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
SP2013-88(2013-12)

# 話者依存型Conditional Restricted Boltzmann Machineによる声質変換

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.scite.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

**あらまし**　本研究では，元の音響特徴量空間よりも音韻性や時間変化性を抑え，話者性を強調させることによって，より入力話者音声の声質を出力話者のものへと変換しやすい話者依存空間を形成することを目的として，話者ごとに conditional restricted Boltzmann machine (CRBM) を用いた声質変換法を提案する．提案手法ではまず初めに，話者ごとに用意した学習データ（パラレルデータである必要は無い）を用いて，入力話者，出力話者の CRBM を独立に学習させる．次に，少量のパラレルデータの音響特徴量を，それぞれの CRBM を通して話者依存高次元空間へ写像（CRBM の前方推論）し，その高次特徴量同士を Neural Network (NN) を用いて変換させる．NN の変換で得られた特徴量は，CRBM の後方推論によって元の音響特徴量へ逆変換することが可能である．評価実験では，従来の GMM や NN，DBN を用いた声質変換法に比べて，主観的にも客観的にも良い精度が得られたことを確認した．
**キーワード**　声質変換，conditional restricted Boltzmann machine, deep learning, 話者強調

# Speaker-dependent conditionl restricted Boltzmann machine for 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.scite.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

**Abstract**　In this paper, we present a voice conversion (VC) method that utilizes conditional restricted Boltzmann machines (CRBMs) for each speaker to obtain time-invariant speaker-independent spaces where voice features are converted more easily than those in an original acoustic feature space. First, we train two CRBMs for a source and target speaker independently using speaker-dependent training data (without the need to parallelize the training data). Then, a small number of parallel data are fed into each CRBM and the high-order features produced by the CRBMs are used to train a concatenating neural network (NN) between the two CRBMs. Finally, the entire network (the two CRBMs and the NN) is fine-tuned using the acoustic parallel data. Through voice-conversion experiments, we confirmed the high performance of our method in terms of objective and subjective evaluations, comparing it with conventional GMM, NN, and speaker-dependent DBN approaches.
**Key words**　Voice conversion, conditional restricted Boltzmann machine, deep learning, speaker specific features

## 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 GMM の過学習問題に対応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では，準線形変換をベースにしており，いわば “浅い” 変換に基づいているため，特徴間の詳細な対応付けが艱難である等，変換の精度には限界があった．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 “深い” 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 [11] 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 [12]) を用いた多層型声質変換法 [13]，Wu らによる CRBM (conditional restricted Boltzmann machine [14]) を用いた非線形声質変換法 [15] 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 [11], [13], [15].

より精度の高い声質変換を実現するためには，音声信号から，(i) 音韻情報に依存せず，(ii) 時間的な揺らぎも抑制し，なるべく (iii) 話者性を強調させた特徴量を抽出できることが望ましい．本稿では，我々の先行研究である話者依存型 RBM (若しくは DBN) に基づく声質変換法 [13] を拡張し，時間的情報と，入力話者-出力話者間の潜在的関係性を一つのネットワークとして体系的に捉えるために，話者依存型 CRBM と高次特徴量間を結ぶ NN を組み合わせた声質変換手法を提案する．CRBM は時系列データを表現するための非線形確率グラフィカルモデルの一つであり，現時刻における可視層と隠れ層との間の無向グラフ，過去の可視層から現時刻の可視層への有向グラフ，過去の可視層から現時刻の隠れ層への有向グラフの 3 つの要素で構成される．本手法を用いることによる利点に関しては後述する．

