Vol.2016-MUS-111 No.21
2016/5/21

Three-way restricted Boltzmann machineによる音声モデリングに基づく話者・音素の同時認識

中鹿 亘[1,a)]　南 泰浩[1,b)]

概要：これまで様々な音声認識（音素認識）技術や話者認識技術が提案されてきたが，それらの要素技術は別々に研究されてきた．本研究では，音響特徴量・潜在的音韻特徴量・話者特徴量の 3 つを変数とする Three-Way Restricted Boltzmann Machine (3WRBM) を用いた音声モデリングによって，話者と音素を同時に認識する手法を提案する．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．このモデルの特徴として，各特徴量の相互条件付確率を容易に計算することができる．これにより，音響特徴量が与えられた時の条件付き確率を最大化する音韻特徴量及び話者特徴量を一つのモデルで同時推定することができる．評価実験では，3WRBM による話者・音素の認識実験の結果を報告し，その有効性について議論する．

Speech modeling using three-way restricted Boltzmann machine for simultaneous speaker-phoneme recognition

Nakashika Toru[1,a)]　Minami Yasuhiro[1,b)]

1. はじめに

現在最も代表的であり効果的な音声認識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，出力確率として，連続確率密度関数である GMM (Gaussian mixture model) が用いられる．つまり，一般的な HMM を用いた音声認識では，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] に基づく音声認識では，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．

一方，話者認識手法として，話者依存 GMM を用いる手法 [3], [4] が一般的に広く用いられている．その応用として，平均声モデル（Universal background model; UBM）からの話者適応に基づく手法 [5]，多空間確率分布を用いた手法 [6]，SVM (support vector machine) を組み合わせた手法 [7] などが挙げられる．また，GMM 以外の代表的な話者認識手法として VQ (vector quantization) を用いた手法 [8]，ANN (artificial neural network) を用いた手法 [9] などがある．しかしこれらの手法ではケプストラムベースの特徴量を用いており，音韻性が含まれた状態でモデルを学習させるため，話者性を適切に捉えた認識モデルになってい

[1] 電気通信大学
The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu, Tokyo 182-8585, Japan
[a)] nakashika@uec.ac.jp
[b)] minami.yasuhiro@is.uec.ac.jp

るとは言い難い．そこでケプストラム特徴量から話者性を効果的に取り出すため MLLR (Maximum likelihood linear regression [10]) に基づく手法 [11], [12]，音素依存 GMM を用いた手法 [13]，PCA (principal componet analysis) による音韻抑制に基づく手法 [14] などが提案された．

以上のように，これまで音声認識・話者認識手法はそれぞれ別々に研究されてきた．例えば多くの話者認識の研究では，話者認識を実現するためのモデリング手法が言及されており，学習済みの話者認識モデルが音声認識へ利用されることは稀である．同一のモデルで音声認識・話者認識を行うことができれば，別々の認識システムを構築する必要がないため，学習コストや記憶領域の大幅な削減が期待される．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 [15] 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関係性を表す結合重みが存在すると仮定している．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．なお，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，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．さらに，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）ことも応用例として考えられる．

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 章では提案モデルを話者・音声認識へ応用する手法について述べる．5 章で話者認識・連続音素認識の評価実験について述べ，6 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 [15]）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 [16]．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 [17]. 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．このとき，エネルギー関数は，RBM のものから

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

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

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

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

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．

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

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

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

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，$\boldsymbol{g}(\cdot)$ 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 [18] 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．

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，

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

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．

さらに，同時に 2 つ以上の音韻が作用していることは現実には起こりえないので，音韻は一つの要素のみがアクティブになるという制約（$\sum_j h_j = 1$）を加える．

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．

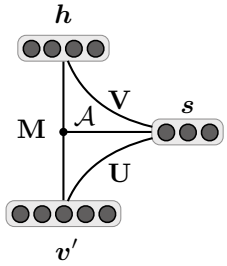

図 1 提案手法における音声モデリングのグラフ構造.
Fig. 1 Graphical representation of the proposed model.

