話者正規化学習に基づく潜在的音韻情報を考慮した音声モデリングによる非パラレル声質変換*

◎中鹿亘 (電通大)，滝口哲也 (神戸大)

1 はじめに

近年，ある入力音声に対し音韻情報を保存したまま話者性のみを変換させる声質変換技術が盛んに研究されている．この背景として，音声合成時における話者性の付与や構音障がい者音声正規化など，様々なタスク [1, 2] への応用が可能であることが挙げられる．これまでの声質変換法として，GMM (Gaussian Mixture Model) を用いた手法 [3, 4]，NMF (Non-negative matrix factorization) による変換 [5]，深層学習に基づく変換 [6] などが提案されてきた．しかしながら，これらの手法では，モデルの学習時にパラレルデータ（入力話者と出力話者の，同一発話内容による音声対）を必要とし，これによって事前処理にコストが掛かる，使用するデータセットが制限される，音声に不自然な変換が加わってしまうなど様々な弊害が生じる．入出力話者間のパラレルデータを必要としない手法として，Eigenvoice を用いた手法がある [7]．これは，予め多数の話者から参照話者へのマッピング関数，参照話者から多数の話者へのマッピング関数を GMM 及び固有声を用いて学習しておくことで，多対多声質変換を実現している．しかしこのアプローチにおいても，GMM の学習時には複数話者のパラレルデータを用意する必要がある．

そこで学習時において全くパラレルデータの必要ない声質変換手法（本稿では非パラレル声質変換法と呼ぶ）として，我々は ARBM (adaptive restricted Boltzmann machine）を用いた手法を提案してきた [8]．ARBM は，特徴ベクトルを表現する可視素子，潜在特徴ベクトルを表す隠れ素子，発話者を特定する識別素子を変数とする確率モデルである．このモデルでは可視素子-隠れ素子間にのみ，話者に依存した（話者に依存しないパラメータを話者固有の行列で射影した）強度（重み）で結合が存在していると仮定している．話者依存・非依存のパラメータを同時学習させ，特定固有の情報を入れ替えることで多対多声質変換を実現している．しかしながら，話者依存パラメータの物理量が曖昧であり，どのような原理で声質が変換されるのか不明瞭であった．そこで ARBM による声質変換法を元に，問題を整理し，再定義するのが本稿の目的である．

2 問題設定

一般に，音声信号に対して話者性に関する情報は乗算的に付与されることが知られている．本稿では，時刻 t における話者 r の音響特徴ベクトル $\hat{\boldsymbol{x}}_{rt} \in \mathbb{R}^D$ は以下のように表されるものとする．

$$\hat{\boldsymbol{x}}_{rt} = \mathbf{A}_r \boldsymbol{x}_t + \boldsymbol{b}_r \tag{1}$$

ただし，$\boldsymbol{x}_t$ は話者正規化された（標準話者の）音響特徴ベクトル，$\mathbf{A}_r \in \mathbb{R}^{D\times D}$ と $\boldsymbol{b}_r \in \mathbb{R}^D$ はそれぞれ話者 r 固有の適応行列（正則行列）およびバイアス項である．式 (1) のバイアス項は，複数の話者による音声データセットでは，話者ごとにマイクの特性や録音環境が異なることを考慮して加えている．ここで，$\boldsymbol{x}_t$ は時間に依存しない分散を持つ各次元独立な多変量正規分布に従うとする．すなわち $\boldsymbol{x}_t \sim \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma})$ であり，$\boldsymbol{\Sigma} = \mathrm{diag}\,(\boldsymbol{\sigma}^2)$ とする．ただし，$\boldsymbol{\mu}_t \in \mathbb{R}^D, \boldsymbol{\sigma}^2 = [\sigma_1^2, \cdots, \sigma_D^2] \in \mathbb{R}^D$ である．このとき，$\hat{\boldsymbol{x}}_{rt}$ も多変量正規分布に従い，

$$\begin{aligned} \hat{\boldsymbol{x}}_{rt} &\sim \mathcal{N}(\hat{\boldsymbol{\mu}}_{rt}, \hat{\boldsymbol{\Sigma}}_r), \\ \hat{\boldsymbol{\mu}}_{rt} &= \mathbf{A}_r \boldsymbol{\mu}_t + \boldsymbol{b}_r \\ \hat{\boldsymbol{\Sigma}}_r &= \mathbf{A}_r \boldsymbol{\Sigma} \mathbf{A}_r^\top \end{aligned} \tag{2}$$

