Vol.2013-SLP-97 No.2
2013/7/25

効率的なサンプリング手法を用いた話者モデリング

俵 直弘[1]　小川 哲司[1]　渡部 晋治[2]　中村 篤[3]　小林 哲則[1]

概要：多重スケール混合分布（Multi-scale mixture model）を推定するための効率的なサンプリング手法を提案する．多重スケール混合分布は，混合分布を要素分布として持つ混合モデルで，本稿では，要素分布として混合ガウス分布（Gaussian mixture model: GMM）を導入したモデルを扱う．複数の話者が発話した音声データの集合に対して本モデルを適用した場合，発話のような数十フレーム程度の比較的短いスケールで観測される話者内変動は，各要素 GMM により表現される．一方で，異なる話者の発話間に含まれ，比較的長いスケールで観測される話者間変動は，多重スケール混合分布全体により表現される．このような階層構造を持つ複雑な分布のモデル構造推定問題では，マルコフ連鎖モンテカルロ（Markov chain Monte Carlo: MCMC）法のような確率論的アプローチに基づくモデル推定の枠組みが有効である．しかし，ギブスサンプリングのような単純な MCMC 法をそのまま適用した場合，本来は階層構造を持つべき長時間スケールの構造と短時間スケールの構造が，どちらも対等にサンプリングされるため，繰り返しを含むモデル推定の過程で，容易に局所解に陥ってしまう．そこで，本研究では，blocked ギブスサンプリングに類する手法を導入することで，モデルの階層構造を考慮できるサンプリング手法を提案する．このとき，Iterative conditional modes (ICM) アルゴリズムを導入し，一部のサンプリングプロセスを決定論的な枠組みに置き換えることにより，全ての分布がひとつの分布に縮退してしまう病的な解が選ばれる現象を回避できることを示す．非定常なノイズを重畳した評価セットに対する話者クラスタリング実験により，提案するサンプリング法に基づく構造推定手法が，従来のサンプリング手法や変分ベイズ法に基づく構造推定手法よりも，高い精度でクラスタリング出来ることを示した．

キーワード：フルベイズアプローチ，blocked Gibbs sampling, iterative conditional modes，多重スケール混合分布，話者クラスタリング

1. はじめに

音声データのように，確率過程から生成されるデータ集合をモデリングする場合観測される個々のデータ（例えば，フレーム特徴量）だけを見ても，そのデータが内包する潜在的な情報（例えば音素情報や話者情報）はわからない場合が多い．このようなデータを取り扱うためには，例えば，発話のような複数のデータの集合を 1 つの単位として，適切なモデリングの方法を考える必要がある．

データの単位として，発話区間検出で切り出された発話を用いた場合，これら発話の間には，話者の違いや感情の変化などに起因する変動成分が存在する．本稿では，このような変動を**発話レベル変動**と呼ぶ．一方，同一話者の発話であっても，それら発話中に含まれる音素の分布や，非定常な背景ノイズのような局所的な音響環境の違いに起因する変動成分が存在する．本稿では，このような変動を**フレームレベル変動**と呼ぶ．発話レベル変動成分は，発話単位の，比較的長時間のスケールで観測される一方で，フレームレベル変動成分は．数十ミリ秒単位の，比較的短時間のスケールで観測される．

このように，異なるスケールの変動成分を含むデータをモデリングする場合，データをそれぞれのスケールで適切に表現するために，階層構造を持った確率モデルを考える必要がある．近年，このような階層構造を持つデータのモデリング問題において，潜在変数モデルと呼ばれるモデルが，主に自然言語処理の分野を中心に発展している．例えば，潜在的意味解析（probabilistic latent semantic analysis: pLSA）[1] や，そのフルベイズ拡張である潜在ディリクレ割当て（latent Dirichlet allocation: LDA）[2] が，文書と単語のように複数の異なるスケールに基づいた離散データのモデリングのために提案されている．一方，音声処理の分野においても，連続値を扱うためにガウス分布を生成分布として導入した，多重スケール混合モデルが提案されている [3], [4], [5].

[1] 早稲田大学基幹理工学研究科
[2] Mitsubishi Electric Research Laboratories (MERL)
[3] NTT コミュニケーション科学基礎研究所

このような生成モデルに基づくモデリング問題において，観測データに含まれる外れ値やノイズに対して頑健な推定を行うためには，ベイズ推定が有効である．音声発話を対象としたモデリングにおいても，これまでに多くのベイズ的アプローチが適用されてきた．例えば，事後確率最大化（maximum a posterior: MAP）基準に基づいた手法 [6] や，変分ベイズ（variational Bayesian: VB）基準に基づいたモデリング手法 [7] が，音声認識 [8] や，話者クラスタリング [3], [9] のために提案され，その有効性が示されている．

ここで挙げた従来研究の大半は，期待値最大化（expectation maximization: EM）法のように，決定論的なアプローチに基づく手法であった．対して，我々はこれまでに，マルコフ連鎖モンテカルロ（Markov chain Monte Carlo: MCMC）のような確率的手法に基づくモデル推定手法の実現を模索してきた [4], [5]．文献 [5] では，単純なギブスサンプリング（Gibbs sampling）を，多重スケール混合分布のモデル構造推定に導入することで，特にデータ量が著しく制限された条件下において，従来の VB 法に基づく手法に比べて，より頑健にモデリングが行えることを明らかにした．このアプローチでは，フレームレベル潜在変数と，発話レベル潜在変数のサンプリングを交互に行うことで，真の事後分布からのサンプリングを行ったが，このような実装は，計算コストが比較的小さい一方で，発話レベル潜在変数のサンプリングにおいて，フレームレベル潜在変数のサンプリングで得られた値で強く制限されるため，局所解に容易に陥ってしまうという問題があった．本稿では，このような従来のサンプリング手法において問題となる発話・フレームレベル潜在変数の局所解の問題を，新たなサンプリング手法を提案することにより解決する．

本稿の以降の構成は以下の通りである．**2.** では，多重スケール混合モデルの定式化を行う．**3.** では，従来のギブスサンプリングに基づいてモデル推定を行う手法について説明した後に（**3.1** – **3.3**），提案する新しいサンプリング方法に基づいてモデル推定を行う手法を説明する（**3.4**）**4.** では，話者クラスタリング実験により，提案するサンプリング手法が，特に，非定常なノイズを多く含むデータに対して，従来のサンプリングに基づく手法よりも，頑健にモデルを推定できることを示す．最後に，**5.** では，まとめと今後の課題について述べる．

