音声言語情報処理 34－20
(2000. 12. 21)

音声認識精度向上のための信頼度尺度の比較

緒方 淳　　有木 康雄

龍谷大学 理工学部

〒 520-2194 大津市瀬田大江町横谷 1-5
Tel: 077-543-7427
E-mail: ogata@arikilab.elec.ryukoku.ac.jp, ariki@rins.st.ryukoku.ac.jp

あらまし 本報告では，高精度な音声認識を目的とし，ワードグラフから算出される種々の信頼度尺度(Confidence Measure)について比較検討を行った．信頼度尺度の算出法としては，ワードグラフ中の候補数に基づくものや，単語事後確率に基づくものについて検討した．本研究では，信頼度とワードグラフの再構築を行う繰り返しデコーディング法においてそれらを実装し，新聞記事読み上げディクテーションタスクにて評価を行った．

キーワード：大語彙連続音声認識，信頼度，ワードグラフ，単語事後確率

A Comparison of Confidence Measures for Improved Speech Recognition

Jun Ogata　　Yasuo Ariki

Faculty of Science and Technology, Ryukoku University

1-5 Yokotani, Oe-cho, Seta, Otsu-shi, 520-2194 Japan
Tel: +81-77-543-7427
E-mail: ogata@arikilab.elec.ryukoku.ac.jp, ariki@rins.st.ryukoku.ac.jp

Abstract In this paper, we investigate some confidence measures calculated from word graphs for improved speech recognition. In confidence estimation, mainly two methods are compared; one is based on number of hypothesis in word graphs and the other is based on word posterior probabilities. We implemented them in an iterative decoding method based on the confidence estimation and the word graph re-construction, and evaluated them in LVCSR task.

Key words : LVCSR, confidence measure, word graph, word posterior probability

1 はじめに

近年，音声ディクテーション，すなわち大語彙連続音声認識の研究が活発に行われ，様々なシステムで成功を納めている．しかし，実際的な使用においては，音声認識結果は必ずしも信頼できるものではないことや，システムが想定外のことを，ユーザーが喋った場合に対処できないといった問題がある．そのような観点から，音響モデルや言語モデルの高精度化とは別に，音声認識結果(途中結果)の信頼度(Confidence Measure)に基づくアプローチが活発に研究されている．

代表的な研究としては，音響的な信頼度をもとに発話検証を行うというものがある[1]-[3]．また，これとは別の研究として，音声対話システムにおいて，音声認識結果の信頼度を用いることで，より頑健な対話を実現する研究[4]などがある．[5]では，音声認識結果の信頼度を推定することによって，より高精度なオンライン話者適応を実現している．

我々は，認識精度の向上を目的として，このような音声認識結果の信頼度を，既存の大語彙連続音声認識に組み込む方法について検討している[6]．本報告では，単語レベルの信頼度尺度として，種々の算出方法について検討し，それらを以前我々が提案した，信頼度とワードグラフの再構築に基づく繰り返しデコーディング法において比較検討を行う．

2 認識システムの構成と信頼度の算出法

ここでは，本研究で用いたベースラインの認識システムと，検討する種々の単語レベルの信頼度(Confidence Measure:CM)の算出法について述べる．

2.1 認識システムの構成

ベースラインの認識システムとしては，ワードグラフを中間結果とする2-pass 構成のシステムを用いている(図1)[7][8][9]．まず，1st-passで単語bigramを用いたlexical tree searchを行ない，認識結果をもとにワードグラフを作成する．このとき，最もスコアの高い単語にのみback-off接続を行う最尤単語back-off接続を用いることにより，認識精度を落すことなく処理時間を大幅に削減している[8][9]．2nd-passでは，ワードグラフに登録された1st-passの音響尤度とtrigramを用いてリスコアリングを行なう．

