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

2007－ＳＬＰ－69 (43)
2007／12／21

変分ベイズ法に基づく声質変換

丸目 雅浩† 南角 吉彦† 酒向 慎司† 徳田 恵一† 北村 正†

† 名古屋工業大学大学院 工学研究科 情報工学専攻
〒466–8555 名古屋市 昭和区 御器所町

あらまし 音声合成の需要の高まりにより，多様な話者性や発話スタイルを持った音声の合成が望まれている．しかし，このような音声の合成には，話者や発話スタイルに応じてモデルを用意する必要があり現実的ではない．そこで，少量の学習データにより，多様な話者性を持つ音声の合成を可能とする混合ガウスモデル（GMM）に基づく声質変換が提案されている．しかし，従来の GMM に基づく声質変換では，尤度最大化（ML）基準によりモデルパラメータを点推定しているため，学習データが十分に得られない場合，モデルの推定精度が低下する可能性がある．そこで，GMM に基づく声質変換に変分ベイズ法を適用し，ベイズ基準による声質変換を行う．提案法では，ML 基準に比べて，声質変換の音質と話者性において，品質向上が確認でき，推定精度の高いモデルが得られることがわかった．
キーワード 変分ベイズ法，声質変換，混合ガウスモデル

Voice Conversion based on Variational Bayesian Method

Masahiro MARUME†, Yoshihiko NANKAKU†, Shinji SAKO†, Keiichi TOKUDA†,
and Tadashi KITAMURA†

† Depertment of Computer Science and Engineering, Nagoya Institute of Technology
Gokiso-cho, Showa-ku, Nagoya, 466–8555 Japan

Abstract It is desired a technique for synthesizing speech with various speaker characteristics and speaking styles, by increasing the demand of speech synthesis. However, a large amount of training data is required to construct the system for each characteristics and speaking styleVoice conversion based on Gaussian Mixture Model(GMM) is one of techniques which can solve this problem. GMM is estimated from a small amount of training data based on the Maximam Likelihood(ML) criterion. However, the GMM based voice conversion technique still suffers from the overfitting problem due to insufficient training data and a point estimation of the ML criterion. To improve this problem, we applied the varational Bayes method to the GMM based voice conversion. In experiments, it was confirmed that the proposed technique improves the quality of converted voice, because of its higher generalization ability than the conventional ML based approach.
Key words Variational bayesian method, Voice conversion, Gaussian mixture model

1. はじめに

近年の音声合成技術の発展により，テキスト音声合成の品質は向上し，カーナビゲーションシステムやロボットなどに用いられているが，音声合成技術が人と機械のインタフェースとして，広く普及するためには，合成音声の高品質化だけでなく，多様な話者性や発話スタイルを持つ音声の合成が必要となる．しかし，現在の音声合成システムでは，話者や発話スタイルに応じてモデルを用意する必要があり，現実的ではない．そのため，多様な話者性を持つ音声を合成する手法として，モデル適応やスペクトル変換などの様々な適応・変換手法が研究されている．その中の一つとして，スペクトル変換に基づいた声質変換 [1] がある．これは，ある話者（元話者）の音声を別の話者（目標話者）の音声へと変換できる技術である．言語情報を用いることなく，2 話者間のスペクトル特徴量の対応関係をモデル化し，それを用いて変換を行う．声質変換の手法として，コードブックマッピング法による声質変換や GMM に基づく声質変換 [2] が挙げられるが，GMM に基づく声質変換はコードブックマッピング法に比べ，品質の良い変換が可能であり，広く用いられている．

従来の GMM に基づく声質変換は，音韻バランスを考慮した数十文章の学習データを用いて行われるが，スペクトル特徴

量のみを変換するため，変換性能を劣化させず，学習データ量を削減できる可能性があると考えられる．しかし，従来は，モデルの学習や目標話者の特徴量の推定をML基準により行っており，モデルパラメータを点推定しているため，学習データが十分に得られない場合にモデルの推定精度が低下してしまう．それに対して，ベイズ基準では，モデルパラメータを確率変数として，学習データに対する事後分布を推定することで，高い汎化性能を持つモデルが得られる．さらに，学習前の事前知識を事前分布として与えることができるため，適切な事前分布を与えることで，極少量の学習データにより，推定精度の高いモデルが得られると考えられる．しかし，事後分布の推定には困難な積分計算が含まれており，近似が必要となる．従来は，マルコフ連鎖モンテカルロ（MCMC）法によるサンプリング手法やラプラス近似法を用いてガウス関数近似を行う手法が用いられていたが，それぞれMCMC法は膨大な計算量，ラプラス近似は少量の学習データにおける近似精度に問題があった．この問題に対し，変分ベイズ法[3]が近年提案され，様々なモデルにベイズ基準の適用が可能となった．また，音声認識や音声合成の分野においては，HMMによる音響モデルに適用されている[4],[5]．そこで，本稿ではGMMに基づく声質変換に変分ベイズ法を適用し，ベイズ基準による声質変換を行う．

以下，**2.** で，従来のGMMに基づく声質変換について述べ，**3.** で，提案法である変分ベイズ法に基づく声質変換について述べる．また，**4.** で提案法を用いた声質変換について述べ，最後に **5.** で本稿のまとめと今後の課題について述べる．

2. GMMに基づく声質変換

GMMに基づく声質変換では，GMMによって2話者間のスペクトル特徴量 $\boldsymbol{X}$，$\boldsymbol{Y}$ の対応関係をモデル化し，そのモデルに基づいて，元話者の特徴量に対応する目標話者の特徴量を推定する．