2. 定式化

本節では，GMM を混合要素分布として持つ多重スケール混合モデルを説明する．

$\mathbf{o}_{ut} \in \mathcal{R}^D$ を u 番目の発話の t フレームの D 次元特徴ベクトル（以下では，フレーム特徴量と呼ぶ）とし，$\mathbf{O}_u \triangleq \{\mathbf{o}_{ut}\}_{t=1}^{T_u}$ を T_u 個のフレーム特徴量から構成される u 番目の発話，$\mathcal{O} \triangleq \{\mathbf{O}_u\}_{u=1}^{U}$ を U 個の発話の集合とする．

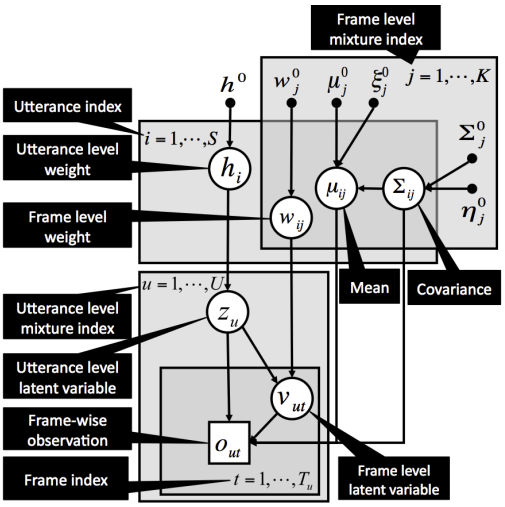


図 1 多重スケール混合モデルのグラフィカル表現．

これら発話を生成するモデルとして，以下のモデルを考える．まず，各発話内のフレーム特徴量系列を，M 個のガウス分布から構成される D 変量 GMM によりモデル化する．さらに，これら GMM を混合要素として S 個持つ混合分布（mixture of GMMs: MoGMMs）を考える．S を話者（クラスタ）数とするならば，この MoGMMs は，それ自身は話者空間全体における話者間変動を表現し，その要素 GMM は，話者内変動を表現していると解釈できる．MoGMMs を多重スケール混合モデルと呼ぶ．

ここで，この階層的混合モデルを解析的に扱うために，2 種類の潜在変数を導入する．まず，$\mathcal{Z} = \{z_u\}_{u=1}^{U}$ は，**発話レベル潜在変数**で，各発話が MoGMM のどの要素（すなわち，どの話者 GMM）に属するかを表す．そして，$\mathcal{V} = \{\{v_{ut}\}_{t=1}^{T_u}\}_{u=1}^{U}$ が，**フレームレベル潜在変数**で，u 番目の発話に含まれる t 番目のフレーム特徴量が，どの混合要素分布に割当てられるのかを表す．これら潜在変数を用いて，本モデルの尤度関数を以下で定義する[*1]．

$$p(\mathcal{O}|\mathcal{Z},\mathcal{V},\mathbf{\Theta}) = \prod_{u=1}^{U} h_{z_u} \prod_{t=1}^{T_u} w_{z_u v_{ut}} \mathcal{N}(\mathbf{o}_{ut}|\boldsymbol{\mu}_{z_u v_{ut}}, \mathbf{\Sigma}_{z_u v_{ut}}) \quad (1)$$

$$P(\mathcal{V}|\mathcal{Z},\mathbf{w}) = \prod_{u=1}^{U} \prod_{t=1}^{T_u} \mathcal{M}(v_{ut}|\mathbf{w}_{z_u}) \quad (2)$$

$$P(\mathcal{Z}|\mathbf{h}) = \prod_{u=1}^{U} \mathcal{M}(z_u|\mathbf{h}) \quad (3)$$

ただし，$\mathcal{M}(\cdot|\mathbf{w}_{z_u})$，$\mathcal{M}(\cdot|\mathbf{h})$ は，それぞれ，$\mathbf{w}_{z_u} = \{w_{z_u 1}, \cdots, w_{z_u M}\}$，$\mathbf{h} = \{h_1, \cdots, h_S\}$ をパラメタとする多項分布とし，$\mathcal{N}(\cdot|\boldsymbol{\mu}_{z_u v_{ut}}, \mathbf{\Sigma}_{z_u v_{ut}})$ は，平均ベクトル $\boldsymbol{\mu}_{z_u v_{ut}} \in \mathbb{R}^D$ と分散共分散行列 $\mathbf{\Sigma}_{z_u v_{ut}} \in \mathbb{R}^{D \times D}$ をパラメタとするガウス分布とする．ただし，本稿では，$\mathbf{\Sigma}_{ij}$ は，d 行 d 列目の要素が $\sigma_{ij,d}$ で表される対角共分散行列とする．

本モデルに対するフルベイズ的な扱いを可能にするため

[*1] $p(\cdot)$ は連続値上の確率密度関数を表すとし，$P(\cdot)$ は離散値上の確率密度関数を表すとする．

に，モデルパラメタ $\boldsymbol{\Theta} \triangleq \left\{h_i, \left\{w_{ij}, \boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij}\right\}_{j=1}^{M}\right\}_{i=1}^{S}$ について，以下の共役事前分布を導入する．

$$p(\boldsymbol{\Theta}|\boldsymbol{\Theta}^0) = \begin{cases} \mathbf{h} & \sim \mathcal{D}(\mathbf{h}^0) \\ \mathbf{w}_i & \sim \mathcal{D}(\mathbf{w}^0) \\ \{\mu_{ij,d}, \sigma_{ij,d}\} & \sim \mathcal{NG}(\xi^0, \eta^0, \mu^0_{j,d}, \sigma^0_{j,d}) \end{cases} \quad (4)$$

ただし，$\mathcal{D}\left(\mathbf{h}^0\right)$ と $\mathcal{D}\left(\mathbf{w}^0\right)$ は，それぞれ，$\mathbf{h}^0$ と $\mathbf{w}^0$ をハイパーパラメタとするディリクレ分布を表し，$\mathcal{NG}\left(\xi^0, \eta^0, \mu^0_{j,d}, \sigma^0_{j,d}\right)$ は，ξ^0，η^0，$\mu^0_{j,d}$，$\sigma^0_{j,d}$ をハイパーパラメタとするガウス–ガンマ分布を表す．図 1 に本モデルのグラフィカル表現を示す．

3. モデル推定

前節で定義した混合ガウス分布に基づく多重スケール混合分布について，ベイズ推定に基づいて，その構造を推定する方法を述べる．ここで中心となるタスクは，観測データ $\mathcal{O}$ が与えられたときの，潜在変数 $\{\mathcal{V}, \mathcal{Z}\}$ とモデルパラメタ $\boldsymbol{\Theta}$ の事後分布

$$p(\mathcal{V}, \mathcal{Z}, \boldsymbol{\Theta}|\mathcal{O}) = \frac{1}{H_p} p(\mathcal{O}, \mathcal{V}, \mathcal{Z}, \boldsymbol{\Theta}) \quad (5)$$

