声質変換のための Restricted Boltzmann Machine を用いたパラレル辞書の学習法*

☆中鹿亘，滝口哲也，有木康雄 (神戸大)

1 はじめに

近年，入力音声信号に含まれる音韻情報を残したまま，特定話者の声質のみを変換する技術（声質変換）の研究が盛んに行われている．この背景として，音声認識における話者性の制御による認識精度の向上や発話困難な障がい者の発話支援などへの応用が挙げられる．これまで声質変換は GMM (Gaussian Mixture Model) を用いた手法 [1, 2] が広く用いられてきたが，比較的制約の少ない統計的モデルのために over-fitting や over-smoothing の問題が生じてしまう [3, 4, 5]．そこで近年，over-fitting や over-smoothing の少ないスパース表現に基づく手法として，NMF (Non-negative Matrix Factorization) [6, 7] や US (Unit Selection) [5] による声質変換が提案されている．

スパース表現に基づく声質変換法では一般に，入力話者音声と出力話者音声のパラレルデータから基底（パラレル辞書）を学習し，テストデータが与えられたときに適切な (入力話者の) 辞書を選択して，(対となる出力話者の辞書を用いて) 出力話者の音声を生成する．この変換精度は，学習時における辞書の作成精度と，テスト時における辞書の選択精度によって決まる．パラレル辞書 NMF を用いた手法 [6] では，学習時には入力話者と出力話者のパラレルデータをそのままパラレル辞書として用意するので，辞書作成の誤差は生じないが，テスト時において，入力話者と出力話者のアクティビティ行列の不一致により誤差が生じてしまい，変換精度を下げる要因となる．また，NMF 声質変換においてアクティビティ行列が一致するように辞書を作成する手法 [7] も提案されているが，逆に辞書の作成精度が下がってしまい，変換精度に影響を与える．

本稿では，スパース表現に基づく声質変換において，パラレル辞書の作成・選択を統一的な枠組みで行うために，結合型 RBM(restricted Boltzmann machine) [8] を用いた声質変換法を提案する．

2 結合型 RBM による声質変換

声質変換では，入力話者の音響特徴ベクトル $\boldsymbol{x} \in \mathbb{R}^D$ を出力話者の音響特徴ベクトル $\boldsymbol{y} \in \mathbb{R}^D$ へ変換する．一般的なスパース表現に基づく手法では，予め

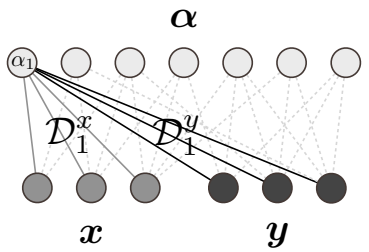

Fig. 1 Combined RBM for voice conversion. Each hidden unit connects with the dictionaries of the source and the target speakers.

K 個の入力話者の辞書 $\mathcal{D}_x \in \mathbb{R}^{D\times K}$ と出力話者の辞書 $\mathcal{D}_y \in \mathbb{R}^{D\times K}$ のペア（パラレル辞書）を学習しておき，変換時には

$$\boldsymbol{x} \approx \mathcal{D}_x \boldsymbol{\alpha} \tag{1}$$

となる辞書選択重み $\boldsymbol{\alpha} \in \mathbb{R}^K$ を何らかの手法で求め，

$$\boldsymbol{y} \approx \mathcal{D}_y \boldsymbol{\alpha} \tag{2}$$

を計算することで出力話者のベクトル $\boldsymbol{y}$ を得る．

本研究では式 (1)(2) の計算及び辞書の作成を同時に行うために，Fig. 1 に示される結合型 RBM (restricted Boltzmann machine) を用いる．RBM は可視層，隠れ層からなる 2 層ネットワークであり，層内は無く層間のみユニットの双方向結合が存在するという特徴がある．図のように，可視層ユニットを $\boldsymbol{x}$ と $\boldsymbol{y}$ の結合ベクトルとすれば，各重みのユニット $\boldsymbol{\alpha}_i$ が $\boldsymbol{x}$ と $\boldsymbol{y}$ へ，それぞれ i 番目の辞書 $\mathcal{D}_x^i$, $\mathcal{D}_y^i$ の重みが掛かっているネットワークと見ることもできる．

今，学習用のパラレルデータ $(\boldsymbol{x}, \boldsymbol{y})$ が与えられたとき，$\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\alpha}$ の同時尤度を以下のように表す．

$$p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\alpha}; \mathcal{D}_x, \mathcal{D}_y) = \frac{1}{Z} e^{-E(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\alpha}; \mathcal{D}_x, \mathcal{D}_y)} \tag{3}$$

$$E(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\alpha}; \mathcal{D}_x, \mathcal{D}_y) = -\|\boldsymbol{x}\|^2 - \|\boldsymbol{y}\|^2 - \boldsymbol{c}^t \boldsymbol{\alpha} - \boldsymbol{x}^T \mathcal{D}_x \boldsymbol{\alpha} - \boldsymbol{y}^T \mathcal{D}_y \boldsymbol{\alpha} \tag{4}$$

ただし，Z は正規化項，$\boldsymbol{c}$ はバイアスパラメータを表す．また，$\boldsymbol{x}, \boldsymbol{y}$ は要素ごとに平均 0, 偏差 1 となるように正規化されているものとする．辞書 $\mathcal{D}_x, \mathcal{D}_y$ (と $\boldsymbol{c}$) は (3) 式を最大化するように，CD (Contrastive Divergence) 法によって求められる．(3) 式に示され

*Parallel-dictionary learning using restricted Boltzmann machine for voice conversion. by Toru NAKASHIKA, Tetsuya TAKIGUCHI, Yasuo ARIKI (Kobe University)

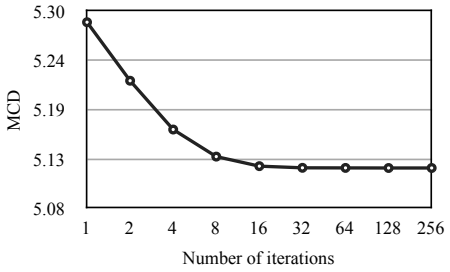

Fig. 2　Average MCDs when changing the number of iterations ($N = 100k, K = 192$).

