Vol.2017-MUS-115 No.40
2017/6/18

テキスト・音声間の双方向変換に基づく DNN音声認識・合成のための事前学習法

曾根 健太郎[1,a)]　中鹿 亘[1,b)]　南 泰浩[1]

概要：統計的パラメトリック音声合成への従来のアプローチは，テキスト情報から音声パラメータを生成するために，決定木を用いてクラスタリングされた文脈依存隠れマルコフモデル（Hidden Markov Model; HMM）を用いる．しかし，決定木は，複雑なコンテキストの依存関係を効率的にモデル化できないことが知られている．その問題を解決するため，深層ニューラルネットワーク（Deep Neural Network; DNN）を用いて決定木を代替する手法がある．この手法により，テキスト情報から音声パラメータへの依存関係を効率的に表現することができるが，この手法では音声合成しか行うことができない．音声認識を行いたい場合は，音声認識器をまた別に用意して学習する必要があり，その場合学習コストが余分にかかってしまう．そこで，本研究では，学習コストの削減を目的とし，生成モデルである Deep Relational Model（DRM）を用いてテキスト・音声間の双方向の関係性を表現することで，DNN 音声認識器と DNN 音声合成器を同時に事前学習することができる手法を提案する．評価実験では，音声認識と音声合成の 2 つのタスクにおいて，提案手法により事前学習された DNN が，パラメータの初期値をランダムに与えた DNN よりも優れていることが示された．

キーワード：音声認識，音声合成，生成モデル，ボルツマンマシン，事前学習

Pre-training Method for DNN-based Speech Recognition and Synthesis Based on Bidirectional Conversion between Text and Speech

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

Abstract: Conventional approaches to statistical parametric speech synthesis use context-dependent hidden Markov models (HMMs) clustered using decision trees to generate speech parameters from linguistic features. However, decision trees are not always appropriate to model complex context dependencies efficiently. Although an alternative scheme based on a deep neural network (DNN) was presented as a possible way to overcome the difficulty, this approach has a restriction that it is applied for only speech synthesis; for example, this approach cannot be applied for speech recognition. Thus, systems for speech recognition requires training cost separately. This paper proposes a pre-training method for DNN based speech recognition and synthesis by capturing deep relationships between text and speech using deep relational model (DRM) to reduce training costs. Experimental results show that pre-trained DNN-based systems using the proposed method outperformed randomly initialized DNN-based systems.

Keywords: Speech Recognition, Speech Synthesis, Generative Model, Boltzmann Distribution, Pre-training Method

1. はじめに

これまで，音声認識では隠れマルコフモデル（Hidden Markov Model; HMM）を用いた統計的手法 [1] が研究さ

[1] 電気通信大学大学院情報理工学研究科
情報・ネットワーク工学専攻
UEC, Chofu, Tokyo 182–8585, Japan
[a)] sone@is.sd.uec.ac.jp
[b)] nakashika@uec.ac.jp

れている．HMM は入力となる音声信号の特徴量と，自身の内部状態との関係を Gaussian Mixture Model（GMM）を用いて表現し，学習を行う．また，GMM の代わりに深層ニューラルネットワーク（Deep Neural Network; DNN）を用いて，音声の特徴量から HMM の内部状態を推定する方法が提案されている [2]．この手法はハイブリッド方式と呼ばれ，従来の GMM を用いた HMM 音声認識よりも高い性能を示すことが知られている．

一方，音声合成では HMM を用いた統計的パラメトリック音声合成 [3] が盛んに研究されてきた．この手法は音声のピッチや長さを同時にモデル化できる点 [4] や，話者適応が可能 [5] という柔軟性から，波形接続型の音声合成 [6] よりも優れた点をもつ．しかし，HMM 音声合成では合成音声の質が問題となる [7]．その理由のひとつとして，決定木を用いてコンテキストクラスタリングを行う文脈依存 HMM では，音素や言語特徴量とそれらに対応する音声パラメータの依存関係を効率的にモデル化できないことがあげられる．そこで，Zen ら [8] は決定木を DNN で代替することで，合成音声の質を向上させる手法を提案した．

しかし，これらの音声認識・音声合成手法はいずれも音声認識器・音声合成器をそれぞれ独立に学習しており，両方のタスクに適用することができない．したがって，認識器を構築したあとに音声合成を行う場合，別途合成器を学習する必要があり，学習コストがかかってしまう．そこで，本研究では，2 変数間の双方向変換を可能にする Deep Relational Model（DRM）[9] を応用してテキスト・音声間の双方向の関係性を表現することで，DNN 音声認識器および DNN 音声合成器を同時に事前学習する手法を提案する．

2. DNN 音声合成

この章では，Zen らの DNN を用いた音声合成手法について説明する．音声合成を行う DNN は，言語特徴量に対応する入力層と音響特徴量に対応する出力層で，いくつかの隠れ層を挟んだ構造をもつ．入力テキストは形態素解析器にかけられ言語特徴量 $\boldsymbol{w}^t$ へと変換される．ここで，$\boldsymbol{w}^t$ はフレーム t における言語特徴ベクトルを表し，one-hot 表現により音素を表すバイナリ値とコンテキストを表す連続値（当該音素における当該フレームの相対的な位置，当該音素の継続長など）からなる．