提案手法では，まず入力話者音声，出力話者音声の音響特徴量（例えば MFCC など）を用いて，それぞれの話者ごとに独立して CRBM を学習させる．次に，入力・出力話者でフレーム対応のとれた音響特徴量（パラレルデータ）を，それぞれの CRBM を用いて高次元特徴空間へ射影（前方推論）し，入力話者の高次特徴量を入力信号，出力話者の高次特徴量を教師信号として，高次特徴量間の非線形変換 NN（結合 NN）を学習させる．変換時にはこの NN の出力ベクトルから CRBM の逆射影（後方推論）を用いることで，元の音響特徴量へ戻すことで，音声信号を得る．結局，入力話者の音響特徴量は，入力話者 CRBM の前方推論，結合 NN，出力話者 CRBM の後方推論を経ることで出力話者の音響特徴量へ変換することができるが，これらの流れは一つの非線形多層ネットワークとみなすこともでき，パラメータの微調整が可能である．ここで，話者依存型 CRBM の学習に用いる音声信号は，(単一話者の音声であるため) 話者性は変わらずに，様々な音韻情報が含まれているため，それぞれの CRBM の隠れ層は，限られた数の素子で学習データを最大限に表現する（話者性を豊富に含み，音韻情報を抑制した）潜在特徴を捉えようとすることが期待される．さらに，それぞれの CRBM には時系列特徴ベクトルの集合を入力するため，過去から現時刻への有向グラフが特徴ベクトルの時間変化に関する情報を吸収し，その結果，可視層-隠れ層間の無向グラフではそれ以外の（つまり，時間変化には依存しない）潜在特徴を捉えることに集中できると期待できる．このようにして，本研究では，時間非依存でかつ話者性を強調させた高次元特徴空間において特徴ベクトルを変換させることで，より精度の高い声質変換を目指す．

## 2. Conditional Restricted Bolzmann Machine

提案手法では，CRBM (conditional restricted Boltzmann machine) を用いて，声質の変換の行い易い高次元空間で特徴変換を行う．本章ではまず，基礎技術となる RBM (restricted Boltzmann machine) について述べ，続いて CRBM について説明する．

### 2.1 RBM

RBM は特殊な構造を持つ 2 層ネットワークであり，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 [16]．RBM では，2 値の可視素子 $\boldsymbol{v} = [v_1, \cdots, v_I]^T, v_i \in \{0,1\}$ と隠れ素子 $\boldsymbol{h} = [h_1, \cdots, h_J]^T, h_j \in \{0,1\}$ の同時確率 $p(\boldsymbol{v}, \boldsymbol{h})$ は，以下のように表される．

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

$$E(\boldsymbol{v}, \boldsymbol{h}) = -\boldsymbol{b}^T \boldsymbol{v} - \boldsymbol{c}^T \boldsymbol{h} - \boldsymbol{v}^T \boldsymbol{W} \boldsymbol{h} \tag{2}$$

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

ここで，$\boldsymbol{W} \in \mathbb{R}^{I \times J}$，$\boldsymbol{b} \in \mathbb{R}^{I \times 1}$，$\boldsymbol{c} \in \mathbb{R}^{J \times 1}$ はそれぞれ可視層-隠れ層間の重み行列，可視層のバイアス，隠れ層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 $p(\boldsymbol{h}|\boldsymbol{v})$，$p(\boldsymbol{v}|\boldsymbol{h})$ (厳密にはそれぞれの素子が発火する条件付き確率) は以下の様な単純な関数で表現される．

$$p(h_j = 1|\boldsymbol{v}) = \sigma(c_j + \boldsymbol{v}^T \boldsymbol{W}_{:j}) \tag{4}$$

$$p(v_i = 1|\boldsymbol{h}) = \sigma(b_i + \boldsymbol{h}^T \boldsymbol{W}_{i:}^T) \tag{5}$$

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

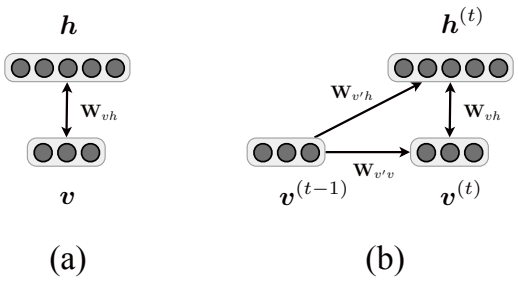


図 1　Comparison between an RBM (a) and a CRBM (b).

