音声言語情報処理 22−17
（1998. 7. 25）

# 情報量に基づく trigram パラメータの逐次的削減手法

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において、統計的言語モデルである N-gram は非常に効果的であるが、膨大なパラメータが必要であるという問題がある。本稿では、従来法である cutoff よりも言語モデルに与える影響が少ない N-gram モデルのパラメータ削減手法を提案する。提案手法は最尤推定に基づいており、学習セットパープレキシティーを最小にするパラメータで trigram モデルを実現する。提案手法の効果をパープレキシティーと単語誤り率の観点から検証し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　パープレキシティー　パラメータ削減

# Reduction Algorithm of Trigram Parameters Based on Information Amount

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**　In large vocabulary continuous speech recognition, N-gram models, are effective, but they need a huge number of parameters.

In this paper we propose an algorithm for reducing the size of a trigram model, with less affecting its performance than the traditional *cutoff* method. This algorithm is based on Maximum Likelihood(ML) estimation. Here ML estimation can realize the trigram language model with a small set of trigram probability parameters that minimize the training set perplexity. To confirm the effectiveness of our algorithm, experiments in terms of perplexity and word error rate are carried out.

**Key words**　large vocabulary continuous speech recognition,　N-gram,　perplexity, 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るようになった。この音声ディクテーション技術は、不特定話者大語彙連続音声認識技術を集大成したものと考えることができ、音声分析、音響モデル、形態素解析、統計的言語モデル、連続音声認識アルゴリズムなど、音声認識の基本技術がすべて含まれている [1]。

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効率的に縮小する必要がある。かつては、文法に基づく言語モデルが主流であったが、

(1) 半自動的に構築できる
(2) モデル作成に必要なテキストデータベースの整備が進んでいる
(3) 統計的な音響モデルとの相性が良い

などの理由から、現在は統計的言語モデルである N-gram モデルが主に研究され、用いられている [2, 3, 4, 5]。

N-gram モデルを用いたディクテーション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。タスクが複雑になり、N-gram の学習に用いるデータベースが大きくなるにつれて、これらの問題は重要になる。これまでにいくつかの N-gram モデルのパラメータ削減手法が提案されているが、それぞれ問題点を抱えていたり、あるいは厳密な定式化は行なわれていない。

本研究では、新しいパラメータ削減手法を提案し、言語モデルの性能を表わす尺度であるエントロピー、パープレキシティーとの関係について定式化を行なう。また、大語彙連続音声認識の実験を行ない提案手法の効果を検証する。

# 2 N-gram モデル

N-gram モデルは単語列の出現確率を統計的に推定するモデルである。単語列 $P(\boldsymbol{w})$ の生起確率は、

$$\begin{aligned} P(\boldsymbol{w}) &= P(w_1 w_2 \cdots w_n) \\ &= P(w_1) \prod_{i=2}^{n} P(w_i | w_1 \cdots w_{i-1}) \quad (1) \end{aligned}$$

で与えられる。実際に式 (1) の確率を求めるのは不可能である。そこで、$i$ 番目の単語 $w_i$ の生起確率が、直前の (N-1) 単語 $w_{i-N+1} \cdots w_{i-2} w_{i-1}$ だけに依存すると考え、

$$P(w_1 w_2 \cdots w_n) = \prod_{i=1}^{n} P(w_i | w_{i-N+1} \cdots w_{i-1}) \quad (2)$$

という近似を行なうモデルが、N-gram モデルである。

## 2.1 確率の推定

単語列 $w_{i-N+1} \cdots w_{i-1}$ を $w_{i-N+1}^{i-1}$ 、その出現回数を $C(w_{i-N+1}^{i-1})$ と表わすと、条件付き確率 $P(w_i | w_{i-N+1}^{i-1})$ は、最尤推定によると、

$$P(w_i | w_{i-N+1}^{i-1}) = \frac{C(w_{i-N+1}^{i})}{C(w_{i-N+1}^{i-1})} \quad (3)$$

で推定できる。特に、bigram、trigram の場合は、各々次の式で確率が推定できる。

$$P(w_i | w_{i-1}) = \frac{C(w_{i-1}^{i})}{C(w_{i-1})} \text{ (bigram) } (4)$$

$$P(w_i | w_{i-2} w_{i-1}) = \frac{C(w_{i-2}^{i})}{C(w_{i-2}^{i-1})} \text{ (trigram) } (5)$$

