一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
EA2016-140,SIP2016-195,SP2016-135 (2017-03)

構音障害者音声認識のための 適応型restricted Boltzmann machineを用いた特徴量抽出

高島　悠樹† 　中鹿　　亘†† 　滝口　哲也† 　有木　康雄†

† 神戸大学大学院システム情報学研究科 〒657-8501 兵庫県神戸市灘区六甲台町 1-1
†† 電気通信大学大学院情報理工学研究科 〒182-8585 東京都調布市調布ヶ丘 1-5-1
E-mail: †y.takashima@me.cs.scitec.kobe-u.ac.jp, ††nakashika@uec.ac.jp, †††{takigu,ariki}@kobe-u.ac.jp

あらまし　本報告では，アテトーゼ型脳性麻痺による構音障害者の音声特徴量の検討を行う．意図的な動作時や緊張状態にある場合に筋肉の不随意運動が生じるため，彼らの発話は不安定となる．従来研究として，畳み込みニューラルネットワークを用いた特徴量抽出法が提案されているが，アテトーゼ型脳性麻痺による構音障害者は発話による身体への負担が大きいために，大量の学習データを用意することが困難である．また，正確なアライメントを取ることが難しいため，ネットワークを学習するための教師信号を用意することも困難である．本研究では，適応型 RBM (adaptive restricted Boltzmann machine) を用いた特徴量抽出法を提案する．この手法の特徴として，教師なし学習と話者正規化学習が挙げられる．適応型 RBM 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明示的に分離しながら学習が行なわれるため, 話者毎の学習データ量を削減することができる．また, 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, 不明瞭な発話からラベル情報を得る必要がない．本稿では，提案手法により得られる特徴量を用いた単語認識実験の結果を報告する．

キーワード　構音障害, 特徴量抽出，　restricted Boltzmann machine, 話者適応

Feature Extraction Using Adaptive Restricted Boltzmann Machine for Dysarthric Speech Recognition

Yuki TAKASHIMA†, Toru NAKASHIKA††, Tetsuya TAKIGUCHI†, and Yasuo ARIKI†

† Graduate School of System Informatics, Kobe University
1-1 Rokkodaicho, Nada-ku, Kobe, Hyogo, 657-8501 Japan
†† Graduate School of Information Systems, The University of Electro-Communications
1-5-1 Chofugaoka, Chofu, Tokyo, 182-8585 Japan
E-mail: †y.takashima@me.cs.scitec.kobe-u.ac.jp, ††nakashika@uec.ac.jp, †††{takigu,ariki}@kobe-u.ac.jp

Abstract　We investigate in this paper a feature extraction method for a person with an articulation disorder resulting from athetoid cerebral palsy. In the case of a person with this type of articulation disorder, the articulation of the utterances tends to become unstable due to the strain placed on the speech-related muscles. In our previous work, the feature extraction method using a convolutional neural network was proposed. In general, neural networks require sufficient amount of training data. However, generally speaking, the amount of speech data obtained from a person with an articulation disorder is limited because their burden is large due to strain on the speech muscles. Because the dysarthric speech fluctuates every utterance, it is difficult to obtain the correct alignment. In this paper, we propose a feature extraction method using adaptive restricted Boltzmann machine (ARBM). Because an ARBM is trained separating the speaker-independent parameters and the speaker-dependent parameters expressly, the amount of the training data can be reduced. The parameters of an ARBM are estimated using unsupervised learning. Therefore, it is not necesary to use incorrect label information. In this paper, we report our experimental results of speech recognition using the features extracted from our proposed method.

Key words　Articulation disorders, feature extraction, restricted Boltzmann machine, speaker adaptation

1. はじめに

我が国においては，平成 28 年 4 月より「障害を理由とする差別の解消の推進に関する法律」(以下，「障害者差別解消法」) が施行され，全ての国民が，相互に人格と個性を尊重し合いながら共生する社会の実現が目指されている．障害者差別解消法は，社会的障壁の除去を怠ることによる権利侵害の防止を定めている．典型的な例として，窓口で障害者の障害の特性に応じたコミュニケーション手段の提供が挙げられる．「声」は代表的なコミュニケーション手段である．近年，音声認識技術は盛んに研究され，日常生活に普及している．しかし，多くの音声認識技術は健常者を対象としており，障害者を対象としたものは非常に少ない．社会的障壁の除去のために，障害者を対象とした音声認識技術の研究の必要性があると考えられる．日本では，障害者手帳を持つ人口が約 500 万人を越えている [1]．また，近年の超高齢社会において，加齢に伴う身体機能の低下や障害を持つ人は数千万人にのぼる．このような背景からも，福祉分野における情報技術の重要性が高まってきている [2], [3]．英語圏においては，構音障害者音声のデータベースが作成されている [4]〜[6]．脳性麻痺の種類や音声の明瞭性が明記されており，日本よりも盛んに研究が行なわれている．