入力された言語特徴量は，DNN によって出力となる音響特徴量 $\boldsymbol{a}^t$ へ変換される．ここで，$\boldsymbol{a}^t$ はフレーム t における音響特徴ベクトルを表す．音響特徴量は音声のスペクトルとその導関数（動的特徴量）[10] からなる．DNN の重みは，学習データの入力特徴量とそれに対応する出力特徴量を用いて学習される．HMM 音声合成同様，DNN から出力された音響特徴量を平均，全学習データから計算した分散を共分散行列として，Parameter generation algorithm[11] を用いて音声パラメータを生成する．そして最後に，生成した音声パラメータから音声波形を合成する．

本研究では，フレーム独立のモデリングに着目している．したがって，特に記載がない限り，以降添え字 t を省略する．

3. Deep Relational Model

この章では，中鹿らの Deep Relational Model（DRM）について説明する．DRM は Restricted Boltzmann Machine（RBM）[12] や Deep Boltzmann Machine（DBM）[13] 同様，エネルギー関数に基づくモデルである．したがって，DRM はある可視変数 $\boldsymbol{x} \in \{0,1\}^I$ と別の可視変数 $\boldsymbol{y} \in \{0,1\}^K$，そして隠れ変数 $\boldsymbol{h}^{(l)} \in \{0,1\}^{J_l} (l = 1, ..., L)$ より与えられる結合確率分布で定義される．ここで，L は隠れ層の層数を表す．DRM は RBM や DBM 同様，各ユニットは隣接する層のユニットのみと結合をもち，同じ層のユニットとは結合をもたない．DRM の結合確率分布は以下のように表される．

$$p(\boldsymbol{x}, \boldsymbol{y}; \theta) = \sum_{\forall \boldsymbol{h}^{(l)}} p(\boldsymbol{x}, \boldsymbol{y}, \forall \boldsymbol{h}^{(l)}; \theta) \tag{1}$$

$$p(\boldsymbol{x}, \boldsymbol{y}, \forall \boldsymbol{h}^{(l)}; \theta) = \frac{1}{Z(\theta)} \exp\{-E(\boldsymbol{x}, \boldsymbol{y}, \forall \boldsymbol{h}^{(l)}; \theta)\} \tag{2}$$

ここで，E はエネルギー関数であり

$$E\ (\boldsymbol{x}, \boldsymbol{y}, \forall \boldsymbol{h}^{(l)}; \theta) = -\boldsymbol{b}^T \boldsymbol{x} - \sum_{l=1}^{L} \boldsymbol{c^{(l)}}^T \boldsymbol{h}^{(l)} - \boldsymbol{d}^T \boldsymbol{y}$$

$$- \boldsymbol{x}^T \boldsymbol{W}^{(1)} \boldsymbol{h}^{(1)} - \sum_{l=2}^{L} \boldsymbol{h}^{(l-1)T} \boldsymbol{W}^{(l)} \boldsymbol{h}^{(l)} - \boldsymbol{h}^{(L)T} \boldsymbol{W}^{(L+1)} \boldsymbol{y}$$

で定義される．ただし，$\boldsymbol{b} \in \mathbb{R}^I$，$\boldsymbol{c}^{(l)} \in \mathbb{R}^{J_l}$，$\boldsymbol{d} \in \mathbb{R}^K$ はそれぞれ 1 つめの可視層，隠れ層 l，2 つめの可視層のバイアスを，$\boldsymbol{W}^{(1)} \in \mathbb{R}^{I \times J_1}$，$\boldsymbol{W}^{(l)} \in \mathbb{R}^{J_{l-1} \times J_l}$，$\boldsymbol{W}^{(L+1)} \in \mathbb{R}^{J_L \times K}$ はそれぞれ 1 つめの可視層と隣接する隠れ層，隠れ層 $l-1$ と隠れ層 l，隠れ層 L と 2 つめの可視層間の重みを表し，学習により推定されるパラメータである．

(1)，(2) の定義より，各層の条件付き確率分布は以下で与えられる．

$$p\ (x_i = 1 | \boldsymbol{h}^{(1)}) = \sigma(b_i + \boldsymbol{W}_{i:}^{(1)} \boldsymbol{h}^{(1)}) \tag{3}$$

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

$$p\ (y_k = 1 | \boldsymbol{h}^{(L)}) = \sigma(d_k + \boldsymbol{W}_{:k}^{(L)T} \boldsymbol{h}^{(L)}) \tag{5}$$

ただし，$\sigma(\cdot)$ はシグモイド関数を表す．また，式 (4) において $\boldsymbol{h}^{(0)} = \boldsymbol{x}, \boldsymbol{h}^{(L+1)} = \boldsymbol{y}$ としている．