## 2.2 back-off smoothing

単純な最尤推定では、学習データに現れなかった単語列には確率値として 0 を与えてしまう。また、出現頻度の小さな単語列に対しても、統計的に信頼性のある確率を推定することは難しい。

以上の問題は、スパースネスの問題 (sparseness problem) と言われている。この問題の対処法として、線形補間法や back-off smoothing 法が提案されている。ここでは、back-off smoothing を考えることにする。

back-off smoothing は、学習データに出現しなかった N-gram 確率を (N-1)-gram から推定

する手法であり、trigram の場合は bigram から推定することになる。

最尤推定による確率を

$$P(w_i|w_{i-2}^{i-1}) = \frac{C(w_{i-2}^{i})}{C(w_{i-2}^{i-1})} \quad (6)$$

とするとき、back-off smoothing により、N-gram 確率 $P^*(w_i|w_{i-N+1\cdots w_{i-1}})$ は次の式で推定される。

$$P^*(w_i|w_{i-2}^{i-1}) = \begin{cases} \lambda(w_{i-2}^{i}) \cdot P(w_i|w_{i-2}^{i-1}) \\ \quad \text{if } C(w_{i-2}^{i}) > 0 \\ \alpha(w_{i-2}^{i-1}) \cdot P^*(w_i|w_{i-1}) \\ \quad \text{else if } C(w_{i-2}^{i-1}) > 0 \\ P(w_i|w_{i-1}) \\ \quad \text{otherwise} \end{cases} \quad (7)$$

式中の $\lambda$ はディスカウント係数 (discount coefficient) といい、学習データに出現した N-gram の確率値を割り引く働きをしている。また、$\alpha(w_{i-2}^{i-1})$ は、確率の総和を 1 にする back-off 係数で、

$$\alpha(w_{i-2}^{i-1}) = \frac{1 - \sum_{C(w_{i-2}^{i})>0} P^*(w_i|w_{i-2}^{i-1})}{1 - \sum_{C(w_{i-2}^{i})>0} P^*(w_i|w_{i-1})} \quad (8)$$

である。

### 2.3 perplexity

言語モデルの性能を表す尺度として、perplexity が用いられる。テキストが単語列 $w_1^n$ からなるとき、単語列のエントロピー $\mathcal{H}$ は、

$$\mathcal{H} = -\frac{1}{n} \log P(w_1 \cdots w_n) \quad (9)$$

であり、テストセットパープレキシティーは

$$\mathcal{PP} = 2^{\mathcal{H}} = P(w_1 \cdots w_n)^{-\frac{1}{n}} \quad (10)$$

で算出される。

## 3 従来の N-gram モデルのパラメータ削減手法

N-gram モデルのパラメータ削減手法としてこれまでに、いくつかの方法が提案されている。本章では、従来法の問題点について考察する。

### 3.1 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、(N-1)-gram や他の N-gram との相関性についてはあまり考慮されておらず、局所的に経験則が破綻するケースや、他の N-gram に対する影響が大きい場合には問題がある。

### 3.2 情報量に基づく手法

情報量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 [10, 11]。

これらの手法の問題は、N-gram パラメータを一つずつ削除できない、頻度を考慮していないという 2 点にある。

### 3.3 back-off 確率との比

trigram 確率と back-off により推定される確率との比に着目する手法が提案されている [12]。そこでは、

$$K * (\log(\text{元の確率}) - \log(\text{back-off確率})) \quad (11)$$

を尺度としている。ここで $K$ は Good-Turing の推定によってディスカウントされた N-gram カウントである。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、2 つの確率の比が同じであれば区別されない。

## 4 提案法

前節で述べた従来方法と比較して、提案手法の利点は、

○ パラメータ削減の影響を定量的に評価して、最も影響の小さいものを削除する (最尤推定) ことで、逐次削減が可能、すなわち任意の大きさのモデルを作成できる

という点である。プロセスは、以下の通りである。

1. N-gram モデルを作成する (back-off,cutoff 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
   a. 元のモデルの条件付き確率 (N-gram) 分布 $p$
   b. N-gram パラメータを back-off により推定する場合に得られる確率分布 $p'$ (更新された back-off 係数を用いて算出)
3. 以上の二つの分布と、新しい back-off 係数を用いて、エントロピー (あるいは perplexity) 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削除し、必要であれば back-off 係数を更新する。

