音声言語情報処理 27−18
(1999. 7. 24)

N-gram モデルのエントロピーに基づくパラメータ削減に関する検討

踊堂 憲道　鹿野 清宏　中村 哲

奈良先端科学技術大学院大学 情報科学研究科

〒630-0101 奈良県生駒市高山町 8916-5

E-Mail:{norimi-y,shikano,nakamura}@is.aist-nara.ac.jp

概要　大語彙連続音声認識 (ディクテーション) 技術は，キーボード入力の省力化や，様々な環境下における人間とコンピュータ間の音声インタフェースの実現のために必要不可欠な技術であり，活発に研究が行なわれている．認識システムには，人間の言語知識の役割を果たす言語モデルが組み込まれており，一般的には統計的言語モデルである N-gram が用いられている．しかし，数千語〜数万語を対象とする場合，N-gram モデルのパラメータが指数関数的に増大し，システム構築に際して，大きな障害が生じることになる．本論文では，これまでに提案された種々の N-gram モデルのパラメータ削減手法の比較を行なう．また，我々が提案する削減手法を (N-1)-gram に適用するための予備実験を行なったので，その結果について報告する．

キーワード　大語彙連続音声認識　N-gram　エントロピー　パラメータ削減

A Study on Entropy-based Compression Algorithms for N-gram parameters

Norimichi Yodo　Kiyohiro Shikano　Satoshi Nakamura

Graduate School of Information Science, Nara Institute of Science and Technology

8916-5 Takayama-cho, Ikoma-shi, Nara, 630-0101 JAPAN

E-Mail:{norimi-y,shikano,nakamura}@is.aist-nara.ac.jp

Abstract　Large vocabulary continuous speech recognition (LVCSR), which is simply called as dictation, is an essential technology for the realization of voice typing and interface between human being and a computer in various conditions. An LVCSR system reduces search space using language models, where statistical N-gram models are generally used. However, they need a huge number of parameters that grow exponentially with N and the vocabulary size. Especially in the task with large vocabulary (from a few thousand of words to several ten thousands of words), their huge memory requirement results in the system implementation difficulty. In this paper we compare algorithms for reducing the number of parameters of an N-gram model. Preliminary experiments on the augmentation of our compression algorithm to deal with (N-1)-gram are carried out.

Key words　large vocabulary continuous speech recognition　N-gram　entropy　parameter reduction

1　はじめに

連続音声認識技術は年々進歩を遂げ，最近では人間の自然な発話をコンピュータに認識（文字化）させる大語彙連続音声認識（ディクテーション）の研究が盛んに行なわれている．例えば，情報処理振興事業協会（IPA:Information Technology Promotion「日本語ディクテーション基本ソフトウエアの開発」プロジェクト [1, 2] は複数の大学や国立研究所が連携して研究開発し，技術を一般に公開することで，我国の音声認識研究のレベル向上に貢献している．また，ＰＣ上で動くソフトウェアがＩＢＭ [3] やＮＥＣ [4] などから市販されており，音声認識技術の認知度もずいぶんと上がってきている．この技術の進歩と認知度の向上は，音声ワープロやニュース原稿書き起こし，あるいは障害者のコミュニケーション手段や自然発話によるカーナビゲーションシステムなどの実現への追い風となるだろう．

さて，音声認識は一種の探索問題であり，大語彙

連続音声認識などの複雑なタスクでは言語モデルを用いて探索空間を縮小する．音声認識における言語モデルと言えば，かつては文法に基づくものが専らであったが，最近では，

1. 半自動的に構築できる (ツールも提供されている [5])
2. モデル作成に必要なテキストデータベースの整備が行なわれた [6]
3. 統計的な音響モデルとの相性が良い

などの理由から，統計的言語モデルである N-gram モデルが主流である [7, 8, 9].

