音声言語情報処理 16－4
（1997. 5. 26）

単語 N-gram と段階的探索を用いた大語彙連続音声認識

李晃伸　河原達也　堂下修司
京都大学 工学研究科 情報工学専攻
〒606-01 京都市 左京区 吉田本町

あらまし

大語彙の連続音声認識においては，広大な探索空間を効率良く探索するための手法が必要不可欠である．本研究では，統計的言語モデルである単語 N-gram を用いた連続音声認識における段階的探索手法について考察を行ない，京都大学で実装中の 5,000 語の認識システムの構成と動作例を報告する．まず使用する単語 N-gram 言語モデルの諸元について述べ，話題別 N-gram の導入についても検討する．次に種々の段階的探索手法について比較を行う．特に，高い探索精度を実現するために，単語グラフ等を介さない効率的なマルチパス探索手法を提案する．

和文キーワード　大語彙連続音声認識，単語 N-gram，段階的探索，話題別 N-gram，毎日新聞データベース

Multi-Pass Search for Large Vocabulary Continuous Speech Recognition using Word N-gram

Akinobu Lee　Tatsuya Kawahara　Shuji Doshita
Department of Information Science, Kyoto University
Sakyo-ku, Kyoto 606-01, Japan
e-mail: ri@kuis.kyoto-u.ac.jp

Abstract

In large vocabulary continuous speech recognition (LVCSR), an efficient search algorithm is essential. We examine multi-pass search methods for LVCSR using word N-gram, and overview 5000 word CSR system currently developed at Kyoto university. First, the specifications of word N-gram stochastic language model estimated from Mainichi newspaper database are shown. Topic-dependent N-gram model is also studied. Then, we compare various multi-pass search methods, and propose an efficient multi-pass search that does not use intermediate representation like word graph and realizes accurate search.

Keywords　large vocabulary continuous speech recognition, word N-gram, multi-pass search, topic-dependent N-gram, Mainichi newspaper database

1 はじめに

近年，大語彙連続音声認識の研究が多くの研究機関で進められている．特に新聞記事の読み上げタスクは，大量のコーパスを電子化された新聞記事から容易に得ることができることから統計的言語モデルを用いた大語彙連続音声認識の研究に適しており，海外の様々な研究機関で盛んに研究されている [1][2].

日本語においては，通常単語の分かち書きが行なわれないために単語 N-gram の推定には形態素解析が必要であるなどの理由から研究が立ち遅れていたが，近年大語彙音声認識の研究のための研究機関を越えて共有できる言語データベースの構築が進められている [3][4] など，日本語における本格的な大語彙連続音声認識の研究 [5][6] のための環境が整いつつある．

大語彙の連続音声認識においては，出現し得る仮説の数は極めて膨大であるため，無駄な仮説を展開することがなく効率よく最尤仮説を見つけるための探索手法が重要である．大語彙の条件下でより効率よく精度の高いモデルを用いる方法として，探索を複数のパスに分け，粗いモデルから高精度のモデルまで順に段階的に適用するという段階的探索の手法がある [7].

本研究では単語 N-gram を用いた大語彙連続音声認識における種々の段階的探索手法を比較し，新たな探索アルゴリズムを提案する．まず認識に用いる言語モデルとして，毎日新聞データベース [3] から構築した単語 N-gram 言語モデルについて諸元を示す．話題別 N-gram の構築についても言及する．そして種々の段階的探索手法について比較を行い，その上でヒューリスティック探索において展開仮説数を効果的に削減する手法を検討する．トレリス探索は，単語グラフなどの中間結果を単語の形で介する手法に比べて精度の高いスコアが得られるという長所があるが，一方で探索空間が語彙に比例して大きくなり，大語彙では破綻する．これに対して，2-gram をヒューリスティック制約として 3-gram を用いて best-first に探索を行うことにより無駄な仮説を展開せずに効率よく解を探索する．加えて，第 1 パスの音響的マッチングの結果に基づいて予測単語を絞り込む手法を提案する．最後に提案手法を実装したシステムの構成を述べ，毎日新聞記事読み上げタスクにおける動作例を報告する．