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 [7] 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．また，構音障害は脳性麻痺のみならず，健常者であっても，高齢による発話機能低下，また脳血管障害によっても起こる場合がある．本研究は，高齢化の進む日本において，高齢者とのコミュニケーション支援にも応用が可能である．

構音障害者の発話スタイルは，筋肉の付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) [8] 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傾向がある．従来研究として，DNN (deep neural network) を用いた手法 [9] や CNN (convolutional neural network [10]) を用いた発話変動にロバストな音声特徴量抽出法 [11] が提案されてきた．これら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．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 [12]) を用いた音声特徴量抽出法を提案する．このモデルは RBM (restricted Boltzmann machine [13]) 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 [14]〜[16] 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 $\boldsymbol{v} \in \mathbb{R}^I$ と隠れ素子 $\boldsymbol{h} \in \mathbb{B}^J$ ($\mathbb{B}$ 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 [17](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．

$$
\begin{aligned}
p(\boldsymbol{v}, \boldsymbol{h}) &= \frac{1}{Z} \mathrm{e}^{-E_{RBM}(\boldsymbol{v}, \boldsymbol{h})} \\
E_{RBM}(\boldsymbol{v}, \boldsymbol{h}) &= \left|\left| \frac{\boldsymbol{v} - \boldsymbol{b}}{2\boldsymbol{\sigma}} \right|\right|^2 - \left(\frac{\boldsymbol{v}}{\boldsymbol{\sigma}^2}\right)^{\mathrm{T}} \mathbf{W}\boldsymbol{h} - \boldsymbol{c}^{\top}\boldsymbol{h} \qquad (1)
\end{aligned}
$$

ここで，$||\cdot||^2$ は L2 ノルム，括線は要素除算を表す．$\boldsymbol{b} \in \mathbb{R}^I$, $\boldsymbol{c} \in \mathbb{R}^J$, $\mathbf{W} \in \mathbb{R}^{I \times J}$, $\boldsymbol{\sigma}^2 \in \mathbb{R}^I$ 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．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，

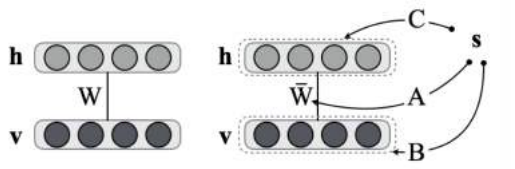

図 1　RBM と ARBM のグラフ構造

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

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．

$$p(v_i = v|\boldsymbol{h}) = \mathcal{N}(v|b_i + \boldsymbol{W}_{i:}\boldsymbol{h}, \sigma_i^2) \tag{2}$$

$$p(h_j = 1|\boldsymbol{v}) = \mathcal{S}(c_j + \left(\frac{\boldsymbol{v}}{\boldsymbol{\sigma}^2}\right)^{\mathrm{T}} \boldsymbol{W}_{:j}) \tag{3}$$

ここで，$\boldsymbol{W}_{i:}$ と $\boldsymbol{W}_{:j}$ は $\mathbf{W}$ の第 i 行ベクトル，第 j 列ベクトルを表す．また，$\mathcal{N}(\cdot|\mu, \sigma^2)$ 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 のパラメータ $\Theta = \{\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{\sigma}\}$ 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．この対数尤度をそれぞれのパラメータで偏微分すると，

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}} = \langle \left(\frac{\boldsymbol{v}}{\boldsymbol{\sigma}^2}\right) \boldsymbol{h}^{\mathrm{T}} \rangle_{\mathrm{data}} - \langle \left(\frac{\boldsymbol{v}}{\boldsymbol{\sigma}^2}\right) \boldsymbol{h}^{\mathrm{T}} \rangle_{\mathrm{model}}, \tag{4}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{b}} = \langle \left(\frac{\boldsymbol{v}}{\boldsymbol{\sigma}^2}\right) \rangle_{\mathrm{data}} - \langle \left(\frac{\boldsymbol{v}}{\boldsymbol{\sigma}^2}\right) \rangle_{\mathrm{model}}, \tag{5}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{c}} = \langle \boldsymbol{h} \rangle_{\mathrm{data}} - \langle \boldsymbol{h} \rangle_{\mathrm{model}}, \tag{6}$$