DRM のパラメータ $\theta = \{\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{d}\}$ は，$p(\boldsymbol{x}, \boldsymbol{y}; \theta)$ の対数尤度 $\mathcal{L} = \log \prod_t p(\boldsymbol{x}^t, \boldsymbol{y}^t; \theta)$ が最大となるように推定される．対数尤度のパラメータ θ に関する偏微分は

$$\frac{\partial \mathcal{L}}{\partial \theta} = \langle -\frac{\partial E}{\partial \theta} \rangle_d - \langle -\frac{\partial E}{\partial \theta} \rangle_m \tag{6}$$

と計算される．ここで，$\langle\cdot\rangle_d, \langle\cdot\rangle_m$ はそれぞれデータの期待値，モデルの期待値を表す．前者は観測データの平均を計算することで得られるが，後者の計算は組合せ爆発の問題が生じる．そこで，DRM の学習では，平均場近似を用いてモデルの期待値を計算する．すなわち，観測データ $(\boldsymbol{x}, \boldsymbol{y})$ を用いて式 (4) により隠れ層の値 $\boldsymbol{h}^{(l)}$ を T 回更新することで $\langle \boldsymbol{h}^{(l)} \rangle_m$ を得る．また，得られた $\langle \boldsymbol{h}^{(l)} \rangle_m$ を用いて式 (3), (5) により $\langle \boldsymbol{x} \rangle_m, \langle \boldsymbol{y} \rangle_m$ を得る．

また，DRM の学習によりパラメータを推定する前に，RBM を用いて事前学習を行う．DRM の事前学習では Deep Belief Network（DBN）[12] 同様，2 層ずつ，逐次貪欲的に学習していく．まず，最上位および最下位のパラメータを可視変数 $\boldsymbol{x}, \boldsymbol{y}$ を用いて学習する．次に，先ほど得られた RBM についてそれぞれ $\boldsymbol{x}, \boldsymbol{y}$ が与えられたときの隠れ変数のサンプル値を可視変数とみなして RBM を構築し，パラメータを逐次学習していく．このように，RBM を用いて外側の層から内側の層へと事前学習を行う．

4. DRM の音声認識・音声合成への応用

この章では，3 章で説明した DRM を連続値へと拡張した Gaussian-Categorical DRM（GCDRM）を定義し，音声認識・音声合成へ応用する手法について述べる．

4.1 Gaussian-Categorical DRM の定義

2 章で述べたとおり，テキストを解析して得られる言語特徴量および音声パラメータを生成するために必要な音響特徴量はバイナリ値と連続値からなる．したがって，提案モデルでは，従来の DRM のように Bernoulli 分布だけでなく，音素を表すカテゴリカル分布と，テキストのコンテキストおよび音声パラメータを表す正規分布を同時に扱う．エネルギー関数に基づくモデルを用いて連続値を扱う手法として Gaussian-Bernoulli RBM （GBRBM）[14] が提案されている．しかし，この手法では分散項の影響で学習が不安定になるため，Improved GBRBM[15] が提案された．また，音声合成のための DNN を事前学習する手法として，Mixed GBRBM および Mixed Categorical-Bernoulli RBM（Mixed CBRBM）[16] が提案された．本研究では Mixed GBRBM，Mixed CBRBM および Improved GBRBM のエネルギー関数を参考に，GCDRM のエネルギー関数を次のように定義する．

$$
\begin{aligned}
E(\boldsymbol{x}^c,\ \boldsymbol{x}^g\ , \boldsymbol{y}^c, \boldsymbol{y}^g, \forall \boldsymbol{h}^{(l)}; \theta) = & -\boldsymbol{b}^{cT}\boldsymbol{x}^c - \boldsymbol{x}^{cT}\boldsymbol{W}^{(1)c}\boldsymbol{h}^{(1)} \\
& - \frac{(\boldsymbol{x}^g - \boldsymbol{b}^g)^T(\boldsymbol{x}^g - \boldsymbol{b}^g)}{2\boldsymbol{\sigma}^{(x)g2}} - \Big(\frac{\boldsymbol{x}^g}{\boldsymbol{\sigma}^{(x)g2}}\Big)^T \boldsymbol{W}^{(1)g}\boldsymbol{h}^{(1)} \\
& - \sum_{l=1}^{L} \boldsymbol{c}^{(\boldsymbol{l})^T}\boldsymbol{h}^{(l)} - \sum_{l=2}^{L} \boldsymbol{h}^{(l-1)T}\boldsymbol{W}^{(l)}\boldsymbol{h}^{(l)} \\
& - \boldsymbol{d}^{cT}\boldsymbol{y}^c - \boldsymbol{h}^{(L)T}\boldsymbol{W}^{(L+1)c}\boldsymbol{y}^c \\
& - \frac{(\boldsymbol{y}^g - \boldsymbol{d}^g)^T(\boldsymbol{y}^g - \boldsymbol{d}^g)}{2\boldsymbol{\sigma}^{(y)g2}} - \boldsymbol{h}^{(L)T}\boldsymbol{W}^{(L+1)g}\frac{\boldsymbol{y}^g}{\boldsymbol{\sigma}^{(y)g2}}
\end{aligned}
$$

