Factored 3-Way Restricted Boltzmann Machine を用いたマルチモーダル音声認識の検討*

☆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．本研究は，コミュニケーション手段として口話を用いる重度難聴者を対象として，音声と唇形状によるマルチモーダル音声認識を実現し，ユビキタス社会における彼らの生活の支援をすることを目的としている．

人間は発話内容を理解する際，種々の情報を統合的に利用している．音声が聞き取り難い場合，発話者の顔，特に唇の動きに注目して発話内容を理解しようとし，逆に，唇の動きと音声が不一致の場合，唇の動きに影響されて発話内容を誤って理解してしまうこともある．これは，McGurk effect(マガーク効果) と呼ばれ，音韻知覚が音声の聴覚情報のみで決まるのではなく，唇の動きといった視覚情報からも影響を受けることが報告されている [3]．このように人間による発話内容の理解には，唇の画像と音声の情報の統合的利用が極めて重要である．

唇の動きからの発話内容の読み取りは，リップリーディング (読唇) と呼ばれ，聴覚障害者にとって重要なコミュニケーション手段の一つである．リップリーディングは，背景雑音に影響されることがないため，計算機上での実現が期待されている．例えば，監視カメラに収録された会話映像のように音声が聞き取りにくい場合であっても，リップリーディングであれば発話内容の分析が可能であり，犯罪の防止や抑止に繋がると考えられる．そのため，音声の雑音に対して頑健な発話認識を行う手法の一つとして，音声情報に唇動画像情報を併用して認識を行うマルチモーダル音声認識が注目され，研究が進められている [4, 5, 6]．

重度難聴者は耳で音を聞くことができないため，正確な発音をすることが難しく，発話スタイルが健常者と異なる．彼らのコミュニケーション手段の一つとして口話があり，訓練により意図した発話の唇の形状を作ることが可能である．そこで，彼らの音声を認識するために，唇画像を併用した音声認識システムの構築が望まれる．重度難聴者を対象としたマルチモーダル音声認識として，CNN (convolutional neural network) を用いた手法 [7] が提案されている．しかし，この手法はニューラルネットワークを用いているためパラメータ数が多く，比較的大量のデータを必要とする．また，音声と唇画像のモダリティギャップが考慮されていない．

本研究では，音声と唇画像を統合した特徴量抽出のために，factored 3-way restricted Boltzmann machine (F3WRBM [8]) を用いる．このモデルは RBM (restricted Boltzmann machine [9]) 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．近年，RBM を用いた特徴抽出 (特徴学習) 法 [8, 10, 11] 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．F3WRBM は 2 つの観測変数と 1 つの潜在変数からなるモデルであり，音響特徴量と画像特徴量を観測変数とすることで，音声と唇画像を相補的に考慮できる特徴が潜在変数として現れると期待できる．

2 Restricted Boltzmann Machine

RBM は，Fig. 1 左図に示すように，可視素子 $\boldsymbol{v} \in \mathbb{R}^D$ と隠れ素子 $\boldsymbol{h} \in \mathbb{B}^H$ ($\mathbb{B}$ 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 [12](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．

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

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

ここで，$[b_1, \cdots, b_D] \in \mathbb{R}^D$，$[c_1, \cdots, c_H] \in \mathbb{R}^H$，$\{W_{ij}\}_{D \times H} \in \mathbb{R}^{D \times H}$，$[\sigma_1^2, \cdots, \sigma_D^2] \in \mathbb{R}^D$ 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クスである．ここで，

*Multi-modal speech recognition using factored 3-way restricted Boltzmann machine, by Yuki Takashima (Kobe University), Toru Nakashika (UEC), Tetsuya Takiguchi (Kobe University/JST PRESTO), Yasuo Ariki (Kobe University)

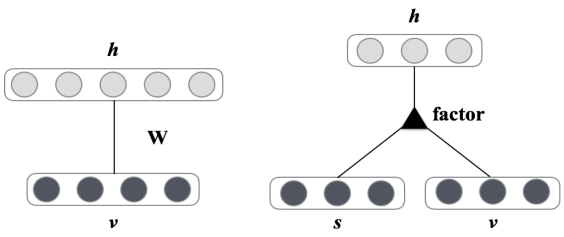

Fig. 1 Graphical representation of an RBM (left) and factored 3-way RBM (right)