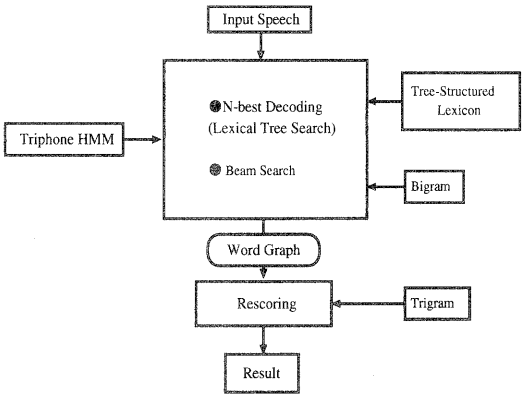

図 1: 認識システム

2.2 ワードグラフからのCMの計算

本研究では，1st-passの認識結果として出力されるワードグラフをもとにCMを計算する．ワードグラフやN-bestリストから算出されるCMは，間接的ではあるが，音響・言語の両方が考慮されていると考えることができる[10][11][12]．したがって，通常のサブワードデコーダー等の併用による音響尤度比に比べて，より高性能なCMが算出可能である[11]．本研究では，ワードグラフからのCMとして以下のように2 種類検討した．

2.2.1 エッジ接続数によるCM

図2にワードグラフの例を示す．グラフのエッジ部分は各単語候補を表しており，ノード部分は単語の分岐点を表している．ワードグラフは，音声認識デコーダーが出力するN-bestの候補をコンパクトにまとめたものであり，グラフのエッジ部分(各単語)の接続数が多い候補ほど，結果的にN-best文候補に多く含まれ，入力発話に対して信頼度が高いといえる．そこで，ここではグラフ中の単語の接続数をもとにCMを求める[6]．

まず，グラフ中の異なり単語リストを$(w_1, w_2, \ldots, w_N)$とする．ここでは，ある単語w_nのCMを求めることにする．グラフにおいて，w_nに対する直前単語からの接続数を$f_{in}(w_n)$，後続単語への接続数を$f_{out}(w_n)$とすると(図3参照)，ある単語w_nの信頼度$CM(w_n)$を以下のように算出する．

$$CM_{connect}(w_n) = \frac{f_{in}(w_n) f_{out}(w_n)}{\sum_{i=1}^{N} f_{in}(w_i) f_{out}(w_i)} \qquad (1)$$

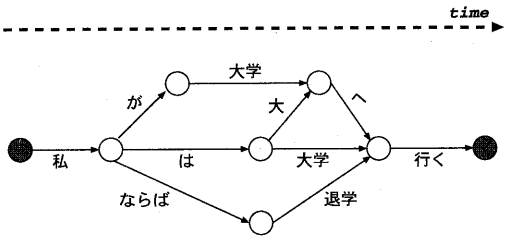

図 2: ワードグラフ

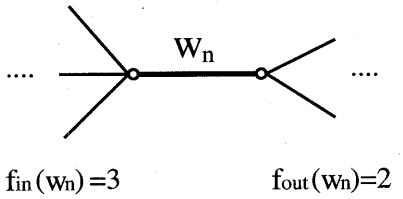

図 3: グラフ中の単語 w_n の接続例

2.2.2 単語事後確率による CM

ワードグラフから各単語の事後確率を計算し，CMを求める．

まず，ワードグラフのそれぞれのエッジに対し，forward-backwardアルゴリズムにより，事後確率を以下のように計算する [13][12].

$$P(w_n|X) = \frac{\alpha(s(W_n))P_{ac}(W_n)P_{lm}(W_n)\beta(e(W_n))}{P(X)} \tag{2}$$

