制限ボルツマンマシンの複素数拡張と音声合成への応用

代表研究者	中　鹿　亘	電気通信大学	情報理工学研究科	助教
共同研究者	高　木　信　二	国立情報学研究所	コンテンツ科学研究系	特任助教
共同研究者	山　岸　順　一	国立情報学研究所	コンテンツ科学研究系	准教授

1 はじめに

近年，ディープラーニングを用いた手法が飛躍的に高い精度を上げ，画像認識や音声認識など，幅広い分野において盛んに研究されている[1]．これまでに数多くのディープラーニング手法が提案されているが，最も代表的なモデルとして，制限ボルツマンマシン（restricted Boltzmann machine; RBM）を多層に積み重ねた Deep Belief Net (DBN)が挙げられる[2]．また，RBM 単独でも特徴量抽出手法としてしばしば用いられ，様々な RBM の拡張モデルも提案されている[3,4,5]．

以上のように，RBM はこれまで様々な分野で用いられてきたにもかかわらず，いずれのアプローチでも入力特徴量はバイナリまたは実数値が仮定されてきた[2,6,7]．しかし，音声の複素スペクトル，MRI 画像，音響インテンシティなど，実データに基づいた画像認識や音声信号処理では複素数データを取り扱う場合が多い．音声認識や音声合成では，位相よりも振幅スペクトルの方が認識や合成に効果的であることが知られているため，音響特徴量として MFCC やメルケプストラム特徴量，STRAIGHT スペクトルなどの振幅スペクトルに基づいた特徴量がしばしば利用される．

RBM を用いた特徴抽出でも入力には実数値のメルケプストラム特徴量が利用される場合が多い．しかしこれらは位相情報が欠落しており，元の音声データに対して少なからず情報の損失が存在するため，位相情報を含めた複素数データをそのまま表現することが重要となる．

本研究では，複素数データ（複素スペクトル）から直接潜在的な特徴量を抽出する，ディープラーニングや特徴抽出手法の根幹をなす RBM の拡張モデル（complex-valued RBM; CRBM; 複素 RBM）を新たに定義し，複素 RBM による音声モデリング手法について検討する．本研究では，人工データを用いた実験及び音声符号化・復号化実験を通じ，複素 RBM の有用性について調査した．複素 RBM は複素入力の各素子の実部と虚部間の接続（または複素入力とその共役との接続）を考慮しているため，従来の RBM では表現できなかった，実部と虚部に相関のある複素数データをより正確に表現することができる．また，従来の RBM と同様，可視素子間及び隠れ素子間には接続がないと仮定しているため，Gibbs サンプリングや CD (contrastive divergence)法を用いてパラメータを効率よく推定することもできる．

また本研究では，複素 RBM を揺るぎない手法として確立させるため，1)複素 PCA による複素スペクトル圧縮，2) 複素 MLPG による複素時系列データ生成，3)複素 Adam による高速学習アルゴリズムの３つの改善手法と，複素スペクトル系列の時間相関を表現する複素 RBM の拡張モデル(複素 TRBM)についても検討を行った．

2 複素制限ボルツマンマシン（複素 RBM）

RBM(Restricted Boltzmann machine)は，入力データを表現する可視素子と，潜在的な情報を表現する隠れ素子の間に双方向な接続重みが存在する（ただし可視素子間または隠れ素子間には接続はない）と仮定した確率モデルである．従来，RBM はバイナリ値の入力が想定されていた[6]が，実数値データを扱うために Gaussian-Bernoulli RBM が後に提案された[7]．その後様々な RBM の拡張モデルが提案されてきたが，複素数の入力を仮定するモデルはこれまで提案されなかった．そこで本研究では，先駆的に複素数の入力を仮定する RBM の拡張モデル（複素制限ボルツマンマシン；複素 RBM）を提案し，その有効性を調査した．

2-1 モデルの定義

確率密度関数（パラメータ推定時におけるコスト関数）は従来同様実数であるが，入力が実数ではなく複素数となる RBM の拡張モデル（Complex-valued RBM；複素 RBM と呼ぶ）を考える．本研究では，I 次元の複素数データ $\boldsymbol{z} \in \mathbb{C}^I$ を可視素子とする複素 RBM を以下のように定義する．

$$
\begin{aligned}
p(\boldsymbol{z};\boldsymbol{\theta}) =& \sum_{\boldsymbol{h}} p(\boldsymbol{z},\boldsymbol{h};\boldsymbol{\theta}) \\
p(\boldsymbol{z},\boldsymbol{h};\boldsymbol{\theta}) =& \frac{1}{U(\boldsymbol{\theta})} e^{-E(\boldsymbol{z},\boldsymbol{h};\boldsymbol{\theta})} \\
E(\boldsymbol{z},\boldsymbol{h};\boldsymbol{\theta}) =& \frac{1}{2} \begin{bmatrix} \boldsymbol{z} \\ \bar{\boldsymbol{z}} \end{bmatrix}^H \boldsymbol{\Phi}^{-1} \begin{bmatrix} \boldsymbol{z} \\ \bar{\boldsymbol{z}} \end{bmatrix} \\
& - \begin{bmatrix} \boldsymbol{b} \\ \bar{\boldsymbol{b}} \end{bmatrix}^H \boldsymbol{\Phi}^{-1} \begin{bmatrix} \boldsymbol{z} \\ \bar{\boldsymbol{z}} \end{bmatrix} - 2\boldsymbol{c}^\top \boldsymbol{h} \\
& - \begin{bmatrix} \boldsymbol{z} \\ \bar{\boldsymbol{z}} \end{bmatrix}^H \boldsymbol{\Phi}^{-1} \begin{bmatrix} \mathbf{W} \\ \bar{\mathbf{W}} \end{bmatrix} \boldsymbol{h} \\
U(\boldsymbol{\theta}) =& \int \sum_{\boldsymbol{h}} e^{-E(\boldsymbol{z},\boldsymbol{h};\boldsymbol{\theta})} d\boldsymbol{z}
\end{aligned}
$$

ただし，$\boldsymbol{b} \in \mathbb{C}^I$, $\boldsymbol{c} \in \mathbb{R}^J$はそれぞれ可視素子と隠れ素子のバイアス，$\mathbf{W} \in \mathbb{C}^{I \times J}$は可視素子-隠れ素子間の複素結合重みを表し，オーバーラインは複素共役，$.^H$はエルミート転置を表す．また，

$$
\boldsymbol{\Phi} \triangleq \begin{bmatrix} \boldsymbol{\Gamma} & \mathbf{C} \\ \mathbf{C}^H & \boldsymbol{\Gamma}^H \end{bmatrix}
$$

であり，

$\boldsymbol{\Gamma} \triangleq \Delta(\boldsymbol{\gamma}), \quad \boldsymbol{\gamma} \in \mathbb{R}+^I$

$\mathbf{C} \triangleq \Delta(\boldsymbol{\delta}), \quad \boldsymbol{\delta} \in \mathbb{C}^I$

はそれぞれ複素数 z の分散と擬似分散（共役複素数との共分散）を表すパラメータとする．

結局複素 RBM のパラメータは$\boldsymbol{\theta} = \{\boldsymbol{b}, \boldsymbol{c}, \mathbf{W}, \boldsymbol{\gamma}, \boldsymbol{\delta}\}$となる．ここで，

$$
\boldsymbol{p} \triangleq \frac{\boldsymbol{\gamma}}{\boldsymbol{\gamma}^2 - |\boldsymbol{\delta}|^2} \in \mathbb{R}^I
$$

$$
\boldsymbol{q} \triangleq -\frac{\boldsymbol{\delta}}{\boldsymbol{\gamma}^2 - |\boldsymbol{\delta}|^2} \in \mathbb{C}^I
$$