ルをあ表す．また，$\sigma(x)$ は要素ごとのシグモイド関数を表す $(\sigma(\boldsymbol{x}) = \mathbf{1} \oslash (\mathbf{1} + e^{-\boldsymbol{x}}))$．

それぞれの RBM のパラメータ $\Theta = \{\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{c}\}$ 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，

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{ij}} = \langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{model}} \tag{6}$$

$$\frac{\partial \mathcal{L}}{\partial b_i} = \langle v_i \rangle_{\text{data}} - \langle v_i \rangle_{\text{model}} \tag{7}$$

$$\frac{\partial \mathcal{L}}{\partial c_j} = \langle h_j \rangle_{\text{data}} - \langle h_j \rangle_{\text{model}} \tag{8}$$

が得られる．ただし，$\langle \cdot \rangle_{\text{data}}$ と $\langle \cdot \rangle_{\text{model}}$ 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 (4)(5) によって得られる再構築したデータの期待値 $\langle \cdot \rangle_{\text{recon}}$ が用いられる（CD: Contrastive Divergence 法 [12]）．

それぞれのパラメータは式 (6)(7)(8) 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，

$$\Theta^{(\text{new})} = \Theta^{(\text{old})} - \alpha \frac{\partial \mathcal{L}}{\partial \Theta} \tag{9}$$

ここで，$\alpha$ は学習率を表す．

### 2.2　CRBM

CRBM は Taylor ら [14] によって提案された RBM の拡張モデルであり，時系列データを取り扱うことに適している．RBM における可視層-隠れ層間の無向グラフに加えて，CRBM では過去 $P$ フレーム前までの可視層集合 $\{\boldsymbol{v}^{(p)}\}_{p=t-P}^{t}, \boldsymbol{v}^{(p)} = [v_1^{(p)}, \cdots, v_I^{(p)}]^T, v_i^{(p)} \in \{0,1\}$ から，現時刻 $t$ における隠れ層 $\boldsymbol{h}^{(t)} = [h_1^{(t)}, \cdots, h_J^{(t)}]^T, h_j^{(t)} \in \{0,1\}$ への有向グラフを考慮したモデルとなっている（RBM との違いを図 1 に示す）．簡単のため，本研究では $P = 1$ とする．このモデルでは，図 1(b) にように，3 種類の推定すべき重みパラメータ：$\boldsymbol{W}_{v'v} \in \mathbb{R}^{I \times I}$（$\boldsymbol{v}^{(t-1)}$ から $\boldsymbol{v}^{(t)}$ への有向重み行列），$\boldsymbol{W}_{v'h} \in \mathbb{R}^{I \times J}$（$\boldsymbol{v}^{(t-1)}$ から $\boldsymbol{h}^{(t)}$ への有向重み行列），$\boldsymbol{W}_{vh} \in \mathbb{R}^{I \times J}$（$\boldsymbol{v}^{(t)}$ と $\boldsymbol{h}^{(t)}$ の間の無向重み行列）が存在する．これらの重みパラメータは，RBM と同様に，CD 法を用いて最適化される．CRBM の場合，過去のデータを観測した時の条件付き確率密度は以下のように表される．

$$p(\boldsymbol{v}^{(t)} | \boldsymbol{v}^{(t-1)}) = \frac{1}{Z} \sum_{\boldsymbol{h}^{(t)}} e^{-E(\boldsymbol{v}^{(t)}, \boldsymbol{h}^{(t)} | \boldsymbol{v}^{(t-1)})} \tag{10}$$

ここで $Z$ は正規化項（式 (3)）を表し，$E$ は以下のエネルギー関数を示している．

$$\begin{aligned} E(\boldsymbol{v}^{(t)}, \boldsymbol{h}^{(t)} | \boldsymbol{v}^{(t-1)}) = &-\boldsymbol{b}^T \boldsymbol{v}^{(t)} - \boldsymbol{c}^T \boldsymbol{h}^{(t)} - \boldsymbol{v}^{(t)^T} \boldsymbol{W}_{vh} \boldsymbol{v}^{(t)} \\ &- \boldsymbol{v}^{(t-1)^T} \boldsymbol{W}_{v'v} \boldsymbol{v}^{(t)} - \boldsymbol{v}^{(t-1)^T} \boldsymbol{W}_{v'h} \boldsymbol{h}^{(t)} \end{aligned} \tag{11}$$

RBM の場合と同様に，観測データの対数尤度 $\mathcal{L} = \log \prod_t p(\boldsymbol{v}^{(t)} | \boldsymbol{v}^{(t-1)})$ をそれぞれのパラメータで偏微分すると，

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{v'v_{i'i}}} = \langle v_i^{(t)} v_{i'}^{(t-1)} \rangle_{\text{data}} - \langle v_i^{(t)} v_{i'}^{(t-1)} \rangle_{\text{model}} \tag{12}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{v'h_{i'j}}} = \langle v_{i'}^{(t-1)} h_j^{(t)} \rangle_{\text{data}} - \langle v_{i'}^{(t-1)} h_j^{(t)} \rangle_{\text{model}} \tag{13}$$