ここで，$s(W_n)$はエッジ(単語)W_nの始端ノード，$e(W_n)$は終端ノードをそれぞれ表す(図4参照)．$\alpha(s(W_n))$はエッジW_nのforward確率を表し，グラフの始端ノードから$s(W_n)$までの全てのパスの音響・言語尤度の総和で求められる．同様に，$\beta(s(W_n))$はbackward確率を表し，$e(W_n)$からグラフの終端ノードまでの全てのパスの音響，言語尤度の総和で求められる．また，$P(X)$はワードグラフ内の全ての文候補の尤度の総和で与えられる．$P_{ac}(W_n)$は単語W_nの音響尤度であり，$P_{lm}(W_n)$はW_nに関する言語尤度(bigram，trigram)である．式(2)により，それぞれのエッジの事後確率が求められると，単語ごとに足し合わせることによって，最終的にCMが以下のように求められる．

$$CM_{posterior}(w_n) = \sum_{w=W_n} P(w|X) \tag{3}$$

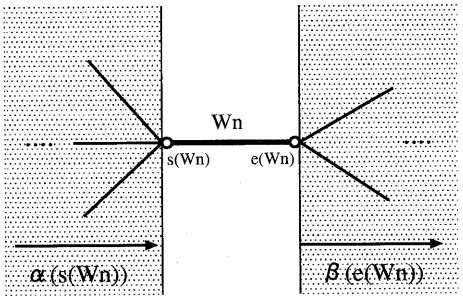

図 4: エッジ事後確率の計算

3 ワードグラフリスコアリングにおける比較

前節で述べた単語レベルのCMをデコーディングに組み込むことを考える．次に，ワードグラフリスコアリング(2nd-pass)においての，CMの比較実験を行う．

3.1 CMを組み込んだワードグラフリスコアリング

ベースラインの認識システムにおいては，ワードグラフリスコアリングは，1st-passで求めた単語音響尤度とtrigramを用いて，単語単位にDynamic Programmingにより探索を進める．本研究では，単語音響尤度とtrigram確率の他に，単語レベルのCMもスコア計算に加える．すなわち，探索中のある単語Wのスコアを$Score(W)$とすると，$Score(W)$は次のように表される．

$$Score(W) = acous(W) + lm(W) + w_{cm} logCM(W) \tag{4}$$

ここで，$acous(W)$は単語Wの音響対数尤度，$lm(W)$は単語Wに関する言語対数尤度，$CM(W)$は単語Wに対するCM，w_{cm}はCMに対する重みである．実際の認識においては，trigramに対しても言語重み，単語挿入ペナルティが与えられる．式(4)は，単語の音響的な確からしさ(音響尤度)と隣接単語との接続関係(言語モデル)に加え，その単語が入力発話に対してどの程度の重要さ(CM)を持っているか考慮したスコアであると考えることができる．

3.2 リスコアリングにおける比較実験

3.2.1 音響モデル

音響モデルとしては，前後の音素環境を考慮したtriphone HMMを用いた．音響モデルの学習には，まずA-TR連続音声データベースa～jセットから6名分のデー

タとその視察ラベルを用いて初期モデルを作成した．次に日本音響学会新聞記事読み上げコーパス(JNAS)のうち，男性話者137名分の21782発話を用いて連結学習を行なった．音響分析条件とHMMのトポロジーを表1に示す．

表 1: 音響分析とHMM

	サンプリング周波数	16kHz
音	特徴パラメータ	MFCC (39次元)
響	フレーム長	20ms
分	フレーム周期	10ms
析	窓タイプ	ハミング窓
	状態数	5状態3ループ
H	タイプ	Triphone HMM
M	混合数	12
M	学習方法	連結学習

3.2.2 言語モデル

言語モデルには，IPAモデル98年度版のうち，語彙数20K，cut-offはbigram，trigramそれぞれに対して4-4のモデルを用いている．言語モデルの学習データは，毎日新聞記事75ヶ月分である[14].

3.2.3 評価用データ

評価用データには，IPA-98-TestSetのうち，男性23名が発声したデータ100文を用いている．未知語率は0.44%である．また，1文あたりの平均発声時間は5.8secである[14].