が得られる．ここで，$\langle \cdot \rangle_{\mathrm{data}}$ と $\langle \cdot \rangle_{\mathrm{model}}$ 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 (CD 法) [18] を用いて求められる．また，分散パラメータを非負値に制約し，学習を安定させるため $\sigma_i^2 = \mathrm{e}^{z_i}$ と置換し，z_i を更新することにより σ_i^2 の推定を行なう．z_i の勾配は以下のように計算される．

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial z_i} =& \mathrm{e}^{-z_i} \langle \frac{(v_i - b_i)^2}{2} - v_i \boldsymbol{W}_{i:}\boldsymbol{h} \rangle_{\mathrm{data}} \\ & - \mathrm{e}^{-z_i} \langle \frac{(v_i - b_i)^2}{2} - v_i \boldsymbol{W}_{i:}\boldsymbol{h} \rangle_{\mathrm{model}}\end{aligned} \tag{7}$$

各パラメータは式 (4)∼(7) から，確率的勾配法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

本節では，前節で述べた RBM を拡張したモデルとして，適応型 RBM (adaptive restricted Boltzmann machine; ARBM) について説明し，音声認識への応用として，適応型 RBM を用いた特徴量抽出法について述べる．

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 $\boldsymbol{s} = [s_1, \cdots, s_R]^T$ を持つ (R は識別素子の数とする). 例えば，入力 $\boldsymbol{v}$ 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の間の結合重み，可視素子及び隠れ素子のバイアスは識別素子 $\boldsymbol{s}$ で制御される．この結合重み $\mathbf{W}(\boldsymbol{s})$，可視素子及び隠れ素子のバイアス $\boldsymbol{b}(\boldsymbol{s}), \boldsymbol{c}(\boldsymbol{s})$ はそれぞれ以下のように定義される．

$$\mathbf{W}(\boldsymbol{s}) = \sum_r \mathbf{A}_r s_r \bar{\mathbf{W}} \tag{8}$$

$$\boldsymbol{b}(\boldsymbol{s}) = \bar{\boldsymbol{b}} + \sum_r \boldsymbol{b}_r s_r = \bar{\boldsymbol{b}} + \mathbf{B}\boldsymbol{s} \tag{9}$$

$$\boldsymbol{c}(\boldsymbol{s}) = \bar{\boldsymbol{c}} + \sum_r \boldsymbol{c}_r s_r = \bar{\boldsymbol{c}} + \mathbf{C}\boldsymbol{s} \tag{10}$$

ただし, $\bar{\mathbf{W}} \in \mathbb{R}^{D \times H}$, $\bar{\boldsymbol{b}} \in \mathbb{R}^D$, $\bar{\boldsymbol{c}} \in \mathbb{R}^H$ 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，$\mathbf{A}_r \in \mathbb{R}^{D \times D}$，$\boldsymbol{b}_r \in \mathbb{R}^D (\mathbf{B} = [\boldsymbol{b}_1\ \boldsymbol{b}_2\ \cdots\ \boldsymbol{b}_R] \in \mathbb{R}^{D \times R})$, $\boldsymbol{c}_r \in \mathbb{R}^H (\mathbf{C} = [\boldsymbol{c}_1\ \boldsymbol{c}_2\ \cdots\ \boldsymbol{c}_R] \in \mathbb{R}^{H \times R})$ 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，$\boldsymbol{b}_r$, $\boldsymbol{c}_r$ はそれぞれ，話者 r 固有のバイアスを表す. 話者適応行列の集合を，$\mathcal{A} = \{\mathbf{A}_r\}_{r=1}^R$ と表記することとする．

構音障害者は筋肉の不随意運動により安定した発声が困難であり，発話の変動が生じやすい傾向がある．この症状の程度は構音障害者それぞれ異なるため，話者毎の分散を考慮したモデル化を行なう必要があると考えられる．本研究では，話者固有の分散パラメータ $\boldsymbol{\sigma}^2(\boldsymbol{s})$ を以下の式で定義する．

$$\boldsymbol{\sigma}^2(\boldsymbol{s}) = \mathrm{e}^{\bar{\boldsymbol{z}} + \sum_r \boldsymbol{d}_r s_r} = \mathrm{e}^{\bar{\boldsymbol{z}} + \mathbf{D}\boldsymbol{s}} \tag{11}$$

