一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SP2014-34(2014-05)

スパース表現に基づく声質変換のための結合型restricted Boltzmann machine

中鹿　亘†　滝口　哲也††　有木　康雄††

† 神戸大学大学院システム情報学研究科
〒 657-8501 兵庫県神戸市灘区六甲台町 1-1
†† 神戸大学自然科学系先端融合研究環
〒 657-8501 兵庫県神戸市灘区六甲台町 1-1
E-mail: †nakashika@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

あらまし　近年，声質変換の研究分野において，over-fitting や over-smoothing の生じにくいスパース表現に基づく手法が注目を浴びている．スパース表現に基づく声質変換法では，予め入力話者・出力話者のパラレル辞書を求めておき，スパースな辞書選択重みを用いて適切な辞書を選択することで声質変換を実現する．この手法は主に 2 つのアプローチに分けることができる．1 つ目はパラレル辞書として，学習データの音響特徴量をそのまま辞書として用いるアプローチであり，もう 1 つは，パラレル辞書そのものを何らかの手法で学習させるアプローチである．本研究では，後者のアプローチに基づき，近年注目を浴びている Deep Learning の基礎技術となる restricted Bolzmann machine (RBM) を用いて，入力話者・出力話者のパラレル辞書を体系的に求める手法を提案する．評価実験では，代表的な手法である Gaussian mixture model (GMM) だけでなく，従来のスパース表現に基づく手法である non-negative matrix factorization (NMF) による声質変換法に比べて高い精度が得られたことを確認した．

キーワード　声質変換，restricted Boltzmann machine, スパース表現, パラレル辞書学習

A joint restricted Boltzmann machine for dictionary learning in sparse-representation-based voice conversion

Toru NAKASHIKA†, Tetsuya TAKIGUCHI††, and Yasuo ARIKI††

† Graduate School of System Informatics, Kobe University
1-1 Rokkodaicho, Nada-ku, Kobe, Hyogo, 657-8501 Japan
†† Organization of Advanced Science and Technology, Kobe University
1-1 Rokkodaicho, Nada-ku, Kobe, Hyogo, 657-8501 Japan
E-mail: †nakashika@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

Abstract　In voice conversion, sparse-representation-based methods have recently been garnering attention because they are, relatively speaking, not affected by over-fitting or over-smoothing problems. In these approaches, voice conversion is achieved by estimating a sparse vector that determines which dictionaries of the target speaker should be used, calculated from the matching of the input vector and dictionaries of the source speaker. The sparse-representation-based voice conversion methods can be broadly divided into two approaches: 1) an approach that uses raw acoustic features in the training data as parallel dictionaries, and 2) an approach that trains parallel dictionaries from the training data. Our approach belongs to the latter; we systematically estimate the parallel dictionaries using a restricted Boltzmann machine, a fundamental technology commonly used in deep learning. Through voice-conversion experiments, we confirmed the high-performance of our method, comparing it with the conventional Gaussian mixture model (GMM)-based approach, and a non-negative matrix factorization (NMF)-based approach, which is based on sparse-representation.

Key words　Voice conversion, restricted Boltzmann machine, sparse representation, parallel dictionary learning

1. はじめに

近年，声質変換法（入力話者音声の音韻情報を残し，話者性のみを出力話者のものへ変換させる手法）が，音声信号処理の分野で盛んに研究されている．その背景として，雑音環境下 [1] や感情音声 [2] の音声認識精度の向上，発話困難な障がい者のためのアシスタント [3]，その他様々なタスク [4], [5] への応用が可能であることが挙げられる．入力話者の音声を出力話者の声質をもつ音声へ変換するためには主に F0 特徴とスペクトル情報を変換する必要があるが，多くの声質変換に関する研究では，F0 ではなくスペクトルの変換法に注力しており，本研究においてもこれに準ずる．

文献 [6], [7] にも述べられているように，声質変換法としてこれまでに様々な統計的アプローチが研究されてきた．中でも GMM (Gaussian Mixture Model) を用いた手法 [8] が最も広く用いられており，様々な改良がなされてきた．例えば，戸田ら [9] は，動的特徴量とグローバルバリアンス (GV) を導入して，変換精度を向上させる手法を提案した．また，Helander ら [10] は PLS (Partial Least Squares) を用いて精度を改善する手法を提案した．

しかしながら，GMM に基づくアプローチは，over-fitting や over-smoothing の問題が報告されている [10]〜[13]．これは，GMM がその性質上ノイズの影響を受けやすく，また変換式が平均値に基づく線形変換となっていることに起因する．そこで近年，このような問題の少ないスパース表現に基づく声質変換法が提案されており，NMF (non-negative matrix factorization) [14] を用いた手法 [15], [16] や US (unit selection) を用いた手法 [12] が挙げられる．

