Vol.2014-SLP-102 No.9
2014/7/26

# 行列変量正規分布の混合モデルとその声質変換への応用

齋藤 大輔[1,a)]　土井 秀信[1,b)]　峯松 信明[2,c)]　広瀬 啓吉[1,d)]

**概要**：本稿では，行列を確率変量とする確率分布を利用した声質変換の枠組みを提案する．声質変換においては，1) 入力・出力話者双方の特徴量空間の精緻なモデル化，2) これらの特徴量空間の変換関係の適切なモデル化の二つを考慮する必要がある．ガウス混合モデル（Gaussian Mixture Model; GMM）に基づく声質変換は，その柔軟性から広く用いられている． 通常，GMM に基づく手法では，入力および出力の特徴量を連結した結合ベクトル空間を最初に構築し，この結合ベクトル空間上において GMM によって結合確率密度関数を表現する事で，入力および出力特徴量の同時確率をモデル化する．このとき結合ベクトルに基づく手法では，主に「結合」特徴量空間の精緻なモデル化を行っていると考えられ，必ずしも入出力特徴量空間の関係性を適切にモデル化しているとは言えない．本稿における提案法では，入出力特徴量の同時確率を行列変量空間における GMM としてモデル化する事で，この問題に対処する．行列空間における行方向および列方向は，声質変換における 2 つのモデル化すべき機能を明示的に捉えており，提案法は入出力双方の特徴量空間の精緻なモデル化と両空間の関係性の適切なモデル化を同時に実現しうるものである．声質変換の実験の結果，提案法が変換性能を向上させる事を示す．

**キーワード**：声質変換, ガウス混合モデル, 行列変量確率分布, 行列変量正規分布, 行列変量ガウス混合モデル

# Mixture Model of Matrix Variate Normal Distribution and its Application to Voice Conversion

Daisuke Saito[1,a)]　Hidenobu Doi[1,b)]　Nobuaki Minematsu[2,c)]　Keikichi Hirose[1,d)]

***Abstract:*** This paper describes a novel approach to voice conversion utilizing probability density functions (PDF) of matrix variate. In voice conversion studies, two important functions should be realized: 1) precise modeling of both the source and target feature spaces, and 2) construction of a proper transform function between these spaces. Voice conversion based on Gaussian mixture model (GMM) is the state-of-the-art standard because of their flexibility and easiness in handling. In GMM-based approaches, a joint vector space of the source and target is first constructed, and the joint PDF of the two vectors is modeled as GMM in the joint vector space. The joint vector approach mainly focuses on precise modeling of the 'joint' feature space, and does not always construct a proper transform between two feature spaces. In contrast, the proposed method constructs the joint PDF as GMM in a matrix variate space whose row and column respectively correspond to the two functions, and it has potential to precisely model both the characteristics of the feature spaces and the relation between the source and target spaces. Experimental results show that the proposed method contributes to improve the performance of voice conversion.

***Keywords:*** voice conversion, Gaussian mixture model, matrix variate distribution, matrix variate normal, matrix variate Gaussian mixture model

[1] 東京大学 大学院情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113–0033, Japan
[2] 東京大学 大学院工学系研究科
Graduate School of Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113–0033, Japan
a) dsk_saito@gavo.t.u-tokyo.ac.jp
b) hdoi@gavo.t.u-tokyo.ac.jp
c) mine@gavo.t.u-tokyo.ac.jp
d) hirose@gavo.t.u-tokyo.ac.jp

© 2014 Information Processing Society of Japan

## 1. はじめに

声質変換は，入出力の対応関係を記述する変換モデルに基づいて，任意の文に対して入力音声の声質を所望の声質へ変換する技術である [1]．声質変換は，広義には異なる二つの特徴量空間のマッピング技術と考えられ，音声を入力とする話者変換に加えて，テキスト音声合成における話者性や，雑音環境下音声の音声特徴量強調など様々な音声技術に応用されている [2], [3]．話者変換のための統計的変換手法は盛んに研究されており，その中でもガウス混合モデル（Gaussian Mixture Model; GMM）に基づく変換法はその柔軟性から広く用いられている [3], [4]．

GMM に基づく変換法では，入力特徴ベクトル，もしくは入力と出力の特徴ベクトルを連結した結合ベクトルに対して，その確率密度分布を GMM によってモデル化する．この GMM を用いて，それぞれの正規分布に対応する線形変換を入力特徴ベクトルの事後確率で重み付けした重み付き線形和として，入出力間の対応関係を導出できる．これらの手法は，確率密度分布のモデル化に際してガウス分布を用いている事から，その学習アルゴリズムを容易に導出可能である．加えて，音声認識で提案されているような，最尤線形回帰（Maxmum likelihood linear regression; MLLR）や事後確率最大化基準（Maximum a Posteriori; MAP）による適応などの各種の話者適応手法の導入 [5], [6]，および出力話者の確率モデルを変換時の事前分布として用いる手法など [7]，柔軟な運用が可能である．