ここで，話者非依存な分散を $\bar{\boldsymbol{\sigma}}^2 = \mathrm{e}^{\bar{\boldsymbol{z}}}$ と置換することとし，$\boldsymbol{d}_r \in \mathbb{R}^I (\mathbf{D} = [\boldsymbol{d}_1\ \boldsymbol{d}_2\ \cdots\ \boldsymbol{d}_R] \in \mathbb{R}^{I \times R})$ は話者 r 固有のパラメータを表す．

適応型 RBM では，式 (8)∼ (11) 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．

$$\begin{aligned} p(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{s}) &= \frac{1}{Z_A} \mathrm{e}^{-E_A(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{s})} \\ E_A(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{s}) &= \left|\left| \frac{\boldsymbol{v} - \boldsymbol{b}(\boldsymbol{s})}{2\boldsymbol{\sigma}(\boldsymbol{s})} \right|\right|^2 - \left(\frac{\boldsymbol{v}}{\boldsymbol{\sigma}^2(\boldsymbol{s})}\right)^{\mathrm{T}} \mathbf{W}(\boldsymbol{s})\boldsymbol{h} - \boldsymbol{c}(\boldsymbol{s})^{\top}\boldsymbol{h} \end{aligned} \tag{12}$$

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．

$$p(v_i = v|\boldsymbol{h}, \boldsymbol{s}) = \mathcal{N}(v|b_i(\boldsymbol{s}) + \boldsymbol{W}_{i:}(\boldsymbol{s})\boldsymbol{h}, \sigma_i^2(\boldsymbol{s})) \tag{13}$$

$$p(h_j = 1|\boldsymbol{v}, \boldsymbol{s}) = \mathcal{S}(c_j(\boldsymbol{s}) + \left(\frac{\boldsymbol{v}}{\boldsymbol{\sigma}^2(\boldsymbol{s})}\right)^{\mathrm{T}} \boldsymbol{W}_{:j}(\boldsymbol{s})) \tag{14}$$

適応型 RBM のパラメータ $\Theta = \{\bar{\mathbf{W}}, \bar{\boldsymbol{b}}, \bar{\boldsymbol{c}}, \bar{\boldsymbol{\sigma}}, \mathcal{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}\}$ は，従来の RBM と同様に，N 個の観測データを $\{\boldsymbol{v}_n, \boldsymbol{s}_n\}_{n=1}^N$ とするとき，この確率変数の対数尤度 $\mathcal{L} = \log \prod_n p(\boldsymbol{v}_n|\boldsymbol{s}_n) = \log \prod_n \sum_{\boldsymbol{h}} p(\boldsymbol{v}_n, \boldsymbol{h}_n|\boldsymbol{s}_n)$ 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 $\Theta = \{\bar{\mathbf{W}}, \mathbf{A}_r, \mathbf{B}, \mathbf{C}, \mathbf{D}\}$ で偏微分すると，

$$\frac{\partial \mathcal{L}}{\partial \bar{\mathbf{W}}} = \langle \sum_r \mathbf{A}_r^{\mathrm{T}} \boldsymbol{v}' \boldsymbol{h}^{\mathrm{T}} s_r \rangle_{\mathrm{data}} - \langle \sum_r \mathbf{A}_r^{\mathrm{T}} \boldsymbol{v}' \boldsymbol{h}^{\mathrm{T}} s_r \rangle_{\mathrm{model}} \tag{15}$$