が得られる．無向グラフに関するその他のパラメータの偏微分（$\boldsymbol{W}_{vh}$，$\boldsymbol{b}$，$\boldsymbol{c}$）に関してはそれぞれ式 (6)(7)(8) と同様にして導出される．

一度パラメータが推定されれば，CRBM の前方推論（$\boldsymbol{v}^{(t)}$ と $\boldsymbol{v}^{(t-1)}$ が与えられたときの $\boldsymbol{h}^{(t)}$ の条件付き確率）と後方推論（$\boldsymbol{h}^{(t)}$ と $\boldsymbol{v}^{(t-1)}$ が与えられたときの $\boldsymbol{v}^{(t)}$ の条件付き確率）は，それぞれ以下のように計算される．

$$p(h_j^{(t)} = 1 | \boldsymbol{v}^{(t)}, \boldsymbol{v}^{(t-1)}) = \sigma(c_j + \boldsymbol{v}^{(t)^T} \boldsymbol{W}_{vh_{:j}} + \boldsymbol{v}^{(t-1)^T} \boldsymbol{W}_{v'h_{:j}}) \tag{14}$$

$$p(v_i^{(t)} = 1 | \boldsymbol{h}^{(t)}, \boldsymbol{v}^{(t-1)}) = \sigma(b_i + \boldsymbol{h}^{(t)^T} \boldsymbol{W}_{vh_{i:}}^T + \boldsymbol{v}^{(t-1)^T} \boldsymbol{W}_{v'v_{:j}}) \tag{15}$$

## 3.　提 案 手 法

一般に，音声信号には音韻情報と発話者の話者性に関する情報が含まれている．声質変換の目的は，入力話者音声の音韻情報は変更せずに，話者性のみを出力話者の声質へ変換することであるので，音声信号から，話者性のみを強調させた特徴量を抽出できることが望ましい．本研究では，話者依存型 CRBM (conditional restricted Boltzmann machine) と高次特徴変換 NN (Neural Network) を組み合わせて，音響特徴量から特定話者の話者性を強調させた高次元特徴量抽出，話者性に関する情報の変換，高次元特徴量から音韻情報を含む元の音響特徴量へ逆変換というプロセスを，統一的な枠組みで実行する手法を提案する．

提案手法による声質変換の概要を図 2 に示す．我々のアプローチでは，予め話者ごとに独立して，その発話者のみの音声データを使って CRBM の特定話者モデルを学習させておく（図 2 (a)）．変数 $\boldsymbol{x}^{(t)}$ と $\boldsymbol{y}^{(t)}$（もしくは $\boldsymbol{x}^{(t-1)}$ と $\boldsymbol{y}^{(t-1)}$）は時刻 $t$（もしくは $t-1$）での，入力話者と出力話者の CRBM における可視素子であり，例えば MFCC などの音響特徴量ベクトルを表す．2.2 節で述べた CRBM は 2 値データを入力していたが，一般に音響特徴量が 2 値ベクトルであるとは限らない．そこで本研究では，音響特徴量ベクトルを予めシグモイド関数を用いて，近似的に 2 値化しておく．

例えば入力話者に関しては，パラメータ行列 $\boldsymbol{W}_{xh}$ は $T$ 個のセグメント型学習サンプルの尤度 $p(\boldsymbol{x}) = \prod_{t=1}^{T} p(\boldsymbol{x}^{(t)}|\boldsymbol{x}^{(t-1)})$ を最大化するように推定される（ただし $\boldsymbol{x}^{(0)} = \boldsymbol{0} \in \mathbb{R}^{I\times 1}$）．それぞれの隠れ素子 $\boldsymbol{h}_x^{(t)}$ は互いに条件付き独立なので，可視素子（学習データ）に見られる共通した特徴を表すように作用する．今回用いる学習データは音韻に関する情報は多様に変化するが，話者に依存した情報は変化しない．それゆえに，抽出された高次特徴量 $\boldsymbol{h}_x^{(t)}$ は，音韻情報を抑制し，（どんな学習サンプルにも共通した）話者情報を捉えた特徴量であると期待できる．さらに，式 (11) で示されるように，有向重み $\boldsymbol{W}_{x'h}$ と $\boldsymbol{W}_{x'x}$ は無向重み $\boldsymbol{W}_{xh}$ と同時に推定されるので，有向重みが学習データの時間変化に関する情報を吸収し，その結果 $\boldsymbol{W}_{xh}$ による射影によって，それ以外の情報（時間変化に関係のない話者性）を強調させる手助けができると考えられる．

