

適応型 Gaussian-Gaussian RBM を用いた構音障害者音声認識*

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

現在, 我が国の障害者手帳を持つ 18 歳以上の人口は 350 万人を超えており, 聴覚・言語障害者の数は 36 万人とされている [1]. 文献 [2] では, 構音障害者音声を対象とした音響モデル適応の検証を行っているが, 言語障害者などの障害者を対象としている研究は非常に少ない.

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, 意図的な動作を行う際に筋肉の不随意運動が発生するため, 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, 手話認識や音声合成システムを使用することでコミュニケーションをとることは可能であるが, 脳性麻痺患者の多くは手足が不自由であり, 音声に頼るしかない状況が考えられる. そのため, 構音障害者のための音声認識には十分なニーズがあり, 研究の必要性があるといえる. 音声認識技術を用いることで, 発話内容を聞き取ることが困難な健常者とのコミュニケーションが円滑になり, 障害者の就業機会の増加や講演時の補助への活用などが期待される.

構音障害者の発話スタイルは, 筋肉の付随運動により健常者と大きく異なるため, 従来の不特定話者音響モデルは役に立たず, 認識精度が著しく低下する. 話者性を音響モデルに反映させる手法として, MLLR (maximum likelihood linear regression) [3] を用いた話者適応が挙げられる. しかし, 健常者と構音障害者の発話スタイルは大きく異なるため, 十分な精度の向上は望めない. また, 発話内容が同一であっても, 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 [4]) を用いた発話変動にロバストな音声特徴量抽出法 [5] が提案されてきた. この手法の問題点として, 大きく 2 つ挙げられる. 1 つ目は, 教師信号に HMM (hidden Markov model) による強制アライメントの結果を用いている点である. 構音障害音声のスペクトルは変動が大きいので, 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. 2 つ目は, ニューラルネットワークの学習に大量のデータを必要とする点である. 構音障害者は筋肉の不随意運動による身体への負担が大きく, 大量の音声データを収録することは困難であり, 話者毎の学習データ量は限られてくる.

本研究では, 上述の問題点を克服するために, 適応型 RBM (adaptive restricted Boltzmann machine; ARBM [6]) を用いた音響特徴量抽出法を提案する. このモデルは RBM (restricted Boltzmann machine [7]) を拡張したものであり, エネルギー関数に基づく確率モデルである. 声質変換において, 音韻情報と話者情報を分離したモデル化を実現した. 近年, RBM を用いた特徴抽出 (特徴学習) 法 [8, 9] が多く提案されており, 物体認識や音素認識に応用されている. 適応型 RBM は, 複数話者の音声データ集合から, 話者に依存しない情報と話者に依存する情報に分離しながら, 潜在的な特徴を抽出する確率モデルである. 話者共通のパラメータと話者固有のパラメータを明示的に分離しながらモデルの学習が行なわれるため, 潜在特徴量が音韻情報を表現すると仮定し, 本研究では, この潜在特徴量を音響特徴量として音声認識に利用する. また, 適応型 RBM は教師なし学習によりパラメータの推定が行なわれるため, 誤ったアライメント情報を用いる必要がない. 学習データに存在しない未知話者に対しては, 未知話者のデータ (適応データ) を用いて, 話者非依存パラメータを固定しながら話者依存パラメータのみを推定するため, 話者毎のデータ量を少なく抑えることができる. 以下, 第 2 章で通常の RBM について述べ, 第 3 章で適応型 RBM と音声認識への応用について説明する. 第 4 章で従来の MLLR による話者適応と比較し, 第 5 章で本稿をまとめる.

