テンソル表現に基づく任意話者声質変換における話者正規化学習の検討*

○齋藤大輔, 峯松信明, 広瀬啓吉 (東大)

1 はじめに

本稿では，任意話者声質変換における話者正規化学習の導入について議論する．特にテンソル表現に基づいて話者空間を構築した場合の話者正規化学習について提案し，その効果を検証する．

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．特に入力発声の話者性を制御する話者変換はテキスト音声合成への応用を目的として数多く研究されている [1, 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，その中でも混合正規分布モデル（GMM）に基づく変換法はその柔軟性から近年主流となっている [1, 3]．

しかし，変換モデルの構築の際には，基本的に同一発話内容の入出力音声対からなるパラレルデータを用いる必要がある．また，変換モデルの利用は学習時の入出力話者対に限定される．すなわち話者性を柔軟に制御することは声質変換における重要な課題といえる（任意話者声質変換）．任意話者声質変換の実現として，音声認識における話者適応に基づくいくつかの手法がある [4, 5]．そのなかでも，話者空間の構築に基づく固有声変換法（Eigenvoice conversion; EVC）が提案されている [5]．複数のパラレルデータより得られた結合確率密度分布から，事前収録話者の話者 GMM をそれぞれ抽出し，ガウス分布の平均ベクトルを連結した GMM スーパーベクトルを用いて話者空間を構築する．話者認識の場合と同様に，任意の話者はこの話者空間の一点で表され，基底に対する少数の重みパラメータを推定することで，話者性を柔軟に制御することができる．よって変換性能の向上には，精緻な話者空間の構築が重要である．しかし，GMM スーパーベクトルによる話者空間表現は，複数要因からの音響的な変動を一つの特徴量空間に含んでいる．このため，EVC において適応データ量に対する拡張性は限定的である．

我々はこの問題の解決として，テンソル解析を基盤とした話者空間表現を提案した [6]．この手法では，任意の話者はスーパーベクトルではなく，行および列がそれぞれ GMM の要素と平均ベクトルに対応する行列の形で表現される．このような話者表現を用いることで事前収録話者のデータセットが 3 階のテンソルで表現でき，テンソル解析の導入により話者空間を構築することができる．上述の表現に基づいた，テンソル表現に基づく任意話者声質変換（Tensor-based Arbitrary Speaker Conversion; TASC）の有効性が一対多声質変換において示されている [6]．

テンソル解析に基づく話者表現は，様々な手法と統合的に用いることが可能である．本稿では，テンソル表現に基づく任意話者声質変換と話者正規化学習を統合し，その有効性を検証する．話者正規化学習は，規範的な話者非依存モデルを学習可能であり [7]，任意話者声質変換における有効性が示されている [8]．テンソルに基づく話者空間表現と話者正規化学習を併せることで，より柔軟で精緻な話者変換の実現が期待される．

2 固有声変換法（EVC）

本章では，一対多 EVC について概説する．今，参照話者の音響特徴量を $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \boldsymbol{\Delta x}_t]^\top$，$s$ 番目の事前収録話者の音響特徴量を $\boldsymbol{Y}_t^{(s)} = [\boldsymbol{y}_t^{(s)^\top}, \boldsymbol{\Delta y}_t^{(s)^\top}]^\top$ と表す．ただし $^\top$ は転置を表す．ここで，音響特徴量は D 次元の静的および動的特徴量を結合した $2D$ 次元の音響特徴量となる．参照話者と事前収録話者の結合確率密度は，EV-GMM として以下のようにモデル化される．

$$P(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}|\boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}^{(s)}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}([\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^{(s)^\top}]^\top; \boldsymbol{\mu}_m^{(Z)}(\boldsymbol{w}^{(s)}), \boldsymbol{\Sigma}_m^{(Z)}) \quad (1)$$

$$\boldsymbol{\mu}_m^{(Z)}(\boldsymbol{w}^{(s)}) = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{B}_m \boldsymbol{w}^{(s)} + \boldsymbol{b}_m^{(0)} \end{bmatrix}, \boldsymbol{\Sigma}_m^{(Z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix} \quad (2)$$

