# Deep Boltzmann machine を用いた音素ラベル情報推定*

☆高島悠樹, 中鹿亘, 滝口哲也, 有木康雄 (神戸大)

## 1 はじめに

近年，音声認識技術は広く普及し，人々の生活の助けとなっている．スマートフォンを例に挙げると，端末に対して発話を行うことで通話やメールを行うことができる．また，これまでの多くは成人を対象としたものだったが，現在では高齢者や子供など発話スタイルの異なる音声の場合でも高い精度が得られており，利用できる対象は広がりつつある．しかし，これらは言語障害などのない人々を対象としており，構音障害などの言語障害を患う方を対象とした音声認識は非常に少ない．言語障害には様々な種類の症状があるが，本研究では，アテトーゼ型の脳性麻痺による構音障害者を対象としている．アテトーゼ型の脳性麻痺では，大脳基底核に損傷を受けたことによる筋肉の不随意運動 (アテトーゼ) のために，筋肉の動きを正常に制御できない症状が現れる．この症状は，緊張時や意図的な動作を行おうとする際に多く生じるため，発話時に筋肉の緊張が起こり正しく構音できない場合がある．発話が困難な方でも，手話や音声合成システム [1] を用いて会話を行うことも可能であるが，脳性麻痺患者の多くは手足が不自由であり，音声に頼らざるを得ない状況が考えられる．構音障害者の音声認識の実現により，健常者とのコミュニケーションを円滑にし，また，障害者の就業機会の増加や生活の補助などが期待される．そのため，構音障害者の音声に対する研究の必要性は高いといえる．

構音障害者の発話スタイルは，筋肉の不随意運動により健常者と大きく異なるため，従来の不特定話者モデルでは認識精度が著しく低下する．そのため，構音障害者特有のモデルを用意する必要がある．また，発話内容が同じであっても，発話のばらつきが健常者と比べて大きくなるという課題が考えられる．従来研究として，CNN (Convolutional Neural Network [2, 3, 4]) を用いた発話変動にロバストな音声特徴量抽出法 [5] がある．CNN 特有の畳み込み操作とプーリング操作により，構音障害者特有の発話変動によるスペクトルの微小な変化に対して頑健な特徴量抽出を行うことができる．この手法は，ネットワークの学習に Back propagation を用いており，教師信号として HMM による強制アライメントの結果を用いている．しかし，構音障害者の音声スペクトルは変動が大きいため，精度の良いアライメントをとることができない．そのため，ネットワークの学習に用いる教師信号は誤りを含むことになり，より有効な特徴量抽出を阻害していると考えられる．実際，我々の予備実験により，HMM アライメントを人手で修正したアライメントを用いてネットワークを学習すると音声認識精度が向上することを確認した．このことから，さらなる構音障害者音声認識精度の向上のために，より精度の高いアライメント情報を得る必要があるといえる．強制アライメントが乱れる原因として，構音障害者音声特有の音素の欠落や置換によるラベル情報の誤りが挙げられる．そこで，DBM (Deep Boltzmann machine [6]) を用いた教師なしのラベル情報推定を行うことにより，正しいラベル情報の取得を試みる．

第 2 章で構音障害者音声特有の課題およびラベル情報推定の基本要素について述べ，第 3 章で実験と評価を行ったのち，特徴量抽出についても検討する．最後に，第 5 章で本研究の成果をまとめるとともに，今後の研究課題について述べる．

## 2 確率モデルベースのラベル情報推定

本節ではまず，構音障害者音声の特徴について述べる．次に，提案手法の DBM (Deep Boltzmann Machine) を用いたラベル情報の推定の基礎要素である RBM (restricted Boltzmann machine [7])，DBN (Deep Belief Networks [8]) および DBM について述べ，続いて提案の流れについて説明する．