ここで，$\boldsymbol{x}^c \in \{0,1\}^{X^c}$, $\boldsymbol{x}^g \in \mathbb{R}^{X^g}$ はそれぞれ可視変数 $\boldsymbol{x}$ のうち，カテゴリカル分布に従うユニット，正規分布に従うユニットを表し，$\boldsymbol{y}^c \in \{0,1\}^{Y^c}$, $\boldsymbol{y}^g \in \mathbb{R}^{Y^g}$ もそれぞれ可視変数 $\boldsymbol{y}$ のうち，同様のユニットを表す（$X^g + X^c = I$, $Y^g + Y^c = K$, $\boldsymbol{x} = [\boldsymbol{x}^{gT}\ \boldsymbol{x}^{cT}]^T$, $\boldsymbol{y} = [\boldsymbol{y}^{gT}\ \boldsymbol{y}^{cT}]^T$）．そして，$\boldsymbol{W}^{(l)c} \in \mathbb{R}^{X^c \times J_1}, \boldsymbol{W}^{(L+1)c} \in \mathbb{R}^{J_L \times Y^c}, \boldsymbol{b}^c \in \mathbb{R}^{X^c}, \boldsymbol{d}^c \in \mathbb{R}^{Y^c}$ はそれぞれ可視変数のうちカテゴリカル分布に従うユニットに対応するパラメータを，そして，$\boldsymbol{W}^{(1)g} \in \mathbb{R}^{X^g \times J_1}, \boldsymbol{W}^{(L+1)g} \in \mathbb{R}^{J_L \times Y^g}, \boldsymbol{b}^g \in \mathbb{R}^{X^g}, \boldsymbol{d}^g \in \mathbb{R}^{Y^g}$ はそれぞれ可視変数のうち正規分布に従うユニットに対応するパラメータを，$\boldsymbol{\sigma}^{(x)g} \in \mathbb{R}^{X^g}, \boldsymbol{\sigma}^{(y)g} \in \mathbb{R}^{Y^g}$ はそれぞれ可視変数 $\boldsymbol{x}^g, \boldsymbol{y}^g$ の偏差を表し，いずれも推定すべきパラメータである．また，式中の除算は要素ごとの除算を表す．

エネルギー関数を新たに定義したことにより，可視層の条件付き確率分布は次のようになる．

$$
p(x_i^c = 1|\boldsymbol{h}^{(1)}) = \frac{\exp(b_i^c + \boldsymbol{W}_{i:}^{(1)c}\boldsymbol{h}^{(1)})}{\sum_{i'} \exp(b_{i'}^c + \boldsymbol{W}_{i':}^{(1)c}\boldsymbol{h}^{(1)})} \tag{7}
$$

$$
p(x_i^g = x|\boldsymbol{h}^{(1)}) = \mathcal{N}\Big(x|b_i^g + \boldsymbol{W}_{i:}^{(1)g}\boldsymbol{h}^{(1)}, \sigma_i^{(x)2}\Big) \tag{8}
$$

$$
p(y_k^c = 1|\boldsymbol{h}^{(L)}) = \frac{\exp(d_k^c + \boldsymbol{W}_{:k}^{(L+1)cT}\boldsymbol{h}^{(L)})}{\sum_{k'} \exp(d_{d'}^c + \boldsymbol{W}_{:k'}^{(L+1)cT}\boldsymbol{h}^{(L)})} \tag{9}
$$

$$
p(y_k^g = y|\boldsymbol{h}^{(L)}) = \mathcal{N}\Big(y|d_k^g + \boldsymbol{W}_{:k}^{(L+1)gT}\boldsymbol{h}^{(L)}, \sigma_i^{(y)2}\Big) \tag{10}
$$

ここで，$\mathcal{N}(\cdot|\mu, \sigma^2)$ は平均 μ，分散 σ^2 の正規分布を表す．また，1 層目，L 層目の隠れ層について，条件付き確率分布はそれぞれ

$$
\begin{aligned}
p(\ \boldsymbol{h}_j^{(1)}\ = 1|\boldsymbol{x}, \boldsymbol{h}^{(2)}) = & \\
& \sigma(c_j^{(1)} + \boldsymbol{W}_{:j}^{(1)T}\frac{\boldsymbol{x}}{\boldsymbol{\sigma}^{(x)2}} + \boldsymbol{W}_{j:}^{(2)}\boldsymbol{h}^{(2)})
\end{aligned} \tag{11}
$$

$$
\begin{aligned}
p(\ \boldsymbol{h}_j^{(L)}\ = 1|\boldsymbol{y}, \boldsymbol{h}^{(L-1)}) = & \\
& \sigma(c_j^{(L)} + \boldsymbol{W}_{:j}^{(L)T}\boldsymbol{h}^{(L-1)} + \boldsymbol{W}_{j:}^{(L+1)T}\frac{\boldsymbol{y}}{\boldsymbol{\sigma}^{(y)2}})
\end{aligned} \tag{12}
$$