統計的声質変換においては，1) 入力・出力話者双方の特徴量空間の精緻なモデル化 および 2) 入力および出力特徴量空間の変換関係の適切なモデル化 という二つの観点から変換モデルを構築する必要がある．前述の GMM に基づく声質変換法のうち，結合ベクトルに基づくアプローチにおいては，入力および出力の特徴量を連結した結合ベクトル空間を最初に構築し，この単一ベクトル空間の確率密度関数として二つの特徴量の同時分布をモデル化する．すなわち，結合ベクトルに基づくアプローチは，声質変換における前述の二つのモデル化をベクトルの連結操作によって暗に実現していると考える事ができる．ひとたび入力と出力の特徴量ベクトルを連結すれば，同時分布の学習に際して，入力および出力の特徴量空間の特徴は明示的には扱われない．この手法においては，「結合」特徴量空間の精緻なモデル化を行っていると解釈可能である．しかし，入力および出力の特徴量空間に比べて結合特徴量空間は，その次元が大きくなる（通常は 2 倍になる）ため，モデルの複雑度が適切でない場合に，より過学習の影響を受けやすいと考えられる．この場合，例えば相互共分散行列が対角行列となるといった制約を仮定することにより，過学習の影響を低減できるが，この制約が必ずしも声質変換のモデル化において適切とは限らない．声質変換において，上述の二つのモデル化を実現するためには，「対象とする特徴量空間内の相関関係」および「入力および出力空間の間の対応関係」という複数の要因を明示的に捉えた結合モデルの学習を行う必要があると考えられる．

任意話者間の話者変換の実現において，複数の音響的変動要因のモデル化に行列表現を導入する事の有効性が示されている．この観点から，我々はこれまでに任意話者声質変換において，テンソル解析に基づく新しい話者空間表現を提案した [8]．この手法では，各話者は，GMM スーパーベクトルとしてではなく，行と列がそれぞれ GMM の各要素分布と特徴量空間の次元とに対応した行列の形式で表現される．本研究ではこのアプローチに着想を得て，同時分布のモデル化そのものに行列形式の表現を導入する事を検討する．提案法においては，入出力特徴量の同時確率を行列変量空間における GMM としてモデル化する．これにより，入出力双方の特徴量空間の精緻なモデル化と両空間の関係性の適切なモデル化を同時に実現する．

## 2. 結合ベクトルを用いた同時分布のモデル化

本章では，結合ベクトルを用いた GMM に基づく声質変換法について述べる．今，入力話者の発話を表す特徴量系列を $\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n_x}]$，同一発話内容の出力話者の特徴量系列を $\boldsymbol{y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{n_y}]$ とする．動的計画法を用いてこれらの系列をフレーム毎に対応づけることで，結合ベクトル $\boldsymbol{z} = [\boldsymbol{x}^\top, \boldsymbol{y}^\top]^\top$ の特徴量系列 $\boldsymbol{z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n]$ を得る．ただし $^\top$ は転置を表す．この特徴量系列を用いて，以下の式で表される GMM のパラメータを推定し，ベクトル $\boldsymbol{z}_t$ の確率密度をモデル化する．

$$P(\boldsymbol{z}_t|\boldsymbol{\lambda}^{(z)}) = \sum_{m=1}^{M} w_m \mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}) \tag{1}$$

ここで $\mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)})$ は，平均ベクトルを $\boldsymbol{\mu}_m^{(z)}$，分散共分散行列を $\boldsymbol{\Sigma}_m^{(z)}$ とする $m$ 番目の正規分布を表し，$w_m$ は各分布の重みを表す．$\boldsymbol{\lambda}^{(z)}$ は結合ベクトルの GMM の一連のモデルパラメータを表すものとする．$\boldsymbol{z}_t$ のベクトル空間は，その部分空間として入力および出力話者の特徴ベクトルを含むため，これらのモデルパラメータは以下のように表すことができる．

$$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix} \tag{2}$$

ただし $\boldsymbol{\mu}_m^{(x)}$, $\boldsymbol{\Sigma}_m^{(xx)}$, $\boldsymbol{\mu}_m^{(y)}$, $\boldsymbol{\Sigma}_m^{(yy)}$ は $m$ 番目の正規分布における入力または出力話者の平均ベクトルおよび分散共分散行列である．また $\boldsymbol{\Sigma}_m^{(xy)}$ および $\boldsymbol{\Sigma}_m^{(yx)}$ は，入出力話者間の相互共分散行列を表す．過学習の抑制のため，これらの行列が対角行列となるという制約がしばしば用いられる [9]．