を推定することである．ただし，H_p は，事後分布の $\{\mathcal{V}, \mathcal{Z}, \boldsymbol{\Theta}\}$ に関する積分が 1 になるように定めた定数で，以下で定義される．

$$H_p \triangleq p(\mathcal{O}) = \sum_{\mathcal{V},\mathcal{Z}} \int p(\mathcal{O}, \mathcal{V}, \mathcal{Z}, \boldsymbol{\Theta}) \mathrm{d}\boldsymbol{\Theta} \quad (6)$$

話者クラスタリングに問題を限定した場合，各発話を最適な話者クラスタに割当てる問題は，(5) で定義された事後分布を用いて，発話レベル潜在変数 $\mathcal{Z}$ の最適な実現値を推定する問題に帰着される．しかし，この事後分布を解析的に求めることは一般に困難であるため，何らかの近似が必要となる．

文献 [3] では，本モデルとほぼ等価な構造を持つモデルについて，VB 法に基づいて最適な事後分布を近似的に推定する手法が提案されている．しかし，VB 法では全ての未知変数について事後分布を推定する必要があるため，推定すべきパラメタ数が多く，特に観測データ $\mathcal{O}$ の数が制限されている場合において，過学習の問題が指摘されている [5]．

このような過学習の問題は，以下のように，事後分布 (5) についてモデルパラメタ $\boldsymbol{\Theta}$ に関する周辺化を行った周辺化事後分布を考えることで回避できる．

$$P(\mathcal{V}, \mathcal{Z}|\mathcal{O}) = \frac{1}{H_p} \int p(\mathcal{O}, \mathcal{V}, \mathcal{Z}, \boldsymbol{\Theta}) \mathrm{d}\boldsymbol{\Theta} \quad (7)$$

このように，モデルパラメタを事前に周辺化消去した場合，モデルパラメタの事後分布を推定せずに，潜在変数の事後分布を直接推定できるため，観測されたデータが少ない場合においても，頑健に事後分布を推定できる．しかし，VB 法の枠組みの中で，このような周辺化を行うためには，対数関数の凸性に関する近似を仮定しなければならず [10]，特に，本モデルのように階層的なモデル構造に関しては適用が困難である．そこで，VB 法の代わりに MCMC 法を導入することで，周辺化された事後分布 (7) を推定する手法が提案されている [4], [5]．

3.1 MCMC 法に基づくモデル推定

MCMC 法に基づくアプローチでは，未知の確率変数の事後分布を直接推定する代わりに，これら確率変数の実現値をその事後分布から直接サンプリングする．このとき，MCMC に基づくアプローチでは，VB 法とは異なり，正規化項 (6) の具体的な値を評価する必要がないため，モデルパラメタに関する周辺化 (7) を容易に行うことができる．

多重スケール混合モデルにおける MCMC 法を構築するため，完全データの対数尤度関数として，以下の関数を定義する．

$$\begin{aligned} H(\boldsymbol{\Psi}) &\triangleq -\log p_\beta(\mathcal{O}, \mathcal{V}, \mathcal{Z}) \\ &= -\log p(\mathcal{O}|\mathcal{V}, \mathcal{Z}) - \frac{1}{\beta} \log P(\mathcal{V}, \mathcal{Z}) \end{aligned} \quad (8)$$

ただし，β は逆温度と呼ばれ，後述する焼きなまし法 (simulated annealing: SA) のために導入する．この変数は，MCMC 法の収束の速さを決定する．このとき，事後分布は，

$$\begin{aligned} P(\mathcal{V}, \mathcal{Z}|\mathcal{O}) &= \frac{1}{H_p(\beta)} p(\mathcal{V}, \mathcal{Z}) p(\mathcal{O}|\mathcal{V}, \mathcal{Z})^\beta \\ &= \frac{1}{H_p(\beta)} \exp\left\{-\beta H(\boldsymbol{\Psi})\right\} \end{aligned} \quad (9)$$

と書ける．ただし，$H_p(\beta)$ は $\{\mathcal{V}, \mathcal{Z}\}$ に関する積分を 1 にするための正規化項である．MCMC 法では，発話およびフレームレベル潜在変数に関する適当な初期状態 $\boldsymbol{\Psi}^0 \triangleq \left\{\mathcal{V}^0, \mathcal{Z}^0\right\}$ を設定し，遷移確率 $q(\boldsymbol{\Psi}^{t+1}|\boldsymbol{\Psi}^t)$ に基づいて，現状態 $\boldsymbol{\Psi}^t$ を条件とした次状態 $\boldsymbol{\Psi}^{t+1}$ の条件付き分布からサンプリングを繰り返すことにより，目的の事後分布 (9) からのサンプリング値 $\boldsymbol{\Psi}$ を取得する．

3.2 完全データの周辺化対数尤度

MCMC 法の具体的なアルゴリズムを構築するために必要な完全データの周辺化対数尤度関数 $\log p(\mathcal{O}, \mathcal{V}, \mathcal{Z})$ を導出する．完全データを扱う場合，全ての潜在変数 $\{\mathcal{V}, \mathcal{Z}\}$ の値が観測変数 $\mathcal{O}$ の値と共に与えられるため，各発話に含まれる全てのフレーム特徴量について，どの話者 GMM のどの要素ガウス分布に対して割当てられるのかがわかる．すなわち，各潜在変数の事後確率 $(P(z_u = i|\cdot), P(v_{ut} = j|\cdot), \forall i, j, u, t)$ は，対応するフレーム特徴量の割当てに従って，それぞれ 0 または 1 の値をとるため，本モデルの十分統計量は以下の形で書くことができる．

$$\begin{cases} c_i & = \sum_u \delta(z_u, i), \\ n_{ij} & = \sum_{u,t} \delta(z_u, i) \cdot \delta(v_{ut}, j), \\ \mathbf{m}_{ij} & = \sum_{u,t} \delta(z_u, i) \cdot \delta(v_{ut}, j) \cdot \mathbf{o}_{ut}, \\ r_{ij,d} & = \sum_{u,t} \delta(z_u, i) \cdot \delta(v_{ut}, j) \cdot (o_{ut,d})^2 \end{cases} \tag{10}$$

ただし，$\delta(a,b)$ はクロネッカーのデルタ関数で，$a=b$ のときのみ 1 となり，それ以外では 0 をとる関数である．c_i は MoGMM の i 番目の話者 GMM に割当てられた発話数，n_{ij} は MoGMM の i 番目の話者 GMM 内の j 番目の要素ガウス分布に割当てられたフレーム数に相当し，$\mathbf{m}_{ij}$，r_{ij} は，それぞれ一次，二次十分統計量に相当する．