を導入する（ただし分数線は要素除算を表す）と，

$$
\boldsymbol{\Phi}^{-1} = \begin{bmatrix} \Delta(\boldsymbol{p}) & \Delta(\boldsymbol{q}) \\ \Delta(\bar{\boldsymbol{q}}) & \Delta(\boldsymbol{p}) \end{bmatrix}
$$

となることから，上述のエネルギー関数 E は

$$
\begin{aligned}
& E(\boldsymbol{z},\boldsymbol{h};\boldsymbol{\theta}) = \\
& \boldsymbol{z}^H \Delta(\boldsymbol{p}) \boldsymbol{z} + \Re(\boldsymbol{z}^H \Delta(\boldsymbol{q}) \bar{\boldsymbol{z}}) - 2\Re(\boldsymbol{z}^H \Delta(\boldsymbol{p}) \boldsymbol{b}) \\
& - 2\Re(\boldsymbol{z}^H \Delta(\boldsymbol{q}) \bar{\boldsymbol{b}}) - 2\boldsymbol{c}^\top \boldsymbol{h} - 2\Re(\boldsymbol{z}^H \Delta(\boldsymbol{p}) \mathbf{W}) \boldsymbol{h} \\
& - 2\Re(\boldsymbol{z}^H \Delta(\boldsymbol{q}) \bar{\mathbf{W}}) \boldsymbol{h}
\end{aligned}
$$

と書き直すことができ，エネルギー関数は実数値となること，複素可視素子 z の各次元は共役複素数との結合が存在するが，通常の RBM のように次元間の結合は存在しないことが確認できる．さらに，

$$
\boldsymbol{b}' \triangleq \Delta(\boldsymbol{p}) \boldsymbol{b} + \Delta(\boldsymbol{q}) \bar{\boldsymbol{b}}
$$

$$
\mathbf{W}' \triangleq \Delta(\boldsymbol{p}) \mathbf{W} + \Delta(\boldsymbol{q}) \bar{\mathbf{W}}
$$

と置き換えることで,

$$E(\boldsymbol{z},\boldsymbol{h};\boldsymbol{\theta}) =$$
$$\frac{1}{2}\boldsymbol{z}^H\Delta(\boldsymbol{p})\boldsymbol{z} + \frac{1}{2}\bar{\boldsymbol{z}}^H\Delta(\boldsymbol{p})\bar{\boldsymbol{z}} + \boldsymbol{z}^H\Delta(\boldsymbol{q})\bar{\boldsymbol{z}}$$
$$+ \bar{\boldsymbol{z}}^H\Delta(\bar{\boldsymbol{q}})\boldsymbol{z} - \boldsymbol{z}^H\boldsymbol{b}' - \bar{\boldsymbol{z}}^H\bar{\boldsymbol{b}}' - 2\boldsymbol{c}^\top\boldsymbol{h}$$
$$- \boldsymbol{z}^H\mathbf{W}'\boldsymbol{h} - \bar{\boldsymbol{z}}^H\bar{\mathbf{W}}'\boldsymbol{h}$$

となり，図 1 で示すように z と h，z の共役と h の関係性は互いに共役空間を挟んで鏡像の関係にあることが分かる．さらに，z とその共役は，q の重みで接続されている．

以上の定義により，隠れ素子が与えられたときの可視素子の条件付き確率，および可視素子が与えられたときの隠れ素子の条件付き確率をそれぞれ以下のように表すことができる．

$$p(\boldsymbol{z}|\boldsymbol{h}) = \mathcal{CN}(\boldsymbol{z};\boldsymbol{b}+\mathbf{W}\boldsymbol{h},\boldsymbol{\Gamma},\mathbf{C})$$
$$p(\boldsymbol{h}|\boldsymbol{z}) = \mathcal{B}(\boldsymbol{h};\boldsymbol{f}(2\boldsymbol{c}+2\Re(\mathbf{W}'^H\boldsymbol{z})))$$

ただし，$\mathcal{CN}(\cdot;\boldsymbol{\mu},\boldsymbol{\Gamma},\mathbf{C})$は平均$\boldsymbol{\mu}$，分散共分散行列$\boldsymbol{\Gamma}$，擬似分散共分散行列$\mathbf{C}$の多変量複素正規分布：

$$p(\boldsymbol{z}) = \frac{1}{\pi^D\sqrt{\det(\boldsymbol{\Gamma})\det(\mathbf{Q})}}$$
$$\cdot\exp\left\{-\frac{1}{2}\begin{bmatrix}\boldsymbol{z}-\boldsymbol{\mu}\\ \bar{\boldsymbol{z}}-\bar{\boldsymbol{\mu}}\end{bmatrix}^H\begin{bmatrix}\boldsymbol{\Gamma} & \mathbf{C}\\ \mathbf{C}^H & \boldsymbol{\Gamma}^H\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{z}-\boldsymbol{\mu}\\ \bar{\boldsymbol{z}}-\bar{\boldsymbol{\mu}}\end{bmatrix}\right\}$$

$\mathbf{Q} = \bar{\boldsymbol{\Gamma}} - \mathbf{C}^H\boldsymbol{\Gamma}^{-1}\mathbf{C}$
を表す.

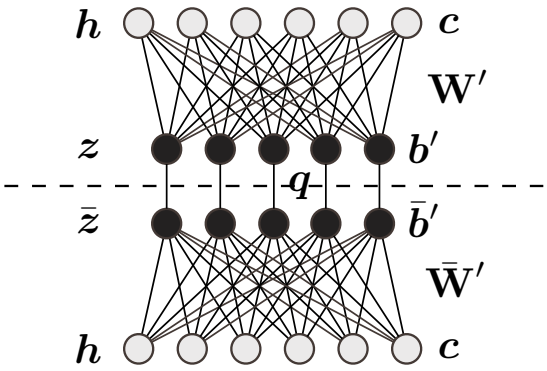

図 1 複素 RBM のグラフ図.

2-2 モデルの学習

複素 RBM の各パラメータは，N 個の独立な複素数データ集合$\{\boldsymbol{z}_n\}_{n=1}^N$の対数尤度：

$$L(\boldsymbol{\theta}) = \log\prod_{n=1}^N p(\boldsymbol{z}_n;\boldsymbol{\theta})$$
$$= \sum_{n=1}^N \log\sum_{\boldsymbol{h}_n} p(\boldsymbol{z}_n,\boldsymbol{h}_n;\boldsymbol{\theta})$$

を最大化するように複素勾配降下法によって更新する（または，次節で述べる複素 Adam で更新する）．
複素勾配降下法は学習率$\alpha>0$を用いて

$$\boldsymbol{\theta}^{(\mathrm{new})} \leftarrow \boldsymbol{\theta}^{(\mathrm{old})} + 2\alpha \frac{\partial L}{\partial \bar{\boldsymbol{\theta}}}$$

を繰り返し計算することでパラメータを更新する．ただし，上式における複素数の偏微分はウェルティンガーの微分を表す．一方，対数尤度に対する偏微分は

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = \mathbb{E}_{p(\boldsymbol{h}|\boldsymbol{z})}\left[-\frac{\partial E(\boldsymbol{z},\boldsymbol{h})}{\partial \boldsymbol{\theta}}\right] - \mathbb{E}_{p(\boldsymbol{z}',\boldsymbol{h}')}\left[-\frac{\partial E(\boldsymbol{z}',\boldsymbol{h}')}{\partial \boldsymbol{\theta}}\right]$$

