社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2003－ＳＬＰ－49　（８）
2003／12／18

変分ベイズ法を用いた逐次状態分割法

實廣　貴敏†　　中村　　哲†

† ATR 音声言語コミュニケーション研究所
〒 619–0288 「けいはんな学研都市」光台二丁目２番地２
E-mail: †{takatoshi.jitsuhiro,satoshi.nakamura}@atr.co.jp

あらまし　音素環境依存型音響モデルを作成する場合，一般に学習データを用いて，小規模なモデルからある規準でモデルを選択しつつ，より大規模なモデルを構築する．この際，よく使われているのが，ゆう度最大化 (Maximaum Likelihood, ML) 規準である．しかし，一般的なモデル選択での問題であるように，ML 規準は過学習になりやすいという問題がある．これを避けるために情報量規準が使われる場合が多い．しかし，一般的な情報量規準は理論的には HMM のような複雑なモデルを厳密には扱うことができない．そこで，より正確に扱うことのできるものとして，変分ベイズ法が近年，機械学習の分野で提案され，多くの分野で応用されてきている．本報告では変分ベイズ法を用い，時間方向の状態数が非均一な環境依存型 HMM を自動で作成する手法を提案する．音響モデル構造を構築する方法としては音素環境方向だけでなく，時間方向分割も考慮できる逐次状態分割法を用い，分割条件の規準および停止条件に変分ベイズ法を適用する．評価実験として，音素識別実験を行い，主に母音で効果が見られた．また，連続音声認識実験ではベースラインに対し，状態数が約 60%のモデルでさらに若干良い性能が得られた．
キーワード　音声認識, 音響モデル, 構造学習, SSS アルゴリズム, 変分ベイズ法

The Succesive State Splitting Algorithm based on the Variational Bayesian Approach

Takatoshi JITSUHIRO† and Satoshi NAKAMURA†

† ATR Spoken Language Translation Research Laboratories
2–2–2 Hikaridai, "Keihanna Science City" Kyoto 619–0288 Japan
E-mail: †{takatoshi.jitsuhiro,satoshi.nakamura}@atr.co.jp

Abstract　We propose a method applying the Variational Bayesian (VB) approach to automatically creating non-uniform, context-dependent HMM topologies. The Maximum Likelihood (ML) criterion is generally used to create HMM topologies. However, it has an over-fitting problem. Information criteria have been used to overcome this problem, but theoretically they cannot be applied to complicated models like HMMs. Recently, to avoid these problems, the VB approach has been developed in the machine-learning field. We introduce the VB approach to the Successive State Splitting (SSS) algorithm, which can create both contextual and temporal variations for HMMs. We define the prior and posterior probability densities and free energy with latent variables as split and stop criteria. Experimental results show that the proposed method can automatically create a more efficient model and obtain better performance, especially for vowels, than the original method.
Key words　speech recognition, acoustic model, topology training, SSS algorithm, variational Bayesian approach

1. はじめに

本研究では，音声認識に用いる音響モデル作成において，より複雑でより適切なモデルを自動的に学習データから得られる手法について検討を行っている．

音響モデル作成における，HMM の状態共有構造を自動生成する方法として，代表的なものに音素決定木クラスタリング [1]，逐次状態分割法 (Successive State Splitting (SSS) algorithm) [2] がある．前者は音声学的知識による音素カテゴリを利用して音素環境のクラスタリングを行っているのに対し，後者は音素環境方向と時間方向との二方向の状態分割をデータに依存して行う．

音素決定木クラスタリングによる方法では，従来，ゆう度最大 (Maximum Likelihood (ML)) 規準で分割する状態を選択する．しかし，一般にモデルパラメータ数が増加すれば，ゆう度は増

加するため，ML 規準のみを停止条件にすることは困難であった．一般的なモデル選択において ML 規準では，過学習になることが指摘されている．

そこで，音素決定木クラスタリングによる方法において，Minimum Description Length (MDL) 規準[4]，Bayesian Information Criterion (BIC) [5] [6] といった情報量規準を分割条件および停止条件として用いる手法が提案されている．これらの手法では情報量規準の値が改善されなくなるまで状態を分割する．情報量規準により，学習サンプル数およびモデルパラメータ数とモデルの対数尤度との関係を考慮することで，過学習を回避できる．

一方，SSS アルゴリズム [2] は初め，話者依存モデルの HMM 状態ネットワーク (隠れマルコフ網: HMnet) 作成方法として提案され，現在は ML-SSS 法 [7] として不特定話者モデル作成方法へと拡張されている．これらの方法は音素決定木クラスタリングと同様，ML 規準を分割条件に用いており，従来の音素決定木クラスタリングによる方法と同じ問題を含んでいた．そこで，著者らは ML-SSS 法において，情報量規準として MDL 規準を分割および停止条件として導入し，MDL 規準を音素環境方向および時間方向の両分割に用いた方法を提案した [8]．