スパース表現に基づく声質変換法では一般に，入力話者音声と出力話者音声のパラレルデータから基底（パラレル辞書）を学習しておき，テストデータが与えられたときに，入力話者の辞書と照らし合わせることでスパース重みを求め（辞書の選択），対となる出力話者の辞書を用いて出力話者の音声を生成する．この変換精度は，学習時における辞書の作成精度と，テスト時における辞書の選択精度によって決まる．パラレル辞書 NMF を用いた手法 [15] では，学習時には入力話者と出力話者のパラレルデータをそのままパラレル辞書として用意するので，辞書作成時における誤差は生じないが，テスト時において，入力話者と出力話者のアクティビティ行列の不一致により誤差が生じてしまい，変換精度を下げる要因となる．また，アクティビティ行列が一致するようにパラレル辞書を学習する手法 [16] も提案されているが，逆に辞書の変換精度が下がってしまい，変換精度に悪影響を与える．

また，これまで述べてきた声質変換法はいずれも線形変換をベースとしており，この制約によって変換精度には限界がある．通常人間の声道形状は非線形的であり，非線形ベースの変換手法の方が音声信号の変換の際にはより適切であると考えられる．音声信号に含まれる声質の特性をより正確に捉えるためには，複数の非線形層を持つ変換構造にすることが望まれる．このアプローチの例として，Desai らによる多層 NN (Neural Networks) を用いた声質変換法 [17] や，我々が提案してきた話者依存型 RBM (restricted Boltzmann machine) 若しくは DBN (deep belief networks [18]) を用いた多層型声質変換法 [19]，話者依存型 CRBM (conditional restricted Boltzmann machine) [20] を用いた手法 [21]，Wu らによる CRBM を用いた非線形声質変換法 [22]，Joint 型 RBM を用いた手法 [23] が挙げられる．いずれの手法においても，非線形変換に基づくアプローチでは，線形変換ベースの手法と比べて比較的高い精度が得られていることが報告されている [17], [19], [21], [22].

こうした背景を踏まえて，本研究ではスパース表現に基づく声質変換において，RBM によるパラレル辞書学習 [24] を用いることで非線形変換に基づく声質変換を実現する．RBM は可視層と隠れ層からなる無向グラフの確率モデルであり，同一層内のユニット間結合はなく，異なる層のユニット間のみ結合が存在する，といった特徴がある．これらの結合強度（重み）は教師なし学習で推定することができる．RBM は近年特に注目を浴びている Deep Learning の基礎技術となっており，これを多段に積み重ねたネットワーク (DBN) を用いた手法が，手書き文字認識 [18] や 3 次元物体認識 [25]，機械翻訳 [26]，音声認識 [27] など，多岐にわたる分野において高い精度を示している．本研究では，入力・出力話者でフレーム対応のとれた，MFCC などの音響特徴量（パラレルデータ）を結合したベクトルを RBM の入力とする．この状態で RBM を学習させることで，隠れユニットを介して入力話者・出力話者の特徴ベクトルの共起性が学習されるため，推定される結合重みがパラレル辞書を表す．類似研究として Ling らによる RBM を用いた声質変換 [23] が挙げられるが，GMM による声質変換法の拡張（代替手法）として RBM を用いており，スパース表現に基づく声質変換においてパラレル辞書の学習手法に RBM を用いる本研究とは本質が異なる．また，我々の先行研究 [19], [21] では，話者ごとの学習データを用いて学習を行った RBM（もしくは DBN）で話者固有の潜在空間を求め，話者性を強調させた潜在特徴量同士をニューラルネットワークにより変換することで声質変換を行っていたが，本研究では入力話者と出力話者の結合ベクトルを RBM の可視層において学習を行う．

以下，2. 章ではスパース表現に基づく声質変換法とその一例として NMF を用いた手法について述べ，3. 章では本研究で用いる確率モデルである RBM について述べる．4. 章では RBM を用いたスパース表現に基づく声質変換提（提案手法）について述べる．5. 章で従来のスパース表現に基づく手法と比較する評価実験について述べ，6. 章で本論文をまとめる．

2. スパース表現に基づく声質変換

声質変換では，入力話者の音響特徴ベクトル $\boldsymbol{x} \in \mathbb{R}^D$ を出力話者の音響特徴ベクトル $\boldsymbol{y} \in \mathbb{R}^D$ へ変換する．一般的なスパース表現に基づく手法では，予め K 個の入力話者の辞書 $\mathcal{D}_x \in \mathbb{R}^{D\times K}$ と出力話者の辞書 $\mathcal{D}_y \in \mathbb{R}^{D\times K}$ のペア（パラレル辞書）を学習しておき，変換時には

$$\boldsymbol{x} \approx \mathcal{D}_x \boldsymbol{\alpha} \tag{1}$$

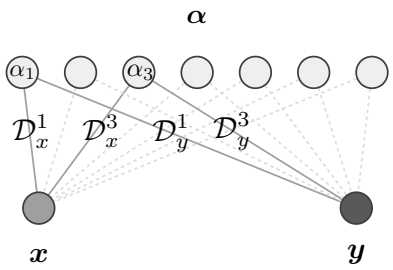

図 1　スパース表現に基づく声質変換の概要.