DPマッチングによりフレーム毎に対応付けられた元話者，目標話者のスペクトル特徴量をそれぞれ $\boldsymbol{X}_n = [\boldsymbol{x}_n^\top, \Delta\boldsymbol{x}_n^\top]^\top$，$\boldsymbol{Y}_n = [\boldsymbol{y}_n^\top, \Delta\boldsymbol{y}_n^\top]^\top$ とした場合の結合ベクトル $\boldsymbol{Z} = [\boldsymbol{Z}_1^\top, \boldsymbol{Z}_2^\top, \ldots, \boldsymbol{Z}_N^\top]^\top$，$\boldsymbol{Z}_n = [\boldsymbol{X}_n^\top, \boldsymbol{Y}_n^\top]^\top$ をGMMの学習データとする．ただし，$\Delta\boldsymbol{x}_n, \Delta\boldsymbol{y}_n$ はそれぞれの話者の動的特徴量である．この学習データを用いて，GMMによるモデル化を行う．

$$P(\boldsymbol{Z}|\boldsymbol{\lambda}) = \prod_{n=1}^{N}\sum_{i=1}^{M} w_i \mathcal{N}(\boldsymbol{Z}_n|\boldsymbol{\mu}_i^{(Z)}, \boldsymbol{\Sigma}_i^{(Z)}) \tag{1}$$

$$\boldsymbol{\mu}_i^{(Z)} = \begin{bmatrix}\boldsymbol{\mu}_i^{(X)} \\ \boldsymbol{\mu}_i^{(Y)}\end{bmatrix}, \quad \boldsymbol{\Sigma}_i^{(Z)} = \begin{bmatrix}\boldsymbol{\Sigma}_i^{(XX)} & \boldsymbol{\Sigma}_i^{(XY)} \\ \boldsymbol{\Sigma}_i^{(YX)} & \boldsymbol{\Sigma}_i^{(YY)}\end{bmatrix} \tag{2}$$

ただし，w_i は混合重み，$\boldsymbol{\mu}_i^{(Z)}$ は平均ベクトル，$\boldsymbol{\Sigma}_i^{(Z)}$ は共分散行列，M は混合数である．ML基準に基づく声質変換では，次式のように，学習データ $\boldsymbol{Z}$ の尤度を最大にするモデルパラメータ $\boldsymbol{\lambda}$ を推定する．

$$\boldsymbol{\lambda}^{(ML)} = \arg\max_{\boldsymbol{\lambda}} P(\boldsymbol{Z}|\boldsymbol{\lambda}) \tag{3}$$

これは，EMアルゴリズムにより実現可能である．

変換時については，入出力話者のスペクトル特徴量系列をそれぞれ $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top, \ldots, \boldsymbol{X}_N^\top]^\top$，$\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \boldsymbol{Y}_2^\top, \ldots, \boldsymbol{Y}_N^\top]^\top$ とすると，$\boldsymbol{Y}$ の最適な静的特徴量系列 $\boldsymbol{y} = [\boldsymbol{y}_1^\top, \boldsymbol{y}_2^\top, \ldots, \boldsymbol{y}_N^\top]^\top$ は式(3)を用いて以下のように表される．

$$\boldsymbol{y}^{(ML)} = \arg\max_{\boldsymbol{y}} P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}^{(ML)}) \tag{4}$$

式(4)の $\boldsymbol{X}$ が与えられたときの $\boldsymbol{Y}$ の事後確率 $P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}^{(ML)})$ は以下のようになる．

$$P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}^{(ML)}) = \sum_{\text{all}\,\boldsymbol{m}} P(\boldsymbol{m}|\boldsymbol{X}, \boldsymbol{\lambda}^{(ML)}) P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{m}, \boldsymbol{\lambda}^{(ML)}) \tag{5}$$

ここで，$\boldsymbol{m} = \{m_n\}_{n=1}^N$ は混合要素系列である．ただし，本稿では式(5)を $P(\boldsymbol{m}|\boldsymbol{X}, \boldsymbol{\lambda}^{(ML)})$ を最大とするような単一の $\boldsymbol{m}$ により近似する．このとき式(5)の対数は以下のようになる．

$$\begin{aligned}&\log P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{m}, \boldsymbol{\lambda}^{(ML)}) \\ &= -\frac{1}{2}\boldsymbol{Y}^\top \boldsymbol{D}_{\boldsymbol{m}}^{-1}\boldsymbol{Y} + \boldsymbol{Y}^\top \boldsymbol{D}_{\boldsymbol{m}}^{-1}\boldsymbol{E}_{\boldsymbol{m}} + K\end{aligned} \tag{6}$$

ただし，

$$\boldsymbol{E}_{\boldsymbol{m}} = [\boldsymbol{E}_1(m_1), \boldsymbol{E}_2(m_2), \cdots, \boldsymbol{E}_N(m_N)] \tag{7}$$

$$\boldsymbol{D}_{\boldsymbol{m}}^{-1} = \mathrm{diag}[\boldsymbol{D}^{-1}(m_1), \boldsymbol{D}^{-1}(m_2), \cdots, \boldsymbol{D}^{-1}(m_N)] \tag{8}$$

$$\boldsymbol{E}_n(i) = \boldsymbol{\mu}_i^{(Y)} + \boldsymbol{\Sigma}_i^{(YX)}\boldsymbol{\Sigma}_i^{(XX)^{-1}}(\boldsymbol{X}_n - \boldsymbol{\mu}_i^{(X)}) \tag{9}$$

$$\boldsymbol{D}(i) = \boldsymbol{\Sigma}_i^{(YY)} - \boldsymbol{\Sigma}_i^{(YX)}\boldsymbol{\Sigma}_i^{(XX)^{-1}}\boldsymbol{\Sigma}_i^{(XY)} \tag{10}$$