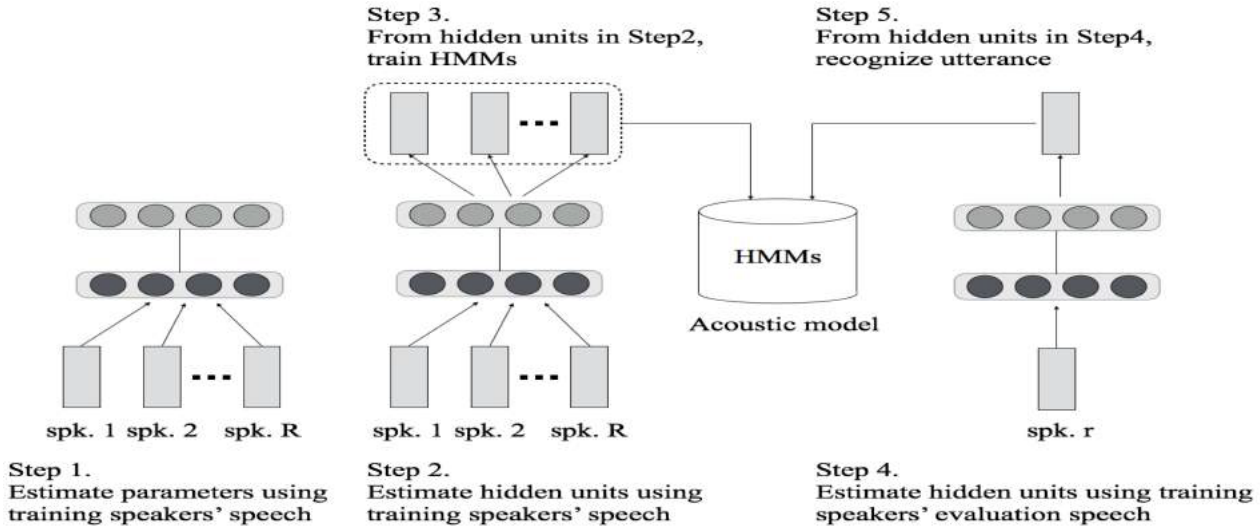

図 2　適応型 RBM を用いた音声認識の流れ

Fig. 2　Procedure of speech recognition using an ARBM. "spk." indicates speaker.

$$\frac{\partial \mathcal{L}}{\partial \mathbf{A}_r} = \langle \boldsymbol{v}' \boldsymbol{h}^{\mathrm{T}} \bar{\mathbf{W}}^{\mathrm{T}} s_r \rangle_{\mathrm{data}} - \langle \boldsymbol{v}' \boldsymbol{h}^{\mathrm{T}} \bar{\mathbf{W}}^{\mathrm{T}} s_r \rangle_{\mathrm{model}}, \tag{16}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{B}} = \langle \boldsymbol{v}' \boldsymbol{s}^{\mathrm{T}} \rangle_{\mathrm{data}} - \langle \boldsymbol{v}' \boldsymbol{s}^{\mathrm{T}} \rangle_{\mathrm{model}}, \tag{17}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{C}} = \langle \boldsymbol{h} \boldsymbol{s}^{\mathrm{T}} \rangle_{\mathrm{data}} - \langle \boldsymbol{h} \boldsymbol{s}^{\mathrm{T}} \rangle_{\mathrm{model}}, \tag{18}$$

$$\frac{\partial \mathcal{L}}{\partial d_i} = \frac{1}{\sigma_i^2(\boldsymbol{s})} \langle \sum_r \{ \frac{(v_i - b_i)^2}{2} - v_i \boldsymbol{W}_{i:} \boldsymbol{h} \} s_r \rangle_{\mathrm{data}} - \frac{1}{\sigma_i^2(\boldsymbol{s})} \langle \sum_r \{ \frac{(v_i - b_i)^2}{2} - v_i \boldsymbol{W}_{i:} \boldsymbol{h} \} s_r \rangle_{\mathrm{model}} \tag{19}$$

ここで，$\boldsymbol{v}' = \frac{\boldsymbol{v}}{\boldsymbol{\sigma}^2(\boldsymbol{s})^2}$ とした．他のパラメータ $\Theta = \{\bar{\boldsymbol{b}}, \bar{\boldsymbol{c}}, \bar{\boldsymbol{\sigma}}\}$ に関しては，それぞれ式 (5)(6)(7) と同様にして求められる．適応型 RBM においても，CD 法を適用できるため，観測データの期待値をモデルの期待値として計算することで，効率よくパラメータを推定することができる．

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，Fig. 2 のようにまず複数 (R 人) の学習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，学習話者のデータに対して，次式のように潜在特徴量 (隠れ素子) の期待値を計算する．

$$\hat{\boldsymbol{h}}_n = \mathbb{E}_{p(\boldsymbol{h}|\boldsymbol{v}_n, \boldsymbol{s}_n)}[\boldsymbol{h}] = \mathcal{S}(\boldsymbol{c}(\boldsymbol{s}_n) + \left(\frac{\boldsymbol{v}_n}{\boldsymbol{\sigma}^2(\boldsymbol{s})} \right)^2 \mathbf{W}(\boldsymbol{s}_n)) \tag{20}$$

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．学習話者の発話を用いて認識を行なう場合，話者に依存しないパラメータと認識話者の適応パラメータを用いて式 (20) から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

本章ではまず，構音障害者音声の分析を行なう．従来の特徴量である MFCC (mel-frequency cepstrum coefficient) を使用し，健常者音声から作成された不特定話者音響モデルを用いた認識実験を行なう．さらに，MLLR による話者適応を行い，認識性能を評価する．次に，多次元情報を 2 次元平面へ可視化する COSMOS (acoustic space map of sound) 法 [19] を用いた音響空間の分析を行なう．最後に，音響空間の近い話者を用いて提案手法の評価を行なう．