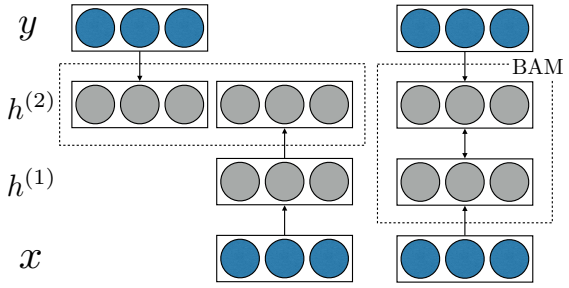
となる．ただし，簡単のため重み $\boldsymbol{W}^{(1)}$, $\boldsymbol{W}^{(L+1)}$ について $\boldsymbol{W}^{(1)} = [\boldsymbol{W}^{(1)gT}\ \boldsymbol{W}^{(1)cT}]^T$, $\boldsymbol{W}^{(L+1)} = [\boldsymbol{W}^{(L+1)gT}\ \boldsymbol{W}^{(L+1)cT}]^T$ としている．また，偏差 $\boldsymbol{\sigma}^{(x)}, \boldsymbol{\sigma}^{(y)}$ について，

$$
\sigma_i^{(x)} = \begin{cases} \sigma_i^{(x)g} & (1 \le i \le X^g) \\ 1 & (X^g < i \le I) \end{cases} \tag{13}
$$

$$
\sigma_k^{(y)} = \begin{cases} \sigma_k^{(y)g} & (1 \le k \le Y^g) \\ 1 & (Y^g < k \le K) \end{cases} \tag{14}
$$

としている．$2, \ldots, L-1$ 層目の隠れ層の条件付き確率分布は式 (4) と同様である．

$\boldsymbol{b} = [\boldsymbol{b}^{gT}\ \boldsymbol{b}^{cT}]^T$, $\boldsymbol{d} = [\boldsymbol{d}^{gT}\ \boldsymbol{d}^{cT}]^T$ とすれば，GCDRM のパラメータ $\theta = \{\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{d}, \boldsymbol{\sigma}^{(x)g}, \boldsymbol{\sigma}^{(y)g}\}$ は，従来の DRM と同様に対数尤度 $\mathcal{L}$ が最大となるように推定される．それ



(a) RBM (b) BAM

図 1 (a) RBM のみ，(b) BAM を用いた事前学習
Fig. 1 Pre-training methods of using (a) only RBMs, and (b) RBMs and BAM.

ぞれのパラメータに関する勾配は

$$\frac{\partial \mathcal{L}}{\partial b_i} = \langle \frac{1}{\sigma_i^{(x)2}} x_i \rangle_d - \langle \frac{1}{\sigma_i^{(x)2}} x_i \rangle_m \tag{15}$$

$$\frac{\partial \mathcal{L}}{\partial c_j^{(l)}} = \langle h_j^{(l)} \rangle_d - \langle h_j^{(l)} \rangle_m \tag{16}$$

$$\frac{\partial \mathcal{L}}{\partial d_k} = \langle \frac{1}{\sigma_k^{(y)2}} y_k \rangle_d - \langle \frac{1}{\sigma_k^{(y)2}} y_k \rangle_m \tag{17}$$

$$\frac{\partial \mathcal{L}}{\partial W_{ij}^{(l)}} = \tag{18}$$

$$\begin{cases} \langle \frac{1}{\sigma_i^{(x)2}} x_i h_j^{(1)} \rangle_d - \langle \frac{1}{\sigma_i^{(x)2}} x_i h_j^{(1)} \rangle_m & (l=1) \\ \langle h_i^{(l-1)} h_j^{(l)} \rangle_d - \langle h_i^{(l-1)} h_j^{(l)} \rangle_m & (l=2,\ldots,L) \\ \langle \frac{1}{\sigma_k^{(y)2}} h_i^{(L)} y_k \rangle_d - \langle \frac{1}{\sigma_k^{(y)2}} h_i^{(L)} y_k \rangle_m & (l=L+1) \end{cases}$$

となる．それぞれのパラメータは式 (15) から式 (18) を用いて繰り返し更新される．また，分散パラメータについては，常に非負値となるように，Improved GBRBM と同様，その対数 $z_i^{(x)} = \log \sigma_i^{(x)2}, z_k^{(y)} = \log \sigma_k^{(y)2}$ を更新することで推定する．

4.2 BAM を用いた事前学習

3 章で述べたとおり，従来の DRM の事前学習では，RBM を用いて可視層側から隠れ層側へ，すなわち外側から内側へと層ごとに行っていく．2 つの隠れ層間のパラメータを事前学習する際は，1 つ前のステップで得られた RBM からサンプルした隠れ変数を擬似的な可視変数とみなす．このとき，$\boldsymbol{x}$ 側からの RBM と $\boldsymbol{y}$ 側からの RBM が，それぞれ異なる隠れ変数を表すことになる（**図 1** a）．それを避けるため，2 つの可視変数間の重みを学習する手法である Bidirectional Associative Memory（BAM）[17] を用いて事前学習を行なう（図 1 b）．