と導けるので，各パラメータのエネルギー関数に対する偏微分を用いて計算される．ただし， 右辺第１項は観測データが与えられた時の隠れ素子の期待値，第２項は モデルの期待値を表す．モデルの期待値は一般に計算困難であるが，p(z|h) と p(h|z)の計算及びそれらに従う乱数サンプリングが容易であるため，従来の RBM と同様，CD 法（contrastive divergence）を用いて近似することができる．具体的には，モデルの期待値を，あるデータ z_n が与えられたときの隠れ素子の条件付き確率 $p(h_n|z_n)$に従う乱数を生成し，その隠れ素子が与えられたときの可視素子の条件付き確率に従う乱数（再構築データ）が与えられた時の隠れ素子の期待値として計算する．エネルギー関数に対するパラメータの偏微分は解析的に求めることができ，それぞれ以下の通りになる．

$$-\frac{\partial E}{\partial \boldsymbol{b}} = \Delta(\boldsymbol{p})\bar{\boldsymbol{z}} + \Delta(\bar{\boldsymbol{q}})\boldsymbol{z}$$
$$-\frac{\partial E}{\partial \boldsymbol{c}} = \boldsymbol{h}$$
$$-\frac{\partial E}{\partial \mathbf{W}} = (\Delta(\boldsymbol{p})\bar{\boldsymbol{z}} + \Delta(\bar{\boldsymbol{q}})\boldsymbol{z})\boldsymbol{h}^{\top}$$
$$-\frac{\partial E}{\partial \boldsymbol{\gamma}} = (\boldsymbol{p}^2 + |\boldsymbol{q}|^2) \circ \frac{\partial E}{\partial \boldsymbol{p}} + 2\Re(\boldsymbol{p} \circ \boldsymbol{q} \circ \frac{\partial E}{\partial \boldsymbol{q}})$$
$$-\frac{\partial E}{\partial \boldsymbol{\delta}} = \boldsymbol{p}^2 \circ \frac{\partial E}{\partial \boldsymbol{q}} + \bar{\boldsymbol{q}}^2 \circ \frac{\partial E}{\partial \bar{\boldsymbol{q}}} + 2\boldsymbol{p} \circ \bar{\boldsymbol{q}} \circ \frac{\partial E}{\partial \boldsymbol{p}}$$

ただし，

$$\frac{\partial E}{\partial \boldsymbol{p}} = \frac{1}{2}|\boldsymbol{z}|^2 - \Re(\boldsymbol{z} \circ (\bar{\boldsymbol{b}} + \bar{\mathbf{W}}\boldsymbol{h}))$$
$$\frac{\partial E}{\partial \boldsymbol{q}} = \frac{1}{2}\bar{\boldsymbol{z}}^2 - \bar{\boldsymbol{z}} \circ (\bar{\boldsymbol{b}} + \bar{\mathbf{W}}\boldsymbol{h})$$

である．分散および擬似分散の更新は，他のパラメータと比較してスケールが異なるため，安定して学習させるために実際には指数をとった変数でパラメータ更新を行う．

3 音声モデリングにおける複素 RBM 学習アルゴリズムの改善

前節で述べたモデルを，音声信号の複素スペクトルのモデリングにそのまま適用するだけでは不十分であり，新モデルを揺るぎない音声モデリング手法として確立させるため，本研究ではさらに，1)複素 PCA による複素スペクトル圧縮，2)MLPG(最尤系列パラメータ生成)による複素時系列データ生成，3)複素 Adam による高速学習アルゴリズムの３つの改善手法を検討・評価した．

3-1 複素 Adam による最適化

ニューラルネットワークのパラメータ最適化手法として Adam（adaptive moment estimation）[8] が提案され，勾配降下法よりも収束が早く，性能が高いことが報告されている．本研究では，実数値で定義されていた Adam を複素数に拡張した複素 Adam（CAdam）を提案する．

複素 Adam では，前述した複素勾配降下法の代わりに，ハイパーパラメータである減衰率 $\beta_1 \in [0,1)$，$\beta_2 \in [0,1)$ を用いて

$$
\begin{aligned}
\beta_1^{(\mathrm{new})} &\leftarrow \beta_1 \cdot \beta_1^{(\mathrm{old})} \\
\beta_2^{(\mathrm{new})} &\leftarrow \beta_2 \cdot \beta_2^{(\mathrm{old})} \\
\boldsymbol{m}^{(\mathrm{new})} &\leftarrow \beta_1^{(\mathrm{new})} \boldsymbol{m}^{(\mathrm{old})} \\
&\quad + (1-\beta_1^{(\mathrm{new})}) \left(\frac{\partial L}{\partial \Re(\boldsymbol{\theta})} + i \frac{\partial L}{\partial \Im(\boldsymbol{\theta})} \right) \\
\boldsymbol{v}^{(\mathrm{new})} &\leftarrow \beta_2^{(\mathrm{new})} \boldsymbol{v}^{(\mathrm{old})} \\
&\quad + (1-\beta_2^{(\mathrm{new})}) \left(\left(\frac{\partial L}{\partial \Re(\boldsymbol{\theta})} \right)^2 + \left(\frac{\partial L}{\partial \Im(\boldsymbol{\theta})} \right)^2 \right)
\end{aligned}
$$

$$
\boldsymbol{\theta}^{(\mathrm{new})} \leftarrow \boldsymbol{\theta}^{(\mathrm{old})} + \alpha \cdot \frac{\sqrt{1-\beta_2^{(\mathrm{new})}}}{1-\beta_1^{(new)}} \cdot \frac{\boldsymbol{m}^{(new)}}{\sqrt{\boldsymbol{v}^{(\mathrm{new})}}}
$$

と各パラメータを繰り返し更新する.

3-2 複素 PCA による次元削減

一般に,1フレームあたりの音声複素スペクトルの次元数は多くなる（例えば分析窓長が1,024の場合，複素スペクトルの次元数は 513）ため，動的特徴量やセグメント特徴量を使用すると膨大な次元数となってしまう．そこで本研究では複素 PCA[9]を用いて複素数データの次元圧縮を行った.

フレーム t の複素スペクトルを o_t とすると、複素 PCA で複素特徴量 z_t は

$$
\boldsymbol{z}_t = \boldsymbol{\Lambda}_{1:P}^{-\frac{1}{2}} \mathbf{U}_{:,1:P}^H \boldsymbol{o}_t
$$

と表される．ただし$\boldsymbol{\Lambda}_{1:P}^{-\frac{1}{2}}$は複素スペクトルの共分散行列の上位 P 個の固有値の平方の逆数を対角成分とする対角行列を表し，$\mathbf{U}_{:,1:P}$は複素スペクトルの共分散行列の複素固有ベクトルのうち，上位 P 個の固有値に対応する複素固有ベクトルを列ベクトルとした複素行列を表す．逆に，複素特徴量から複素スペクトルを復元する場合は,

$$
\boldsymbol{o}_t = \mathbf{U}_{:,1:P} \boldsymbol{\Lambda}_{1:P}^{\frac{1}{2}} \boldsymbol{z}_t
$$

を計算すれば良い.