変換関数 $\mathcal{F}(\cdot)$ は，入力ベクトル $\boldsymbol{x}_t$ が与えられた場合の $\boldsymbol{y}_t$ の条件付き確率密度に基づいて導出することができる．

© 2014 Information Processing Society of Japan

この確率密度は上述の GMM のモデルパラメータ $\boldsymbol{\lambda}^{(z)}$ によって表現でき，以下のようになる．

$$P(\boldsymbol{y}_t|\boldsymbol{x}_t,\boldsymbol{\lambda}^{(z)})=\sum_{m=1}^{M}P(m|\boldsymbol{x}_t,\boldsymbol{\lambda}^{(z)})P(\boldsymbol{y}_t|\boldsymbol{x}_t,m,\boldsymbol{\lambda}^{(z)}) \quad (3)$$

ここで

$$P(m|\boldsymbol{x}_t,\boldsymbol{\lambda}^{(z)}) = \frac{w_m\mathcal{N}(\boldsymbol{x}_t;\boldsymbol{\mu}_m^{(x)},\boldsymbol{\Sigma}_m^{(xx)})}{\sum_{m=1}^{M}w_m\mathcal{N}(\boldsymbol{x}_t;\boldsymbol{\mu}_m^{(x)},\boldsymbol{\Sigma}_m^{(xx)})} \quad (4)$$

$$P(\boldsymbol{y}_t|\boldsymbol{x}_t,m,\boldsymbol{\lambda}^{(z)}) = \mathcal{N}(\boldsymbol{y}_t;\boldsymbol{E}_{m,t}^{(y)},\boldsymbol{D}_m^{(y)}) \quad (5)$$

$$\boldsymbol{E}_{m,t}^{(y)} = \boldsymbol{\mu}_m^{(y)}+\boldsymbol{\Sigma}_m^{(yx)}\boldsymbol{\Sigma}_m^{(xx)-1}(\boldsymbol{x}_t-\boldsymbol{\mu}_m^{(x)}) \quad (6)$$

$$\boldsymbol{D}_m^{(y)} = \boldsymbol{\Sigma}_m^{(yy)}-\boldsymbol{\Sigma}_m^{(yx)}\boldsymbol{\Sigma}_m^{(xx)-1}\boldsymbol{\Sigma}_m^{(xy)} \quad (7)$$

となる．最小平均二乗誤差基準に基づく変換関数は以下のようになる．

$$\mathcal{F}(\boldsymbol{x}_t)=\sum_{m=1}^{M}P(m|\boldsymbol{x}_t,\boldsymbol{\lambda}^{(z)})\boldsymbol{E}_{m,t}^{(y)} \quad (8)$$

一方，最尤変換に基づくパラメータ生成を導入した場合は，式 (7) における分散共分散行列を考慮し，以下のようなパラメータ生成のための更新式を得る [9].

$$\hat{\boldsymbol{y}}_t=\left(\sum_{m=1}^{M}\gamma_{m,t}\boldsymbol{D}_m^{(y)-1}\right)^{-1}\left(\sum_{m=1}^{M}\gamma_{m,t}\boldsymbol{D}_m^{(y)-1}\boldsymbol{E}_{m,t}^{(y)}\right)$$
$$\gamma_{m,t}=P(m|\boldsymbol{x}_t,\boldsymbol{y}_t,\boldsymbol{\lambda}^{(z)}) \quad (9)$$

式 (9) について，式 (8) と比較すると，各ガウス分布の分散共分散行列の逆行列（精度行列）が含まれており，この行列がそれぞれの条件付きガウス分布の信頼度を表していると考える事ができる．

## 3. 行列変量ガウス混合モデル

### 3.1 行列変量正規分布

本章では，行列変量に基づく統計的モデリングについて述べ，その声質変換への応用を検討する．まずはじめに行列変量の確率分布に関していくつかの基礎を導入する [10]. 今，ランダム行列 $\boldsymbol{X}$ を考え，その行と列のサイズをそれぞれ $n$ および $p$ とする（$\boldsymbol{X}\in\mathcal{R}^{n\times p}$）．ここで $\boldsymbol{M}$, $\boldsymbol{U}$, $\boldsymbol{V}$ というそれぞれのサイズが $n\times p$, $n\times n$, $p\times p$ となる行列を考える．ここで $\boldsymbol{U}$ および $\boldsymbol{V}$ は正定値行列とする．確率変数である $\boldsymbol{X}$ が次式で表されるモーメント母関数を持つとき，$\boldsymbol{X}$ は行列変量の正規分布に従うという．

$$M_X(\boldsymbol{T})=\exp\left\{\mathrm{tr}\left(\boldsymbol{M}^\top\boldsymbol{T}\right)+\frac{1}{2}\mathrm{tr}\left(\boldsymbol{T}^\top\boldsymbol{U}\boldsymbol{T}\boldsymbol{V}\right)\right\} \quad (10)$$