提案手法では，このようにして得られた話者性が強調された特徴量同士（$\boldsymbol{h}_x^{(t)}$ から $\boldsymbol{h}_y^{(t)}$ へ）を，図 2 (b) のように NN を用いて変換させる（便宜上 $L$ を NN の隠れ層の数とし，$0 \leqq L \leqq 1$ とする）．この NN の学習には，$T'$ フレーム(注1)のパラレルデータ $\{\boldsymbol{x}_t, \boldsymbol{y}_t\}_{t=0}^{T'}$ を用いる．この NN は，入力話者音響特徴量を入力話者 CRBM で射影した高次特徴量 $\boldsymbol{h}_x^{(t)}$ を入力，出力話者音響特徴量を出力話者 CRBM で射影した高次特徴量 $\boldsymbol{h}_y^{(t)}$ を教師信号として学習を行うが，それらの特徴量は式 (14)(15) から以下のように計算される．

$$\boldsymbol{h}_x^{(t)} = \sigma(\boldsymbol{W}_{xh}\boldsymbol{x}^{(t)} + \boldsymbol{W}_{x'h}\boldsymbol{x}^{(t-1)} + \boldsymbol{c}_x) \tag{16}$$

$$\boldsymbol{h}_y^{(t)} = \sigma(\boldsymbol{W}_{yh}\boldsymbol{y}^{(t)} + \boldsymbol{W}_{y'h}\boldsymbol{y}^{(t-1)} + \boldsymbol{c}_y) \tag{17}$$

ただし，$\boldsymbol{c}_x$ と $\boldsymbol{c}_y$ は入力話者 CRBM，出力話者 CRBM の前方推論の際のバイアスベクトルを表す．NN のパラメータ $\{\boldsymbol{W}_l, \boldsymbol{d}_l\}_{l=0}^{L}$ は典型的な NN の枠組みと同様に，NN の出力ベクトル $\eta(\boldsymbol{h}_x^{(t)})$ と教師ベクトル $\boldsymbol{h}_y^{(t)}$ の差を最小化するように推定される．一度パラメータが推定されれば，入力ベクトル $\boldsymbol{h}_x^{(t)}$ は以下のように出力話者の特徴量へ変換される．

$$\eta(\boldsymbol{h}_x^{(t)}) = \bigodot_{l=0}^{L} \eta_l(\boldsymbol{h}_x^{(t)}) \tag{18}$$

$$\eta_l(\boldsymbol{h}_x^{(t)}) = \sigma(\boldsymbol{W}_l\boldsymbol{h}_x^{(t)} + \boldsymbol{d}_l) \tag{19}$$

ただし，$\bigodot_{l=0}^{L}$ は $L+1$ 個の合成関数を表す（例えば隠れ層を 1 つ持つ NN の場合，$\bigodot_{l=0}^{1} \eta_l(\boldsymbol{z}) = \sigma(\boldsymbol{W}_1\sigma(\boldsymbol{W}_0\boldsymbol{z} + \boldsymbol{d}_0) + \boldsymbol{d}_1)$）．

NN の出力ベクトルから出力話者の音響特徴量へ逆射影するには，式 (15) による CRBM の後方推論によって以下のように計算される．

$$p(\boldsymbol{y}^{(t)}|\boldsymbol{h}_y^{(t)}, \boldsymbol{y}^{(t-1)}) = \sigma(\boldsymbol{W}_{yh}^{T}\boldsymbol{h}_y^{(t)} + \boldsymbol{W}_{y'y}\boldsymbol{y}^{(t-1)} + \boldsymbol{b}_y) \tag{20}$$

