論文 / 著書情報
Article / Book Information

論題(和文)	音声認識におけるモデル間スケーリング係数の自動推定
Title(English)	Efficient Estimation Method of Scaling Factors among Probabilistic Models in Speech Recognition
著者(和文)	大西祥史, 江森正, 越仲孝文, 篠田浩一
Authors(English)	Yoshifumi Onishi, Tadashi Emori, Takafumi Koshinaka, Koichi Shinoda
出典(和文)	電子情報通信学会論文誌, Vol. J95-D, No. 5, pp. 1276-1285
Citation(English)	, Vol. J95-D, No. 5, pp. 1276-1285
発行日 / Pub. date	2012, 5
URL	http://search.ieice.org/
権利情報 / Copyright	本著作物の著作権は電子情報通信学会に帰属します。 Copyright (c) 2012 Institute of Electronics, Information and Communication Engineers.

論　文

音声認識におけるモデル間スケーリング係数の自動推定

大西　祥史[†]　　江森　正[††*]　　越仲　孝文[†]　　篠田　浩一[†††]

Efficient Estimation Method of Scaling Factors among Probabilistic Models in Speech Recognition

Yoshifumi ONISHI[†], Tadashi EMORI[††*], Takafumi KOSHINAKA[†], and Koichi SHINODA[†††]

あらまし　音声認識における確率モデル間のスケーリング係数を効率的に推定する枠組みを提案する．音声認識システムは音響モデル，言語モデルなどの複数のモデルで構成される．モデルごとの出力値の乗算を行う際に，出力確率値の各々を異なる指数（スケーリング係数）でべき乗した上で行うと性能が向上することが経験的に知られている．従来，このスケーリング係数は，その値を変化させて対象の音声データを認識する処理を繰り返し，認識率が高くなる点を選択するという，アドホックな方法で最適化されてきた．本論文では，このスケーリング係数を，対数線形モデルの重みパラメータとみなし，最小単語誤り基準を用いて推定する方法を提案する．提案手法では計算量を低減するために単語ラティスを導入するが，それにより生じる推定値の初期値への依存性を軽減するために，単語ラティス生成とこう配法を用いた係数推定とを交互に繰り返し行う．日本語話し言葉コーパスを用いて評価を行い，提案手法が，最も単語正解精度が高くなるスケーリング係数を初期値に依存せず推定することを確認した．

キーワード　音声認識，スケーリング係数，対数線形モデル，単語ラティス

1. まえがき

音声認識は音響信号の観測系列が与えられたときの条件付き確率（事後確率）が最大となる単語系列を選ぶ問題として定義される．通常，この事後確率の最大化はベイズの定理を用いて逆問題として定式化し，単語系列の出現確率（言語モデル）と，単語系列が与えられたときの観測系列の生成確率（音響モデル）との積を最大にする単語列を求めることとなる（生成モデルによるアプローチ）．

これら音響モデルや言語モデルの真のモデルは知られておらず，現状では前者に混合正規分布を出力分布とする隠れマルコフモデル (HMM) を，後者に単語 N グラムモデルや単語挿入ペナルティー等の確率モデルを用いることが多い．このとき，生成モデルの枠組みで，そのままこれらの出力値の積をとるよりも，各々の出力値を異なる指数でべき乗して用いる方が認識性能が良くなることが経験的に知られている．この指数をここではスケーリング係数と呼ぶ．対数領域では，べき乗の処理は対数確率の重み付和をとることに相当し，そこでの重み係数がスケーリング係数となる．この性能向上の理由は，直接的には，モデルごとに大きく異なる出力値の値域がこの処理により補正されるためと考えられる．スケーリング係数は認識精度向上には欠かせない重要なパラメータとなっている [1]．

しかし，このスケーリング係数は生成モデルの枠組みから外れているため，最ゆう推定により求めることはできない．従来は，あらかじめ異なる候補値を複数用意して，対象となる音声データに対する認識を行い，認識率が最高になる値を選択するというアドホックな方法が用いられてきた．しかしながら，この方法では，組み合わせる確率モデルの数が増えた場合に演算量が指数的に増大するという問題がある [2], [16]．

この問題を解決するために，識別的方法によるス

† 日本電気株式会社，川崎市
NEC Corporation, 1753 Shimonumabe, Nakahara-ku, Kawasaki-shi, 211–8666 Japan
†† （株）NEC 情報システムズ，川崎市
NEC Informatec Systems, Ltd., 2–6–1 Kitamikata, Takatsu-ku, Kawasaki-shi, 213–8511 Japan
††† 東京工業大学，東京都
Tokyo Institute of Technology, 2–12–1 Ookayama, Meguro-ku, Tokyo, 152–8552 Japan
* 現在，ヤフー株式会社

ケーリング係数の推定方法がいくつか提案されている．例えば，Beyerlein [3] は，N ベストリストにおける単語列を対立候補として用い，平滑化誤り数基準 (smoothed error count measure) を最小化することでスケーリング係数の推定を行っている．しかし，N ベストリストは単語列の重複が多く，対立候補の表現として効率的ではない．また，N ベストリストは，その生成に用いた音声認識システムにおけるスケーリング係数の値に依存して変化する．そのため，その N ベストリストを用いて推定されるスケーリング係数はその初期値に依存することとなる．すなわち，この方法は初期値の違いに対して頑健ではない．また，Mak ら [17] は，スケーリング係数を線形計画法で推定している．この手法では，目的関数がスケーリング係数の線形関数である必要があり，汎用性に欠ける．

近年，音響モデルや言語モデルのパラメータ推定において，連鎖構造を対数線形モデルで表現した条件付き確率場 (CRF) を用いる方法が提案されている [4], [5]．これらの方法では，対立候補の表現として単語ラティスを用いている．単語ラティスは，N ベストリストよりも対立候補の重複が少なく，計算コスト・記憶コストが少ない．識別的方法によるスケーリング係数推定においても，対立候補の表現として単語ラティスを用いることで，より効率的な推定が可能になることが期待できる．