ここで $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ は，平均ベクトルを $\boldsymbol{\mu}$，分散共分散行列を $\boldsymbol{\Sigma}$ とする正規分布を表す．m 番目の要素の重みは α_m で表し，混合数を M とする．EV-GMM では，S 人の事前収録話者を利用して，出力話者の平均ベクトル $\boldsymbol{\mu}_m^{(Y)}$ をバイアスベクトル $\boldsymbol{b}_m^{(0)}$ と $J(< S)$ 個の表現ベクトルの線形結合で表す．このとき，出力話者の話者性は J 次元の重みベクトル $\boldsymbol{w}^{(s)}$ で制御できる．すなわち話者空間が J 個の基底スーパーベクトル $\boldsymbol{B} = [\boldsymbol{B}_1^\top, \boldsymbol{B}_2^\top, \ldots, \boldsymbol{B}_m^\top]^\top \in \mathcal{R}^{2DM \times J}$ とバイアススーパーベクトル $\boldsymbol{b} = \left[\boldsymbol{b}_1^{(0)^\top}, \boldsymbol{b}_2^{(0)^\top}, \ldots, \boldsymbol{b}_m^{(0)^\top}\right]^\top \in \mathcal{R}^{2DM \times 1}$ によって構築される．話者空間は主成分分析（PCA）に基づいて以下のように構築される．最初に出力話者非依存の GMM（TI-GMM）を，全ての参照話者と事前収録話者とのパラレルデータを用いて学習する．次に，対応するパラレルデータを用いて TI-GMM の出力話者の平均ベクトルを更新するこ

* “Speaker adaptive training for tensor-based arbitrary speaker conversion”
by SAITO Daisuke, MINEMATSU Nobuaki, and HIROSE Keikichi (The Univ. of Tokyo)

とで，話者依存のモデルを得る．話者空間の特徴量ベクトルとして，事前収録話者の GMM の各要素の平均ベクトルを連結し，スーパーベクトルを生成する．得られたスーパーベクトルを用いて PCA を行うことで，バイアスベクトル $\boldsymbol{b}$ と表現ベクトル $\boldsymbol{B}$ を得ることができる．一方，任意話者声質変換は，出力話者の音響特徴量系列 $\boldsymbol{Y}^{(tar)}$ を用いて，以下に示す最尤基準に基づいて重みベクトル $\boldsymbol{w}$ を推定することで実現できる [5].

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmax}} \int P(\boldsymbol{X}, \boldsymbol{Y}^{(tar)} | \boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}) d\boldsymbol{X} \quad (3)$$

3 テンソル表現に基づく話者空間表現

本章では，テンソル解析に基づく話者空間の構築について述べる [6]．EVC では，GMM スーパーベクトルを話者の表現として用いる．しかし GMM スーパーベクトルによる話者表現は，複数の音響的変動要因を内包する表現であり，非常に高次元の話者空間を構成し，柔軟な制御を抑制する一因となる．この問題を解決するため，本研究では，各話者を行および列がそれぞれ GMM の要素と平均ベクトルに対応するような行列の形で表現する．EVC では PCA を用いて GMM スーパーベクトルのセットから話者空間を構築するのに対して，本研究では話者行列のセットに対して，Tucker 分解を適用してこれを実現する [11]. Tucker 分解は，行列代数における特異値分解の拡張と解釈でき，複数要因からの変動を適切に扱うことができる．

任意話者声質変換において，Tucker 分解に基づいて話者空間を構築するため，各事前収録話者を $M \times D'$ の行列で表現する [12]．ここで M は混合数であり，$D' = 2D$ とする．まずはじめに全データ行列の平均を求め，バイアス行列 $\boldsymbol{b}' = \left[\boldsymbol{b}_1^{(0)}, \boldsymbol{b}_2^{(0)}, \ldots, \boldsymbol{b}_m^{(0)}\right]^\top$ とする．これを各事前収録話者の行列からあらかじめ減算しておく．ここで事前収録話者の話者数を S とすると，話者空間を構築するデータセットは 3 階のテンソル $\mathcal{M} \in \mathcal{R}^{M \times D' \times S}$ で表される．$\mathcal{M}$ を Tucker 分解することによって，基底行列として $\boldsymbol{U}^{(M)} \in \mathcal{R}^{M \times M}$, $\boldsymbol{U}^{(D')} \in \mathcal{R}^{D' \times D'}$, $\boldsymbol{U}^{(S)} \in \mathcal{R}^{S \times S}$ を抽出する．これらの行列は，それぞれ GMM の混合要素，平均ベクトルの次元，話者インデックスの効果を捉えている．話者空間の基底として複数の候補が考えられるが，本研究では GMM の混合要素の関係性に着眼し，[12] と同様に $\boldsymbol{U}^{(M)}$ を基底として用いる．最終的に縮約された基底行列を用いて，任意話者を表現する話者行列を以下のように表す．

$$\boldsymbol{\mu}^{(new)} = \boldsymbol{U}^{(M)} \boldsymbol{W}_{(new)}^\top + \boldsymbol{b}' \quad (4)$$