となる．

各時刻における標準話者の音声は，観測はできないが潜在的に存在する音韻情報によって決定されるはずである．そこで標準話者の平均ベクトル $\boldsymbol{\mu}_t$ は，潜在的音韻特徴ベクトル $\boldsymbol{h}_t \in \mathbb{B}^H$ ($\mathbb{B}$ は 0 または 1 のみを取り得る空間) を用いて，以下のように定まるとする．

$$\boldsymbol{\mu}_t = \mathbf{W}\boldsymbol{h}_t + \boldsymbol{b} \tag{3}$$

ここで，$\mathbf{W} \in \mathbb{R}^{D\times H}$ と $\boldsymbol{b} \in \mathbb{R}^D$ は潜在特徴から音響特徴へ変換する射影行列およびバイアス項である．同様に，音韻特徴ベクトルも音響特徴ベクトルによって決定されると考える．つまり，$\boldsymbol{h}_t \sim \mathcal{B}(\boldsymbol{\pi}_t(\boldsymbol{x}_t))$ で表されるとする．ただし $\mathcal{B}(\cdot)$ は多変数ベルヌーイ分布であり，$\pi_{jt} \in \boldsymbol{\pi}_t$ $(j = 1, \cdots, H)$ は変数 $h_{jt} \in \boldsymbol{h}_t$ が値 1 となる確率（すなわち $\pi_{jt} = p(h_{jt} = 1)$）を表すパラメータである．音韻特徴ベクトルから音響特徴ベクトルを決定する際，共通のパラメータを用い

*Non-parallel voice conversion using combination of restricted Boltzmann machine and speaker-adaptive training. by Toru NAKASHIKA (UEC), Tetsuya TAKIGUCHI (Kobe University)

る方がパラメータ数を削減する上で都合が良い．そこで $\boldsymbol{\pi}_t$ を以下のように定義する．

$$\boldsymbol{\pi}_t = \mathcal{S}(\mathbf{W}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_t + \boldsymbol{c}) \tag{4}$$

ここで $\mathcal{S}(\cdot)$ は要素ごとのシグモイド関数を表す．また，$\boldsymbol{c} \in \mathbb{R}^H$ は時刻に寄らない音韻情報に関するバイアス項である．

ところで，$\boldsymbol{h}_t$ が既知であるときの $\hat{\boldsymbol{x}}_{rt}$ の確率（すなわち $p(\hat{\boldsymbol{x}}_{rt}|\boldsymbol{h})$）を考える．式 (2) より，変数 $\boldsymbol{x}_{rt}$ に関して整理すると，

$$\begin{aligned} p(\hat{\boldsymbol{x}}_{rt}|\boldsymbol{h}_t) &= \mathcal{N}(\hat{\boldsymbol{\mu}}_{rt}, \hat{\boldsymbol{\Sigma}}_r) \\ &\propto e^{-\frac{1}{2}(\hat{\boldsymbol{x}}_{rt}-\hat{\boldsymbol{\mu}}_r)^\top \hat{\boldsymbol{\Sigma}}_r^{-1}(\hat{\boldsymbol{x}}_{rt}-\hat{\boldsymbol{\mu}}_r)} \\ &\propto e^{-\{\frac{1}{2}(\hat{\boldsymbol{x}}_{rt}-\hat{\boldsymbol{b}}_r)^\top \hat{\boldsymbol{\Sigma}}_r^{-1}(\hat{\boldsymbol{x}}_{rt}-\hat{\boldsymbol{b}}_r) - \hat{\boldsymbol{x}}_{rt}^\top \hat{\boldsymbol{\Sigma}}_r^{-1} \hat{\mathbf{W}}_r \boldsymbol{h}\}} \end{aligned} \tag{5}$$