本論文では，音声認識において，効率的，かつ，初期値の違いに対し頑健にスケーリング係数を推定する手法を提案する．本手法では，対立候補として単語ラティスを用い，こう配法によりスケーリング係数を推定することで，効率的な推定を可能とする．また，単語ラティスの生成とスケーリング係数の推定とを交互に繰り返す過程を導入することで（例えば [16]），初期値に依存しない頑健なパラメータ推定を行う．

一般に，音声認識の評価尺度としては，単語正解精度が用いられており，パラメータ推定の基準としてはこれを直接最大化するものが望ましい．先のスケーリング係数推定手法 [3] では平滑化誤り数基準が，対数線形モデルを用いた音響・言語モデルのパラメータ推定手法 [4], [5] では最大相互情報量 (Maximum Mutual Information; MMI) 基準が用いられている [6]．また，同じ目的に最小分類誤り (Minimum Classification Error; MCE) 基準もしばしば用いられる [7]．これらの基準は，単語正解精度を直接に最大化するものではなく，それを直接最大化する，最小単語誤り率 (Minimum Word Error; MWE) 基準 [9] の方がパラメータ推定基準として適していると考えられる．そこで，我々は提案手法で用いるパラメータ推定基準として，MMI [10]，MCE，MWE の 3 種類を試行し，評価実験においてその得失を明らかにする．

以下，**2.** で，一般的な対数線形モデルとその重みパラメータの最適化について説明する．続く **3.** で，対数線形モデルの音声認識への適用と，そこにおけるスケーリング係数の推定手法を提案する．**4.** で，スケーリング係数の初期値依存を避けるための繰返し過程を述べたのち，**5.** で，提案手法の有効性を評価実験により示す．

2. 対数線形モデル

2.1 定　　義

観測系列 $\boldsymbol{x} = (x^{(1)}, \cdots, x^{(|\boldsymbol{x}|)})$ が与えられたとき，ラベル列 $\boldsymbol{y} = (y^{(1)}, \cdots, y^{(|\boldsymbol{y}|)})$ の事後確率を対数線形モデルを用いて次のように表現する．

$$p_{\boldsymbol{\Lambda}}(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp\left(\sum_{k=1}^{K} \lambda_k F_k(\boldsymbol{x}, \boldsymbol{y})\right)}{Z_{\boldsymbol{\Lambda}}(\boldsymbol{x})} \tag{1}$$

ここで，$F_k(\boldsymbol{x}, \boldsymbol{y})$ は，観測系列 $\boldsymbol{x}$ とラベル列 $\boldsymbol{y}$ が入力されたときの k 番目の素性関数，λ_k は k 番目の素性関数の重み係数，K は素性関数の数である．また，$\boldsymbol{\Lambda}$ は重み係数の集合 $\boldsymbol{\Lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_K)$ を表し，$|\boldsymbol{x}|$，$|\boldsymbol{y}|$ はそれぞれ，観測系列とラベル列の個数である．$Z_{\boldsymbol{\Lambda}}(\boldsymbol{x})$ は，出現可能な全ラベル列の確率の和を 1 にするための正規化項（分配関数）で，$Z_{\boldsymbol{\Lambda}}(\boldsymbol{x}) = \sum_{\boldsymbol{y}' \in \mathcal{Y}} \exp\left(\sum_{k=1}^{K} \lambda_k F_k(\boldsymbol{x}, \boldsymbol{y}')\right)$ である．ここで，$\mathcal{Y}$ は全ての可能なラベル列を表す．ラベル列の分類問題は観測系列が与えられたときに事後確率が最大となるラベル列を選ぶ問題として定義される．

2.2 パラメータ推定基準

対数線形モデルの重み係数集合 $\boldsymbol{\Lambda}$ は，学習サンプル $\{(\boldsymbol{x}_1, \boldsymbol{y}_1), \cdots, (\boldsymbol{x}_R, \boldsymbol{y}_R)\}$ を用い，事後確率最大基準で推定される．ここでは，事後確率の対数に -1 を乗じたものを目的関数として，その最小化を行うこととする．これは，音声認識の分野における MMI 基準に相当する．また，それ以外に MCE 基準や MWE 基準もしばしば用いられる．以下，それぞれの基準の定義，及び，各々の基準を用いる際のパラメータ推定方法について述べる(注1)．

(注1)：いずれの目的関数においても，式 (1) の事後確率を用いることで，分母の正規化項を通じて対立候補（ラベル列）の確率が $\boldsymbol{\Lambda}$ の推定に用いられることとなる．

MMI，MCE，MWE の目的関数はそれぞれ

$$L_{\mathrm{MMI}}(\boldsymbol{\Lambda}) = -\sum_{r=1}^{R} \log p_{\boldsymbol{\Lambda}}(\boldsymbol{y}_r|\boldsymbol{x}_r) \tag{2}$$

$$L_{\mathrm{MCE}}(\boldsymbol{\Lambda}) = \sum_{r=1}^{R} (1 - p_{\boldsymbol{\Lambda}}(\boldsymbol{y}_r|\boldsymbol{x}_r)) \tag{3}$$

$$L_{\mathrm{MWE}}(\boldsymbol{\Lambda}) = \sum_{r=1}^{R} \sum_{\boldsymbol{y}\in\mathcal{Y}} l(\boldsymbol{y}, \boldsymbol{y}_r) p_{\boldsymbol{\Lambda}}(\boldsymbol{y}|\boldsymbol{x}_r) \tag{4}$$

と定義される．ここで，式 (3) は，文献 [7] において，ラベル列 $\boldsymbol{y}_r$ をクラスとみなし，以下の式 (5)〜(8) を用いることで導出される [20]．

$$g_r(\boldsymbol{x}) = \log p_{\boldsymbol{\Lambda}}(\boldsymbol{x}, \boldsymbol{y}_r) \tag{5}$$

$$d_r(\boldsymbol{x}) = -g_r(\boldsymbol{x}) + \log \left[\sum_{r', r'\neq r} \exp g_{r'}(\boldsymbol{x}) \right] \tag{6}$$

$$= \log \frac{1 - p_{\boldsymbol{\Lambda}}(\boldsymbol{y}_r|\boldsymbol{x})}{p_{\boldsymbol{\Lambda}}(\boldsymbol{y}_r|\boldsymbol{x})}$$

$$l_r(d_r) = \frac{1}{1 + \exp(-d_r)} = 1 - p_{\boldsymbol{\Lambda}}(\boldsymbol{y}_r|\boldsymbol{x}) \tag{7}$$

$$L_{\mathrm{MCE}}(\boldsymbol{\Lambda}) = \sum_{r=1}^{R} l_r(d_r) = \sum_{r=1}^{R} (1 - p_{\boldsymbol{\Lambda}}(\boldsymbol{y}_r|\boldsymbol{x}_r)) \tag{8}$$