これら十分統計量と，尤度関数 (1) – (3)，モデルパラメタの事前分布 (4) を用いることで，完全データの周辺化尤度関数は以下のように解析的に導出できる．

$$\begin{aligned} &\log p(\mathcal{O}, \mathcal{V}, \mathcal{Z}) \\ &= \log \int p(\mathcal{O}, \mathcal{V}, \mathcal{Z}|\mathbf{\Theta}) p(\mathbf{\Theta}) \mathrm{d}\mathbf{\Theta} \\ &= \log \frac{\Gamma(h^0) \prod_i \Gamma(\tilde{h}_i)}{\Gamma(h^0)^S \Gamma(\sum_i \tilde{h}_i)} + \log \prod_i \frac{\Gamma(\sum_j w_j^0) \prod_j \Gamma(\tilde{w}_{ij})}{\prod_j \Gamma(w_j^0) \Gamma(\sum_j \tilde{w}_{ij})} \\ &+ \beta \log \prod_{i,j} (2\pi)^{-\frac{n_{ij} D}{2}} \frac{(\xi^0)^{\frac{D}{2}} \left(\Gamma\left(\frac{\eta_j^0}{2}\right) \right)^{-D} \left(\prod_d \sigma_{j,dd}^0 \right)^{\frac{\eta_j^0}{2}}}{(\tilde{\xi}_{ij})^{\frac{D}{2}} \left(\Gamma\left(\frac{\tilde{\eta}_{ij}}{2}\right) \right)^{-D} \left(\prod_d \tilde{\sigma}_{ij,dd} \right)^{\frac{\tilde{\eta}_{ij}}{2}}} \end{aligned} \tag{11}$$

ただし，$\tilde{\mathbf{\Theta}}_{ij} \triangleq \left\{ \tilde{h}_i, \tilde{w}_{ij}, \tilde{\xi}_{ij}, \tilde{\eta}_{ij}, \tilde{\mu}_{ij,d}, \tilde{\sigma}_{ij,d} \right\}$ は，周辺化尤度関数のハイパーパラメタで，次式で定義した．

$$\begin{cases} \tilde{h}_i & = h^0 + c_i, \\ \tilde{w}_{ij} & = w_j^0 + n_{ij}, \\ \tilde{\xi}_{ij} & = \xi^0 + n_{ij}, \\ \tilde{\eta}_{ij} & = \eta^0 + n_{ij}, \\ \tilde{\boldsymbol{\mu}}_{ij} & = \tilde{\xi}_{ij}^{-1} \left(\xi^0 \boldsymbol{\mu}_j^0 + \mathbf{m}_{ij} \right), \\ \tilde{\sigma}_{ij,d} & = \sigma_{j,d}^0 + r_{ij,d} + \xi^0 (\mu_{j,d}^0)^2 - \tilde{\xi}_{ij} (\tilde{\mu}_{ij,d})^2 \end{cases} \tag{12}$$

3.3 (Collapsed) ギブスサンプリング

[4], [5] では, 多重スケール混合モデルの構造を推定するために，代表的な MCMC 法の一つである (collapsed) [*1] ギブスサンプリング [11] を導入した．ギブスサンプリングの各ステップでは，任意の潜在変数を 1 つ選択し，それ以外の全ての潜在変数の値を現状態の値 $\mathbf{\Psi}^t$ に固定して，これを条件とした条件付き分布を，そのステップにおける遷移確率 $q(\mathbf{\Psi}^{t+1}|\mathbf{\Psi}^t)$ として用いる．例えば， z_u をサンプリング対象として選択した場合，$\mathcal{Z}$ から z_u を除いた集合を $\mathcal{Z}_{\backslash u}$ ($\mathcal{Z}_{\backslash u} = \{z_{u'} | \forall u' \neq u\}$) としたとき，$z_u$ の実現値として，条件付き事後分布 $p(z_u|\mathcal{O}, \mathcal{Z}_{\backslash u})$ からサンプルを 1 つ抽出し，これを次状態の実現値とする．以上の手続きを全変数 $\mathcal{Z}$ について繰り返し行うことで，目的の事後分布 $P(\mathcal{Z}|\mathcal{O})$ からのサンプルを抽出する．

[*1] *collapsed* は，潜在変数のサンプルはモデルパラメタ $\mathbf{\Theta}$ に関する周辺化を行った周辺化事後分布から得られることを示している．以下，ギブスサンプリングと記載した場合は Collapsed ギブスサンプリングを指す．

Algorithm 1 従来のギブスサンプリングと SA 法に基づくモデル推定手法のアルゴリズム．

```
Initialize {z_u, v_ut : u = 1, ··· , U, t = 1, ··· , T_u}, β = β_init.
repeat
  for all utterances u and frames t do
    for all components j do
      Compute γ_{v_ut=j|z_u=i}(β) by Eq. (13).
    end for
    Draw frame-level latent variable (fLV), v*_ut, from its conditional posterior distribution, M( γ_{v_ut=·|z_u=i}(β) / Σ_j γ_{v_ut=j|z_u=i}(β) ).
  end for
  for all utterances u do
    for all speakers i do
      Compute γ_{z_u=i}(β) by Eq. (14)
    end for
    Draw utterance-level latent variable (uLV), z*_u, from its conditional posterior distribution, M( γ_{z_u=·}(β) / Σ_i γ_{z_u=i}(β) ).
  end for
  Update β with SA scheduler.
until some condition is met
```

多重スケール混合モデルの場合，フレーム・発話レベル潜在変数は，それぞれ，以下の条件付き事後分布に基づいて，各変数のサンプリングを順次行う．

[Frame-level latent variables]

$$\begin{aligned} &p(v_{ut} = j' | \mathcal{O}, \mathcal{V}_{\backslash t}, \mathcal{Z}_{\backslash u}, z_u = i) \\ &\propto \frac{p(\mathcal{O}, \mathcal{V}_{\backslash t}, v_{ut} = j', \mathcal{Z}_{\backslash u}, z_u = i)}{p(\mathcal{O}_{\backslash t}, \mathcal{V}_{\backslash t}, \mathcal{Z}_{\backslash u}, z_u = i)} \\ &\propto \exp\left\{ -\beta \left(H(\tilde{\mathbf{\Psi}}_{i,j'}) - H(\tilde{\mathbf{\Psi}}_{i,j'\backslash t}) \right) \right\} \\ &\triangleq \gamma_{v_{ut}=j'|z_u=i}(\beta) \end{aligned} \tag{13}$$