と書き表すことができる．ただし，$\hat{\boldsymbol{b}}_r = \mathbf{A}_r \boldsymbol{b} + \boldsymbol{b}_r$，$\hat{\mathbf{W}}_r = \mathbf{A}_r \mathbf{W}$ と置いた．一方，$\hat{\boldsymbol{x}}_r$ が既知のとき，$\boldsymbol{x}_t = \mathbf{A}_r^{-1}(\hat{\boldsymbol{x}}_{rt} - \boldsymbol{b}_r)$，$\boldsymbol{\Sigma}^{-1} = \mathbf{A}_r^\top \hat{\boldsymbol{\Sigma}}_r^{-1} \mathbf{A}_r$ であることに留意して，$p(\boldsymbol{h}_t|\hat{\boldsymbol{x}}_{rt})$ は

$$\begin{aligned} p(\boldsymbol{h}_t|\hat{\boldsymbol{x}}_{rt}) &= \mathcal{B}(\boldsymbol{\pi}_t(\mathbf{A}_r^{-1}(\hat{\boldsymbol{x}}_{rt} - \boldsymbol{b}_r))) \\ &\propto e^{(\mathbf{W}^\top \boldsymbol{\Sigma}^{-1} \mathbf{A}_r^{-1}(\hat{\boldsymbol{x}}_{rt}-\boldsymbol{b}_r)+\boldsymbol{c})^\top \boldsymbol{h}_t} \\ &= e^{-(-\hat{\boldsymbol{x}}_{rt}^\top \hat{\boldsymbol{\Sigma}}_r^{-1} \hat{\mathbf{W}}_r \boldsymbol{h} - \hat{\boldsymbol{c}}_r^\top \boldsymbol{h})} \end{aligned} \tag{6}$$

となる．ただし，$\hat{\boldsymbol{c}}_r = \boldsymbol{c} - \hat{\mathbf{W}}_r^\top \hat{\boldsymbol{\Sigma}}_r^{-1} \boldsymbol{b}_r$ と置いた．

今，$\hat{\boldsymbol{x}}_{rt}$ と $\boldsymbol{h}_t$ の同時確率分布を考える．式 (5)(6) に着目すると，べき乗の中で共通して表れる項 $(-\hat{\boldsymbol{x}}_{rt}^\top \hat{\boldsymbol{\Sigma}}_r^{-1} \hat{\mathbf{W}}_r \boldsymbol{h})$ が存在していることが分かる．そこで同時確率分布 $p(\hat{\boldsymbol{x}}_{rt}, \boldsymbol{h}_t)$ を以下のように定義すると式 (5)(6) を満たす．

$$\begin{aligned} p(\hat{\boldsymbol{x}}_{rt}, \boldsymbol{h}_t) =& \frac{1}{Z} e^{-E(\hat{\boldsymbol{x}}_{rt}, \boldsymbol{h}_t)} \\ E(\hat{\boldsymbol{x}}_{rt}, \boldsymbol{h}_t) =& \frac{1}{2}(\hat{\boldsymbol{x}}_{rt} - \hat{\boldsymbol{b}}_r)^\top \hat{\boldsymbol{\Sigma}}_r^{-1}(\hat{\boldsymbol{x}}_{rt} - \hat{\boldsymbol{b}}_r) \\ & - \hat{\boldsymbol{x}}_{rt}^\top \hat{\boldsymbol{\Sigma}}_r^{-1} \hat{\mathbf{W}}_r \boldsymbol{h}_t - \hat{\boldsymbol{c}}_r^\top \boldsymbol{h}_t \end{aligned} \tag{7}$$

ただし，$Z = \int^D \sum_{\boldsymbol{h}_t} e^{-E(\hat{\boldsymbol{x}}_{rt}, \boldsymbol{h}_t)} d^D \hat{\boldsymbol{x}}_{rt}$ は全域での確率を 1 にするための正規化項である．なお，式 (7) において式 (1) を代入すると，

$$\begin{aligned} p(\boldsymbol{x}_t, \boldsymbol{h}_t) &= \frac{1}{Z} e^{-E(\boldsymbol{x}_t, \boldsymbol{h}_t)} \\ E(\boldsymbol{x}_t, \boldsymbol{h}_t) &= \frac{\|\boldsymbol{x}_t - \boldsymbol{b}\|_2^2}{2\boldsymbol{\sigma}^2} - \left(\frac{\boldsymbol{x}_t}{\boldsymbol{\sigma}^2}\right)^\top \mathbf{W} \boldsymbol{h}_t - \boldsymbol{c}^\top \boldsymbol{h}_t \end{aligned} \tag{8}$$