2 Restricted Boltzmann Machine

RBM は, Fig. 1 左図に示すように, 可視素子 $v \in \mathbb{R}^D$ と隠れ素子 $h \in \mathbb{B}^H$ ($\mathbb{B}$ 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 [10] (以下, この IGB-RBM を単に

* Dysarthric speech recognition using Gaussian-Gaussian RBM, by Yuki Takashima (Kobe University), Toru Nakashika (UEC), Tetsuya Takiguchi (Kobe University/JST PRESTO), Yasuo Ariki (Kobe University)

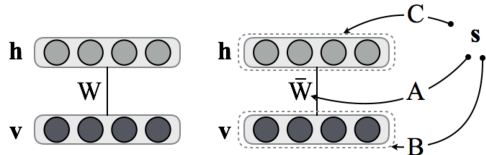


Fig. 1 Graphical representation of an RBM (left) and adaptive RBM (right)

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.

$$p(\mathbf{v}, \mathbf{h}) = \frac{1}{Z} e^{-E_{RBM}(\mathbf{v}, \mathbf{h})}$$

$$E_{RBM}(\mathbf{v}, \mathbf{h}) = \sum_i \frac{(v_i - b_i)^2}{2\sigma_i^2} - \sum_{i,j} \frac{v_i}{\sigma_i^2} W_{ij} h_j - \sum_j c_j h_j \quad (1)$$

ここで, $[b_1, \dots, b_D] \in \mathbb{R}^D$, $[c_1, \dots, c_H] \in \mathbb{R}^H$, $\{W_{ij}\}_{D \times H} \in \mathbb{R}^{D \times H}$, $[\sigma_1^2, \dots, \sigma_D^2] \in \mathbb{R}^D$ 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, 可視素子の分散を表し, いずれも推定すべきパラメータである. i, j はそれぞれ, 可視素子及び隠れ素子のインデックスである. ここで, $Z = \int^D \sum_{\mathbf{h}} e^{-E(\mathbf{v}, \mathbf{h})} d^D \mathbf{v}$ 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.

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

$$p(h_j = 1 | \mathbf{v}) = \mathcal{S}(c_j + \sum_i W_{ij} \frac{v_i}{\sigma_i^2}) \quad (3)$$

ここで, $\mathcal{N}(\cdot | \mu, \sigma^2)$ は平均 μ , 分散共分散 σ^2 の正規分布, $\mathcal{S}(\cdot)$ 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 $\{\mathbf{v}_n\}_{n=1}^N$ 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 θ で偏微分すると,

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

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 [11] 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 適応型 RBM による特徴量抽出

3.1 適応型 RBM

適応型 RBM は, Fig. 1 右図に示すように, 前節で述べた RBM を拡張したモデルであり, 新たに識別素子 $\mathbf{s} = [s_1, \dots, s_R]^T$ を持つ (R は識別素子の数とする). 例えば, 入力 $\mathbf{v}$ が話者 r の発話であることを示す場合, $s_r = 1, \forall s_{r'} = 0 (r' \neq r)$ となる. このモデルでは, 可視素子と隠れ素子間の結合重み, 可視素子及び隠れ素子のバイアスは識別素子 $\mathbf{s}$ で制御される. この結合重み $\mathbf{W}(\mathbf{s})$, 可視素子及び隠れ素子のバイアス $\mathbf{b}(\mathbf{s})$, $\mathbf{c}(\mathbf{s})$ はそれぞれ以下のように定義される.

$$\mathbf{W}(\mathbf{s}) = \sum_r \mathbf{A}_r s_r \bar{\mathbf{W}} \quad (5)$$

$$\mathbf{b}(\mathbf{s}) = \bar{\mathbf{b}} + \sum_r \mathbf{b}_r s_r = \bar{\mathbf{b}} + \mathbf{B} \mathbf{s} \quad (6)$$

$$\mathbf{c}(\mathbf{s}) = \bar{\mathbf{c}} + \sum_r \mathbf{c}_r s_r = \bar{\mathbf{c}} + \mathbf{C} \mathbf{s} \quad (7)$$