$\boldsymbol{U}^{(M)} \in \mathcal{R}^{M \times K} (K \leq S)$, $\boldsymbol{W}_{(new)} \in \mathcal{R}^{D' \times K}$ はそれぞれ，表現行列および重み行列となる．ゆえに提案法では，J 次元の重みベクトルを推定する EVC と異なり，$D' \times K$ の重み行列を推定することになる．

任意話者の重み行列は，適応データ $\boldsymbol{Y}^{(tar)}$ を用いて，最尤基準を導入し，以下の更新式により推定する．

$$\mathrm{vec}(\boldsymbol{W}) = \left(\sum_{m=1}^{M} \overline{\gamma}_m^{(tar)} \boldsymbol{U}_m^\top \boldsymbol{U}_m \otimes \boldsymbol{\Sigma}_m^{(YY)^{-1}} \right)^{-1} \mathrm{vec}(\boldsymbol{C}) \quad (5)$$

$$\boldsymbol{C} = \sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{\Sigma}_m^{(YY)^{-1}} (\boldsymbol{Y}_t^{(tar)} - \boldsymbol{b}_m^{(0)}) \boldsymbol{U}_m \quad (6)$$

$$\boldsymbol{U}_m = \boldsymbol{U}^{(M)}(m, :) \in \mathcal{R}^{1 \times K} \quad (7)$$

$$\gamma_{m,t} = P(m | \boldsymbol{Y}_t^{(tar)}, \boldsymbol{\lambda}, \boldsymbol{W}), \overline{\gamma}_m^{(tar)} = \sum_{t=1}^{T} \gamma_{m,t} \quad (8)$$

vec() は行列を列ベクトルに展開する演算子である．

4 話者正規化学習の導入

本章では，テンソル表現に基づく任意話者変換のための話者正規化学習について述べる．話者正規化学習をテンソル解析に基づく話者表現に適用することで，より柔軟で精緻な声質変換モデルの構築が期待できる．

EVC における話者正規化学習と同様に [8]，共有パラメータはすべての事前学習話者に対するモデルの尤度を最大化する基準で推定する．

$$\hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\mathcal{W}}}_1^S) = \underset{\lambda, \mathcal{W}_1^S}{\operatorname{argmax}} \prod_{s=1}^{S} \prod_{t_s=1}^{T_s} P(\boldsymbol{Z}_{t_s}^{(s)} | \boldsymbol{\lambda}(\boldsymbol{W}_s)), \quad (9)$$

ここで，$\boldsymbol{Z}_{t_s}^{(s)} = [\boldsymbol{X}_{t_s}^\top, \boldsymbol{Y}_{t_s}^{(s)^\top}]^\top$ であり，$\boldsymbol{\lambda}(\boldsymbol{W}_s)$ は重み行列 $\boldsymbol{W}_s$ で表される s 番目の事前学習話者に適応された変換モデルである．$\boldsymbol{\mathcal{W}}_1^S$ はすべての事前学習話者の重み行列 $(\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W_S})$ を集めたテンソル表現である．話者正規化学習では最尤基準に基づいて，共有パラメータと $\boldsymbol{\mathcal{W}}_1^S$ が推定される．これを実現するため，以下の補助関数を導入する．

$$Q(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}}) = \sum_{s=1}^{S} \sum_{m=1}^{M} \overline{\gamma}_m^{(s)} \log P\Big(\boldsymbol{Z}^{(s)}, m | \hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{W}}_s)\Big) \quad (10)$$

$$\gamma_{m,t_s}^{(s)} = P\left(m | \boldsymbol{Z}_{t_s}^{(s)}, \boldsymbol{\lambda}(\boldsymbol{W}_s)\right), \overline{\gamma}_m^{(s)} = \sum_{t_s=1}^{T_s} \gamma_{m,t_s}^{(s)} \quad (11)$$

[8] で言及されているように，すべてのパラメータを上記の補助関数に基づいて更新するのは，パラメータの相互依存性から困難である．よって，以下のように逐次的にパラメータを更新する．