となり，これは Gaussian-Bernoulli RBM (restricted Boltzmann machine [9]) に他ならない（$\frac{\cdot}{\cdot}$ は要素除算を表す）．すなわち式 (7) によるモデル化は，標準話者の音響特徴ベクトルを可視素子，潜在的音響特徴ベクトルを隠れ素子とした RBM において式 (1) により話者適応を施したモデルとみなすことができる (Fig. 1).

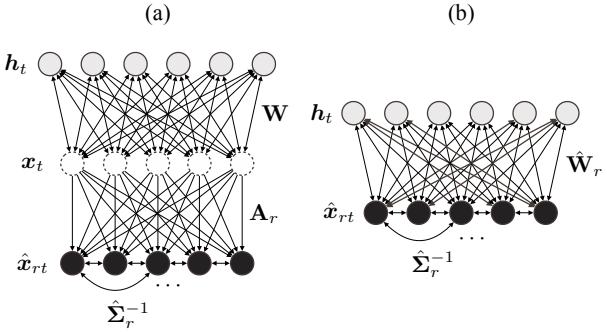

Fig. 1 (a) Proposed model: SATBM (speaker-adaptive-trainable Boltzmann machine) and (b) its simplified representation, which can be seen as a sort of semi-RBM.

また Fig. 1 (b) で示すように $\hat{\boldsymbol{x}}_{rt}$ と $\boldsymbol{h}_t$ の素子間には $\hat{\mathbf{W}}_r$ の接続重みが共有されており，$\boldsymbol{h}_t$ 間には接続はなく $\hat{\boldsymbol{x}}_{rt}$ 間には接続があるため一種の semi-RBM とみなすこともできる．しかし，Fig. 1 (a) にあるように，話者に寄らない音響特徴ベクトル $\boldsymbol{x}_t$ の存在を仮定し，パラメータを細分化することで話者正規化学習を可能にしている点で純粋な semi-RBM と異なる．式 (7) で表される確率モデルを本稿では SATBM (speaker-adaptive-trainable Boltzmann machine) と呼ぶ．我々の先行研究で提案した ARBM (Adaptive restricted Boltzmann machine) はモデル空間における話者適応であったのに対し，式 (1) に示すように SATBM はモデル空間であり且つ特徴量空間での変換でもある．話者適応を用いた音声認識では，モデル空間の変換に基づく MLLR (maximum likelihood linear regression) よりも，特徴量空間での変換でもある CMLLR (constrained MLLR) の方が歪みの少ない変換であると考えられ，高い精度を上げていることが報告されている [10]．SATBM と ARBM の声質変換における比較においても同様の理由により，SATBM の方が高い精度を上げることが期待される．

3 話者正規化学習に基づくパラメータ推定

前節で定義した SATBM は，話者正規化学習 (SAT; speaker adaptive training [11]) に基づいてパラメータを推定することができる．SATBM のパラメータは話者に依存するもの $\boldsymbol{\Theta}^{SD} = \{\mathbf{A}_r, \boldsymbol{b}_r\}_{r=1}^R$ と話者に依存しないもの $\boldsymbol{\Theta}^{SI} = \{\mathbf{W}, \boldsymbol{\sigma}^2, \boldsymbol{b}, \boldsymbol{c}\}$ に分けることができる．これらは R 人の話者による音声データ $\boldsymbol{X} = \{\boldsymbol{X}_r\}_{r=1}^R$，$\boldsymbol{X}_r = \{\hat{\boldsymbol{x}}_{rt}\}_{t=1}^{T_r}$ に対する尤度を最大化するように同時に推定される．すなわち，

$$(\hat{\boldsymbol{\Theta}}^{SD}, \hat{\boldsymbol{\Theta}}^{SI}) \triangleq \underset{(\boldsymbol{\Theta}^{SD}, \boldsymbol{\Theta}^{SI})}{\operatorname{argmax}} \prod_{r=1}^{R} \prod_{t=1}^{T_r} p(\hat{\boldsymbol{x}}_{rt}) \tag{9}$$

