2L2-OS-09a-2

クラスタ適応制限ボルツマンマシンを用いた話者クラスタリングと声質変換への応用

Automatic speaker clustering using cluster-adaptive restricted Boltzmann machine and its application to voice conversion

中鹿　亘，南　泰弘 *1
NAKASHIKA Toru, MINAMI Yasuhiro

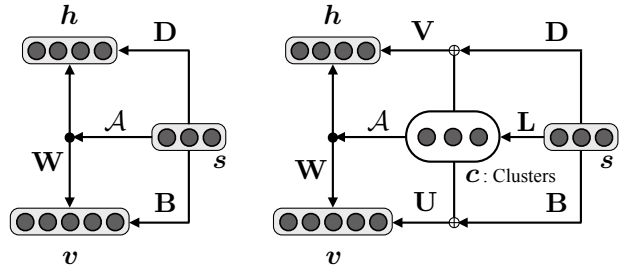
*1電気通信大学
The University of Electro-Communications

In this paper, a new energy-based probabilistic model, called CAB (Cluster Adaptive restricted Boltzmann machine), is proposed for voice conversion (VC) that does not require parallel data during the training and requires small amount of speech data during the adaptation. Most of the existing VC methods require parallel data in the training. Recently, VC methods that do not require parallel data have been also proposed, and garnering much attention because they have attractive advantages of no need to use parallel speech corpora. The proposed CAB is aimed for statistical non-parallel VC based on cluster adaptive training (CAT), where speaker identities will be represented as cluster vectors that determine the adaptation matrix that projects bidirectional weights of RBM. Since the number of clusters is generally smaller than the number of speakers, we can reduce the number of parameters, which enables speaker adaptation with small amount of data.

1. はじめに

これまで声質変換（入力話者音声の音韻情報を保存したまま，話者性に関する情報のみを出力話者のものへ変換させる技術）の分野において，パラレルデータ（入力話者と目標話者の，同一発話内容による音声対）を使用するアプローチ（パラレル声質変換）が主流であり，GMM に基づく手法 [d] など，様々な統計的アプローチが提案されてきたが，近年モデルの学習時にパラレルデータを使用しないアプローチ（非パラレル声質変換）が注目を浴びている [a, c]．何故ならば，パラレル声質変換では学習データの入力・目標話者の発話内容が一致している必要があるため利便性が損なわれてしまう一方，非パラレル声質変換では自由発話を用いて学習を行うことができるため，利便性や実用性が格段に向上するからである．我々が提案した，統計的な非パラレル声質変換アプローチである ARBM (adaptive restricted Boltzmann machine) に基づく声質変換 [c] では，複数の話者による音声データから自動的にそれぞれの話者固有の適応行列と，音響特徴量（メルケプストラム）から話者に依存しない潜在特徴（以降，潜在的な音韻または単に音韻と呼ぶ）への射影行列を同時推定することにより，入力話者の音声および入力話者の適応行列から計算した潜在的な音韻と，目標話者の適応行列を用いて音響特徴量を計算することで目標話者に近い音声を得る．また一度学習によって潜在的な音韻を得るための射影行列が推定されれば，新たな入力話者・目標話者に対してそれぞれの適応行列のみを推定（このステップを適応と呼ぶ）することで変換が可能となる．しかし，話者固有の適応行列は音響特徴量の二乗個のパラメータを含むため，音響特徴量の次元数や話者数が増えるほどパラメータ数が膨大となり，学習コストが掛かってしまう上，適応時に必要となるデータ数が多くなり，事前に学習していない話者のその場での変換が困難となってしまう．そこで本研究では，話者クラスタ学習 [b] に着目し，各話者の発話について少ないデータ数で適応可能な話者クラスタ適応 RBM (Cluster Adaptive restricted Boltzmann machine; CAB) を用いた統計的非パラレル声質変換手法を提案する．本モデルによって自動的に形成されるクラスタは，いくつかの話者をまとめて表現した「代表話者」のようなものを表し，例えば「男らしい声」「明るい声」「籠もった声」など，それぞれ性質の異なる声を表現する．変換時において代表話者を適切に設定することで，例えば「少し籠もった男性の声」など任意の声を新たに生み出すこともできると期待される．

連絡先: 中鹿亘，電気通信大学，〒 182-8585 東京都調布市調布ヶ丘１丁目５－１，042-443-5602，042-443-5602，nakashika@uec.ac.jp



(a) Conventional model (ARBM)　(b) Proposed model (CAB)

図 1: 従来モデル（a）及び提案モデル (b) のグラフ構造比較.