この N-gram モデルを用いたシステムの実用性は，実行時間や必要なメモリ容量に大きく依存する．タスクが複雑になる程，N-gram の学習に用いるデータベースが大きくなる程，これらの問題は重要になる．この理由は，N-gram モデルは N が大きくなると，必要なパラメータが指数関数的に増えるからである．解決策として，これまでに，cutoff 手法や情報量尺度に基づいた N-gram モデルのパラメータ削減手法がいくつか提案されている．

我々は先ほど触れたＩＰＡのプロジェクトにおいて，言語モデルの圧縮に関する研究を行なっており，新しいパラメータ削減手法を提案し [10-12]，その効果が確認された [1, 2] ．その手法は，最尤推定に基づいており，一つのパラメータを削除する際の言語モデルの性能劣化を定量的に評価し，削除する優先度を決定する．そして，優先度の高いパラメータから削除し，back-off 係数を更新していくことで，任意のパラメータ数での言語モデルの作成を可能にした．

しかし，語彙を 2 万から 4 万単語に増やす今年度計画や，N-gram モデルの拡張を考慮すると，さらなる圧縮の可能性を検討する必要がある．本稿では，2 万語を用いた連続音声認識実験におけるパラメータ削減手法の比較結果，および提案手法の拡張として，N-gram と (N-1)-gram パラメータを同時に削減する効果について報告する．

2 従来の N-gram モデルのパラメータ削減手法

この章では，これまでに提案されてきた，N-gram モデルのパラメータ削減手法について考察する．なお，未知の単語系列は，back-off smoothing を用いて推定する back-off N-gram を対象とする．

2.1 cutoff 手法

パラメータ削減手法として最もよく使われる cutoff 手法は，学習データに出現する回数がある値よりも小さいものをカウントしない方法である．単純な手法であるが，パラメータ削減の効果は大きい．しかし，以下の 2 点が考慮されていない．

1. 元の確率値と，back-off smoothing によって推定される値が大きく異なる場合，すなわち「N-gram は (N-1)-gram と似ている」という経験則が破綻する場合
2. 出現回数が同じでも確率値，すなわちコンテキストに対する単語の出現確率が異なる場合 ($p_1 = 1/100$, $p_2 = 1/2$ が同時に削除される)

2.2 back-off 確率との比

元の N-gram 確率値と back-off により推定される確率値との比に着目する手法が提案されている [13]. そこでは，

$$K * (\log(\text{元の確率}) - \log(\text{back-off確率})) \quad (1)$$

を尺度としている．ここで K はディスカウントされた N-gram の出現回数である．この手法は，頻度を考慮していない点が問題である．すなわち，二つのパラメータの値の比が同じであれば区別されないことになる．

2.3 エントロピー尺度 （一括削減）

情報量尺度に基づく削減手法は多数提案されている．基本的には，N-gram の確率分布と (N-1)-gram の確率分布の相関や距離を種々の情報量を用いて評価する考え方である．例えば，"N-gram の確率分布と (N-1)-gram の確率分布の距離"として，相対エントロピーが一般的に用いられている [14, 15, 16]. これらの手法では，N-gram パラメータを (N-1)-gram パラメータで総替えするため，パラメータ毎に見れば，必ずしも性能劣化が少ないものから削除されるというわけではないという問題点がある．

3 エントロピーに基づく逐次削減手法

我々が提案したこの手法は，パラメータ削除による言語モデルの性能低下を，エントロピー尺度に基づいて定量的に，パラメータ毎に評価するものであり [10-12,17]，back-off 係数を更新していく点，任意のパラメータ数を実現できるという点に特徴がある．ここでは，簡単化のために trigram を例にとって考えることにする (図 1). プロセスは，次のようになる．

1. back-off N-gram モデルを作成する（ディスカウンティング手法は任意）.
2. あるコンテキスト，すなわち N-1 単語列を固定して考えた時，次の二つの確率分布を得る
   - $\{p\}$ 元のモデルの条件付き確率分布
   - $\{p'\}$ 一つの N-gram パラメータを削除し，back-off により推定する場合に得られる確率分布
3. 以上の二つの分布と，新しい back-off 係数 (α') を用いて，エントロピーの増加量を求める．その値は二つの分布間の相対エントロピーにコンテキストの頻度をかけることによって求められる．

4. エントロピーの増加量が小さいものから，あるいは与えられたしきい値以下のものを削除し，必要であれば back-off 係数を更新する (終了).