ここで，$g_r(\boldsymbol{x})$ は識別関数，$d_r(\boldsymbol{x})$ は分類誤り尺度，$l_r(d_r)$ は損失関数である．

目的関数 $L_{\mathrm{MWE}}(\boldsymbol{\Lambda})$ における $l(\boldsymbol{y}, \boldsymbol{y}_r)$ は損失関数であり，正解ラベル列 $\boldsymbol{y}_r$ と任意のラベル列 $\boldsymbol{y}$ の違いを定量的に表現する．ここでは，式 (9) を用いる．

$$l(\boldsymbol{y}, \boldsymbol{y}_r) \equiv \sum_{i=1}^{|\boldsymbol{y}|} \begin{cases} -1 & \text{if } y^{(i)} = y_r^{(i)} \\ 1 & \text{if } y^{(i)} \neq y_r^{(i)} \end{cases} \tag{9}$$

ここで，$y^{(i)}$，$y_r^{(i)}$ はそれぞれ，$\boldsymbol{y}$，$\boldsymbol{y}_r$ の i 番目のラベルである．ラベルが正解と一致する場合に -1，しない場合に 1 のスコアを与える．

上式から容易に分かるように，MMI では正解ラベル列とそれ以外のラベル列のゆう度差が大きくなるようにパラメータの値が推定される．MCE では学習データセットで与えられるそれぞれのサンプルごとのラベル列 $\boldsymbol{y}_r$ の誤りの期待値を最小にするようにパラメータの値が推定される．MWE ではラベル列 $\boldsymbol{y}_r$ における各々の構成要素 $y_r^{(i)}$ における誤りの期待値を最小にするようにパラメータの値が推定される．また，MCE と MWE は，誤り期待値計算の単位が違うだけで，$L_{\mathrm{MWE}}(\boldsymbol{\Lambda})$ の損失関数を次のように定義すると同じ種類の目的関数であることが分かる．

$$l(\boldsymbol{y}, \boldsymbol{y}_r) \equiv \begin{cases} 1 & \text{if } \boldsymbol{y} = \boldsymbol{y}_r \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

2.3 重み係数の推定

本節では，式 (2)，(3)，(4) の各々の目的関数を最小化する重み係数 $\boldsymbol{\Lambda}$ の推定方法について述べる．共通の枠組みとしてこう配法が用いられる．こう配法における t 回目の重み係数の値の更新は次のように表される．

$$\boldsymbol{\Lambda}^{(t)} = \boldsymbol{\Lambda}^{(t-1)} + \eta \triangle \boldsymbol{\Lambda}^{(t-1)} \tag{11}$$

ここで，t は繰返し数，$\boldsymbol{\Lambda}^{(t)}$ は t 回目のパラメータ値，η は学習係数である．$\triangle\boldsymbol{\Lambda}$ は $t-1$ 回目と t 回目のこう配で，共役こう配法等では導関数 $\nabla L(\boldsymbol{\Lambda})|_{\boldsymbol{\Lambda}^{(t-1)}}$ を用い，シンプレックス法のような導関数を用いない方法では $\boldsymbol{\Lambda}^{(t-1)}$ からの差分になる．

重み係数 $\boldsymbol{\Lambda}$ は，式 (12) で表される収束条件を満足するまで更新が繰り返される．

$$\left| \frac{L(\boldsymbol{\Lambda}^t) - L(\boldsymbol{\Lambda}^{t-1})}{L(\boldsymbol{\Lambda}^{t-1})} \right| \leq D \tag{12}$$

3. 対数線形モデルを用いた音声認識

3.1 導　　入

音声認識の問題は，音響信号の観測系列 $\boldsymbol{o}$ が得られたときの単語列 $\boldsymbol{w}$ の事後確率 $p(\boldsymbol{w}|\boldsymbol{o})$ が最大となる単語列 $\hat{\boldsymbol{w}}$ を求めることと定義される．このとき，事後確率は直接計算できないため，ベイズの定理を用いて生成モデルである音響モデルと言語モデルの確率の積を計算する（生成モデルによるアプローチ）．

$$\hat{\boldsymbol{w}} = \operatorname*{argmax}_{\boldsymbol{w}} p(\boldsymbol{w}|\boldsymbol{o}) = \operatorname*{argmax}_{\boldsymbol{w}} \frac{p(\boldsymbol{o}|\boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{o})} \tag{13}$$

ここで，$p(\boldsymbol{o}|\boldsymbol{w})$ は単語列 $\boldsymbol{w}$ が得られたときに観測系列 $\boldsymbol{o}$ が生成される確率で，隠れマルコフモデル (HMM) で表される．$p(\boldsymbol{w})$ は単語列 $\boldsymbol{w}$ が観測される確率で，N グラムモデルや，それに加え単語の挿入誤りを減らすための単語挿入ペナルティーが用いられる．更に，各モデルの出力値の値域の大小のバランスをとるためのスケーリング係数が導入され，式 (13) は次式のように再定義される．

$$\hat{\boldsymbol{w}} \simeq \operatorname*{argmax}_{\boldsymbol{w}} \frac{p(\boldsymbol{o}|\boldsymbol{w})p(\boldsymbol{w})^{\kappa_1} e^{\kappa_2 N}}{p(\boldsymbol{o})} \tag{14}$$

ここで，N は $\boldsymbol{w}$ に含まれる単語数，$e^{\kappa_2 N}$ は単語挿入ペナルティー，κ_1，κ_2 はスケーリング係数を表す．**1.** で述べたように，これらのスケーリング係数は最ゆう推定で求めることができない．そこで，前章で説明した対数線形モデルを導入する．