1. 現在の共有パラメータと式 (11) を用いて，$\gamma_{m,t_s}^{(s)}$ 及び $\overline{\gamma}_m^{(s)}$ を求める．
2. $\gamma_{m,t_s}^{(s)}$ 及び $\overline{\gamma}_m^{(s)}$ と現在の共有パラメータを用いて，個々の事前学習話者を表現する重み行列 $\hat{\boldsymbol{W}}_s$ を更新する．

3. 前述の結果を用いて，共有パラメータのうち混合重み $\hat{\alpha}_m$ と基底行列 $\hat{\boldsymbol{U}}^{(M)}$ を更新する．

4. 前記のステップで更新されたパラメータを用いて，共有された分散共分散行列 $\hat{\boldsymbol{\Sigma}}_m^{(ZZ)}$ を更新する．

5. 上記 1. から 4. の手順を，一定の回数繰り返す．

ここで，個々の更新手順において全学習データに対する適応モデルの尤度は単調に増加する．

2. において，更新後の重み行列 $\hat{\boldsymbol{W}}_s$ は以下のように表される．

$$\mathrm{vec}(\hat{\boldsymbol{W}}_s)=\left[\sum_{m=1}^{M}\overline{\gamma}_m^{(s)}\boldsymbol{U}_m^{\top}\boldsymbol{U}_m\otimes\boldsymbol{P}_m^{(YY)}\right]^{-1}\mathrm{vec}\left[\sum_{m=1}^{M}\boldsymbol{C}'_m\boldsymbol{U}_m\right] \quad (12)$$

$$\boldsymbol{C}'_m=\boldsymbol{P}_m^{(YX)}(\overline{\boldsymbol{X}}^{(s)}-\overline{\gamma}_m^{(s)}\boldsymbol{\mu}_m^{(X)})+\boldsymbol{P}_m^{(YY)}(\overline{\boldsymbol{Y}}^{(s)}-\overline{\gamma}_m^{(s)}\boldsymbol{b}_m^{(0)}) \quad (13)$$

$$\overline{\boldsymbol{Z}}_m^{(s)}=\begin{bmatrix}\overline{\boldsymbol{X}}_m^{(s)}\\ \overline{\boldsymbol{Y}}_m^{(s)}\end{bmatrix}=\begin{bmatrix}\sum_{t_s=1}^{T_s}\gamma_{i,t_s}^{(s)}\boldsymbol{X}_{t_s}^{(s)}\\ \sum_{t_s=1}^{T_s}\gamma_{i,t_s}^{(s)}\boldsymbol{Y}_{t_s}^{(s)}\end{bmatrix} \quad (14)$$

$$\boldsymbol{\Sigma}_m^{(ZZ)-1}=\begin{bmatrix}\boldsymbol{P}_m^{(XX)}&\boldsymbol{P}_m^{(XY)}\\ \boldsymbol{P}_m^{(YX)}&\boldsymbol{P}_m^{(YY)}\end{bmatrix}\equiv\boldsymbol{P}_m^{(ZZ)} \quad (15)$$

3. 及び 4. において，共有パラメータは以下のように更新される．

$$\hat{\alpha}_m=\frac{\sum_{s=1}^{S}\overline{\gamma}_m^{(s)}}{\sum_{m=1}^{M}\sum_{s=1}^{S}\overline{\gamma}_m^{(s)}} \quad (16)$$

$$\hat{\boldsymbol{u}}_m=\left(\sum_{s=1}^{S}\overline{\gamma}_m^{(s)}\hat{\boldsymbol{E}}_s^{\top}\boldsymbol{P}_m^{(ZZ)}\hat{\boldsymbol{E}}_s\right)^{-1}\left(\sum_{s=1}^{S}\hat{\boldsymbol{E}}_s\boldsymbol{P}_m^{(ZZ)}\overline{\boldsymbol{Z}}_m^{(s)}\right) \quad (17)$$

$$\boldsymbol{\Sigma}_m^{(ZZ)}=\frac{1}{\sum_{s=1}^{S}\overline{\gamma}_m^{(s)}}\sum_{s=1}^{S}\left\{\overline{\boldsymbol{V}}_m^{(s)}+\overline{\gamma}_m^{(s)}\hat{\boldsymbol{\mu}}^{(s)}\hat{\boldsymbol{\mu}}^{(s)\top}-\left(\hat{\boldsymbol{\mu}}_m^{(s)}\overline{\boldsymbol{Z}}_m^{(s)\top}+\overline{\boldsymbol{Z}}_m^{(s)\top}\hat{\boldsymbol{\mu}}_m^{(s)}\right)\right\} \quad (18)$$