ここで $\boldsymbol{T}$ は $n\times p$ となる行列である．以下本稿ではこの分布を

$$\boldsymbol{X}\sim\mathcal{N}_{\mathrm{mv}}(\boldsymbol{X};\boldsymbol{M},\boldsymbol{U},\boldsymbol{V}) \quad (11)$$

で表すものとする．ベクトル空間における正規分布（ガウス分布）と上記の行列変量正規分布の対応関係については，クロネッカー積とベクトル化演算子を導入する事で導く事ができる．式 (11) は，ベクトル $\mathrm{vec}(\boldsymbol{X})$ に対する次式の確率密度関数と等価である．

$$P(\mathrm{vec}(\boldsymbol{X})|\boldsymbol{\lambda})=\mathcal{N}(\mathrm{vec}(\boldsymbol{X});\mathrm{vec}(\boldsymbol{M}),\boldsymbol{V}\otimes\boldsymbol{U}) \quad (12)$$

ここで vec() は，行列を列ベクトルに展開する演算子であり，$\boldsymbol{\lambda}$ はモデルパラメータを表す．最終的に行列変量 $\boldsymbol{X}$ の確率密度関数は次式のようにあらわされる．

$$P(\boldsymbol{X}|\boldsymbol{\lambda})=c^{-1}\exp\left[-\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{U}^{-1}(\boldsymbol{X}-\boldsymbol{M})\,\boldsymbol{V}^{-1}(\boldsymbol{X}-\boldsymbol{M})\right\}\right]$$
$$\text{where } c=(2\pi)^{(1/2)np}|\boldsymbol{U}|^{(1/2)p}|\boldsymbol{V}|^{(1/2)n} \quad (13)$$

いま，それぞれのサンプルが行列で表されるデータ群 $\mathcal{X}=[\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots,\boldsymbol{X}_T]$ が，式 (13) に基づく確率分布から生成された場合に，モデルパラメータである $\boldsymbol{M}$, $\boldsymbol{U}$ および $\boldsymbol{V}$ の最尤推定を考える．これらの最尤推定値は以下のように表される．

$$\hat{\boldsymbol{M}}=\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{X}_t \quad (14)$$

$$\hat{\boldsymbol{U}}=\frac{1}{pT}\sum_{t=1}^{T}(\boldsymbol{X}_t-\hat{\boldsymbol{M}})\hat{\boldsymbol{V}}^{-1}(\boldsymbol{X}_t-\hat{\boldsymbol{M}})^\top \quad (15)$$

$$\hat{\boldsymbol{V}}=\frac{1}{nT}\sum_{t=1}^{T}(\boldsymbol{X}_t-\hat{\boldsymbol{M}})^\top\hat{\boldsymbol{U}}^{-1}(\boldsymbol{X}_t-\hat{\boldsymbol{M}}) \quad (16)$$

式 (12) に着目すると，行列変量正規分布は，自由な構造の分散共分散行列をもつガウス分布とは異なり，クロネッカー積に基づいて一定の構造を規定した確率分布であると解釈できる．このクロネッカー積による「分離可能な」構造によって，分散共分散行列は，観測された行列の行および列に対して，明示的に異なる特性を与える事ができる．すなわち分散共分散行列を規定するモデルパラメータである $\boldsymbol{U}$ および $\boldsymbol{V}$ は，それぞれ行および列方向に対する分散共分散構造を表している．加えて式 (15)(16) に着目すると，パラメータ推定について一つの利点が見えてくる．上述の最尤推定において，推定に用いるサンプル数は $T$ であるが，$\boldsymbol{U}$ および $\boldsymbol{V}$ を推定する際の実効的なサンプル数が，それぞれ $pT$ および $nT$ となっていることがわかる．すなわち，行列変量を導入した統計的モデリングによってより効率的で精緻な推定が可能になる事が期待される．