$Z = \int^D \Sigma_{\boldsymbol{h}} \mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h})} d^D\boldsymbol{v}$ 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，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．

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

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

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 $\{\boldsymbol{v}_n\}_{n=1}^N$ 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．この対数尤度をパラメータ θ で偏微分すると，

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

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 [13] 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 提案手法

F3WRBM はバイナリな可視素子を仮定し定義されている [8]．本節では，可視素子が正規分布に従うと仮定し，モデルを再定義する．さらに，F3WRBM を用いた特徴量抽出と，音声認識への応用について述べる．

3.1 Factored 3-way RBM

RBM は 2 つの確率変数からなるモデルであるが，これは一般的により高次なモデルへと拡張が可能である [14]．本研究では，音響特徴量を表す $\boldsymbol{s} = [s_1,\cdots,s_D] \in \mathbb{R}^D$，画像特徴量を表す $\boldsymbol{v} = [v_1,\cdots,v_L] \in \mathbb{R}^L$，潜在特徴量を表す $\boldsymbol{h} = [h_1,\cdots,h_H] \in \mathbb{B}^H$ の 3 変数間の関係性を 3-way RBM を用いて表現する．このとき，3 つの確率変数の結合を表現するエネルギー関数は以下のように定義される．

$$E_{3WRBM}(\boldsymbol{s},\boldsymbol{v},\boldsymbol{h}) = -\sum_{i,j,k} s_i v_j h_k W_{ijk} \quad (5)$$

ここで，i，j，k はそれぞれ $\boldsymbol{s},\boldsymbol{v},\boldsymbol{h}$ のインデクスである．$\{W_{ijk}\}_{D\times L\times H} \in \mathbb{R}^{D\times L\times H}$ は 3 階のテンソルで表現されており，非常に膨大なパラメータ数となる．そこで，3 つの確率変数の間に “ファクター” と呼ばれる概念を設け，結合行列を $W_{ijk} = \sum_f C_{if}K_{jf}P_{kf}$ と近似する．ここで，$\{C_{if}\}_{D\times F} \in \mathbb{R}^{D\times F}$，$\{K_{jf}\}_{L\times F} \in \mathbb{R}^{L\times F}$，$\{P_{kf}\}_{H\times F} \in \mathbb{R}^{H\times F}$ はそれぞれ，音響特徴量-ファクター，画像特徴量-ファクター，ファクター-潜在特徴量の間の結合行列であり，F はファクターの数である．これにより，パラメータの数を削減することができ，式 (5) のエネルギー関数は以下のように表現される．

$$\begin{aligned} &-E_{F3WRBM}(\boldsymbol{s},\boldsymbol{v},\boldsymbol{h}) \\ &= \sum_f (\sum_i s_i C_{if})(\sum_j v_j K_{jf})(\sum_k h_k C_{kf}) \end{aligned} \quad (6)$$

音響特徴量及び画像特徴量をファクター空間へ射影してから結合するため，このファクター空間が 2 つのモダリティギャップを埋める空間の役割を果たすことが期待される． (Fig. 1 右図) さらに，本稿では音響特徴量と画像特徴量の確率変数が正規分布に従うと仮定し，エネルギー関数を以下のように定義する．

$$\begin{aligned} &-E(\boldsymbol{s},\boldsymbol{v},\boldsymbol{h}) = \\ &-\sum_i \frac{(s_i-b_i)^2}{2\sigma_i^2} - \sum_j \frac{(v_j-d_j)^2}{2\delta_j^2} + \sum_k c_k h_k + \\ &\sum_f \left(\sum_i \frac{s_i}{\sigma_i^2} C_{if}\right)\left(\sum_j \frac{v_j}{\delta_j^2} K_{jf}\right)\left(\sum_k h_k C_{kf}\right) \end{aligned} \quad (7)$$