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 (あるいは perplexity) の増加量の算出法について整理する。なお、ここでは trigram を例にとって考えることにする。

## 4.1 back-off 係数の更新法

今、2 単語列 $xy$ によってモデルの空間が分割されており、各空間は他に対して独立であると考える。この場合の (部分) 空間を $Q_{xy}$ で表わすことにする。

次単語 $z_i$ を予測する場合に、trigram 確率 $p_k = P(z_k|xy)$ と back-off により bigram 確率 $q_k = P(z_k|y)$ から推定される $p'_k = \alpha'(xy) \cdot q_k$ を考える。

この場合、trigram が学習データになかった (未知の単語列) と仮定するので、未知の trigram に与えられる確率 $p_{unks}$ が、$p_{unks} + p_k$ に、対応する bigram 確率の和は $q_{unks} + q_k$ になり、この値を用いて back-off 係数を更新する。

初期状態での trigram 、それに対応する bigram の確率および back-off 係数は、$\mathbf{c}_+$ を $C(w_x w_y w_{z_i}) > 0$ である $i$ の集合とすると、

$$p_{unks} = 1 - \sum_{i \in \mathbf{c}_+} p_i \tag{12}$$

$$q_{unks} = 1 - \sum_{i \in \mathbf{c}_+} q_i \tag{13}$$

$$\alpha = \frac{p_{unks}}{q_{unks}} \tag{14}$$

となる ($\alpha$ は (8) 式と同じ表現)。$p'_k$ を推定する場合のそれぞれの値は、

$$p'_{unks} = p_{unks} + p_k \tag{15}$$

$$q'_{unks} = q_{unks} + q_k \tag{16}$$

$$\alpha' = \frac{p'_{unks}}{q'_{unks}} \tag{17}$$

となる。図 1は、この手法の考え方を示したものである。

結局、初期状態の確率分布 $\{p\}$ と、分布 $\{p'\}$

$$\{p'\} = \begin{cases} p_i & i \neq k, i \in \mathbf{c}_+ \\ \alpha' \cdot q_k & i = k \\ \alpha' \cdot q_i & i \notin \mathbf{c}_+ \end{cases} \tag{18}$$

を得ることができる。

## 4.2 エントロピー変化量

### 4.2.1 相対エントロピー

空間 $Q_{xy}$ における、$\{p\}$ と $\{p'\}$ の確率分布の類似度を相対エントロピー (relative entropy、あるいは Kullback-Liebler 距離) で与える。相対エントロピー $\mathcal{D}(p||q)$ は、二つの確率分布 $\{p\}$ と $\{q\}$ に対し、

$$\mathcal{D}(p||q) = \sum_i p_i \log \frac{p_i}{q_i} \tag{19}$$

で定義される。すべての $i$ において $p_i = q_i$ ならば $D(p||q) = 0$ である。ここでは、

$$\begin{aligned} \mathcal{D}(p||p') &= \sum_i p_i \log \frac{p_i}{p'_i} \\ &= p_k \log \frac{p_k}{\alpha' q_k} + p_{unks} \log \frac{\alpha}{\alpha'} \end{aligned} \tag{20}$$

で求めることができる。

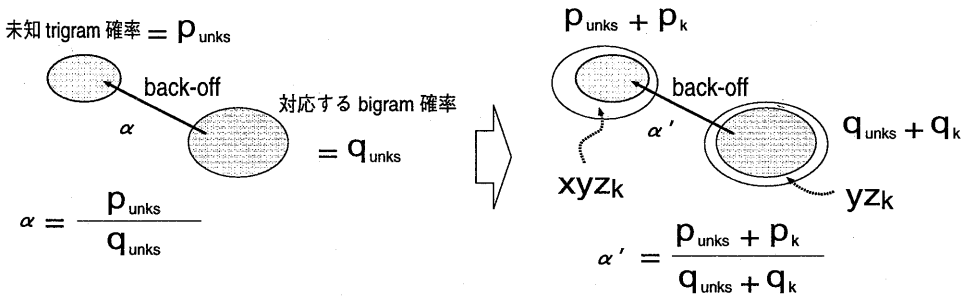


図 1: 提案手法の考え方

#### 4.2.2 エントロピー

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。単語列 $w_i^k = w_i w_j w_k$ の生起確率を $p(w_i^k)$ で表すと、エントロピーは、