3.2 素性関数

本節では，音声認識の各モデルを対数線形モデルで表現するための，重み係数の定義について述べる．

素性関数を，式 (14) で用いられる HMM と N グラムモデル，単語挿入ペナルティーを用いて次式のように定義する．

$$\begin{aligned} F_1(\boldsymbol{o},\boldsymbol{w}) &= \log p(\boldsymbol{o}|\boldsymbol{w}) \\ F_2(\boldsymbol{o},\boldsymbol{w}) &= \log p(\boldsymbol{w}) \\ F_3(\boldsymbol{o},\boldsymbol{w}) &= N \end{aligned} \tag{15}$$

式 (1) のラベル列 $\boldsymbol{y}$ は単語列 $\boldsymbol{w} = (w_1, \cdots, w_N)$ に，観測時系列 $\boldsymbol{x}$ は $\boldsymbol{o} = (o_1, \cdots, o_N)$ に置き換えられている．ここで o_i は単語 w_i に対応する観測時系列，N は単語列 $\boldsymbol{w}$ に含まれる単語数である．$F_1(\boldsymbol{o},\boldsymbol{w})$，$F_2(\boldsymbol{o},\boldsymbol{w})$，$F_3(\boldsymbol{o},\boldsymbol{w})$ は，それぞれ音響確率，言語確率，単語挿入ペナルティーの対数値を表す．重み係数 λ_1 は 1 に固定する．λ_2 は言語モデルに対するスケーリング係数 κ_1，λ_3 は単語挿入ペナルティーに対するスケーリング係数 κ_2 とする．すなわち次式を用いる．

$$\lambda_1 = 1, \quad \lambda_2 = \kappa_1, \quad \lambda_3 = \kappa_2. \tag{16}$$

音響モデルとして HMM，言語モデルとしてトライグラムモデルを用いた場合，素性関数はそれぞれ次のように表される．

$$F_1(\boldsymbol{o},\boldsymbol{w}) = \sum_{i=1}^{N} \log p_h(o_i|w_i) \tag{17}$$

$$F_2(\boldsymbol{o},\boldsymbol{w}) = \sum_{i=1}^{N} \log p_n(w_i|w_{i-1}, w_{i-2}) \tag{18}$$

ここで，$p_h(o_i|w_i)$ は単語 w_i が得られたときに o_i が出現する確率を表し，o_i の出現確率は w_i にのみ依存すると仮定している．$p_n(w_i|w_{i-1}, w_{i-2})$ はトライグラム確率であり，2 単語 w_{i-1}，w_{i-2} が得られたときの単語 w_i の出現確率である．

3.3 単語ラティスを用いた目的関数の計算

2. で，対数線形モデルの重み係数推定基準の代表的なものとして，MMI，MCE，MWE について説明した．どの基準を用いる場合でも，目的関数が求まれば，スケーリング係数はその変化に対する目的関数のこう配を計算することでその局所解を推定できる（**2.3** を参照）．

音声認識の場合，MMI は文集合全体の認識率，MCE は文認識率，MWE は単語認識精度を最大にする基準である．通常，音声認識の性能尺度としては単語正解精度が用いられる．したがって，スケーリング係数推定も単語正解精度を直接最大化する MWE を用いる場合に最も推定精度が高くなることが期待できる．そこで，MWE についてその目的関数の計算方法を詳しく述べた上で，MMI，MCE の場合についても簡単に触れる．

まず，言語モデルとしてトライグラムを用いる場合，式 (4) は，

$$L_{\mathrm{MWE}}(\boldsymbol{\Lambda}) = \sum_{r=1}^{R} \sum_{\omega \in \Omega} l(\omega, w_r) \gamma_{\boldsymbol{\Lambda}}(\omega) \tag{19}$$

となる．ここで，$\boldsymbol{y}$，$\boldsymbol{y}_r$ に対応する 3 単語連鎖部分をそれぞれ ω，w_r と表した．Ω は出現可能な単語列に存在する全ての 3 単語連鎖の集合である．また，$\gamma_{\boldsymbol{\Lambda}}(\omega)$ は ω の占有確率であり，$p_{\boldsymbol{\Lambda}}(\boldsymbol{y}|\boldsymbol{x}_r)$ を ω 以外の $\boldsymbol{y}$ について周辺化したものである．

式 (19) における，3 単語連鎖 ω の占有確率 $\gamma_{\boldsymbol{\Lambda}}(\omega)$ は，理想的には出現可能な単語列の組合せ全てを用いて計算される必要があるが，現実には不可能である．そこで，ここでは認識によって得られる単語ラティスに含まれる単語列を用いた近似を行う [11]．単語ラティスは多くの単語列の組合せをコンパクトに表現できるため，N ベストリストを用いるよりも多くの単語列の仮説を効率的に扱うことが可能である．

今，単語ラティスにおける始端と終端のノードをそれぞれ B，E と表す．単語 w_i^j はノード i，j 間のアークに対応する．o_i^j をノード h，f 間の遷移に対応する観測系列とする．ノード h からノード f までの 3 単語連鎖を，ω_h^f と定義する（図 1）．

これまでに単語ラティス上の単語（連鎖）占有確率を効率的に計算する方法がいくつか提案されている [8], [12]．これらは，Baum-Welch アルゴリズムにおける前向き，後ろ向き確率を用いた計算と同様の方法である．そこでは，単語連鎖 ω_h^f の占有確率は次のように表される．

$$\gamma_{\boldsymbol{\Lambda}}(\omega_h^f) = \frac{\alpha_{\boldsymbol{\Lambda}}^h \beta_{\boldsymbol{\Lambda}}^f \exp\left(\sum_{k=1}^{K} \lambda_k F_k(o_i^j, \omega_h^f)\right)}{Z_{\boldsymbol{\Lambda}}(\boldsymbol{o})} \tag{20}$$

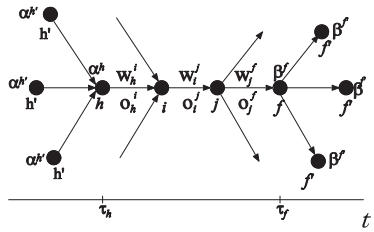


図 1　単語ラティスの例．点 h'，h，i，j，f，f' はノードを表す．τ_h と τ_f はそれぞれノード h と f に対応する時刻を表す．o_i^j と w_i^j はそれぞれノード i とノード j 間のアークに対応する観測量と単語を表す．α^h と β^f はそれぞれ，ノード h の前向きスコア，ノード f の後ろ向きスコアを表す．

Fig. 1　Sample word lattice. Bold points h', h, i, j, f, f' denote nodes. τ_h and τ_f are time for nodes h and f, respectively. o_i^j and w_i^j denote, respectively, an observation and a word corresponding to a directed arc having a beginning node i and an ending node j. α^h and β^f are, respectively, a forward score for node h and a backward score for node f.

