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
SP2015-71(2015-12)

制約付き Three-Way Restricted Boltzmann Machine を用いた音響・音韻・話者情報の同時モデリング

中鹿　亘†　滝口　哲也††

† 電気通信大学情報システム学研究科
〒 182-8585 東京都調布市調布ケ丘 1-5-11
†† 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@uec.ac.jp, ††takigu@kobe-u.ac.jp

あらまし　本研究では，音響特徴量・音韻特徴量・話者特徴量の 3 つを変数とする Three-Way Restricted Boltzmann Machine (3WRBM) を用いて音声モデリングを試みる．3WRBM はそれぞれの変数のユーナリーポテンシャル，2 変数間のペアワイズポテンシャル，そして 3 変数間の Three-way ポテンシャルを総和したエネルギーに基づく確率密度関数である．本研究では，音響・音韻・話者特徴量の Three-way ポテンシャルを話者正規化学習・話者適応の観点から適切に設計する．一度モデルの学習が終われば 3 変数間の関係性が捉えられ，各特徴量の相互条件付確率を簡単に計算することができる．3WRBM による音声モデリングの性能を評価するために，本稿では声質変換実験と話者認識実験の結果を報告する．話者認識実験における話者特徴量は与えられた音響特徴量から尤度最大下基準により推定することで求めることができ，声質変換は，推定された音韻情報と，切り替えた話者情報から音響特徴量を推定することで実現される．

キーワード　音声モデリング，Restricted Boltzmann machine，話者正規化学習，話者認識，声質変換

Simultaneous Modelling of Acoustic, Phonetic, Speaker Features Using Improved Three-Way Restricted Boltzmann Machine

Toru NAKASHIKA† and Tetsuya TAKIGUCHI††

† Graduate School of Information Systems, University of Electro-Communications
1-5-1 Chofugaoka, Chofu, Tokyo, 182-8585 Japan
†† Organization of Advanced Science and Technology, Kobe University
1-1 Rokkodaicho, Nada-ku, Kobe, Hyogo, 657-8501 Japan
E-mail: †nakashika@uec.ac.jp, ††takigu@kobe-u.ac.jp

Abstract　In this paper, we argue the way of modelling speech signals using improved three-way restricted Boltzmann machine (3WRBM) where acoustic features, latent phonological features, and speaker-identity features are considered. The 3WRBM is an energy-based probabilistic model that includes three kinds of potentials: unary potentials of each variable, pairwise potentials of every two variables, and three-way potentials of the three variables. In our approach, we design the three-way potentials properly in the speaker-adaptive training (SAT) manner. The optimized model captures the relationships between the variables, enables to compute conditional probabilities of each variables, and is appliable to many tasks in speech signal processing. For example, estimating speaker-identity features given acoustic features is used for speaker recognition. Another example is estimating acoustic features from the phonological features that are estimated given source speaker's acoustic features and the desired speaker-identity features; that is voice conversion. In our experiments, we evaluate the effectiveness of the speech modelling through a voice conversion task and a speaker recognition task.

Key words　Speech modelling, restricted Boltzmann machine, speaker-adaptive training, speaker recognition, voice conversion

1. はじめに

現在最も代表的であり効果的な音声モデリング手法として HMM (hidden Markov model) が挙げられる．HMM は状態遷移確率と観測特徴量の出力確率で構成される．この出力確率には，通常，連続確率密度関数として GMM (Gaussian mixture model) が用いられる．つまり，一般的な HMM を用いた音声モデリングでは，あるフレーム（状態）における音響特徴量は GMM を用いてモデル化されている．しかし GMM は多数の観測データを表層的にクラスタリングして表現するモデリング手法であり，潜在的に存在する特徴の内部構造まで捉えることができない．一方近年盛んに研究されているディープラーニング [1] に基づく音声モデリングでは，高次の潜在特徴間の関係性を考慮しており，実際音声認識タスクでは GMM と比較して高い精度を上げている [2]．しかしながら，勾配法や EM アルゴリズムに基づいた学習では局所解に陥る場合が多く，特にフリーパラメータを多く含み，自由度の高いディープラーニングに基づくモデリングでは，事前学習をしているとはいえ，必ずしも潜在特徴間の関係性を適切に学習できるとは限らない．局所解を防ぎ，より真の解に近い解を得るためには，適切な制約を設け，自由度を抑えることが重要だと考えられる．

本稿ではこうした背景を踏まえて，音響特徴量，話者特徴量，潜在的な音韻特徴量の 3 つのファクターの関係性を考慮した three-way restricted Boltzmann machine (3WRBM) を用いた音声モデリング手法を提案する．このモデルは 3 次までのポテンシャルを考慮したエネルギー関数に基づく確率モデルであり，2 層の RBM [3] と同様に，同一ファクターユニット間には結合は存在せず，異なるファクターユニット間のみ双方向の結合重みが存在すると仮定している．この結合重みが各特徴間の関係性を表している．本研究では，音響特徴量は，話者に依存しない音韻特徴量と強い繋がりのある標準音響特徴量に，話者特徴量と繋がりのある話者適応行列を乗じることで得られるという仮定をおいて結合重みパラメータに制約を加えている．また，本稿ではフレームレベルの音声モデリングを対象としており，HMM のような時系列モデリングは取り扱わない．