Fig. 1　Example of voice conversion based on sparse representation.

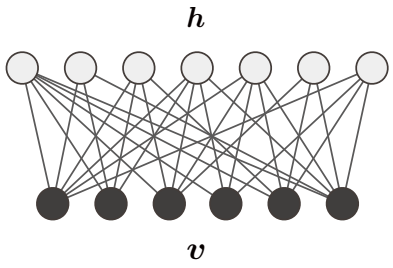

図 2　RBM のグラフ構造.

Fig. 2　Graphical representation of an RBM.

となるスパースな辞書選択重み $\boldsymbol{\alpha} \in \mathbb{R}^K, \|\boldsymbol{\alpha}\|_0 \ll K$ を何らかの手法で求め,

$$\boldsymbol{y} \approx \mathcal{D}_y \boldsymbol{\alpha} \tag{2}$$

を計算することで出力話者のベクトル $\boldsymbol{y}$ を得る（図 1）.

辞書 $\mathcal{D}_x, \mathcal{D}_y$ には，一般には入力話者・出力話者の学習データそのものを用いる場合が多い [15], [28], [29]．スパース重み $\boldsymbol{\alpha}$ の算出方法に関しては L1 正則化 [28] や K 近傍法 [29] など，様々な手法がこれまで用いられてきたが，近年ではより制約の強いスパース NMF を用いた手法 [15] も存在する．また，スパース NMF を用いる手法では，辞書行列に学習データをそのまま使うのではなく，学習によって得られる行列を用いる手法（SMNMF）も提案されている [16].

3. Restricted Boltzmann Machine

Restricted Boltzmann machine (RBM) は特殊な構造を持つ 2 層ネットワークであり，図 2 のように，可視層と隠れ層の確率変数分布を表現する無向グラフィカルモデルである [30]．元々 RBM はバイナリデータを入力させるモデルとして提案されていたが，後に連続値を入力させるモデル（Gaussian-Bernoulli RBM; GBRBM [31]）が考案された．しかしながらこのモデルは，分散項の影響で学習が不安定になるという問題 [18], [32] があったため，Cho らによって GBRBM の改良版 (Improved GBRBM; IGBRBM [33]) が提案された．この IGBRBM では，連続値の可視素子 $\boldsymbol{v} = [v_1, \cdots, v_I]^T, v_i \in \mathbb{R}$ と 2 値の隠れ素子 $\boldsymbol{h} = [h_1, \cdots, h_J]^T, h_j \in \{0, 1\}$ の同時確率 $p(\boldsymbol{v}, \boldsymbol{h})$ は，以下のように表される.

$$p(\boldsymbol{v}, \boldsymbol{h}) = \frac{1}{Z} e^{-E(\boldsymbol{v}, \boldsymbol{h})} \tag{3}$$

$$E(\boldsymbol{v}, \boldsymbol{h}) = \left\| \frac{\boldsymbol{v} - \boldsymbol{b}}{2\boldsymbol{\sigma}} \right\|^2 - \boldsymbol{c}^T \boldsymbol{h} - \left(\frac{\boldsymbol{v}}{\boldsymbol{\sigma}^2} \right)^T \boldsymbol{W} \boldsymbol{h} \tag{4}$$

$$Z = \sum_{\boldsymbol{v}, \boldsymbol{h}} e^{-E(\boldsymbol{v}, \boldsymbol{h})} \tag{5}$$

ここで，$\|\cdot\|^2$ は L2 ノルム，括線は要素除算を表す．$\boldsymbol{W} \in \mathbb{R}^{I \times J}$，$\boldsymbol{\sigma} \in \mathbb{R}^{I \times 1}$，$\boldsymbol{b} \in \mathbb{R}^{I \times 1}$，$\boldsymbol{c} \in \mathbb{R}^{J \times 1}$ はそれぞれ可視層-隠れ層間の重み行列，可視素子の偏差，可視素子のバイアス，隠れ素子のバイアスを示しており，いずれも推定すべきパラメータである.

RBM では可視素子間，または隠れ素子間の接続は存在しないため（つまり，それぞれの可視素子，隠れ素子は互いに条件付き独立であるため），それぞれの条件付き確率 $p(\boldsymbol{h}|\boldsymbol{v})$，$p(\boldsymbol{v}|\boldsymbol{h})$ は以下の様な単純な関数で表現される.

$$p(h_j = 1|\boldsymbol{v}) = \mathcal{S}(c_j + (\frac{\boldsymbol{v}}{\boldsymbol{\sigma}^2})^T \boldsymbol{W}_{:j}) \tag{6}$$

$$p(v_i = v|\boldsymbol{h}) = \mathcal{N}(v|b_i + \boldsymbol{W}_{i:}\boldsymbol{h}, \sigma_i^2) \tag{7}$$

ここで，$\boldsymbol{W}_{:j}$ と $\boldsymbol{W}_{i:}$ は $\boldsymbol{W}$ の第 j 行ベクトル，第 i 列ベクトルを表す．また，$\mathcal{S}(\cdot)$ は要素ごとのシグモイド関数 ($\mathcal{S}(\boldsymbol{x}) = \boldsymbol{1} \oslash (\boldsymbol{1} + e^{-\boldsymbol{x}})$)，$\mathcal{N}(\cdot|\mu, \sigma^2)$ は平均 μ，分散 σ^2 の正規分布を表す.