2 単語 N-gram 言語モデル

単語 N-gram 言語モデルは統計的言語モデルの一つであり，大量のテキストデータから統計的に単語列の頻度を推定する．n 個の単語からなる文仮説 $W = w_1, w_2, ..., w_n$ の生起確率 $P(W)$ は一般に次式で与えられる．

$$P(W) = P(w_1)P(w_2|w_1)P(w_3|w_1^2) P(w_4|w_1^3)...P(w_n|w_1^{n-1})$$

ただし (w_i^j) は $\{w_i, w_{i+1}, ..., w_j\}$ を表す．任意の長さの単語列を考慮するのは統計上不可能なので，参照する過去の履歴を一定長 N に限定する (単語 N-gram).

$$P(W) = P(w_1)P(w_2|w_1)P(w_3|w_2)... P(w_i|w_{i-2}^{i-1})...P(w_n|w_{n-2}^{n-1}) \quad (N = 3\text{の場合})$$

単純な N-gram の構築には膨大な統計量が必要であり，学習データの絶対的不足から，統計量のないデータに対しては平滑化 (スムージング) を行うことで近似的に頻度を与えることが多い．

大語彙音声認識システムを構築するに当たって，単語 2-gram と 3-gram の構築を行った．学習と評価用のコーパスとして毎日新聞 CD-ROM の 91 年から 94 年までの記事データベースを用いた．学習には 91 年から 94 年 9 月までの記事を用い，94 年 10 月から 12 月までの記事を用いて評価する．その形態素解析結果としては，RWCP が提供している毎日新聞形態素解析データを使用する．なお学習元の新聞記事は括弧などの読み上げが困難な記号が存在するが，N-gram 構築に際してそれらを排除するなどの処理は行っていない．語彙は単語カバー率と認識器で扱い得る語彙数の制限とを考え合わせ，出現頻度の上位 5000 語とした．また構築後の N-gram の大きさと頻度の推定精度の兼ね合いから，学習時に統計量が一定頻度以下の単語の組 (tuple) を切り捨る cut-off を行った．そのしきい値は N (2,3-gram に対してそれぞれ 2,3) とした．

表 1: 単語 N-gram の諸元

種類	tuple 数	testset での評価		
		perp.	補正 perp.	cover.
1-gram	5000	238.6	1002.0	83.3%
2-gram	562365	57.2	254.5	67.6%
3-gram	2181112	32.7	144.6	46.1%

(注) 未知語の種類数： perp.=1, 補正 perp=25000

構築した N-gram 言語モデルの諸元を表 1に示す．なお N-gram 中にない単語 tuple の頻度推定には二項事後分布に基づく back-off 平滑化を用いる [8]．また語彙でカバーされない単語は全て未知語として扱う

表 2: 話題別 N-gram の testset perplexity と coverage (N=3)

記事分類		学習セットの形態素数	話題別 3-gram			全体の 3-gram		
			perp.	補正 perp.	cover.	perp.	補正 perp.	cover.
話題別	政治 (01-03)	23812719	36.5	119.0	49.8%	35.1	116.2	51.9%
	国際 (07)	6303389	36.5	131.0	42.4%	35.8	129.0	49.6%
	経済 (08)	6935160	35.6	137.4	41.4%	36.1	139.8	47.5%
	文化 (14)	1523307	43.2	274.0	25.4%	34.7	210.4	38.7%
	科学 (16)	483033	55.1	252.6	24.0%	39.6	169.3	43.5%
	芸能 (18)	2306740	32.4	223.6	28.5%	34.7	227.1	37.0%
	スポーツ (35)	7707627	20.2	92.6	46.0%	25.2	113.3	48.2%
	社会 (41)	24834446	29.5	142.4	43.4%	30.4	144.9	45.1%
全体		102202879	—			32.7	144.6	46.1%

が，この場合 tuple のカバー率が下がるにつれてテストセットパープレキシティが下がるという矛盾が生じる．そこで未知語をカテゴリ UNK として扱いその種類数 $unknum$ を仮定することで補正を行う．1-gram における補正後の未知語の出現確率は，未知語カテゴリの頻度 $C(UNK)$ から以下の式で求める。

$$P(UNK) = \frac{C(UNK)}{W \cdot unknum} \quad (W : \text{総単語数})$$

2-gram 以上においても同様である．ここでは，未知語の出現頻度が上位 5000 語中の最低頻度以下となるように，種類数を 25000 とした．表 1には参考のため補正なしの場合のパープレキシティを併記してある．