本研究で提案するモデルは音韻情報と話者情報を考慮した生成モデルであるため，様々な音声信号処理タスクへ応用することができる．例えば学習済みのモデルを用いて，入力音響特徴量から話者特徴量を推定することで，フレーム毎の話者認識を行うことができる．また，入力音響特徴量から推定された話者特徴量のみを切り替えて音響特徴量を生成することで，入力音声を任意の話者の音声へ変換することができる（声質変換）．さらに，本稿のスコープではないが，音韻特徴量は話者に依存しない情報と仮定しているため，推定された音韻特徴量を用いて音声認識器にかければ，話者普遍性から音声認識精度が向上すると考えられる．

特に本モデリング手法は声質変換タスクにおいて効果を発揮する．声質変換とは入力音声の音韻情報を残したまま話者性のみを対象者のものへ変換させる技術であるが，様々なタスクへ応用可能である [4]〜[8] ことから，近年盛んに研究されている．話者性の中にはスペクトル特徴だけではなく F0 やデュレーション，発話スタイルなども含まれるが，多くの声質変換に関する研究ではスペクトルの変換のみ言及されており，本研究においてもスペクトルの変換を対象とする．これまでの声質変換手法として，GMM に基づく手法 [9]〜[13]，NMF (non-negative matrix factorization) に基づく手法 [14], [15]，ディープラーニングに基づく手法 [16]〜[21] など，様々な統計的アプローチが試みられてきた．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，新たな話者対に対して既存の変換モデルが利用できないなど様々な弊害や問題が生じる．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 [22] や話者適応に基づく手法 [23]，MAP に基づく手法 [24] がある．これらは，予め多数の参照話者音声を用いて，参照話者間のマッピング関数を学習し，入力話者と出力話者に適応させることで入出力話者間のパラレルデータを必要としない声質変換を実現している．しかし依然として参照話者間のモデルの学習にはパラレルデータを用意する必要がある．そこで我々の先行研究では，入力・出力話者間だけでなく参照話者間のモデルの学習時においてもパラレルデータを必要としない声質変換手法を提案してきた [25], [26]．一方，本モデリングでは音声信号から自動的に音韻情報と話者情報を抽出するため，入力・出力話者間だけでなく参照話者間のモデルの学習時においてもパラレルデータを必要としない．また，フレーム単位で入力話者・出力話者の音響特徴量間のマッピングを行う従来型の声質変換法と異なり，音声信号から話者情報のみを切り替えて音声を生成するという提案法は極めて自然なアプローチである．

また，本稿で提案するモデルは，音声信号から自動的に音韻情報と話者情報を抽出するという点で我々がこれまでに提案してきた適応型 RBM [25] や SATBM (speaker-adaptive-trainable Boltzmann machine) [26] と類似している．これらと提案モデルとの最大の違いは，提案モデルでは話者識別素子を変数とみなし，サンプリング可能にしている点である．これにより話者認識へ応用可能となる．本稿では声質変換タスクおよび話者認識タスクを通じて，提案モデルの有効性を検証する．

以下，2. 章では基礎モデルの RBM と，その拡張モデルである 3WRBM について述べる．3. 章では提案する音声モデル（音韻・話者因子の分解を考慮して 3WRBM に制約を加えたモデル）とパラメータ推定法について述べる．4. 章で声質変換および話者認識の評価実験について述べ，5. 章で本論文をまとめる．

2. Energy-based models

本稿で提案するモデルは Energy-based models (EBMs) の一種として定義される．EBM は変数 $\boldsymbol{x}$ に関する個々の要素エネルギーの総和 $E(\boldsymbol{x})$ を考慮した確率モデルであり，一般的に

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

$$Z = \int e^{-E(\boldsymbol{x})} d\boldsymbol{x} \tag{2}$$

と書ける．式 (1) にもあるように，$\boldsymbol{x}$ の尤度を最大化させることと，総エネルギー $E(\boldsymbol{x})$ を最小化させることは等しい．代表的な EBM として MRF (Markov random field) や CRF (conditional random field)，RBM (restricted Boltzmann machine) などが挙げられる．以下 RBM（実数値の観測特徴量を表現できるように拡張した Gaussian-Bernoulli RBM [3]）と，3 変数へ拡張した Three-way RBM (3WRBM) について順に述べる．

2.1 RBM

Restricted Boltzmann machine (RBM) は特殊な構造を持つ 2 層ネットワークであり，D 個の実数値の可視変数 $\boldsymbol{v} = [v_i]_i \in \mathbb{R}^D$ と H 個のバイナリ値の隠れ変数 $\boldsymbol{h} = [h_j]_j \in \{0,1\}^H$ の確率分布を表現する無向グラフィカルモデルである [27]．RBM では可視変数 $\boldsymbol{v}$ と隠れ変数 $\boldsymbol{h}$ からなる総エネルギー $E(\boldsymbol{v}, \boldsymbol{h})$ は以下のように定義される．

$$E(\boldsymbol{v}, \boldsymbol{h}) = \frac{1}{2} \left\| \frac{\boldsymbol{v} - \boldsymbol{b}}{\boldsymbol{\sigma}} \right\|^2 - \boldsymbol{c}^\top \boldsymbol{h} - \left(\frac{\boldsymbol{v}}{\boldsymbol{\sigma}^2} \right)^\top \mathbf{W} \boldsymbol{h} \tag{3}$$