2. 提案モデル：CAB

我々の先行研究 [c] で提案した ARBM は，図 1 (a) に示すように，音響（メルケプストラム）特徴量 $\boldsymbol{v} = [v_1, \cdots, v_I] \in \mathbb{R}^I$ と，潜在特徴量 $\boldsymbol{h} = [h_1, \cdots, h_J] \in \{0,1\}^J, \sum_j h_j = 1$ の間には，話者特徴量 $\boldsymbol{s} = [s_1, \cdots, s_R] \in \{0,1\}^R, \sum_r s_r = 1$ に依存した双方向な接続重み $\tilde{\mathbf{W}} \in \mathbb{R}^{I \times J}$ が存在すると仮定した確率密度関数である．このとき，話者依存の適応行列集合 $\mathcal{A} = \{\mathbf{A}_r\}_{r=1}^R, \mathbf{A}_r \in \mathbb{R}^{I \times I}$ のパラメータ数は I^2R となるが，音響特徴量の二乗（I^2）が比較的大きいため，話者数が増加するほど推定すべきパラメータ数が膨大となり計算コストが掛かってしまう．そこで本研究では，各話者の特徴を複数のクラスタの荷重和で表現するクラスタ適応技術に着目し，話者クラスタ適応 RBM（CAB）を提案する．CAB では，図 1 (b)

表 1: 適応話者の変換精度比較（dB）.

# sent.	0.2	0.5	1	10	40
ARBM	2.48	3.25	3.21	3.41	3.45
CAB	**3.14**	**3.54**	**3.63**	**3.60**	**3.58**

に示すように，話者クラスタ $\boldsymbol{c} \in \mathbb{R}^K$（$K < R$ はクラスタ数）を導入し，$\boldsymbol{c} \triangleq \mathbf{L}\boldsymbol{s}$ と恒等的に表現されるとする．ただし $\mathbf{L} \in \mathbb{R}^{K\times R} = [\boldsymbol{\lambda}_1 \ \cdots \ \boldsymbol{\lambda}_R]$ の各列ベクトル $\boldsymbol{\lambda}_r$ は，それぞれの話者のクラスタへの重みを表す非負パラメータであり，$\|\boldsymbol{\lambda}_r\|_1 = 1, \forall r$ の制約を課す．このとき，$\boldsymbol{s}$ が与えられた時の $\boldsymbol{v}, \boldsymbol{h}$ の条件付き確率密度関数を以下の式で定義する．

$$p(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{s}) = \frac{1}{Z} e^{-E(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{s})} \tag{1}$$

$$E(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{s}) = \frac{1}{2}\left\|\frac{\boldsymbol{v}-\tilde{\boldsymbol{b}}}{\boldsymbol{\sigma}}\right\|_2^2 - \tilde{\boldsymbol{d}}^\top \boldsymbol{h} - (\frac{\boldsymbol{v}}{\boldsymbol{\sigma}^2})^\top \tilde{\mathbf{W}}\boldsymbol{h} \tag{2}$$

ただし，Z は正規化項目，$\boldsymbol{\sigma} \in \mathbb{R}^I$ は音響特徴量の偏差を表すパラメータ，$\frac{\cdot}{\cdot}, \cdot^2$ はそれぞれ要素ごとの除算，要素ごとの二乗を表し，$\tilde{\mathbf{W}}$，$\tilde{\boldsymbol{b}}$ および $\tilde{\boldsymbol{d}}$ はそれぞれ以下の式で定義される：

$$\tilde{\mathbf{W}} \triangleq \mathcal{A} \circ_3^1 \boldsymbol{c}\mathbf{W} \tag{3}$$

$$\tilde{\boldsymbol{b}} \triangleq \boldsymbol{b} + \mathbf{U}\boldsymbol{c} + \mathbf{B}\boldsymbol{s} \tag{4}$$

$$\tilde{\boldsymbol{d}} \triangleq \boldsymbol{d} + \mathbf{V}\boldsymbol{c} + \mathbf{D}\boldsymbol{s} \tag{5}$$

ここで $\mathbf{W} \in \mathbb{R}^{I\times J}, \boldsymbol{b} \in \mathbb{R}^I, \boldsymbol{d} \in \mathbb{R}^J$ は話者･クラスタ非依存パラメータ，$\boldsymbol{b}_r \in \mathbb{R}^I (\mathbf{B} = [\boldsymbol{b}_1 \cdots \boldsymbol{b}_R]), \boldsymbol{d}_r \in \mathbb{R}^J (\mathbf{D} = [\boldsymbol{d}_1 \cdots \boldsymbol{d}_R])$ は話者依存パラメータ，$\mathbf{A}_k \in \mathbb{R}^{I\times I} (\mathcal{A} = \{\mathbf{A}_k\}_{k=1}^K), \mathbf{U} \in \mathbb{R}^{I\times K}, \mathbf{V} \in \mathbb{R}^{J\times K}$ はクラスタ依存パラメータである．CAB の全てのパラメータ $\Theta = \{\mathbf{W}, \mathbf{U}, \mathbf{V}, \mathcal{A}, \mathbf{L}, \mathbf{B}, \mathbf{D}, \boldsymbol{b}, \boldsymbol{d}, \boldsymbol{\sigma}\}$ は，確率的勾配法を用いて同時に更新・推定することが可能である．