それぞれの RBM のパラメータ $\Theta = \{\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{\sigma}, \boldsymbol{c}\}$ は，N 個の観測データを $\{\boldsymbol{v}_n\}_{n=1}^N$ とするとき，この確率変数の対数尤度 $\mathcal{L} = \log \prod_n p(\boldsymbol{v}_n)$ を最大化するように推定される．この対数尤度をそれぞれのパラメータで偏微分すると,

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{ij}} = \langle \frac{v_i h_j}{\sigma_i^2} \rangle_{data} - \langle \frac{v_i h_j}{\sigma_i^2} \rangle_{model} \tag{8}$$

$$\frac{\partial \mathcal{L}}{\partial b_i} = \langle \frac{v_i}{\sigma_i^2} \rangle_{data} - \langle \frac{v_i}{\sigma_i^2} \rangle_{model} \tag{9}$$

$$\frac{\partial \mathcal{L}}{\partial c_j} = \langle h_j \rangle_{data} - \langle h_j \rangle_{model}, \tag{10}$$

が得られる．ただし，$\langle\cdot\rangle_{\mathrm{data}}$ と $\langle\cdot\rangle_{\mathrm{model}}$ はそれぞれ，観測データ，モデルデータの期待値を表す．しかし，一般に後者の期待値に関しては計算困難であるため，代わりに式 (6)(7) によって得られる再構築したデータの期待値 $\langle\cdot\rangle_{\mathrm{recon}}$ が用いられる（CD: Contrastive Divergence 法 [18]）．また，IGBRBM では分散を非負値に制約し，学習を安定化させるため $\sigma_i^2 = e^{z_i}$ と置き換える．これにより，z_i に関する勾配は以下のように計算される.

$$\frac{\partial \mathcal{L}}{\partial z_i} = e^{-z_i} \langle \frac{(v_i - b_i)^2}{2} - v_i \boldsymbol{W}_{i:} \boldsymbol{h} \rangle_{data} - e^{-z_i} \langle \frac{(v_i - b_i)^2}{2} - v_i \boldsymbol{W}_{i:} \boldsymbol{h} \rangle_{model} \tag{11}$$

それぞれのパラメータは式 (8)(9)(10) から，確率的勾配法を用いて繰り返し更新される（初期値はランダムに設定される）．すなわち,

$$\Theta^{(\mathrm{new})} = \Theta^{(\mathrm{old})} - \gamma \frac{\partial \mathcal{L}}{\partial \Theta} \tag{12}$$

ここで，γ は学習率を表す.

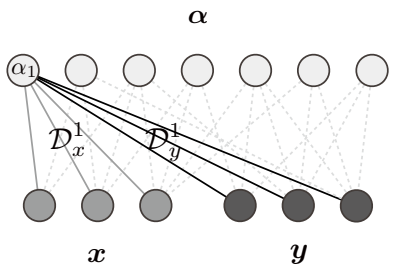

図 3　結合型 RBM を用いた声質変換の概要．それぞれの辞書選択重み（隠れ素子）が入力話者，出力話者の辞書と繋がっている．

Fig. 3　A joint RBM for voice conversion. Each hidden unit connects with the dictionaries of the source and the target speaker.

Input: Dictionaries $\mathcal{D}_x$, $\mathcal{D}_y$, a source speaker's vector $\boldsymbol{x}$ and an initial target vector $\boldsymbol{y}_0$

Output: Estimated target speaker's vector $\hat{\boldsymbol{y}}$

Initialize: Set the initial values as $\hat{\boldsymbol{y}} = \boldsymbol{y}_0$.

Repeat the following updates R times:

（1）　$\hat{\boldsymbol{\alpha}} \triangleq E\{\boldsymbol{\alpha}\}_{p(\boldsymbol{\alpha}|\boldsymbol{x},\hat{\boldsymbol{y}})} = \mathcal{S}(\mathcal{D}_x^T(\frac{\boldsymbol{x}}{\boldsymbol{\sigma}_x^2}) + \mathcal{D}_y^T(\frac{\hat{\boldsymbol{y}}}{\boldsymbol{\sigma}_y^2}) + \boldsymbol{c})$

（2）　$\hat{\boldsymbol{y}} \triangleq E\{\boldsymbol{\alpha}\}_{p(\boldsymbol{y}|\hat{\boldsymbol{\alpha}})} = \mathcal{D}_y\hat{\boldsymbol{\alpha}} + \boldsymbol{b}_y$

図 4　結合型 RBM を用いた出力話者ベクトルの反復的推定法．

Fig. 4　Iterative estimation algorithm of the target vector using a joint RBM.