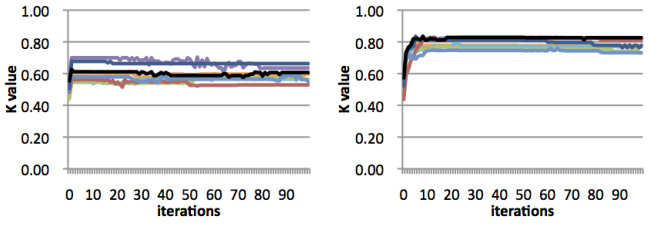
[Utterance-level latent variable]

$$\begin{aligned} &p(z_u = i' | \mathcal{O}, \mathcal{V}, \mathcal{Z}_{\backslash u}) \\ &\propto \frac{p(\mathcal{O}, \mathcal{V}, \mathcal{Z}_{\backslash u}, z_u = i')}{p(\mathcal{O}_{\backslash u}, \mathcal{V}_{\backslash u}, \mathcal{Z}_{\backslash u})} \\ &\propto \exp\left\{ \log \frac{\Gamma(\sum_j \tilde{w}_{i'\backslash u,j})}{\Gamma(\sum_j \tilde{w}_{i',j})} - \beta \sum\nolimits_j \left(H(\tilde{\mathbf{\Psi}}_{i',j}) - H(\tilde{\mathbf{\Psi}}_{i'\backslash u,j}) \right) \right\} \\ &\triangleq \gamma_{z_u=i'}(\beta) \end{aligned} \tag{14}$$

ただし，$H(\tilde{\mathbf{\Psi}}_{i,j})$ は，$\{\mathcal{O}, \mathcal{Z}, \mathcal{V}\}$ に関する完全データ対数尤度を表し，以下で定義した．

$$\begin{aligned} H(\tilde{\mathbf{\Psi}}_{i,j}) &\triangleq \log p(\mathcal{O}, \mathcal{V}_{\backslash ut}, v_{ut} = j, \mathcal{Z}_{\backslash u}, z_u = i) \\ &\propto \log \Gamma(\tilde{w}_{ij}) - \frac{D}{2} \log \tilde{\xi}_{ij} \\ &\quad + D \log \Gamma\left(\frac{\tilde{\eta}_{ij}}{2} \right) - \frac{\tilde{\eta}_{ij}}{2} \sum\nolimits_d \log \tilde{\sigma}_{ij,d} \end{aligned} \tag{15}$$

$\tilde{h}_i$，$\tilde{w}_{ij}$，$\tilde{\xi}_{ij}$，$\tilde{\eta}_{ij}$，$\tilde{\boldsymbol{\mu}}_{ij}$，$\tilde{\sigma}_{ij,d}$ は，周辺化尤度関数のハイパーパラメタで，(12) で定義した．同様に，(13)，(14) 中の $H(\tilde{\mathbf{\Psi}}_{i,j\backslash t})$，$H(\tilde{\mathbf{\Psi}}_{i\backslash u,j})$ はそれぞれ，$\left\{\mathcal{O}_{\backslash t}, \mathcal{Z}, \mathcal{V}_{\backslash t}\right\}$ と $\left\{\mathcal{O}_{\backslash u}, \mathcal{Z}_{\backslash u}, \mathcal{V}_{\backslash u}\right\}$ に関する完全データ対数尤度を表す．多



(a) $\beta^{init} = 1$　　(b) $\beta^{init} = 50$

図 2 ギブスサンプリングにおける SA の初期温度と K 値の関係．各図が，それぞれ，初期温度 β^{init} が，(a) $\beta^{init} = 1$，(b) $\beta^{init} = 50$ のときの結果に対応する．8 本の直線が，それぞれ，8 回の試行における K 値を示す．

重スケール混合モデルにおけるギブスサンプリングでは，まず，全フレームレベル潜在変数のサンプリングを行った後に，全発話レベル潜在変数のサンプリングを行う．

3.3.1 焼きなまし法

本研究では，焼きなまし（Simulated annealing: SA）法 [12] で局所解の低減を図る．例えば，u 番目の発話の第 t フレームが j 番目の混合要素に割当てられている状態を現状態とする．このとき，このフレームに対応するフレームレベル潜在変数 v_{ut} のサンプリングにおいて，現在の割り当て j を選択したときの事後確率 $\gamma_{v_{ut}=j|z_u=i}$ が，それ以外の割当て $j' \neq j$ を選択したときの事後確率 $\gamma_{v_{ut}=j'|z_u=i}$ よりも著しく高い値をとる場合，現在の状態から他の状態への有効な遷移がほとんど発生しない．これは，ギブスサンプリングの各ステップでは，1 変数の更新しか行われないため，ある潜在変数の状態 $\boldsymbol{\Psi}$ の事後確率が，その周辺の組み合わせの事後確率よりも高い場合，他の状態に遷移するためには，より低い事後確率をとる組み合わせを経由しなければ，他の状態へと遷移できず，状態間の遷移が非常に起こり難くなることを示している．この問題は，一般に，ポテンシャル障壁に関する問題として知られており，回避するためには，温度 β を高い値に設定する必要がある．すなわち，温度 β を高く設定した場合，$\gamma_{v_{ut}=j'|z_u=i}$ は全ての組み合わせ j' についてほぼ一様な値となり，その結果，サンプリング系列はランダムウォークに近い挙動をとり，探索空間を大きく移動できるため，このような局所解を抜け出すことが可能になる．しかし，一方で，高い温度の下ではサンプリング系列が真の事後分布に収束することは保証されず，得られるサンプリング系列が不安定になってしまう．そのため，SA 法では初期温度として比較的高い温度を設定し，イタレーション毎に，特定の冷却スケジュールに従って，温度を徐々に下げることにより，局所解の回避と収束性を同時に保証する．本手法では，t 回目のイタレーションについて，

$$\beta^{t+1} \leftarrow \begin{cases} \gamma\beta^t, & if\ \beta^t > 1 \\ 1, & otherwise \end{cases} \tag{16}$$

とする等比級数スケジューリングを用いた．ただし，γ は $0 < \gamma < 1$ を満たす定数とする．

Algorithm 2 提案する blocked ギブスサンプリングと ICM 近似に基づくモデル推定手法のアルゴリズム

```
Initialize {z_u, v_ut : u = 1, ··· , U, t = 1, ··· , T_u}.
repeat
  for all utterances u do
    for all speakers i do
      for all frames t do
        for all components j do
          Compute γ_{v_ut=j|z_u=i}(1) by Eq. (13).
        end for
        Decide the value of fLV, v*_ut, from its posterior
        probability by v*_ut = arg max_j γ_{v_ut=j|z*_u}(1)
      end for
      Compute γ_{z_u=i|V*}(1) by Eq. (14) conditioning on the
      sampled fLVs, {v*_ut}_{t=1}^{T_u}.
    end for
  end for
  for all utterances u do
    Draw the value of uLV, z*_u, from its posterior distribu-
    tion by z*_u ~ M( γ_{z_u=·|V*}(1) / Σ_i γ_{z_u=i|V*}(1) )
  end for
until some condition is met
```