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

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

$$p(\boldsymbol{v}|\boldsymbol{h},\boldsymbol{s}) = \mathcal{N}(\boldsymbol{v} \mid \boldsymbol{b} + \mathbf{U}^\top\boldsymbol{s} + \mathbf{A}_{\boldsymbol{s}}\mathbf{M}\boldsymbol{h}, \boldsymbol{\sigma}^{\boldsymbol{2}}) \tag{13}$$
$$p(\boldsymbol{h}|\boldsymbol{s},\boldsymbol{v}) = \mathcal{B}(\boldsymbol{h} \mid \boldsymbol{g}(\boldsymbol{c} + \mathbf{V}\boldsymbol{s} + \mathbf{M}^\top\mathbf{A}_{\boldsymbol{s}}^\top\boldsymbol{v}') \tag{14}$$
$$p(\boldsymbol{s}|\boldsymbol{v},\boldsymbol{h}) = \mathcal{B}(\boldsymbol{s} \mid \boldsymbol{g}(\mathbf{U}\boldsymbol{v}' + \mathbf{V}^\top\boldsymbol{h} + [\boldsymbol{v}'^\top\mathbf{A}_k]\mathbf{M}\boldsymbol{h})) \tag{15}$$

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，

$$\frac{\partial\mathcal{L}}{\partial\mathbf{M}} = \langle\sum_k \mathbf{A}_k^\top\boldsymbol{v}'\boldsymbol{h}^\top s_k\rangle_{\text{data}} - \langle\sum_k \mathbf{A}_k^\top\boldsymbol{v}'\boldsymbol{h}^\top s_k\rangle_{\text{model}} \tag{17}$$
$$\frac{\partial\mathcal{L}}{\partial\mathbf{A}_k} = \langle\boldsymbol{v}'\boldsymbol{h}^\top s_k\mathbf{M}^\top\rangle_{\text{data}} - \langle\boldsymbol{v}'\boldsymbol{h}^\top s_k\mathbf{M}^\top\rangle_{\text{model}} \tag{18}$$
$$\frac{\partial\mathcal{L}}{\partial\mathbf{U}} = \langle\boldsymbol{s}\boldsymbol{v}'^\top\rangle_{\text{data}} - \langle\boldsymbol{s}\boldsymbol{v}'^\top\rangle_{\text{model}} \tag{19}$$
$$\frac{\partial\mathcal{L}}{\partial\mathbf{V}} = \langle\boldsymbol{h}\boldsymbol{s}^\top\rangle_{\text{data}} - \langle\boldsymbol{h}\boldsymbol{s}^\top\rangle_{\text{model}} \tag{20}$$
$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{b}} = \langle\boldsymbol{v}'\rangle_{\text{data}} - \langle\boldsymbol{v}'\rangle_{\text{model}} \tag{21}$$
$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{c}} = \langle\boldsymbol{h}\rangle_{\text{data}} - \langle\boldsymbol{h}\rangle_{\text{model}} \tag{22}$$
$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\sigma}} = \frac{1}{\boldsymbol{\sigma}^3}\circ\Big(\langle\boldsymbol{v}\circ\boldsymbol{v} - 2\boldsymbol{v}\circ\big(\boldsymbol{b} + \mathbf{U}^\top\boldsymbol{s} + \mathbf{A}_{\boldsymbol{s}}\mathbf{M}\boldsymbol{h}\big)\rangle_{\text{data}} - \langle\boldsymbol{v}\circ\boldsymbol{v} - 2\boldsymbol{v}\circ\big(\boldsymbol{b} + \mathbf{U}^\top\boldsymbol{s} + \mathbf{A}_{\boldsymbol{s}}\mathbf{M}\boldsymbol{h}\big)\rangle_{\text{model}}\Big)$$