4.1　従来法による音声認識

4.1.1　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 [20] 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 MG02, MM03, MK04 は各単語 5 発話分，話者 MK01, MY05, MN06 は各単語 3 発話分の音声データを収録した．音声の標本化周波数は 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．音響特徴量には，MFCC12 次元とその動的特徴量を用いた．さらに，収録雑音の軽減のため，CMN (cepstrum mean normalization) を施した．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．

学習データとして，評価話者以外の話者の全ての音声を用いた．適応データ及び評価データとして，評価話者の第 1 発話及び第 5 発話を選択した．

4.1.2　実験結果と考察

Fig. 3 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 MY05，MN06 の精度が著しく低い理由として，音声が含む無音区間の長さが考えられる．この 2 話者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，この 2 話者の音声は無音区間をあまり含んでいない．そのため，CMN を施した際に，平均の値が他の話者のものと大きく異なったと考えられる．MLLR による話者適応を行なうと，

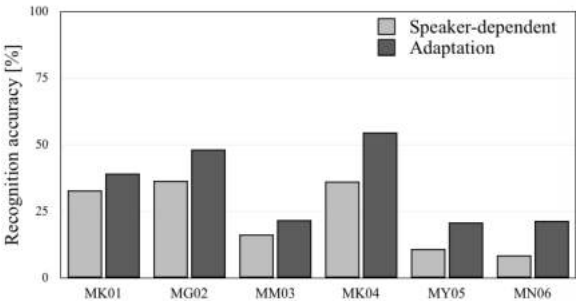

図 3　不特定話者モデルを用いた音声認識実験結果

Fig. 3　Word recognition accuracy [%] for each speaker using the speaker-dependent model.

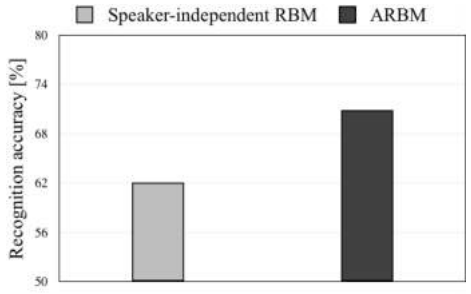

図 5　話者 MG02 に対する RBM と ARBM を用いた単語認識実験結果

Fig. 5　Word recognition accuracy [%] for speaker MG02 using an RBM and an ARBM.

何れの話者も精度が改善された．

4.1.3　音響空間の可視化

認識精度に影響を与える要因として，話者毎の持つ音響空間の距離が挙げられる．例えば，ある特定話者音響モデルがある時に，その話者と近い音響空間を持つ話者の音声の認識精度は高いが，遠い空間の話者の音声は認識が難しくなることが考えられる．話者適応をする際にも，音響空間の上で距離の近い話者への適応は容易だが，遠い距離の話者への適応は困難であることが考えられる．そこで，本節では COSMOS (acoustic space map of sound) 法 [19] を用いた音響空間の分析を行う．

COSMOS 法は，多次元情報を低次元空間に可視化する方法，すなわち多次元尺度構成法 (multi dimensional scaling; MDS) の一つであり，HMM による音響モデルの集合を 2 次元平面上に非線形写像するように，Sammon 法 [21] を拡張したものである．Sammon 法は，高次元空間上の高次元情報の相互距離の総和と低次元空間上の写像位置座標の相互ユークリッド距離の総和が最小となるように，最急降下法により低次元空間上の写像位置座標を最適化する非線形写像手法である．高次元空間上での相互距離関係を低次元空間上でも維持しながら，全ての高次元情報を低次元空間に射影する．

ATR 研究用日本語データベース (A セット) [20] から選択した男性話者 10 名，構音障害者男性 6 名の音声から作成された特定話者音響モデルから作成された COSMOS を Fig. 4 に示す．

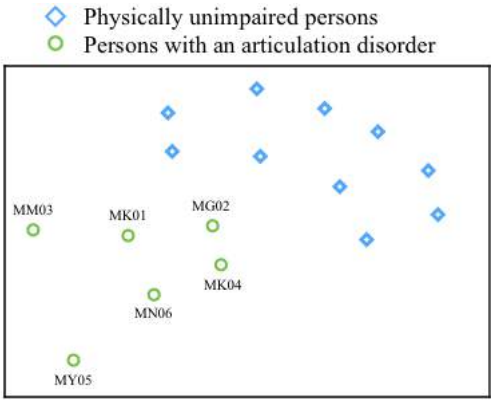

図 4　話者 COSMOS

Fig. 4　COSMOS for each speaker.