後述の評価実験では,分析窓長 256 の複素スペクトルに対して上記の複素 PCA を施し P=40 次元に圧縮した静的な複素特徴量 z_t と，その動的特徴量$\Delta \boldsymbol{z}_t$とを連結した複素特徴量$\boldsymbol{Z}_t \triangleq [\boldsymbol{z}_t^H \ \Delta \boldsymbol{z}_t^H]^H$を複素 RBM の入力特徴量とする.

3-3 複素 MLPG による複素時系列データ生成

一般に音声信号は時相関を持つため,フレーム独立なパラメータ生成は適切ではない．そこで本研究では，最尤パラメータ生成法（maximum likelihood parameter generation; MLPG）[10]を複素数表現に拡張し，フレーム間の相関を考慮しながら複素時系列データを復元する手法を提案・検討した．MLPG は，モデルから推定されたフレーム静的特徴量とフレーム動的特徴量から，最尤法により最適な系列特徴量を推定する手法である．従来の MLPG は実数の特徴量を対象としてきたが，そのまま複素数の特徴量に対して適用することができないため，本研究では以下で述べる定式化を行った.

複素 RBM を学習後，フレーム数 T の評価用複素入力特徴量系列$\boldsymbol{Z}_{1:T} \triangleq [\boldsymbol{Z}_1^H \ \boldsymbol{Z}_2^H \ \cdots \ \boldsymbol{Z}_T^H]^H$をエンコードして得られた隠れ素子の期待値系列$\hat{\boldsymbol{h}}_{1:T}$から，最適な複素静的特徴量系列$\hat{\boldsymbol{z}}_{1:T} \triangleq [\hat{\boldsymbol{z}}_1^H \ \hat{\boldsymbol{z}}_2^H \ \cdots \ \hat{\boldsymbol{z}}_T^H]^H$を推定す

ることを考える．$\hat{\boldsymbol{z}}_{1:T}$は，条件付き確率$p(\boldsymbol{Z}_{1:T}|\hat{\boldsymbol{h}}_{1:T},\boldsymbol{\theta})$を最大とする系列であり，以下の通り定式化される．

$$\hat{\boldsymbol{z}}_{1:T} = \underset{\boldsymbol{z}_{1:T}}{\operatorname{argmax}}\, p(\boldsymbol{Z}_{1:T}|\hat{\boldsymbol{h}}_{1:T},\boldsymbol{\theta})$$

ここで，$\mathcal{L} \triangleq \log p(\boldsymbol{Z}_{1:T}|\hat{\boldsymbol{h}}_{1:T},\boldsymbol{\theta})$ とおくと，

$$\boldsymbol{Z}_{1:T} = \boldsymbol{S}\boldsymbol{z}_{1:T}$$
$$\boldsymbol{S} \triangleq [\boldsymbol{S}_1\ \boldsymbol{S}_2\ \cdots \boldsymbol{S}_T]^\top \otimes \boldsymbol{I}_{P\times P}$$
$$\boldsymbol{S}_t \triangleq [\boldsymbol{s}_t^{(1)}\ \boldsymbol{s}_t^{(2)}]$$

となることから，

$$\mathcal{L} = -\boldsymbol{z}_{1:T}^\top \boldsymbol{S}^\top \mathrm{diag}(\tilde{\boldsymbol{q}})\boldsymbol{S}\boldsymbol{z}_{1:T} - \boldsymbol{z}_{1:T}^\top \boldsymbol{S}^\top \mathrm{diag}(\tilde{\boldsymbol{p}})\boldsymbol{S}\bar{\boldsymbol{z}}_{1:T} + \boldsymbol{z}_{1:T}^\top \boldsymbol{S}^\top \boldsymbol{\mu}_{1:T} + K$$
$$\boldsymbol{\mu}_{1:T} \triangleq [\boldsymbol{\mu}_1^H\ \boldsymbol{\mu}_2^H\ \cdots\ \boldsymbol{\mu}_T^H]^H$$
$$\boldsymbol{\mu}_t \triangleq \boldsymbol{b}' + \mathbf{W}'\hat{\boldsymbol{h}}_t$$

と書き表すことができる．

ただしKは系列推定には関係のない定数，$\boldsymbol{s}_t^{(1)} \in \mathbb{R}^T$はtフレーム目のみ1，他は0となるスパースベクトル，$\boldsymbol{s}_t^{(2)} \in \mathbb{R}^T$は(t-1)フレーム目は-0.5，(t+1)フレーム目は0.5，他は0となるスパースベクトルを表す．また，チルダは同じベクトルをTフレーム繰り返し並べたベクトルを表す．本稿では，複素最急降下法を用いて最尤系列$\hat{\boldsymbol{z}}_{1:T}$を推定する．具体的には，$\hat{\boldsymbol{z}}_t$の初期値をフレームワイズ推定値：

$$\underset{\boldsymbol{Z}_t}{\operatorname{argmax}}\, p(\boldsymbol{Z}_t|\hat{\boldsymbol{h}}_t,\boldsymbol{\theta}) = \boldsymbol{b} + \mathbf{W}\hat{\boldsymbol{h}}_t$$

の静的特徴量とし，学習率α>0を用いて

$$\boldsymbol{z}_{1:T}^{(\mathrm{new})} \leftarrow \boldsymbol{z}_{1:T}^{(\mathrm{old})} + 2\alpha\frac{\partial\mathcal{L}}{\partial\bar{\boldsymbol{z}}_{1:T}}$$

を繰り返し更新する（本実験ではα=0.01，繰り返し回数を100とした）．ただし，上式における複素数の偏微分はウェルティンガー微分を表す．Lの$\boldsymbol{z}_{1:T}$に対する偏微分は，以下の通り計算される．

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{z}_{1:T}} = -2\boldsymbol{S}^\top \mathrm{diag}(\tilde{\boldsymbol{q}})\boldsymbol{S}\boldsymbol{z}_{1:T} - \boldsymbol{S}^\top \mathrm{diag}(\tilde{\boldsymbol{p}})\boldsymbol{S}\bar{\boldsymbol{z}}_{1:T} + \boldsymbol{S}^\top\tilde{\boldsymbol{\mu}}$$

4 リカレント構造を持つ複素RBMの拡張モデル

前述の複素RBMでは，全てのフレームの複素特徴量$\boldsymbol{z}_t \in \boldsymbol{z}_{1:T}\ (1 \le t \le T)$について対応する隠れ素子$\boldsymbol{h}_t \in \boldsymbol{h}_{1:T}$を用いて同一の確率分布$p(\boldsymbol{z}_t,\boldsymbol{h}_t)$を与え，各フレーム特徴量が独立，すなわち

$$p(\boldsymbol{z}_{1:T},\boldsymbol{h}_{1:T}) = \prod_{t=1}^{T} p(\boldsymbol{z}_t,\boldsymbol{h}_t)$$

を仮定していた．しかし，音声信号は時間的に相関のある系列データであり，そこから得られる複素スペクトルなどの特徴量も時間的に依存関係を持つが，複素RBMではそのような関係性を考慮していなかった．そこで本研究では，時間的に依存関係のある複素特徴量系列を表現するため，複素TRBM（complex-valued temporal restricted Boltzmann machine；CTRBM)と呼ぶ新たな確率モデルについても検討した．

複素 TRBM では，T フレームの潜在変数系列 $\boldsymbol{h}_{1:T}$ をマルコフ過程と仮定し，同時確率を

$$p(\boldsymbol{z}_{1:T}, \boldsymbol{h}_{1:T}) = \prod_{t=1}^{T} p(\boldsymbol{z}_t, \boldsymbol{h}_t | \boldsymbol{h}_{t-1})$$