ここで，$\|\cdot\|^2$ は L2 ノルム，括線は要素除算を表す．$\mathbf{W} \in \mathbb{R}^{D \times H}$，$\boldsymbol{\sigma} \in \mathbb{R}^D$，$\boldsymbol{b} \in \mathbb{R}^D$，$\boldsymbol{c} \in \mathbb{R}^H$ はそれぞれ可視変数と隠れ変数間の重み行列，可視変数の偏差，可視変数のバイアス，隠れ変数のバイアスを表すパラメータである．式 (3) において第 1 項，第 2 項はそれぞれ変数 $\boldsymbol{v}$，$\boldsymbol{h}$ の個々のエネルギー（ユーナリーポテンシャル）を表しており，第 3 項は $\boldsymbol{v}$ と $\boldsymbol{h}$ 間の結合エネルギー（ペアワイズポテンシャル）を表している．ユーナリーポテンシャル項を $U(\boldsymbol{v}, \boldsymbol{h})$，ペアワイズポテンシャル項を $P(\boldsymbol{v}, \boldsymbol{h})$ とすると，式 (3) は

$$E(\boldsymbol{v}, \boldsymbol{h}) = U(\boldsymbol{v}, \boldsymbol{h}) + P(\boldsymbol{v}, \boldsymbol{h}) \tag{4}$$

$$U(\boldsymbol{v}, \boldsymbol{h}) = \frac{1}{2} \left\| \frac{\boldsymbol{v} - \boldsymbol{b}}{\boldsymbol{\sigma}} \right\|^2 - \boldsymbol{c}^\top \boldsymbol{h} \tag{5}$$

$$P(\boldsymbol{v}, \boldsymbol{h}) = -\boldsymbol{v}'^\top \mathbf{W} \boldsymbol{h} \tag{6}$$

と書き直すことができる．ただし $\boldsymbol{v}' = \frac{\boldsymbol{v}}{\boldsymbol{\sigma}^2}$ と置いた．なお RBM では式 (1)(2) において $\boldsymbol{x} = [\boldsymbol{v}^\top \ \boldsymbol{h}^\top]^\top$ としている．

2.2 3WRBM

前節の RBM は 2 変数間の結合エネルギーを考慮したモデルであったが，これを 3 変数へ拡張し，3 変数間の結合エネルギー（Three-way ポテンシャル）を考慮したモデル (Three-way RBM; 3WRBM）を定義することができる．一般的にはさらに高次へ拡張することもできる [29]．本研究では音響特徴量を表す $\boldsymbol{v}$，潜在特徴量 $\boldsymbol{h}$，話者特徴量 $\boldsymbol{s} = [s_k]_k \in \{0,1\}^R, \sum_k s_k = 1$ の 3 変数の関係性を 3WRBM を用いて表現する．本研究では様々な話者によるクリーンな音声を対象とし，話者による変動成分は話者特徴量 $\boldsymbol{s}$ によって捉えるため，音声信号から観測はできないがその背後に存在する特徴量として音韻情報が考えられる．そこで $\boldsymbol{h}$ を本稿では音韻特徴量と呼ぶことにする．$\boldsymbol{h}$ と $\boldsymbol{s}$ はバイナリベクトルであり，諸要素がオン（アクティブ）になっている状態を 1 で表す．例えば音声信号に対して音韻要素 j が作用していること表す場合，$h_j = 1$ となり，話者 k の発話であることを表す場合，$s_k = 1, \forall s_{k'} = 0 \ (k' \neq k)$ となる．音韻は部分基底の組み合わせで表現されることを考慮して $\boldsymbol{h}$ には制約を加えない．このとき，エネルギー関数は，RBM のものから

$$E(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{s}) = U(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{s}) + P(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{s}) + T(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{s}) \tag{7}$$

$$U(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{s}) = \frac{1}{2} \left\| \frac{\boldsymbol{v} - \boldsymbol{b}}{\boldsymbol{\sigma}} \right\|^2 - \boldsymbol{c}^\top \boldsymbol{h} - \boldsymbol{d}^\top \boldsymbol{s} \tag{8}$$

$$P(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{s}) = -\boldsymbol{v}'^\top \mathbf{W} \boldsymbol{h} - \boldsymbol{h}^\top \mathbf{V} \boldsymbol{s} - \boldsymbol{s}^\top \mathbf{U} \boldsymbol{v}' \tag{9}$$

$$T(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{s}) = -\sum_{i,j,k} v'_i h_j s_k Z_{ijk} \tag{10}$$