しかし，従来の情報量規準はいくつかの仮定の上で導出されており，ニューラルネットワークや HMM のような複雑な統計モデルは，その仮定を満たしていない [9]．したがって，情報量規準では厳密には，そのようなモデルを扱うことができない．ただし，多くの研究報告では効果があることから，実際には大まかな近似としてのみ，利用できると考えられる．

ML 学習において問題となる過学習を緩和することができるベイズ学習において，変分法を用いて実現する変分ベイズ法 (Variational Bayesian Approach: VB) が，近年，機械学習の分野において提案されている [10] [11]．現在，多くの分野での応用が試みられてきており，音声処理においても検討されつつある．また，文献 [12] などのように，この手法をモデル選択に用いることが可能である．変分ベイズ法による目的関数をモデル選択の規準に用いることで，複雑な統計モデルであっても大きな制約なしに評価することができる．その利点を生かして，さらに音声認識の分野において，決定木クラスタリングを用いた音響モデル構造作成方法に応用されている [13]．しかし，この文献 [13] による方法では，従来の決定木クラスタリング通り，状態ごとの区間を既知として定義しており，変分ベイズ法で一般に使用される潜在変数は導入されていない．

そこで，我々はより複雑なモデルを作成可能なアルゴリズムとして，ML-SSS 法を基に変分ベイズ法を導入した方法を提案する．音素環境方向および時間方向分割を有する構造作成方法において変分ベイズ法の定式化を行うことになる．本手法では時間方向分割も行うことから，音素区間のみ既知で状態区間が未知であることが前提のため，潜在変数を導入する必要がある．

以下，本報告では，2. でベースラインとなる ML-SSS アルゴリズムに触れた後，3. で提案手法である変分ベイズ法を用いた逐次状態分割法について，その概要，事前分布や事後分布の定義を述べる．4. では，提案手法を 2 種類の実験で評価し，5. で結論を述べる．

2. ML-SSS アルゴリズム

2.1 ML-SSS アルゴリズムの概略と問題点

まず，従来法として ML-SSS 法 [7] について概説する．ML-SSS 法では各状態を音素環境方向と時間方向に分割を試み，ML 規準で分割する状態を選択する．これを総状態数 N_s に達するまで繰り返すことで，状態共有構造を生成する．音素決定木クラスタリングでは環境方向の分割に音素カテゴリに関する質問で分割する．これに対して ML-SSS では，triphone データごとにクラスタリングを行い，ML 規準で分割を決定している．時間方向分割では，分割対象の状態の分布パラメータを複製して 2 状態を時間方向に連結し，分布パラメータを forward-backward アルゴリズムで再推定する．しかし，理論上，状態を分割するたびに分布パラメータを求めるには，その前後の状態を含めて再推定を行う必要がある．それでは計算量が膨大になるため，ML-SSS では，近似的に対象状態のみについてゆう度計算をし，分割前の状態滞在確率で重みをかけた値を用いる．

さらに ML-SSS では総状態数 N_s を停止条件として必要とする．また，時間方向分割に対して HMnet 内のパス長制限，つまり triphone での時間方向の状態数制限 N_p を必要とする．この制約がないと ML 規準の性質から，時間方向にはデータのある限り長くなる傾向がある．したがって，あるデータに対して最適なモデルをひとつ決めるためには，一般にこの 2 つの値を実験的に求める必要がある．学習データ，評価データが変われば，その度ごとに調整しなければならない．

3. 変分ベイズ法を用いた逐次状態分割法

図 1 に提案手法である VB-SSS アルゴリズムを示す．初期モデル構造をセットし，HMM のパラメータを推定しておく．次にハイパーパラメータの初期値を設定する．各状態において，変分ベイズ法を用いて事後ハイパーパラメータを推定し，分割に対する目的関数 $\mathcal{F}^{(VB)}$ を計算しておく．分割前後でこの目的関数のゲインが大きい分割を選択する．状態分割は ML-SSS 法と同様で音素環境方向と時間方向に分割を行う．それぞれの分割において，事後ハイパーパラメータを推定し，目的関数を計算，分割前との目的関数のゲイン，環境方向分 $\Delta\mathcal{F}_c^{(VB)}$ と時間方向分 $\Delta\mathcal{F}_t^{(VB)}$ を計算する．このゲインが最大となる分割状態および分割方向を選択する．目的関数が減少あるいは収束した場合，分割を停止する．そうでない場合は，HMM パラメータを再推定し，分割を繰り返す．この規準を用いることにより，N_s および N_p を停止条件とせず，自動でモデル構造を決定できる．この報告では，分割時の事後ハイパーパラメータの推定にはすべてのデータ，状態の計算を行った．