この過程における，back-off 係数の更新方法および，エントロピーの増加量の算出法について整理する．

3.1 back-off 係数の更新法

trigram モデルを考える場合，コンテキストである 2 単語列によってモデルの空間，すなわち全ての trigram パラメータを分類することができる．ここでは，3 単語列 $xyz_1, xyz_2, \ldots$ は，「部分空間 Q_{xy} に含まれる」という表現を用いて，「各部分空間は他の部分空間に対して独立である」と仮定する．今，部分空間 Q_{xy} だけを考えて，削除前の trigram パラメータ $xyz_i (i = 1, 2, \ldots)$ の確率分布を $\{p\}$，対応する bigram パラメータ yz_i の確率分布を $\{q\}$ とする．また，未知の trigram および bigram に与えられる確率値の総和をそれぞれ p_{unks}, q_{unks}[1] とする．このとき，back-off 係数 α は，c_+ を $C(xyz_i) > 0$ である i の集合とすると，

$$\alpha = \frac{p_{unks}}{q_{unks}} = \frac{1 - \sum_{i \in c_+} p_i}{1 - \sum_{i \in c_+} q_i} \tag{2}$$

となる (図 1(a)).

今，trigram パラメータ xyz_k を削除する場合，$p_k = P(z_k|xy)$ を bigram から back-off smoothing により推定することになる．このとき，新しい back-off 係数を以下の値に更新する (図 1(b)).

$$\alpha' = \frac{p'_{unks}}{q'_{unks}} = \frac{p_{unks} + p_k}{q_{unks} + q_k} \tag{3}$$

最終的に，初期状態の確率分布 $\{p\}$ と，(一つの) パラメータ削除後の分布 $\{p'\}$

$$\{p'\} = \begin{cases} p_i & i \neq k, i \in c_+ \\ \alpha' \cdot q_k & k \in c_+ \\ \alpha' \cdot q_i & i \notin c_+ \end{cases} \tag{4}$$

を得ることになる．

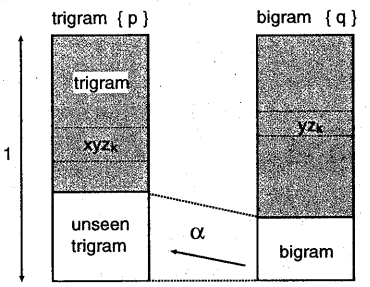
3.2 エントロピー変化量

一般に，「言語モデルが表現する言語の複雑さ」を表わす尺度として，エントロピーが用いられる．3 単語列 xyz の生起確率を $p(xyz)$ で表すと，エントロピーは，

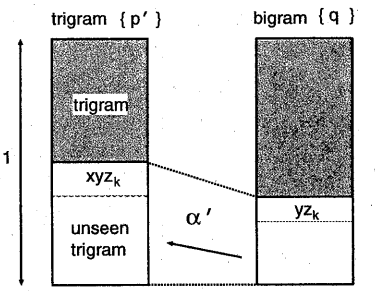
$$\mathcal{H} = -\sum_{xyz} p(xyz) \log p(xyz) \tag{5}$$

で表され，最尤推定式と等価である．部分空間 Q_{xy} の生起確率を $P(Q_{xy})$ とすると，式 (5) は，trigram

[1]"unks" は unknown words を意味する



(a) 初期確率分布　(α: back-off 係数)



(b) パラメータ削除後の確率分布 (α': back-off 係数)

図 1 back-off 係数の更新

確率（条件付き確率）$p_i = P(z_i|xy)$ を用いて，

$$= -\sum_{xy} P(Q_{xy}) \log P(Q_{xy}) + \sum_{xy} P(Q_{xy})(-\sum_i p_i \log p_i)$$

と書き直すことができる．

さて，全空間でのエントロピーの変化量は，部分空間 Q_{xy} における変化量と等しいと仮定しているので，$C(\cdot)$ を事象の生起回数として，

$$\begin{aligned} \Delta\mathcal{H} &= \mathcal{H}' - \mathcal{H} \\ &= P(Q_{xy}) \sum_i p_i \log \frac{p_i}{p'_i} \\ &= \frac{C(xy)}{C(all)} \times \mathcal{D}(p||p') \end{aligned} \tag{6}$$

となる．ここで，$C(all)$ は，bigram の総出現回数であり，$\mathcal{D}$ は，相対エントロピーである．

同様に，パープレキシティーの変化量は，空間 Q_{xy} だけで考えると，

$$\begin{aligned} \log(\Delta\mathcal{PP}) &\propto \sum_i C(xyz_i) \log \frac{p_i}{p'_i} \\ &= C(xy) \times \mathcal{D}(p||p') \end{aligned} \tag{7}$$