### 3.2 行列変量ガウス混合モデルに基づく声質変換

ベクトル変量において，単一のガウス分布を混合ガウスモデルに拡張する場合と同様に，行列変量正規分布についても混合モデルを構築する事が可能である [11]．本稿では以降これを行列変量ガウス混合モデル（matrix variate Gaussian mixture mode; MV-GMM）と呼ぶ．本節では MV-GMM に基づく同時分布のモデル化による声質変換の

© 2014 Information Processing Society of Japan

枠組みについて論じる．結合ベクトルの GMM に基づく声質変換と同様に，入力話者の発話を表す特徴量系列を $\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n_x}]$，同一発話内容の出力話者の特徴量系列を $\boldsymbol{y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{n_y}]$ とする．提案法においては，動的計画法に基づくフレーム間の対応付けを行ったのちに，新しい特徴量系列を，結合行列 $\boldsymbol{Z}_t = [\boldsymbol{x}_t, \boldsymbol{y}_t] \in \mathcal{R}^{D\times S}$ の系列 $\boldsymbol{Z} = [\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_n]$ として考える．ここで $S$ は話者数を表し，ここでは $S = 2$ である．なお，提案法において，従来の結合ベクトルに基づくアプローチと異なり，複数話者を用いたモデル学習へと容易に拡張する事が可能であり，その場合 $S > 2$ となる．これについては次節で述べる．ここで，MV-GMM によって結合行列 $\boldsymbol{Z}_t$ をモデル化した場合の結合確率密度は以下のように表される．

$$P(\boldsymbol{Z}_t|\boldsymbol{\lambda}^{(Z)}) = \sum_{m=1}^{M} w_m \mathcal{N}_{\mathrm{mv}}(\boldsymbol{Z}_t; \boldsymbol{M}_m, \boldsymbol{U}_m, \boldsymbol{V}_m) \quad (17)$$

式 (17) より，結合確率密度は要素分布となる個々の行列変量正規分布の重み付き和で表現されることがわかる．ここで式 (17) における表記は各要素分布の平均に相当するパラメータ $\boldsymbol{M}_m$ が行列である点と，二つの行列 $\boldsymbol{U}_m$ および $\boldsymbol{V}_m$ を除いて，式 (1) と同様である．行列 $\boldsymbol{U}_m \in \mathcal{R}^{D\times D}$ は，第 $m$ 番目の要素分布において，特徴量空間の分散構造を表現する分散共分散行列であり，行列 $\boldsymbol{V}_m \in \mathcal{R}^{S\times S}$ は，入出力話者の間の相関関係を表現している分散共分散行列である．これらのモデルパラメータの推定は EM アルゴリズムを用いる事ができ，各更新式は式 (14)–(16) に類似した，以下の形で表される．

$$\gamma_{m,t} = \frac{w_m \mathcal{N}_{\mathrm{mv}}(\boldsymbol{Z}_t; \boldsymbol{M}_m, \boldsymbol{U}_m, \boldsymbol{V}_m)}{\sum_{m=1}^{M} w_m \mathcal{N}_{\mathrm{mv}}(\boldsymbol{Z}_t; \boldsymbol{M}_m, \boldsymbol{U}_m, \boldsymbol{V}_m)} \quad (18)$$

$$\hat{\boldsymbol{M}}_m = \frac{1}{T_m} \sum_{t=1}^{T} \gamma_{m,t} \boldsymbol{Z}_t \quad (19)$$

$$\hat{\boldsymbol{U}}_m = \frac{1}{S T_m} \sum_{t=1}^{T} \gamma_{m,t} (\boldsymbol{Z}_t - \hat{\boldsymbol{M}}_m) \hat{\boldsymbol{V}}_m^{-1} (\boldsymbol{Z}_t - \hat{\boldsymbol{M}}_m)^\top \quad (20)$$

$$\hat{\boldsymbol{V}}_m = \frac{1}{D T_m} \sum_{t=1}^{T} \gamma_{m,t} (\boldsymbol{Z}_t - \hat{\boldsymbol{M}}_m)^\top \hat{\boldsymbol{U}}_m^{-1} (\boldsymbol{Z}_t - \hat{\boldsymbol{M}}_m) \quad (21)$$

$$T_m = \sum_{t=1}^{T} \gamma_{m,t} \quad (22)$$

ここで $T_m$ は $m$ 番目の要素分布に対応する実効的なサンプル数を表している．式 (20) および (21) をみると，単一の行列変量正規分布の場合と同様に，分散共分散行列の最尤推定に用いる実効的なサンプル数が増加していると解釈でき，効率的な推定が実現されていると考えられる．