であり，K は $\boldsymbol{Y}$ とは独立な定数である．よって，ML基準における最適な $\boldsymbol{y}$ は式(6)を最大化することで得られる．ここで，静的特徴量系列 $\boldsymbol{y}$ と静的・動的特徴量系列 $\boldsymbol{Y}$ の間には，$\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}$ の関係が成り立つ[6]．ただし，$\boldsymbol{W}$ は静的特徴量系列に動的特徴量を付加する行列である．これより，以下のような動的特徴量を考慮した滑らかなスペクトル特徴量系列を得ることができる．

$$\boldsymbol{y}^{(ML)} = (\boldsymbol{W}^\top \boldsymbol{D}_{\boldsymbol{m}}^{-1}\boldsymbol{W})^{-1}(\boldsymbol{W}^\top \boldsymbol{D}_{\boldsymbol{m}}^{-1}\boldsymbol{E}_{\boldsymbol{m}}) \tag{11}$$

また，GMMによるモデル化を行わない F_0 とパワー項は元話者の値に基づき，それぞれ以下の線形方程式により変換する．

$$p_y = \frac{p_x - \mu_x}{\sigma_x} \times \sigma_y + \mu_y \tag{12}$$

ただし，p_x, p_y は変換前と変換後の F_0 またはパワー項の値，μ_x, σ_x は変換前の F_0 または パワー項の平均と分散，μ_y, σ_y は変換後の平均と分散である．

3. 変分ベイズ法に基づく声質変換

3.1 変分ベイズ法による事後分布の近似

ML基準では，モデルパラメータを点推定するのに対して，ベイズ基準では，学習データに対する事後分布の推定を行う．よって，GMMに基づく声質変換にベイズ基準を適用する場合，目標話者の静的特徴量系列 $\boldsymbol{y}$ は以下の式より得られる．

$$\boldsymbol{y}^{(Bayes)} = \arg\max_{\boldsymbol{y}} \int P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{m}, \boldsymbol{\lambda}) P(\boldsymbol{\lambda}|\boldsymbol{Z}) d\boldsymbol{\lambda} \tag{13}$$

ただし，混合要素系列 $\boldsymbol{m}$ はあらかじめ与えられると仮定する．式 (13) は，$\boldsymbol{Y}$ の事後確率 $P(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{m},\boldsymbol{\lambda})$ をモデルパラメータの事後分布 $P(\boldsymbol{\lambda}|\boldsymbol{Z})$ により，重み付き平均しており，これにより高い汎化性能が得られる．しかし，ベイズ基準における事後分布 $P(\boldsymbol{\lambda}|\boldsymbol{Z})$ の推定には，困難な積分計算が必要となるため，変分ベイズ法によって，事後分布の近似分布を推定する．

変分ベイズ法では，全ての未知パラメータを周辺化した以下の周辺尤度を考える．

$$P(\boldsymbol{Z}) = \sum_{\boldsymbol{m}} \int P(\boldsymbol{Z},\boldsymbol{m}|\boldsymbol{\lambda})P(\boldsymbol{\lambda})d\boldsymbol{\lambda} \tag{14}$$

この対数を取った対数周辺尤度に対し，事後分布 $P(\boldsymbol{m},\boldsymbol{\lambda}|\boldsymbol{Z})$ の近似分布 $Q(\boldsymbol{m},\boldsymbol{\lambda})$ を用いて，下限 $\mathcal{F}$ を定義する．

$$\begin{aligned}\log P(\boldsymbol{Z}) &= \log \sum_{\boldsymbol{m}} \int Q(\boldsymbol{m},\boldsymbol{\lambda}) \frac{P(\boldsymbol{Z},\boldsymbol{m}|\boldsymbol{\lambda})P(\boldsymbol{\lambda})}{Q(\boldsymbol{m},\boldsymbol{\lambda})} d\boldsymbol{\lambda} \\ &\geq \left\langle \log \frac{P(\boldsymbol{Z},\boldsymbol{m}|\boldsymbol{\lambda})P(\boldsymbol{\lambda})}{Q(\boldsymbol{m},\boldsymbol{\lambda})} \right\rangle_{Q(\boldsymbol{m},\boldsymbol{\lambda})} \\ &= \mathcal{F} \end{aligned} \tag{15}$$

ここで $\langle\cdot\rangle_{Q(\boldsymbol{x})}$ は $Q(\boldsymbol{x})$ に関する期待値を表す．次に，$Q(\boldsymbol{m},\boldsymbol{\lambda})$ に積分計算が可能となるような拘束条件を与える．

$$Q(\boldsymbol{m},\boldsymbol{\lambda}) = Q(\boldsymbol{m})Q(\boldsymbol{\lambda}) \tag{16}$$

これより，式 (15) の $\mathcal{F}$ を最大化するような分布 $Q(\boldsymbol{m}), Q(\boldsymbol{\lambda})$ が事後分布 $P(\boldsymbol{m},\boldsymbol{\lambda}|\boldsymbol{Z})$ の最適な近似分布となる．変分近似を用いて，$Q(\boldsymbol{m}), Q(\boldsymbol{\lambda})$ は次式で表される．

$$Q(\boldsymbol{m}) \propto \exp\langle \log P(\boldsymbol{Z},\boldsymbol{m}|\boldsymbol{\lambda})\rangle_{Q(\boldsymbol{\lambda})} \tag{17}$$

$$Q(\boldsymbol{\lambda}) \propto P(\boldsymbol{\lambda}) \exp\langle \log P(\boldsymbol{Z},\boldsymbol{m}|\boldsymbol{\lambda})\rangle_{Q(\boldsymbol{m})} \tag{18}$$