$$\mathcal{H} = -\sum_{w_i^k} p(w_i^k) \log p(w_i^k) \qquad (21)$$

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 (21) は、trigram 確率 (条件付き確率) $p_i = P(z_i|xy)$ を用いて、

$$\begin{aligned} &= -\sum_{x,y} P(Q_{xy}) \log P(Q_{xy}) \\ &\quad + \sum_{x,y} P(Q_{xy})(-\sum_i p_i \log p_i) \end{aligned}$$

と書き直すことができる。

#### 4.2.3 エントロピー変化量

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、

$$\begin{aligned} \Delta\mathcal{H} &= \mathcal{H}' - \mathcal{H} \\ &= P(Q_{xy})(-\sum_i p_i \log p_i') \\ &\quad - P(Q_{xy})(-\sum_i p_i \log p_i) \\ &= P(Q_{xy}) \sum_i p_i \log \frac{p_i}{p_i'} \\ &= P(Q_{xy}) \cdot \mathcal{D}(p||p') \\ &= \frac{C(xy)}{const} \times \mathcal{D}(p||p') \qquad (22) \end{aligned}$$

となる。すなわち、相対エントロピーに事象の回数の重みをかけたものがエントロピーの変化量である。

全事象が $\mathcal{T}$ である trigram モデルにおいては、

$$\log \mathcal{PP} = - \sum_{xyz \in \mathcal{T}} C(xyz) \log P(z|xy) \qquad (23)$$

であるから、変化量は、空間 $Q_{xy}$ だけで考えると、

$$\begin{aligned} \log(\Delta\mathcal{PP}) &\propto \sum_i C(xyz_i) \log \frac{p_i}{p_i'} \\ &= C(xy) \\ &\quad \times \sum_i \frac{C(xyz_i)}{C(xy)} \log \frac{p_i}{p_i'} \\ &= C(xy) \sum_i p_i \log \frac{p_i}{p_i'} \\ &= C(xy) \times \mathcal{D}(p||p') \qquad (24) \end{aligned}$$

という式で求めることができる。

## 5 実験

提案手法の効果を検証するために、テストセットパープレキシティーを求め、連続音声認識実験を行なった。

表 1: 学習データ解析結果 (45 ケ月)

| | |
|---|---|
| 総文数 | 2371K |
| 総単語数 (語) | 65.3M |
| 異なり単語数 (語) | 290K |
| 品詞種類 | 402 |
| 一文当りの単語数 | 27.55 |
| 平均単語長 (字) | 1.659 |
| 5K, 単語被覆率 | 85.8% |
| 10K, 単語被覆率 | 91.5% |
| 15K, 単語被覆率 | 94.1% |
| 20K, 単語被覆率 | 95.7% |
| 90%被覆単語数 | 8.1K |
| 95%被覆単語数 | 17.5K |
| 97%被覆単語数 | 27.6K |

K ($\times 10^3$)
M ($\times 10^6$)

## 5.1 テストセットパープレキシティー

RWC テキストデータベース (毎日新聞記事 4 年分の形態素解析結果) [6] のうち、45 カ月分を言語モデルの学習用、残りを評価用とした。学習データの形態素解析結果を表 1に示す。以下では、形態素を単に単語と呼ぶことにする。

このデータの頻度上位 5000 単語、および 20000 単語を用いて cutoff によるモデル、ならびに提案手法によるモデルを作成し、trigram パラメータの数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 および表 5.1にその結果を示す。なお、言語モデルの作成には、CMU-Cam-Toolkit[8] を用い、back-off 手法の際の discounting は Witten-Bell 法で行なっている。

テストセットパープレキシティーでは、提案手法の効果が顕著に現われている。最も差のみられる点では、 cutoff の約 1/3 の trigram パラメータ数で同じパープレキシティーを示している。

表 2: 言語モデルの比較

| model | | perplexity | trigram | file size |
|---|---|---|---|---|
| 5K | cutoff | 62.9 | 226.3K | 16.6MB |
| | proposed | 62.2 | 69.5K | 11.8MB |
| 20K | cutoff | 87.1 | 228.3K | 28.7MB |
| | proposed | 86.7 | 74.5K | 23.8MB |