とする．話者正規化学習の考え方から，話者に起因する変動は $\boldsymbol{\Theta}^{SD}$ に，それ以外の音韻に起因する変動は $\boldsymbol{\Theta}^{SI}$ によって捉えられる．さらに提案法では式 (3) により，標準話者の音響特徴量と音韻情報の関係性をモデル化しており，SAT+MLLR に基づく話者適応よりも音声データに適合する可能性を示唆している．

勾配法によってパラメータを更新するため，パラメータ θ に対する対数尤度の偏微分を考える．対数尤度 $l = \log \prod_r \prod_t p(\hat{\boldsymbol{x}}_{rt}) = \sum_r \sum_t \log \sum_{\boldsymbol{h}} p(\hat{\boldsymbol{x}}_{rt}, \boldsymbol{h}_t))$ であることから，式 (7) より，

$$\frac{\partial l}{\partial \theta} = \sum_r (\langle \frac{\partial E(\hat{\boldsymbol{x}}_{rt}, \boldsymbol{h}_t)}{\partial \theta} \rangle_{\text{data}} - \langle \frac{\partial E(\hat{\boldsymbol{x}}_{rt}, \boldsymbol{h}_t)}{\partial \theta} \rangle_{\text{model}}) \tag{10}$$

が導ける．ただし，$\langle \cdot \rangle_{\text{data}}$，$\langle \cdot \rangle_{\text{model}}$ はそれぞれ話者 r のデータ（$p(\boldsymbol{h}_t|\hat{\boldsymbol{x}}_{rt})$）に対する期待値，モデル（$p(\hat{\boldsymbol{x}}_{rt}, \boldsymbol{h}_t)$）の期待値を表す．モデルに対する期待値は計算困難だが通常の RBM と同様 CD（contrastive divergence）法 [12] を適用することで，効率よくパラメータを推定することができる．各パラメータの偏微分値 $\frac{\partial E(\hat{\boldsymbol{x}}_{rt}, \boldsymbol{h}_t)}{\partial \theta}$ を計算すると以下の式が得られる．

$$\begin{aligned}
\frac{\partial E(\hat{\boldsymbol{x}}_{rt}, \boldsymbol{h}_t)}{\partial \mathbf{A}_r} &= -\frac{1}{2}(\mathbf{A}_r^{-1}\mathbf{C}_{rt}\hat{\boldsymbol{\Sigma}}_r^{-1} + \hat{\boldsymbol{\Sigma}}_r^{-1}\mathbf{D}_{rt}\mathbf{A}_r^{-\top}) \\
\frac{\partial E(\hat{\boldsymbol{x}}_{rt}, \boldsymbol{h}_t)}{\partial \boldsymbol{b}_r} &= -\hat{\boldsymbol{\Sigma}}_r^{-1}(\hat{\boldsymbol{x}}_{rt} - \hat{\boldsymbol{b}}_r - \hat{\mathbf{W}}_r\boldsymbol{h}_t) \\
\frac{\partial E(\hat{\boldsymbol{x}}_{rt}, \boldsymbol{h}_t)}{\partial \mathbf{W}} &= -\mathbf{A}_r^{\top}\hat{\boldsymbol{\Sigma}}_r^{-1}(\hat{\boldsymbol{x}}_{rt} - \boldsymbol{b}_r)\boldsymbol{h}_t^{\top} \\
\frac{\partial E(\hat{\boldsymbol{x}}_{rt}, \boldsymbol{h}_t)}{\partial \boldsymbol{\sigma}^2} &= -\frac{1}{2}\text{diag}(\mathbf{A}_r^{\top}\hat{\boldsymbol{\Sigma}}_r^{-1}\mathbf{E}_{rt}\hat{\boldsymbol{\Sigma}}_r^{-1}\mathbf{A}_r) \\
\frac{\partial E(\hat{\boldsymbol{x}}_{rt}, \boldsymbol{h}_t)}{\partial \boldsymbol{b}} &= -\mathbf{A}_r^{\top}\hat{\boldsymbol{\Sigma}}_r^{-1}(\hat{\boldsymbol{x}}_{rt} - \hat{\boldsymbol{b}}_r) \\
\frac{\partial E(\hat{\boldsymbol{x}}_{rt}, \boldsymbol{h}_t)}{\partial \boldsymbol{c}} &= -\boldsymbol{h}_t
\end{aligned}$$

