Restricted Boltzmann Machine を用いた話者性・雑音を考慮したモデリングの検討*

☆高島悠樹 (神戸大), 中鹿亘 (電通大), 滝口哲也 (神戸大, JST さきがけ), 有木康雄 (神戸大)

1 はじめに

近年，スマートフォンの普及に伴い，実環境下での音声認識の利用機会は増加している [1]．音声認識の研究は盛んに行なわれ，その精度は向上しているが，一般に雑音環境下においては精度が著しく低下することが知られている．実環境下においては，背景雑音等の影響により，音声の品質は静音下と比較して劣化する．そのため，実環境で利用する音声認識システムは耐雑音性を備えていることが望まれる．

雑音環境下音声認識手法として，音響モデルを雑音環境に適応させるモデル適応と，雑音重畳音声に対して雑音成分を抑制する雑音抑制がある．本研究では，後者の手法について検討を行う．雑音抑制の手法として，SS (spectral subtraction)，NMF (nonnegative matrix factorization) を用いた雑音分離 [2, 3]，denoising autoencoder を用いた雑音抑制 [4, 5] などが挙げられる．しかしながら，これらの手法では，雑音データやパラレルデータ (クリーン音声と雑音重畳音声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かかる，使用するデータセットが制限される，音声に不自然な変換が加わってしまうなど様々な弊害が生じる．パラレルデータを必要としない手法として，noise adaptive training [6] による雑音適応がある．これは，クリーン音声パラメータと雑音パラメータを分離して学習する手法であり，後に述べる話者適応学習と同様の手法である．

また，一般に音声認識精度を劣化させる原因として学習データと評価データ話者の違いによるモデルのミスマッチも挙げられる．このミスマッチへの対処法として，MLLR (maximum likelihood linear regression) や CMLLR (constrained MLLR) による音響モデル，特徴量の話者適応が広く知られている．さらに，話者依存項と話者非依存項を用意して音響モデルの学習を行う話者正規化学習 (SAT; speaker adaptive training [7]) が提案されている．

本研究では，音響モデリングの観点から音声認識に優位な特徴量を検討する．具体的には，複数話者の雑音重畳音声を，RBM (restricted Boltzmann machine [8]) を用いて, 話者性及び雑音の両方を考慮しつつモデリングを行う．RBM は，観測ベクトルを表現する可視素子，潜在情報を表す隠れ素子，可視素子-隠れ素子間の結合重みを変数とする確率モデルである．これまでの RBM ベースの音響モデリングとして，ARBM (adaptive restricted Boltzmann machine [9]) を用いた手法，SATBM (speaker-adaptive-trainable Boltzmann machine [10]) を用いた手法が提案されてきた．前者は，話者に依存した結合重みが存在すると仮定し，話者非依存の結合重みを話者固有の行列により射影することによりモデルに話者性を反映させた．後者は，隠れ素子が音韻性を表現すると仮定し，物理量を考慮したモデル化，問題の再設定を行なった．これらの手法により (特に後者)，RBM を用いたモデル化により音声信号から話者性及び音韻性を識別的にモデリングできる可能性が示唆されている．いずれの手法も声質変換をタスクとしていたが，本稿では雑音環境下音声認識を目標として，認識に優位な特徴量の獲得を目指す．

2 Restricted Boltzmann Machine

RBM は，Fig. 1 に示すように，可視素子 $\boldsymbol{v} \in \mathbb{R}^D$ と隠れ素子 $\boldsymbol{h} \in \mathbb{B}^H$ ($\mathbb{B}$ は 0 または 1 のみを取り得る空間) からなる無向グラフィカルモデルである．入力として連続値を定義した IGB (Improved Gaussian-Bernoulli)-RBM [11](以下，この IGB-RBM を単に RBM とする) では，その同時確率とエネルギー関数は以下の式で表される．

$$
\begin{aligned}
p(\boldsymbol{v}, \boldsymbol{h}) &= \frac{1}{Z} \mathrm{e}^{-E(\boldsymbol{v}, \boldsymbol{h})} \\
E(\boldsymbol{v}, \boldsymbol{h}) &= \frac{1}{2}(\boldsymbol{v}-\boldsymbol{b})^{\top} \boldsymbol{\Sigma}^{-1}(\boldsymbol{v}-\boldsymbol{b}) - \boldsymbol{v}^{\top}\boldsymbol{\Sigma}^{-1}\mathbf{W}\boldsymbol{h} - \boldsymbol{c}^{\top}\boldsymbol{h}
\end{aligned}
\tag{1}
$$