という式で求めることができる．

4 パラメータ削減手法の評価 (パープレキシティー)

前章で取り上げた種々の方法の比較を行なう．ここでは，CD-ROM版毎日新聞記事から言語モデル (back-off trigram) を作成し，テストセットパープレキシティーを評価尺度に用いた．

学習には 45 ケ月分を，テストセットには 3 ケ月分のデータを用いた．なお，言語モデルの学習には不要と思われるような記事や文はあらかじめ除いてある [18]．学習データの形態素解析結果を表 1 に示す．本実験では，形態素，その原形，品詞番号を考慮して単語を識別している．すなわち，表 2 の下線部の形態素はすべて区別される．以下では，形態素を単に単語と呼ぶことにする．

4.1 trigram パラメータの削減

このデータの頻度上位 20000 単語を用いて cutoff 手法，相対エントロピーによる一括削減，逐次削減モデルを作成し，パラメータの数 (bigram + trigram) とテストセットパープレキシティーとの関連を調べた．図 2 にその結果を示す．図中の () 内の数字は trigram パラメータの削減率を示している．グラフ中の最も右のポイントに対応するモデルは，bigram に対して 1，trigram に対して 1 の cutoff を行なっている．なお，言語モデルの作成には，CMU-Cam-Toolkit[5] を用い，back-off smoothing の際のディスカウンティングは Witten-Bell 法で行なっている [19]．同じパラメータ数では，逐次的削減手法によるモデルが最も良い性能を示し，次に cutoff によるモデル，一括削減によるモデルとなっている．とりわけ逐次削減手法では，trigram パラメータを 1/2 に削減しても精度を維持している．

4.2 逐次削減手法の bigram パラメータへの適用

前節の結果から，削減手法としては，相対エントロピーに基づく逐次削減が最も効果的であることが分かった．さらなる言語モデル圧縮を実現するために，(N-1)-gram（ここでは bigram ）にもこの削減手法を適用し，同様にテストセットパープレキシティーを求めた．図 3 に結果を示す．図の太線は，trigram パラメータだけを削減したときのパープレキシティーの変化を示している．そして，太線から分岐する各線（＋の列）は trigram パラメータ数を固定して，bigram パラメータを減らしていったときの変化を示している．

trigram パラメータの削減率が小さいとき（多くが残っている）は，bigram パラメータの削減が言語モデルの性能を大きく低下させることが分かる．しかし，trigram パラメータの削減率が大きくなってくると，ある領域では，bigram パラメータを削減した方が高い性能が得られることが分かる．

表 1 学習データ解析結果（45 ケ月）

総文数	約 237 万文
総単語数	約 6,623 万語
異なり単語数	約 19 万語
一文当りの単語数	28.16 語
5K, 単語被覆率	88.23%
20K, 単語被覆率	96.47%
90%被覆単語数	6,315
97%被覆単語数	22,959

表 2 単語の定義

用例	形態素, 原形, 品詞
私は 知ら ない.	(知ら, 知る, 連用形ナイ接続)
そうとは 知ら ず	(知ら, 知る, 連用形ズ接続)
100円	(円, 円, 助数詞)
円 が回復	(円, 円, 名詞)