これらの近似分布は相互に関係しているため，$Q(\boldsymbol{m})$ と $Q(\boldsymbol{\lambda})$ の更新を交互に繰り返すことで $\mathcal{F}$ を極大に導くことができる．また，各更新において，$\mathcal{F}$ は必ず増加するため，収束性が保証されている．

ここで，事前分布 $P(\boldsymbol{\lambda})$ を次式により定義する．

$$P(\boldsymbol{\lambda}) = P(\boldsymbol{w}) \prod_{i=1}^{M} P(\boldsymbol{b}_i) \tag{19}$$

ただし，$\boldsymbol{w} = \{w_i\}_{i=1}^{M}$ は混合重み，$\boldsymbol{b}_i = \{\boldsymbol{\mu}_i^{(Z)}, \boldsymbol{S}_i^{(Z)}\}$ は出力確率パラメータであり，$\boldsymbol{\mu}_i^{(Z)}$ は平均ベクトル，$\boldsymbol{S}_i^{(Z)}$ は共分散行列の逆行列を表す．よって，$Q(\boldsymbol{\lambda})$ は以下のように表すことができる．

$$Q(\boldsymbol{\lambda}) = Q(\boldsymbol{w}) \prod_{i=1}^{M} Q(\boldsymbol{b}_i) \tag{20}$$

$$\begin{aligned} Q(\boldsymbol{w}) &\propto P(\boldsymbol{w}) \exp\langle \log P(\boldsymbol{m}|\boldsymbol{\lambda})\rangle_{Q(\boldsymbol{m})} \\ &= P(\boldsymbol{w}) \exp\left\{ \sum_{i=1}^{M} \sum_{n=1}^{N} \langle m_n^i \rangle \log w_i \right\} \\ Q(\boldsymbol{b}_i) &\propto P(\boldsymbol{b}_i) \exp\langle \log P(\boldsymbol{Z}|\boldsymbol{m},\boldsymbol{\lambda})\rangle_{Q(\boldsymbol{m})} \\ &= P(\boldsymbol{b}_i) \exp\left\{ \sum_{n=1}^{N} \langle m_n^i \rangle \log \mathcal{N}(\boldsymbol{Z}_n|\boldsymbol{\mu}_i^{(Z)}, \boldsymbol{S}_i^{(Z)^{-1}}) \right\} \end{aligned} \tag{21}$$

ただし，$m_n^i = \delta(m_n, i)$ であり，$m_n = i$ の場合は 1，それ以外は 0 となる．ここで，

$$\langle m_n^i \rangle = \sum_{\boldsymbol{m}} Q(\boldsymbol{m}) m_n^i \tag{22}$$

である．また，これらの近似分布を用いることで，$Q(\boldsymbol{m})$ は以下のようになる．

$$\begin{aligned} Q(\boldsymbol{m}) = &\exp \prod_{i=1}^{M} \prod_{n=1}^{N} \langle \log w_{m_n} \rangle_{Q(w_{m_n})} \\ &\times \exp \prod_{n=1}^{N} \langle \log \mathcal{N}(\boldsymbol{Z}_n|\boldsymbol{\mu}_{m_n}^{(Z)}, \boldsymbol{S}_{m_n}^{(Z)^{-1}}) \rangle_{Q(b_{m_n})} \end{aligned} \tag{23}$$

式 (23) は GMM の完全データの尤度関数と同じ形になるため，式 (22) の期待値は，EM アルゴリズムを用いて計算することができる．

3.2 事前分布の設定

本稿では，モデルパラメータの事前分布 $P(\boldsymbol{\lambda})$ として，共役事前分布を用いる．共役事前分布とは，事前分布 $P(\boldsymbol{\lambda})$ と事後分布 $P(\boldsymbol{\lambda}|\boldsymbol{Z})$ が同じ分布族となる事前分布のことである．ここでは，$P(\boldsymbol{w})$ に Dirichlet 分布，$P(\boldsymbol{b}_i)$ に Gauss-Wishart 分布を用いる．よって，事後分布の近似分布である $Q(\boldsymbol{w})$ と $Q(\boldsymbol{b}_i)$ がそれぞれ Dirichlet 分布と Gauss-Wishart 分布となるような事前分布を定義する．

$$P(\boldsymbol{w}) = \mathcal{D}(\boldsymbol{w}|\boldsymbol{\phi}_i) = \frac{\Gamma\left(\sum_{i=1}^{M} \phi_i\right)}{\prod_{i=1}^{M} \Gamma(\phi_i)} \prod_{i=1}^{M} w_i^{\phi_i - 1} \tag{24}$$

$$P(\boldsymbol{b}_i) = \mathcal{N}(\boldsymbol{\mu}_i^{(Z)}|\boldsymbol{\nu}_i, (\xi_i \boldsymbol{S}_i^{(Z)})^{-1})\, \mathcal{W}(\boldsymbol{S}_i^{(Z)}|\eta_i, \boldsymbol{B}_i) \tag{25}$$

$$\begin{aligned} &\mathcal{N}(\boldsymbol{\mu}_i^{(Z)}|\boldsymbol{\nu}_i, (\xi_i \boldsymbol{S}_i{}^{(Z)})^{-1}) = C_{\mathcal{N}_i} |\boldsymbol{S}_i^{(Z)}|^{\frac{1}{2}} \\ &\quad \times \exp\left\{ -\frac{1}{2} \mathrm{Tr}\left(\xi_i \boldsymbol{S}_i^{(Z)} (\boldsymbol{\mu}_i^{(Z)} - \boldsymbol{\nu}_i)(\boldsymbol{\mu}_i^{(Z)} - \boldsymbol{\nu}_i)^T \right) \right\} \end{aligned} \tag{26}$$