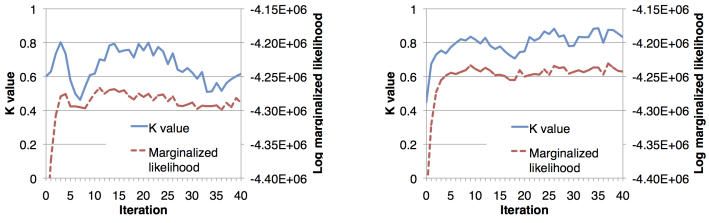
図 2 に初期温度 β^{init} を，それぞれ，1，50 として，初期値と乱数のシードを変えた 8 回の試行について，100 回繰り返したときの，各イタレーションで得られるサンプリング結果を，K 値で評価した結果を示す．ただし，評価セットとして，後述する B1 + noise 2 を用い，各話者 GMM の混合数は 8 とした．この図から，初期温度を $\beta^{init} = 1$（すわなち，SA を行わない）とした場合，全ての試行において比較的早くに収束し，このときの K 値はいずれも低く，また，試行ごとに収束値が大きく異なることから，多くの試行において局所解に陥っていることがわかる．一方，初期温度を $\beta^{init} = 50$ とした場合，SA を行わなかった場合に比べ，結果のばらつきが少なくなり，かつ，高い K 値が収束値として得られていることがわかる．以上の結果は，従来のギブスサンプリングを本モデルに適用するためには，局所解を回避するために SA 法が必須であることを示している．

以上の手続きに基づいて，多重スケール混合分布の潜在変数を推定するアルゴリズムを Algorithm 1 に示す．

3.4 Blocked ギブスサンプリング

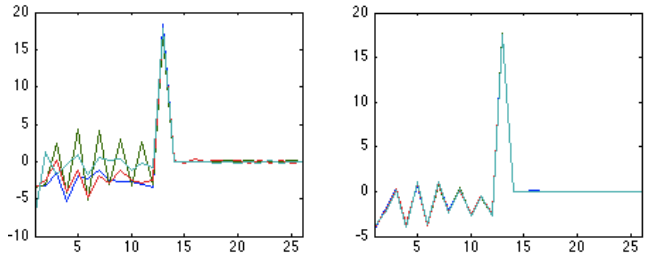
3.1 では，ギブスサンプリングに基づいて多重スケール混合モデルを推定する手法を説明し，この時発生するポテンシャル障壁に関する問題を，SA 法を用いて解決できることを示した．しかし，多重スケール混合モデルのような階層構造を持つ分布を MCMC 法により推定する場合，未解決の問題が更にもう一つ存在する．

Algorithm 1 で示した従来のギブスサンプリングに基づく手法では，全フレームに対してフレームレベル潜在変数



(a) (b)

図 3 *(a) blocked* ギブスサンプリング，*(b) iterated conditional modes* (*ICM*) により得られる K 値と周辺化対数尤度．



(a) (b)

図 4 話者 GMM に含まれる 4 つのガウス分布に割当てられた 26 次元の音響特徴量（MFCCs）の平均ベクトル．それぞれ，(a) 5 回目のイタレーション，(b) 6 回目のイタレーション時の状態を示す．各直線が各ガウス分布に対応する．

のサンプリングを行った後に，発話レベル潜在変数のサンプリングを行う．このとき，発話レベル潜在変数は，前のステップでサンプリングされたフレームレベル潜在変数の値を条件とした事後分布からサンプリングされる．これはすなわち，(14) で表される発話レベル潜在変数の事後分布を評価する際に，全話者 GMM に対して，同一のフレームレベル潜在変数の値に基づいて，評価されることに相当する．このような強い制約は，特に混合数が多い複雑な分布において，解を収束させるまでの時間が膨大になる．すなわち，より多くのフレーム・発話レベル潜在変数の組み合わせを評価するためには，たくさんのサンプルが必要となり，それゆえに，得られるサンプル系列が真の事後分布に収束するまでに相当にたくさんのサンプリングが必要となる．これは，一般に，エントロピー障壁に関する問題として知られる問題であり，サンプリングの効率が著しく低下する要因となる．3.3.1 で述べたように，SA 法を導入し，高い初期温度を設定することで，このような問題はある程度緩和できると考えられるが，ランダムウォークに近い挙動を長く続けるサンプリング系列が真の分布へと収束するためには，より多くのイタレーションが必要になる．

この問題を解決するために，各イタレーションにおいて，より多くの仮説を評価できる新しいサンプリング法を提案する．提案する手法では，u 番目の発話に対応した発話レベル潜在変数をサンプリングする際に，この発話に対応した発話・フレームレベル潜在変数 $\{z_u, \mathcal{V}_u\}$ を，その同時事後分布 $P(z_u, \mathcal{V}_u | \mathcal{Z}_{\backslash u}, \mathcal{V}_{\backslash u}, \mathcal{O})$ からまとめてサンプリングする．このとき，この同時事後分布は以下の形に分解できる．

$$P(z_u, \mathcal{V}_u | \mathcal{Z}_{\backslash u}, \mathcal{V}_{\backslash u}, \mathcal{O}) = P(z_u | \mathcal{Z}_{\backslash u}, \mathcal{V}_u, \mathcal{V}_{\backslash u}, \mathcal{O}) P(\mathcal{V}_u | z_u, \mathcal{Z}_{\backslash u}, \mathcal{V}_{\backslash u}, \mathcal{O}) \quad (17)$$

そこで，伝承サンプリング [13] の考え方を導入することによりこの分解された分布からサンプルを抽出する．伝承サンプリングでは，まず，式 (17) の右辺第二項からフレームレベル潜在変数をサンプリングするために，新たなギブスサンプリングを導入し，以下のサンプルを抽出する．

$$\mathcal{V}_u^* \sim P(\mathcal{V}_u^* | z_u, \mathcal{Z}_{\backslash u}, \mathcal{V}_{\backslash u}, \mathcal{O}) \quad (18)$$

この新たなギブスサンプリングは，式 (13) で既に定義したように，各フレームレベル潜在変数に関して，それ以外の全フレームレベル潜在変数を条件とした時の事後確率を評価することにより構築することができ，得られたサンプル系列は以下の関係に従うことが保証される．

$$z_u \sim P(z_u | \mathcal{Z}_{\backslash u}, \mathcal{V}_u^*, \mathcal{V}_{\backslash u}, \mathcal{O}) \quad (19)$$

この新しいサンプリング手法と，従来のサンプリング法で最も大きく異なる点は，従来のサンプリング法では発話レベル潜在変数とフレームレベル潜在変数がそれぞれ交互にサンプリングされているのに対し，提案するサンプリング法ではこれらを同時にサンプリングするという点である．このように，提案法では複数の潜在変数が同時にサンプリングされるため，これを blocked ギブスサンプリング [14] の一種とみなすことができる．