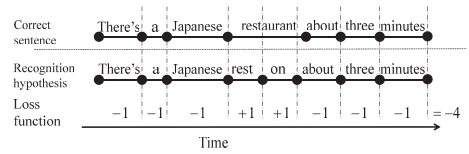


図 2　損失関数の計算例．横軸は時間を表す．黒丸は単語の始端と終端のノードを表す．単語の下の線は単語アークを表す．損失関数は認識結果の単語ごとに正解と照合する．認識した単語の開始時間と終了時間の間に正解単語が一部でも重なれば，正解として数える．

Fig. 2　Example of a computation of the loss function. The horizontal axis denotes time frames. A bold point denotes the starting/ending node of a word, and an arc connecting nodes denotes a word. The loss function is calculated by comparing each word in the recognition hypothesis with its corresponding word in the correct sequence. A word in the hypothesis is *correct* when it overlaps with the same word in the correct sentence in equal or more than one frame.

ここで，$\alpha_{\mathbf{\Lambda}}^h$ はノード h における前向きスコア，$\beta_{\mathbf{\Lambda}}^f$ はノード f における後ろ向きスコアである．前向きスコアは，単語ラティスの先頭から再帰的に計算される．ノード h の前向きスコアは次のようになる．

$$\alpha_{\mathbf{\Lambda}}^h = \sum_{h'} \sum_{h''} \alpha_{\mathbf{\Lambda}}^{h'} \exp\left(\sum_k^K \lambda_k F_k(o_h^i, \omega_{h''}^i) \right) \tag{21}$$

h'' はノード h' に接続されるアークの始端ノードを表す．この式は，ノード h に接続する全てのアークの前向きスコアを全て足し合わせることで前向きスコアが計算できることを表す．後ろ向きスコアは同様の手順で単語ラティスの最後尾のノード E から計算される．式 (20) 中の分配関数 $Z_{\mathbf{\Lambda}}(\boldsymbol{o})$ は，前向きスコアと後ろ向きスコアを用いて次のように表される．

$$Z_{\mathbf{\Lambda}}(\boldsymbol{o}) = \alpha_{\mathbf{\Lambda}}^E = \beta_{\mathbf{\Lambda}}^B \tag{22}$$

次に式 (19) の損失関数 $l(\omega, w_r)$ の計算について説明する．この式では，単語ラティスに出現する各単語と正解単語列中の対応する単語（列）とを比較し，式 (9) で示されるように正解の場合は -1，不正解の場合 1 とする．ここで，単語ごとのスコアの計算を図 2 を使って具体的に説明する．認識結果「There's」「a」「Japanese」「about」「three」「minutes」はそれぞれの単語が出現する時刻の正解単語と一致するため，損失関数の値を -1 とする．一方，「rest」「on」は対応する正解単語列が「restaurant」となっているため，損失関数の値を 1 とする．このように計算することで，「rest」「on」が置換誤り，挿入誤りのいずれでも共に損失となる．

以上で求められた個々の単語（列）の占有確率 $\gamma_{\mathbf{\Lambda}}(\omega_h^f)$ と，その単語の損失関数 $l(\omega, w_r)$ を掛け，その総和をとることで，式 (19) の MWE の目的関数 $L_{\mathrm{MWE}}(\mathbf{\Lambda})$ を計算する．

MCE の目的関数は，正解単語列のゆう度と式 (22) で計算される分配関数との比を計算し，全ての学習サンプルについてその和をとったものである．MMI の目的関数は，やはり正解単語列のゆう度と分配関数との比を計算した上で，全ての学習サンプルについてその積をとったものである．

Mak ら [17] は，スケーリング係数を線形計画法で推定している．この方法は，こう配法を用いる本手法に比べ一般に計算量は少ない．しかし，識別学習の目的関数がスケーリング係数の線形関数でなければならないという制約がある．本研究で用いている 3 種類の目的関数はいずれもスケーリング係数の非線型関数であり，この手法をそのまま用いることはできない．

提案手法は，組み合わせるモデル数が増えてもその演算量にべき乗の増加は生じず，更にモデル数を増やした場合に対しても容易に適用可能である．例えば，言語モデルを複数用いる方法 [18], [19] への適用が考えられる．

4. 繰返し過程の導入

前章で述べたように，単語ラティスには多くの異なる単語列が含まれており N ベストリストを用いるよりも多くの単語列の仮説を効率的に扱うことが可能である．しかしながら，異なる単語列の数は理想的な場合と比較して依然少ないため，スケーリング係数の推定値は単語ラティスに含まれる単語列に依存する．一方，単語ラティスは，その生成時に用いられたスケーリング係数の値に依存する．すなわち，単語ラティスとスケーリング係数は相互に依存している．このことは，スケーリング係数の推定値がその初期値に依存することを意味する．

この初期値への依存性を軽減するために，単語ラティスの生成とスケーリング係数の推定を交互に行う，繰返し過程を導入する（例えば[16]）．スケーリング係数の推定により認識精度が向上し，それにより単語ラティスにより正解に近い単語列が競合仮説として含まれるようになり，正の循環が起きて，初期値の影響を受けずに推定ができることが期待される．