$$\begin{aligned} &\mathcal{W}(\boldsymbol{S}_i^{(Z)}|\eta_i, \boldsymbol{B}_i) = \\ &\quad C_{\mathcal{W}_i} |\boldsymbol{S}_i^{(Z)}|^{\frac{\eta_i - d - 1}{2}} \exp\left\{ -\frac{1}{2} \mathrm{Tr}(\boldsymbol{S}_i^{(Z)} \boldsymbol{B}_i) \right\} \end{aligned} \tag{27}$$

$$\bar{C}_{\mathcal{N}_i} = (2\pi)^{-\frac{d}{2}} \xi_i^{\frac{d}{2}} \tag{28}$$

$$\bar{C}_{\mathcal{W}_i} = \frac{|\boldsymbol{B}_i|^{\frac{\eta_i}{2}}}{2^{\frac{\eta_i d}{2}} \pi^{\frac{d(d-1)}{4}} \prod_{j=1}^{d} \Gamma\left(\frac{\eta_i + 1 - j}{2}\right)} \tag{29}$$

ここで，$\Gamma(\cdot)$ はガンマ関数，d はスペクトル特徴量の次元数である．事前分布を表すパラメータは $\{\phi_i, \xi_i, \eta_i, \boldsymbol{\nu}_i, \boldsymbol{B}_i\}_{i=1}^{M}$ となり，これをハイパーパラメータと呼ぶ．また，事前分布として，共役事前分布を用いているため，事後分布のハイパーパラメータも同様のパラメータセットにより表すことができる．ここでは，$\{\bar{\phi}_i, \bar{\xi}_i, \bar{\eta}_i, \bar{\boldsymbol{\nu}}_i, \bar{\boldsymbol{B}}_i\}_{i=1}^{M}$ とする．

事前分布は学習前の事前情報として任意の分布を与えることができるため，そのハイパーパラメータの値は事後分布の推定に影響を与えるといえる．よって，適切なハイパーパラメータを設定することで，モデルの推定精度を向上させられると考えられる．本稿では，事前分布をあらかじめ用意されたデータ $\{\boldsymbol{Z}_t\}_{t=1}^{T_i}$ (ここでは事前データと呼ぶ) を用いて設定する．GMM

の尤度関数における出力確率分布を以下のように展開し，

$$\prod_{t=1}^{T_i}\mathcal{N}(\boldsymbol{Z}_t|\boldsymbol{\mu}_i^{(Z)},\boldsymbol{S}_i^{(Z)^{-1}})$$
$$\propto \mathcal{N}(\boldsymbol{\mu}_i^{(Z)}|\bar{\boldsymbol{\mu}}_i,(T_i\boldsymbol{S}^{(Z)})^{-1})\mathcal{W}(\boldsymbol{S}_i^{(Z)}|T_i+d,(T_i\bar{\boldsymbol{\Sigma}}_i)) \quad (30)$$

これを事前分布とみなすと，ハイパーパラメータはそれぞれ以下のように与えられる．

$$\xi_i = T_i,\quad \eta_i = T_i + d,\quad \boldsymbol{\nu}_i = \bar{\boldsymbol{\mu}}_i,\quad \boldsymbol{B}_i = T_i\bar{\boldsymbol{\Sigma}}_i \quad (31)$$

ただし，

$$\bar{\boldsymbol{\mu}}_i = \frac{1}{T_i}\sum_{t=1}^{T_i}\boldsymbol{Z}_n,\quad \bar{\boldsymbol{\Sigma}}_i = \frac{1}{T_i}\sum_{t=1}^{T_i}\boldsymbol{Z}_n\boldsymbol{Z}_n^\top - \bar{\boldsymbol{\mu}}_i\bar{\boldsymbol{\mu}}_i^\top \quad (32)$$

である．ここで，T_i は学習データ量，$\bar{\boldsymbol{\mu}}_i$，$\bar{\boldsymbol{\Sigma}}_i$ はそれぞれ平均ベクトルと共分散行列を表している．本稿では，事前データにより学習された GMM の混合重み，平均ベクトル，共分散行列を用いる．ただし，T_i に関しては，GMM の混合重みを α_i として，以下のように定義する．

$$T_i = \alpha_i T \quad (33)$$

よって，この T を調整パラメータとして用いることとする．

3.3 事後分布の推定

事後分布を推定するためのハイパーパラメータの具体的な更新式について以下で説明する．本稿では，まず，$Q(\boldsymbol{m})$ の更新を行う．

$$\langle m_n^i\rangle \propto \exp\langle\log w_i\rangle_{Q(w)} \times \exp\langle\log\mathcal{N}(\boldsymbol{Z}_n|\boldsymbol{\mu}_i^{(Z)},\boldsymbol{S}_i^{(Z)^{-1}})\rangle_{Q(b_i)} \quad (34)$$

$$\langle\log w_i\rangle_{Q(w)} = \Psi(\bar{\phi}_i) - \Psi\left(\sum_{i=1}^{M}\bar{\phi}_i\right) \quad (35)$$

$$\langle\log\mathcal{N}(\boldsymbol{Z}_n|\boldsymbol{\mu}_i^{(Z)},\boldsymbol{S}_i^{(Z)^{-1}})\rangle_{Q(b_i)}$$
$$= -\frac{1}{2}\left[d\log\pi + \frac{d}{\bar{\xi}_i} - \sum_{j=1}^{d}\Psi\left(\frac{\bar{\eta}_i+1-j}{2}\right) + \log|\bar{\boldsymbol{B}}_i| + Tr\left\{\bar{\eta}_i\bar{\boldsymbol{B}}_i^{-1}(\boldsymbol{Z}_n-\bar{\boldsymbol{\nu}}_i)(\boldsymbol{Z}_n-\bar{\boldsymbol{\nu}}_i)^\top\right\}\right] \quad (36)$$