3.4.1 Iterated conditional modes (ICM) アルゴリズム

図 3 (a) に，日本語話し言葉コーパス（Corpus of Spontaneous Japanese; CSJ) から作成した評価セットに対し，blocked ギブスサンプリングを適用し，得られた発話レベル潜在変数を，一般的なクラスタリング尺度である K 値で評価した結果を示す．また，同じ図に，そのときの周辺化対数尤度 (8) を示した．この図から，K 値が 5 回目のイタレーションで急激に低下し，その後一時的に上昇するが，すぐに再び低下していることがわかる．この理由を明らかにするために，図 4 に，話者 GMM に含まれる 4 つのガウス分布に割当てられた 26 次元の音響特徴量の平均ベクトルを示す．それぞれ，(a) 5 回目のイタレーション，(b) 6 回目のイタレーション時の状態である．この例では，話者 GMM の混合数は 4 とした．この図から，全 GMM の全ての混合要素分布が同じ分布に縮退していることがわかる．しかし，評価セットとして与えたデータは多峰性の分布に従うデータであるため，本来このような単峰の分布が推定されるべきではない．このような病的な解が得られた理由は，以下の様なサンプリングの性質によるものだと考えられる．すなわち，複数の要素分布が互いに同じようなパラメタを持っているとき，確率的な割当ての決定過程は，

フレーム特徴量をほぼランダムに割当てる．このような例は，特にフレームレベル潜在変数のサンプリング時に発生する場合が多い．なぜならば，各話者は，特徴量空間において，比較的小さな分散を持っていると考えられ，このため，本来は複数の分布に分割されるべき分布が，しばしば単一の分布に同一視されてしまったと考えられる．この問題は，ギブスサンプリングにおいて大量のサンプリングを繰り返すことにより回避できる可能性があるが，計算量の観点からそのような解決法は現実的ではない．そこで，フレームレベル潜在変数のサンプリングにおいてのみ，温度 $\beta \to 0$ とした極限を設定した．これは，フレームレベル潜在変数をその条件付き事後分布からサンプリングする代わりに，条件付き事後確率が最大となる分布に割り当てることに相当する．このとき，u 番目の発話の t 番目のフレーム特徴量に対応したフレームレベル潜在変数の値は，以下の式により更新される．

$$v_{ut}^* = \arg\max_j \gamma_{v_{ut}=j} | z_u^* \tag{20}$$

このような考え方は，iterated conditional modes（ICM）アルゴリズム [15] として知られる決定論的な手法に基づいたフレームレベル潜在変数の割り当てに相当する．ICM に基づいた近似では，事後確率がほとんど同じ値をとるフレームレベル潜在変数のペアに対しても，異なる混合要素分布に割り振ることができる．図 3 (b) に，フレームレベル潜在変数の推定にのみ ICM を適用した結果を示す．この結果が示すように，ICM を用いた場合では，複数の峰を持つ分布が正しく推定されており．分布の縮退を回避するために ICM が有効であることがわかる．提案する ICM 近似に基づいた blocked ギブスサンプリングのアルゴリズムを Algorithm 2 に示す．

4. 評価実験

4.1 実験条件

TIMIT データベースと日本語話し言葉コーパス（CSJ）から作成した評価セットに対する話者クラスタリング実験により，以下の三つのモデル推定手法を比較した．

- **b-Gibbs (proposed):** 提案する blocked ギブスサンプリングを用いた MCMC 法に基づくモデル推定
- **Gibbs:** 従来のギブスサンプリングを用いた MCMC 法に基づくモデル推定手法 [4], [5]
- **VB:** VB 法に基づいたモデル推定手法 [3]

すべての実験は，TIMIT および CSJ から作成した 8 種類の評価セットに対して行った．TIMIT からは core test set（以下，T1）と，complete test set から T1 に含まれるデータを除いたデータ（T2）を用いて評価セットを作成した．T1 には，24 人が発話した 192 発話が含まれ，T2 には，T1 とは重複しない 144 人が発話した 1,152 発話が含まれる．CSJ からは，以下の手順で，6 種類の評価セットを作成した．まず、コーパスに含まれる全講演に対して，無音区間 500 ms 以上を基準として発話単位に区切ったとき，発話長が 5 秒以上 10 秒以下の発話を抽出した．次に，5 話者をランダムに選択し，各人のランダムに選択された 5，10，20 発話を含む評価セットをそれぞれ A1，A2，A3 とした．同様に，先に選択した話者とは異なる 10 話者をランダムに選択し，各人のランダムに選択された 5，10，20 発話を含む評価セットをそれぞれ B1，B2，B3 とした．このとき，各評価セット毎に，選択する話者を変えた 5 種類の組み合わせを用意した．各評価セットに対する結果は，この 5 種類の組み合わせに対する結果の平均値である．ノイズを含むデータに対する各手法の頑健性を評価するため，CSJ データセットに対し，電子協騒音データベース付属の (10) 人混み，および (9) 幹線道路，交差点騒音を SNR 10 dB で重畳した．これらノイズをそれぞれ noise 1，noise 2 とする．特徴量は，音響特徴量 MFCC（12 次元）に対数エネルギーと Δ パラメタを加えた計 26 次元である．フレーム長は 25 ms，フレーム周期は 10 ms とした．

表 1 評価セットの詳細.

Test set	number of speakers	number of utterances	average total duration [min.]
T1	24	192	9.7
T2	144	1152	58.8
A1	5	25	2.8
A2	5	50	5.6
A3	5	100	11.1
B1	10	50	5.6
B2	10	100	11.3
B3	10	200	22.5

評価尺度には，一般的なクラスタリング尺度の一つである K 値を用いた．K 値は，平均話者純度と平均クラスタ純度の幾何平均として定義される [16]．各評価セットごとに，初期値と乱数のシードを変えて同じ実験を 8 回繰り返し，周辺化対数尤度（8）が最大となる結果を選択した．(12) で定義したハイパーパラメタは，以下のように定めた．$w^0 = 1$ とし，全混合要素について，$\mathbf{w}^{(0)} = \{w^0, \cdots, w^0\}$ とした．同様に，$h^0 = 1$ とし，全話者 GMM について，$\mathbf{h}^{(0)} = \{h^0, \cdots, h^0\}$ とした．また，$\eta^{(0)} = 1$，$\xi^{(0)} = 1$ とした．$\boldsymbol{\mu}^{(0)}$ と $\boldsymbol{\Sigma}^{(0)}$ については，それぞれ評価セットに含まれる全発話の平均ベクトルと分散共分散行列とした．混合要素数は，いずれの手法も，T1 と T2 については 4 とし，残りの評価セットについては全て 8 とした．

4.2 実験結果

表 2 に，クリーンな評価セットに対して 3 手法を適用して得られる K 値を示す．ここで，CSJ から作成した評価セット（A1 から B2）は比較的単鋒な分布に従う一方で，TIMIT から作成した評価セット（T1 と T2）は比較的多