4. RBM によるパラレル辞書学習と反復推定

本研究では式 (1)(2) の計算及び辞書の作成を同時に行うために，図 3 に示される結合型 RBM を用いる．前章で述べたように，RBM は可視層，隠れ層からなる 2 層ネットワークであり，層内は無く層間のみユニットの双方向結合が存在するという特徴がある．図のように，可視層ユニットを入力話者の音響特徴量 $\boldsymbol{x}$ と出力話者の音響特徴量 $\boldsymbol{y}$ の結合ベクトルとすれば，辞書選択重み $\boldsymbol{\alpha}_i$ が $\boldsymbol{x}$ と $\boldsymbol{y}$ へ，それぞれ i 番目の辞書 $\mathcal{D}_x^i$, $\mathcal{D}_y^i$ の重みが掛かっているネットワークと見ることができる．

今，学習用のパラレルデータ $(\boldsymbol{x}, \boldsymbol{y})$ が与えられたとき，$\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\alpha}$ の同時確率を以下のように表す．

$$p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\alpha}; \mathcal{D}_x, \mathcal{D}_y) = \frac{1}{Z} e^{-E(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\alpha}; \mathcal{D}_x, \mathcal{D}_y)} \tag{13}$$

$$E(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\alpha}; \mathcal{D}_x, \mathcal{D}_y) = \left\| \frac{\boldsymbol{x} - \boldsymbol{b}_x}{2\boldsymbol{\sigma}_x} \right\|^2 + \left\| \frac{\boldsymbol{y} - \boldsymbol{b}_y}{2\boldsymbol{\sigma}_y} \right\|^2 - \boldsymbol{c}^T\boldsymbol{\alpha} - \left(\frac{\boldsymbol{x}}{\boldsymbol{\sigma}_x^2}\right)^T \mathcal{D}_x\boldsymbol{\alpha} - \left(\frac{\boldsymbol{y}}{\boldsymbol{\sigma}_y^2}\right)^T \mathcal{D}_y\boldsymbol{\alpha} \tag{14}$$

ただし，$Z = \sum_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{\alpha}} e^{-E(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\alpha})}$ は正規化項，$\boldsymbol{c}$ は辞書選択重みのバイアスパラメータを表す．また，$\boldsymbol{b}_x$，$\boldsymbol{\sigma}_x$ はそれぞれ入力話者音響特徴量のバイアス，偏差パラメータを表し，$\boldsymbol{b}_y$，$\boldsymbol{\sigma}_y$ はそれぞれ出力話者音響特徴量のバイアス，偏差パラメータを表す．辞書 $\mathcal{D}_x, \mathcal{D}_y$ (とその他のパラメータ) は (13) 式を $\boldsymbol{\alpha}$ で周辺化した尤度 $\mathcal{L} = p(\boldsymbol{x}, \boldsymbol{y}) = \sum_{\boldsymbol{\alpha}} p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\alpha})$ を最大化するように求められる．この尤度をそれぞれのパラメータで偏微分すると，

$$\frac{\partial \mathcal{L}}{\partial \mathcal{D}_{x\,dk}} = \langle \frac{x_d \alpha_k}{\sigma_{x_d}^2} \rangle_{data} - \langle \frac{x_d \alpha_k}{\sigma_{x_d}^2} \rangle_{model} \tag{15}$$

$$\frac{\partial \mathcal{L}}{\partial \mathcal{D}_{y\,dk}} = \langle \frac{y_d \alpha_k}{\sigma_{y_d}^2} \rangle_{data} - \langle \frac{y_d \alpha_k}{\sigma_{y_d}^2} \rangle_{model} \tag{16}$$

が得られるが，実際は前章で述べたように CD 法によって第二項を近似する．その他のパラメータに関しても，(9)(10)(11) 式と同様に求めることができる．

変換の際には，図 4 に示すように，RBM の前方推論・後方推論の繰り返しによって出力話者の音響特徴量 $\hat{\boldsymbol{y}}$ を推定する．まず，初期値 $\boldsymbol{y}_0$ を任意に決めておき，それぞれの辞書が選択される確率

$$p(\boldsymbol{\alpha} = \mathbf{1}|\boldsymbol{x}, \boldsymbol{y}) = \mathcal{S}(\mathcal{D}_x^T(\frac{\boldsymbol{x}}{\boldsymbol{\sigma}_x^2}) + \mathcal{D}_y^T(\frac{\boldsymbol{y}}{\boldsymbol{\sigma}_y^2}) + \boldsymbol{c}) \tag{17}$$