ただし，

$$\begin{aligned}
\mathbf{C}_{rt} &= (\hat{\boldsymbol{x}}_{rt} - \boldsymbol{b}_r)(\hat{\boldsymbol{x}}_{rt} - \hat{\boldsymbol{b}}_r - 2\hat{\mathbf{W}}_r\boldsymbol{h}_t)^{\top} \\
\mathbf{D}_{rt} &= (\hat{\boldsymbol{x}}_{rt} - \hat{\boldsymbol{b}}_r)(\hat{\boldsymbol{x}}_{rt} - \boldsymbol{b}_r)^{\top} \\
\mathbf{E}_{rt} &= (\hat{\boldsymbol{x}}_{rt} - \hat{\boldsymbol{b}}_r)(\hat{\boldsymbol{x}}_{rt} - \hat{\boldsymbol{b}}_r)^{\top} - 2(\hat{\boldsymbol{x}}_{rt} - \boldsymbol{b}_r)(\hat{\mathbf{W}}_r\boldsymbol{h}_t)^{\top}
\end{aligned}$$

と置いた．なお，上記によってパラメータを更新する度に逆行列 $\hat{\boldsymbol{\Sigma}}_r^{-1}$ を計算する必要があり，このままでは計算コストが高い．そこで本稿で述べる実験では $\mathbf{A}_r$ を対角行列，もしくは三重対角行列とする．これは推定するパラメータ数を大幅に削減することができるため，特定の話者の発話データが少ないケースなどを考慮すれば，過学習の抑制や学習の安定化の面でも都合が良い．特に音響特徴量をケプストラムとした場合，ターゲット話者のケプストラムへのワーピング行列は対角成分（第 0 対角）と第 1 対角，第 -1 対角のみで十分であり，他の成分は無視できる程小さくなることが知られている [13]．なお三重対角行列を用いた場合，Thomas アルゴリズムによって高速に $\mathbf{A}_r\boldsymbol{y} = \boldsymbol{d}$ となる解 $\boldsymbol{y}$ を計算することができる．これにより 2 回の Thomas アルゴリズムの実行によって $\hat{\boldsymbol{\Sigma}}_r^{-1}$ を高速に求めることができる．

4 声質変換への応用

SATBM を用いて声質変換を行う場合，まず事前学習として複数（R 人）の参照話者によるデータを用いて式 (9) により各パラメータを同時推定する．これにより話者正規化されたパラメータ集合 $\hat{\boldsymbol{\Theta}}^{SI}$ が得られる．次に，$\hat{\boldsymbol{\Theta}}^{SI}$ を固定して，入力話者と出力話者の適応データ（$\{\hat{\boldsymbol{x}}_{it}\}_{t=1}^{Ti}, \{\hat{\boldsymbol{x}}_{ot}\}_{t=1}^{To}$）を用いてそれぞれの話者依存パラメータ $\boldsymbol{\Theta}_i^{SD} = \{\mathbf{A}_i, \boldsymbol{b}_i\}$，$\boldsymbol{\Theta}_o^{SD} = \{\mathbf{A}_o, \boldsymbol{b}_o\}$ を推定する．すなわち

$$\hat{\boldsymbol{\Theta}}_r^{SD} \triangleq \underset{\boldsymbol{\Theta}_r^{SD}}{\text{argmax}} \prod_{t=1}^{T_r} p(\hat{\boldsymbol{x}}_{rt}; \boldsymbol{\Theta}_r^{SD}, \hat{\boldsymbol{\Theta}}^{SI}),\ r \in \{i, o\} \tag{11}$$