ただし，$\Psi(\cdot)$ は digamma 関数である．式 (34) を用いて求められる期待値は以下のようになる．

$$\bar{N}_i = \sum_{n=1}^{N}\langle m_n^i\rangle \quad (37)$$

$$\bar{\boldsymbol{Z}}_i = \frac{1}{\bar{N}_i}\sum_{n=1}^{N}\langle m_n^i\rangle\boldsymbol{Z}_n \quad (38)$$

$$\bar{\boldsymbol{C}}_i = \frac{1}{\bar{N}_i}\sum_{n=1}^{N}\langle m_n^i\rangle(\boldsymbol{Z}_n-\bar{\boldsymbol{Z}}_i)(\boldsymbol{Z}_n-\bar{\boldsymbol{Z}}_i)^\top \quad (39)$$

これらの期待値を用いて，$Q(\boldsymbol{\lambda})$ のハイパーパラメータを以下のように更新する．

$$\bar{\phi}_i = \phi_i + \bar{N}_i,\quad \bar{\xi}_i = \xi_i + \bar{N}_i,\quad \bar{\eta}_i = \eta_i + \bar{N}_i \quad (40)$$

$$\bar{\boldsymbol{\nu}}_i = \frac{\bar{N}_i\bar{\boldsymbol{Z}}_i + \xi_i\boldsymbol{\nu}_i}{\bar{N}_i + \xi_i} \quad (41)$$

$$\bar{\boldsymbol{B}}_i = \bar{N}_i\bar{\boldsymbol{C}}_i + \boldsymbol{B}_i + \frac{\bar{N}_i\xi_i}{\bar{N}_i+\xi_i}(\bar{\boldsymbol{Z}}_i-\boldsymbol{\nu}_i)(\bar{\boldsymbol{Z}}_i-\boldsymbol{\nu}_i)^\top \quad (42)$$

表 1 実験条件
Table 1 Experimental condiction

学習データ	ATR 日本語データベース B_set 1 文章, 3 文章
元話者 / 目標話者	mtk / mht
混合数	1, 2, 4, 8
スペクトル特徴量	0 次を除いた 24 次メルケプストラム + Δ
共分散行列	全共分散行列

3.4 ベイズ基準による変換特徴量の生成

変分ベイズ法に基づく声質変換では，元話者の特徴量系列 $\boldsymbol{X}$ に対する目標話者の静的特徴量系列 $\boldsymbol{y}$ が以下の式より得られる．

$$\boldsymbol{y}^{(Bayes)} \simeq \arg\max_{\boldsymbol{y}}\int P(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{m},\boldsymbol{\lambda})Q(\boldsymbol{\lambda})d\boldsymbol{\lambda} \quad (43)$$

これは，式 (13) の事後分布 $P(\boldsymbol{\lambda}|\boldsymbol{Z})$ を近似分布 $Q(\boldsymbol{\lambda})$ に置き換えたものとなっている．ただし，本稿では，簡単のため MAP(事後確率最大化) 近似を用いる．すなわち，推定された事後分布 $Q(\boldsymbol{\lambda})$ を最大化するモデルパラメータ $\boldsymbol{\lambda}$ を求める．

$$\boldsymbol{\lambda}^{(MAP)} \simeq \arg\max_{\boldsymbol{\lambda}}Q(\boldsymbol{\lambda}) \quad (44)$$

得られたモデルパラメータを用いて，従来と同様の方法 (式 (4)) によって変換特徴量の生成を行う．

$$\boldsymbol{y}^{(MAP)} = \arg\max_{\boldsymbol{y}}P(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{m},\boldsymbol{\lambda}^{(MAP)}) \quad (45)$$

これまでの式から変分ベイズ法に基づく声質変換の手順をまとめると以下のようになる．

- **学習部**
1. 事前分布のハイパーパラメータを設定 (式 (31))
2. $Q(\boldsymbol{m})$ の更新 (式 (34) – 式 (36))
3. $Q(\boldsymbol{m})$ を用いた期待値計算 (式 (37) – 式 (39))
4. $Q(\boldsymbol{\lambda})$ の更新 (式 (40) – (42))
5. $\mathcal{F}$ が最大になるまで，2. – 4. を繰り返し更新

- **変換部**
1. $Q(\boldsymbol{\lambda})$ の MAP 近似 (式 (44))
2. 変換特徴量の生成 (式 (45))

4. 声質変換実験

GMM に基づく声質変換はスペクトル変換に基づく手法であるため，少量の学習データによる変換が可能である．そこで，少量の学習データにおいて，GMM に基づく声質変換に変分ベイズ法を適用した場合の有効性の検討を行う．

4.1 モデル探索に関する検討

変分ベイズ法では，$\mathcal{F}$ を最大化するような事後分布の近似分布を推定している．また，ベイズ基準においては，$\mathcal{F}$ が最大となるモデル構造が最適なものとして選択される [7]．これより，$\mathcal{F}$ が最大となるように事前分布を与えることが重要であると考えられる．よって，ここでは，事前分布の調整パラメータである T を変化させた場合の $\mathcal{F}$ についての検討を行う．

実験条件を表 1 に示す．また，本稿では，事前分布を ATR

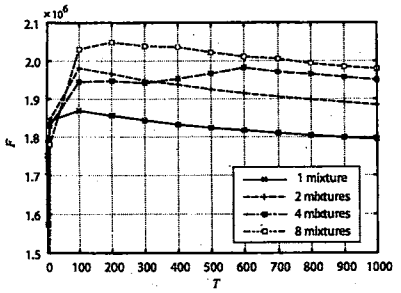

図 1 学習データ 1 文章における $\mathcal{F}$(手法 1)
Fig. 1 $\mathcal{F}$ in 1 sentence(method 1)