ここで，$\boldsymbol{b} \in \mathbb{R}^D$, $\boldsymbol{c} \in \mathbb{R}^H$, $\mathbf{W} \in \mathbb{R}^{D \times H}$, $\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\sigma}^2)$ はそれぞれ，可視素子バイアス，隠れ素子バイアス，可視層-隠れ層間の結合重み，可視素子の分散共分散行列を表し，いずれも推定すべきパラメータである．また，$\boldsymbol{\sigma}^2 = [\sigma_1^2, \cdots, \sigma_D^2]$ とする．ここで，$Z = \int^D \Sigma_{\boldsymbol{h}} \mathrm{e}^{-E(\boldsymbol{v}, \boldsymbol{h})} d^D\boldsymbol{v}$ は全域での確率を 1 にするための正規化項である．

RBM では，可視素子間, 及び隠れ素子間の接続は存在せず，可視素子，隠れ素子は互いに条件付き独立

---

* Acoustic modeling using restricted Boltzmann machine considering speaker and noise, by Yuki Takashima (Kobe University), Toru Nakashika (UEC), Tetsuya Takiguchi (Kobe University, JST PRESTO), Yasuo Ariki (Kobe University)

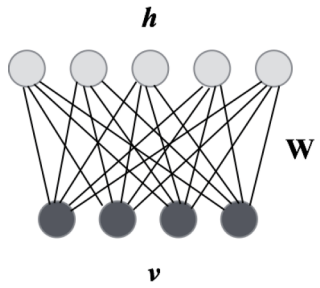

Fig. 1　Graphical representation of an RBM

であるため，それぞれの条件付き確率は以下のような単純な式で表現される．

$$p(\boldsymbol{v}|\boldsymbol{h}) = \mathcal{N}(\boldsymbol{v}|\mathbf{W}\boldsymbol{h} + \boldsymbol{b}, \boldsymbol{\Sigma}) \tag{2}$$

$$p(\boldsymbol{h}|\boldsymbol{v}) = \mathcal{S}(\mathbf{W}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{v} + \boldsymbol{c}) \tag{3}$$

ここで，$\mathcal{N}(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ は平均 $\boldsymbol{\mu}$, 分散共分散 $\boldsymbol{\Sigma}$ の各次元独立な多変量正規分布，$\mathcal{S}(\cdot)$ は要素ごとのシグモイド関数を表す．

RBM の各パラメータは，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．この対数尤度をパラメータ θ で偏微分すると，

$$\frac{\partial \mathcal{L}}{\partial \theta} = \langle \frac{\partial E(\boldsymbol{v}, \boldsymbol{h})}{\partial \theta} \rangle_{\text{data}} - \langle \frac{\partial E(\boldsymbol{v}, \boldsymbol{h})}{\partial \theta} \rangle_{\text{model}}, \tag{4}$$

が得られる．ここで，$\langle\cdot\rangle_{\text{data}}$ 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，後者は一般に計算困難なため Contrastive Divergence 法 [12] を用いて求められる．各パラメータは式 (4) から，確率的勾配法 (SGD) を用いて繰り返し更新される．

3　提案手法と音声認識への応用

本稿では，前節で述べた RBM を拡張し，話者性及び背景雑音を考慮したモデルを定義する．さらに，提案するモデルを用いて音声認識タスクへ応用する手法について述べる．