と表現する．ここで，条件付き確率分布 $p(\boldsymbol{z}_t, \boldsymbol{h}_t | \boldsymbol{h}_{t-1})$ は，複素 TRBM のパラメータ θ を用いて以下の通り定義される．

$$\begin{aligned}
p(\boldsymbol{z}_t, \boldsymbol{h}_t | \boldsymbol{h}_{t-1}; \boldsymbol{\theta}) =& \frac{1}{U_{\boldsymbol{h}_{t-1}}(\boldsymbol{\theta})} e^{-E(\boldsymbol{z}_t, \boldsymbol{h}_t | \boldsymbol{h}_{t-1}; \boldsymbol{\theta})} \\
E(\boldsymbol{z}_t, \boldsymbol{h}_t | \boldsymbol{h}_{t-1}; \boldsymbol{\theta}) =& \frac{1}{2} \begin{bmatrix} \boldsymbol{z}_t \\ \bar{\boldsymbol{z}}_t \end{bmatrix}^H \boldsymbol{\Phi}^{-1} \begin{bmatrix} \boldsymbol{z}_t \\ \bar{\boldsymbol{z}}_t \end{bmatrix} \\
& - \begin{bmatrix} \boldsymbol{b} \\ \bar{\boldsymbol{b}} \end{bmatrix}^H \boldsymbol{\Phi}^{-1} \begin{bmatrix} \boldsymbol{z}_t \\ \bar{\boldsymbol{z}}_t \end{bmatrix} \\
& - \boldsymbol{c}^\top \boldsymbol{h}_t - \boldsymbol{h}_{t-1}^\top \mathbf{U} \boldsymbol{h}_t \\
& - \begin{bmatrix} \boldsymbol{z}_t \\ \bar{\boldsymbol{z}}_t \end{bmatrix}^H \boldsymbol{\Phi}^{-1} \begin{bmatrix} \mathbf{W} \\ \bar{\mathbf{W}} \end{bmatrix} \boldsymbol{h}_t \\
U_{\boldsymbol{h}_{t-1}}(\boldsymbol{\theta}) =& \int \sum_{\boldsymbol{h}_t} e^{-E(\boldsymbol{z}_t, \boldsymbol{h}_t | \boldsymbol{h}_{t-1}; \boldsymbol{\theta})} d\boldsymbol{z}_t
\end{aligned}$$

$\mathbf{U} \in \mathbb{R}^{J \times J}$ は隣接する隠れ素子間の結合重みパラメータを表す．複素 RBM のエネルギー関数と比較すると，複素 TRBM では隠れ素子の再帰項 $-\boldsymbol{h}_{t-1}^\top \mathbf{U} \boldsymbol{h}_t$ が追加されていることが分かる．複素 TRBM では，複素 RBM と異なり，一つ前の時刻の隠れ素子に依存して現在の時刻の確率が定義される．また，t=1 のときの確率分布を以下のように定義する．

$$\begin{aligned}
p(\boldsymbol{z}_1, \boldsymbol{h}_1; \boldsymbol{\theta}) =& \frac{1}{U_{\boldsymbol{h}_0}(\boldsymbol{\theta})} e^{-E(\boldsymbol{z}_1, \boldsymbol{h}_1; \boldsymbol{\theta})} \\
E(\boldsymbol{z}_1, \boldsymbol{h}_1; \boldsymbol{\theta}) =& \frac{1}{2} \begin{bmatrix} \boldsymbol{z}_1 \\ \bar{\boldsymbol{z}}_1 \end{bmatrix}^H \boldsymbol{\Phi}^{-1} \begin{bmatrix} \boldsymbol{z}_1 \\ \bar{\boldsymbol{z}}_1 \end{bmatrix} \\
& - \begin{bmatrix} \boldsymbol{b} \\ \bar{\boldsymbol{b}} \end{bmatrix}^H \boldsymbol{\Phi}^{-1} \begin{bmatrix} \boldsymbol{z}_1 \\ \bar{\boldsymbol{z}}_1 \end{bmatrix} \\
& - \boldsymbol{c}_0^\top \boldsymbol{h}_1 \\
& - \begin{bmatrix} \boldsymbol{z}_1 \\ \bar{\boldsymbol{z}}_1 \end{bmatrix}^H \boldsymbol{\Phi}^{-1} \begin{bmatrix} \mathbf{W} \\ \bar{\mathbf{W}} \end{bmatrix} \boldsymbol{h}_1 \\
U_{\boldsymbol{h}_0}(\boldsymbol{\theta}) =& \int \sum_{\boldsymbol{h}_1} e^{-E(\boldsymbol{z}_1, \boldsymbol{h}_1 | \boldsymbol{h}_0; \boldsymbol{\theta})} d\boldsymbol{z}_1
\end{aligned}$$

ただし，$\boldsymbol{c}_0 \in \mathbb{R}^J$ は隠れ素子の初期バイアスパラメータである．ここで $\boldsymbol{h}_0 \triangleq \mathbf{U}^{-\top}(\boldsymbol{c}_0 - \boldsymbol{c})$ とすると，t=1 のときも確率分布を上で表すことができる．

以上の定義式より，一つ前の時刻の隠れ素子が与えられたときの可視素子および隠れ素子の条件付き確率をそれぞれ以下の計算できる．

$$
\begin{aligned}
p(\boldsymbol{z}_t|\boldsymbol{h}_t, \boldsymbol{h}_{t-1}) &= p(\boldsymbol{z}_t|\boldsymbol{h}_t) \\
&= \mathcal{N}_c(\boldsymbol{z}_t; \boldsymbol{b} + \mathbf{W}\boldsymbol{h}_t, \mathbf{\Gamma}, \mathbf{C}) \\
p(\boldsymbol{h}_t|\boldsymbol{z}_t, \boldsymbol{h}_{t-1}) &= \mathcal{B}(\boldsymbol{h}_t; \boldsymbol{\rho}(\boldsymbol{c} + \mathbf{U}^\top \boldsymbol{h}_{t-1} + 2\Re(\mathbf{W}'^H \boldsymbol{z}_t)))
\end{aligned}
$$

複素 TRBM のパラメータ推定についても，複素 RBM 同様勾配法に基づく最優推定によって行う．学習用の複素系列データ $\boldsymbol{z}_{1:T}$ の対数尤度 L は

$$
L = \log p(\boldsymbol{z}_{1:T}) = \log \sum_{\boldsymbol{h}_{1:T}} p(\boldsymbol{z}_{1:T}, \boldsymbol{h}_{1:T})
$$

と表される．ここで，どのフレームについても正規化項 $U_{\boldsymbol{h}_{t-1}}$ が等しい，と条件を緩和すると，パラメータ θ に関する偏微分は

$$
\frac{\partial L}{\partial \theta} = \mathbf{E}_{p(\boldsymbol{h}_{1:T}|\boldsymbol{z}_{1:T})}[-\frac{\partial E}{\partial \theta}] - \mathbf{E}_{p(\boldsymbol{h}_{1:T}, \boldsymbol{z}_{1:T})}[-\frac{\partial E}{\partial \theta}]
$$