ただし, $\bar{\mathbf{W}} \in \mathbb{R}^{D \times H}$, $\bar{\mathbf{b}} \in \mathbb{R}^D$, $\bar{\mathbf{c}} \in \mathbb{R}^H$ はそれぞれ, 話者非依存な結合重み, 可視素子及び隠れ素子の話者非依存なバイアスを表す. また, $\mathbf{A}_r \in \mathbb{R}^{D \times D}$, $\mathbf{b}_r \in \mathbb{R}^D$ ($\mathbf{B} = [\mathbf{b}_1 \ \mathbf{b}_2 \ \dots \ \mathbf{b}_R] \in \mathbb{R}^{D \times R}$), $\mathbf{c}_r \in \mathbb{R}^H$ ($\mathbf{C} = [\mathbf{c}_1 \ \mathbf{c}_2 \ \dots \ \mathbf{c}_R] \in \mathbb{R}^{H \times R}$) はそれぞれ, 話者 r 固有のパラメータを表す. つまり, $\mathbf{A}_r$ は, 話者非依存な結合重み $\bar{\mathbf{W}}$ を話者 r へ特定化 (適応) するための適応行列であり, $\mathbf{b}_r$, $\mathbf{c}_r$ はそれぞれ, 話者 r 固有のバイアスを表す.

適応型 RBM では, 式 (5)~(7) で定義したパラメータを用いて, 式 (1) より同時確率及びエネルギー関数は以下のように定義される.

$$p(\mathbf{v}, \mathbf{h}, \mathbf{s}) = \frac{1}{Z_A} e^{-E_A(\mathbf{v}, \mathbf{h}, \mathbf{s})}$$

$$E_A(\mathbf{v}, \mathbf{h}, \mathbf{s}) = \sum_i \frac{(v_i - b_i(\mathbf{s}))^2}{2\sigma_i^2} - \sum_{i,j} \frac{v_i}{\sigma_i^2} W_{ij}(\mathbf{s}) h_j - \sum_j c_j(\mathbf{s}) h_j \quad (8)$$

$$Z_A = \int^D \sum_{\mathbf{h}, \mathbf{s}} e^{-E(\mathbf{v}, \mathbf{h}, \mathbf{s})} d^D \mathbf{v} \quad (9)$$

ここで, これらの定義により, 条件付き確率も通常の RBM と同様に以下のように計算できる.

$$p(v_i = v | \mathbf{h}, \mathbf{s}) = \mathcal{N}(v | b_i(\mathbf{s}) + \sum_j h_j W_{ij}(\mathbf{s}), \sigma_i^2) \quad (10)$$

$$p(h_j = 1 | \mathbf{v}, \mathbf{s}) = \mathcal{S}(c_j(\mathbf{s}) + \sum_i W_{ij}(\mathbf{s}) \frac{v_i}{\sigma_i^2}) \quad (11)$$

適応型 RBM の各パラメータは, 従来の RBM と同様に, N 個の観測データを $\{\mathbf{v}_n, \mathbf{s}_n\}_{n=1}^N$ とするとき, この確率変数の対数尤度 $\mathcal{L} = \log \prod_n p(\mathbf{v}_n, \mathbf{s}_n) =$

$\log \prod_n \sum_h p(v_n, h_n, s_n)$ を最大化するように推定される。

3.2 適応型 RBM を用いた音声認識

適応型 RBM を音声認識へ応用する場合、まず複数 (R 人) の話者によるデータを用いて適応型 RBM の各パラメータを同時推定する。そして、推定されたパラメータを用いて、複数の話者によるデータに対して、次式のように潜在特徴量 (隠れ素子) を推定する。

$$\hat{h}_n = \arg \max_h p(h|v_n, s_n) = S(c(s_n) + \frac{v_n^2}{\sigma^2} W(s_n)) \quad (12)$$

得られた潜在特徴量を音響特徴量とみなし、音響モデルを学習する。次に、 $\bar{W}$ など話者に依存しないパラメータを固定して、適応データを用いて評価話者の適応パラメータ A_{R+1}, b_{R+1} を推定する。そして、得られた適応パラメータと話者に依存しないパラメータを用いて式 (12) と同様に潜在特徴量を推定し、学習された音響モデルを用いて認識を行う。