3.2.4 実験結果及び考察

ワードグラフリスコアリングによる実験結果を表2に示す．ここで，"trigram"は通常のワードグラフリスコアリングを，"$CM_{connect}$"はグラフエッジの接続数によるCM，"CM_{post}-2gram"はbigramをベースにした事後確率によるCM，"CM_{post}-3gram"はtrigramをベースにした事後確率によるCMをそれぞれ表している．

実験結果より，CMを組み込んだリスコアリングは，どのCMにおいても通常のtrigramリスコアリングより認識精度が上回っていることがわかる．ただし，その性能改善は全体的に僅かなものであった．グラフエッジの接続数に基づくCM($CM_{connect}$)と事後確率に基づくCM(CM_{post})では，認識精度の差はほとんどなかった．これは，両者の違いが，音響尤度・言語尤度を直接的に考慮しているか，間接的に考慮しているかということであり，どちらのCMも基本的に本質的な差はなかったものと考えられる．また，bigramベースの事後確率によるCMと，trigramベースの事後確率によるCMでは，リスコアリングにおいては，認識精度の面では，両者に差は見られなかった．

表 2: リスコアリング実験結果

	WAC(%)
3-gram	93.3
$CM_{connect}$	93.7
CM_{post}-2gram	94.0
CM_{post}-3gram	94.0

4 繰り返しデコーディングによる比較

ここでは，以前我々が提案した繰り返しデコーディング[6]において，CMの比較実験を行う．まず，繰り返しデコーディングを実現するための，CMを組み込んだ1st-passの探索法について述べ，次に繰り返しデコーディングの詳細について述べる．

4.1 CMを組み込んだ1st-pass

これまで，beam searchの枠組みに直接CMを組み込んだデコーディング法が提案されており，発話検証，未知語リジェクション等のタスクにおいてその有効性が報告されている[1]．ここでは，音声認識精度の向上を目的として，単語レベルのCMを，1st-passにおけるbeam searchに組み込むことを考える．基本的には，ワードグラフにおいて，CMが推定された単語に関してはその値を探索中のスコアに加え，CMが推定されてない単語(CMを計算した際にワードグラフ内に存在しなかった単語)に関しては，ある一定のペナルティを与えるものとする．

ただし，lexical tree searchにおいては，語頭部分のノードを複数単語で共有するため，単語終端ノード(辞書木のリーフ)に至るまで現単語を特定できない．そのため，bigram確率，あるいは単語レベルのCMを，辞書木内のノードに関して一意に確定することができないといった問題が生じる．探索中のすべてのノードに対して，bigram確率やCMが与えられない場合は，beam幅によるpruningが効果的に働くことができず，認識率を落す原因になる．そこで，bigramに関しては，辞書木の各ノードにおいて，そのノードを共有するすべての単語のうちで最大のbigram確率を予測値として与える，いわゆるbigram factorizationが行われる(図5)．本研究においても，このようなbigram factorizationを行い，CMに関しては，bigram factorizationを行う際に確定した

単語のCM(最大のbigram確率を持つ単語に対するCM)を，辞書木内のノードに与えるようにする．

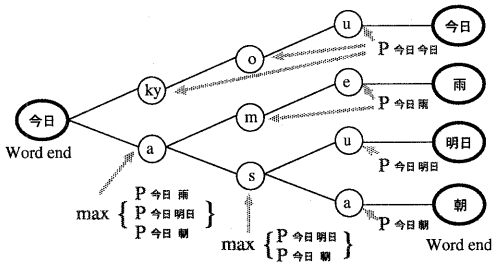

図 5: bigram factorization

4.2 繰り返しデコーディングシステムの構成

前節で述べたCMを組み込んだ1st-passによるワードグラフの再構築と，3.1節で述べたCMを組み込んだワードグラフリスコアリングを用いて，繰り返しデコーディングを行う．そのフローチャートを図6に示す．まずbigramを用いたlexical tree search(1st-pass)を実行し，ワードグラフを作成する．作成されたワードグラフより，単語レベルのCMを推定し，そのリストを保持する．次に，推定されたCMリストをもとに，再び1st-passを実行する．このとき，前節で述べたCMを組み込んだlexical tree searchにより，ワードグラフを再構築する．再構築されたワードグラフをもとにCMを求め，1st-passにフィードバックする．このような繰り返しにより，次々とワードグラフが更新され，徐々に認識精度が上昇することが期待できる．