アルゴリズムの概要を図 3 に示す．図 3 中の Step 1 で，スケーリング係数の初期値を任意の値に設定する．Step 2 で，音声認識システムを用い，全ての学習音声データについて単語ラティスを生成する．Step 3 では，式 (11) のこう配法でスケーリング係数を更新する．Step 4 では，こう配法で更新された目的関数の値が式 (12) の収束条件を満たす場合，推定処理を終える．満たさない場合は Step 5 に移る．Step 5 では，式 (12) の収束条件を満たす場合，スケーリング係数を更新して単語ラティスを生成し，Step 2 へ飛ぶ．満たさない場合，Step 3 へ飛ぶ．ここで D_{out} は，ある時点で生成された単語ラティスを用いたときのこう配法の収束条件であり，D_{in} は単語ラティスの更新を含めた繰返し過程全体の収束条件である．D_{in}，D_{out} に対し，$D_{\text{in}} \leq D_{\text{out}}$ の制約を設ける．こう配法，すなわち D_{out} の収束条件を満たさない場合は，既に生成された単語ラティスを用いてスケーリング係数の推定を行うことで効率的な探索を行う．なお，生成された単語ラティスに正解の単語列が含まれていない場合，Step 3 における推定ができない．この問題を避けるため，正解の含まれていない単語ラティスに正解単語列を統合する処理を行う．

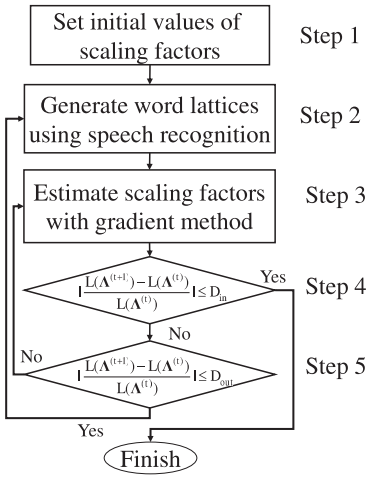


図 3　繰返し過程によるスケーリング係数の推定．
Fig. 3　Iteration process for the estimation of scaling factors.

5. 評価実験

提案手法の評価のために，大規模音声データベースを用いた大語彙連続音声認識の実験を行った．まず，3 種類の推定基準，MMI，MCE，MWE それぞれについて，その目的関数の値と音声認識率との関係を調査した．次に，推定されたスケーリング係数の初期値への依存性を解析した．

5.1 実験条件

データベースとして，日本語話し言葉コーパス (CSJ) を用いた[13]．音響モデルの学習には 540 時間の音声とその書き起こしを用いた．言語モデルの学習には，音響モデルの学習に用いた音声の書き起こしを用いた．総単語数は 560 万単語である．スケーリング係数の推定と認識評価は，学習データとは別に 1.66 時間の音声データ（2614 発声，話者 10 名）を用いて行った．音声分析はフレーム周期 10 ms，窓幅 23 ms で行った．特徴量は，12 次元の MFCC とそれらの 1 次と 2 次の回帰係数，パワーの 1 次と 2 次の回帰係数，調波性特徴量とその 1 次回帰係数の合計 40 次元を用いた[14]．HMM は状態数を 3000 とし，各状態の混合正規分布数を 32 とした．認識は単語ラティスの生成とリスコアの 2 パスで行い，単語ラティスの生成には言語モデルとして語彙数 71730 のバイグラム 45 万個を用いた．また，リスコアには 120 万個のトライグラムを用いた．スケーリング係数の推定にも同じトライグラムを用いた．また，全ての言語モデルにおいて，Katz のバックオフ平滑化を行った．

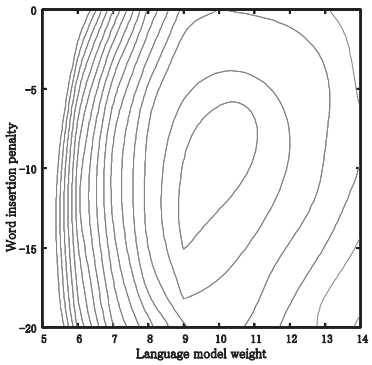


図 4　スケーリング係数と単語正解精度．縦軸は単語挿入ペナルティー，横軸は言語モデル重み．等高線は同じ単語正解精度をもつ点を結んだもの．最も内側の等高線が最高の 80.7%を示し，外側にいくごとに 0.2 ポイント精度が低くなる．

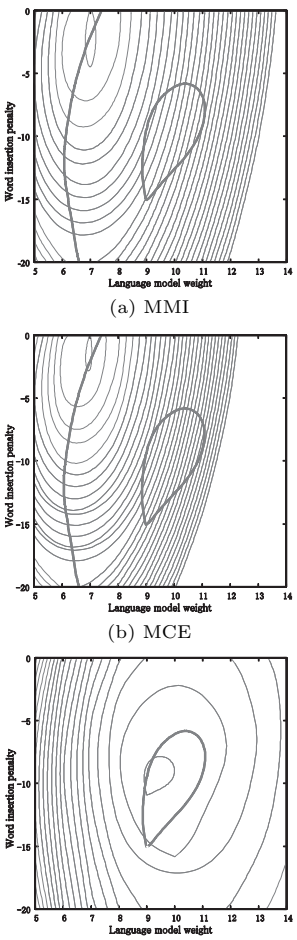
Fig. 4　Relation between word accuracies and scaling factors. Each contour represents a set of points with the same accuracy. The innermost contour indicates the region of the highest word accuracy of 80.7%. The interval in the accuracy between the contours next to each other is 0.2 points.

5.2　スケーリング係数と音声認識率

MMI，MCE，MWE の三つの基準について，目的関数の値と音声認識率の関係を調査した．まず，言語モデルと単語挿入ペナルティーのスケーリング係数の値を，それぞれ等間隔に変化させ，それら格子点上で目的関数の値を計算し，併せて音声認識実験により単語正解精度を求めた．ここでは，言語モデルのスケーリング係数（以下，言語モデル重み）を 5 から 20，単語挿入ペナルティーのスケーリング係数（以下，単語挿入ペナルティー）を −20 から 0 として間隔はいずれも 1 とした．目的関数値の計算には音声認識における第 1 パスで生成された単語ラティスを用いた．この単語ラティスは言語モデル重みを 10，単語挿入ペナルティーを 0 として生成した．図 4 に単語正解精度，図 5 (a) と図 5 (b)，図 5 (c) にそれぞれの目的関数の値を示す．

図 5 (a) と図 5 (b) から，MMI と MCE の目的関数の最小値の領域は，単語正解精度が最高となる領域から離れていることが分かる．目的関数が最小のときの単語正解精度は 78.9%であり，最高の単語正解精度と比較して 1.8 ポイント低いことが分かった．