と自然に拡張することができる．ただし $\mathcal{Z} \in \mathbb{R}^{D \times H \times K}$ の要素 Z_{ijk} は Three-way ポテンシャル $T(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{s})$ における 3 変数要素 v_i, h_j, s_k 間の結合重み，$\boldsymbol{d} \in \mathbb{R}^R$ は $\boldsymbol{s}$ に関するバイアス，$\mathbf{V} \in \mathbb{R}^{H \times R}$ と $\mathbf{U} \in \mathbb{R}^{R \times D}$ はそれぞれ $\boldsymbol{h}, \boldsymbol{s}$ 間と $\boldsymbol{s}, \boldsymbol{v}$ 間のペアワイズ結合重みを表す．$\boldsymbol{x} = [\boldsymbol{v}^\top \ \boldsymbol{h}^\top \ \boldsymbol{s}^\top]^\top$ とおけば $\boldsymbol{v}$，$\boldsymbol{h}$，$\boldsymbol{s}$ の同時確率密度関数は式 (1)(2) で表すことができる．RBM と同様に，各変数間にはその関係性の度合いを示す双方向の結合重みが存在し，それぞれの変数の要素同士（例えば s_k と $s_{k'}$）には結合が存在しないと仮定している．この性質のおかげで，$\boldsymbol{v}$，$\boldsymbol{h}$，$\boldsymbol{s}$ の条件付き確率をそれぞれ以下のように簡単に計算することができる．

$$p(\boldsymbol{v}|\boldsymbol{h}, \boldsymbol{s}) = \mathcal{N}(\boldsymbol{v} \mid \boldsymbol{b} + \mathbf{W}\boldsymbol{h} + \mathbf{U}^\top \boldsymbol{s} + \sum_{j,k} h_j s_k \mathcal{Z}_{:jk}, \boldsymbol{\sigma}^{\boldsymbol{2}})$$

$$p(\boldsymbol{h}|\boldsymbol{s}, \boldsymbol{v}) = \mathcal{B}(\boldsymbol{h} \mid \boldsymbol{f}(\boldsymbol{c} + \mathbf{V}\boldsymbol{s} + \mathbf{W}^\top \boldsymbol{v}' + \sum_{i,k} v'_i s_k \mathcal{Z}_{i:k}))$$

$$p(\boldsymbol{s}|\boldsymbol{v}, \boldsymbol{h}) = \mathcal{B}(\boldsymbol{s} \mid \boldsymbol{m}(\boldsymbol{d} + \mathbf{U}\boldsymbol{v}' + \mathbf{V}^\top \boldsymbol{h} + \sum_{i,j} v'_i h_j \mathcal{Z}_{ij:}))$$

ただし $\mathcal{N}(\cdot)$ は次元独立の多変量正規分布，$\mathcal{B}(\cdot)$ は多次元ベルヌーイ分布，$\boldsymbol{f}(\cdot)$ は要素ごとのシグモイド関数，$\boldsymbol{m}(\cdot)$ は要素ごとの softmax 関数を表す．また $\mathcal{Z}_{:jk}$，$\mathcal{Z}_{i:k}$，$\mathcal{Z}_{ij:}$ はそれぞれ $\mathcal{Z}$ の第 1 モード，第 2 モード，第 3 モードの部分ベクトルを表す．3WRBM は文献 [28] で定義される factored 3WRBM と類似しているが，factored 3WRBM は 1 種類の可視変数と隠れ変数間の関係性を 3 次でモデル化しているのに対して，本稿で述べる 3WRBM は性質の異なる 2 種類の可視変数と隠れ変数の関係性をモデル化している (可視変数内の結合は存在しないと仮定している点で異なる)．

3. 音韻・話者因子に関する制約

前節で述べた Three-way RBM (3WRBM) はパラメータの数が膨大となり，モデルの自由度が必要以上に高く，うまく学習されない可能性がある．そこで何らかの制約を加え，パラメータ数を抑えることが望ましい．また，適切な構造化・制約は局所解を防ぎ，より質の高い解を得ることができると考えられる．本研究では,「音声らしさ」に着目した構造化や制約を加える．

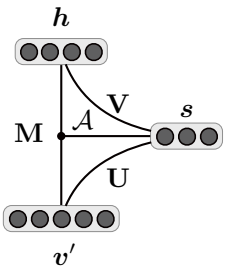

図 1　提案手法における音声モデリングのグラフ構造.
Fig. 1　Graphical representation of the proposed model.

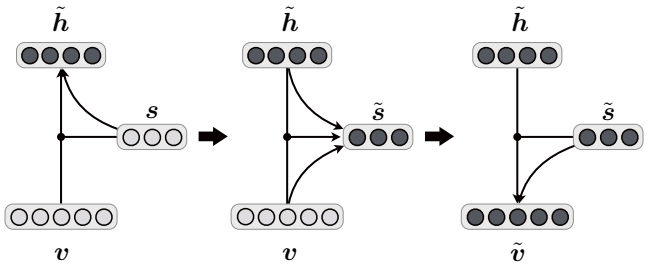

図 2　提案モデルにおける 3-step サンプリング.
Fig. 2　3-step sampling used in the training.

まず，Three-way ポテンシャルについて考察する．Three-way ポテンシャルの一部，$\mathcal{Z}_{:jk}$ に関するエネルギーは音韻要素 j と話者 k が作用しているとき，$T(\boldsymbol{v}, h_j = 1, s_k = 1) = -\boldsymbol{v}'^{\top}\mathcal{Z}_{:jk}$ と計算され，このエネルギーは正規化された音響特徴量（観測ベクトル）$\boldsymbol{v}'$ が $\mathcal{Z}_{:jk}$ と類似するとき小さな値をとる．言い換えれば，安定状態（エネルギーの小さい状態）では $\boldsymbol{v}'$ は $\mathcal{Z}_{:jk}$ と類似しているため，$\mathcal{Z}_{:jk}$ は音韻要素 j，話者 k に依存した，観測データの中に出現する音響特徴量パターンを表していると考えられる．ここで $\mathcal{Z}_{:jk}$ を，音韻と話者の因子に分解することを考え，