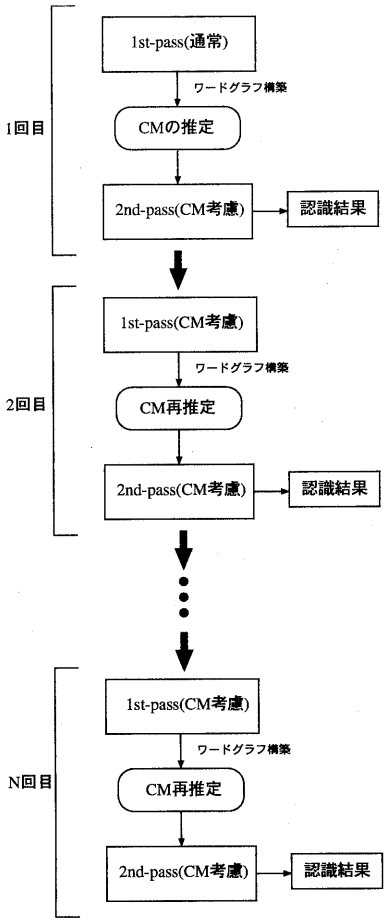

図 6: 繰り返しデコーディングのフローチャート

4.3 繰り返しデコーディング法の評価実験

3.2節と同様の条件にて，繰り返しデコーディング法の評価実験を行った．1st-passにおけるbest候補の認識結果を図7に，2nd-passのリスコアリング結果を図8に示す．ここで，1st-passにおけるbeam幅は1000，繰り返し回数は最大7とした．

実験結果より，1st-passの結果，2nd-passの結果ともに，繰り返しデコーディングを行うことによって，最大約1%の認識精度向上が見られた．図7より，1st-passのlexical tree searchの枠組みにCMを組み込むことは有効であることがわかる．認識精度は，繰り返しを重ねるごとに上昇し，ある一定の回数までくると飽和状態となった．

全体的に，ワードグラフのエッジ接続数に基づくCMと事後確率に基づくCMでは，事後確率に基づくCMの方がより高精度に認識できていた．繰り返しデコーディングにおいては，事後確率に基づくCMのように，音響・言語尤度を直接考慮した，より詳細な指標が有効に働いていることがわかる．

5 おわりに

本研究では，音声認識の精度向上を目的として，認識結果の信頼度尺度(CM)を組み込んだデコーディング法について比較検討を行った．

まず，ワードグラフをもとに単語レベルのCMを推定し，ワードグラフリスコアリング(2nd-pass)における比較実験を行った．その際，信頼度尺度としては，グラフエッジの接続数をもとにしたCM，事後確率をもとにしたCMを用いた．比較実験の結果，通常のtrigramリスコアリングに比べて，CMを考慮することで僅かながら認識精度の改善が見られた．しかし，CMの比較では，

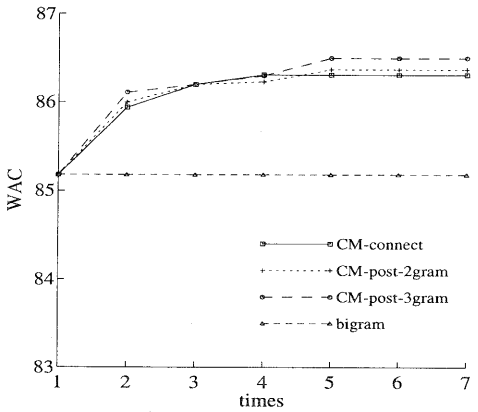

図 7: 繰り返しデコーディング結果 (1st-pass)