を用いて $\boldsymbol{\alpha}$ の期待値を計算する．次に，この $\hat{\boldsymbol{\alpha}}$ を用いて $\boldsymbol{y}$ の期待値を計算する．ただし，$\boldsymbol{y}$ の条件付き確率は RBM の後方推論によって以下のように表される．

$$p(\boldsymbol{y}|\boldsymbol{\alpha}) = \mathcal{N}(\boldsymbol{y}|\mathcal{D}_y\boldsymbol{\alpha} + \boldsymbol{b}_y, \boldsymbol{\sigma}_y^2) \tag{18}$$

以上の手順（$\boldsymbol{\alpha}$ と $\boldsymbol{y}$ の推定）を R 回繰り返し，反復的に $\hat{\boldsymbol{y}}$ を求める．初期値 $\boldsymbol{y}_0$ として，入力話者音声のベクトル $\boldsymbol{x}$ を用いる手法，GMM など他の手法の推定結果を用いる手法などが考えられるが，本研究では要素が全てゼロのベクトル $\mathbf{0}$ を用いる．

(13) 式に示されるように，提案法では学習データの尤度だけでなく，辞書選択重みの尤度も考慮して同時に最適化を行っている．この点は従来のスパース表現に基づく手法である SMNMF と同様だが，第一に辞書選択重みの推定法が大きく異る．SMNMF では線形変換により辞書選択重みを推定しているが，提案法では (17) 式のように非線形変換を用いている．また，SMNMF では入力データが非負値に限定されていたが，本手法ではそういった制約はないため，実数値を取るデータを入力させることができる．特に，分布が単峰的になりやすい音声の MFCC は，入力データにガウス分布を仮定している提案法との相性が良い．

5. 評 価 実 験

5.1 実 験 条 件

本実験では，ATR の日本語音声データベース A セット [34] を用いて，提案手法である結合型 RBM を用いた手法（“JRBM”）と，従来のスパース表現に基づく声質変換としてサンプルベースの NMF を用いた手法 (“EXNMF”)，学習ベースの NMF を用いた手法 (“SMNMF”) [16] との間で，声質変換精度の比較を行った．また，参考として代表的な声質変換手法である GMM(64 混合) とも比較した．このデータベースから，入力話者として男性話者（MMY），出力話者として女性話者（FTK）を選んだ．提案手法と GMM の入力（出力）音響特徴量として，STRAIGHT スペクトル [35] から計算された 24 次元の MFCC

表 1　各手法の客観評価値.

Table 1　Performance of each method.

	Joint RBM (Proposed)			Spectral mapping (NMF)		Exemplar-based (NMF)		GMM
Method	JRBM-96	JRBM-192	JRBM-384	SMNMF-1000	SMNMF-2500	EXNMF-1000	EXNMF-58426	GMM-64
SDIR (dB)	5.21	**5.32**	4.45	5.14	4.68	4.91	5.23	4.11

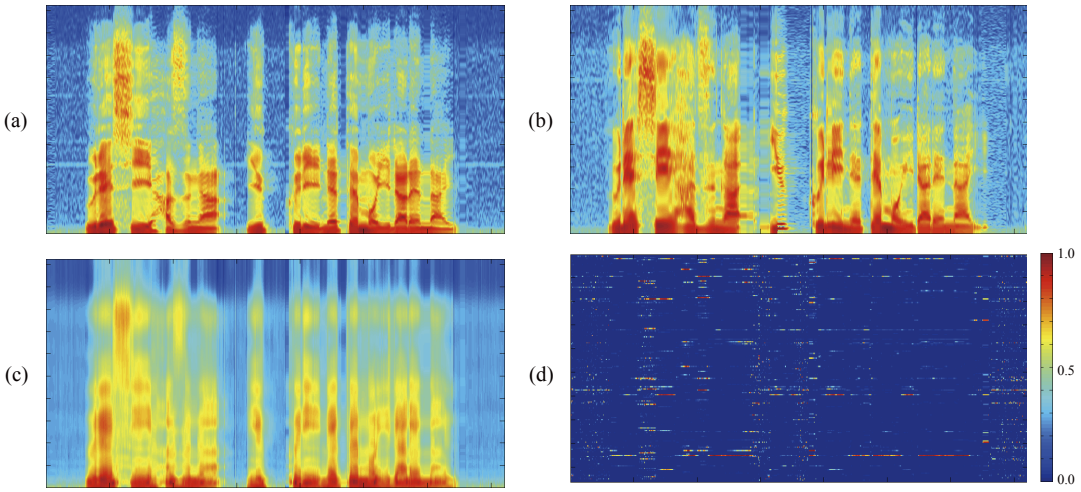

Fig. 4　Examples of voice conversion using an utterance “ureshii hazuga yukkuri netemo irarenai”. (a) Spectrogram of the male source speaker, (b) spectrogram of the female target speaker, (c) converted spectrogram obtained by JRBM-192, and (d) expectation values of the estimated $\boldsymbol{\alpha}$ (vertical and horizontal axes indicate the index and the time, respectively).