$$\mathcal{Z}_{:jk} = \mathbf{A}_k \boldsymbol{m}_j \tag{11}$$

とおく．ただし $\boldsymbol{m}_j \in \mathbb{R}^D$ は音韻 j に依存した作用素，$\mathbf{A}_k \in \mathbb{R}^{D\times D}$ は話者 k に依存した作用素を表す．式 (11) は，$\mathcal{Z}_{:jk}$ は音韻 j の特徴ベクトル $\boldsymbol{m}_j$ を話者 k の行列 $\mathbf{A}_k$ で射影した音響特徴量パターンであることを示している．一般に音響特徴量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ため，式 (11) によるモデル化は妥当であると考えられる．したがって $\boldsymbol{m}_j$ は音韻 j の話者に依存しない音響特徴量パターン（標準音響特徴量），$\mathbf{A}_k$ は標準音響特徴量を話者 k の空間へ射影する適応行列を表すと考えられる．この $\boldsymbol{m}_j$ によって音韻 j と音響特徴量の関係性をモデル化できるため，$\mathbf{W}_{:j} = \mathbf{0}$ とする．

また，話者 k のバイアス d_k は，データ全体の中で話者 k が出現する頻度のようなものを表している．しかしそれぞれの話者を対等に取り扱うという目的で，本研究では $\boldsymbol{d} = \mathbf{0}$ とする．

以上をまとめて，本稿では，音声モデリングのためのエネルギー関数を以下で定義する．

$$\begin{aligned} &E(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{s}) \\ &= \frac{1}{2}\left\|\frac{\boldsymbol{v}-\boldsymbol{b}}{\boldsymbol{\sigma}}\right\|^2 - \boldsymbol{c}^{\top}\boldsymbol{h} - \boldsymbol{h}^{\top}\mathbf{V}\boldsymbol{s} - \boldsymbol{s}^{\top}\mathbf{U}\boldsymbol{v}' - \boldsymbol{v}'^{\top}\mathbf{A}_{\boldsymbol{s}}\mathbf{M}\boldsymbol{h} \end{aligned} \tag{12}$$

ただし，$\mathbf{A}_{\boldsymbol{s}} = \sum_k \mathbf{A}_k s_k$，$\mathbf{M} = [\boldsymbol{m}_1 \ \cdots \ \boldsymbol{m}_H]$ とおいた．また便宜上 $\mathcal{A} = \{\mathbf{A}_k\}_k$ とする．このとき，条件付き確率は

$$p(\boldsymbol{v}|\boldsymbol{h}, \boldsymbol{s}) = \mathcal{N}(\boldsymbol{v} \mid \boldsymbol{b} + \mathbf{U}^{\top}\boldsymbol{s} + \mathbf{A}_{\boldsymbol{s}}\mathbf{M}\boldsymbol{h}, \boldsymbol{\sigma}^{\boldsymbol{2}}) \tag{13}$$

$$p(\boldsymbol{h}|\boldsymbol{s}, \boldsymbol{v}) = \mathcal{B}(\boldsymbol{h} \mid \boldsymbol{f}(\boldsymbol{c} + \mathbf{V}\boldsymbol{s} + \mathbf{M}^{\top}\mathbf{A}_{\boldsymbol{s}}^{\top}\boldsymbol{v}) \tag{14}$$

$$p(s_k|\boldsymbol{v}, \boldsymbol{h}) = \mathcal{B}(s_k \mid m(\mathbf{U}_{k:}\boldsymbol{v}' + \mathbf{V}_{:k}^{\top}\boldsymbol{h} + \boldsymbol{v}'\mathbf{A}_k\mathbf{M}\boldsymbol{h})) \tag{15}$$

となる．式 (12) が示す 3 変数 $\boldsymbol{v}$（正確には $\boldsymbol{v}'$），$\boldsymbol{h}$，$\boldsymbol{s}$ の関係性をグラフで表現すると，Fig. 1 のようになる．

3.1　パラメータ推定

提案モデルのパラメータ $\Theta = \{\mathbf{M}, \mathcal{A}, \mathbf{U}, \mathbf{V}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{\sigma}\}$ は，R 人の話者による N フレームの音声データ $\{\boldsymbol{v}_n, \boldsymbol{s}_n\}_{n=1}^N$ に対する対数尤度

$$\mathcal{L} = \log \prod_n p(\boldsymbol{v}_n, \boldsymbol{s}_n) = \sum_n \log \sum_{\boldsymbol{h}} p(\boldsymbol{v}_n, \boldsymbol{h}_n, \boldsymbol{s}_n) \tag{16}$$