ここで，$\boldsymbol{b}_y$ は出力話者 CRBM の後方推論時のバイアスベクトルを表す．

以上の議論をまとめると，過去の特徴ベクトル $\boldsymbol{x}^{(t-1)}$，$\boldsymbol{y}^{(t-1)}$ を観測したときに，時刻 $t$ の入力話者音声の音響特徴量 $\boldsymbol{x}^{(t)}$ から出力話者音声の音響特徴量 $\boldsymbol{y}^{(t)}$ へ変換する，提案法による声質変換式は以下のように表すことができる．

$$\boldsymbol{y}^{(t)} = \bigodot_{k=0}^{L+2} \sigma(\boldsymbol{W}_{(k)}\boldsymbol{x}^{(t)} + \boldsymbol{a}_{(k)}(\boldsymbol{x}^{(t-1)}, \boldsymbol{y}^{(t-1)})) \tag{21}$$

ここで，$\boldsymbol{W}_{(k)}$ と $\boldsymbol{a}_{(k)}(\boldsymbol{x}^{(t-1)}, \boldsymbol{y}^{(t-1)})$ は動的パラメータ $\Theta^{(t)} = \{\boldsymbol{W}, \boldsymbol{a}^{(t)}\}$ の要素：

$$\boldsymbol{W} = \{\boldsymbol{W}_{(k)}\}_{k=0}^{L+2} \tag{22}$$

$$= \{\boldsymbol{W}_{xh}, \boldsymbol{W}_0, \cdots, \boldsymbol{W}_L, \boldsymbol{W}_{yh}{}^{T}\} \tag{23}$$

$$\boldsymbol{a}^{(t)} = \{\boldsymbol{a}_{(k)}(\boldsymbol{x}^{(t-1)}, \boldsymbol{y}^{(t-1)})\}_{k=0}^{L+2} \tag{24}$$

$$= \{\boldsymbol{W}_{x'h}\boldsymbol{x}^{(t-1)} + \boldsymbol{c}_x, \boldsymbol{a}_0, \cdots, \boldsymbol{a}_L, \boldsymbol{W}_{y'y}\boldsymbol{y}^{(t-1)} + \boldsymbol{b}_y\} \tag{25}$$

を表す．

式 (21) で表せられる変換式は，$L+4$ 個のシグモイド関数層を持つ NN の動的モデルを示唆している．そのため，2 つの異なる CRBM と中間の NN を合わせて 1 つのネットワークとみなすことができ，音響特徴量のパラレルデータを用いてそれぞれのパラメータを微調整することが可能である．

また，式 (21) が示すように，時刻 $t$ の出力話者ベクトルを得るためには，現在の入力話者ベクトル，一つ前のフレームの入力話者ベクトルと出力話者ベクトルが必要である．しかしながら，一般には過去の正しい出力話者ベクトルが分からないため，本稿では，最後に推定された過去の出力話者ベクトルを用いて繰り返し推定する（初期値はゼロベクトル）．この繰り返し推定法は，我々の予備実験において効果的であることが示されているが，具体的な実験内容に関しては論旨の都合上省略させていただく．

一方，混合数 $M$ の GMM に基づく声質変換法 [9] では，入力話者特徴ベクトル $\boldsymbol{x}$ は以下のように変換される．

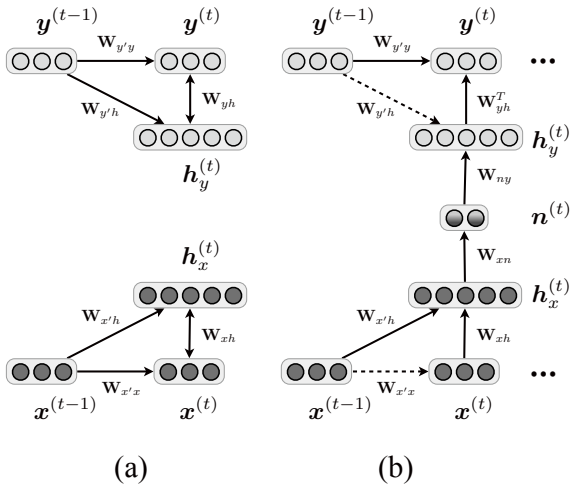


図 2　(a) CRBMs for a source speaker (below) and a target speaker (above), (b) our proposed voice conversion architecture combining two speaker-dependent CRBMs with a NN.