$$\overline{\boldsymbol{V}}_m^{(s)}=\sum_{t_s=1}^{T}\gamma_{m,t_s}^{(s)}\boldsymbol{Z}_{t_s}^{(s)}\boldsymbol{Z}_{t_s}^{(s)\top} \quad (19)$$

$$\hat{\boldsymbol{\mu}}_m^{(s)}=\hat{\boldsymbol{E}}_s\hat{\boldsymbol{u}}_m=\begin{bmatrix}\hat{\boldsymbol{\mu}}_m^{(X)}\\ \hat{\boldsymbol{W}}_s\hat{\boldsymbol{U}}_m^{\top}+\hat{\boldsymbol{b}}_m^{(0)}\end{bmatrix} \quad (20)$$

$$\hat{\boldsymbol{u}}_m=\left[\hat{\boldsymbol{\mu}}_m^{(X)\top},\hat{\boldsymbol{b}}_m^{(0)\top},\hat{\boldsymbol{U}}_m\right]^{\top}\in\mathcal{R}^{(2D'+K)\times 1} \quad (21)$$

$$\hat{\boldsymbol{E}}_s=\begin{bmatrix}\boldsymbol{I}&\boldsymbol{O}&\boldsymbol{O}\\ \boldsymbol{O}&\boldsymbol{I}&\hat{\boldsymbol{W}}_s\end{bmatrix}\in\mathcal{R}^{2D'\times(2D'+K)} \quad (22)$$

5 声質変換実験

5.1 実験条件

提案手法の有効性と話者正規化学習の効果を確かめるため，一対多声質変換の実験を行った．参照話者として ATR 日本語音声データベース [13] から男性 1 名のデータを用いた．また事前収録話者として JNAS から男性話者 137 名，女性話者 136 名の計 273 名の発声を用いた [14]．各事前収録話者は 50 文を読み上げている．評価対象話者として男女 3 名ずつを選んだ．適応文数を 1 文から 16 文で変化させ，各話者 21 文を評価に用いた．

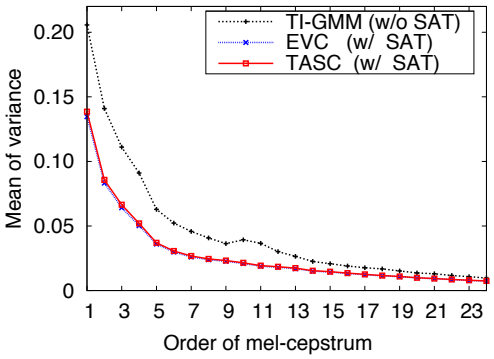

Fig. 1 出力話者モデルの分散項の平均値

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 [15]．また GMM の混合数（M）は 128 とした．

EVC における話者正規化学習では，表現ベクトルの数は $J=272$ とし，提案法における基底行列のサイズは $K=80$ とした．それぞれの値は，話者正規化学習を用いない実験の結果決定した．話者正規化学習の更新回数は 7 回とした．

変換性能について，EVC および TASC のそれぞれについて話者正規化学習を用いた場合と用いなかった場合について比較を行った．また参考として従来のパラレル学習に基づく声質変換の性能についても検証した [9]．EVC 及び TASC のそれぞれについて，適応時には正規化学習時の基底数から変化させた基底数とした（$J'\leq J$, $K'\leq K$）．

5.2 話者正規化学習による分散の縮退効果

出力話者の分散共分散行列 $\boldsymbol{\Sigma}_m^{(YY)}$ の対角成分について，TI-GMM，話者正規化学習を用いた EVC，話者正規化学習を用いた提案法のモデルを比較した．Fig. 1 は，それぞれのモデルについて，個々のガウス分布の分散の平均を示している．TI-GMM の分散共分散行列の対角成分は，話者正規化学習を用いたモデルに比べて大きな値となっている．この結果は話者正規化学習によって多数の話者データを用いて学習した際のばらつきが適切に縮退されていることを示している．一方，EVC と TASC を比べると，分散共分散行列の対角成分の値に大きな差はなかった．

5.3 客観評価実験

メルケプストラム歪みに基づく客観評価の結果を Fig. 2 に示す．Fig. 2 は適応データ数に対するメルケプストラム歪みの変化を示している．従来のパラレル学習の結果は，それぞれのデータ数で最適な混合数を選択している．話者正規化学習の適用によって EVC，TASC ともに性能の改善が得られた．すなわち話者正規化学習によって効果的に分散が縮退され，