BAM のエネルギー関数は次のように定義される．

$$E_{BAM}(\boldsymbol{x}, \boldsymbol{y}) = -\boldsymbol{b}^T \boldsymbol{x} - \boldsymbol{d}^T \boldsymbol{y} - \boldsymbol{x}^T \boldsymbol{W} \boldsymbol{y} \tag{19}$$

ここで，$\boldsymbol{W}$ は重み，$\boldsymbol{b}, \boldsymbol{d}$ はバイアス項である．Chen ら [18] は，BAM を確率密度関数として解釈し，RBM と同様，CD（Contrastive Divergence）法を用いて学習する手法を提案している．BAM の結合確率分布は

$$P(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{Z_{BAM}} \exp(-E_{BAM}(\boldsymbol{x}, \boldsymbol{y})) \tag{20}$$

となる．ただし $Z_{BAM} = \sum_{\boldsymbol{x},\boldsymbol{y}} \exp(-E_{BAM}(\boldsymbol{x}, \boldsymbol{y})$ は分配関数である．

4.3 GCDRM を用いた音声認識・音声合成

GCDRM を音声認識・音声合成へ応用する場合，言語特徴量 $\boldsymbol{w}$ を 1 つ目の可視変数 $\boldsymbol{x}$，音響特徴量 $\boldsymbol{a}$ を 2 つ目の可視変数 $\boldsymbol{y}$ とみなす．言語特徴量は one-hot 表現により音素を表すバイナリ値と，コンテキストを表す連続値からなる．一方，音響特徴量は連続値のみから構成されるため，式 (9) およびエネルギー関数中の $\boldsymbol{y}^c$ に関する項を無視することができる．

GCDRM の学習の後，得られたパラメータを初期値とする DNN を構築し，誤差伝播法を用いてパラメータの fine-tuning を行う．その際，音声認識器を構築したい場合は，入力を音響特徴量，出力を言語特徴量とする．また，音声合成器を構築したい場合は，入力を言語特徴量，出力を音響特徴量とする．認識器，合成器のどちらを構築する場合でも，DNN の初期値として同じパラメータを与える．

5. 実験

5.1 実験条件

本実験では，HTS ワーキンググループ [19] が公開している音声コーパスである NIT ATR503 M001 を用いて，提案手法を用いた音声認識・音声合成の精度を評価した．この音声コーパスは，単一の男性話者によって，音素バランスがとれた文章を読み上げた音声と，読み上げた文章を形態素解析したコンテキストラベルデータが含まれている．コーパスはセット A からセット J までの 10 セットで構成され，セット A からセット I まではそれぞれ 50 文，セット J は 53 文を含む．このコーパス内で使用されている音素の種類は 43 種類であり，コンテキストの種類は単語の活用形を表す ID やフレーズのアクセントタイプを表す ID などからなる 47 種類である．言語特徴量として先行，当該，後続の 3 音素を表す one-hot ベクトルおよびコンテキストを用いた．また，音響特徴量として，35 次元の MFCC とその一次微分，二次微分を用いた．

5.2 音声合成タスク

まず，提案手法の最適なパラメータ数を決定するため，層数および隠れ層のユニット数を変化させ，音声合成の精度を比較した．実験は音声コーパスからセット A，B の 100 文を用いて学習し，セット J の 53 文を用いてテストを行なった．言語特徴量を入力として推定された音響特徴

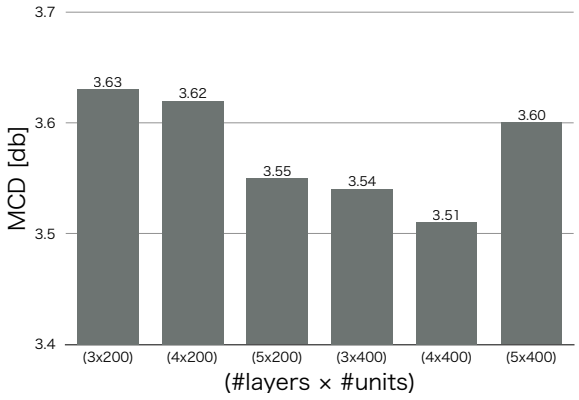


図 2 提案手法において層数，ユニット数を変化させたときの音声合成の精度（MCD [db]）

Fig. 2 Performance of our method when changing the numbers of hidden layers and hidden units at each hidden layer (MCD [db]).

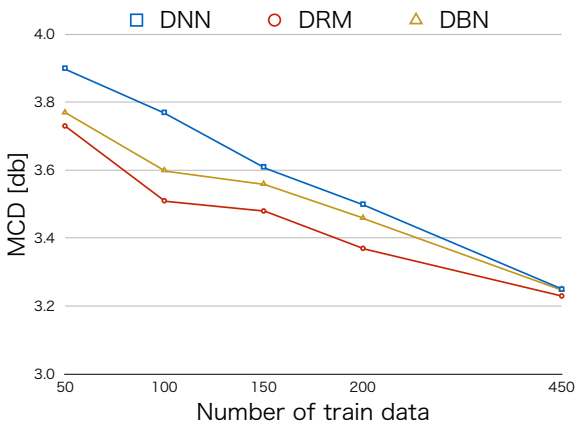


図 3 各手法による合成音声・自然音声間の平均 MCD [db] 比較

Fig. 3 Comparison of MCD [db] between the generated speech and the target speech obtained by each method.

