2 結果と考察

開発用 650 文に対する結果を表 3 に示す．表中のビーム幅は，単語内ビーム幅である．ビーム幅固定で行なう実験にくらて，第 1 パスの WER はほぼ同等の結果が得られた．一方，第 2 パスの WER は，劣化している．探索空間が基となる単語グラフ内に限られているためだと考えられる．WGD はビーム幅固定のときよりかなり大きくなっている．これは，主に factoring の影響だと考えられる．単語グラフ内に全てのノードが残っているわけではないので，閾値の推定が適当に行なえず，枝刈の効果がでていない．

評価用 100 文に対する結果を表 4 に示す．第 1 パスの WER 結果は，ビーム幅を固定するとき同様 open 実験のため劣化している．しかし，第 2 パスの WER は，同等以上の結果を得られている．また，ビーム幅を固定

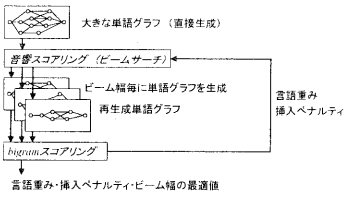

図 4: ビーム幅を考慮した最適化手順

するときとの比較では，同程度の結果が得られている．

8 まとめ

マルチパスサーチに基づく音声認識において，単語グラフ作成のためのパラメータ推定法について検討した．本研究では，単語グラフを利用することで，言語重み・挿入ナルティ，ビーム幅などのパラメータを効率よく推定することがことを示した．

ビーム幅を固定した場合，第1パスと同じ条件で単語グラフをリスコアすることで1，2回の繰り返し操作で最適な言語重み・挿入ペナルティを求められる．ビーム幅を含めた最適化では，あらかじめ，大きなサイズのグラフを用意することでパラメータの最適化が行なえる．また，十分な開発用データが与えられているとき closed 実験よりも良好な結果を得られることを示した．

参考文献

[1] 堀，岡，加藤，伊藤，好田："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"，情報処理学会論文誌，vol40,No4,pp.1365-1393(1999-4).

[2] 斎院，加藤，伊藤，好田："単語グラフ生成の言語重み・挿入ペナルティの最適化の検討"，信学技報，SP2000-26,pp.75-82(2000-6).

[3] 加藤，斎院，伊藤，好田："単語グラフ生成におけるパラメータ最適化法の検討"，日本音響学会，秋季研究発表会，1-5-17,pp33-34(2000-9).

[4] 伊藤，好田："N-best 候補からの言語重みと挿入ペナルティの最適化に関する検討"，情処研法，99-SLP-28-6(2000-3).

[5] S.Ortmanns, H.Ney, A.Eiden: "Language-model look-ahead for large vocabulary speech recognition". Proc.ICSLP'96. Vol.4,pp.2095-2098 (1996).

表 3: 最適パラメータの性能評価:開発用 650 文

評価文		WER (%)		GER	WGD
	b/w/p	1st	2nd	(%)	
1600. 4	380, 20, 0	10.3	6.9	1.2	333.8
1600. 8	360, 20, 0	8.7	6.4	0.6	450.7
1600.16	300, 20, −10	8.3	6.3	0.7	303.2
2000. 4	340, 22, 0	9.4	7.1	1.1	193.2
2000. 8	280, 18, −5	8.7	6.3	0.9	205.5
2000.16	300, 22, −5	8.1	6.1	0.7	259.5

表 4: 最適パラメータの性能評価:評価用 100 文

評価文		WER (%)		GER	WGD
st.mix	b/w/p	1st	2nd	(%)	
1600. 4	300, 18, −40	7.1	5.9	1.0	24.2
	380, 20, 0	9.2	5.7	0.6	234.8
1600. 8	280, 18, −35	7.8	6.1	0.2	38.8
	360, 20, 0	8.5	5.8	0.3	222.3
1600.16	360, 18, −35	7.7	5.8	0.1	77.6
	300, 20, −10	7.3	5.8	0.4	30.2
2000. 4	300, 24, −10	8.2	6.2	0.9	35.0
	340, 22, 0	9.8	5.8	0.9	60.3
2000. 8	340, 18, −35	8.6	5.7	0.3	61.2
	280, 18, −5	8.1		1.9	259.7
2000.16	380, 24, −25	7.3	5.4	0.1	52.5
	300, 22, −5	6.7	5.7	0.7	19.7

上段は 100 文から求めた最適値．

下段は 650 文から求めた最適値．

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