### 2.1 構音障害者音声の特徴

構音障害者はアテトーゼと呼ばれる筋肉の不随意運動を伴うため，常に意図した構音ができるとは限らない．つまり，同一話者の同一発話であっても，そのばらつきは健常者と比べて大きくなるという傾向がある．現在，音声認識で広く用いられている MFCC などのケプストラム特徴量は，音声信号のフレームごとに独立して計算されるため，上述の課題を十分に考慮しているとは言えない．そのため，特徴量として MFCC を用いて強制アライメントを行った場合，その結果は必ずしも期待したものになるとは限らない．

Fig. 1 は，健常者発話/hyoujun/のスペクトル，Fig. 2 は，障害者発話/hyoujun/の強制アライメント結果と修正アライメント結果の比較を示す．Figs. 1，2 より，健常者発話に比べて，障害者発話はフォルマントがはっきりせず，特徴を掴みづらいことが確認でき

*Phoneme Estimation using Deep Boltzmann Machine, by Yuki Takashima, Toru Nakashika, Tetsuya Takiguchi, Yasuo Ariki (Kobe University)

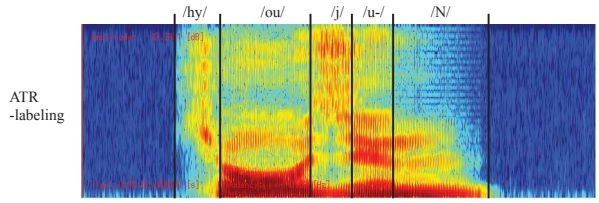


Fig. 1　Example of a spectrogram spoken by a physically unimpaired person /hyoujun/

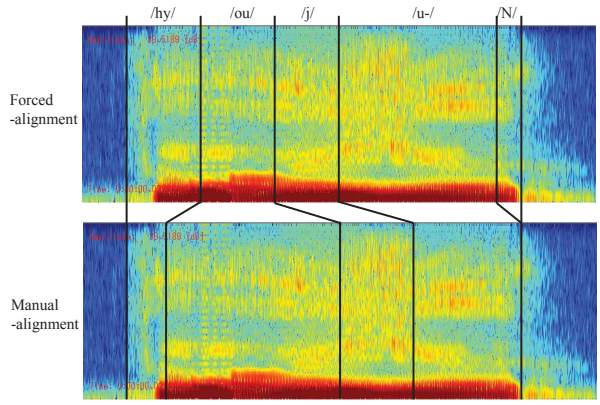


Fig. 2　Example of a spectrogram spoken by a person with an articulation disorder /hyoujun/

る．Fig. 2 の/hyoujun/は，最後の'N' の音素が欠落してしまっているために，'ou' が短すぎたり'u-' が長すぎたりする例である．構音障害者は，発話による負担が健常者より大きいため，発話の最後で息漏れを起こすなどして，音素の欠落が生じる．また，意図した構音ができないために，音素が部分的に変わったり，別の音素が挿入されたりといった問題もある．このような音素の欠落や置換を検出するために，本研究ではRBM を用いて入力発話内の音素ラベル情報の取得を試みる．

### 2.2　RBM

RBM は，可視素子 $v_i$ と隠れ素子 $h_j \in \{0,1\}$ 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 $v_i \in \mathbb{R}$ を定義した IGB (Improved Gaussian-Bernoulli)-RBM [9](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．

$$p(\mathbf{v},\mathbf{h}) = \frac{1}{Z}\exp\Big(-E(\mathbf{v},\mathbf{h})\Big), \tag{1}$$

$$E_{\mathrm{RBM}}(\mathbf{v},\mathbf{h}) = \sum_i \frac{(v_i - b_i)^2}{2\sigma_i^2} - \sum_{i,j} v_i W_{ij} h_j - \sum_j c_j h_j, \tag{2}$$

$$Z = \sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})} \tag{3}$$

ここで，$\mathbf{v} \in \mathbb{R}^{I\times 1}$，$\mathbf{h} \in \mathbb{R}^{J\times 1}$，$b$，$c$，$W$，$\sigma$ はそれぞれ可視素子，隠れ素子，可視層バイアス，隠れ層バイアス，可視層-隠れ層間の結合重みであり，いずれも推定すべきパラメータである．$i,j$ は，それぞれ可視素子および隠れ素子のインデクスであり，$I,J$ は，それぞれ可視素子および隠れ素子の次元数である．可視素子および隠れ素子の条件付き確率は以下の式で表現される．