3 話題別単語 N-gram

本研究で扱う新聞記事読み上げタスクは，一般の日常会話タスクと比べて以下の特徴を持つ．

- 毎日データベースには元記事の掲載面ラベルが付与されており，話題別分類が容易に行える．
- 一発話ごとの話題のドメインが明確である．

したがって，話題ごとに記事を分類しそれぞれ独立に N-gram の tuple の頻度を推定することで，話題別の知識や意味を反映したよりドメインに特化した言語モデルが得られると考えられる．

また，近年ニュース音声の話題同定の研究がなされている [9]．これは入力となるニュース朗読音声に対して認識処理を行い，抽出されたキーワードから話題を同定する試みである．この話題同定機構と話題別の言語モデルによる認識器を組み合わせることで，入力音声を認識しながらその話題のドメインを同定し，その話題に対応する N-gram を選択していくことでより精度の良い認識が実現できると期待される．

そこで，話題別の単語 N-gram の構築を行った．まずデータベースを各記事にタグづけされている掲載面コードに基づいて分類する．そのうち掲載面と話題の対応が比較的とれていると思われる分類について，それぞれの学習セットから 3-gram の頻度推定を行う．ただしコード 01 から 03(元の新聞記事の 1 面〜3 面) は，「政治」に関する話題としてまとめた．語彙は全体の N-gram と同一とし，cut-off は 2,3-gram に対して 1, 1 とした．その他の推定の条件は全体で構築する場合と同一である．

この話題別 N-gram の各話題ごとのテストセットに対する評価を表 2に示す．全体の N-gram での各値と比較すると，幾つかの分野には話題別 N-gram の効果が認められたが，多くは全体の場合とほとんど変らなかった．これは一分野当りの学習セットが少なすぎたためであると考えられる．

4 連続音声認識のための探索手法の比較

連続音声認識は，与えられた制約の下で入力音声に対し最も尤度の高い単語列のパスを発見する探索問題であるととらえられる．言語モデルの精度を高めるほど精度の高い認識が期待できるが，その構成や参照が複雑になるに伴って探索のコストは膨大になる．そこで，探索を複数のパスに分け，まず入力音声に対して粗い比較的簡単なモデルを適用して中間結果を出力し，それを元により高精度でコストの高いモデルを用いて仮説候補を探索するという段階的探索手法が有効である．以下に幾つかの段階的探索手法について，パス間の中間結果の形式に基づいて分類し比較する．

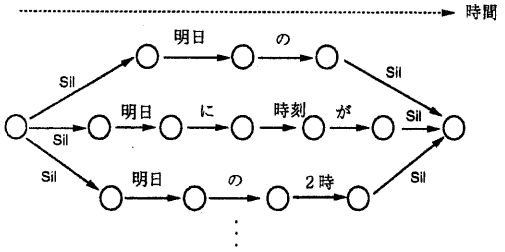

図 1: N-best 候補

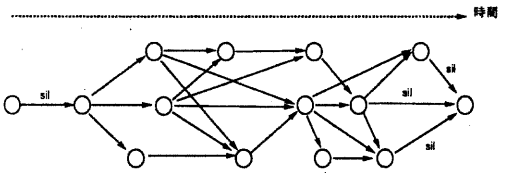

図 2: 単語グラフ

N-best candidates 最初のパスで認識処理を行い，結果として尤度の上位 N 個の文候補を出力する (図 1)．次のパスで高次の言語モデルを用いて尤度の再計算を行う．段階的探索手法としては最も単純な手法である．第 2 パスの探索空間は N 個の文仮説をあらわす単語列である．類似した文候補が大量に得られるので，最初のパスで有意な結果を得るためには N の値を大きくする必要がある．

word graph 最初のパスで認識処理を行いながら，途中部分文仮説を各単語の終了時刻を決定しながら単語単位で残していく．その際仮説間でそれまでの単語列が同じ部分を共有化することで，単語を arc とする仮説のグラフを生成する (図 2)．次のパスでそのグラフ上を高次の言語モデルを用いて探索を行う．第 2 パスの探索空間はこの単語グラフ上である．N-best 候補を単語ごとにマージしたものととらえることができ，N-best 形式に比べコンパクトで高い表現力を持つ．

