TR-IT-0186

単語グラフ密度を用いた連続音声認識系の性能評価法の検討

A study on performance estimation of CSR using word graph density

横山 真男
Masao Yokoyama

清水 徹
Toru Shimizu

1996.9.13

単語グラフを用いた連続音声認識系における、認識性能評価法の検討とその結果を述べる。計算量やbeam探索幅も考慮に入れた多面的な認識性能の評価方法を検討した。その結果、認識率は単語グラフ密度の対数と相関が高いこと、単語グラフ密度の対数をbeam幅と言語尤度の重みで正規化した量を用いて、大まかな認識率を予想できることがわかった。また、この結果を用いて、単語グラフ密度を用いた音響モデルの性能評価を試みた。

目次

1 概要 **2**

2 はじめに **2**

3 単語グラフを用いた連続音声認識系の構成 **2**
3.1 単語グラフの生成 2
3.2 実験条件 3

4 単語グラフ密度と単語認識率・計算量の関係 **3**
4.1 実験条件 3
4.2 考察 5

5 言語重み係数・ビーム幅で補正した単語グラフ密度と単語認識率・計算量の関係 **5**
5.1 実験条件 5
5.2 考察（補正方法） 5

6 単語グラフを用いた音響モデルの評価 **7**
6.1 実験条件 7
6.2 考察 7

7 まとめ **9**

1 概要

本研究報告書では、単語グラフを用いた連続音声認識系における、認識性能評価法の検討とその結果を述べる。従来、認識性能が単語認識率のみで評価されていたが、認識にかかる計算量やbeam探索幅も考慮に入れた多面的な性能評価方法を検討した。その結果、認識率は単語グラフ密度の対数と相関が高いこと、単語グラフ密度の対数をbeam幅と言語尤度の重みで正規化した量を用いて、大まかな認識率を予想できることがわかった。また、この結果を用いて、単語グラフ密度を用いた音響モデルの性能評価を試みた。

2 はじめに

近年、大語彙連続音声認識における効率的な探索手法として単語グラフを用いた認識手法が提案されている[1]-[4]。これまで認識性能の評価は、音響モデル、言語モデルなどの改良の評価は、単語認識率によって行われてきた。また、探索において仮説数を削減し、高速化を計る手法として単語グラフが用いられているが、効率的な探索のための知識源の配分についてはほとんど検討がなされていない。探索にかかった計算時間や、語彙数、beam幅、言語ゆう度比によっても認識率は左右されるので、単語認識率といった一面的な性能評価では不十分で、単語認識率以外のパラメタも考慮に入れた多面的な評価法が必要といえる。つまり、単語認識率、単語latticeの複雑さ、計算時間などの多面的な特徴をとらえることで、単語認識率と非常に関係が深いパラメータを求めることができれば、そのパラメータによって性能評価ができると考えられる。

単語グラフを用いたmulti pass searchにおける認識系において、一定のbeam幅に入る単語仮説数は、認識器の弁別性能が高いほど少ないと予想される。その探索時の単語仮説数は音声認識の出力である単語グラフの規模に反映される。一発声あたりの単語仮説数である単語グラフ密度は次の式で求められる。

$$\text{単語グラフ密度} = \frac{\text{単語グラフ内総単語仮説数}}{\text{発声単語数}}$$

そこで本報告ではまず、単語グラフ密度と先ほど挙げたパラメータとの関係を調査し、それを元に単語グラフ密度による認識性能の評価法を検討した。また、その結果を用いて音響モデルの評価を行なった。

3 単語グラフを用いた連続音声認識系の構成

本認識系は、次に述べる時間同期に処理を行なう第一パスと時間非同期の第二パスから構成されるmulti pass searchである[5]。

3.1 単語グラフの生成

つぎの2つのパスによって作成される。

- 第一パス：単語仮説の生成
  - 音素環境依存音素HMMと可変長N-gram生成法により作成された単語辞書・クラスバイグラム確率値を用いて、時間同期ビームサーチを行なう。
  - 詳細な音響モデルを用い、言語モデルは制約の緩いものを用いる。