$$p(v_i = v|\mathbf{h}) = \mathcal{N}(v|b_i + \sum_j h_j W_{ij}, \sigma_i^2), \tag{4}$$

$$p(h_j = 1|\mathbf{v}) = sigm(c_j + \sum_i W_{ij}\frac{v_i}{\sigma_i^2}), \tag{5}$$

$$sigm(x) = \frac{1}{1+e^{-x}} \tag{6}$$

RBM の各パラメータは，$N$ 個の観測データを $\{v^{(n)}\}_{n=1}^N$ とするとき，この確率変数の対数尤度 $\mathcal{L} = \log\prod_n p(\mathbf{v}_n)$ 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 $\theta$ で偏微分すると，

$$\frac{\partial \log p(\mathbf{v}^{(n)})}{\partial\theta} = \langle\frac{\partial E_{RBM}(\mathbf{v}^{(n)},\mathbf{h})}{\partial\theta}\rangle_{data} + \langle\frac{\partial E_{RBM}(\mathbf{v},\mathbf{h})}{\partial\theta}\rangle_{model}, \tag{7}$$

が得られる．ここで，$\langle\cdot\rangle_{data}$ と $\langle\cdot\rangle_{model}$ 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 [8] を用いて求められる．各パラメータは式 (7) から，確率的勾配法 (SGD) を用いて繰り返し更新される．

### 2.3　DBN

RBM はそれ単体では表現力に限りがあるため，RBM を積み重ね深いネットワーク (DBN) にすることで，表現力を高めることができると考えられる [10]．RBM と同様に，DBN の場合も隣接した層の間は全結合があり，層内の結合は存在しない．ネットワークの学習には greedily algorithm が用いられる．これは，任意の RBM の学習を前段の RBM のアクティベーションを入力として行う学習方法である．

### 2.4　DBM

DBM は，DBN と同じく RBM を積み重ねたものであるが，その違いは，DBN がボトムアップパスのみであることに対し，DBM はトップダウンパスを許していることである．以下に 3 層 DBM のエネルギー関数と隠れ素子の条件付き確率を示す．可視層の条件付き確率は DBN と同様に求められる．

$$E_{\mathrm{DBM}}(\mathbf{v},\mathbf{h}^1,\mathbf{h}^2)=\sum_i\frac{(v_i-b_i)^2}{2\sigma_i^2}-\sum_{i,j}v_iW_{ij}^1h_j^1-\sum_{j,m}h_jW_{jm}^2h_m^2, \tag{8}$$

$$p(h_j^1=1|\mathbf{v},\mathbf{h}^2)=sigm(\sum_i W_{ij}^1\frac{v_i}{\sigma_i^2}+\sum_m W_{jm}^2h_m^2), \tag{9}$$

$$p(h_m^2=1|\mathbf{h}^1)=sigm(\sum_j W_{jm}^2h_j^1). \tag{10}$$

ここで，$\mathbf{h}^1\in\mathbb{R}^{J\times1}$，$\mathbf{h}^2\in\mathbb{R}^{M\times1}$，$m$，$M$ はそれぞれ，隠れ第 1・2 素子，隠れ第 2 素子のインデクス，次元数を表す．隠れ層のバイアスは省略した．

### 2.5 提案手法

まず，音声データを観測データとしてモデルに与え，パラメータを学習する．隠れ素子に音素を対応させ，再び同一データをモデルに与えると，そのフレームに対応する音素が隠れ素子に出現すると仮定する．隠れ素子の条件付き確率を求め，もっとも値の大きい素子に対応する音素を推定されたラベル情報とする．通常，RBM を用いて推定された隠れ素子間には相関はないが，1 つの音声フレームには 1 つの音素が対応しているとする仮定して，下記のモデルを使用する．