同様に MV-GMM でモデル化された同時確率から，条件付き確率 $P(\boldsymbol{y}_t|\boldsymbol{x}_t)$ に基づく変換関数を導出する事ができる．MV-GMM は明示的に分離可能な分散共分散構造を有しているため，MV-GMM の $m$ 番目の要素分布から導出される条件付き確率は，以下のようにより単純な形で表現される．

$$P(\boldsymbol{y}_t|\boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(Z)}) = \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{E}_{m,t}^{(y)}, \boldsymbol{D}_m^{(y)}) \quad (23)$$

$$\boldsymbol{E}_{m,t}^{(y)} = \boldsymbol{\mu}_m^{(y)} + \frac{v_m^{(yx)}}{v_m^{(xx)}} (\boldsymbol{x}_t - \boldsymbol{\mu}_m^{(x)}) \quad (24)$$

$$\boldsymbol{D}_m^{(y)} = \left( v_m^{(yy)} - \frac{v_m^{(yx)} v_m^{(xy)}}{v_m^{(xx)}} \right) \boldsymbol{U}_m \quad (25)$$

ここで，$\boldsymbol{M}_m = [\boldsymbol{\mu}_m^{(x)}, \boldsymbol{\mu}_m^{(y)}]$ とし，$v_m^{(\cdot)}$ は行列 $\boldsymbol{V}_m$ における対応する要素である．式 (24) および (25) から，$m$ 番目の要素分布に対応する変換関数が，話者の相関関係を表す行列 $\boldsymbol{V}_m$ のパラメータから導出されることがわかる．式 (6) および (7) と比較すると，MV-GMM に基づくパラメータ生成は逆行列演算を必要としないため，より高速に実行可能である事がわかる．

### 3.3 複数話者を用いたモデル学習

前節で述べた通り，提案法による MV-GMM を用いたモデリングでは，結合特徴量空間の分散共分散構造を明示的に分離して捉えているため，結合特徴量空間の拡張が容易であるという利点がある．例えば，特徴量となるベクトルを入出力とは異なる別の話者から抽出し，これを列ベクトルとして結合行列に加える事で多人数話者によるモデル学習を実現できる．従来の結合ベクトルに基づく手法でこれを行う場合，複数の特徴ベクトルを連結し高次元の結合特徴量ベクトルを構築する必要がある．このようなモデル化による複数話者を用いたモデル学習では，過学習の問題が顕著となる．一方，提案法においては明示的に分離された構造と式 (20) のような効率的な学習によって，追加した話者の特徴量を利用して，本来の入出力話者間の特徴量空間をより精緻にモデル化し，結果的に変換性能の向上につながる可能性がある．これは分散共分散構造に一種の制約を持たせた話者正規化学習と解釈可能である [12], [13].

## 4. 声質変換実験による評価

### 4.1 実験条件

提案法に基づく声質変換性能の評価と複数話者を用いたモデル学習の効果について検証するため，2 種類の声質変換実験を実施した．本実験の目的は 2 つあり，1 つは MV-GMM に基づく提案手法が従来手法に比べて，より効果的に特徴量空間の特性および入出力空間の関係性をモデル化しているかの検証，もう一つは入出力の対象話者とは異なる話者を学習に加えた場合の効果についての検証である．

一つ目の実験においては，CMU ARCTIC データベース [14] の二人の男性話者のデータ（bdl および rms）を用いて声質変換実験を行った．話者 bdl を入力話者，話者 rms を出力話者とした．モデルの学習には a0001 から a0256 までの 256 文を学習データとして用い，a0544 から a0593 ま

© 2014 Information Processing Society of Japan

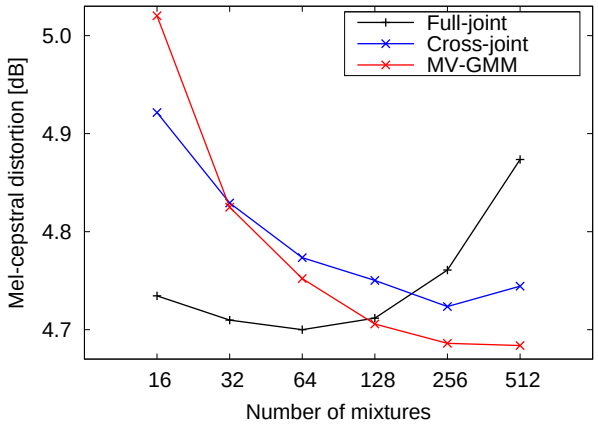


**図 1** Results of objective evaluations by mel-cepstral distortion (MCD).

での 50 文を評価セットとして選択した．

二つ目の実験について，音声データとして，ATR 日本語音声データベース [15] の B セットから 3 名の男性話者を選択した（MHT, MMY, MSH）．このうち話者 MHT を入力話者，話者 MMY を出力話者とし，話者 MSH を複数話者を用いたモデル学習における追加話者とした．モデル学習に際して，サブセット A から E までの 250 文を用い，評価セットとしてサブセット J の 53 文を選択した．この実験においては，まず MHT と MMY の間で動的計画法に基づくアラインメントを行った．その後，このアラインメント情報に基づき，2 話者の間の平均的な特徴量系列を算出した．最後にこの平均的な特徴量系列と追加話者 MSH の間で動的計画法に基づくアラインメントを算出する事で，提案法における結合行列を構築した．