を最大化するように同時に推定することが可能である．それぞれのパラメータで対数尤度 $\mathcal{L}$ を偏微分すると，

$$\frac{\partial \mathcal{L}}{\partial \mathbf{M}} = \langle \sum_k \mathbf{A}_k^{\top}\boldsymbol{v}'\boldsymbol{h}^{\top}s_k \rangle_{\text{data}} - \langle \sum_k \mathbf{A}_k^{\top}\boldsymbol{v}'\boldsymbol{h}^{\top}s_k \rangle_{\text{model}} \tag{17}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{A}_k} = \langle \boldsymbol{v}'\boldsymbol{h}^{\top}s_k\mathbf{M}^{\top} \rangle_{\text{data}} - \langle \boldsymbol{v}'\boldsymbol{h}^{\top}s_k\mathbf{M}^{\top} \rangle_{\text{model}} \tag{18}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{U}} = \langle \boldsymbol{s}\boldsymbol{v}'^{\top} \rangle_{\text{data}} - \langle \boldsymbol{s}\boldsymbol{v}'^{\top} \rangle_{\text{model}} \tag{19}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{V}} = \langle \boldsymbol{h}\boldsymbol{s}^{\top} \rangle_{\text{data}} - \langle \boldsymbol{h}\boldsymbol{s}^{\top} \rangle_{\text{model}} \tag{20}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{b}} = \langle \boldsymbol{v}' \rangle_{\text{data}} - \langle \boldsymbol{v}' \rangle_{\text{model}} \tag{21}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{c}} = \langle \boldsymbol{h} \rangle_{\text{data}} - \langle \boldsymbol{h} \rangle_{\text{model}} \tag{22}$$

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \boldsymbol{\sigma}} = &\frac{1}{\boldsymbol{\sigma}^3} \circ \Big(\langle \boldsymbol{v} \circ \boldsymbol{v} - 2\boldsymbol{v} \circ \big(\boldsymbol{b} + \mathbf{U}^{\top}\boldsymbol{s} + \mathbf{A}_{\boldsymbol{s}}\mathbf{M}\boldsymbol{h}\big) \rangle_{\text{data}} \\ &- \langle \boldsymbol{v} \circ \boldsymbol{v} - 2\boldsymbol{v} \circ \big(\boldsymbol{b} + \mathbf{U}^{\top}\boldsymbol{s} + \mathbf{A}_{\boldsymbol{s}}\mathbf{M}\boldsymbol{h}\big) \rangle_{\text{model}} \Big) \end{aligned} \tag{23}$$

が得られる．ただし，各偏微分項右辺の $\langle\cdot\rangle_{\text{data}}$，$\langle\cdot\rangle_{\text{model}}$ はそれぞれデータに対する期待値，モデルの期待値を表す．モデルに対する期待値は項数が膨大となり計算困難だが，CD（contrastive divergence）法 [1] 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．CD 法は $\langle\cdot\rangle_{\text{model}}$ を Gibbs サンプリングによる再構築データの期待値 $\langle\cdot\rangle_{\text{recon.}}$ で近似する．本稿では Fig. 2 に示すように $\boldsymbol{h}$，$\boldsymbol{v}$，$\boldsymbol{s}$ を順にサンプリングする．例えば $\boldsymbol{h}$ のサンプリングでは，式 (14) を用いて，$\tilde{\boldsymbol{h}} \sim p(\boldsymbol{h}|\boldsymbol{s}, \boldsymbol{v})$ とすることでサンプル $\tilde{\boldsymbol{h}}$ を得る．このようにすることで，既知の特徴量 $\boldsymbol{v}$，$\boldsymbol{s}$ から $\tilde{\boldsymbol{h}} \sim p(\boldsymbol{h}|\boldsymbol{s}, \boldsymbol{v})$，$\tilde{\boldsymbol{s}} \sim p(\boldsymbol{s}|\boldsymbol{v}, \tilde{\boldsymbol{h}})$，$\tilde{\boldsymbol{v}} \sim p(\boldsymbol{v}|\tilde{\boldsymbol{h}}, \tilde{\boldsymbol{s}})$，$\tilde{\tilde{\boldsymbol{h}}} \sim p(\boldsymbol{h}|\tilde{\boldsymbol{s}}, \tilde{\boldsymbol{v}}) \cdots$ と Gibbs チェインを繋げていくことができる．

4.　音声モデリングの検証実験

提案モデルの音声モデリングの効果を確かめるため，話者認識と声質変換実験を行った．両実験とも日本音響学会研究用連

表 1　話者認識実験の結果.

Table 1　Speaker recognition accuracy of each method.

Method	GMM	GMM(UBM)	Our	Our($\boldsymbol{h}$ known)
Acc.[%]	85.9	83.2	78.1	**90.6**

表 2　SVM による話者認識における特徴量の違い.

Table 2　Speaker recogniton accuracies using various features in SVM.

Features	$\boldsymbol{v}$ (mcep)	$\boldsymbol{h}$	$\boldsymbol{s}$
# dims.	32	20	8
Acc.[%]	82.0	42.2	**78.7**

続音声データベース（ASJ-JIPDEC）の中からランダムに男性 4 名女性 4 名 ($R = 8$) を選び，40 発話分の音声データを学習に，別の 10 発話分の音声データを評価に用いた．分析合成ツールの WORLD [30] によって得られたスペクトルから計算した 32 次元のメルケプストラムを入力特徴量に用いた ($D = 32$)．また，潜在特徴量の数を $H = 20$ とした．学習率 0.01，モーメント係数 0.9，バッチサイズ 100，繰り返し回数 50 の確率的勾配法を用いてモデルを学習した．