Gaussian-softmax RBM[11] は通常の RBM に以下の制約を与える．

$$\sum_j h_j\le1 \tag{11}$$

この制約により，隠れ素子の条件付き確率は以下のように表現される．

$$p(h_j=1|\mathbf{v})=\frac{\exp(c_j+\sum_i W_{ij}\frac{v_i}{\sigma_i^2})}{1+\sum_j\exp(c_j+\sum_i W_{ij}\frac{v_i}{\sigma_i^2})} \tag{12}$$

この制約は全ての隠れ素子が 0，あるいは 1 つの素子のみが 1 になるということを意味する．無音区間にもラベルを付与できるように，我々は上記の制約を以下のように修正して用いる．また，その時の隠れ素子の条件付き確率も示す．

$$\sum_j h_j=1 \tag{13}$$

$$p(h_j=1|\mathbf{v})=\frac{\exp(c_j+\sum_i W_{ij}\frac{v_i}{\sigma_i^2})}{\sum_j\exp(c_j+\sum_i W_{ij}\frac{v_i}{\sigma_i^2})} \tag{14}$$

DBN，DBM の最上位 RBM を Gaussian softmax RBM として評価を行う．

## 3 評価実験

提案手法の有効性を確認するために，まず健常者音声に対して実験を行った．本実験では ATR 研究用日本語データベース (A set)[12] を用いて提案手法の効果を確認した．このデータベースから，男性話者 1 名 (MMY) を選び，モデルの学習・評価に使用した．音素バランス 216 単語から，5 母音 a,i,u,e,o をそれぞれ 1000 フレームずつ取り出して使用した．学習・評価用のデータはいずれも同じで，39 次元のメル周波数スペクトラムを用いた．推定された隠れ素子の条件付き確率は連続値のため，2 値化して正解率を算出した．また，切り出した発話 (母音) の総フレーム数の半分以上のフレームが正しくアクティベートされている場合に，その発話を正解として，発話毎の正解率を算出した．モデルには，RBM，隠れ層 2 層の DBN と DBM を使用した．RBM の隠れ素子数は 5，DBN と DBM の隠れ素子数は可視層に近い方から 25，5 とした．Fig. 3 に，本実験のモデルを示す．また，比較として 5 混合 GMM(Gaussian mixture model) を用いた実験も行った．GMM の各ガウス分布が母音を表現していると考え，事後確率が最大の分布に対応する音素にラベリングされたと考える．

RBM と DBN の場合，Figs. 4，5 より，どの母音も総じて高い精度を示していることが確認できる．本実験で使用したモデルはフレーム単位で推定を行うため，信号が定常状態にある時は識別可能だが，他の音素との境界部分はスペクトルが大きく変化しているため，正しいラベル推定が難しいと考えられる．Fig. 5 より，u の判別率が他に比べて劣っていることが確認できる．DBM の場合は，どの音素も前述の 2 つのモデルよりも精度が劣っている．特に e に関しては著しく低くなっており，モデルの見直しをする必要があると考えられる．Table 1 に，DBN を使用した時の confusion matrix を示す．これより，u は i に誤り易いことが分かる．これは，本実験で切り出した u の音声信号が半母音や，定常状態が極端に短いものを多く含んでいたために生じていると考えられる．他にもスペクトルに違いが大きく現れる無声化子音に関しても隠れ素子の追加をするなど検討する必要があると考えられる．

## 4 おわりに

本稿では，構音障害者の正確な音素ラベル情報の取得を目指して，RBM ベースの確率モデルを用いた音素ラベル推定を提案し，健常者音声 5 母音に対して，その有効性を確認した．DBM に関しては，特にモデルの検討の必要があることが確認できた．また，