3. 声質変換への応用

CAB を声質変換へ応用する場合，ある入力話者の音声の音響特徴量 $\boldsymbol{v}^{(i)}$ 及び話者特徴量 $\boldsymbol{s}^{(i)}$，目標話者の話者特徴量 $\boldsymbol{s}^{(i)}$ が与えられたとき，最も確率の高い音響特徴量 $\boldsymbol{v}^{(o)}$ が目標話者の音響特徴量であると定式化する．すなわち，

$$\hat{\boldsymbol{v}}^{(o)} \triangleq \operatorname*{argmax}_{\boldsymbol{v}} p(\boldsymbol{v}|, \boldsymbol{v}^{(i)}, \boldsymbol{s}^{(i)}, \boldsymbol{s}^{(o)}) \tag{6}$$

$$\simeq \boldsymbol{b} + \mathbf{B}\boldsymbol{s}^{(o)} + \mathbf{U}\mathbf{L}\boldsymbol{s}^{(o)} + \mathcal{A} \circ_3^1 \mathbf{L}\boldsymbol{s}^{(o)}\mathbf{W}\hat{\boldsymbol{h}} \tag{7}$$

ただし，$\hat{\boldsymbol{h}}$ は入力話者の音響特徴量及び話者特徴量が与えられたときの $\boldsymbol{h}$ の条件付き期待値：

$$\hat{\boldsymbol{h}} \triangleq \mathbb{E}[\boldsymbol{h}|\boldsymbol{v}^{(i)}, \boldsymbol{s}^{(i)}] \tag{8}$$

$$= \boldsymbol{f}(\boldsymbol{d} + \mathbf{V}\mathbf{L}\boldsymbol{s}^{(i)} + \mathbf{D}\boldsymbol{s}^{(i)} + \mathbf{W}^\top (\mathcal{A} \circ_3^1 \mathbf{L}\boldsymbol{s}^{(i)})^\top \frac{\boldsymbol{v}^{(i)}}{\boldsymbol{\sigma}^2})$$

である．ただし，$\boldsymbol{f}(\cdot)$ は要素ごとの softmax 関数を表す．

4. 評価実験

提案モデルの有効性を確かめるため，以下で述べる声質変換実験を行った．モデルの学習には日本音響学会研究用連続音声データベース（ASJ-JIPDEC）の中からランダムに $R = 16$ 名の話者を選び，40 センテンスの音声データを用いた．学習話者の評価には，男性 1 名（ECL0001）を入力話者，女性 1 名（ECL1003）を目標話者とし，学習データとは別の 10 センテンスの音声データを用いた．モデルの適応には，学習時に含まれない女性話者（ECL1004），男性話者（ECL0002）をそれぞれ入力話者，目標話者とし，適応データのセンテンス数を 0.2 から 40 まで変えて評価を行った．分析合成ツールの WORLD によって得られたスペクトルから計算した 32 次元のメルケプストラムを入力特徴量に用いた ($I = 32$)．また，潜在音韻特徴量の数を $J = 16$，クラスタの数を $K = 3$ とした．実際に推定された各話者のクラスタ重み $\boldsymbol{\lambda}_r$ の分布を図 2 に示す．図 2 より，性別の教師を与えていないにも関わらず，男性のクラスタ (Cluster 1) と女性のクラスタ（Cluster 2）が自動的に形成され，それ以外に男女が混ざった別のクラスタ（Cluster 3）が形成されているのが分かる．次に，提案モデルと従来モデルの，適応話者を用いた声質変換精度（MDIR; mel-cepstral distortion improvement ratio）を比較した（Table 1）．Table 1 から示されるように，いずれの適応センテンス数でも提案モデルが優っているが，特にセンテンス数が少ない時（1 以下の時）提案モデルの有効性が顕著に現れた．

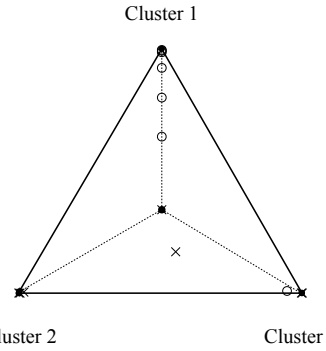


図 2: $K = 3, R = 16$ のときの，各話者のクラスタ重み分布．○ は男性話者，× は女性話者を表す．

5. おわりに

本稿では，少ないパラメータ数で適応可能となることを目的とし，従来の ARBM を拡張して，話者クラスタを自動的に形成する仕組みを追加し，声質変換へ応用する手法を提案した．評価実験では，提案モデルは，特に適応データ数が少ない場合において従来の ARBM よりも高い精度を示した．

参考文献

[a] INCA Algorithm for Training Voice Conversion Systems From Nonparallel Corpora, Audio (2010)

[b] Cluster adaptive training of hidden Markov models, year, IEEE Transactions on Speech and Audio Processing, pp. 417–428 (2000)

[c] Non-Parallel Training in Voice Conversion Using an Adaptive Restricted Boltzmann Machine, year, year, IEEE/ACM Transactions on Audio, Speech, and Language Processing, pp. 2032–2045 (2016)

[d] Continuous probabilistic transform for voice conversion, year, year, IEEE Transactions on Speech and Audio Processing, pp. 131–142 (1998)