るように，提案法では学習データの尤度だけでなく，辞書選択重みの尤度も考慮して同時に最適化を行っている．

変換の際には，RBM の前方推論・後方推論を利用する．まず，$\boldsymbol{y} = \mathbf{0}$ とおき，以下の式によって重み $\boldsymbol{\alpha}$ を求める．

$$p(\boldsymbol{\alpha} = \mathbf{1}|\boldsymbol{x}, \boldsymbol{y}) = \sigma(\mathcal{D}_x^T \boldsymbol{x} + \mathcal{D}_y^T \boldsymbol{y} + \boldsymbol{c}) \tag{5}$$

ただし，$\sigma(\cdot)$ は要素ごとのシグモイド関数を表す．辞書選択重みとして，この確率値を用いる．これによって，$\boldsymbol{y}$ を次式で求めることができる．

$$\boldsymbol{y} = \underset{y_0}{\operatorname{argmax}}\, p(\boldsymbol{y} = \boldsymbol{y}_0|\boldsymbol{\alpha}) \tag{6}$$

$$= \underset{y_0}{\operatorname{argmax}}\, \mathcal{N}(\boldsymbol{y}_0|\mathcal{D}_y\boldsymbol{\alpha},\ \mathbf{1}) \tag{7}$$

$$= \mathcal{D}_y\boldsymbol{\alpha} \tag{8}$$

実際には，(8) 式で推論された $\boldsymbol{y}$ はパワーの小さいベクトルとなるため，推論された $\boldsymbol{y}$ を用いて再度 (5) 式で評価し重みを求め，(8) 式で強調された $\boldsymbol{y}$ を得る．このように，(5) 式と (8) 式を R 回繰り返し，反復的に $\boldsymbol{y}$ を求める．

3　評価実験

本実験では ATR 研究用日本語音声データベース (A Set) [9] を用いた声質変換を行い，提案手法の効果を確認した．このデータベースから，入力話者として男性話者 1 名 (MMY) を，出力話者として女性話者 1 名 (FTK) を選んだ．学習・評価用のデータは STRAIGHT パラメータから抽出した 0 次元を除く $D = 24$ 次元の MFCC 特徴を用いた．学習用のデータとして，216 単語の音声から動的計画法を用いて作成したパラレルデータのうち，ランダムに選んだ $N(k)$ フレームを使用した．また，評価用のデータとして 20 文の音声を用いて，平均 MCD (Mel cepstral distortion) を算出し，提案手法を評価した．また，RBM の隠れ層のユニット数 K や繰り返し回数 R を変えて，提案法の比較を行った．

Table 1　MCDs in each condition ($R = 32$).

(N,K)	(100,192)	(200,96)	(300,96)	(300,192)
MCD	5.11	4.93	4.91	**4.88**

繰り返し回数を変えた時の実験結果を Fig. 2 に示す．この図より，繰り返し回数を増加させていくと，変換精度が向上していくことが確認できる．また，隠れユニット数，学習データ数を変えた時の結果を Table 3 に示す．この表より，N，K を増加させていくと精度が向上することが分かる．

4　おわりに

本稿では，パラレル辞書の作成．選択を同時に求めるフレームワークとして，入力話者・主力話者の結合型 RBM を用いた声質変換法を提案した．また，辞書選択重み，出力話者音響特徴量を繰り返し的に求める手法を提案し，その効果を確認した．

参考文献

[1] Y. Stylianou, et al., “Continuous probabilistic transform for voice conversion,” IEEE Trans. Speech and Audio Processing, vol. 6, no. 2, pp. 131—142, 1998.
[2] T. Toda, et al., “Voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” IEEE Trans. Audio, Speech, and Language Processing, vol. 15, no. 8, pp. 2222—2235, 2007.
[3] Z. H. Ling, et al., “Modeling Spectral Envelopes Using Restricted Boltzmann Machines and Deep Belief Networks for Statistical Parametric Speech Synthesis,” IEEE Trans. Audio, Speech, and Language Processing, vol. 21, no. 10, pp. 2129–2139, 2013.
[4] E. Helander, et al., “Voice conversion using dynamic kernel partial least squares regression,” IEEE Trans. Audio, Speech, and Language Processing, vol. 20, no. 3, pp. 806–817, 2012.
[5] Z. Wu, et al., “Exemplar-based unit selection for voice conversion utilizing temporal information,” Proc. Interspeech 2013, pp. 3057–3061, 2013.
[6] R. Takashima, et al., “Exemplar-Based Voice Conversion Using Sparse Representation in Noisy Environments,” IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences, vol. E96-A, no. 10, pp. 1946–1953, 2013.
[7] R. Takashima, et al., “Noise-Robust Voice Conversion Based on Spectral Mapping on Sparse Space,” SSW8, pp. 71–75, 2013.
[8] J. Gao, et al., “Restricted Boltzmann Machine Approach to Couple Dictionary Training for Image Super-resolution,” IEEE ICIP, pp. 499–503, 2013.
[9] A. Kurematsu, et al., “ATR Japanese speech database as a tool of speech recognition and synthesis,” Speech Communication, vol. 9, pp. 357–363, 1990.