3.1　提案モデルの定義

一般に，音声信号に対して話者性に関する情報は乗算的に，背景雑音に関する情報は加算的に付与されることが知られている．時刻 t において，音声信号 $x(t)$，乗算性雑音 (話者性) $a(t)$，加算性雑音 (背景雑音) $q(t)$ とすると，観測信号 $o(t)$ は以下のように表される．

$$o(t) = a(t) * x(t) + q(t) \tag{5}$$

ここで，$*$ は畳み込み演算を表す．周波数ドメインにおいて，式 (5) は以下の式で記述される．

$$O_t(\omega) = A(\omega)X_t(\omega) + Q(\omega) \tag{6}$$

ここで，ω は周波数ビンのインデクスを表す．また，本稿において，話者性及び背景雑音は時不変の変数であると仮定する．

ここで，時刻 t における話者 r，背景雑音 n の観測信号を，周波数ドメインにおいて，式 (6) より以下のように表現する．

$$\boldsymbol{o}_{rtn} = \mathbf{A}_r \boldsymbol{x}_t + \boldsymbol{q}_n \tag{7}$$

ここで，$\boldsymbol{o} \in \mathbb{R}^D$，$\boldsymbol{x}_t \in \mathbb{R}^D$，$\boldsymbol{q}_n \in \mathbb{R}^D$ はそれぞれ各次元が周波数ビンに対応するベクトルであり，$\mathbf{A}_r = \text{diag}(\boldsymbol{a}_r)$ は話者行列である．ただし，$\boldsymbol{a}_r = [a_1^r, \cdots, a_D^r] \in \mathbb{R}^D$ である．

ここで，中鹿ら [10] の知見から，音声信号が RBM でモデル化されると仮定すると，RBM の定義より，$\boldsymbol{x}_t \sim \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma})$ と表される．ここで，$\boldsymbol{\mu}_t$, $\boldsymbol{\Sigma}$ はそれぞれ，音声信号の平均ベクトル，分散共分散行列 (対角) を表す．また，雑音信号が正規分布に従うと仮定し，$\boldsymbol{q}_n \sim \mathcal{N}(\boldsymbol{\xi}_n, \boldsymbol{\Delta}_n)$ と表現されるものとする．ここで，$\boldsymbol{\xi}_n \in \mathbb{R}^D$,$\boldsymbol{\Delta}_n = \text{diag}(\boldsymbol{\delta}_n^2)$ はそれぞれ，雑音 n の平均ベクトル，分散共分散行列を表す．ただし，$\boldsymbol{\delta}_n^2 = [{\delta_1^{(n)}}^2, \cdots, {\delta_D^{(n)}}^2]$ とする．これらの仮定のもとで，観測ベクトル $\boldsymbol{o}_{rtn}$ は以下の正規分布に従う．

$$\boldsymbol{o}_{rtn} \sim \mathcal{N}(\mathbf{A}_r\boldsymbol{\mu}_t + \boldsymbol{\xi}_n, \mathbf{A}_r\boldsymbol{\Sigma}\mathbf{A}_r^\top + \boldsymbol{\Delta}_n) \tag{8}$$

さらに，RBM の定義より，$\boldsymbol{\mu}_t = \mathbf{W}\boldsymbol{h}_t + \boldsymbol{b}$ と表現されるため，式 (8) は以下のように展開できる．

$$\mathcal{N}\left(\mathbf{A}_r(\mathbf{W}\boldsymbol{h}_t + \boldsymbol{b}) + \boldsymbol{\xi}_n, \mathbf{A}_r\boldsymbol{\Sigma}\mathbf{A}_r^\top + \boldsymbol{\Delta}_n\right) \tag{9}$$

$$= \mathcal{N}\left(\mathbf{A}_r\mathbf{W}\boldsymbol{h}_t + \mathbf{A}_r\boldsymbol{b} + \boldsymbol{\xi}_n, \mathbf{A}_r\boldsymbol{\Sigma}\mathbf{A}_r^\top + \boldsymbol{\Delta}_n\right) \tag{10}$$