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．

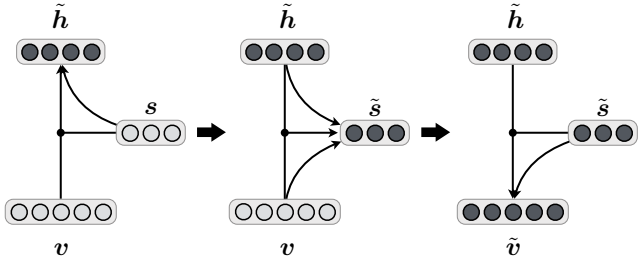

図 2 提案モデルにおける 3-step サンプリング.
Fig. 2 3-step sampling used in the training.

4. 音声・話者認識への応用

提案モデルは音韻・話者情報を考慮した生成モデルであるため，一度学習が終われば様々な音声信号処理タスクへ応用することができる．まず，提案モデルを音声・話者の同時認識へ応用する手法について述べる．

評価データのフレーム音響特徴量 $\boldsymbol{v}$ が与えられたとき，音韻 h_j と話者 s_k が同時にアクティブとなる条件付確率はそれぞれ以下のように計算される．

$$p(h_j = 1, s_k = 1|\boldsymbol{v}) = \frac{p(\boldsymbol{v}, h_j = 1, s_k = 1)}{\sum_{\boldsymbol{h},\boldsymbol{s}} p(\boldsymbol{v},\boldsymbol{h},\boldsymbol{s})} = g(c_j + V_{jk} + \mathbf{U}_{k:}\boldsymbol{v}' + \boldsymbol{v}'^\top\mathbf{A}_k\boldsymbol{m}_j) \tag{23}$$

ただし V_{jk}，$U_{k:}$ はそれぞれ $\mathbf{V}$ の第 j,k 要素，$\mathbf{U}$ の第 k 行ベクトルを表す．ここで，

$$f_{jk} \triangleq c_j + V_{jk} + \mathbf{U}_{k:}\boldsymbol{v}' + \boldsymbol{v}'^\top\mathbf{A}_k\boldsymbol{m}_j \tag{24}$$

とすると，最適な音韻 $h_{\hat{j}}$ と話者 $s_{\hat{k}}$ を以下のように推定することができる．

$$(\hat{j},\hat{k}) = \underset{(j,k)}{\text{argmax}}\, p(h_j = 1, s_k = 1|\boldsymbol{v}) = \underset{(j,k)}{\text{argmax}}\, f_{jk} \tag{25}$$

すなわち，音声・話者を同時認識するためには，各要素が式 (24) で表される行列 $\mathbf{F} = [f_{jk}]$ を計算し，最も値の大きい要素インデックスを求めればよい．

表 1　話者認識実験の結果.
Table 1　Speaker recognition accuracy of each method.

Method	GMM	UBM	3WRBM	3WRBM (ideal)
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 recognition accuracies using various features.

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

なお，本研究における音素認識に関する実験では，効率良く評価するために，既存の音声認識システムを用いる．音韻情報 $\boldsymbol{h}$ は，話者情報と切り離されているため，話者情報を含んだ音響特徴量 $\boldsymbol{v}$ よりも音声認識システムの入力特徴量として適切であると考えられる．音響特徴量 $\boldsymbol{v}$ が与えられたときに得られる音韻情報の期待値は以下のように計算される．

$$\begin{aligned}\mathbf{E}[\boldsymbol{h}|\boldsymbol{v}] &= [p(h_j=1|\boldsymbol{v})] \\ &= \left[\frac{\sum_{\boldsymbol{s}} p(\boldsymbol{v}, h_j=1, \boldsymbol{s})}{\sum_{\boldsymbol{h}',\boldsymbol{s}'} p(\boldsymbol{v}, \boldsymbol{h}', \boldsymbol{s}')}\right] \\ &= \boldsymbol{c} + \boldsymbol{q}(\mathbf{V} + \boldsymbol{v}'^{\top}\mathbf{U}^{\top} + \mathbf{M}^{\top}[\mathbf{A}_k^{\top}\boldsymbol{v}'])\end{aligned} \tag{26}$$