ここで，$[b_1,\cdots,b_D] \in \mathbb{R}^D$，$[d_1,\cdots,d_D] \in \mathbb{R}^L$ はそれぞれ $\boldsymbol{s}$，$\boldsymbol{v}$ に関するバイアス，$[\sigma_1^2,\cdots,\sigma_D^2] \in \mathbb{R}^D$，$[\delta_1^2,\cdots,\delta_L^2] \in \mathbb{R}^L$ はそれぞれ $\boldsymbol{s}$，$\boldsymbol{v}$ に関する分散を表す．RBM と同様に，それぞれの変数の要素同士には結合が存在しないと仮定している．そのため，$\boldsymbol{s}$，$\boldsymbol{v}$，$\boldsymbol{h}$ の条件付き確率はそれぞれ以下のように簡単に計算することができる．

$$\begin{aligned} &p(s_i = s|\boldsymbol{v},\boldsymbol{h}) \\ &= \mathcal{N}\big(s|b_i + \sum_f C_{if}(\sum_j \frac{v_j}{\delta_j^2}K_{jf})(\sum_k h_k P_{kf}), \sigma_i^2\big) \end{aligned} \quad (8)$$

$$\begin{aligned} &p(v_j = v|\boldsymbol{s},\boldsymbol{h}) \\ &= \mathcal{N}\big(v|d_i + \sum_f K_{jf}(\sum_i \frac{s_i}{\sigma_i^2}C_{if})(\sum_k h_k P_{kf}), \delta_j^2\big) \end{aligned} \quad (9)$$

$$p(h_k=1|\boldsymbol{s},\boldsymbol{v}) = \mathcal{S}\big(c_k+\sum_f P_{kf}(\sum_i \frac{s_i}{\sigma_i^2}C_{if})(\sum_j \frac{v_j}{\delta_j^2}K_{jf})\big) \quad (10)$$

3.2 パラメータ推定

提案モデルのパラメータは，同期の取れたNフレームの音声データと画像データに対する対数尤度

$$\mathcal{L}' = \log\prod_n p(\boldsymbol{s}_n,\boldsymbol{h}_n) = \sum_n \log\sum_{\boldsymbol{h}} p(\boldsymbol{s}_n,\boldsymbol{h}_n,\boldsymbol{v}_n) \quad (11)$$

を最大化するように同時に推定することが可能である．本研究では確率的勾配法を用いて各パラメータを更新する．式 (11) を各パラメータ θ に関して偏微分を計算すると以下の式が得られる．

$$\frac{-\partial E(\boldsymbol{s},\boldsymbol{v},\boldsymbol{h})}{\partial C_{i'f'}} = \frac{s_{i'}}{\sigma_{i'}^2}(\sum_j \frac{v_j}{\delta_j^2}K_{jf'})(\sum_k h_k P_{kf'})$$

$$\frac{-\partial E(\boldsymbol{s},\boldsymbol{v},\boldsymbol{h})}{\partial P_{k'f'}} = h_{k'}(\sum_i \frac{s_i}{\sigma_i^2}C_{if'})(\sum_j \frac{v_j}{\delta_j^2}K_{jf'})$$

$$\frac{-\partial E(\boldsymbol{s},\boldsymbol{v},\boldsymbol{h})}{\partial b_{i'}} = \frac{s_{i'}-b_{i'}}{\sigma_{i'}^2}$$

$$\frac{-\partial E(\boldsymbol{s},\boldsymbol{v},\boldsymbol{h})}{\partial c_{k'}} = h_{k'}$$

$K_{j'f'}$，$d_{j'}$ に関しても同様である．分散パラメータは学習の安定化と非負制約のため，$\sigma_i^2=e^{z_i}$，$\delta_j^2=e^{p_j}$ とおき，それぞれ z_i，p_j で偏微分しパラメータの更新を行う．

$$\frac{-\partial E(\boldsymbol{s},\boldsymbol{v},\boldsymbol{h})}{\partial z_{i'}} = \mathrm{e}^{-z_{i'}}\Big(\frac{1}{2}(s_{i'}-b_{i'})^2 - s_{i'}\big(\sum_f C_{i'f}(\sum_j \frac{v_j}{\delta_j^2}K_{jf})(\sum_k h_k P_{kf})\big)\Big)$$