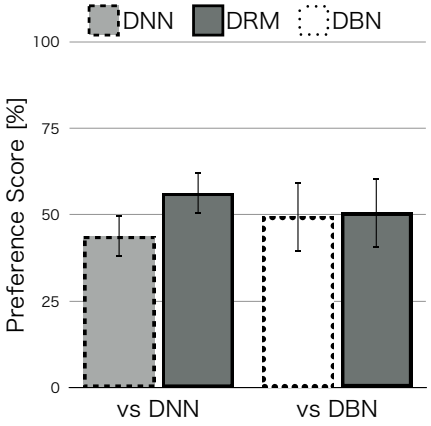


図 4 各手法を用いた主観評価結果 [%]

Fig. 4 Subjective preference scores [%] of speech samples obtained each method.

量を平均，全学習データから計算した分散を共分散行列として，Parameter generation algorithm を用いて MFCC を出力した．また，音声合成の精度を評価する指標として，ターゲットとなる音声と生成した音声の MFCC 間のユークリッド距離における近さを表すメルケプストラム歪み（MCD）[20] を使用し，セット J53 文でテストを行なった平均を比較した．その結果を**図 2** に示す．例えば (3×200) はユニット数が 200 の隠れ層が 3 層のモデルを示す．

図 2 から，(4×400) のモデルが，音声合成タスクにおいて最も MCD が低くなることがわかる．また，ユニット数が 200 のときは層数を増やすと精度が上がることがわかる．しかし，ユニット数が 400 のときは，層数を 5 まで増やすと精度が下がる．この理由として，推定すべきパラメータの増加することで，モデルがうまく学習できないためであると考えられる．

5.2.1 客観評価

次に，学習データ数を 50 文，100 文，150 文，200 文，450 文と変化させ，提案手法を DNN，DBN と比較した．テストには学習データに含まれない 53 文を使用した．それぞれの手法について，(4×400) のモデルを用いた．DBN および DRM は，学習したパラメータを DNN のパラメータの初期値として fine-tuning を行い，DNN のパラメータはランダムな初期値を与えて学習を行なった．実験の結果を**図 3** に示す．図 3 から，いずれの学習データ数においても，合成精度が向上していることがわかる．学習データ数が 450 文のときは各手法についてほとんど結果の差がない．これは，DNN の学習において，学習データを十分に与えればパラメータをうまく最適化することができるためであると考えられる．したがって，提案手法は学習データ数が少ないときに有用であるといえる．

5.2.2 主観評価

次に，提案手法の性能を評価するため，200 文で学習した DNN，DBN，提案手法による合成音声の自然性を，XAB テストにより評価した．テストに使用した音声は，学習データに含まれない 53 文の中からランダムに選んだ 20 文である．被験者は 20 代の学生 8 名である．DNN と提案手法，DBN と提案手法について，どちらの合成音声が自然音声に近いかという基準で評価した．いずれの手法も (4×400) のモデルを用いた．また，音声の F0 および状態継続長は，自然音声から抽出したものを使用し，MFCC のみ，各手法から出力されたものを使用し合成音声を生成した．

XAB テストによる合成音声の自然さに関する評価結果を**図 4** に示す．誤差範囲は 95%信頼区間を示す．図 4 の結果から，DNN と提案手法では，提案手法が有効であることがわかる．一方，DBN と提案手法では，わずかに提案手法のほうが良い結果が得られているが，両手法の自然性に 5%の有意水準で有意差は認められなかった．しかし，提案手法は DBN と違い，音声合成器と音声認識器を同時に事前学習することができる点において有用であるといえる．

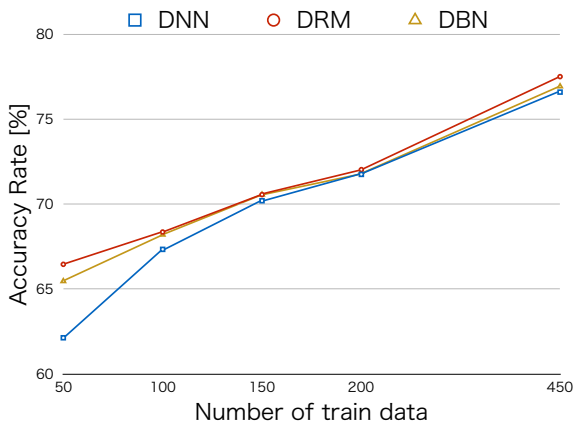


図 5 各手法による当該音素の正解率 [%]

Fig. 5 Accuracy rate [%] of the current phoneme obtained by each method.