4 評価実験

4.1 実験条件

データセットとして、構音障害者の男性 6 名の音声データを収録し使用した。発話内容は、ATR 音素バランス単語 A セット [12] から 216 単語を選択した。障害の程度により単語の発話数が異なり、話者 B, C, D は各単語 5 発話分、話者 A, D, E は各単語 3 発話分の音声データを収録した。音声の標準化周波数は 16kHz、語長 16bit であり、音響分析には Hamming 窓を用いた。STFT におけるフレーム幅、シフト幅はそれぞれ 25ms, 10ms である。本稿で用いる音響モデルは、54 音素の monophone-HMM で、各 HMM の状態数は 5、状態あたりの混合分布数は 6 である。

ARBM への入力音響特徴量として 32 次元のメルケプストラムを用いた。話者により各単語の発話数が異なるが、評価話者以外の 1~3 発話目の音声データを ARBM の学習に使用した。評価話者については次節で述べる。また、潜在特徴量の数を 32 とした。学習率 0.01、モーメント係数 0.9、バッチサイズ 512、繰り返し回数 500 の確率的勾配法を用いてモデルを学習した。

4.2 MLLR による話者適応

まず、ベースラインとして、従来の MLLR による話者適応の実験を行なった。音響特徴量は、12 次元の MFCC とそのデルタ成分とし、cepstral mean normalization を施したものを使用した。ある話者の評価を行う際には、当該話者以外の全データを学習に、当

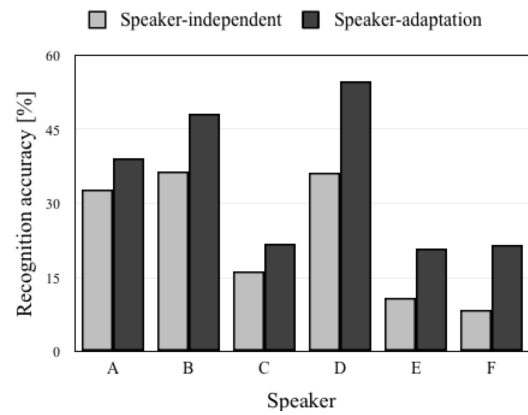


Fig. 2 Word recognition accuracies of speaker-independent acoustic model and speaker-adaptation acoustic models.

該話者の各単語の 1 発話目を適応に、当該話者の各単語の最終発話 (3 あるいは 5 発話目) を評価に用いた。Fig. 2 に実験結果を示す。複数の障害者音声のみを用いて不特定話者音響モデルを構築しても、話者間で認識精度にばらつきがあることが分かる。話者 E, F の精度が著しく低い理由として、音声を含む無音区間の長さが考えられる。話者 E, F 以外の話者の音声は無音区間を比較的長く含んでいるが、話者 E, F の音声は無音区間をあまり含んでいない。そのため、cepstral mean normalization を施した際に、平均の値が他の話者のものと大きく異なると考えられる。さらに、MLLR による話者適応を行なうと、何れの話者も精度が改善された。話者 D の精度が最も高かったため、ARBM の実験の評価データとして話者 D を選択する。

4.3 ARBM による話者適応

本稿では、評価話者 D による実験は今後の研究課題とし、学習話者 B, C の 5 発話目を用いた評価を行う。Fig. 3 に提案手法による音声認識実験結果を示す。提案手法は通常の RBM を用いて特徴量抽出を行なった場合と比較して、話者 B は 19.91%, 話者 C は

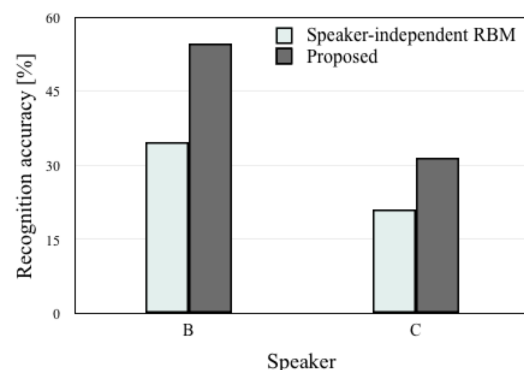


Fig. 3 Word recognition accuracies of each method.