と計算できる．各パラメータに関するエネルギー関数の偏微分については複素 RBM 同様解析的に求まる．なお，右辺第一項は学習データが与えられた時の隠れ素子系列についての期待値で，

$$
p(\boldsymbol{h}_{1:T}|\boldsymbol{z}_{1:T}) = \prod_{t=1}^{T} p(\boldsymbol{h}_t|\boldsymbol{z}_t, \boldsymbol{h}_{t-1})
$$

となり，確率自体は容易に導出できる．ただし，フレーム t における期待値は一つ前の時刻の隠れ素子の値が必要となる．RTRBM (recurrent TRBM [11])のようにバイナリ値の潜在変数と，その期待値を表す変数を用いて学習し，推論時に期待値を伝播させる方法も考えられるが，本研究では単純に推論時のみ隠れ素子の期待値を伝播させて近似計算する．また，右辺第二項は膨大な数の項が含まれ計算困難であるが，複素 RBM と同様 CD 法によって近似することができる．具体的には，各フレーム t について，上式右辺第一項では

$$
\begin{aligned}
\hat{\boldsymbol{h}}_t &\triangleq \mathbb{E}_{p(\boldsymbol{h}_t|\boldsymbol{z}_t, \hat{\boldsymbol{h}}_{t-1})}[\boldsymbol{h}_t] \\
&= \boldsymbol{\rho}(\boldsymbol{c} + \mathbf{U}^\top \hat{\boldsymbol{h}}_{t-1} + 2\Re(\mathbf{W}'^H \boldsymbol{z}_t))
\end{aligned}
$$

から得られる隠れ素子を用い，第二項では

$$
\tilde{\boldsymbol{h}}_t \sim p(\boldsymbol{h}_t|\boldsymbol{z}_t, \tilde{\boldsymbol{h}}_{t-1})
$$

$$
\tilde{\boldsymbol{z}}_t \sim p(\boldsymbol{z}_t|\tilde{\boldsymbol{h}}_t)
$$

から得られるサンプルおよび期待値を用いて勾配を計算する．

5 評価実験

5-1 人工データを用いた複素 RBM の検証実験

提案手法である複素 RBM の有効性を確認するため，まず初めに，人工的に生成した N=2000 個の 1 次元複素数データを用いた実験を行った．この人工データは，図２（黒点）に示すように，実部と虚部に相関を持つように生成された．本実験では，複素 RBM による人工データのモデリングと，実部・虚部を表す 2 次元の可視素子を持つ RBM によるモデリングの結果を比較した．両モデルとも隠れ素子数 2 とし，学習率 0.01，モーメント係数 0.1，バッチサイズ 20，繰り返し回数 200 の確率的勾配降下法によって学習を行った．その後，両方のモデルについて，学習データを入力し，隠れ素子の期待値を計算（エンコード），この隠れ素子から逆に可視素子の確率を計算（デコード）し，乱数を生成させた．複素 RBM 及び RBM によって生成された乱数をそれぞれ図２左図（赤点）及び右図（赤点）に示す．図２より，複素 RBM の方が RBM よりも，元の複素人工データ分布をよく表現できていることが分かる．これは複素 RBM では実部と虚部の相関を表現する構造を持っているのに対し，RBM では相関を表現することができないからだと考えられる．

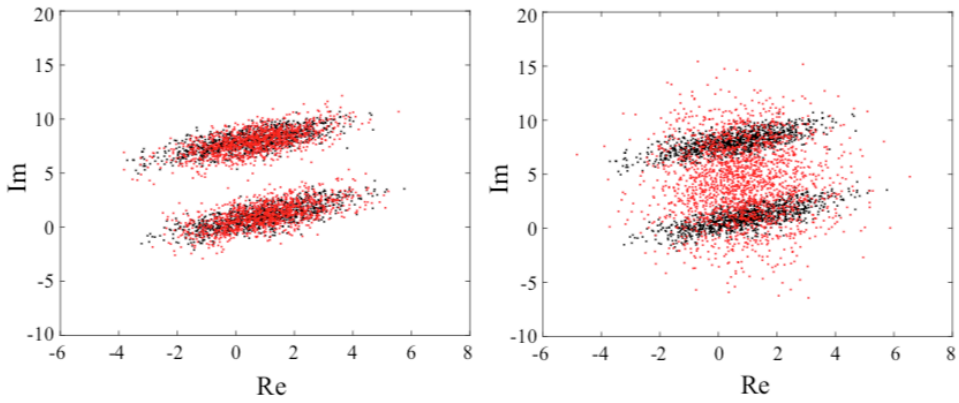

図 2 人工的に生成した 1 次元複素数データ(黒点)と，学習された複素 RBM(左図)及び RBM(右図)からランダムに生成されたデータ(赤点)を示す.

5-2 音声データを用いた複素 RBM の評価実験

次に，音声データを用いて提案法である複素 RBM の評価実験を行なった．本実験では ATR 音声データベースセット A 女性アナウンサー（FTK）50 文音声（約 4.2 分，サンプリングレート 20kHz を 16kHz にダウンサンプリング）をモデルの学習に，同じ話者の別の 53 文を評価に用いた．それぞれの音声データについて，窓幅 256，64 サンプルオーバーラップの短時間フーリエ変換を施した複素スペクトル(129 次元)を，複素 PCA で 40 次元に圧縮した特徴量を複素 RBM の可視素子とした．なお，複素 PCA で 40 次元に圧縮したデータから複素スペクトルを復元し，得られた音声の PESQ（Perceptual evaluation of speech quality）値が 4.46 であり，複素 PCA による品質劣化が少ないことから，本実験では次元数削減のため，可視素子として複素スペクトルではなく複素 PCA を用いた．まず，学習率 0.01，モーメント係数 0.1，バッチサイズ 100，繰り返し回数 200 の複素確率的勾配降下法と，学習率 0.001，減衰率 $\beta 1 = 0.9$，$\beta 2 = 0.999$，バッチサイズ 100，繰り返し回数 200 の複素 Adam の 2 つのモデル（隠れ素子数は 1,000）を学習させた．また比較手法として，同じ複素数データの実部と虚部を連結したベクトルを可視素子とした RBM（隠れ素子数は同様に 1000）を，同様の条件の確率的勾配降下法と Adam を用いて学習させた．それぞれの手法について，学習時の更新ステップ（Epoch）毎の再構築エラー（MSE）をプロットしたものを図 3 に示す．まず，複素確率的勾配降下法を用いた複素 RBM（CARBM+CSGD）と確率的勾配降下法を用いた RBM（RBM+SGD）を比較すると，前者の方が収束が早く，収束時におけるエラーが低いことが分かる．また，複素 Adam を用いた複素 RBM（CARBM+CAdam）と Adam を用いた RBM（RBM+Adam）を比較すると，わずかながら複素 RBM の方が収束時の再構築エラーが低くなっている．複素 Adam 及び Adam を用いた方が高精度であったことから，以降の実験ではこれらを用いる．

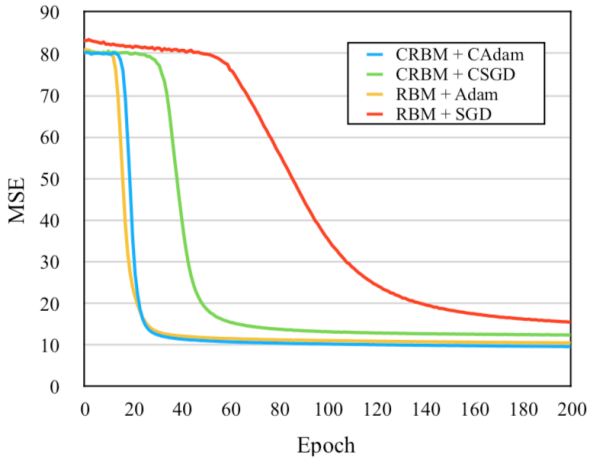

図 3 各繰り返しステップにおける複素 RBM(複素確率的勾配法を用いた場合と複素 Adam を用いた場合)と RBM(確率的勾配降下法を用いた場合と Adam を用いた場合)の再構築エラー.