図 5 (c) の MWE の場合，単語正解精度の高い領域に目的関数の最も小さい領域が重なっている．このことから，MWE を基準とすることで最も単語正解精度



(a) MMI

(b) MCE

(c) MWE

図 5　目的関数の値と単語正解精度．MMI，MCE，MWE のそれぞれの基準について示す．図の縦軸は単語挿入ペナルティー，横軸は言語モデル重み．細線は目的関数の値が同じ点を結んだ等高線で，外側にいくほど目的関数の値は大きくなる．太線は単語正解精度の同じ点を結んだ等高線．中央の囲まれた領域は最も単語正解精度の高い領域．外側の太い線は単語正解精度の等高線のうち，目的関数が最も小さくなる点を通るものである．

Fig. 5　Relations between the estimates of objective functions and word accuracies for three different criteria, (a) MMI, (b) MCE and (c) MWE. Each contour in thin line represents a set of points with the same estimated value of the objective function, where contours closer to the center have smaller estimates. Each contour in thick line represents a set of points with the same word accuracy. The center contour in thick line surrounds the region with the highest word accuracy. The outer contour in thik line passes through the point with the smallest value of the objective function estimates.

を高くするスケーリング係数の値を得ることができることが分かる．

以上から，MMI，MCE で推定されるスケーリング係数は最も高い単語正解精度は得られないが，1.8 ポント程度の誤差で推定可能であることが分かった．一方，MWE では，単語正解精度の最も高くなるスケーリング係数の値を得ることが可能であり，この基準がスケーリング係数の推定に最も適していることが，実験で裏付けられた．

5.3 繰返し過程の評価

本節では，前節でスケーリング係数の推定に最も適していることが分かった MWE を用い，**4.** で述べた繰返し過程を用いたスケーリング係数推定を評価する．ここではこう配法として Nelder-Mead によるシンプレックス法を用いた [15]．

図 6 に繰返しの過程における目的関数の値の変化の様子を示す．収束の判定に用いるしきい値は $D_{\text{in}} = 10^{-5}$，$D_{\text{out}} = 10^{-4}$ とした．四つの異なる初期値から推定を開始した結果を示す．図 6 に矢印で示した点において Step 5 の条件が満たされ，単語ラティスが更新された．初期値がいずれの場合においても，その次の繰返し過程で目的関数の値が大幅に小さくなっていることが分かる．Step 4 の条件を満たし，目的関数の変化が十分に小さくなると，収束したと判定される．今回の実験ではいずれの場合にも，単語ラティスの更新は 1 回で処理を終えた．

次に，図 7 に繰返し過程における推定されたスケーリング係数値の変化の様子を示す．最適値から極端に離れた 4 組の異なる初期値（図の 4 隅の値）を用いて行ったにもかかわらず，全て単語正解精度が最高となる値に収束していることが確認できた．この結果，提案手法が初期値の違いに頑健であることが確認できた．

加藤ら [16] は，やはり単語ラティスを用いて，スケーリング係数の複数の候補値に対しリスコアリングを行い，単語誤り率を最小にする値を選択する方法を提案している．提案手法では，式 (19) の計算が必要なため，一つの候補値に対する計算量は，この手法より大きい．今回の実験条件下で実測したところ，平均 8 倍程度である．しかし，加藤らの手法では，延べ 336 個の候補値に対する計算が必要であるのに対し，本手法の探索過程における候補値は延べ 14 個以下であり，提案手法の計算量の方が少ない．

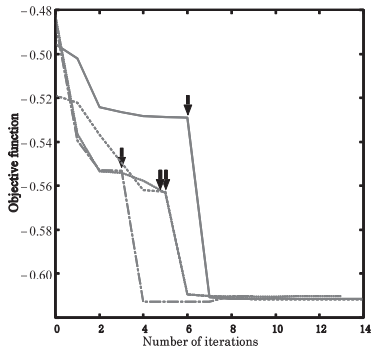


図 6 繰返し過程を用いたスケーリング係数推定における目的関数の値．横軸はこう配法の繰返し回数，縦軸は $L_{\text{MWE}}(\mathbf{\Lambda})$ の値．四つの異なる初期値から推定を開始した場合の結果を 4 本の線で示す．矢印はこう配法 (Step 5) が収束した繰返し回数を表し，単語ラティスの更新が行われた．

Fig. 6 The values of the objective function $L_{\text{MWE}}(\mathbf{\Lambda})$ in the iterative estimation process of the scaling factors. Each of the four lines represents one iteration procedure starting from one of the four different initial settings. The point with an arrow indicates the iteration process in which the gradient method (Step 5) converged. In the next process, the word lattice was updated.

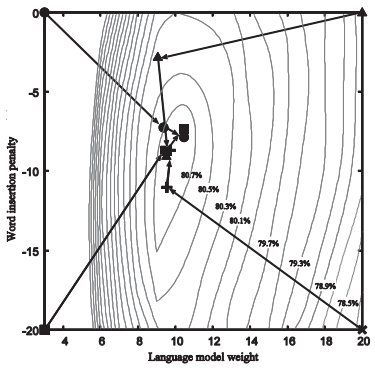


図 7 繰返し過程におけるスケーリング係数値の変化と単語正解精度．横軸は言語モデル重み，縦軸は単語挿入ペナルティー．細線は単語正解精度が同一の点を結んだ等高線で，内側ほど単語正解精度が高い．○，△，□，×は，それぞれ違った初期値を用いて推定されたスケーリング係数の値を示す．それぞれの矢印の先頭の点は，矢印の根元の値を用いてこう配法により推定された値である．

Fig. 7 Results of scaling factor estimation. Each contour represents a set of points with the same word accuracy. Contours closer to the center have higher recognition accuracies. Each of the symbols, ○, △, □, ×, represents the scaling factor estimated from one of the four different initial values. The head of an allow indicates the scaling factor value estimated from the value at its tail in one iteration step.