全ての実験において，スペクトル特徴量として，STRAIGHT 分析に基づくスペクトルから得られた 24 次のメルケプストラムを用いた（$D$=24）[16]．パラメータ生成手法として，最小平均二乗誤差基準ではなく，式 (9) の条件付き最尤基準に基づく変換を行った．ただしパラメータの動的特徴については今回の実験では考慮しなかった [9]．

CMU ARCTIC を用いた実験では，1) 結合ベクトルに基づく手法において全共分散構造の分散共分散構造を用いたもの （Full-joint），2) 結合ベクトルに基づく手法において分散共分散行列 $\mathbf{\Sigma}_m^{(xx)}$ 及び $\mathbf{\Sigma}_m^{(yy)}$ と相互共分散行列 $\mathbf{\Sigma}_m^{(xy)}$ 及び $\mathbf{\Sigma}_m^{(yx)}$ を対角行列としたもの（Cross-joint），3) 提案法 （MV-GMM） の 3 つの手法を比較した．提案法において，$\boldsymbol{U}_m$ および $\boldsymbol{V}_m$ は全共分散行列とした．混合モデルの要素数（$M$）については，16 から 512 まで変化させた．

ATR 日本語音声データベース を用いた複数話者学習に関する実験においては，MV-GMM において 1) 入出力話者の結合行列のみを用いた場合（$S = 2$）と 2) 入出力話者に加えて追加話者の特徴量を結合した結合行列を用いた場合（$S = 3$）の 2 つの手法を評価した．この実験では混合モデルの要素数（$M$）について，256 とした．

**表 1** Results of objective evaluations by MCD in the optimal conditions. The optimal numbers of mixture components were selected. # of parameters means the number of variance-covariance parameters which should be estimated in the model.

| | MCD [dB] | $M$ | # of parameters |
|---|---|---|---|
| Full-joint | 4.70 | 64 | 75264 |
| Cross-joint | 4.72 | 256 | 18432 |
| MV-GMM | 4.68 | 512 | 155136 |

**表 2** Results of objective evaluations by MCD when model training using multiple speaker is applied.

| | MCD [dB] |
|---|---|
| $S = 2$ | 4.643 |
| $S = 3$ | 4.635 |

### 4.2 客観評価

声質変換の性能について，変換された特徴量と出力対象話者の特徴量との間で，次式で定義されるメルケプストラム歪みを用いて評価した．

$$\text{Mel-CD [dB]} = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{24}\left(mc_d^{(y)} - \hat{mc}_d^{(y)}\right)^2} \quad (26)$$

客観評価の結果を図 1 に示す．混合数が少ない場合においては，分散共分散構造に制約のない結合ベクトルに基づく手法（Full-joint）が最良の結果となっている．一方で混合数が 64 をこえると，Full-joint の性能が大きく劣化していることがわかる．これは制約を持たない分散共分散構造によって，モデルが複雑になった場合に過学習をおこしている事が考えられる．MV-GMM に基づく提案法と対角行列で制約された手法（Cross-joint）を比較した場合，混合数の増加に対して類似した傾向を示した．混合数が 32 をこえた場合について，提案法の性能は Cross-joint の性能を上回っている．これは式 (20) の通り，提案法によって特徴量空間の特性が入出力話者双方の特徴量を効果的に使ってモデル化されているためと考えられる．また Full-joint と比較しても，最適な混合数の条件において，若干の性能改善が見られた．これは行列変量に基づく制約によって，声質変換に必要な「特徴量空間の精緻なモデル化」と「入出力空間の関係性のモデル化」が効果的に実現された結果と考えられる．

表 1 に最適な混合数における客観評価の結果を示す．表 1 から，提案法は最適条件において，分散共分散構造の表現に，最も多くのモデルパラメータを有している．それにも関わらず，提案法は過学習を抑制し，効果的にモデルパラメータを推定できていることがわかる．これは MV-GMM が効率的かつ精緻なパラメータ推定を実現していることを意味している．