最後に，隠れ素子数を 500，1,000，2,000，4,000 と変化させて複素 RBM 及び RBM を学習し，それぞれのモデルで評価用音声を復元した音声について，オリジナル音声を参照した PESQ による客観的比較を行った．図４に示すように，どの隠れ素子数においても，複素 RBM の方が PESQ 値が高いことが分かる．また，複素 RBM と RBM のいずれにおいても隠れ素子数を増やすと PESQ 値が向上しているが，RBM では隠れ素子数が 2,000 の時に飽和している．一方，複素 RBM では 4,000 まで隠れ素子数を増加させても PESQ 値の向上が見られる．

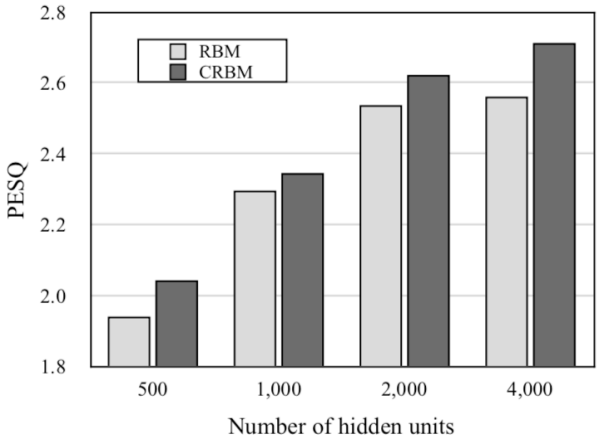

図 4 隠れ素子数を変えた場合の，複素 RBM 及び RBM によって復元された音声の PESQ による品質評価.

また，オリジナルの振幅スペクトルと，隠れ素子数が 1,000 の複素 RBM によって復元されたスペクトルを図５に示す．図５より，復元されたスペクトルはオリジナルのスペクトルに近く，フォルマントや基本周波数など，音声の特徴をよく捉えられていることが分かる．

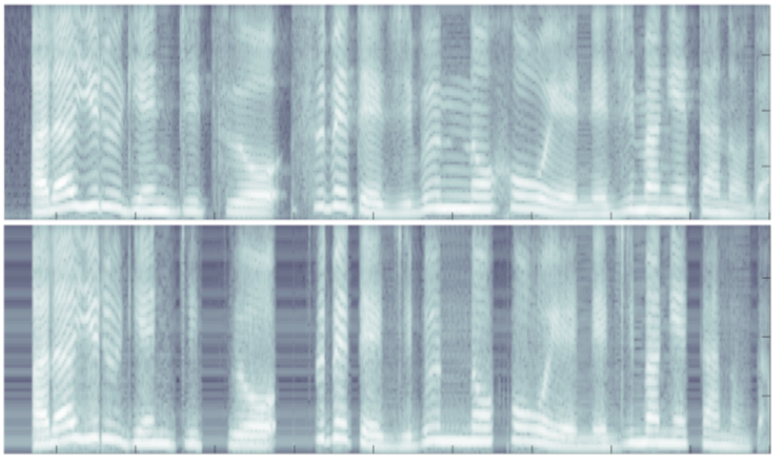

図 5 オリジナルの振幅スペクトル(上)と複素 RBM によって復元されたスペクトル(下)．縦軸は周波数，横軸は時間を表す.

5-3 音声データを用いた複素 TRBM の評価実験

前述の実験同様，ATR 音声データベース・セット A から女性アナウンサー（FTK）の音声を用いて複素 TRBM の品質評価実験を行なった．同セットから 50 文音声（約 4.2 分，サンプリングレート 20kHz を 16kHz にダウンサンプリング）を使用し，窓幅 256，64 サンプルオーバーラップの短時間フーリエ変換を施して得られた複素スペクトル(129 次元)を T=1,000 フレーム連続して並べたものを複素特徴量系列とし,隠れ素子数 J=200 の複素 TRBM の学習を行った.本実験では複素スペクトルの時接続を考慮しているため複素 PCA による次元圧縮や複素 MLPG によるパラメータ生成を使用しない．なお，複素 TRBM は複素 Adam に基づく確率的勾配法によりパラメータを推定した．また，安定して学習を行うため，最初の 50 エポックは U=0 とし（複素 RBM とみな

し）て学習し，その後の 100 エポックで全パラメータを更新した．比較手法として，同じ複素入力特徴量の，実部と虚部を連結したベクトルを可視素子として同様の条件（ただし，複素 Adam ではなく通常の Adam を用いている）で学習させた TRBM，系列でなく各フレーム特徴量とみなし，バッチサイズ 1,000 で学習させた複素 RBM を用いた．モデルの評価として，学習とは異なる別の 53 文の，同じ話者の音声データから計算される複素スペクトルを各モデルでエンコード（隠れ素子を計算）およびデコード（隠れ素子から可視素子を計算）し，復元された音声に対して算出される PESQ（Perceptual evaluation of speech quality）を算出した．

表 1 各手法による PESQ 値．

Method	TRBM	CRBM	CTRBM
PESQ	1.699	1.792	**1.794**

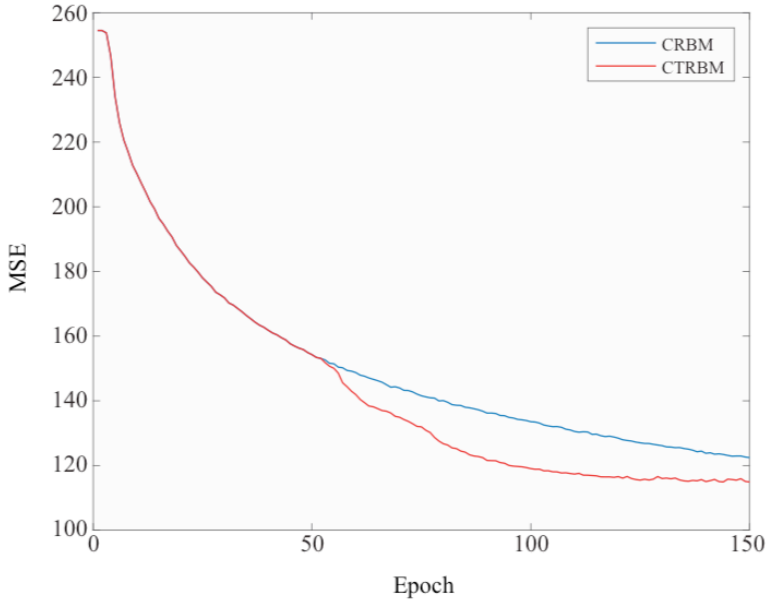

図 6 学習時における MSE の比較．

各手法による復元音声の品質結果を表 1 に示す．表 1 より，実数値表現（TRBM）よりも複素数表現（複素 RBM，複素 TRBM）の方が精度が高く，さらに提案手法が僅かに従来の複素 RBM よりも良い精度を示している．また，学習中の MSE（図 6）から，50 エポック目から開始される複素 TRBM の学習の方が，従来の複素 RBM よりも収束が早く，効率的であることが分かる．