図 5　評価データ「嬉しいはずが，ゆっくり寝てもいられない」の推定結果例．(a) 入力話者音声のスペクトル（変換前のスペクトル），(b) 出力話者音声のスペクトル，(c) 提案法 JRBM-192 による変換後のスペクトル，(d) 推定された辞書選択重み $\boldsymbol{\alpha}$ の期待値（縦軸，横軸はそれぞれ重みのインデックス，時間を表す）.

を用いた．NMF ベースの手法では 513 次元の STRAIGHT スペクトルを用いた．音声合成時には，文献 [36] に述べられているソースフィルターモデルを用いて，MFCC から STRAIGHT スペクトルへ逆変換し，STRAIGHT 合成によって変換音声信号を得た．入力・出力話者のパラレルデータは，動的計画法により作成された．RBM の学習の際の学習率，繰り返し回数はともにそれぞれ 0.01，400 を用いた．隠れ層における素子数を 96，192，384 と変えて比較を行った (それぞれ “JRBM-96”，“JRBM-192”，“JRBM-384” と表記する)．また，変換時における繰り返し回数 R は，十分に速く収束していたため，$R = 5$ とした．学習には A セットの 216 単語（58426 フレーム）を用いている．評価データとして，学習データには含まれていない 25 文の発話音声を用いた．学習ベースの NMF では，基底数を 1,000 (“SMNMF-1000”)，2,500 (“SMNMF-2500”) として比較を行った．サンプルベースの NMF では，学習フレームを全て用いた場合 (“EXNMF-58426”)，ランダムに 1,000 フレーム用いた場合 (“EXNMF-1000”) で比較した．客観評価基準として，SDIR (spectral distortion improvement ratio) を用いた．フレームごとにこの SDIR を求め，全フレームの平均 SDIR を算出することで，各手法による変換精度を比較した．

5.2　実験結果と考察

各手法による SDIR を表 1 にまとめた．表 1 から，提案手法である “JRBM-192” が他の手法と比べて最も高い精度が得られたことが分かる．特に学習ベースの NMF の結果と比較すれば，提案法による，非負値に縛られない非線形変換をベースにした反復推定法が効果的であったと考えられる．提案法の中で比較すると，素子の数が多すぎても少なすぎても精度が低下している．素子の数が少なすぎるとデータ分布を十分に表現できず，多すぎると過剰なモデルとなるからだと考えられ，自動的に適切な素子の数を選ぶ手法に関しては検討中である．

また，図 5 は最も精度の良かった提案法 (JRBM-192) による推定結果の例を示している．特に，図 5(d) に示すように，明示的にスパース制約を用いていないにもかかわらず，推定された辞書選択重みがスパース（ほとんどの値がゼロ）になっていることが分かる．これは，各素子においてなるべく情報が重複しないようにパラメータが推定されるという過程で，自然とスパースになる RBM の性質によるものと考えられる．

6.　おわりに

本研究では，少数の辞書を用いて出力話者音声へ変換するスパース表現に基づく声質変換において，従来の NMF による手法では表現のできないモデルとして，入力話者・出力話者を結合させた RBM を用いた声質変換法を提案した．評価実験では提案法により，従来のスパース表現に基づく手法（サンプルベー

スの NMF と学習ベースの NMF）や GMM よりも客観的に優れた変換音声が得られた．今後の課題として，deep Boltzmann machine (DBM) を用いた，より深い階層構造を持つ声質変換法への拡張を検討していきたい．

文　献

[1] A. Kain and M. W. Macon: “Spectral voice conversion for text-to-speech synthesis”, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 285–288 (1998).

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

[3] K. Nakamura, T. Toda, H. Saruwatari and K. Shikano: “Speaking-aid systems using gmm-based voice conversion for electrolaryngeal speech”, Speech Communication, **54**, 1, pp. 134–146 (2012).

[4] L. Deng, A. Acero, L. Jiang, J. Droppo and X. Huang: “High-performance robust speech recognition using stereo training data”, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 301–304 (2001).

[5] A. Kunikoshi, Y. Qiao, N. Minematsu and K. Hirose: “Speech generation from hand gestures based on space mapping”, Proc. Interspeech, pp. 308–311 (2009).

[6] R. Gray: “Vector quantization”, IEEE ASSP Magazine, **1**, 2, pp. 4–29 (1984).

[7] H. Valbret, E. Moulines and J.-P. Tubach: “Voice transformation using PSOLA technique”, Speech Communication, **11**, 2, pp. 175–187 (1992).

[8] Y. Stylianou, O. Cappé and E. Moulines: “Continuous probabilistic transform for voice conversion”, IEEE Transactions on Speech and Audio Processing, **6**, 2, pp. 131–142 (1998).

[9] T. Toda, A. W. Black and K. Tokuda: “Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory”, IEEE Transactions on Audio, Speech, and Language Processing, **15**, 8, pp. 2222–2235 (2007).