として $\boldsymbol{\Theta}_i^{SD}$ と $\boldsymbol{\Theta}_o^{SD}$ を推定する．

入力話者話者のフレーム音響特徴ベクトル $\boldsymbol{x}_{it}$ を出力話者の音響特徴ベクトル $\boldsymbol{x}_{ot}$ へ変換することを考える．本稿では単純な線形射影に基づく変換と最尤法に基づく変換の 2 つのアプローチを考える．まず線形射影に基づく変換では，以下の式により $\boldsymbol{x}_{ot}$ を推定する．

$$\boldsymbol{x}_{ot} \triangleq \mathbf{A}_o\mathbf{A}_i^{-1}(\boldsymbol{x}_{it} - \boldsymbol{b}_i) + \boldsymbol{b}_o \tag{12}$$

これは $\boldsymbol{x}_t = \mathbf{A}_i^{-1}(\boldsymbol{x}_{it} - \boldsymbol{b}_i) = \mathbf{A}_o^{-1}(\boldsymbol{x}_{ot} - \boldsymbol{b}_o)$ の関係から導いたものである．しかしこれには，真の標準話者特徴ベクトル空間が得られているという前提が存在している．

もう一つのアプローチは $\boldsymbol{x}_i$ が与えられたときの $\boldsymbol{x}_o$ の出現確率が最大となるベクトルを選ぶ方法である．すなわち以下のように定式化される．

$$\begin{aligned}
\boldsymbol{x}_{ot} &\triangleq \underset{\boldsymbol{x}_{ot}}{\text{argmax}}\, p(\boldsymbol{x}_{ot}|\boldsymbol{x}_{it}) \\
&= \underset{\boldsymbol{x}_{ot}}{\text{argmax}} \sum_{\boldsymbol{h}_t} p(\boldsymbol{h}_t|\boldsymbol{x}_{it})p(\boldsymbol{x}_{ot}|\boldsymbol{h}_t) \\
&\simeq \underset{\boldsymbol{x}_{ot}}{\text{argmax}}\, p(\hat{\boldsymbol{h}}_t|\boldsymbol{x}_{it})p(\boldsymbol{x}_{ot}|\hat{\boldsymbol{h}}_t) \\
&= \underset{\boldsymbol{x}_{ot}}{\text{argmax}}\, p(\boldsymbol{x}_{ot}|\hat{\boldsymbol{h}}_t) \\
&= \mathbf{A}_o\mathbf{W}\mathcal{S}(\mathbf{W}^{\top}\boldsymbol{\Sigma}^{-1}\mathbf{A}_i^{-1}(\boldsymbol{x}_{it} - \boldsymbol{b}_i) + \boldsymbol{c}) + \mathbf{A}_o\boldsymbol{b} + \boldsymbol{b}_o
\end{aligned} \tag{13}$$

ただし $\hat{\boldsymbol{h}}_t \triangleq \underset{\boldsymbol{h}_t}{\text{argmax}}\, p(\boldsymbol{h}_t|\boldsymbol{x}_{it})$ とおいた．式 (12) と式 (13) を比べると，確率表現に基づく式 (13) では非線形関数が加わっており，式 (12) よりも高い表現力を持つことが期待される．

Table 1　Average MDIR [dB] of each condition.

cond.	50S-3D-P	50S-3D-L	50S-1D-P	50S-1D-L	5S-3D-P	5S-3D-L	5S-1D-P	5S-1D-L
MDIR	2.66	1.54	1.72	1.56	2.46	1.07	1.13	0.91

5　検証実験

提案手法の SATBM に基づく声質変換の性能を検証するため，日本音響学会研究用連続音声データベース（ASJ-JIPDEC）を用いた実験を行った．セット A の男性 30 名女性 34 名計 64 話者の音声のうち 5 発話もしくは 50 発話のデータ（非パラレルデータ）をモデルの学習に用いた．分析合成ツールの WORLD[14] によって得られたスペクトルから計算された 64 次元のメルケプストラムを入力とし，96 次元の潜在特徴ベクトルを与えた．評価用として男性 1 名を入力話者，女性 1 名を出力話者に選び，50 発話分のパラレルデータによって客観評価を行った．MCD は出力話者音声とのメルケプストラム距離に基づく指標だが，必ずしも声質の変換結果を認知する上で与えられた音声間の距離が最小となれば良いわけではなく，また用いる評価データによって基準となる MCD 値が異なるため，本稿では以下で定義される MDIR (mel-cepstral distortion improvement ratio) を評価基準に用いた．

$$MDIR[dB] = \frac{10\sqrt{2}}{\ln 10}(\|\boldsymbol{m}_o - \boldsymbol{m}_i\|_2 - \|\boldsymbol{m}_o - \boldsymbol{m}_c\|_2)$$

