音声言語情報処理 32－1
（2000．7．14）

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

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指し，単語レベルの信頼度(Confidence Measure)を組み込んだデコーディング法について検討し，その評価を行う．ベースとなるデコーダーは2パスの構成をとっており，中間結果としてワードグラフを出力する．信頼度は，ワードグラフをもとに算出し，ワードグラフのリスコアリングによってその効果を調べた．また，本研究では，信頼度を組み込んだ探索法，及び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を提案する．提案する繰り返しデコーディング法を，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い，その有効性を確認した．

キーワード：大語彙連続音声認識，信頼度，ワードグラフ

A Study on Confidence Based Decoder 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 study on a confidence based decoding method for improved speech recognition, and evaluate it. A word graph is constructed as an intermediate result in our 2-pass decoder. Confidence values are calculated from the word graph, and evaluated in word graph rescoring. In this study, we propose an iterative decoding method incorporating a confidence based search and word graph reconstruction. We evaluated the proposed method in LVCSR task. As a result, a slight improvement was observed in terms of the word accuracy compared to the standard 2-pass method.

Key words : LVCSR, confidence measure, word graph

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．

本報告では，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する．まず，認識システムの中間結果であるワードグラフに基づいて，信頼度を算出する方法について述べ，ワードグラフのリスコアリングによってその効果を調べる．本研究では，信頼度を組み込んだlexical tree search(1st-pass)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を提案し，新聞記事読み上げディクテーションタスクにて評価実験を行い，その効果について報告する．

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，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)[6][7][8]．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[7][8]．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．

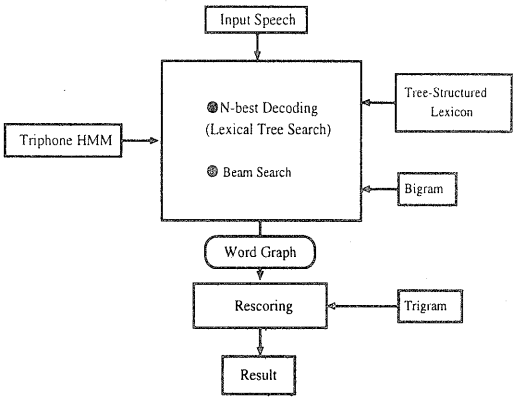

図 1: 認識システム

2.2 ワードグラフからのCMの計算

本研究では，CMの計算を，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行う．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[9][10][11]．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[10]．

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，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．

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を求めることとする．w_nのグラフ中におけ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．

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

3 CMを組み込んだデコーディング

本研究では，音声認識の精度向上を目的として，2.2節で述べた単語レベルのCMを組み込んだデコーディング法について検討する．まず，ここでは，CMを組み込んだワードグラフリスコアリングについて述べ，その効果を調べる．

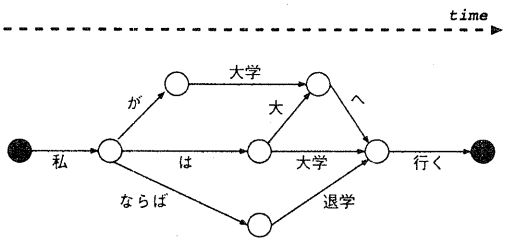

図 2: ワードグラフ

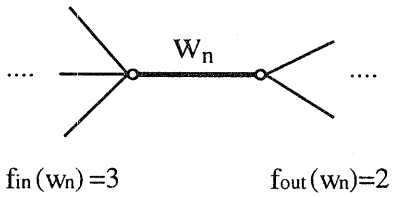

図 3: グラフ中の単語 w_n の接続例

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．ここで，探索中のある単語Wのスコアを$Score(W)$とすると，

$$Score(W) = acous(W) + p(W|U,V) + w_{cm}CM(W) \quad (2)$$

となる．このとき，$acous(W)$は単語Wの音響尤度，$p(W|U,V)$はtrigram，$CM(W)$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．(2)式は，単語の音響的な確からしさ(音響尤度)と近隣単語との接続関係(言語モデル)に加え，その単語が入力発話に対してどれほどの重要さ(CM)があるかを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

通常のリスコアリングとCMを組み込んだリスコアリングとの比較実験を行った．以下に実験条件を示す．

3.2.1 音響モデル

音響モデルには，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ATR連続音声データベースa～jセットの6名分のデータの視察ラベルを用いて初期モデルを作成し，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．音素数は無音も含めて41種類である．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[12]．

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[12]．

3.2.4 実験結果及び考察

ワードグラフのサイズを様々に変えたときのリスコアリング結果を図4に示す．ここで，WACは単語認識精度(Word Accuracy)を表し，WGDはワードグラフ密度(Word Graph Density)を表す．WGDは，1st-passで求めたワードグラフのサイズを表す指標であり，以下のように求められる[13]．

$$WGD = \frac{\text{ワードグラフ内の総単語数}}{\text{正解発話の単語数}} \quad (3)$$

グラフのサイズは，1st-passにおけるbeam幅を様々に変化させることによって調整した．図中，"CM"はCM

を組み込んだリスコアリングを示し，"without CM"はtrigramによる通常のリスコアリングを示す．

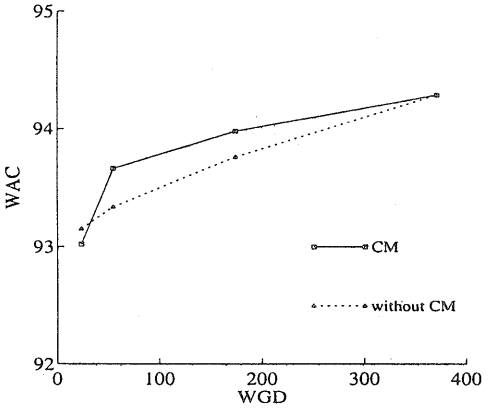

図 4: ワードグラフリスコアリング結果

実験結果より，ほとんどのグラフサイズにおいて，CMを組み込んだリスコアリングは，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．したがって，CMを組み込んだリスコアリングは有効であるといえる．しかし，その差はわずか(最大約0.3%程度)であり，更に精密なCMを推定する必要があろう．