- 単語の終端に達した単語仮説を単語グラフに登録する。

- 第二パス：単語仮説のプルーニング
  - 単語仮説の言語ゆう度の再評価を行なう。
  - 音響ゆう度は第一パスで得られたものを使う。
  - 詳細な言語モデルを用いて、総ゆう度の低いものを単語グラフから削除する。

3.2 実験条件

本実習で用いた音声認識実験条件を表1に示す。

サンプリング周波数	12 [kHz]
分析窓	20 [ms] ハミング窓
フレームシフト	10 [ms]
パラメータ	16次LPCケプストラム + logパワー + 16次Δケプストラム + Δ logパワー
音響モデル	5混合対角共分散連続分布型HMM 状態数401のHMnet, Gender-dependent
学習データ	男性146名 + 女性197名（平均発話数10）
評価データ	テストセットSL3 男性3名 + 女性4名 ATR Travel Arrangement Corpusの”ホテル予約” 100発声, 983語
言語モデル	828対話から713クラスのクラスバイグラムを作成
単語辞書	6,635 words

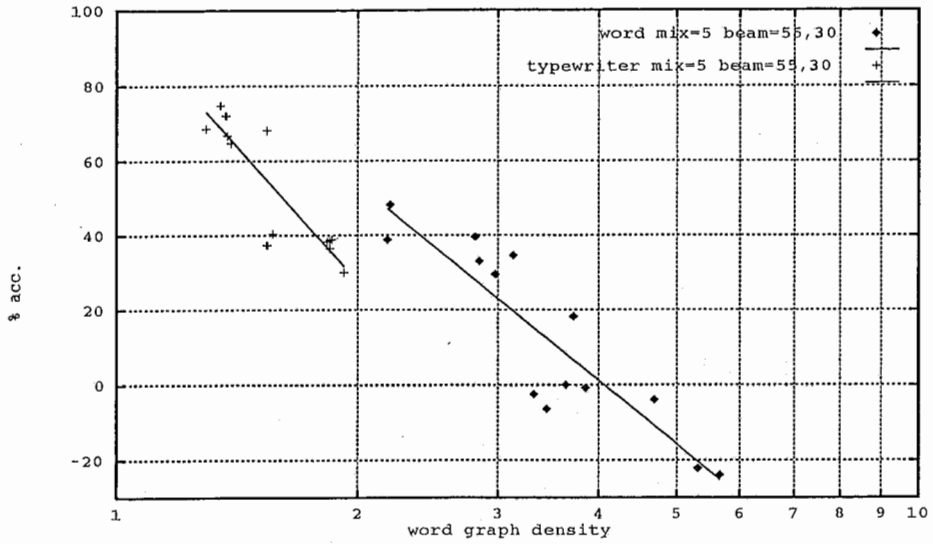
表1: 実験条件

4 単語グラフ密度と単語認識率・計算量の関係

単語グラフ密度と認識率・計算量との間には相関関係があるかどうかを連続単語認識、音素タイプライタで調べた結果を図1,2に示す。

4.1 実験条件

データは男性3名女性4名の各話者の一対話毎の単語グラフ密度・単語認識率・ＣＰＵ時間である。さらに、相関が捉えやすいようにデータを増やすために、男性に女性の音響モデル、女性に男性の音響モデルで認識した結果も加え回帰直線を求めた。連続単語認識の言語重みは(1st pass = 4.0, 2nd pass = 8.0)、連続単語認識・音素タイプライタのbeam幅は(1st pass = 55, 2nd pass = 30)で行なった。



(単語認識 word) $y = -174.19 * log(x) + 106.07$ 相関系数 -0.89

(音素タイプライタ typewriter) $y = -240.28 * log(x) + 99.83$ 相関系数 -0.89

図 1: 単語グラフ密度と単語認識率の関係 (対話毎)