表 2 クリーン評価セットに対する実験結果（K 値）

Evaluation data	b-Gibbs	Gibbs	VB
T1 (spkr:24 utt:192)	**0.87**	0.81	0.71
T2 (spkr:144 utt:1152)	**0.74**	0.52	0.41
A1 (spkr:5 utt:25)	**0.99**	0.92	0.88
A2 (spkr:5 utt:50)	**0.99**	0.91	0.95
A3 (spkr:5 utt:100)	**1.00**	0.90	0.98
B1 (spkr:10 utt:50)	**0.88**	0.89	0.73
B2 (spkr:10 utt:100)	**0.95**	0.90	0.76
B3 (spkr:10 utt:200)	**0.97**	0.90	0.80

峰性を示す分布に従うことに注意する．この結果から，まず，従来の VB 法に基づいた手法（**VB**）は，全評価セットに対して比較的悪い結果を与える．したがって，**VB** を適用するためにはより大量のデータが必要であることが予想される．一方，従来のサンプリングに基づく手法（**Gibbs**）と提案するサンプリング法に基づく手法（**b-Gibbs**）は，CSJ から作成したすべての評価セット（A1 から B3）に対して，高い精度でクラスタリングできることがわかる．一方，TIMIT から作成した評価セット（T1 と T2）に対しては，従来の **Gibbs** では適切にモデリングできなかったことがわかる．表 3 は，非定常なノイズが重畳されたデータに対するクラスタリング性能である．この結果から，提案する **b-Gibbs** は，いずれの評価セットに対しても従来の **Gibbs** と **VB** よりも高い精度でモデリングが可能であることがわかる．ここで，非定常なノイズが重畳されたデータは，多峰性の分布に従うと考えられることに注意する．この結果から，従来の **Gibbs** は，単峰な分布に従うデータに対しては高い精度でモデリングできるものの，多峰性を示す分布に従うデータ（クリーン評価セットにおける T1，T2 と，ノイズ評価セットにおける A1 から B3）に対しては適切にモデリングできないことがわかる．一方，提案する **b-Gibbs** では，このようなデータセットに対しても適切にモデリングが行えることがわかる．

5. まとめと今後の展望

Blocked ギブスサンプリングとその貪欲的な近似である ICM を用いて，多重スケール混合モデルを推定するための手法を提案した．提案するサンプリング手法は，従来のギブスサンプリング法に基づく手法では局所解に陥ってしまう複雑な多峰性の分布に従うデータに対しても，頑健にモデリングできることを示した．

我々は，多重スケール混合分布をノンパラメトリックベイズモデルに拡張することで，最適な話者クラスタ数も同時に推定できる手法を提案している [17]．しかし，この手法では，従来のギブスサンプリングに基づいた手法に基づき，モデル構造の推定を行なっている．そこで，今後の予定として，提案するサンプリング手法を，ノンパラメトリックベイズモデルに適用することを検討している．

表 3 ノイズ評価セットに対する実験結果（K 値）

Evaluation data	b-Gibbs	Gibbs	VB
A1 +noise1 (spkr:5 utt:25)	**0.89**	0.67	0.64
A2 +noise1 (spkr:5 utt:50)	**0.88**	0.71	0.72
A3 +noise1 (spkr:5 utt:100)	**0.84**	0.67	0.74
B1 +noise1 (spkr:10 utt:50)	**0.75**	0.65	0.57
B2 +noise1 (spkr:10 utt:100)	**0.75**	0.66	0.62
B3 +noise1 (spkr:10 utt:200)	**0.77**	0.69	0.74
A1 +noise2 (spkr:5 utt:25)	**0.84**	0.71	0.53
A2 +noise2 (spkr:5 utt:50)	**0.80**	0.66	0.63
A3 +noise2 (spkr:5 utt:100)	**0.88**	0.68	0.72
B1 +noise2 (spkr:10 utt:50)	**0.77**	0.72	0.56
B2 +noise2 (spkr:10 utt:100)	**0.75**	0.61	0.63
B3 +noise2 (spkr:10 utt:200)	**0.74**	0.63	0.71

参考文献

[1] Hofmann, T.: Probabilistic latent semantic indexing, *SIGIR*, New York, NY, USA, ACM, pp. 50–57 (1999).
[2] Blei, D. M. et al.: Latent Dirichlet allocation, *J. Mach. Learn. Res.*, Vol. 3, pp. 993–1022 (2003).
[3] Valente, F. and Wellekens, C. J.: Variational Bayesian adaptation for speaker clustering, *ICASSP* (2005).
[4] Watanabe, S. et al.: Gibbs sampling based Multi-scale Mixture Model for speaker clustering., *ICASSP*, IEEE, pp. 4524–4527 (2011).
[5] Tawara, N. et al.: Fully Bayesian inference of multi-mixture Gaussian model and its evaluation using speaker clustering, *ICASSP*, pp. 5253–5256 (2012).
[6] luc Gauvain, J. and hui Lee, C.: Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains, *IEEE Trans. Speech Audio Process.*, Vol. 2, pp. 291–298 (1994).
[7] Watanabe, S. et al.: Variational Bayesian estimation and clustering for speech recognition, *IEEE Trans. Speech Audio Process.*, Vol. 12, pp. 365–381 (2004).
[8] Reynolds, D. A. et al.: Speaker verification using adapted Gaussian mixture models, *Digital Signal Processing*, p. 2000 (2000).
[9] Valente, F. et al.: Variational Bayesian speaker diarization of meeting recordings., *ICASSP*, IEEE, pp. 4954–4957 (2010).
[10] Sung, J. et al.: Latent-space variational Bayes, *IEEE Tran. on PAMI*, Vol. 30, No. 12, pp. 2236–2242 (2008).
[11] Liu, J. S.: *Monte Carlo strategies in scientific computing*, Springer, corrected edition (2008).
[12] Kirkpatrick, S. et al.: Optimization by simulated annealing, *Science*, Vol. 220, pp. 671–680 (1983).
[13] Bishop, C. M.: *Pattern Recognition and Machine Learning*, Springer-Verlag New York, Inc. (2006).
[14] Jensen, C. S. and Kong, A.: Blocking Gibbs sampling in very large probabilistic expert systems, *Internat. J. Human-Computer Studies*, Vol. 42, pp. 647–666 (1995).
[15] Kittler, J. and Föglein, J.: Contextual classification of multispectral pixel data., *Image Vision Comput.*, Vol. 2, No. 1, pp. 13–29 (1984).
[16] Solomonoff, A. et al.: Clustering speakers by their voices, *ICASSP*, pp. 757–760 (1998).
[17] Tawara, N. et al.: Fully Bayesian speaker clustering based on hierarchically structured utterance-oriented Dirichlet process mixture model, *INTERSPEECH* (2012).