[10] E. Helander, T. Virtanen, J. Nurminen and M. Gabbouj: “Voice conversion using partial least squares regression”, IEEE Transactions on Audio, Speech, and Language Processing, **18**, 5, pp. 912–921 (2010).

[11] Z.-H. Ling, L. Deng and D. Yu: “Modeling spectral envelopes using restricted boltzmann machines and deep belief networks for statistical parametric speech synthesis”, IEEE Transactions on Audio, Speech, and Language Processing, 10, pp. 2129–2139 (2013).

[12] Z. Wu, T. Virtanen, T. Kinnunen, E. S. Chng and H. Li: “Exemplar-based unit selection for voice conversion utilizing temporal information”, Proc. INTERSPEECH, pp. 3057–3061 (2013).

[13] H.-T. Hwang, Y. Tsao, H.-M. Wang, Y.-R. Wang and S.-H. Chen: “Alleviating the over-smoothing problem in gmm-based voice conversion with discriminative training”, Proc. INTERSPEECH, pp. 3062–3066 (2013).

[14] D. D. Lee and H. S. Seung: “Algorithms for non-negative matrix factorization”, Advances in neural information processing systems, pp. 556–562 (2000).

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

[16] R. Takashima, R. Aihara, T. Takiguchi and Y. Ariki: “Noise-robust voice conversion based on spectral mapping on sparse space”, SSW8, pp. 71–75 (2013).

[17] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black and K. Prahallad: “Voice conversion using artificial neural networks”, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)IEEE, pp. 3893–3896 (2009).

[18] G. E. Hinton, S. Osindero and Y.-W. Teh: “A fast learning algorithm for deep belief nets”, Neural computation, **18**, 7, pp. 1527–1554 (2006).

[19] T. Nakashika, R. Takashima, T. Takiguchi and Y. Ariki: “Voice conversion in high-order eigen space using deep belief nets”, Proc. Interspeech, pp. 369–372 (2013).

[20] G. W. Taylor, G. E. Hinton and S. T. Roweis: “Modeling human motion using binary latent variables”, Advances in neural information processing systems, pp. 1345–1352 (2006).

[21] T. Nakashika, T. Takiguchi and Y. Ariki: “Speaker-dependent conditionl restricted boltzmann machine for voice conversion”, **113**, 366, pp. 83–88 (2013).

[22] Z. Wu, E. S. Chng and H. Li: “Conditional restricted boltzmann machine for voice conversion”, IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP) (2013).

[23] L.-H. Chen, Z.-H. Ling, Y. Song and L.-R. Dai: “Joint spectral distribution modeling using restricted boltzmann machines for voice conversion”, Proc. Interspeech, pp. 3052–3056 (2013).

[24] J. Gao, Y. Guo and M. Yin: “Restricted boltzmann machine approach to couple dictionary training for image super-resolution”, IEEE International Conference on Image Processing, pp. 499–503 (2013).

[25] V. Nair and G. E. Hinton: “3D object recognition with deep belief nets.”, NIPS, pp. 1339–1347 (2009).

[26] T. Deselaers, S. Hasan, O. Bender and H. Ney: “A deep learning approach to machine transliteration”, Fourth Workshop on Statistical Machine TranslationAssociation for Computational Linguistics, pp. 233–241 (2009).

[27] A.-r. Mohamed, G. E. Dahl and G. Hinton: “Acoustic modeling using deep belief networks”, IEEE Transactions on Audio, Speech, and Language Processing, **20**, 1, pp. 14–22 (2012).

[28] D. L. Donoho: “Compressed sensing”, IEEE Transactions on Information Theory, **52**, 4, pp. 1289–1306 (2006).

[29] H. Chang, D.-Y. Yeung and Y. Xiong: “Super-resolution through neighbor embedding”, Computer Vision and Pattern Recognition, Vol. 1IEEE, pp. 275–282 (2004).

[30] Y. Freund and D. Haussler: “Unsupervised learning of distributions of binary vectors using two layer networks”, Computer Research Laboratory (1994).

[31] G. E. Hinton and R. R. Salakhutdinov: “Reducing the dimensionality of data with neural networks”, Science, **313**, 5786, pp. 504–507 (2006).

[32] A. Krizhevsky and G. Hinton: “Learning multiple layers of features from tiny images”, Computer Science Department, University of Toronto, Tech. Rep (2009).

[33] K. Cho, A. Ilin and T. Raiko: “Improved learning of gaussian-bernoulli restricted boltzmann machines”, Artificial Neural Networks and Machine Learning–ICANN 2011, Springer, pp. 10–17 (2011).

[34] 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, **9**, 4, pp. 357–363 (1990).

[35] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino and H. Banno: “TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation”, ICASSPIEEE, pp. 3933–3936 (2008).

[36] B. Milner and X. Shao: “Speech reconstruction from mel-frequency cepstral coefficients using a source-filter model”, Proc. Interspeech, pp. 2421–2424 (2002).