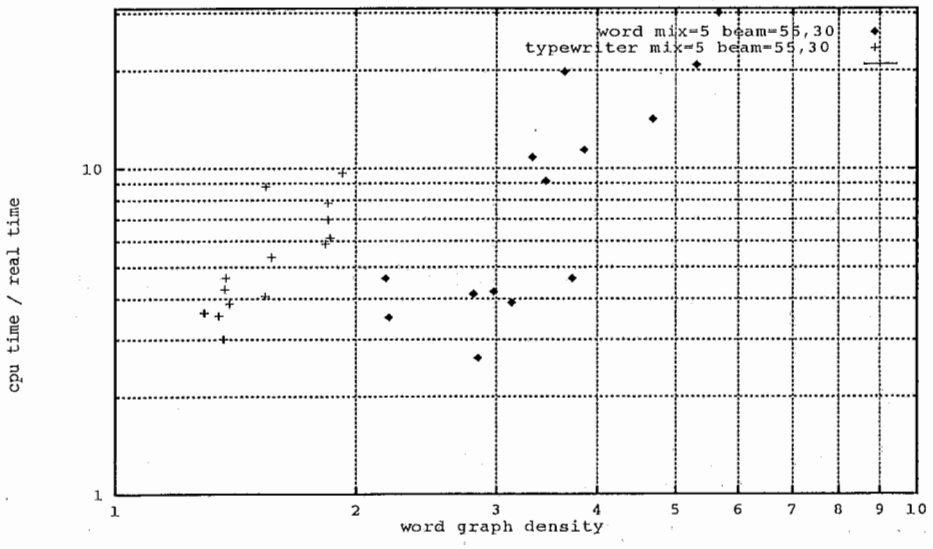


図 2: 単語グラフ密度と CPU 時間の関係 (対話毎)

4.2 考察

図 1,2の結果から、次のことがわかった。

- 単語認識率 (accuracy) は単語グラフ密度に比例して減少する。
- CPU 時間は単語グラフ密度に比例して増加する。

5 言語重み係数・ビーム幅で補正した単語グラフ密度と単語認識率・計算量の関係

前項で単語グラフ密度と認識率・CPU 時間の相関が認められたが、単語グラフ密度は、言語重み係数や、beam 幅でも変化する。言語モデルに依存する 2nd pass において、言語重み係数や、beam 幅を変化させた場合の単語グラフ密度と認識率の関係を調べ、言語重み係数や、beam 幅によらない単語グラフ密度と認識率の関係を導く補正方法を導出してみた。

5.1 実験条件

2nd pass の言語重みと beam 幅の組合せを次に示す通りに変化させて図 3,4の結果を得た。

- 言語重みを 8.0 → 12.0 に増やす
- beam 幅を 30 → 55 に増やす

5.2 考察 (補正方法)

図 3,4の結果より、横軸 log(単語グラフ密度) に対して、言語重み・beam 幅による回帰直線の傾きの変化と元の回帰直線の傾きの比は、言語重み・beam 幅それぞれの変化の比とほぼ一致する。