p_j に関しても同様である．本稿では，$\boldsymbol{h}$，$\boldsymbol{v}$，$\boldsymbol{s}$ を順にサンプリングする．例えば $\boldsymbol{h}$ のサンプリングでは，式 (10) を用いて，$\widetilde{\boldsymbol{h}}\sim p(\boldsymbol{h}|\boldsymbol{s},\boldsymbol{v})$ とすることでサンプル $\widetilde{\boldsymbol{h}}$ を得る．このようにすることで，既知の特徴量 $\boldsymbol{s},\boldsymbol{v}$ から $\widetilde{\boldsymbol{h}}\sim p(\boldsymbol{h}|\boldsymbol{s},\boldsymbol{v}),\widetilde{\boldsymbol{v}}\sim p(\boldsymbol{v}|\boldsymbol{s},\widetilde{\boldsymbol{h}}),\widetilde{\boldsymbol{s}}\sim p(\boldsymbol{s}|\widetilde{\boldsymbol{v}},\widetilde{\boldsymbol{h}}),\widetilde{\widetilde{\boldsymbol{h}}}\sim p(\boldsymbol{h}|\widetilde{\boldsymbol{s}},\widetilde{\boldsymbol{v}})\cdots$ と Gibbs チェインを繋げていくことができる．

3.3 F3WRBM による特徴量抽出

本研究では，F3WRBM を用いて音声情報と画像情報を統合した特徴量の抽出を行う．具体的には，音響特徴量 $\boldsymbol{s}$ と画像特徴量 $\boldsymbol{v}$ から式 (10) を用いて潜在特徴量 $\boldsymbol{h}$ を推定し，これを音響モデル HMM (hidden Marcov model) の入力特徴量とする．一般的なマルチモーダル音声認識手法として，結果統合やマルチストリーム HMM が用いられているが，統合時の重みの調節や計算コストが大きいという問題点がある．本手法では，特徴量抽出器において 2 つのモダリティの統合を行うため，音響モデルは単純な構成で済み計算コストを抑えることができる．また，パラメータを教師なし学習により学習するため，音声と画像の相補的な関係が自動的に学習できると期待される．

4 評価実験

4.1 実験条件

データセットとして，重度難聴者の男性 1 名の音声及び唇画像を収録し使用した．発話内容は，ATR 音素バランス単語 A セット [15] から選択した．F3WRBM 及び音響モデルの学習データとして 2620 単語，評価データとして 216 単語を用いた．重度難聴者の発話スタイルは，健聴者の発話スタイルと大きく異なるため，特定話者モデルにより認識を行う．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，5ms 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．

F3WRBM への入力音響特徴量として 39 次元のメル周波数スペクトル，入力画像特徴量として唇画像に対して DCT を行い上位 30 次元を用いた．また，ファクター数を 96，潜在特徴量の数を 196 とした．学習率 0.001，モーメント係数 0.9，バッチサイズ 128，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 実験結果と考察

Fig. 2 にマルチモーダル音声認識結果を示す．ベースラインとして，従来の音響特徴量 MFCC(12 次元)+ΔMFCC と画像特徴量 DCT(30 次元) を連結して HMM の入力とする初期統合を行った (”MFCC+Δ+DCT”)．また，RBM を用いた単純な統合方法として，メル周波数スペクトル (39 次元) と DCT(30 次元) を連結して RBM を学習し，潜在特徴量を新たな特徴量として認識する手法を行った (”RBM”)．この RBM は潜在特徴量の数を 128 とし，PCA (principal component analysis) を行い 60 次元に圧縮して認識に用いた．”F3WRBM”が提案手法により抽出した特徴量を用いたものであり，これも潜在特徴量に対し PCA を行い，60 次元に圧縮して用いている．Fig. 2 より，提案手法は他の手法より精度が大きく劣化していることが分かる．これは，提案モデルが音声と唇画像を対等に扱っているためだと考えられる．一般に，クリーン環境下においては，音響特徴量と唇画像特徴量を用いて認識を行うと，音響

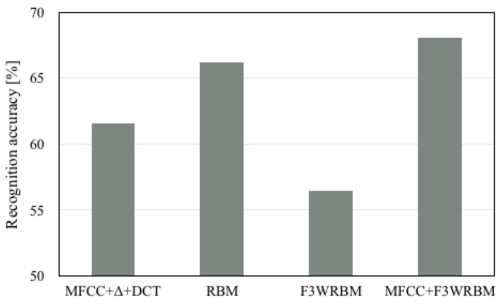

Fig. 2　Word recognition accuracies of each feature.