最後に，提案手法の複素 TRBM により推定されたパラメータ W(絶対値を取ったもの)，U をそれぞれ図 7 (b)，図 8 に示す．比較のため，複素 RBM で推定されたパラメータ W（絶対値を取ったもの）を図 7 (a)に示す．図 7 より，W の各列ベクトルがそれぞれの隠れ素子に対応する複素スペクトル基底だとみたときに，複素 TRBM で得られる基底の方が，複素 RBM の基底よりも互いに異なるものになっている．このことから複素 TRBM は隠れ素子のアクティベーションをよりスパースに導くことができると推察される．また，図 8 より，推定された U の中で僅かに対角要素の値が他の要素の値よりも大きいことから，同じ隠れ素子を時間的に継続させる働きを持つことが分かる．

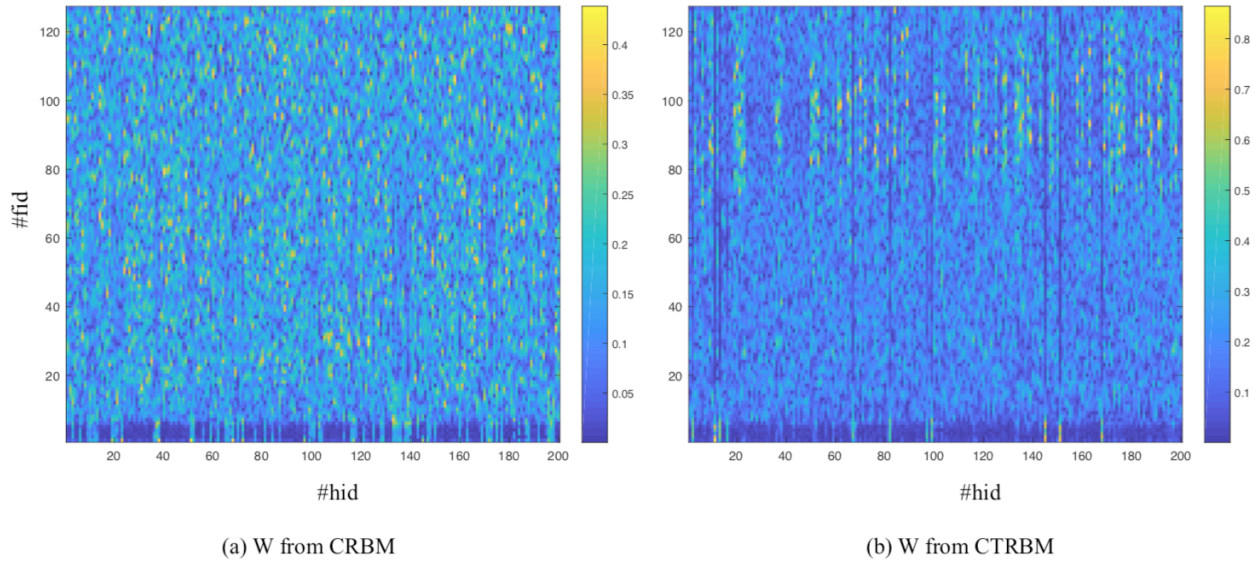

図 7 推定された可視素子・隠れ素子間の結合重みパラメータ W.

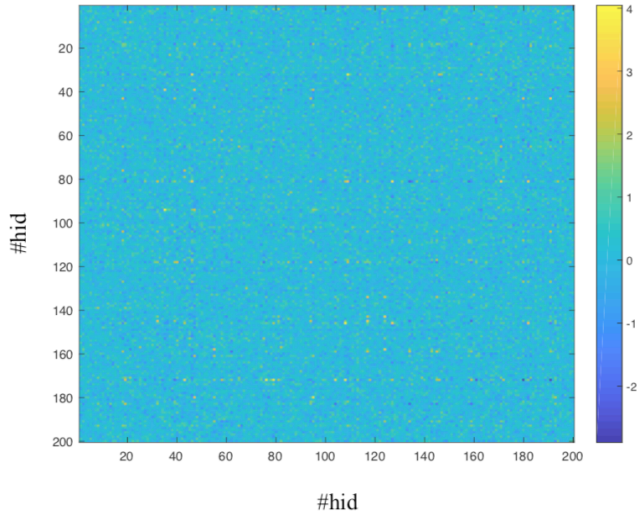

図 8 推定された，隣接する隠れ素子間の結合重みパラメータ U.

6 おわりに

本研究では，複素数データを可視素子とする RBM の拡張モデル（複素 RBM）を提案し，複素数の人工データ及び音声データを用いて性能評価を行なった．特に後者の実験では，従来の RBM によって復元された音声よりも，複素 RBM によって復元された音声の方が PESQ 値が高く，品質向上が見られた．また，複素 RBM の学習では，複素 Adam によってパラメータを推定する方が，勾配降下法よりも効果的であることが確認できた．さらに本研究では時間的に隣接する隠れ素子間の接続を考慮した複素 RBM の拡張モデル（複素 TRBM）についても検証し，客観評価によりその有効性を示した．

複素 RBM は入力データが複素数であれば全ての信号処理分野において用いることができ,応用先は幅広く，信号処理諸分野において革命をもたらす可能性がある．今後は，このような他分野への応用を検討していきたい．

【参考文献】

[1] LeCun, Y., Bengio, Y. and Hinton, G., “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[2] 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.

[3] Krizhevsky, A. and Hinton, G. E., “Factored 3-way restricted Boltzmann machines for modeling natural images,” Journal of Machine Learning Research, 2010.

[4] Salakhutdinov, R. and Hinton, G. E., “Deep Boltzmann Machines,” AISTATS, 2009.

[5] Sohn, K., Zhou, G., Lee, C. and Lee, H., “Learning and selecting features jointly with point-wise gated Boltzmann machines,” ICML, vol. 2, 2013.

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

[7] Lee, H., Ekanadham, C. and Ng, A. Y., “Sparse deep belief net model for visual area V2,” pp. 873–880, 2008.

[8] Kingma, D. and Ba, J., “Adam: A method for stochastic optimization,” ICLR, 2015.

[9] J. Horel, “Complex principal component analysis: Theory and examples,” Journal of climate and Applied Meteorology, vol. 23, no. 12, pp. 1660-1673, 1984.

[10] K. Tokuda et al., “Speech parameter generation algorithms for HMM-based speech synthesis,” ICASSP, pp. 1315-1318, 2000.

[11] Sutskever, I., Hinton, G. E. and Taylor, G. W., “The recurrent temporal restricted Boltzmann machine,” Advances in Neural Information Processing Systems, pp.1601–1608, 2009.

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
Complex-Valued Restricted Boltzmann Machine for Direct Speech Parameterization from Complex Spectra	arXiv	2018 年 3 月
リカレント構造を持つ複素制限ボルツマンマシンによる複素スペクトル系列モデリング	第 120 回音声言語情報処理研究会	2018 年 2 月
複素 RBM を用いた音声スペクトルモデリングの改良と評価	日本音響学会 2017 年秋季研究発表会	2017 年 9 月
Complex-valued restricted Boltzmann machine for direct learning of frequency spectra	Interspeech 2017	2017 年 8 月
複素 RBM : 制限ボルツマンマシンの複素数拡張と音声信号への応用と評価	第 117 回音声言語情報処理研究会	2017 年 7 月