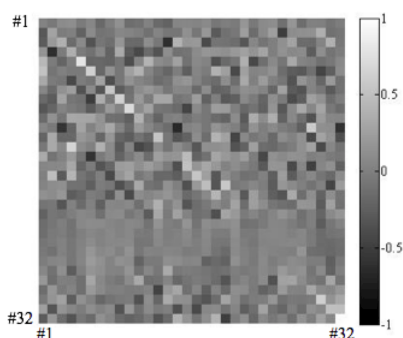


Fig. 4 Estimated adaptation matrix for the speaker B.

10.48%の精度向上が得られた。この理由として、話者に依存しないパラメータと話者に依存するパラメータに分離しながらモデルパラメータを推定したことで、潜在特徴量により明確な音韻情報が現れたためだと考えられる。

提案法によって、実際に推定された話者 B の適応行列 A を Fig. 4 に示す。構音障害者音声は高ケフレンシーが欠落しやすいという傾向があり、適応行列の高次元成分にもそれが現れている。Fig. 5 は、学習に用いた話者 B の /ikioi/ のメルケプストラムである。この図からも、構音障害者音声において高ケフレンシーがばやけていることが確認できる。

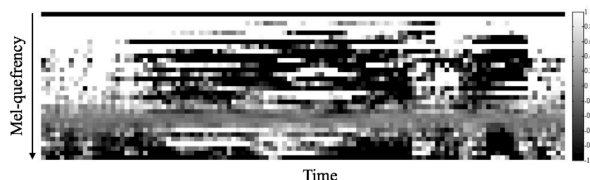


Fig. 5 Mel-cepstral feature of the sentence "ikioi" uttered by speaker B.

5 おわりに

本研究では、話者に依存するパラメータと依存しないパラメータを持つ適応型 RBM を用いた音響特徴量抽出法を提案した。構音障害者音声を用いた単語認識実験により、提案手法の評価を行なった。評価実験において、既知話者に対して提案手法の有効性を示した。今後は、未知話者に対しても実験を行う。また、構音障害者音声は高周波成分が欠落しやすいという特徴があり、これは話者非依存な結合重みを学習する際に不利に働くと考えられる。なぜなら、話者非依存な結合重みは声の基底成分を表現していると考えることができ、それを形成するには、より正確な音声データを用いる方がより自然であるからである。そのため、話者に依存しないパラメータは健常者で学習

し、適応パラメータで障害者特有の基底へ変換することで、より適切なモデル化ができると期待できる。

謝辞 本研究の一部は、JST さきがけの支援を受けたものである。

参考文献

- [1] 内閣省, “平成 25 年版障害者白書,” .
- [2] 中村圭吾 *et al.*, “発話障害者音声を対象にした健常者音響モデルの適応と検証,” 日本音響学会講演論文集, pp. 109–110, 2015.
- [3] M. J. F. Gales, “Maximum likelihood linear transformations for hmm-based speech recognition,” *Computer Speech & Language*, vol. 12, no. 2, pp. 75–98, 1998.
- [4] Y. LeCun *et al.*, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [5] T. Nakashika *et al.*, “Dysarthric speech recognition using a convolutive bottleneck network,” in *ICSP*, 2014, pp. 505–509.
- [6] T. Nakashika *et al.*, “Non-parallel training in voice conversion using an adaptive restricted boltzmann machine,” *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 24, no. 11, pp. 2032–2045, 2016.
- [7] Y. Freund and D. Haussler, “Unsupervised learning of distributions on binary vectors using two layer networks,” *Tech. Rep.*, 1994.
- [8] M. Ranzato *et al.*, “Factored 3-way restricted boltzmann machines for modeling natural images,” in *AISTATS*, Y. W. Teh and D. M. Titterton, Eds. 2010, vol. 9 of *JMLR Proceedings*, pp. 621–628, JMLR.org.
- [9] G. E. Dahl *et al.*, “Phone recognition with the mean-covariance restricted boltzmann machine,” in *NIPS*, J. D. Lafferty *et al.*, Eds. 2010, pp. 469–477, Curran Associates, Inc.
- [10] 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.
- [11] G. E. Hinton *et al.*, “A fast learning algorithm for deep belief nets,” *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, July 2006.
- [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.