特徴量のみを用いて認識を行う場合と比べて精度は劣化し，唇画像は逆にノイズとして影響してしまう．しかし，唇画像も認識に好影響な特徴を持っているはずであり，提案手法はその相補的な関係を学習し，特徴量に反映できると期待したが，本稿のモデルではその効果は見られなかった．原因として，学習された3つの結合重みの学習の程度にばらつきがあり，全体として十分な学習が行えていないことが考えられる．そこで，各結合重みに対して，何らかの正規化を制約として加えることで，安定した学習が行えるのではないかと考えられる．

次に，MFCC と提案手法で得られた特徴量 (”F3WRBM”) を連結して認識を行った (”MFCC+F3WRBM”)．動的特徴量を含むベースラインよりも良い精度が得られていることから，提案手法により得られた特徴量は従来の特徴量にはない特徴が表現されていると考えられる．

5　おわりに

本稿では，factored 3-way RBM を用いて，音声特徴量と画像特徴量を統合した特徴量抽出法を提案した．重度難聴者の男性 1 名の音声と唇画像を用いたマルチモーダル音声認識実験により，提案手法の評価を行った．教師なし学習により，音声と画像の相補的な関係のモデル化を期待したが，評価実験において，提案手法で得られた特徴量のみを用いた場合は良い精度が得られず，効果を確認できなかった．本研究では，パラメータの学習時に制約を入れていないため，3つある結合行列のうち，ファクターと潜在変数間の結合行列に学習が偏る傾向が見られた．今後は，それぞれの結合行列がバランス良く学習できる正規化を導入し，精度の改善を図りたい．

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] H. McGurk and J. MacDonald, “Hearing lips and seeing voices,” Nature, vol. 264, pp. 746–748, 1976.
[4] G. Potamianos *et al.*, “Audio-visual automatic speech recognition: An overview,” Issues in visual and audio-visual speech processing, vol. 22, pp. 23, 2004.
[5] G. Potamianos *et al.*, “Recent advances in the automatic recognition of audiovisual speech,” Proceedings of the IEEE, vol. 91, no. 9, pp. 1306–1326, 2003.
[6] Y. Mroueh *et al.*, “Deep multimodal learning for audio-visual speech recognition.,” in *ICASSP*, 2015, pp. 2130–2134.
[7] Y. Takashima *et al.*, “Audio-visual speech recognition using convolutive bottleneck networks for a person with severe hearing loss,” IPSJ Transactions on Computer Vision and Applications, vol. 7, pp. 64–68, 2015.
[8] M. Ranzato *et al.*, “Factored 3-way restricted boltzmann machines for modeling natural images.,” in *AISTATS*, Y. W. Teh and D. M. Titterington, Eds. 2010, vol. 9 of *JMLR Proceedings*, pp. 621–628, JMLR.org.
[9] Y. Freund and D. Haussler, “Unsupervised learning of distributions on binary vectors using two layer networks,” Tech. Rep., 1994.
[10] 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.
[11] M. Ranzato and G. E. Hinton, “Modeling pixel means and covariances using factorized third-order boltzmann machines.,” in *CVPR*. 2010, pp. 2551–2558, IEEE Computer Society.
[12] 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.
[13] G. E. Hinton *et al.*, “A fast learning algorithm for deep belief nets,” Neural Comput., vol. 18, no. 7, pp. 1527–1554, July 2006.
[14] T. J. Sejnowski, “Higher-order boltzmann machines,” in *AIP Conference Proceedings*, 1986, vol. 151, pp. 398–403.
[15] 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.