ここで，$\hat{\mathbf{W}}_r = \mathbf{A}_r\mathbf{W}$, $\hat{\boldsymbol{b}}_{rn} = \mathbf{A}\boldsymbol{b} + \boldsymbol{\xi}_n$, $\hat{\boldsymbol{\Sigma}}_{rn} = \mathbf{A}_r\boldsymbol{\Sigma}\mathbf{A}_r^\top + \boldsymbol{\Delta}_n$ とおくと，

$$\boldsymbol{o}_{rtn} \sim \mathcal{N}(\hat{\mathbf{W}}_r\boldsymbol{h}_t + \hat{\boldsymbol{b}}_{rn}, \hat{\boldsymbol{\Sigma}}_{rn}) \tag{11}$$

となる．ここで，式 (11) を RBM の可視素子の確率分布に照らし合わせると，式 (1) より，

$$p(\boldsymbol{o}_{rtn}, \boldsymbol{h}_t) = \frac{1}{Z}\mathrm{e}^{-E'(\boldsymbol{o}_{rtn}, \boldsymbol{h}_t)} \tag{12}$$

$$E'(\boldsymbol{o}_{rtn}, \boldsymbol{h}_t) = \frac{1}{2}(\boldsymbol{o}_{rtn} - \hat{\boldsymbol{b}}_{rn})^\top \hat{\boldsymbol{\Sigma}}_{rn}^{-1}(\boldsymbol{o}_{rtn} - \hat{\boldsymbol{b}}_{rn}) - \boldsymbol{o}_{rtn}^\top \hat{\boldsymbol{\Sigma}}_{rn}^{-1}\hat{\mathbf{W}}_r\boldsymbol{h}_t - \boldsymbol{c}^\top\boldsymbol{h}_t \tag{13}$$

と定義することで，話者性及び背景雑音を考慮した音響モデリングを RBM により表現することができる．すなわち，標準話者の音声信号ベクトルを可視素子，潜在的音韻情報ベクトルを隠れ素子とした RBM において，話者固有の射影行列により話者適応を，雑音固有のバイアスにより雑音適応を施したモデルとみなすことができる．可視素子及び隠れ素子の条件付き確率は通常の RBM と同様に計算される．

提案するRBMも，通常のRBMと同様にパラメータを推定することができる．提案するRBMのパラメータは，話者に依存するもの $\mathbf{\Theta}^{SD} = \{\mathbf{A}_r\}_{r=1}^{R}$，雑音に依存するもの $\mathbf{\Theta}^{ND} = \{\boldsymbol{\xi}_n, \boldsymbol{\delta}_n^2\}_{n=1}^{N}$，話者と雑音の両方に依存しないもの $\mathbf{\Theta}^{SNI} = \{\mathbf{W}, \boldsymbol{\sigma}^2, \boldsymbol{b}, \boldsymbol{c}\}$ に分けることができる．これらは R 人の話者による N 個の背景雑音のもとで収録された音声データ $\boldsymbol{X} = \{\boldsymbol{X}_{rn}\}_{r=1,n=1}^{R,N}$, $\boldsymbol{X}_{rn} = \{\boldsymbol{o}_{rtn}\}_{t=1}^{T_{rn}}$ に対する尤度を最大化するように同時に推定される．すなわち，

$$(\hat{\mathbf{\Theta}}^{SD}, \hat{\mathbf{\Theta}}^{ND}, \hat{\mathbf{\Theta}}^{SNI},) \triangleq \underset{\mathbf{\Theta}^{SD}, \mathbf{\Theta}^{ND}, \mathbf{\Theta}^{SNI}}{\arg\max} \prod_{r=1}^{R} \prod_{n=1}^{N} \prod_{t=1}^{T_{rn}} p(\boldsymbol{o}_{rtn}) \tag{14}$$