3.1 環境方向および時間方向分割

各状態 1 個のガウス分布と N_a 個の遷移確率を持つ，総状態数 N_s 個の HMM Θ において，確率密度関数は下記のように表せる．

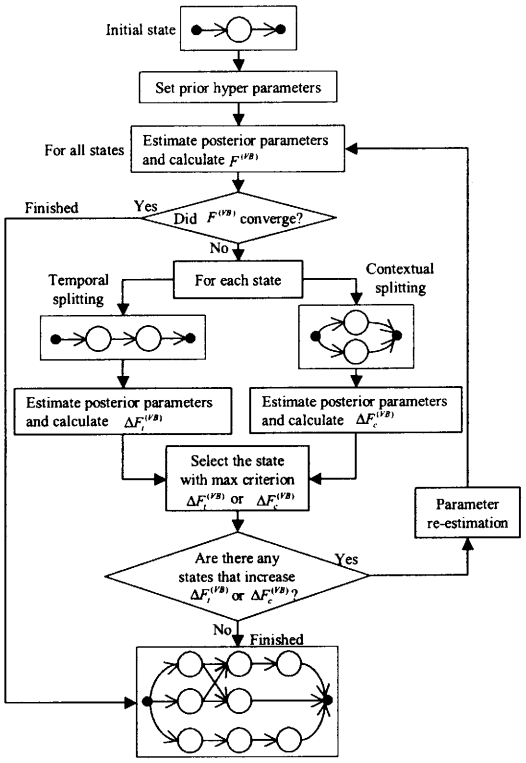

図 1　変分ベイズ逐次状態分割アルゴリズムの流れ図

Fig. 1　Flow of the Variational Bayesian SSS algorithm.

$$p(\boldsymbol{O}|\Theta) = \prod_{t=1}^{T} p(\boldsymbol{o}_t|s_t,\Theta)p(r_{t+1}|s_t,\Theta)$$
$$= \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{o}_t;\boldsymbol{\mu}_{s_t},\boldsymbol{S}_{s_t}^{-1})a_{s_t r_{t+1}} \qquad (1)$$

ここで，$\boldsymbol{O} = \{\boldsymbol{o}_1,\ldots,\boldsymbol{o}_t,\ldots,\boldsymbol{o}_T\}$ は学習データサンプルの集合，s_t は時刻 t での状態番号，$r_t = 1,\ldots,N_a$ は時刻 t での遷移アーク番号を示す．$\boldsymbol{\mu}_{s_t}$ は状態 s_t での出力確率の平均ベクトル，$\boldsymbol{S}_{s_t}$ は精度行列 (共分散行列の逆行列)，$a_{s_t r_{t+1}}$ は遷移確率を示す．共分散行列として対角行列を用いる．遷移確率の数 N_a は最大 N_s 個であるが，SSS 法では 1 状態の遷移確率は自己ループと 1 つの遷移のみ扱えるため，実際には $N_a = 2$ である．

3.2　完全データに対する結合確率分布

SSS アルゴリズムにおける潜在変数を導入した完全データに対する結合確率分布は

$$p(\boldsymbol{O},Z|\Theta) = \prod_{t=1}^{T}\prod_{i=1}^{N_s}\prod_{j=1}^{N_a}\{\mathcal{N}(\boldsymbol{o}_t;\boldsymbol{\mu}_i,\boldsymbol{S}_i^{-1})a_{ij}\}^{z_{ij}^t} \qquad (2)$$

と表せる．ここで，$Z = \{z_{ij}^t\}_{i=1,j=1,t=1}^{N_s,N_a,T}$ は潜在変数の集合である．各サンプル，各状態，各遷移に対応する潜在変数 z_{ij}^t は，第 t 番目サンプルが状態 i の遷移 j に属する確率に相当する．一般的に用いられる音素決定木クラスタリングでは，各状態は単一分布であり，状態単位での境界を固定した上で，さらに音素環境のみ考慮するため，各学習サンプルがどの分布に属するかは一意に決まる．そのため，潜在変数を用いる必要がない．これに対し，SSS アルゴリズムでは，各状態一分布であるが，音素単位での境界を既知とし，状態単位の境界は固定せず，forward-backward アルゴリズムによりパラメータ推定を行う．そのため，例えば時間方向分割の場合，図 2 に示すように各状態に対し，対応する音素境界内の学習サンプルが確率的に属していると考えられる．したがって，潜在変数を用いて各学習サンプルの各分布への寄与度を表現する必要がある．