ただし，$\boldsymbol{q}(\mathbf{X}) = \log\sum_k e^{\mathbf{X}_{:k}}$ は要素ごとの一般化 softplus 関数を表す．式 (26) で計算されるベクトルを既存の音声認識システムの入力に用いて学習・認識を行う．

5. 評価実験

提案モデルの効果を確かめるため，話者認識と音素認識実験を行った．本実験では日本音響学会研究用連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，5 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[19] 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 の確率的勾配法を用いてモデルを学習した．

5.1 話者認識に関する実験結果

まず，話者認識に関する実験結果を報告する．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 $N_{\text{corr.}}$ とすると，$100\cdot N_{\text{corr.}}/N_{\text{all}}$ として話者認識率を算出した．比較手法として GMM による話者認識および UBM に基づく手法を用いた．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，UBM では予め全話者の音声を用いて GMM を学習しておき，その後話者ごとの GMM を再学習させた．

表 3　連続音素認識実験の結果.
Table 3　Continuous phone recognition (correct rate [%]) with changing the number of phonemes to be recognized.

	5 vowels	+5 cons.	+10 cons.
$\boldsymbol{v}$ (mcep)	53.53	43.18	36.52
$\boldsymbol{h}$	**59.34**	41.61	33.03

実験結果を Table 1 に示す．GMM と UBM は識別的なアプローチであり，提案手法である 3WRBM は生成的なアプローチであるため単純に精度のみで比較すべきではないが，本実験では GMM が最も高い精度が得られ，次いで UBM となった．なお，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（3WRBM (ideal)）．

次に，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，$\boldsymbol{v}$ 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．

5.2 音声認識に関する実験結果

最後に，提案手法によって得られる音韻特徴量の効果を調べる実験を行った．本実験で用いるデータセットはフレームレベルの音素ラベルが与えられていないため，フレームレベルの音素認識率を計算することが困難である．音素ラベルの代わりに与えられている発話文ごとの書き起こしテキストを用いて，連続音素 HMM に基づく音素認識実験を行った．各音素 HMM は開始と終了を除く 3 状態とし，各状態の出力確率として 32 混合の GMM を用いた．HMM の入力特徴量として式 (26) で得られる音韻特徴量の期待値を用いた．また，比較のためメルケプストラム特徴量を入力特徴量とした場合の結果を調べた．本実験では「5 母音のみ」「5 母音+5 子音」「5 母音+10 子音」の 3 つのケースについて音素正解率を算出した．実験結果を Table 3 に示す．Table 3 に見られるように，「5 母音のみ」の場合において提案手法（$\boldsymbol{h}$）はメルケプストラム特徴量よりも有効であることが確認できた．これは得られた音韻特

徴量がメルケプストラム特徴量よりも話者に依存しない，音声認識に有効な特徴量となったからだと考えられる．また Table 3 によれば，それ以外の，子音を考慮した場合では音韻特徴量があまり効果的ではなかったことが分かる．これは，子音のスペクトルは一般的に母音と比べて非定常的であり，学習データに頻出するパターンとして音韻特徴量では捉えることができなかったからだと言える．一方メルケプストラム特徴量ではその細かな変動を HMM を通じて捉えることができたためだと考えられる．

6. 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およびそれを用いた音韻・話者の同時認識手法を提案した．話者認識と連続音素認識の 2 つの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．連続音素認識実験では，母音のような定常的な音声パターンに対して本モデルが効果的であることが分かった．今後はフレームレベルで音素認識評価できるデータセットを用いて音韻・話者の同時認識に関する評価を行いたい．また，音韻情報や話者情報に関して一部教師を与えるなど，半教師学習に基づく手法を検討したい．