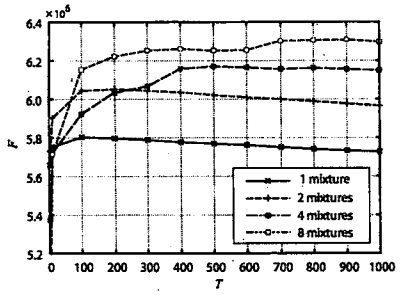

図 2 学習データ 3 文章における $\mathcal{F}$(手法 1)
Fig. 2 $\mathcal{F}$ in 3 sentences(method 1)

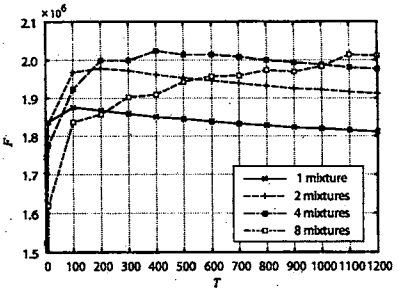

図 3 学習データ 1 文章における $\mathcal{F}$(手法 2)
Fig. 3 $\mathcal{F}$ in 1 sentence(method 2)

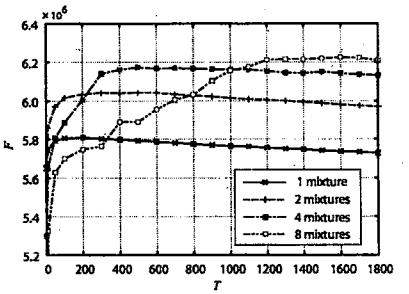

図 4 学習データ 3 文章における $\mathcal{F}$(手法 2)
Fig. 4 $\mathcal{F}$ in 3 sentences(method 2)

日本語データベース C_set の男性話者 40 名の音素連鎖バランス単語 216 語を用いて，以下の 2 種類の方法より求めた．

- **手法 1：不特定話者モデルに基づく事前分布**

1. C_set から任意の 2 話者を 20 通り選択し，選ばれた話者間で結合ベクトルを生成
2. 全ての結合ベクトルを GMM によりモデル化
3. GMM の平均ベクトルと共分散行列から事前分布のハイパーパラメータ $\boldsymbol{\nu}_i$, $\boldsymbol{B}_i$ を設定 (式 (31))

- **手法 2：話者選択に基づく事前分布**

1. C_set の話者 s を GMM $\boldsymbol{\Lambda}$ $(=\{\boldsymbol{\Lambda}^{(s)}\}_{s=1}^{40})$ でモデル化
2. 元話者，目標話者の学習データのスペクトル特徴量 $\boldsymbol{X}$, $\boldsymbol{Y}$ に対する尤度を計算
3. 尤度が最大となる話者 k_x, k_y を選択

$$k_x = \arg\max_{\boldsymbol{X}} P(\boldsymbol{X}|\boldsymbol{\Lambda}^{(s)}) \tag{46}$$

$$k_y = \arg\max_{\boldsymbol{Y}} P(\boldsymbol{Y}|\boldsymbol{\Lambda}^{(s)}) \tag{47}$$

4. 話者 k_x, k_y の 216 語から結合ベクトルを生成
5. 結合ベクトルを GMM によりモデル化
6. GMM の平均ベクトルと共分散行列から事前分布のハイパーパラメータ $\boldsymbol{\nu}_i$, $\boldsymbol{B}_i$ を設定 (式 (31))

実験結果を図 1 ～ 4 に示す．結果より混合数を増やすと，$\mathcal{F}$ が最大となる T の値が増加する傾向があることが分かる．これは，混合数を増やした場合，各混合要素に最適な事前データの情報量を与えるために，より多くの T が必要となるものだと考えられる．また，図 3 で混合数 4，図 1, 2, 4 では，混合数 8 の場合に $\mathcal{F}$ が最大となり，ベイズ基準において，最適なモデル構造であるといえる．この結果から，各混合数において $\mathcal{F}$ を最大とする T のモデルを用いて，次節の主観評価実験を行った．また，$\mathcal{F}$ に基づく最適なモデル構造の選択と合成音声の品質の関係を検討する．

4.2 主観評価実験

GMM に基づく声質変換をそれぞれ ML 基準，ベイズ基準で行った場合の合成音声の品質を比較するため，音質を MOS 評価，話者性を DMOS 評価によりそれぞれ評価した．実験条件は **4.1** と同様に表 1 の条件，事前分布としては手法 1 と手法 2 の 2 種類を用いた．ただし，学習データ 1 文章と 3 文章の評価は別々に行った．また，**4.1** の結果から各混合数において $\mathcal{F}$ を最大とする T のモデルを用いた．被験者 8 名ずつに対して，MOS 評価では，学習に用いない 50 文章より 15 文章，DMOS 評価では，学習に用いない 50 文章より 10 文章をランダムで選び，評価を行った．

実験結果を図 5 ～ 8 に示す．図 5, 6 から学習データが十分に得られない場合，ML 基準では，混合数の増加に伴い，モデルの推定精度が低下し，音質と話者性の両方において，合成音声の品質が劣化していることがわかる．しかし，ベイズ基準では，高い汎化性能を持つモデルが推定でき，さらに，事前分布の影響により，モデルの推定精度が高まるため，音質，話者性ともに向上した．また，手法 1 と手法 2 の比較から，より適切な事前分布を与えた場合に，合成音声の音質と話者性を改善できることが確認できた．図 7, 8 のように，学習データを増加