単語グラフでは，仮説中の単語の開始時間と終了時間を第 1 パスで一意に固定する．このため第 2 パスにおいて各文仮説の言語スコアを計算する際に，仮説の文脈に依存する単語の境界時間の変動を考慮することができず，真のスコアとの間に誤差が生じる．

また word graph の構築においては，効率よく仮説処理を行うための手法として，直前の単語さえ同じであれば仮説をマージしてスコアの高い仮説のみ残す

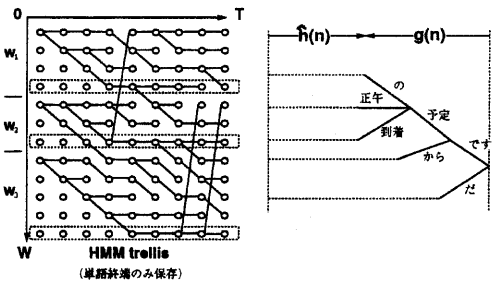

図 3: トレリスとヒューリスティック探索

という単語対近似が有効である．しかし日本語において N-gram 言語モデルの単位を形態素とする場合，1，2 音素からなる短い単語が非常に多いため単語対近似による誤差が大きくなるという問題がある [7]．

trellis 最初のパスの結果を HMM トレリスのまま残し，ヒューリスティックとして次のパスの探索に用いる．第 1 パスは語彙を展開した一つの HMM ネットワーク上で認識処理を行い，トレリスを生成する．第 2 パスは各文仮説の評価値として，第 1 パスで求めたトレリスを接続して得られる尤度を用いる．これにより未探索部の尤度を推定しながら探索を行う．第 1 パスと第 2 パスは逆方向に行う必要がある (図 3)．

第 2 パスで文仮説に対する HMM トレリスを逐一再計算するので word graph より精度の高い仮説尤度が得られる．トレリス計算のオーバヘッドから 1 仮説ごとの計算コストは大きいが，ヒューリスティックによって best-first に展開が進めば少ない展開仮説数で済み，結果的にコストは小さくなる．

しかし，すでに述べた 2 つの手法においては中間結果によって探索空間が限定されその上で第 2 パスの探索が行われるのに対し，この手法では中間結果は展開仮説数の削減に寄与しない．探索の言語的制約として N-gram を用いる場合，仮説を展開するごとに語彙数分の仮説を展開しなければならないという問題が生じる．

5 ヒューリスティック探索における音響マッチングに基づく探索空間の削減

前節で述べたように，トレリスを中間結果とするヒューリティック探索手法では仮説の展開単語として，その仮説の文脈と関係なく語彙中の全単語を候補とする必要がある．これは言語モデルである単語 N-gram が確率的統計言語モデルであり全ての単語連鎖

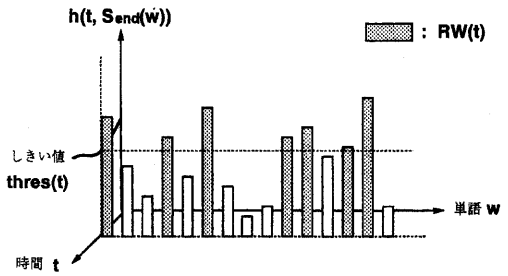

図 4: 第1パスの結果からの出現単語セットの抽出

を確率的に与えるためであるが，単純に語彙中の全単語を常に展開していく方法は大語彙の条件下においては計算量の面から現実的でない．

発話文が全て N-gram 中に存在する tuple のみからなると仮定するようなタスクにおいては，展開仮説は直前の (N-1) 単語との tuple が存在するような単語のみを検索して展開すれば良く，探索空間は非常に限定できる．しかし本研究の対象であるニュース記事読み上げ音声タスクにおいて testset に対する 3-gram のカバー率は 46.1% であり，カバーされない tuple を含む発話は探索に失敗することからも，この仮定は厳しすぎて不適当である．