4.1　話 者 認 識

モデルを学習した後，評価データのフレーム音響特徴量から話者 s_k がアクティブとなる確率

$$p(s_k = 1|\boldsymbol{v}) = m(-C + \mathbf{U}_{k:}\boldsymbol{v}' + \sum_j g(\frac{c_j}{R} + \mathbf{V}_{jk} + \boldsymbol{v}'^{\top}\mathbf{A}_k\boldsymbol{m}_j))$$

を計算し，$\hat{k} = \mathrm{argmax}_k\, p(s_k = 1|\boldsymbol{v})$ として話者を推定した．ただし $g(\cdot)$ は softplus 関数を表し，$C = \sum_j g(\frac{c_j}{R})$ とした．評価データの全フレーム数を N_{all}，正解したフレーム数を $N_{\mathrm{corr.}}$ とすると，$100 \cdot N_{\mathrm{corr.}}/N_{\mathrm{all}}$ として話者認識率を算出した．比較手法として GMM による話者認識を用いた．比較手法では話者ごとに 64 混合の GMM を学習させ，評価データのフレーム特徴量を入力し，最も尤度の高い GMM を選ぶことで話者を推定した．また，予め全話者の音声を用いて GMM を学習 (UBM; universal background model) しておき，その後話者ごとの GMM を再学習させる手法とも比較した．

実験結果を Table 1 に示す．本実験では残念ながら提案手法（“Our”）では GMM よりも高い精度が得られなかった．これは，提案モデルが Fig. 2 に示すサンプリング法に基づいて学習されているため，音響特徴量と話者特徴量の 2 変数を与えないと音韻特徴量をうまく推定することができないからだと考えられる．そこで，話者特徴量を与えて式 (14) より音韻情報を推定し，これを既知として式 (15) より話者特徴量を推定したところ，認識精度が向上することが確認できた（“Our($\boldsymbol{h}$ known)）．

また，音響特徴量から計算される話者特徴量や音韻特徴量の質を調べるために，線形カーネル SVM (support vector machine) を用いて話者認識実験を行った（1 vs. 1 法による認識）．この実験では SVM の入力特徴量として音響特徴量をそのまま用いた場合（つまりメルケプストラム特徴量）と，推定された音韻特徴量 $\boldsymbol{h}$ を用いた場合，推定された話者特徴量 $\boldsymbol{s}$ を用いた場合で精度を比較した．実験結果を Table 2 に示す．Table 2 より，

表 3　声質変換実験の結果.

Table 3　Speaker recognition accuracy of each method.

Method	GMM	ARBM	SATBM	3WRBM	Our
Non-parallel?	No	**Yes**	**Yes**	**Yes**	**Yes**
MIDR[dB]	4.06	2.11	2.66	-0.147	**3.35**

$\boldsymbol{v}$ と $\boldsymbol{s}$ を比較すると，$\boldsymbol{s}$ では次元数が 32 から 8 へ削減されたにも関わらず $\boldsymbol{v}$ と遜色ない結果が得られた．また $\boldsymbol{h}$ は $\boldsymbol{v}$ よりも大幅に認識率が低下していることが分かる．このことから提案モデルは音韻と話者情報をある程度分離でき，その結果 $\boldsymbol{s}$ に話者情報が保存され，$\boldsymbol{h}$ では話者情報が削減されているということが窺える．

4.2　声 質 変 換

声質変換は，入力話者 k_i の音響特徴量 $\boldsymbol{v}_i$ と話者特徴量 $\boldsymbol{s}_i$ を入力し，式 (14) より音韻特徴量を推定した後，その音韻特徴量と出力話者 k_o の話者特徴量 $\boldsymbol{s}_o$ を用いて式 (13) より音響特徴量 $\hat{\boldsymbol{v}}_o$ を推定することで実現される．具体的には，

$$\hat{\boldsymbol{v}}_o = \mathbf{A}_{\boldsymbol{s}_o}\mathbf{M} \circ \boldsymbol{f}(\mathbf{M}^{\top}\mathbf{A}_{\boldsymbol{s}_i}^{\top}(\frac{\boldsymbol{v}_i}{\boldsymbol{\sigma}^2}) + \mathbf{V}\boldsymbol{s}_i + \boldsymbol{c}) + \mathbf{U}^{\top}\boldsymbol{s}_o + \boldsymbol{b}$$

と計算される．ただし，$\boldsymbol{s}_i$ と $\boldsymbol{s}_o$ はそれぞれ $s_{k_i} = 1$，$s_{k_o} = 1$ となる one-hot ベクトルである．

声質変換の精度を測る指標として，以下で定義される MDIR (mel-cepstral distortion improvement ratio) を用いた．

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

ここで $\boldsymbol{m}_o$ は入力話者音声とアライメントをとった出力話者音声のメルケプストラム特徴量を表す．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．

比較手法には，従来のパラレルデータを使用する代表的な声質変換手法である GMM (64 混合)，パラレルデータを使用しない手法として ARBM [25] と SATBM [26] を用いた．また，参考までに 3. 章で述べた制約を加えずに，Three-way ポテンシャルの結合重み $\mathcal{Z}$ をフリーパラメータとした 3WRBM による声質変換とも比較した．GMM との比較のために，それぞれの手法において，入力話者に男性 1 名，出力話者に女性 1 名を選んで，$R = 2$ としてモデルの学習を行っている．