ここで $\boldsymbol{m}_o$，$\boldsymbol{m}_i$，$\boldsymbol{m}_c$ はそれぞれあるフレームにおける出力話者，入力話者，変換後のメルケプストラム特徴ベクトルを表す．MDIR は改善率を表すため，値が大きいほど高い変換精度を示す．

実験結果を Table 1 に示す．実験条件の S と D はそれぞれ学習に用いた発話文数，適応行列の対角数を表す．また，L は線形射影に基づく変換，P は最尤法に基づく変換を表す．表より，三重対角の適応行列を用いることが非常に効果的であることが分かる．また単純な線形変換よりも最尤法に基づいた変換の方が高い精度を示していた．特に三重対角かつ最尤法を用いれば，学習データ数が 5 発話でも 50 発話の場合に匹敵するほどの性能であった．また，従来手法である ARBM (5S-1D-P) の MDIR は 0.82[dB] であった．同条件の提案手法（MDIR 1.13[db]）と比較すると，提案手法の方が高い変換精度を示した．

6　おわりに

本研究では特徴空間における変換に基づく潜在的音響特徴量を考慮した話者正規化学習により音韻と話者情報を分離し，声質変換に適用する手法を提案した．今後はスパース正則化など，音韻情報やパラメータの事前確率を考慮したモデリングを行いたい．

参考文献

[1] C. Veaux and X. Robet, “Intonation conversion from neutral to expressive speech,” Interspeech, pp. 2765–2768, 2011.

[2] K. Nakamura et al., “Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech,” Speech Commun., vol. 54, no. 1, pp. 134–146, 2012.

[3] Y. Stylianou et al., “Continuous probabilistic transform for voice conversion,” IEEE Trans. Speech and Audio Process., vol. 6, no. 2, pp. 131–142, 1998.

[4] T. Toda et al., “Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory,” IEEE Trans. Audio, Speech, and Lang. Process., vol. 15, no. 8, pp. 2222–2235, 2007.

[5] R. Takashima, T. Takiguchi and Y. Ariki: “Exemplar-based voice conversion in noisy environment”, SLT, pp. 313–317 (2012).

[6] T. Nakashika, T. Takiguchi and Y. Ariki: “High-order sequence modeling using speaker-dependent recurrent temporal restricted Boltzmann machines for voice conversion”, Interspeech, pp. 2278–2282 (2014).

[7] T. Toda et al., “Eigenvoice conversion based on Gaussian mixture model,” Interspeech, pp. 2446–2449, 2006.

[8] T. Nakashika, T. Takiguchi and Y. Ariki: “High-order sequence modeling using speaker-dependent recurrent temporal restricted Boltzmann machines for voice conversion”, Interspeech, pp. 2278–2282 (2014).

[9] K. Cho et al., “Improved learning of Gaussian-Bernoulli restricted Boltzmann machines,” ICANN, pp. 10–17, 2011.

[10] M. J. F Gales, “Maximum likelihood linear transformations for HMM-based speech recognition,” Computer Speech and Language, vol. 12, pp. 75–98, 1997.

[11] T. Anastakos et al., “A compact model for speaker-adaptive training,” Int. Conf. Speech Language Processing ’96, vol. 2, pp. 1137–1140, 1996.

[12] G. E. Hinton et al., “A fast learning algorithm for deep belief nets,” Neural computation, vol. 18, no. 7, pp. 1527–1554, 2006.

[13] T. Emori and K. Shinoda, “Vocal tract length normalization using rapid maximum-likelihood estimation for speech recognition,” IEICE Transactions, vol. J83-D-II, no. 11, pp. 2108–2117, 2000.

[14] M. Morise, “An attempt to develop a singing synthesizer by collaborative creation,” SMAC2013, pp. 287-292, 2013.