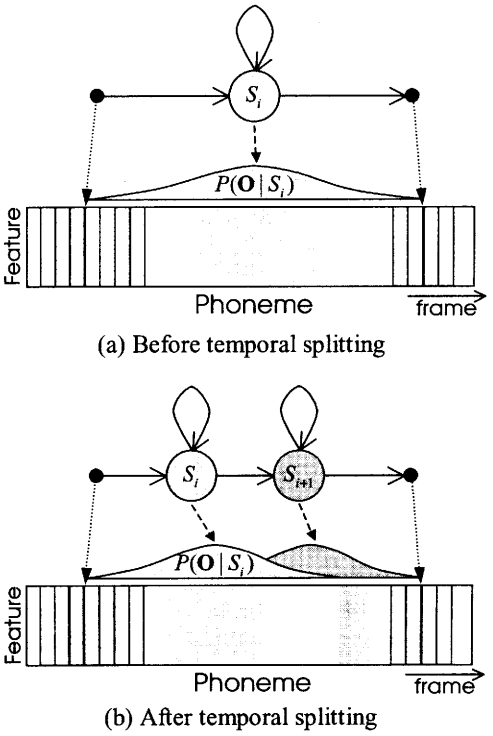

図 2　時間方向分割

Fig. 2　Temporal splitting

HMM パラメータ Θ_i である第 i 状態を第 i_1 状態と第 i_2 状態とに分割し，パラメータ $\hat{\Theta}_i$ が推定されたとき，分割規準は変分ベイズ法の目的関数 $\mathcal{F}^{(VB)}$ を用いて下記のように表される．

$$\Delta\mathcal{F}_{n+1}^{(VB)} = \mathcal{F}_{n+1}^{(VB)}(\hat{\Theta}_i) - \mathcal{F}_n^{(VB)}(\Theta_i) \qquad (3)$$

ここで，n は学習回数である．$\mathcal{F}^{(VB)}$ の定義は 3.5 節で触れる．

3.3　事前分布

確率変数を下記のように分解できると仮定する．

$$p(\Theta) = p(N_s,N_a)p(\boldsymbol{a}|N_s,N_a)p(\boldsymbol{S}|N_s)p(\boldsymbol{\mu}|\boldsymbol{S},N_s) \qquad (4)$$

さらに，遷移確率 $\boldsymbol{a} = \{a_{ij}\}_{i=1,j=1}^{N_s,N_a}$ の事前分布に *Dirichlet* 分布の結合分布，出力確率の平均，共分散 $\{\boldsymbol{\mu},\boldsymbol{S}\} = \{\{\boldsymbol{\mu}_i\}_{i=1}^{N_s},\{\boldsymbol{S}_i\}_{i=1}^{N_s}\}$ の結合事前分布に *normal-Gamma* 分布を

仮定する．

$$p(\boldsymbol{a}|N_s,N_a)=\prod_{i=1}^{N_s}\mathcal{D}(\{a_{ij}\}_{j=1}^{N_a};\phi_0)\propto\prod_{i=1}^{N_s}\prod_{j=1}^{N_a}a_{ij}^{\phi_0-1} \tag{5}$$

$$a_{ij}\geqq 0,\ \sum_{j=1}^{N_a}a_{ij}=1,$$

$$p(\boldsymbol{\mu},\boldsymbol{S}|N_s)=\prod_{i=1}^{N_s}\prod_{k=1}^{D}\mathcal{N}(\mu_{ik};\nu_{0k},\xi_0^{-1}s_{ik}^{-1})\mathcal{G}(s_{ik};\eta_0/2,b_{0k}/2) \tag{6}$$

ここで，D は特徴量次元である．μ_{ik} と s_{ik} はそれぞれ平均ベクトル $\boldsymbol{\mu}_i$ または精度行列 $\boldsymbol{S}_i$ の k 番目の要素．$\mathcal{N}()$ はガウス分布，$\mathcal{G}()$ はガンマ分布を示す．ϕ_0，ν_{0k}，ξ_0，η_0，b_{0k} は各分布の事前ハイパーパラメータ．また，ガンマ分布は下記のように表される．

$$\mathcal{G}(s;\eta,\lambda)=\frac{\lambda^{\eta}}{\Gamma(\eta)}s^{\eta-1}\exp(-\lambda s),\ s>0 \tag{7}$$

ここで，$\Gamma()$ はガンマ関数．

3.4 事後分布

各パラメータの事後分布 (最適変分事後分布) は，下記に示す変分ベイズ法における最適変分事後分布の一般式から導出できる [14].