とする．

通常のRBMと同様に勾配法によって，パラメータを更新するため，パラメータ θ に対する対数尤度 $\mathcal{L}' = \log \prod_r \prod_n \prod_t p(\boldsymbol{o}_{rtn}) = \Sigma_r \Sigma_n \log \Sigma_h p(\boldsymbol{o}_{rtn}, \boldsymbol{h}_t)$ の偏微分を考える．各パラメータの偏微分値 $\frac{\partial E'(\boldsymbol{o}_{rtn}, \boldsymbol{h}_t)}{\partial \theta}$ を計算すると以下の式が得られる．

$$\begin{aligned}
\frac{\partial E'(\boldsymbol{o}_{rtn}, \boldsymbol{h}_t)}{\partial \mathbf{A}_r} &= -\big(\hat{\mathbf{\Sigma}}_{rn}^{-1}\mathbf{C}_{rtn}\hat{\mathbf{\Sigma}}_{rn}^{-1}\mathbf{A}_r\mathbf{\Sigma} \\
&\quad - \hat{\mathbf{\Sigma}}_{rn}^{-1}\mathbf{D}_{rtn}\hat{\mathbf{\Sigma}}_{rn}^{-1}(\mathbf{A}_r\mathbf{\Sigma} - \mathbf{\Delta}_n\mathbf{A}^{-1}) + K\big) \\
\frac{\partial E'(\boldsymbol{o}_{rtn}, \boldsymbol{h}_t)}{\partial \boldsymbol{\xi}_n} &= -\hat{\mathbf{\Sigma}}_{rn}^{-1}(\boldsymbol{o}_{rtn} - \hat{\boldsymbol{b}}_{rn}) \\
\frac{\partial E'(\boldsymbol{o}_{rtn}, \boldsymbol{h}_t)}{\partial \boldsymbol{\delta}^2} &= -\mathrm{diag}(\hat{\mathbf{\Sigma}}_{rn}^{-1}\mathbf{E}_{rtn}\hat{\mathbf{\Sigma}}_{rn}^{-1}) \\
\frac{\partial E'(\boldsymbol{o}_{rtn}, \boldsymbol{h}_t)}{\partial \mathbf{W}} &= -\mathbf{A}_r^{\top}\hat{\mathbf{\Sigma}}_{rn}^{-1}\boldsymbol{o}_{rtn}\boldsymbol{h}_t \\
\frac{\partial E'(\boldsymbol{o}_{rtn}, \boldsymbol{h}_t)}{\partial \boldsymbol{\sigma}^2} &= -\mathrm{diag}(\mathbf{A}_r^{\top}\hat{\mathbf{\Sigma}}_{rn}^{-1}\mathbf{E}_{rtn}\hat{\mathbf{\Sigma}}_{rn}^{-1}\mathbf{A}_r) \\
\frac{\partial E'(\boldsymbol{o}_{rtn}, \boldsymbol{h}_t)}{\partial \boldsymbol{b}} &= -\mathbf{A}_r^{\top}\hat{\mathbf{\Sigma}}_{rn}^{-1}(\boldsymbol{o}_{rtn} - \hat{\boldsymbol{b}}_{rn}) \\
\frac{\partial E'(\boldsymbol{o}_{rtn}, \boldsymbol{h}_t)}{\partial \boldsymbol{c}} &= -\boldsymbol{h}_t
\end{aligned}$$

ただし，

$$\begin{aligned}
\mathbf{C}_{rtn} &= \frac{1}{2}(\boldsymbol{o}_{rtn} - \hat{\boldsymbol{b}}_{rn})(\boldsymbol{o}_{rtn} - \hat{\boldsymbol{b}}_{rn})^{\top} \\
\mathbf{D}_{rtn} &= \mathbf{A}_r\boldsymbol{b}\boldsymbol{o}_{rtn}^{\top} - \hat{\mathbf{W}}_r\boldsymbol{h}_t\boldsymbol{o}_{rtn} \\
\mathbf{E}_{rtn} &= \frac{1}{2}(\boldsymbol{o}_{rtn} - \hat{\boldsymbol{b}}_{rn})(\boldsymbol{o}_{rtn} - \hat{\boldsymbol{b}}_{rn})^{\top} - \hat{\mathbf{W}}_r\boldsymbol{h}_t\boldsymbol{o}_{rtn}^{\top}
\end{aligned}$$