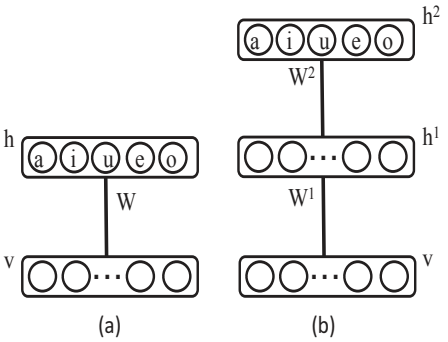


Fig. 3　Graphical representation of (a) an RBM and (b) a deep architecture

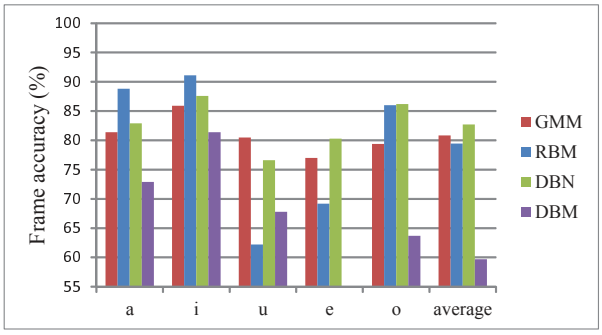


Fig. 4　Accuracy per frame

筋肉の緊張により発話が変動しやすいという構音障害者特有のモデルも検討する予定である.

## 参考文献


[1] C. Veaux *et al.*, "Using HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders," in *INTERSPEECH*. 2012, ISCA.

[2] Y. LeCun and Y. Bengio, "The handbook of brain theory and neural networks," chapter Convolutional Networks for Images, Speech, and Time Series, pp. 255–258. MIT Press, Cambridge, MA, USA, 1998.

[3] Y. LeCun *et al.*, "Gradient-based learning applied to document recognition," in *Intelligent Signal Processing*. 2001, pp. 306–351, IEEE Press.

[4] H. Lee *et al.*, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Advances in Neural Information Processing Systems 22*, pp. 1096–1104. 2009.

[5] T. Nakashika *et al.*, "Dysarthric speech recognition using a convolutive bottleneck network," in *ICSP*, 2014, pp. 505–509.

[6] R. Salakhutdinov and G. Hinton, "Deep Boltzmann machines," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2009, vol. 5, pp. 448–455.

[7] Y. Freund and D. Haussler, "Unsupervised learning of distributions on binary vectors using two layer networks," Tech. Rep., Santa Cruz, CA, USA, 1994.

[8] G. E. Hinton *et al.*, "A fast learning algorithm for deep belief nets," Neural computation, vol. 18, no. 7, pp. 1527–1554, 2006.

[9] 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.

[10] R. R. H. Lee, R. Grosse and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th Annual International Conference on Machine Learning, ser, ICML'09*, 2009, pp. 609–616.

[11] H. L. K. Sohn, D. Y. Jung and A. O. H. lii, "Efficient learning of sparse, distributed, convolutional feature representations for object recognition," in *ICCV*, 2011.

[12] 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.


Table 1　Confusion matrix of the DBN

| | | result | | | | | |
|---|---|---|---|---|---|---|---|
| | | a | i | u | e | o | total |
| correct | *a* | 78 | 0 | 0 | 0 | 0 | 78 |
| | *i* | 0 | 75 | 0 | 0 | 0 | 75 |
| | *u* | 1 | 12 | 58 | 0 | 0 | 71 |
| | *e* | 0 | 2 | 1 | 67 | 0 | 70 |
| | *o* | 0 | 0 | 0 | 0 | 80 | 80 |
| | *total* | 79 | 89 | 59 | 67 | 80 | |

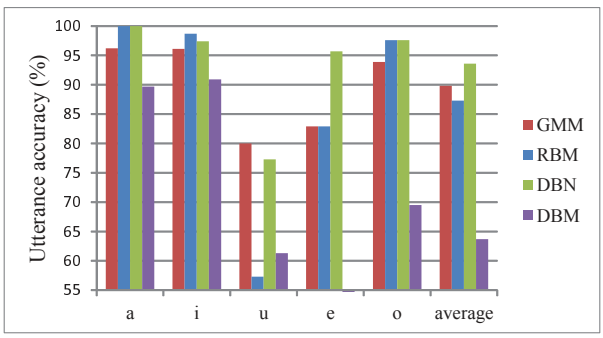


Fig. 5　Accuracy per utterance