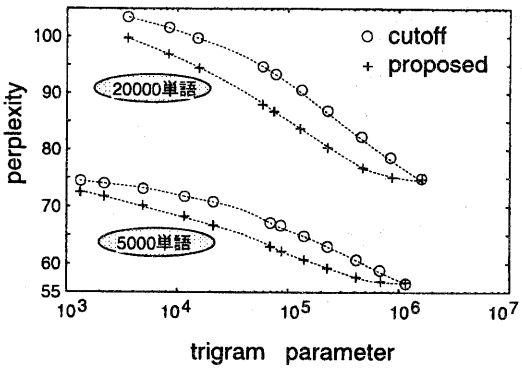


図 2: trigram パラメータ数とテストセットパープレキシティー

## 5.2 認識実験

次に、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,1997 年度版」、および「JNAS データベース (毎日新聞記事読み上げ文)」を用いて 5000 語を対象とする連続音声認識実験を行なった。音響モデルは表 3に示されたものを用いた。

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。

認識エンジン JULIUS[7] 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。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。

cutoff 法、および提案手法により作成された言語モデルの trigram パラメータ数と認識精度 (単語誤り率) の関係は図 3 のようになった。

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] |
| 窓間隔 | 10 [ms] |
| 特徴パラメータ | MFCC(12) + ΔMFCC(12) + Δ Pow (計 25 次) |
| 混合数 | 8 |
| 状態数 | 3000 |
| 学習データ | JNAS + ASJ |

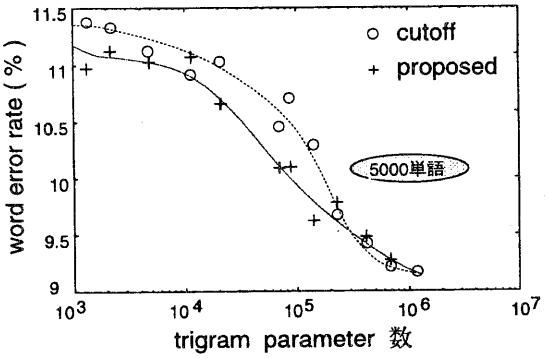


図 3: trigram パラメータ数と単語誤り率

認識実験では、パープレキシティーほどの差は観察されなかった。これは音声データ文の少なさに起因すると思われる。しかし、大局的に見れば提案手法の方がより少ないパラメータで同じ精度を示している。

## 6 おわりに

本稿では、情報量に基づく N-gram モデルのパラメータ削減手法の提案を行ない、 trigram に適用した。パープレキシティーでは cutoff よりも良好な結果を得ることができた。また、認識実験の結果においても効果が確認された。

今後の課題として、提案手法を bigram にも適用することや、可変長 N-gram の作成への適用を検討している。

## 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。関係者の皆様に感謝致します。

## 参考文献

[1] 鹿野, 中川, 「日本語音声ディクテーション基本技術講習会 – ディクテーション技術の概要と動向 –」 (1997 年 8 月).

[2] 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, SLP 11-5, pp.25-30 (1996).

[3] 政瀧他, 「連続音声認識のための可変長連鎖統計言語モデル」, 信学技報, SP95-73, pp.1-6 (1995 年 11 月).

[4] 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 年 12 月).

[5] 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, SLP 14-14, pp.99-104 (1996).

[6] 技術研究組合:新情報処理開発機構, 『RWC テキストデータベース報告書』 (1997 年 3 月).

[7] 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 年 4 月).

[8] P.Clarkson, R.Rosenfeld, "Statistical Language Modeling Using The CMU-Cambridge Toolkit", ESCA Eurospeech 1997, vol.5, pp.2707-2710 (1997).

[9] A.Bonafonte, J.B.Marino, "Language Modeling Using X-grams", Proc.ICSLP-96, vol.1, pp.394-397 (1996).

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

[11] D.Ron,Y.Singer, N.Tishby,"Learning Probabilistic Automata with Variable Memory Length", 7th Anuual ACM Conf. on Computational Learning Theory, pp.35-46 (1994).

[12] K.Seymore, R.Rosenfeld, "Scalable Backoff Language Models", Proc.ICSLP-96, pp.232-235 (1996).