実験結果を Table 3 に示す．全ての手法の中で最も高い精度で変換できたのは GMM であった．しかし，GMM は他の手法と異なり学習時にパラレルデータを使用しなければならない．つまり，GMM と他の手法は単純に変換精度だけで比較することができないため，Table 3 では参考値として載せている．パラレルデータを使用しない手法同士を比較すると，提案手法で最も高い精度が得られた．特に 3WRBM と提案モデルを比較すると，3. 章で述べた構造化や制約が重要であることが分かる．3WRBM では無制約であるため，音韻情報と話者情報の分離が十分に行えず，$\boldsymbol{h}$ の中に話者情報を多く含むようになってしまうため，入力話者と出力話者の音響特徴量から推定される $\boldsymbol{h}$ が大きく異なってしまい，全く声質変換できないという結果となった．一方提案モデルでは GMM に少し劣る程度の変換精度が得られた．

5. おわりに

本稿では音韻・話者因子の分離を考慮した制約付き Three-way RBM (3WRBM) による音声モデリング手法を提案した．また 3-step サンプリングによる 3WRBM のパラメータ推定法を提案した．話者認識と声質変換の 2 つの音声信号処理タスクを通じて提案モデルにおける音声モデリングの性能を検証した．話者認識実験では，音響特徴量から推定される $\boldsymbol{s}$ は話者認識率が高く，$\boldsymbol{h}$ は話者認識率が低いことから，本モデルにはある程度音韻・話者情報の分離能力を持つことが確認できた．声質変換実験では，パラレルデータを使用しない声質変換において提案モデルにより大幅に精度を向上させることができた．また，音韻・話者の分離にはそれぞれの因子を考慮した構造化が非常に有効であることが確認できた．今後音韻特徴量を音声認識に用いるなど，本モデルの可能性について，さらに検証を続けていきたい．

文　献

[1] G.E. Hinton, S. Osindero, and Y.W. Teh, “A fast learning algorithm for deep belief nets,” Neural computation, vol.18, no.7, pp.1527–1554, 2006.

[2] A.r. Mohamed, G.E. Dahl, and G. Hinton, “Acoustic modeling using deep belief networks,” IEEE Transactions on Audio, Speech, and Language Processing, vol.20, no.1, pp.14–22, 2012.

[3] K. Cho, A. Ilin, and T. Raiko, “Improved learning of Gaussian-Bernoulli restricted Boltzmann machines,” Artificial Neural Networks and Machine Learning–ICANN 2011, pp.10–17, Springer, 2011.

[4] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, “Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech,” Speech Communication, vol.54, no.1, pp.134–146, 2012.

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

[6] 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.

[7] 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.

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

[9] Y. Stylianou, O. Cappé, and E. Moulines, “Continuous probabilistic transform for voice conversion,” IEEE Transactions on Speech and Audio Processing, vol.6, no.2, pp.131–142, 1998.

[10] 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, vol.15, no.8, pp.2222–2235, 2007.

[11] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, “Voice conversion using partial least squares regression,” IEEE Transactions on Audio, Speech, and Language Processing, vol.18, no.5, pp.912–921, 2010.

[12] D. Saito, S. Watanabe, A. Nakamura, and N. Minematsu, “Probabilistic integration of joint density model and speaker model for voice conversion,” Proc. Interspeech, pp.1728–1731, 2010.

[13] 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.

[14] 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.

[15] R. Aihara, T. Takiguchi, and Y. Ariki, “Individuality-preserving voice conversion for articulation disorders using phoneme-categorized exemplars,” ACM Transactions on Accessible Computing (TACCESS), vol.6, no.4, p.13, 2015.

[16] 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.

[17] T. Nakashika, T. Takiguchi, and Y. Ariki, “Voice conversion using RNN pre-trained by recurrent temporal restricted Boltzmann machines,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.23, no.3, pp.580–587, 2015.

[18] 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.

[19] 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.

[20] L. Sun, S. Kang, K. Li, and H. Meng, “Voice conversion using deep bidirectional long short-term memory based recurrent neural networks,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.4869–4873, 2015.

[21] 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), pp.3893–3896IEEE, 2009.

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

[23] A. Mouchtaris, J.V. der Spiegel, and P. Mueller, “Nonparallel training for voice conversion based on a parameter adaptation approach,” IEEE Transactions on Audio, Speech, and Language Processing, vol.14, no.3, pp.952–963, 2006.

[24] 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.

[25] T. Nakashika, T. Takiguchi, and Y. Ariki, “Parallel-data-free, many-to-many voice conversion using an adaptive restricted Boltzmann machine,” Proc. Machine Learning in Spoken Language Processing (MLSLP), pp.1–6, 2015.

[26] T. Nakashika, and T. Takiguchi, “Non-parallel voice conversion using combination of restricted Boltzmann machine and speaker-adaptive training,” Proc. Acoustical Society of Japan, pp.223–226, 2015.

[27] Y. Freund, and D. Haussler, Unsupervised learning of distributions of binary vectors using two layer networks, Computer Research Laboratory, 1994.

[28] A. Krizhevsky, G.E. Hinton, et al., “Factored 3-way restricted Boltzmann machines for modeling natural images,” International Conference on Artificial Intelligence and Statistics, pp.621–628, 2010.

[29] T.J. Sejnowski, “Higher-order Boltzmann machines,” AIP Conference Proceedings, vol.151, no.1, pp.398–403, 1986.

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