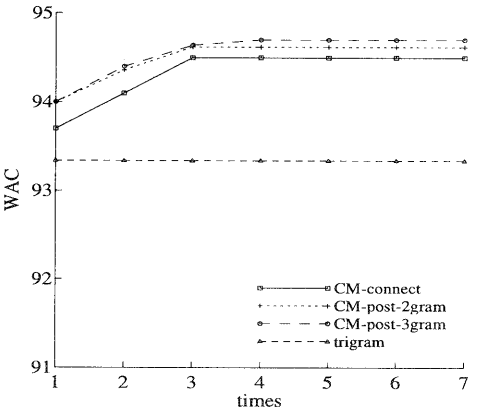

図 8: 繰り返しデコーディング結果 (2nd-pass)

リスコアリングにおいては，両者の差はほとんど見られなかった．

次に，同様の CM を 1st-pass にも組み込み，ワードグラフを再構築することで，繰り返しデコーディング法を実現し，その効果を調べた．実験の結果，1st-pass，2nd-pass ともに，繰り返しを重ねるごとに認識精度は上昇し，ある一定の回数で飽和する傾向が見られた．また，繰り返しデコーディングにおいては，事後確率をもとにした CM がより有効に働き，特に trigram をベースにした事後確率の場合に，最も認識精度が向上した．

今後はより精密な CM を検討することによって，繰り返しデコーディング法により更なる精度向上が期待できる．本報告では，新聞記事読み上げディクテーションタスクにて評価を行ってきたが，今後は，話し言葉，自由発話に対して，提案した繰り返しデコーディング法を適用していく予定である．

参考文献

[1] M.W.Koo, C.H.Lee and B.H.Juang: "A new decoder based on a generalized confidence score", ICASSP'98, pp.213-216, (1998-5).

[2] Sukkar, R.A. and Lee, C.H.: "Vocabulary Independent Discriminative Utterance Verification for Non-Keyword Rejection in Subword based Speech Recognition", IEEE, Trans. Speech & Audio Process.,Vol4, No.6, pp.420-429 (1996).

[3] Kawahara, T.,Lee, C.H. and Juang, B.-H.: "Flexible Speech Understanding based on Combined Key-Phrase Detection and Verification" IEEE, Trans. Speech & Audio Process.,Vol6, No.6, pp.558-568 (1998).

[4] 駒谷和範, 河原達也: "音声認識結果の信頼度を用いた頑健な混合主導対話の実現法", 情処研報, **SLP2000**-30-9, pp.39-44 (2000-2).

[5] T.Zeppenfeld, M.Finke, K.Ries, M.Westphal and A.Waibel: "Recognition of Conversational Telephone Speech Using the Janus Speech Recognition", ICASSP'97, pp.1815-1818, (1997-04).

[6] 緒方淳, 有木康雄: ""信頼度を組み込んだデコーディングによる音声認識の検討", 情処研報, **SLP2000**-32-1, pp.1-6 (2000-7).

[7] 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).

[8] 緒方淳, 有木康雄: "Lexical tree search における探索ネットワーク構造の検討", 信学技法, SP99-142, pp.35-40 (2000-01).

[9] 緒方淳, 有木康雄: "back-off 接続を考慮した大語彙連続音声認識の高速化", 音講論集, pp.43-44 (2000-03).

[10] T.Kemp, T.Schaaf: "Estimating Confidence Using Word Lattice", ICASSP'97, pp.875-878 (1997-04).

[11] D.Willett, A.Worm, C.Neukirchen, G.Rigoll: "Confidence Measures for HMM-Based Speech Recognition", ICSLP'98, pp.3241-3244 (1998-12).

[12] F.Wessel, K.Macherey, H.Ney.: "A Comparison of Word Graph and N-best List based Confidence Measures", EuroSpeech,99, pp.315-318 (1999-09).

[13] G.Evermann, P.C.Woodland: "Large Vocabulary Decoding and Confidence Estimation using Word Posterior Probabilities", ICASSP'2000, pp.2366-2369 (2000-06).

[14] 李晃伸, 河原達也: "大語彙連続音声認識エンジン Julius における A*探索法の改善", 情処研報, **SLP99**-27-5, pp.33-39 (1999-7).