参考文献

[1] Hinton, G. E., Osindero, S. and Teh, Y.-W.: A fast learning algorithm for deep belief nets, *Neural computation*, Vol. 18, No. 7, pp. 1527–1554 (2006).
[2] Mohamed, A.-r., Dahl, G. E. and Hinton, G.: Acoustic modeling using deep belief networks, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 20, No. 1, pp. 14–22 (2012).
[3] Reynolds, D. A. and Rose, R. C.: Robust text-independent speaker identification using Gaussian mixture speaker models, *IEEE Transactions on Speech and Audio Processing*, Vol. 3, No. 1, pp. 72–83 (1995).
[4] Kinnunen, T. and Li, H.: An overview of text-independent speaker recognition: From features to supervectors, *Speech communication*, Vol. 52, No. 1, pp. 12–40 (2010).
[5] Reynolds, D. A. and Dunn, T. F. Q. R. B.: Speaker verification using adapted Gaussian mixture models, *Digital signal processing*, Vol. 10, No. 1, pp. 19–41 (2000).
[6] Miyajima, C., Hattori, Y., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T.: Text-independent speaker identification using Gaussian mixture models based on multi-space probability distribution, *IEICE Transactions on Information and Systems*, Vol. 84, No. 7, pp. 847–855 (2001).
[7] Campbell, W. M., Campbell, J. P., Reynolds, D. A., Singer, E. and Torres-Carrasquillo, P. A.: Support vector machines for speaker and language recognition, *Computer Speech & Language*, Vol. 20, No. 2, pp. 210–229 (2006).
[8] Burton, D. K.: Text-dependent speaker verification using vector quantization source coding, *Acoustics, Speech and Signal Processing, IEEE Transactions on*, Vol. 35, No. 2, pp. 133–143 (1987).
[9] Farrell, K. R., Mammone, R. J. and Assaleh, K. T.: Speaker recognition using neural networks and conventional classifiers, *Speech and Audio Processing, IEEE Transactions on*, Vol. 2, No. 1, pp. 194–205 (1994).
[10] Leggetter, C. J. and Woodland, P. C.: Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models, *Computer Speech & Language*, Vol. 9, No. 2, pp. 171–185 (1995).
[11] Mak, M.-W., Hsiao, R.-C. and Mak, B.: A comparison of various adaptation methods for speaker verification with limited enrollment data, *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, Vol. 1, IEEE, pp. I–I (2006).
[12] Karam, Z. N. and Campbell, W. M.: A new kernel for SVM MLLR based speaker recognition., *INTERSPEECH*, pp. 290–293 (2007).
[13] Castaldo, F., Colibro, D., Dalmasso, E., Laface, P. and Vair, C.: Compensation of nuisance factors for speaker and language recognition, *IEEE Transactions on Audio, Speech, and Language Processing, volume=15, number=7, pages=1969–1978, year=2007, publisher=IEEE.*
[14] Lu, H., Nishida, M., Horiuchi, Y. and Kuroiwa, S.: Text-Independent speaker identification in phoneme-independent subspace using PCA transformation, *International Journal of Biometrics*, Vol. 2, No. 4, pp. 379–390 (2010).
[15] Cho, K., Ilin, A. and Raiko, T.: Improved learning of Gaussian-Bernoulli restricted Boltzmann machines, *Artificial Neural Networks and Machine Learning–ICANN 2011*, Springer, pp. 10–17 (2011).
[16] Freund, Y. and Haussler, D.: *Unsupervised learning of distributions of binary vectors using two layer networks*, Computer Research Laboratory (1994).
[17] Sejnowski, T. J.: Higher-order Boltzmann machines, *AIP Conference Proceedings*, Vol. 151, No. 1, pp. 398–403 (1986).
[18] Krizhevsky, A., Hinton, G. E. et al.: Factored 3-way restricted Boltzmann machines for modeling natural images, *International Conference on Artificial Intelligence and Statistics*, pp. 621–628 (2010).
[19] Morise, M.: An attempt to develop a singing synthesizer by collaborative creation, *SMAC2013*, pp. 287–292 (2013).