【変分ベイズ EM ステップ】

Step 1. 初期分布 $q(\theta|m)^{(0)}=\prod_{i=1}^{I}p(\theta_i|m)^{(0)}$ を設定し，$t=0$ とする．

Step 2. 以下を収束するまで繰り返す．

$$q(Z|m)^{(t+1)}=C\exp<\ln p(\boldsymbol{O},Z|\theta,m)>_{q(\theta|m)^{(t)}} \tag{8}$$

for $i=1,\ldots,I$,

$$q(\theta_i|m)^{(t+1)}=C'\exp<\ln p(\boldsymbol{O},Z|\theta,m)>_{q(Z|m)^{(t+1)},q(\theta_{-i}|m)^{(t)}} \tag{9}$$

$t=t+1$ とする．

それぞれ，式 (8) が潜在変数，式 (9) が確率変数の最適変分事後確率分布の推定式を示している．こられの一般式からモデルに合った事後分布を導出しておく．パラメータ推定時には初期値を与えた後，確率変数を固定しては潜在変数の事後分布を推定し，潜在変数を固定しては確率変数の事後分布を推定する．これを変数の値が収束するまで行う．

以後，式 (8)(9) から導出したそれぞれの事後確率を記しておく．遷移確率の事後確率を下記に示す．

$$q(\boldsymbol{a}|\boldsymbol{O},N_s,N_a)=\prod_{i=1}^{N_s}\mathcal{D}(\{a_{ij}\}_{j=1}^{N_a};\{\phi_{ij}\}_{j=1}^{N_a})\propto\prod_{i=1}^{N_s}\prod_{j=1}^{N_a}a_{ij}^{\phi_{ij}-1} \tag{10}$$

ここで，

$$\phi_{ij}=\phi_0+\bar{N}_{ij},\ \bar{N}_{ij}=\sum_{t=1}^{T}\bar{z}_{ij}^t,\ \bar{z}_{ij}^t=<z_{ij}^t>_{q(Z)}$$

平均ベクトルと共分散行列の結合分布は下記のように表される．

$$q(\boldsymbol{\mu},\boldsymbol{S}|\boldsymbol{O},N_s)=\prod_{i=1}^{N_s}\prod_{k=1}^{D}\mathcal{N}(\mu_{ik};\nu_{ik},\xi_i^{-1}s_{ik}^{-1})\mathcal{G}(s_{ik};\eta_i/2,b_{ik}/2) \tag{11}$$

ここで，

$$\bar{N}_i=\sum_{t=1}^{T}\bar{z}_i^t,\ \bar{z}_i^t=<z_i^t>_{q(Z)},$$

$$\nu_{ik}=\frac{\bar{N}_i\bar{o}_{ik}+\xi_0\nu_{0k}}{\bar{N}_i+\xi_0},\ \xi_i=\xi_0+\bar{N}_i,\ \eta_i=\eta_0+\bar{N}_i,$$

$$b_{ik}=b_{0k}+\bar{c}_{ik}+\frac{\bar{N}_i\xi_0}{\bar{N}_i+\xi_0}(\bar{o}_{ik}-\nu_{0k})^2,$$

$$\bar{\boldsymbol{o}}_i=\frac{1}{\bar{N}_i}\sum_{t=1}^{T}\bar{z}_i^t\boldsymbol{o}_t,\ \bar{c}_{ik}=\sum_{t=1}^{T}\bar{z}_i^t(o_{tk}-\bar{o}_{ik})^2$$

である．潜在変数の変分事後確率分布は，

$$\bar{z}_{ij}^t=\exp(\gamma_{ij}^t)/\sum_{k=1}^{N_s}\sum_{l=1}^{N_a}\exp(\gamma_{kl}^t) \tag{12}$$

$$\gamma_{ij}^t\propto\Psi(\phi_{ij})-\Psi(\sum_{j=1}^{N_a}\phi_{ij})+\frac{D}{2}\Psi\left(\frac{\eta_i}{2}\right)-\frac{1}{2}\sum_{k=1}^{D}\ln\frac{b_{ik}}{2}-\frac{\eta_i D}{2\xi_i(\eta_i-D-1)}-\frac{1}{2}\sum_{k=1}^{D}\frac{\eta_i}{b_{ik}}(o_{tk}-\nu_{ik})^2 \tag{13}$$