健常者と障害者がそれぞれ明瞭なクラスタを形成していることから，それぞれのモデルには明確な隔たりがあることが分かる．不特定話者モデルを用いた認識実験で比較的高い精度が得られた話者は，COSMOS 上でも近い位置に，低い精度が得られた話者は遠い位置にあり，認識結果の妥当性を裏付けている．

4.2　提案手法による音声認識

4.2.1　実 験 条 件

前節の実験より，音響空間で距離の近い話者 MK01, MG02, MY06 を使用する．話者 MK01, MG02, MY06 の第 1 発話から第 3 発話を学習データとして使用する．本稿では，学習話者に対する認識性能を評価するため，MG02 の未知発話である第 5 発話を評価データとする．

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．また，潜在特徴量の数を 32 とした．学習率 0.005，モーメント係数 0.9，バッチサイズ 512，繰り返し回数 200 の確率的勾配法を用いてモデルを学習した．

4.2.2　実験結果と考察

Fig. 5 に提案手法による音声認識実験結果を示す．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，提案手法は 8.79%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．

Fig. 6 に，話者 MG02 の発話/ikioi/のスペクトルグラム，適応型 RBM への入力特徴量，推定された隠れ素子，及び，隠れ素子から再構築された入力特徴量を示す．推定された隠れ素子から，比較的高い再現度で入力特徴量を再構築できていることが分かる．発話/ikioi/には，/i/が 3 回含まれているが，推定された隠れ素子にそれが表現されているとは言い難い．同じ/i/にも様々なパターンがあり，1 つの基底で表現しきれないためだと考えられる．隠れ層が音韻情報をより忠実に制御するためには，さらなる改良が必要だと考えられる．

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 を用いた音声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

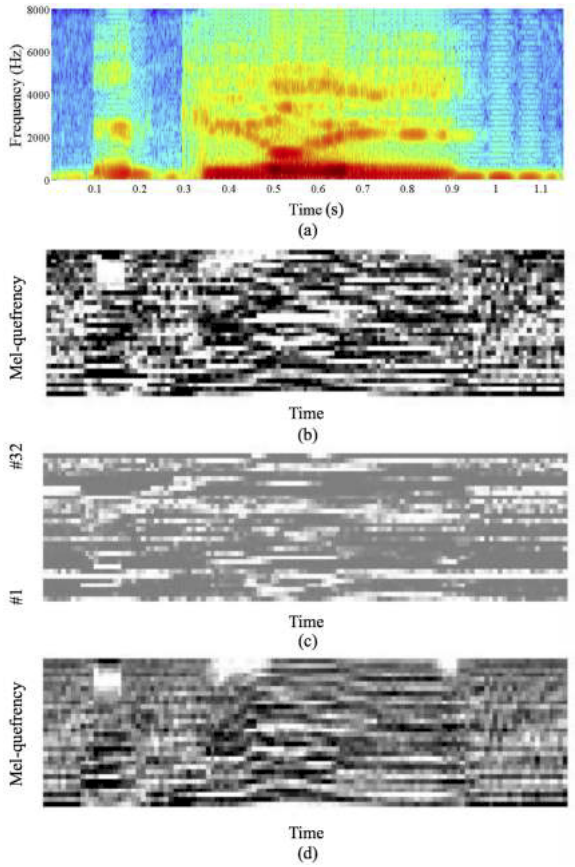

図 6　発話/ikioi/から抽出された特徴量を示す．(a) スペクトルグラム, (b) 適応型 RBM への入力特徴量 $\boldsymbol{v}$, (c) 推定された隠れ素子 $p(\boldsymbol{h}|\boldsymbol{v},\boldsymbol{s})$, (d) 隠れ素子から再構成された入力特徴量 $p(\boldsymbol{v}|\hat{\boldsymbol{h}},\boldsymbol{s})$. 下の 3 つに図において，白色は高い確率値，黒色は低い確率値を示す．

Fig. 6　Examples of extracted feature using an utterance /ikioi/. (a) Spectrogram by a person with an articulation disorder, (b) the input feature $\boldsymbol{v}$ for an ARBM, (c) the probability distribution of hidden units $p(\boldsymbol{h}|\boldsymbol{v},\boldsymbol{s})$ given the input feature, (d) the reconstructed feature $p(\boldsymbol{v}|\hat{\boldsymbol{h}},\boldsymbol{s})$. In the three figure below, the white and the black indicate the high and the low probability, respectively.