そこで、言語重み・beam 幅の違いによる補正方法として次の方法が考えられる。

x : 入力単語グラフ密度、y : % Accracy とし、beam 幅 = B, 言語重み係数 = L としたときの、回帰直線を

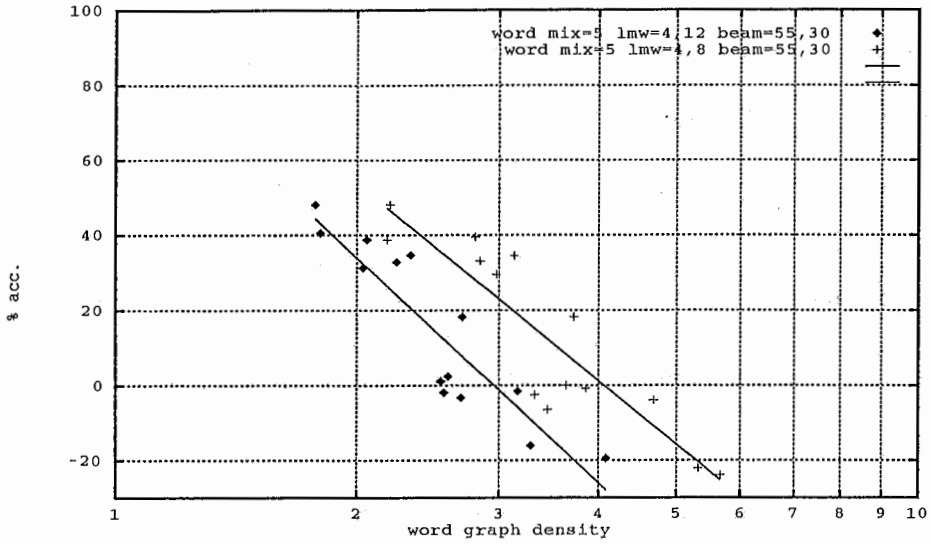
$$y = a * log(x) + b$$

(a,b はグラフから得られる定数)
で与えられるとする。

求めたい入力単語グラフ密度の beam 幅、言語重み係数をそれぞれ、B', L' とすると、得られる回帰直線は

$$y = \frac{B}{B'}\frac{L'}{L} a * log(x) + b$$

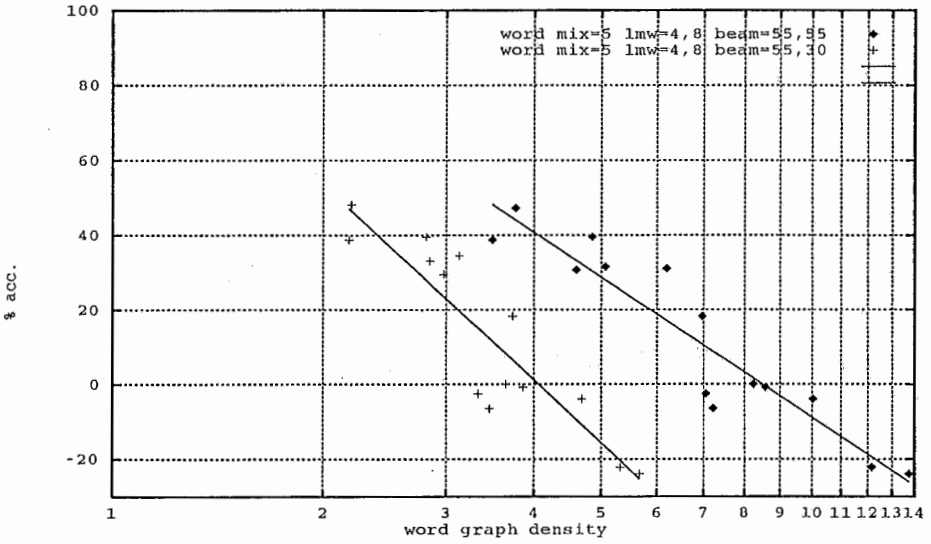
上の式を求めれば beam 幅、言語重み係数に関わらずおおよその認識率を推測することができる。また、補正した結果を図 5に示す。