となる．ここで，$\Psi(x)=\partial\ln\Gamma(x)/\partial x$ は digamma 関数と呼ばれる．

3.5 目的関数

ここでは，$p(\cdot|N_s,N_a)$ の表記を $p(\cdot)$ のように簡便化する．つまり，$p(\Theta|N_s,N_a)\rightarrow p(\Theta)$ とする．目的関数は確率変数間に仮定した式 (4) の関係から以下のような形に変形できる．

$$\begin{aligned}\mathcal{F}^{(VB)}&=\int q(Z)q(\Theta)\ln\frac{p(\boldsymbol{O},Z|\Theta)p(\Theta)}{q(Z)q(\Theta)}dZd\Theta\\&=\int q(Z)q(\Theta)\ln\prod_{t=1}^{T}p(\boldsymbol{o}_t,z_t|\Theta)dZd\Theta\\&\quad-\int q(Z)\ln q(Z)dZ-\int q(\Theta)\ln q(\Theta)d\Theta\\&\quad+\int q(\Theta)\ln p(\Theta)d\Theta\\&=\mathcal{F}_1(\boldsymbol{O},\Theta,Z)+\sum_{i=1}^{N_s}\left\{\mathcal{F}_2(\{a_{ij}\}_{j=1}^{N_a})\right.\\&\quad\left.+\mathcal{F}_3(\boldsymbol{\mu}_i,\boldsymbol{S}_i)+\mathcal{F}_4(\boldsymbol{S}_i)\right\}\end{aligned} \tag{14}$$

ここで，$\mathcal{F}_1,\mathcal{F}_2,\mathcal{F}_3,\mathcal{F}_4$ は詳細は省くが，それぞれの変数に依存した関数となり，具体的に計算できる．これを分割時に毎回計算することになる．

4. 実　　験

4.1 実験条件

この節では，提案手法を切り出し音素認識実験と連続音声認識で評価する．切り出し音素認識実験は音素区間ごとに切り出された区間に対して行う音素識別実験であり，各音素区間に対し，最も良いスコアの音素モデルを認識結果とする．これにより，得られた各音素モデルの性能評価を行う．

提案手法 VB-SSS との比較に，ベースラインとして ML-SSS 法，さらに MDL 規準を分割および停止規準とする MDL-SSS 法を用いた．ML-SSS 法では，最大状態長を 3 または 4 とした 2 種類のモデルを作成した．これをベースラインとする．MDL-SSS 法では，文献 [8] と同様，下記の規準を用いた．

$$G_c^{(MDL)}(S_i) = -G_c^{(ML)}(S_i) + C_c \frac{\alpha_c' - \alpha_c}{2} \log \Gamma(S), \quad (15)$$

$$G_t^{(MDL)}(S_i) = -G_t^{(ML)}(S_i) + C_t \left\{ \frac{\alpha_t'}{2} \log \Gamma'(S) - \frac{\alpha_t}{2} \log \Gamma(S) \right\}, \quad (16)$$

ここで，G_c^{ML} と $G_t^{(ML)}$ はそれぞれ環境方向分割，時間方向分割の対数ゆう度ゲインである．$\Gamma(S) = \sum_i \sum_t \gamma_t(S_i)$ はすべての状態に対するサンプル数の期待値であり，$\gamma_t(S_i)$ は時刻 t に状態 S_i に留まる確率である．$\Gamma'(S)$ は時間方向分割後の値である．α_c, α_c' は分割前後のパラメータ数を示す．重み係数 C_c と C_t は $C_c = 2, C_t = 20$ として用いた．

音響モデル学習データとして，ATR 旅行会話タスク (TRA) データベース [15] より男性 166 人の日本語対話音声，約 2.1 時間を用いた．評価には，同じデータベースから学習データに含まれない 17 人の男性話者による 213 発話を用いた．この論文では，変分ベイズ法を分割および停止時の規準としてのみ用いた．認識実験では従来の HMM による計算を行った．さらに，連続音声認識実験を行った．言語モデルとして，多重クラス複合 bigram モデル [16] を用い，語彙サイズは 5 千単語とした．

サンプリング周波数は 16kHz，フレーム長は 20ms，フレーム周期は 10ms とした．特徴量として 12 次 MFCC，Δ MFCC，Δ 対数パワーを用いた．ケプストラム平均正規化は発話単位で行った．音素体系として 26 種類の音素とひとつの無音を用いた．各音素ごとに音響モデル構造を作成し，初期モデルはそれぞれ 3 状態モデルとした．構造作成中は 1 状態につき 1 ガウス分布を用いた．無音モデルは 3 状態で音素モデルとは別に作成した．