及び K をデータとモデルの期待値に差のない項と置いた．

3.2 提案するRBMを用いた音声認識

提案するRBMを用いて音声認識を行う場合，まず事前学習として複数の参照話者によるクリーン音声データを用いて話者依存パラメータ $\mathbf{\Theta}^{SD}$ と話者と雑音に非依存なパラメータ集合 $\mathbf{\Theta}^{SNI}$ を同時推定する．このとき，雑音パラメータ集合 $\mathbf{\Theta}^{ND}$ は存在しないものとして扱い，学習も行なわない．次に，雑音重畳音声データを用いて，雑音パラメータ $\mathbf{\Theta}^{ND}$ も含めてパラメータ推定を行う．

続いて，得られたパラメータ群を用いて，入力特徴量から式 (2) と同様にして，潜在的特徴量（隠れ素子）を推定する．この時，話者性は $\mathbf{\Theta}^{SD}$ で，雑音成分は $\mathbf{\Theta}^{ND}$ で制御され，隠れ素子は話者及び雑音に依存しない音韻に近い情報を表すと考えられる．そこで，本研究では，この隠れ素子を新たな特徴量として，HMM (hidden Markov model) を用いて音声認識を行う．

4 評価実験

4.1 実験条件

本稿では認識実験は行なわず，提案モデルが雑音成分を推定できることを検証する．

本実験では，ATR 研究用日本語音声データベース [13] より，男性話者 3 名，女性話者 3 名の計 6 名の音声を用いて，提案するモデルの有効性を調べた．このコーパスから，音素バランス単語 216 単語を各話者について用意し，パラメータの学習に用いた．スペクトルドメインにおけるモデリングのため，モデルの学習に用いる入力特徴量として，音声信号から計算される振幅スペクトルを用いた．STFT におけるフレーム幅，シフト幅，周波数ビンの数はそれぞれ 25ms, 10ms, 512 ビンであり，Hamming 窓を使用した．この振幅スペクトルに対し，シミュレーション的にそれぞれ，低域，中域，高域に高い平均を持つ，3 種類の正規乱数 (分散は 1) を付加し雑音環境を想定した．(順に $n = \{1, 2, 3\}$ の雑音とする) 振幅スペクトルは負値を取らないため，生成した乱数が負値の場合には値を 0 に置き換えた．本稿で用いる RBM は，可視素子に正規分布を仮定しており，負値をとらない振幅スペクトルをそのまま入力特徴量として使用すると都合が悪い．そこで，本実験では，振幅スペクトルのガウス性を高めるために，ZCA whitening により正規化を行なった．提案する RBM における学習率，繰り返し回数はそれぞれ 0.01,120 とした．ミニバッチ法を用いてモデルの学習を行い，各バッチは，各コンディション (6 話者*3 雑音=18 通り) から 32 フレームずつ抽出し作成した．提案する RBM の隠れ素子数は特徴量として使うことを考慮し 15 とした．

4.2 実験結果と考察

理想的なパラメータの値と，提案法によって実際に推定されたパラメータの比較を Figs. 2, 3 に示す．それぞれ雑音平均パラメータ $\{\boldsymbol{\xi}_n\}_{n=1}^{3}$ と雑音分散パラメータ $\{\boldsymbol{\delta}_n^2\}_{n=1}^{3}$ であり，左図が理想パラメータ，右図が推定されたパラメータである．青，緑，赤線はそ

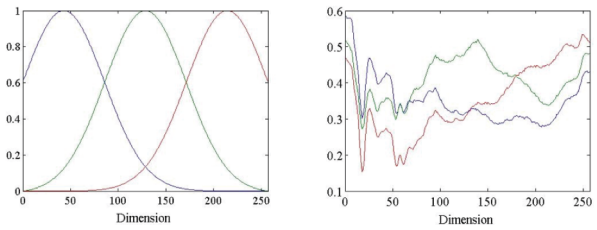

Fig. 2　Ideal parameters and estimated parameters ($\{\boldsymbol{\xi}_n\}_{n=1}^3$)