評価を行なった．評価実験において，学習話者に対して提案手法の有効性を示した．今後は，未知話者に対して認識性能の評価を行なう．

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．

文　献

[1] 厚生労働省，“平成 27 年度福祉行政報告例”.
[2] K.-i. Yabu, T. Ifukube, and S. Aomura, “A speech synthesis method using a pointing device : As a speaking aid for speech disorders,” The journal of Japan Academy of Health Sciences, vol.12, no.1, pp.49–57, jun 2009.
[3] T. KOYAMA and T. SAITOH, “Efficient features for sign language word recognition using kinect,” IEICE technical report. ME and bio cybernetics, vol.114, no.408, pp.117–120, Jan 2015.
[4] X. Menndez-Pidal, J.B. Polikoff, S.M. Peters, J.E. Leonzio, and H.T. Bunnell, “The nemours database of dysarthric speech.,” ICSLP, pp.1962–1965, ISCA, 1996.
[5] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T.S. Huang, K. Watkin, and S. Frame, “Dysarthric speech database for universal access research.,” INTERSPEECH, pp.1741–1744, ISCA, 2008.
[6] F. Rudzicz, A.K. Namasivayam, and T. Wolff, “The torgo database of acoustic and articulatory speech from speakers with dysarthria.,” Language Resources and Evaluation, vol.46, no.4, pp.523–541, 2012.
[7] R. Ueda, T. Takiguchi, and Y. Ariki, “Individuality-preserving voice reconstruction for articulation disorders using text-to-speech synthesis.,” ICMI, pp.343–346, ACM, 2015.
[8] M.J.F. Gales, “Maximum likelihood linear transformations for hmm-based speech recognition,” Computer Speech & Language, vol.12, no.2, pp.75–98, 1998.
[9] M.J. Kim, J. Wang, and H. Kim, “Dysarthric speech recognition using Kullback-Leibler divergence-based hidden Markov model.,” INTERSPEECH, ed. by N. Morgan, pp.2671–2675, ISCA, 2016.
[10] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol.86, no.11, pp.2278–2324, 1998.
[11] T. Nakashika, T. Takiguchi, Y. Ariki, S. Duffner, and C. Garcia, “Dysarthric speech recognition using a convolutive bottleneck network,” ICSP, pp.505–509, 2014.
[12] T. Nakashika, T. Takiguchi, and Y. Minami, “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.
[13] Y. Freund and D. Haussler, “Unsupervised learning of distributions of binary vectors using 2-layer networks.,” NIPS, eds. by J.E. Moody, S.J. Hanson, and R. Lippmann, pp.912–919, Morgan Kaufmann, 1991.
[14] M. Ranzato, A. Krizhevsky, and G.E. Hinton, “Factored 3-way restricted boltzmann machines for modeling natural images.,” AISTATS, eds. by Y.W. Teh and D.M. Titterington, vol.9, pp.621–628, JMLR Proceedings, JMLR.org, 2010.
[15] G.E. Dahl, M. Ranzato, A. rahmanMohamed, and G.E. Hinton, “Phone recognition with the mean-covariance restricted boltzmann machine.,” NIPS, eds. by J.D. Lafferty, C.K.I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, pp.469–477, Curran Associates, Inc., 2010.
[16] M. Ranzato and G.E. Hinton, “Modeling pixel means and covariances using factorized third-order boltzmann machines.,” CVPR, pp.2551–2558, IEEE Computer Society, 2010.
[17] A.L. K. Cho and T. Raiko, “Improved learning of Gaussian-Bernoulli restricted Boltzmann machines,” Artificial Neural Networks and Machine Learning, pp.10–17, 2011.
[18] G.E. Hinton, S. Osindero, and Y.-W. Teh, “A fast learning algorithm for deep belief nets,” Neural Comput., vol.18, no.7, pp.1527–1554, July 2006.
[19] M. SHOZAKAI and G. NAGINO, “Two-dimensional visualization of acoustic space by multidimensional scaling,” 情報処理学会研究報告, vol.109，pp.129–136, jul 2004.
[20] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, “ATR Japanese speech database as a tool of speech recognition and synthesis,” Speech Communication, vol.9, no.4, pp.357–363, 1990.
[21] J. Sammon, “A nonlinear mapping for data structure analysis,” IEEE Transactions on Computers, vol.18, pp.401–409, 1969.