5.3 音声認識タスク

次に，提案手法の音声認識精度を DNN，DBN と比較した．実験は，音響特徴量を入力とし，推定された言語特徴量の中から，当該音素の正解率を評価した．評価の方法として，学習データ数を 50 文，100 文，150 文，200 文，450 文と変化させ，提案手法を DNN，DBN と比較した．テストには学習データに含まれない 53 文を使用し，53 文の平均正解率で評価した．また，それぞれの手法について，(4×400) のモデルを用いた．5.2 節の実験で学習した DRM のパラメータを DNN の初期値とし，fine-tuning を行い，DBN および DNN は 5.2.1 節と同様の条件で学習した．実験の結果を**図 5** に示す．図 5 から，いずれの学習データ数においても，認識精度が向上していることがわかる．学習データ数が 150 文，200 文以降は結果に大きな差が見られない．これは 5.2 節と同様の理由によると考えられる．

6. おわりに

本稿では，DRM を連続値を扱えるように拡張することで，DNN 音声認識器および DNN 音声合成器を同時に事前学習する手法について提案した．音声合成実験において，DNN や DBN を用いた手法よりもターゲット音声に近い音声を生成することができた．また，音声認識実験において，DNN や DBN を用いた手法よりも当該音素に対する正解率が向上した．2 つの実験において，とくにデータ数が少ないときについて精度を向上することができた．

今後は認識および合成の精度向上を目的とし，時系列を考慮した手法について検討したい．

参考文献

[1] Juang, B. H. and Rabiner, L. R.: Hidden Markov models for speech recognition, *Technometrics*, Vol. 33, No. 3, pp. 251–272 (1991).
[2] Hinton, G., Deng, L., Yu, D., Dahl, G., rahman Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. and Kingsbury, B.: Deep neural networks for acoustic modeling in speech recognition, *Signal Processing Magazine*, Vol. 29, No. 6, pp. 82–97 (2012).
[3] Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J. and Oura, K.: Speech synthesis based on Hidden Markov models, *IEEE*, Vol. 101, No. 5, pp. 1234–1252 (2013).
[4] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T.: Simultaneous modeling of spectrum pitch and duration in hmm-based speech synthesis, *Eurospeech*, pp. 2347–2350 (1999).
[5] Tamura, M., Masuko, T., Tokuda, K. and Kobayashi, T.: Speaker adaptation for hmm-based speech synthesis system using MLLR, *ICASSP*, pp. 805–808 (2001).
[6] Hunt, A. J. and Black, A. W.: Unit selection in a concatenative speech synthesis system using a large speech database, *ICASSP*, pp. 373–376 (1996).
[7] Zen, H., Tokuda, K. and Black, A. W.: Statistical parametric speech synthesis, *Speech Commun*, Vol. 51, No. 11, pp. 1039–1064 (2009).
[8] Zen, H., Senior, A. and Schuster, M.: Statistical parametric speech synthesis using deep neural networks, *ICASSP*, pp. 7962–7966 (2013).
[9] Nakashika, T., Takiguchi, T. and Ariki, Y.: Modeling bidirectional relationships for image classification and generation, *ICASSP*, pp. 1327–1331 (2016).
[10] Masuko, T., Tokuda, K., Kobayashi, T. and Imai, S.: Speech synthesis from HMMs using dynamic features, *ICASSP*, pp. 389–392 (1996).
[11] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T. and Kitamura, T.: Speech parameter generation algorithms for HMM-based speech synthesis, *International Conference on Artificial Intelligence and Statistics*, pp. 1315–1318 (2009).
[12] 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).
[13] Salakhutdinov, R. and Hinton, G. E.: Deep boltzmann machines, *ICASSP*, pp. 448–455 (2013).
[14] Hinton, G. E. and Salakhutdinov, R.: Reducing the dimensionality of data with neural networks, *Science*, Vol. 313, No. 5786, pp. 504–507 (2006).
[15] Cho, K., Ilin, A. and Raiko, T.: Improved learning of gaussian-bernoulli restricted boltzmann machines, *ICANN*, pp. 10–17 (2011).
[16] Kang, S., Qian, X. and Meng, H.: Multi-distribution deep belief network for speech synthesis, *ICASSP*, pp. 8012–8016 (2013).
[17] Kosko, B.: Bidirectional associative memories, *IEEE Trans Systems, Man, Cybern.*, Vol. 18, No. 1, pp. 49–60 (1988).
[18] Chen, L.-H., Raitio, T., Valentini-Botinhao, C., Ling, Z.-H. and Yamagishi, J.: A deep generative architecture for postfiltering in statistical parametric speech synthesis, *Technometrics*, Vol. 23, No. 11, pp. 2003–2014 (2015).
[19] HTS ワーキンググループ：HMM-based Speech Synthesis System (HTS), 名古屋工業大学 (online), available from ⟨http://hts.sp.nitech.ac.jp/⟩ (accessed 2017-5-11).
[20] Desai, S., Raghavendra, E. V., Yegnanarayana, B., Black, A. W. and Prahallad, K.: Voice conversion using artificial neural networks, *ICASSP*, pp. 3893–3896 (2009).