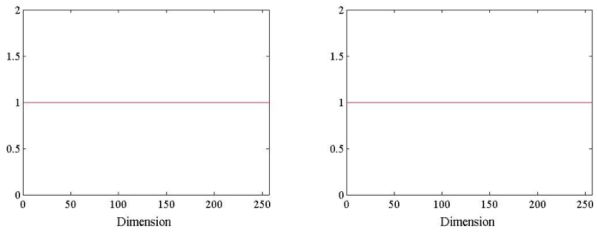

Fig. 3　Ideal parameters and estimated parameters ($\{\boldsymbol{\delta}_n^2\}_{n=1}^3$)

れぞれ，3 種類の雑音環境 ($n = \{1, 2, 3\}$) のパラメータを表す．

図より，雑音平均パラメータは，低，中，高域に高い平均を持って推定されており，ばらつきはあるものの概ね期待通りの推定が行なわれていることが分かる．雑音分散パラメータに関して，推定されたパラメータは理想的なパラメータと完全に一致している．しかし，雑音はピークを持つ帯域から離れるほど付加されなくなる傾向にあるため (平均が 0 に近くなるに従い負値が生成されることが多くなり，0 に置き換えられる値が増えるため)，実際にはピークを持つ帯域から離れるほど分散の値は小さくなることが予想される．推定されたパラメータが上述の結果にならなかった原因として，初期値の問題や勾配が小さかったことが挙げられる．

5　おわりに

本研究では，潜在的な特徴量を抽出する RBM を拡張して，話者項，雑音項，話者と雑音に依存しない項を分離してパラメータを学習するモデリング法を提案した．本稿では，モデルにおける各項の分離可能性を示したが，音声認識実験を行ない，さらなるモデルの検証を行ないたい．また，現状のモデルは，雑音項を時間に依存しないと仮定しているため，非定常な雑音を考慮することは難しいと考えられる．実環境においては，非定常な雑音環境下の場合の方が多いと考えられるため，さらなるモデルの拡張を検討していきたい．

参考文献

[1] 総務省，“平成 26 年版情報通信白書,” .

[2] K. W. Wilson *et al.*, “Speech denoising using nonnegative matrix factorization with priors,” in *ICASSP*. 2008, pp. 4029–4032, IEEE.

[3] F. Weninger *et al.*, “Non-negative matrix factorization for highly noise-robust ASR: To enhance or to recognize?,” in *ICASSP*. 2012, IEEE.

[4] T. Ishii *et al.*, “Reverberant speech recognition based on denoising autoencoder,” in *INTERSPEECH*, F. Bimbot *et al.*, Eds. 2013, pp. 3512–3516, ISCA.

[5] X. Feng *et al.*, “Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition,” in *ICASSP*. 2014, pp. 1759–1763, IEEE.

[6] O. Kalinli *et al.*, “Noise adaptive training for robust automatic speech recognition.,” IEEE Transactions on Audio, Speech & Language Processing, vol. 18, no. 8, pp. 1889–1901, 2010.

[7] A. Anastasakos *et al.*, “A compact model for speaker-adaptive training,” in *ICSLP*, 1996, vol. 2, pp. 1137–1140.

[8] Y. Freund and D. Haussler, “Unsupervised learning of distributions on binary vectors using two layer networks,” Tech. Rep., 1994.

[9] T. Nakashika *et al.*, “Parallel-dictionary-free voice conversion using adaptive restricted boltzmann machine,” in *Acoustical Society of Japan 2015 Spring Meeting*, 2015, pp. 279–282.

[10] T. Nakashika and T. Takiguchi, “Non-parallel voice conversion using combination of restricted boltzmann machine and speaker-adaptive training,” in *Acoustical Society of Japan 2015 Autumn Meeting*, 2015, pp. 223–226.

[11] A. L. K. Cho and T. Raiko, “Improved learning of Gaussian-Bernoulli restricted Boltzmann machines,” in *Artificial Neural Networks and Machine Learning*, 2011, pp. 10–17.

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

[13] A. Kurematsu *et al.*, “ATR Japanese speech database as a tool of speech recognition and synthesis,” Speech Communication, vol. 9, no. 4, pp. 357–363, 1990.