6. むすび

音声認識におけるスケーリング係数の効率的な推定方法を提案した．提案手法では，スケーリング係数を対数線形モデルの重みパラメータとみなし，その最適値を最小単語誤り率基準 (MWE) を用いて推定した．推定には，単語ラティスを用いたこう配法を用いた．こう配法で推定される値が単語ラティスの影響を受けることを考慮して繰返し過程を導入した結果，初期値の違いに対して頑健に推定ができることを確認した．また，スケーリング係数の目的関数として，MMI 及び MCE では，発見的手法で得られた単語正解精度より 1.8 ポイント低かったが，MWE の場合は同一の精度を得ることができ，MWE が有効であることを確認した．

今後は，モデル数が増加しても演算量が急激に増加しないという本手法の利点を生かし，より多数の確率モデルの組合せを用いた場合におけるスケーリング係数の推定と，それによる認識精度の改善を行いたい．

文　献

[1] L.R. Bahl, R. Bakis, F. Jelinek, and R.L. Mercer, “Language-model/acoustic channel balance mechanism,” IBM Technical Disclosure Bulletin, vol.23, no.7B, pp.3464–3465, Dec. 1980.

[2] A. Ito, M. Kohda, and S. Makino, “Fast optimization of language model weight and insertion penalty from n-best candidates,” Acoustical Science and Technology, vol.26, no.4, pp.384–387, 2005.

[3] P. Beyerlein, “Discriminative model combination,” Proc. ICASSP ’98, pp.481–484, 1998.

[4] M. Mahajan, A. Gunawardana, and A. Acero, “Training. Algorithms for hidden conditional random fields,” Proc. ICASSP 2006, pp.273–276, 2006.

[5] G. Heigold, S. Wiesler, M. Nussbaum, P. Lehnen, R. Schlüter, and H. Ney, “Discriminative HMMs, log-linear models, and CRFs: What is the difference?,” Proc. ICASSP 2010, pp.5546–5549, 2010.

[6] L.R. Bahl, P.F. Brown, P.V. de Souza, and R.L. Mercer, “Maximum mutual information estimation of hidden Markov model parameters for speech recognition,” Proc. lCASSP ’86, pp.49–52, 1986.

[7] B.-H. Juang and S. Katagiri, “Discriminative learning for minimum error classification,” IEEE Trans. Signal Process., vol.40, no.12, pp.3043–3054, 1992.

[8] J. Lafferty, A. McCallum, and F. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” Proc. 18th International Conference on Machine Learning, pp.282–289, 2001.

[9] D. Povey, Discriminative Training for Large Vocabulary Speech Recognition, Ph.D. thesis, Cambridge University, 2004.

[10] T. Emori, Y. Onishi, and K. Shinoda, “Automatic estimation of scaling factors among probabilistic models in speech recognition,” Proc. Interspeech 2007, pp.1453–1456, 2007.

[11] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, “The HTK Book. Revised for HTK Version 3.2,” http://htk.eng.cam.ac.uk/, Dec. 2002.

[12] F. Wessel, R. Schlüter, K. Macherey, and H. Ney, “Confidence measures for large vocabulary continuous speech recognition,” IEEE Trans. Speech Audio Process., vol.9, no.3, pp.288–298, March 2001.

[13] S. Furui, K. Maekawa, H. Isahara, H. Shinozaki, and T. Ohdaira, “Toward the realization of spontaneous speech recognition — Introduction of a Japanese priority program and preliminary results,” Proc. ICSLP, vol.3, pp.518–521, 2000.

[14] K. Takagi and T. Watanabe, “Utilization of spectral harmonics structure for speech recognition,” Proc. Meeting of the Acoustical Society of Japan, pp.3–4, Sept. 1997.

[15] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, Numerical Recipes in C, Cambridge University Press, 1992.

[16] 加藤正治，斎院俊典，伊藤彰則，好田正紀，“単語グラフ生成におけるパラメータ最適化の検討，” 情処学研報，SLP，音声言語情報処理，vol.119, pp.107–112, 2000.

[17] B. Mak and T. Ko, “Min-max discriminative training of decoding parameters using iterative linear programming,” Proc. Interspeech 2008, pp.915–918, 2008.

[18] D. Klakow, “Log-linear interpolation of language models,” Proc. ICSLP, vol.5, pp.1695–1698, 1998.

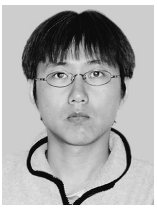
[19] S. Broman and M. Kurimo, “Methods for combining language models in speech recognition,” Proc. Interspeech 2005, pp.1317–1320, 2005.

[20] X. He, L. Deng, and W. Chou, “Discriminative learning in sequential pattern recognition,” IEEE Signal Process. Mag., vol.25, pp.14–36, 2008.

（平成 23 年 8 月 8 日受付，12 月 2 日再受付）

大西　祥史

1995 阪大・基礎工・物性卒．2000 同大大学院博士課程了，理博．同年 NEC 入社．以来，音声認識の研究開発に従事．現在，NEC 情報・メディアプロセッシング研究所勤務．日本音響学会，日本物理学会各会員．



江森　　正

1993 東京理科大・理工・物理卒．1995 慶應大学大学院物理学修士課程了．同年 NEC 入社．以来，音声認識の研究に従事．現在，ヤフー（株）R&D 統括本部勤務．日本音響学会会員．

越仲　孝文　（正員）

1991 京大・工・航空卒．1993 同大大学院修士課程了．同年 NEC 入社．以来，音声・画像パターン認識の研究に従事．現在，NEC 情報・メディアプロセッシング研究所主任研究員．2000 本会学術奨励賞受賞．日本音響学会会員．

篠田　浩一　（正員：シニア会員）

1987 東大・理・物理卒．1989 同大大学院修士課程了．同年 NEC 入社．以来，音声・動画像パターン認識，ヒューマンインタフェースの研究に従事．1997〜1998 米国ルーセントテクノロジ・ベル研究所客員研究員．2001 東京大学大学院情報理工学研究科助教授，2003 東京工業大学大学院情報理工学研究科助教授，国立統計数理研究所客員助教授．現在，東京工業大学大学院情報理工学研究科准教授．1997 日本音響学会栗屋学術奨励賞，1998 本会論文賞各受賞．日本音響学会，IEEE，ACM，情報処理学会，人工知能学会各会員．