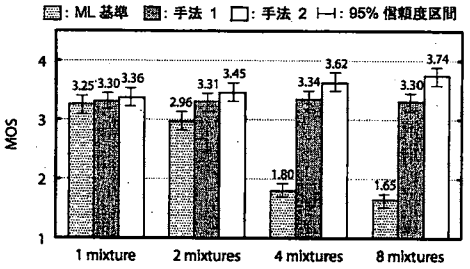

図 5 学習データ 1 文章における MOS

Fig. 5 MOS in 1 sentence

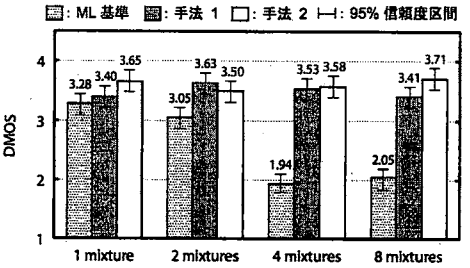

図 6 学習データ 1 文章における DMOS

Fig. 6 DMOS in 1 sentence

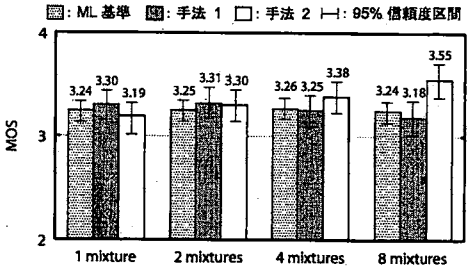

図 7 学習データ 3 文章における MOS

Fig. 7 MOS in 3 sentences

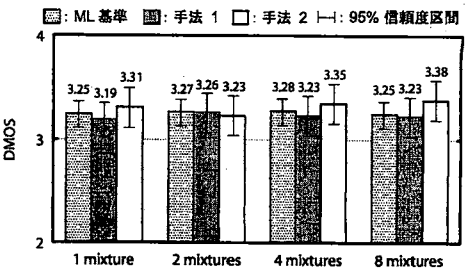

図 8 学習データ 3 文章における DMOS

Fig. 8 DMOS in 3 sentences

させた場合においては，ML 基準に対する手法 1 の有意性がほぼなくなっている．これは，学習データの増加によって，モデルの推定における事前分布の影響が小さくなったことが考えられる．しかし，手法 2 は，手法 1 よりも適切な事前分布を設定しているため，ML 基準に比べ，精度の高いモデルを推定でき，音質，話者性ともに向上した．

$\mathcal{F}$ に基づくモデル構造の選択と合成音声の品質の関係は，手法 2 について，$\mathcal{F}$ に基づき選択された混合数 (1 文章：混合数 4，3 文章：混合数 8) と合成音声の音質，話者性が最も良い混合数が学習データ 3 文章では，一致していることがわかった．また，学習データ 1 文章においても，図 3 より，混合数 4 と 8 の $\mathcal{F}$ の値に差がないため，概ね，一致したといえる．しかし，手法 1 については，$\mathcal{F}$ の傾向 (1 文章：混合数 8，3 文章：混合数 8) と合成音声の品質が一致せず，合成音声の評価は各混合数において，あまり差がないものとなってしまった．このことから，$\mathcal{F}$ に基づいてモデル構造を選択するためには，適切な事前分布を与える必要があると考えられる．

5. むすび

本稿では，GMM に基づく声質変換に変分ベイズ法を適用した．その結果，学習データが十分に得られない場合は，ベイズ基準の汎化性と事前分布の影響によって，ML 基準に比べ，音質と話者性がともに向上することが確認できた．また，学習データを増加させた場合は，事前分布の影響が小さくなるため，不特定話者モデルに基づいて事前分布を与えた場合は，ML 基準に対して，有意な差がほとんど見られなかった．しかし，話者選択に基づき適切な事前分布を設定した場合に，音質と話者性の両方を高めることができた．

$\mathcal{F}$ に基づいたモデル構造の選択と合成音声の品質の関係については，話者選択に基づいて事前分布を与えた場合は，$\mathcal{F}$ が最大となる混合数と合成音声の音質，話者性の評価が最も良い混合数が概ね一致したが，不特定話者モデルに基づき与えた場合は，一致しないという結果になった．

今後の課題としては，複数話者による評価や $\mathcal{F}$ の傾向と合成音声の品質の関係についての検討，ベイズ基準に基づいたスペクトル特徴量の変換が挙げられる．

文　献

[1] 戸田 智基, "最尤特徴量間変換法とその応用," 電子情報通信学会技術研究報告, vol.105, No.571, SP2005-147, pp.49–54, Jan. 2006.

[2] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," IEEE Trans. Speech and Audio Processing, Vol.6, No.2, pp.131–142, Mar. 1998

[3] H.Attias, "Inferring parameters and structure of latent variable models by variational Bayes," *Proc. of the 15th Conference on Uncertainty in Artificial Intelligence*, pp.21–30, 1999.

[4] 南角 吉彦 他, "ベイズ的アプローチに基づく HMM 音声合成," 電子情報通信学会技術研究報告, vol.103, No.264, SP2003–77, pp.19–24, Aug. 2003.

[5] 渡部晋治, "ベイズ法を用いた音声認識," 電子情報通信学会技術研究報告, vol.104, No.470, SP2004–74, pp.13–20, Nov. 2004.

[6] K.Tokuda *et al.*, "Speech parameter generation algorithms for HMM-based speech synthesis," *Proc.of ICASSP*, vol.3, pp.1315–1318, Jun. 2000.

[7] 上田 修功, "ベイズ学習," (全 4 回) 電子情報通信学会誌, vol.85, No.4,6,7,8, 2002.