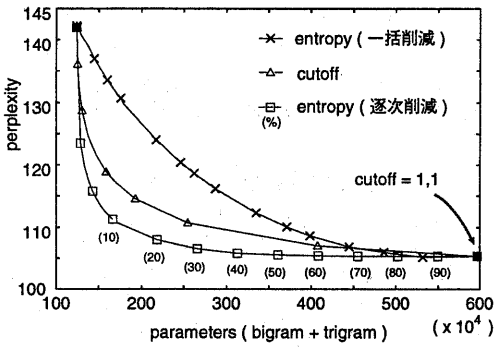


図 2 パラメータ削減手法の比較

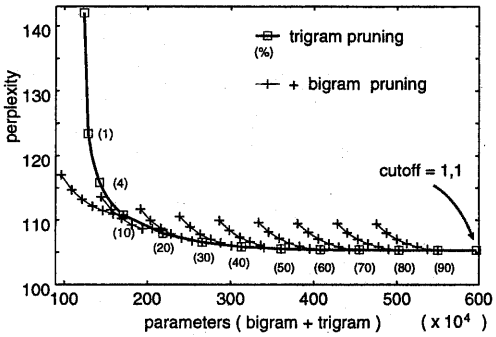


図 3 逐次削減手法による言語モデル性能

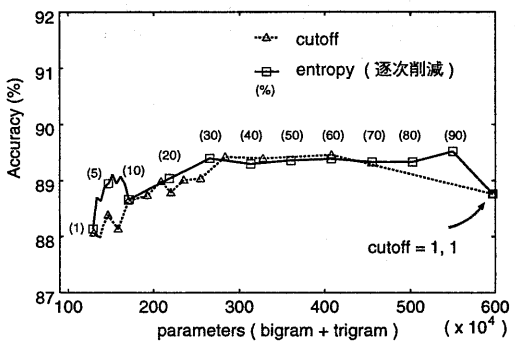


図 4　cutoff と提案手法の比較 (単語正解率)

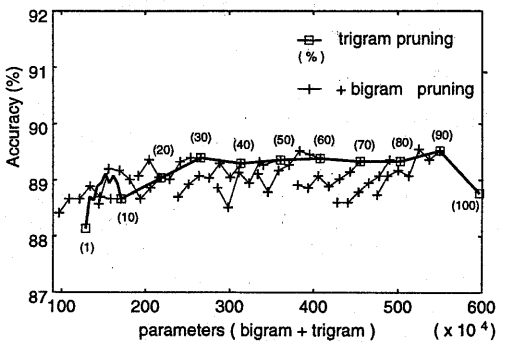


図 5　逐次削減手法によるモデルと単語正解率

5　大語彙連続音声認識による評価

2 万語を対象とする大語彙連続音声認識実験を行い，パラメータの削減が認識精度にあたえる影響を調べる．実験には，情報処理振興事業協会 (IPA) の「日本語ディクテーション基本ソフトウエア,1998 年度版 (CD-ROM)」に納められている音響モデル，言語モデル，認識エンジン JULIUS，および「JNAS データベース (毎日新聞記事読み上げ文)」を用いた．

音響モデルは表 3 に示された triphone を用い，音声データは JNAS データベースから男女各 10 名の発話による 200 文章を選んだ．このデータは音響モデル，言語モデル双方に対してオープンなものである．

京都大学で開発されている認識エンジン JULIUS[20] は，探索を 2 パスで行なっている．第 1 パスでは，left-to-right の bigram モデルを，第 2 パスでは，right-to-left (逆向き) の trigram モデルを用いる仕様となっている．JULIUS の実行に必要なメモリの半分以上が trigram モデルによって占められている．ここでは，第 1 パスの bigram を固定し，第 2 パスで用いる trigram を種々変えて，認識精度の変化を調べた．

5.1　cutoff と逐次削減手法の比較

cutoff と逐次削減手法を用いて，trigram パラメータの削減を行なったところ，図 4 のような結果が得られた．まず，逐次削減法は，trigram パラメータ数を約 1/3 に減らしても精度を維持している．そして，ほとんどすべての点において，逐次削減手法によるモデルの方が，高い単語正解率を示していることがわかる．特に，大幅な削減を行なった場合 (グラフ左側部分) にはその差が大きくなり，パープレキシティーで見られた傾向が反映されている．

5.2　bigram への適用 (逐次削減法)

次に，trigram と bigram の両パラメータを逐次削減手法により同時に削除したモデルを用いて認識実験を行なった (図 5)．太線は，trigram パラメータだけを削減した場合の認識率の変化を示している (前節と同じ)．太線から分岐している各線は，trigram パラメータを固定し bigram パラメータを削減したモデルを用いた場合の結果を示している．ここでも，パープレキシティーに見られたのと同様に，trigram パラメータの削減率が小さいときは bigram パラメータの削減は精度の低下を引き起こし，trigram パラメータの削減率が大きくなってくると，ある領域では bigran パラメータを削減した方が高い認識率を示すことがわかる．