そこで，trellis を用いる方法において第1パスの認識処理結果から展開単語をあらかじめ絞り込む手法を提案する．第1パスのトレリス計算時にビーム探索を行うが，その際に各時間において，その単語の終端にあたるノードがビーム内に残っている単語集合を抽出しておく．単語を w，その終端にあたる HMM の状態を $S_{end}(w)$，時間 t における状態 s の尤度を $h(t,s)$，ビームのしきい値を $thres(t)$ とすると，この単語集合 $RW(t)$ は以下の式で表される (図 4)．

$$RW(t) = \{w : h(t, S_{end}(w)) > thres(t)\}$$

この $RW(t)$ は，第1パスの結果のトレリス内に残った単語であり，これを第2パスにおける展開単語として利用する．

第2パスにおいて単語を展開する際，現在の仮説の終端時間 t_e を決定し，そこに対応する出現単語セット $RW(t_e)$ を展開単語とする．単語同期の探索では t_e は一意に定めることが難しいが，HMM の継続時間長による推定などが考えられる．

さきほどの N-gram 中に存在する tuple のみの接続を許す手法と組み合わせる．つまり展開単語は，その仮説の最後の (N-1) 単語との tuple が存在するような単語集合と，上記の提案手法により求められる単語集合との和集合とする．

以上の展開単語の選択手法は原理的には word graph 生成における単語候補抽出と等価である．違いは，全ての時間 t に対して $RW(t)$ を求めておき，第2パスで仮説に対してトレリスの再計算を行って最尤の t を定めることができる点である．結局，この手法は結果的に単語グラフで単語の始終端にある程度の自由度を持たせたものといえる．

6 認識システムの構成

以上の提案手法を用いて，N-gram 言語モデルに基づく大語彙連続音声認識システムを構築した．探索は2つのパスで構成され，第1パスを入力音声と同じ向き (left-to-right) に処理し，その中間結果を受けて第2パスは逆向き (right-to-left) に探索を行う．以下に各探索パスの処理を詳しく述べる．

第1パス 言語モデルは 2-gram を用いる．辞書内の単語の HMM を展開して並列に並べ，各単語間の遷移を 2-gram による接続確率に従って付け加えることで1つの HMM ネットワークを構成する．これを入力音声に対しフレーム同期のビームサーチを行い，トレリスを生成する．また同時に各時間において，単語の末尾にあたるノードがビーム内に残っているような単語のリスト $RW(t)$ を保存しておく．

辞書は単語の先頭から同じ音素を共有することで木構造化して用いる．単語の prefix を共有するため，単語間の 2-gram による接続確率を静的に記述できない．そのため，言語スコアはトレリス計算時に Viterbi パスが単語の末尾に達したときにはじめて評価を行う．

第2パス スタックデコーディングを用い，第1パスとは逆向きに単語同期に探索を行ない，最尤単語列の推定を行う．仮説の展開は以下のようにして行なう．まず仮説スタックを参照して，展開すべき仮説をスタックから取り出す．この仮説の最後の単語に対するスコアを計算し，その際に仮説の終端時間 t_e を推定する．第1パスの結果から単語セット $RW(t_e)$ をこの仮説に対する予測単語として，それぞれに対して 3-gram に基づき言語スコアを再計算する．これにその仮説の最後の (N-1) 単語との tuple が存在するような単語集合を追加し，第1パスの結果のスコアを加えて新たな仮説の評価値とする．これらのスコアをソートして順次スタックに積んでいく．

7 システム動作例

この文認識システムの動作例を以下に示す．入力音声として，[4] に基づく毎日新聞記事読み上げデー

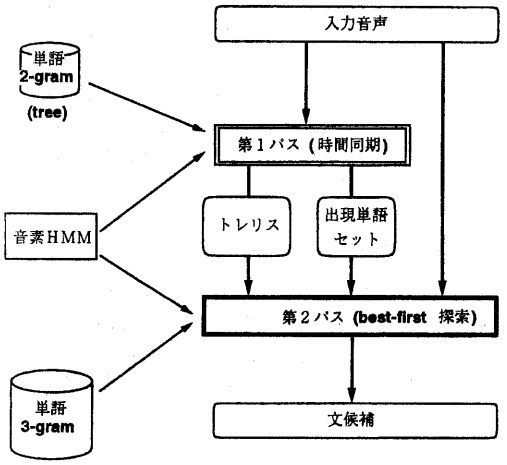

図 5: 認識システムの構成

タベースの調査・収録の協力のために京都大学において収録した新聞音声読み上げデータのうち，男性の読み上げ文から数文を選んで認識を行った．

例1：「多数党の意思がチェックされるからだ．」
NS ta suto boo hi ga cheku sa reru ka da NS
例2：「イラク軍がクウェート国境に終結するなど緊張が高まっているため．」
NS iraku ga kueeto kokyoo ni hi uke NS suN do kiN kyoo waka naki nu ta N NS