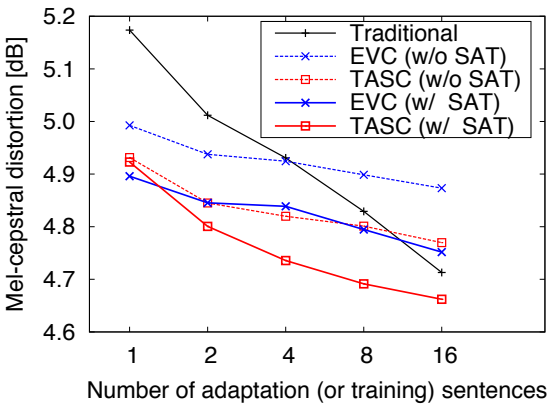

Fig. 2　客観評価実験結果; それぞれの点において混合数や基底サイズはメルケプストラム歪みの基準で最適なものを選択している．

Table 1　各手法における最適な基底サイズ

# of sentences	1	2	4	8	16
EVC w/o SAT (J)			272		
EVC w/ SAT (J')			272		
TASC w/o SAT (K)	20	20	40	80	80
TASC w/ SAT (K')	10	20	30	40	40

より個々の話者依存モデルに近いモデルが構築されたと考えられる．EVC と比較すると，話者正規化学習の有無に関わらず，TASC の性能は EVC を上回っている．これは提案法における話者空間表現の優位性を示しているといえる．話者正規化学習を用いたテンソル表現に基づく提案法は，適応文数が 16 文の場合でも，パラレル学習の結果を上回っており，話者正規化学習とテンソル表現を組み合わせた提案法により効果的に適応データの情報が捉えられていると考えられる．

Table 1 は，それぞれの手法における基底パラメータの数を表している．EVC においては，適応文数が変化してもすべての基底を用いる場合が最適となった．EVC の柔軟性は，GMM スーパーベクトルに基づく高次元表現によって制約されているといえる．一方，TASC の場合は，基底パラメータの数は適応文数に応じて変化している．GMM の混合数は 128 であり，基底行列のサイズは効果的に削減されている．話者正規化学習を導入した場合，最適な基底行列のサイズは若干削減されている．これは話者正規化学習による分散の縮退の効果が反映している可能性が考えられる．

6　おわりに

本稿では，テンソル表現を用いた任意話者声質変換のための話者正規化学習について提案した．テンソル表現を用いた任意話者声質変換においても，話者正規化学習によって精緻なモデルが構築され，声質変換の性能が向上することが示された．今後の課題として，提案法の有効性を大規模な聴取実験によって示す必要がある．また表現行列のサイズの最適化についても検討していく予定である．任意話者声質変換における，話者空間表現，パラメータ構造最適化，適応アルゴリズムを含めたベイズ学習によるモデル化も課題といえる．

参考文献

[1] A. Kain and M. W. Macon, “Spectral voice conversion for text-to-speech synthesis,” Proc. ICASSP, vol. 1, pp. 285–288, 1998.

[2] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, “Voice conversion through vector quantization,” Proc. ICASSP, pp. 655–658, 1988.

[3] 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.

[4] 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.

[5] T. Toda, Y. Ohtani, and K. Shikano, “Eigenvoice conversion based on Gaussian mixture model,” Proc. INTERSPEECH, pp. 2446–2449, 2006

[6] 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.

[7] T. Anastasakos, J. McDonough, R. Schwarts and J. Makhoul, “A compact model for speaker adaptive training,” Proc. ICSLP, vol. 2, pp. 1137–1140, 1996.

[8] 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.

[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] L. De Lathauwer, B. De Moor and J. Vandewalle, “A multilinear singular value decomposition,” SIAM Journal on Matrix Analysis and Applications, vol. 21, No. 4, pp. 1253–1278, 2000.

[11] L. R. Tucker, “Some mathematical notes on three-mode factor analysis,” Psychometrika, vol. 31, no. 3, pp. 279–311, 1966.

[12] Y. Jeong, “Speaker adaptation based on the multilinear decomposition of training speaker models,” Proc. ICASSP, pp. 4870–4873, 2010.

[13] 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.

[14] “Jnas: Japanese newspaper article sentences,” http://www.milab.is.tsukuba.ac.jp/jnas/instruct.html

[15] 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.