(2nd pass beam 幅 = 55) $y = -198.71 * log(x) + 93.64$ 相関係数 -0.88

(2nd pass beam 幅 = 30) $y = -174.19 * log(x) + 106.07$ 相関係数 -0.89

図 3: 言語重み係数を変化させた時の単語グラフ密度と単語認識率



(2nd pass beam 幅 = 55) $y = -124.45 * log(x) + 115.63$ 相関係数 -0.92

(2nd pass beam 幅 = 30) $y = -174.19 * log(x) + 106.07$ 相関係数 -0.89

図 4: beam 幅を変化させた時の単語グラフ密度と単語認識率

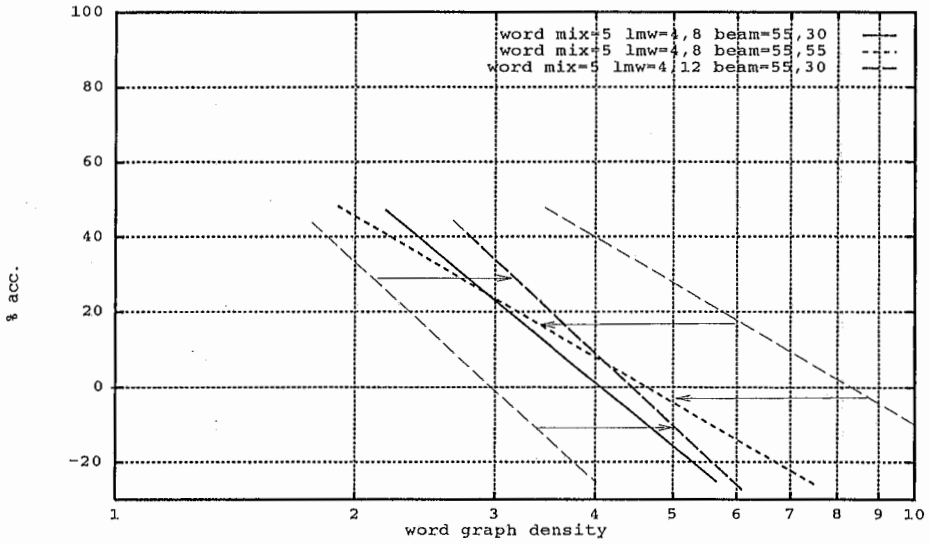


図 5: 言語重み係数・ beam 幅の補正をした回帰直線 (B = 30, L = 8)

6 単語グラフを用いた音響モデルの評価

5節で得られた「単語グラフ密度が小さいほど性能が良い」という結果をもとに混合数の違う音響モデルの性能評価を行なった。

6.1 実験条件

言語重み・ beam 幅を固定して、混合数を 1 , 3 , 5 , 7 , 1 0で変化させる。評価データは男性話者3名で、連続単語認識と音素タイプライタで実験した単語グラフ密度と単語認識率の関係を、図6に示す。また、連続単語認識での混合数と CPU 計算量の関係を図6に示す。

6.2 考察

図6より混合数1は悪く、5・7・10はあまり変わらない傾向が見られた。また、図7より混合数の増加により、音響ゆう度の計算量分に比例して計算量が増加しているが、混合数1・3の CPU 時間では期待される値よりも多くかかっている。