(注1)：簡単のため，評価実験では CRBM の学習と NN の学習は同一のパラレルデータを用いる（$T' = T$）．

$$\boldsymbol{y} = \sum_{m=1}^{M} P(m|\boldsymbol{x})(\boldsymbol{\Sigma}_{yx}^{(m)}\boldsymbol{\Sigma}_{xx}^{(m)-1}(\boldsymbol{x} - \boldsymbol{\mu}_x^{(m)}) + \boldsymbol{\mu}_y^{(m)}) \tag{26}$$

$$P(m|\boldsymbol{x}) = \frac{w^{(m)}\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_x^{(m)}, \boldsymbol{\Sigma}_{xx}^{(m)})}{\sum_{m=1}^{M} w^{(m)}\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_x^{(m)}, \boldsymbol{\Sigma}_{xx}^{(m)})} \tag{27}$$

この式が示すように，GMM の変換式は準線形関数の加算モデルとなっている．式 (21) に基づく我々の変換式は時系列データを入力する複数の異なる非線形関数の合成関数で定式化されているため，従来の GMM に基づく声質変換や，静的な変換を行う他のネットワークベースの手法 [11], [13] に比べて，より詳細な特徴変換が可能であると期待できる．

## 4. 評価実験

### 4.1 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 [17] を用いて，提案手法である話者依存型 CRBM を用いた手法（“Our”）と，従来の GMM を用いた手法（“GMM”），NN を用いた手法（“NN”），我々の先行研究である DBN を用いた手法 [13]（“DBN”）とで，声質変換精度の比較を行った．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．入力（出力）音響特徴量として，STRAIGHT スペクトル [18] から計算された 24 次元の MFCC を用いた．音声合成時には，文献 [19] 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．GMM に基づく変換法（64 混合）では，セグメント特徴を入力させる提案手法との公平な比較のため，静的な MFCC に加え，動的な MFCC（24 次元のデルタ MFCC とデルタデルタ MFCC）を加えた 72 次元のベクトルを用いた．入力・出力話者のパラレルデータは，データセットの 216 単語音声を用いて，動的計画法により作成された．このパラレルデータの同時特徴量は NN と GMM の学習の際に用いられ，2 つの話者依存型 CRBM の学習は独立して行われている．DBN（RBM），CRBM 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．NN のパラメータはランダムに初期化される．いずれの手法もパラメータの微調整を行う．変換音声の評価時には，同データセットからランダムに選んだ 20 文の発話音声（合計でおおよそ 70 秒分）を，各手法によって変換した（いずれも学習データには含まれていない音声）．客観評価基準として，声質変換の分野で一般的に用いられている，メル軸で出力話者音声と変換音声がどれくらい似ているのかを表す尺度である MCD (mel-cepstral distortion) を用いた．フレームごとにこの MCD を求め，全フレームの平均 MCD を算出することで，各手法による変換精度を比較した．

### 4.2 ネットワーク構造の違い

声質変換の本実験に入る前に，まず，ネットワークベースの手法（“Our”，“NN”，“DBN”）について，ネットワークの構造の違いによる変換精度の差をみる予備実験を行った．この予備実験では，表 1 に示されているように，隠れ層の数や素子数を変えた “arc. 1” から “arc. 2” までの 4 種類の異なるネットワーク構造を用いて声質変換の精度を比較した．表 1 の角括弧の中の数値は，入力層から出力層までの各層の素子の数を表している．例えば提案法では，[入力話者 CRBM の第 1 層:第 2 層-中間 NN-出力話者 CRBM の第 1 層:第 2 層] のように記述される．

図 3 から，各手法によって適切なネットワーク構造が異なることが分かる．また，この図から読み取れることとして，提案手法や DBN を用いた手法では，必ずしも隠れ層の多い深い構造を持つネットワーク（例えば “arc. 3”）が良い結果をもたらすとは限らない．結果的に提案手法と DBN による手法では “arc. 2” を用いた場合，NN による手法では “arc. 3” を用いた場合から最も高い変換精度を得たので，以後の本実験においても，それぞれの手法に最適なネットワーク構造を使用する．

### 4.3 実験結果と考察

学習データ数を変えながら，GMM を含めた各手法による声質変換の精度を図 4 にまとめた．また，図 5 は $N = 10k$ のときの変換音声を用いた主観評価実験の結果を表している．MOS 値 (mean opinion score) に基づく主観評価実験では，9 名の被験者に，元の出力話者音声（分析合成された音声）と変換音声のペアを聞いてもらい，話者性と自然性の 2 つの観点から，変換音声がどれだけ出力話者音声に近いかを 5 段階で評価してもらった（5: excellent, 4: good, 3: fair, 2: poor, and 1: bad）．