変分ベイズ法は一般に事前ハイパーパラメータの設定が必要になる．いくつかの予備実験から，ここでは，事前ハイパーパラメータを $\phi_0 = 1.0$, $\xi_0 = 1.0$, $\eta_0 = 2.0$ とした．ν_{0k} と b_{0k} は実際の分布の平均ベクトル，精度行列から設定した．

4.2 音素認識実験

図 3，4 に，それぞれ母音，子音に対する音素認識率を示す．"Phoneme recognition rate" は各音素内で正しく識別された比率を示す．図 5 は母音と子音の認識率の平均値を示す．図はそれぞれ最大状態長が 3 または 4 の ML-SSS 法によるもの，提案手法 VB-SSS 法によるものを示している．図はそれぞれ最大状態長が 3 または 4 の ML-SSS 法によるものと MDL-SSS 法によるもの，提案手法 VB-SSS 法によるものを示している．提案手法は 3 種の手法の中で最も良い性能を示した．特に母音では同じ

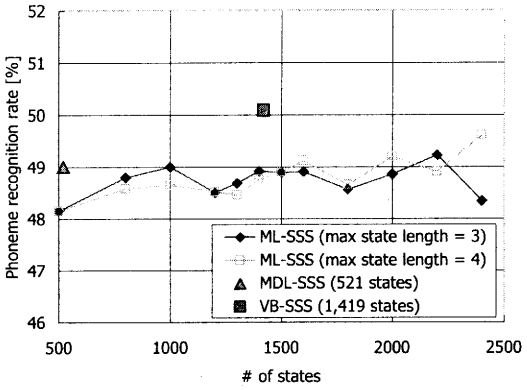

図 3　母音に対する平均音素認識率
Fig. 3　Average of phoneme recognition rates for vowels.

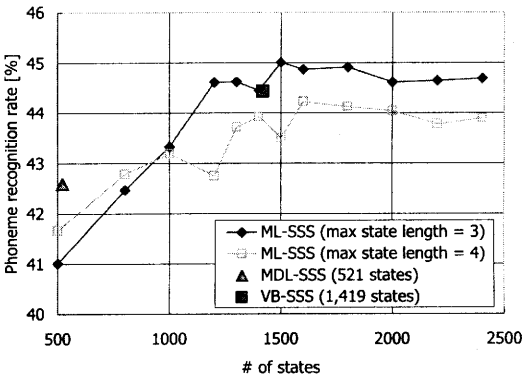

図 4　子音に対する平均音素認識率
Fig. 4　Average of phoneme recognition rates for consonants.

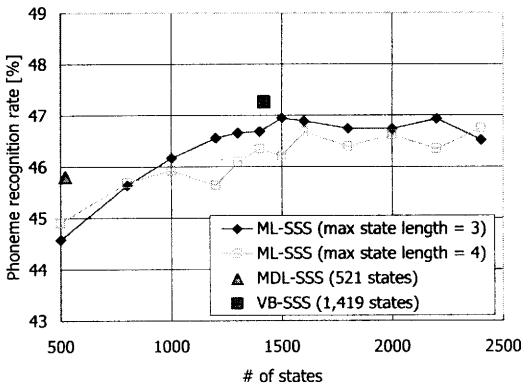

図 5　母音・子音に対する平均音素認識率
Fig. 5　Average of phoneme recognition rates for vowels and consonants.

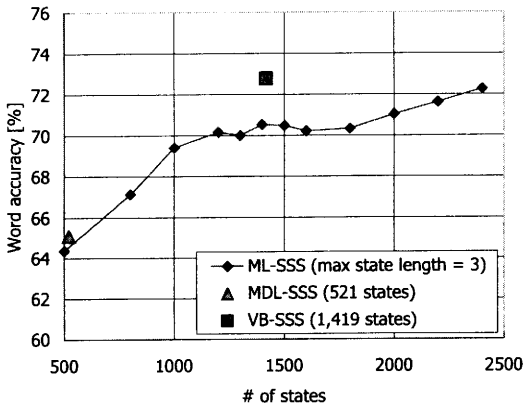

図6　5千単語連続音声認識での単語認識率

Fig. 6　Word accuracy rates for 5k-word continuous speech recognition.