### 4.3 複数話者を用いた学習の効果

表 2 は，学習に用いた話者数を変化させた場合の MV-GMM の性能の変化を示している．表 2 から，追加話者の

© 2014 Information Processing Society of Japan

特徴量を結合して学習した場合に，若干ながら性能向上が確認された．ここで追加された特徴量は入出力いずれの話者の特徴量でもない．すなわち提案法における学習では追加された特徴量を効果的に利用して特徴量空間を学習していると考えられる．その点から提案法は効率的なパラメータ共有を行い，一種の話者正規化学習を実現しているとも解釈できる．

## 5. おわりに

本稿では，行列変量ガウス混合モデルを用いた声質変換の枠組みについて提案した．提案法においては，行列の行方向および列方向の関係性を明示的にモデル化した分散共分散行列を用いる事で，特徴量空間の特性の精緻なモデル化と入出力空間の関係性の適切なモデル化を効果的に実現可能である．加えて提案法において，入出力話者とは異なる話者の追加データによって，変換性能が向上する可能性を示した．今後の課題として，主観評価実験によって提案法の評価を行う必要がある．加えて行列変量による枠組みは時間方向のモデル化についても効果的と考えられるため，長時間特徴量や動的特徴量の行列変量によるモデル化は興味深い方向性である．またその他のパラメータ共有手法と提案手法との統合なども検討課題として挙げられる．

**謝辞** 本研究は科研費・若手研究 (B) (25730105) の助成を受けたものである．

### 参考文献

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, “Voice conversion through vector quantization,” Proc. ICASSP, pp. 655–658, 1988.
[2] L. Deng, A. Acero, L. Jiang, J. Droppo, and X. Huang, “High-performance robust speech recognition using stereo training data,” Proc. ICASSP, pp. 301–304, 2001.
[3] A. Kain and M. W. Macon, “Spectral voice conversion for text-to-speech synthesis,” Proc. ICASSP, vol. 1, pp. 285–288, 1998.
[4] Y. Stylianou, O. Cappé, and E. Moulines, “Continuous probabilistic transform for voice conversion,” IEEE Trans. on Speech and Audio Processing, vol. 6, no. 2, pp. 131–142, 1998.
[5] A. Mouchtaris, J. V. der Spiegel, and P. Mueller, “Non-parallel training for voice conversion based on a parameter adaptation approach,” IEEE Trans. on Audio, Speech, and Language Processing, vol. 14, no. 3, pp. 952–963, 2006.
[6] C. H. Lee and C. H. Wu, “Map-based adaptation for speech conversion using adaptation data selection and non-parallel training,” Proc. INTERSPEECH, pp. 2254–2257, 2006.
[7] D. Saito, S. Watanabe, A. Nakamura and N. Minematsu, “Statistical voice conversion based on noisy channel model,” IEEE Trans. on Audio, Speech, and Language Processing, vol. 20, no. 6, pp. 1784–1794, 2012.
[8] D. Saito, K. Yamamoto, N. Minematsu, and K. Hirose, “One-to-many voice conversion based on tensor representation of speaker space,” Proc. INTERSPEECH, pp. 653–656, 2011.
[9] T. Toda, A. W. Black, and K. Tokuda, “Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory,” IEEE Trans. on Audio, Speech, and Language Processing, vol. 15, no. 8, pp. 2222–2235, 2007.
[10] P. Dutilleul, “The MLE algorithm for the matrix normal distribution,” Journal of Statistical Computation and Simulation, vol. 64, pp. 105–123, 1999.
[11] C. Viroli, “Finite mixture of matrix normal distributions for classifying three-wya data,” Journal of Statistics and Computing, vol. 21, Issue 4, pp. 511–522, 2011.
[12] T. Anastasakos, J. McDonough, R. Schwarts and J. Makhoul, “A compact model for speaker adaptive training,” Proc. ICSLP, vol. 2, pp. 1137–1140, 1996.
[13] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, “Speaker adaptive training for one-to-many eigenvoice conversion based on Gaussian mixture model,” Proc. INTERSPEECH, pp. 1981–1984, 2007.
[14] J. Kominek and A. W. Black, “CMU ARCTIC Databases for Speech Synthesis,” Lang. Technol. Inst., Carnegie Mellon Univ., Pittsburgh, PA. 2003 [Online]. Available: http://festvox.org/cmu_arctic/index.html
[15] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K.Shikano, “ATR Japanese speech database as a tool of speech recognition and synthesis,” Speech Communication, vol.9, pp.357–363, 1990.
[16] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigné, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,” Speech Communication, vol.27, pp.187–207, 1999.

© 2014 Information Processing Society of Japan