8 まとめ

本稿では，単語 N-gram を用いた大語彙連続音声認識における種々の段階的探索手法を比較し，その上でヒューリスティック探索において展開単語を第1パスのビーム内に残った単語に限定する手法を提案し，認識システムの実装を行った．word graph を用いる認識手法は計算効率が高く N-best 候補を用いる手法に比べて高い表現力を持つが，単語対近似などによって仮説スコアに誤差が生じる．一方トレリス計算の手法はスコア精度は高いが，トレリスの再計算や探索空間の広さに問題がある．この問題はスコア精度と計算効率のトレードオフであるといえる．また仮説の終了時間の決定については種々の方法を試す必要がある．

システムを動作させた結果，今回使用した単語 N-gram は日本語特有の問題をかかえていることが判明した．すなわち形態素を「単語」として認識単位としているので「の」「は」「で」などの非常に短い単語が多く，またこれらの同音語も単語頻度の上位を占める．測定したところ上位 5,000 語内の同音語数は 1334 語であった．これらの仮説が上位仮説を占めてしまい仮説のスタックあふれが生じていた．

今後は，単語 N-gram の改善と共に，毎日新聞記事読み上げデータベースを用いた本格的な認識実験を通して提案手法の評価を行う．特に word graph 手法との実験的評価による比較を行う予定である．

謝辞

毎日新聞データベースにおける形態素解析結果を作成し提供して頂いた RWCP データベースワークショップ・テキストグループに感謝致します．形態素出現頻度に関する統計データを提供して下さった電総研の伊藤克亘氏に感謝致します．また，日頃御討論いただく京都大学堂下研究室の皆様に感謝致します．

参考文献

[1] P.C.Woodland, M.J.F.Gales, D.Pye and V.Valtchev: The HTK Large Vocaburary Recognition System for the 1995 ARPA H3 Task, *Proc. DARPA Speech Recognition Workshop*, pp. 99–104 (1996).

[2] J.L.Gauvain, L.Lamel, G.Adda and D.Matrouf: The LIMSI 1995 Hub3 System, *Proc. DARPA Speech Recognition Workshop*, pp. 105–111 (1996).

[3] 武田一哉, 伊藤克亘, 松岡達雄, 竹沢寿幸, 鹿野清宏: 大語彙連続音声認識研究のためのテキストデータ整備, 情報処理学会研究報告, 96-SLP-11-9 (1996).

[4] 伊藤克亘, 武田一哉, 竹沢寿幸, 松岡達雄, 鹿野清宏: 大語彙連続音声認識のための読み上げ文の選択, 音響学会春季論文集, Vol. 2-Q-7, pp. 153–154 (1997).

[5] 大附克年, 吉田航太郎, 松岡達雄, 古井貞熙: 高次 n-gram を用いた大語彙連続音声認識の検討, 音響学会春季論文集, Vol. 2-6-2, pp. 47–48 (1997).

[6] 周旻, 堤真理子, 中川聖一: 確率モデルによる大語彙連続音声認識の評価, 情報処理学会研究報告, 96-SLP-11-6 (1996).

[7] 野田喜昭, 松永昭一, 嵯峨山茂樹: 単語グラフを用いた大語彙連続音声認識における近似演算手法の検討, 信学技報, SP96-102 (1997).

[8] 川端豪, 田本真詞: 二項事後分布に基づく n-gram 言語モデルの back-off 平滑化, 音響学会秋季論文集, Vol. 1-2-14, pp. 27–28 (1995).

[9] 横井謙太朗, 河原達也, 堂下修司: 単語の共起情報を用いたニュース朗読音声の話題同定機構, 電子情報通信学会技術研究報告, SP96-105 (1997).