パラメータ数のモデルと比べて1%の向上が見られた．いくつかの子音では精度低下が見られたが，全体的にはほぼ同等の精度が得られた．triphoneで見た場合，分割が進んでいくうちに各triphoneに割り当てられるサンプル数が少量になっても，従来法より比較的ロバストに推定できていると考えられる．これに対し，子音ではもともとデータ量が少ないため，効果が見られないと考えられる．また，ここで，認識率自体が比較的低いが，これは単一分布モデルでの評価であることと，認識対象が対話音声であり，一般に音声認識が難しいタスクであることが原因である．さらに，MDL-SSS法では状態数が小さいモデルが推定され，性能は他のモデルに比べ劣る．これはMDL規準は他の研究で報告されるように，少量学習データでは有効でないことを示していると考えられる．我々の以前の検討[8]では，MDL-SSS法はML-SSS法と同等の性能を得ることができていたが，今回の報告では，学習データ量がより小さく(およそ15分の1)，理論的な仮定から大きく外れ，効果的なパラメータ推定ができなかったと考えられる．

4.3　連続音声認識実験

さらに同じ評価データを用いて語彙5千単語での連続音声認識実験を行った．前節と同じ単一分布モデルで評価した．図6に3種類の方法による単語認識率を示す．提案手法VB-SSSはML-SSS法に対して約60%のモデルサイズで，さらに若干良い性能が得られた．

5. ま と め

時間方向に非均一な音素環境依存HMM構造を自動的に得る方法として，変分ベイズ法を逐次状態分割法へ取り入れた手法を提案した．音素環境方向および時間方向の分割に対する変分ベイズ法による規準を定義した．音素識別実験では，特に母音に対して従来法よりよい性能が得られた．5千単語連続音声認識実験では，モデルサイズが従来法の約60%であるにもかかわらず，若干良い性能が得られた．今後は，事後分布パラメータ推定および目的関数の計算量を減らし，より大量の学習データでの評価を行う予定である．

謝　辞

本研究は通信・放送機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである．

文　献

[1] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modeling," Proc. of the ARPA Workshop on Human Language Technology, pp. 307–312, 1994.

[2] J. Takami, S. Sagayama, "A Successive State Splitting Algorithm for Efficient Allophone Modeling," Proc. ICASSP'92, vol. 1, pp. 573–576, 1992.

[3] K. Shinoda, T. Watanabe, "Acoustic Modeling Based on the MDL Principle for Speech Recognition," Proc. of EUROSPEECH'97, pp. 99–102, 1997.

[4] Koichi Shinoda, Takao Watanabe, "MDL-based context-dependent subword modeling for speech recognition," The Journal of the Acoustical Society of Japan (E), vol. 21, no. 2, pp. 79–86, 2000.

[5] S. S. Chen, R. A. Gopinath, "Model Selection in Acoustic Modeling," Proc. of EUROSPEECH'99, vol. 3, pp. 1087–1090, 1999.

[6] Wu Chou, W. Reichl, "Decision Tree State Tying Based on Penalized Bayesian Information Criterion," Proc. of ICASSP'99, vol. I, pp. 345–348, 1999.

[7] M. Ostendorf, H. Singer, "HMM topology design using maximum likelihood successive state splitting," Computer Speech and Language, vol. 11, pp. 17–41, 1997.

[8] T. Jitsuhiro, T. Matsui, S. Nakamura, "Automatic Generation of Non-Uniform Context-Dependent HMM Topologies Based on the MDL Criterion," Proc. of EUROSPEECH'03, vol. 4, pp. 2721–2724, 2003.

[9] 渡辺澄夫, "データ学習アルゴリズム," データサイエンス・シリーズ6, 共立出版, 2001.

[10] S. R. Waterhouse, D. MacKay, A. J. Robinson, "Bayesian methods for mixture of experts," Advances in Neural Information Processing Systems (NIPS), vol. 8, MIT Press, 1996.

[11] H. Attias, "Inferring parameters and structure of latent variable models by variational Bayes," Proc. of Uncertainty in Artificial Intelligence, 1999.

[12] N. Ueda, Z. Ghahramani, "Optimal model inference for Bayesian mixture of experts," Proc. of IEEE Neural Networks for Signal Processing (NNSP), pp. 145–154, 2000.

[13] S. Watanabe, Y. Minami, A. Nakamura and N. Ueda, "Constructing shared-state HMMs based on a Bayesian approach," Proc. of ICSLP, vol. 4, pp. 2669–2672, 2002.

[14] 上田 修功, "Tutorial Series: ベイズ学習 [III]," 電子情報通信学会誌, vol. 85, no. 7, pp. 504–509, 2002.

[15] T. Takezawa, T. Morimoto, Y. Sagisaka, "Speech and Language Databases for Speech Translation Research in ATR," Proc. of the 1st International Workshop on East-Asian Language Resources and Evaluation (EALREW'98), 1998.

[16] H. Yamamoto, Y. Sagisaka, "Multi-Class Composite N-gram Based on Connection Direction," Proc. of ICASSP'99, vol. 1, pp. 533–536, 1999.