以上の結果から、次のことがいえる。

- 混合数を上げてもある程度以上では計算時間がかかるだけで、認識率は上がりにくい。
- 混合数が少ないと認識率は低下し計算量は相対的に大きくなる。

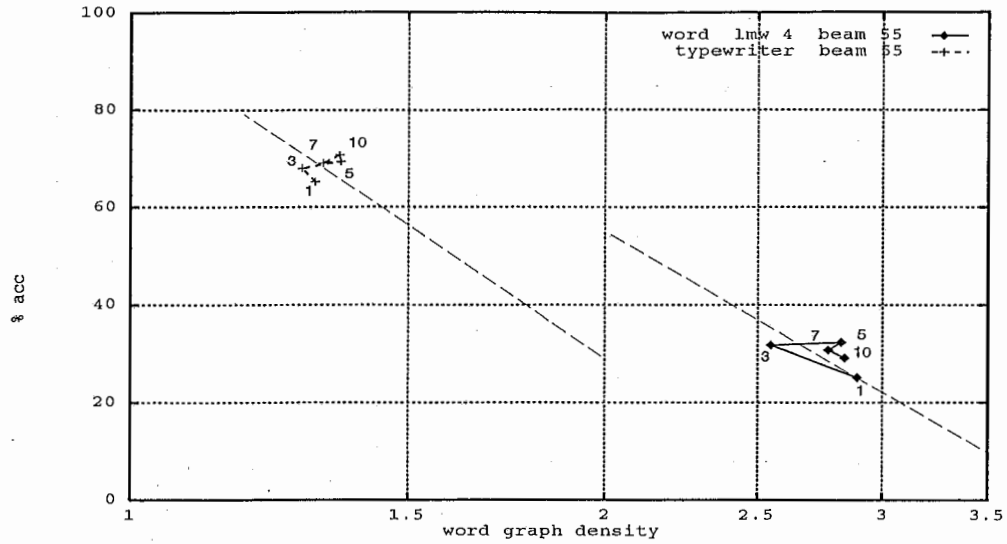


図 6: 混合数の違いによる音響モデルの性能比較

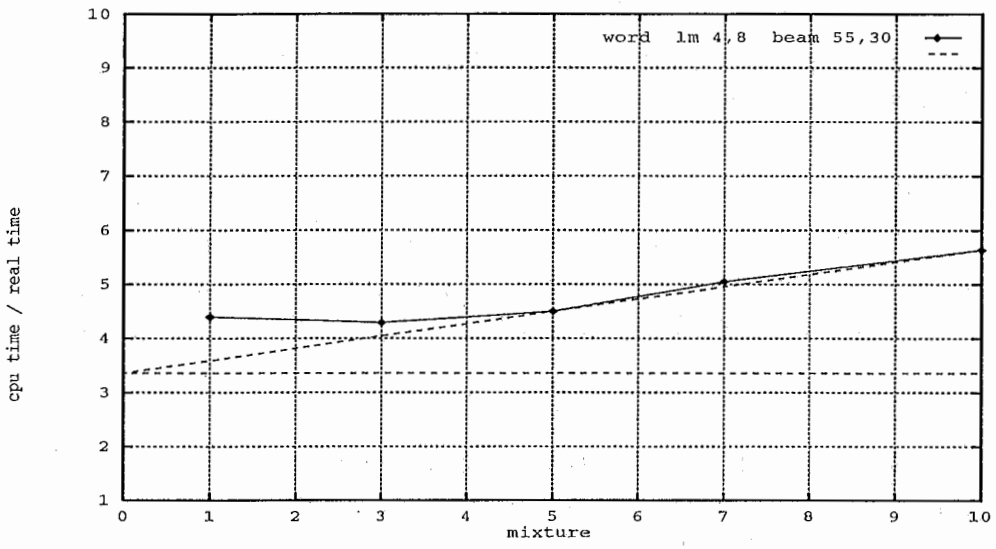


図 7: 混合数と CPU 時間の関係（男性話者 3 名の平均）

7 まとめ

本報告では、単語グラフを用いた multi pass search における認識性能について、単語認識率、単語グラフ密度、計算時間などの多面的な評価法を検討した。結果として、言語重みや beam 幅に関わらず単語グラフ密度から正解がわからない場合のおおよその認識率が予測できることがわかった。

また、混合数の違いによる音響モデルの評価を行なった。その結果、混合数を上げてもある程度以上では計算時間がかかるだけで、認識率は上がりにくく、逆に混合数が少ないと認識率は低下し計算量は相対的に大きくなることがわかった。

今後の課題として、発声毎の実験では、分散が大きく、傾向をとらえることができなかったので、今後この点について検討をしていく。

謝辞

今回の実験にツールを提供して頂いた ATR 音声翻訳通信研究所第一研究室の山本氏、外村氏に感謝の意を表します。

また、快適な実習環境を提供して頂いた ATR 音声翻訳通信研究所第一研究室の研究員諸氏、本実習の機会を与えて下さった ATR 音声翻訳通信研究所の山崎泰弘社長、並びに ATR 音声翻訳通信研究所第一研究室の匂坂芳典室長に深く感謝致します。

参考文献

[1] H. Murveit, J. Butzberger, V. Digalakis, M. Weintraub: "Large Vocabulary Dictation using SRI's Decipher Speech Recognition System: Progressive Search Techniques," Proc. of ICASSP'93, pp.119-122, (1993).

[2] J.L. Gauvain, L.F. Lamel, G. Adda, M. Adda-Decker: "The LIMSI Continuous Speech Dictation Syatem: Evaluation on the ARPA Wall Street Journal Task," Proc. of ICASSP'94, pp.557-560, (1994).

[3] H.Ney, X. Aubert: "A Word Graph Algorithm for Large Vocabulary, Continuous Speech Recognition," Proc. of ICSLP'94, pp.1355-1358, (1994).

[4] P.C. Woodland, C.J. Leggetter, J.J. Odell, V. Valtchev, S.J. Young: "The 1994 HTK Large Vocabulary Speech Recognition System," Proc. of ICASSP'95, pp.73-76, (1995).

[5] 清水, 山本, 政瀧, 松永, 匂坂 " 単語グラフと可変長 N-gram を用いた大語彙自然発話音声認識", 音講論集 1-P-18,H8-3