表 3　音響モデル

サンプリング周波数	16 [kHz]
プリエンファシス	0.97
分析窓	Hamming Window
分析窓長, 窓間隔	25 [ms],10 [ms]
特徴パラメータ	MFCC(12) + ΔMFCC(12) + Δ Pow (計 25 次)
混合数	16
状態数	2000

6　おわりに

本稿では，種々の N-gram モデルのパラメータ削減手法を trigram モデルに適用し，パープレキシティーを比較した．結果，我々が提案する手法の効果を確認した．また，削減手法の拡張として，trigram パラメータと bigram パラメータを同時に削減する予備実験を行ない，ある領域では，trigram パラメータを減らすよりも bigram を減らした方がよいという事実を確認した．さらに，提案手法を用いて 20000 単語を対象とする大語彙連続音声認識実験を行ない，パープレキシティーと同様の効果を確認した．

以上の結果から，パープレキシティーを尺度として，逐次削減法により N-gram と (N-1)-gram パラメータを同時に削減することで，N-gram のみを削除するよりも高い精度を得られることを確認した．

謝辞

本研究は情報処理振興事業協会 (IPA) の「日本語ディクテーション基本ソフトウエア」プロジェクトの支援を受けて行なわれた．関係者の皆様，とりわけデータや有益な

助言を提供していただいた電総研・伊藤克亘氏にこの場を借りてお礼申し上げます.

参考文献

[1] 情報処理振興事業協会: 日本語ディクテーション基本ソフトウェアの開発, 調査研究報告書,1998 年度 (1999).

[2] 河原他: 日本語ディクテーション基本ソフトウェア (98 年度版) の性能評価情報処理学会研究報告, SLP26-6, pp.39-46 (1999).

[3] http://www.ibm.co.jp/voiceland/

[4] http://www.nec.co.jp/japanese/product/kiban/control/product/ninsiki/mt.html

[5] P.Clarkson et al. : Statistical Language Modeling Using The CMU-Cambridge Toolkit, ESCA Eurospeech 1997, vol.5, pp.2707-2710 (1997).

[6] 武田他: 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, SLP11-9, pp.49-32 (1999).

[7] 伊藤他: 大語彙言語データベースからの N-gram 構築とタスク適応の検討, 情報処理学会研究報告, SLP11-5, pp.25-30 (1996).

[8] 大附他: 新聞記事を用いた大語彙連続音声認識の検討, 信学技報, SP95-90, pp.63-68 (1995).

[9] 吉田他: 単語 trigram を用いた大語彙連続音声認識, 情報処理学会研究報告, SLP14-14, pp.99-104 (1996).

[10] 踊堂: 形態素単位の N-gram モデルの構築と圧縮に関する研究, 奈良先端大・修士論文,1998 年 2 月.

[11] 踊堂他: 情報量に基づく trigram パラメータの逐次的削減手法, 情報処理学会研究報告, SLP22-17, pp.91-96 (1998).

[12] N.Yodo et al. : Compression Algorithm of Trigram Language Models Based on Maximum Likelihood Estimation, Proc.ICSLP-98,pp.716-719 (1998).

[13] K.Seymore et al. : Scalable Backoff Language Models, Proc.ICSLP-96, pp.232-235 (1996).

[14] A.Bonafonte et al. : Language Modeling Using X-grams, Proc.ICSLP-96, vol.1, pp.394-397 (1996).

[15] R.Kneser: Statistical Language Modeling Using a Variable Context Length, Proc.ICSLP-96, vol.1, pp.494-497 (1996).

[16] D.Ron, et al. : Learning Probabilistic Automata with Variable Memory Length, 7th Anuual ACM Conf. on Computational Learning Theory, pp.35-46 (1994).

[17] A.Stolcke: Entropy-based pruning of back-off language models, Proc.Broadcast News Transcription and Understanding Workshop, pp.27 0-274 (1998).

[18] 伊藤他: 大語彙連続音声認識のためのテキストデータ処理, 音学講論,pp.105-106,1996 年 9 月.

[19] P.Placeway et al. : The Estimation of Powerful Language Models from Small and Large Cor pora, Proc.ICASSP-93, vol.II, pp.33-36 (1993).

[20] 李他: 単語トレリスインデックスを用いた大語彙連続音声認識エンジン JULIUS, 信学技報,SP98-3, pp.17-24 (1998).