これらの図から，提案手法である “Our” が主観的にも客観的にも，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（0.001 水準で有意差が認められた）．また，“NN” と “DBN” の違いは，一部の重みパラメータの初期値として話者依存型 RBM（DBN）を用いるかどうかであり（パラメータの微調整を行うため），パラメータの初期値によって大きく結果が異なっていることが分かる．このことから，繰り返し更新によってある程度パラメータが収束しても，大域最適解とは程遠い局所最適解に陥っており，初期値の選び方は重要であることが分かる．言い換えれば，初期値が大域最適解に近ければ近いほど，ネットワークの構造は同じでも精度は向上すると言える．同様の理由で，“Our” の結果が “DBN” に比べて改善したのは，話者依存型 DBN の代わりに話者依存型 CRBM を用いることで，より時間変化に依存しない話者性を強調した高次元空間が形成されたためであると考えられる（得られた空間への写像行列が結果的に大域最適解に近いものであると考えられるため）．

## 5. おわりに

本研究では，話者性が強調される高次元空間を形成することを目的として話者依存型 CRBM を用いた声質変換手法を提案した．また，提案手法では入力話者の CRBM の高次空間射影，高次空間における特徴変換，出力話者の CRBM の音響特徴量空間への逆射影を得て出力話者音声へ変換するが，これらのプロセスが体系的に 1 つのネットワークとして表現できることを示した．評価実験では提案法により，従来の GMM や，他のネットワークベースの手法よりも客観的かつ主観的に優れた変換音声が得られた．

表 1 Various architectures used for the preliminary experiment.

| Arcitectures | NN | DBN [13] | Our method | Layers |
|---|---|---|---|---|
| arc. 1 | [24-24-24-24] | [24:24-24:24] | [(24,24):24-(24,24):24] | 4 |
| arc. 2 | [24-48-48-24] | [24:48-48:24] | [(24,24):48-(48,24):24] | 4 |
| arc. 3 | [24-24-24-24-24-24] | [24:24:24-24:24:24] | [(24,24):24:24-24:(24,24):24] | 6 |
| arc. 4 | [24-48-24-24-48-24] | [24:48:24-24:48:24] | [(24,24):48:24-24:(48,24):24] | 6 |

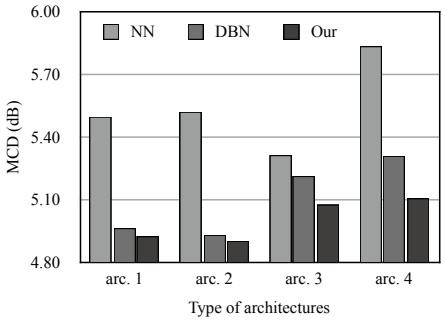


図 3 Averaged mel-cepstral distortion with changing network architetures ($N = 10,000$).

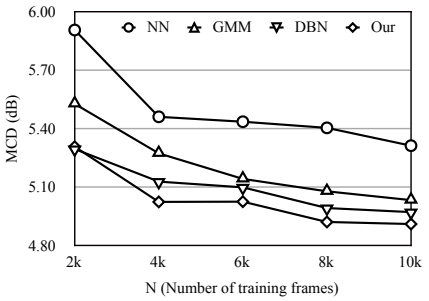


図 4 Averaged mel-cepstral distortion for each method.

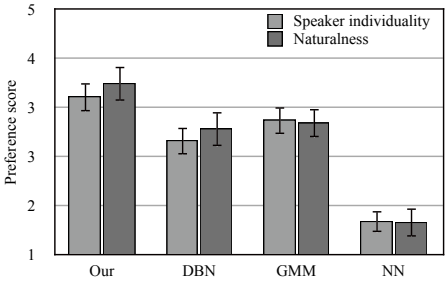


図 5 MOS scores w.r.t. speaker individuality and naturalness. The error bars show 95% confidence intervals.

## 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] 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).

[12] 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).

[13] 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).

[14] 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).

[15] 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).

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

[17] 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).

[18] 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", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)IEEE, pp. 3933–3936 (2008).

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