4 CMとワードグラフの再構築による繰り返しデコーディング法

本研究では，これまでに述べたCMを用いてワードグラフの再構築を行う，繰り返しデコーディングという認識手法を提案する．基本的な考え方は，まず通常の1st-passを実行し，ワードグラフを作成する．次に，2.2節で述べたCMを計算し，再び1st-passへフィードバックし，CMを組み込んだ1st-passによりワードグラフを再構築する．このようなサイクルを繰り返しながら，徐々に認識精度の向上を図るものである．まず，繰り返しデコーディングを行うため，CMを組み込んだ1st-passの探索法について検討し，次に提案法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．

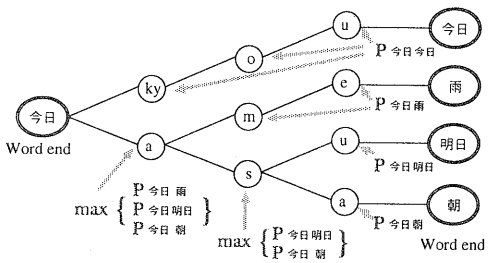

図 5: bigram factorization

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．

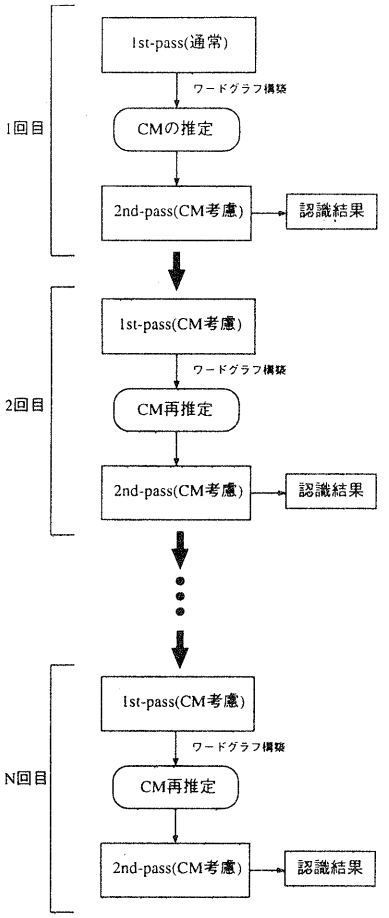

図 6: 繰り返しデコーディングのフローチャート

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．また，図中，"iterative"は提案手法である繰り返しデコーディングを表し，"bigram"，"trigram"はそれぞれ，通常の1st-passの認識結果，通常のワードグラフリスコアリング結果を表している．

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．本実験においては，繰り返し3回目あたりでワードグラフが収束し(CMがほとんど更新されず)，認識精度は一定となった．その原因として，本研究で用いたワードグラフをもとにしたCMが非常に単純なものであったためと考えられる．したがって今後，より精密なCMの推定を行う必要がある．

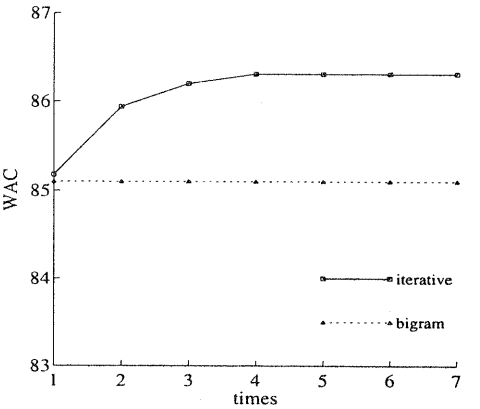

図 7: 繰り返しデコーディング結果(1st-pass)

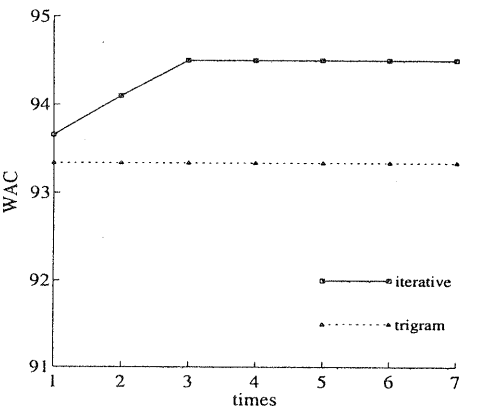

図 8: 繰り返しデコーディング結果(2nd-pass)

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(CM)を組み込んだデコーディング法について検討を行った．

まず，ワードグラフをもとに単語レベルのCMを推定し，それを2-passシステムのリスコアリングに組み込み，その効果を調べた．通常のtrigramによるリスコアリングとの比較実験の結果，認識精度において，わずかながらであるが向上した．

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．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る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] 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).

[7] 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).

[8] 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).

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

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

[11] 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).

[12] 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).

[13] 清水徹, 山本博史, 政瀧浩和, 松永昭一, 匂坂芳典 : "大語い連続音声認識のための単語仮説数削減", 信学論, Vol.79-D-II, No.12, pp.2117-2124 (1996).