時間変化を考慮したDeep Learningを用いた声質変換*

☆中鹿亘，滝口哲也，有木康雄（神戸大）

1　はじめに

入力音声から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) が，音声認識における話者性の制御や発話支援などへ応用できるとして，近年盛んに研究されている．これまでの声質変換に関する代表的な手法として，GMM (Gaussian Mixture Model) を用いた声質変換法 [1] が広く用いられており，多くの改良がなされている．例えば，戸田ら [2] は従来の GMM による声質変換法に動的特徴と Global Variance を導入し，より自然な音声へ変換する手法を提案している．また，別の統計的アプローチとして，Neural Networks (NN) を用いた手法が Desai らによって提案されている [3]．GMM に基づく変換では，入力音声と出力音声の同時確率を予め推定しておき，条件付き確率を最大化するように出力音声を生成していたが，NN による変換法では，入力話者の特徴ベクトルを出力話者の特徴ベクトルへ直接変換するモデルを学習する．

一方，特徴抽出器を複数積み重ねた深い構造を持つ機械学習器 (Deep Learning) を用いた手法が，音声認識などの分野において高い精度を上げることで近年特に注目を浴びており [4]，我々は Deep Learning 手法の一つである Deep Belief Nets (DBNs) を用いた精度の高い声質変換法を提案してきた [6]．

上で述べた従来手法をはじめ，多くの声質変換法では，音声信号は時間的に変化するものであるにもかかわらず，モデルの中にデータの時間遷移が考慮されておらず，音声信号の時間的な変化を正しく捉えることは困難であった．そこで本研究では，Conditional Restricted Boltzmann Machine (CRBM) [7] を用いて音声の時間的変化を捉え，Deep Learning の枠組みで声質変換を行う手法を提案する．

2　提案手法

本研究では，話者ごとに CRBM (Conditional Restricted Boltzmann Machine) を用いて時間変化を考慮した高次元特徴抽出を行い，(1) なるべく音韻性・時間性を除外し，(2) 本質的な話者性のみを残した空間で変換を行うことによって高精度な声質変換法を実現することを目指す．CRBM を用いてこれらが実現できる根拠については後述する．

提案法によるシステムの概要を Fig. 1 に示す．図中 (a) 下はソース話者の CRBM の構造を示してあり，

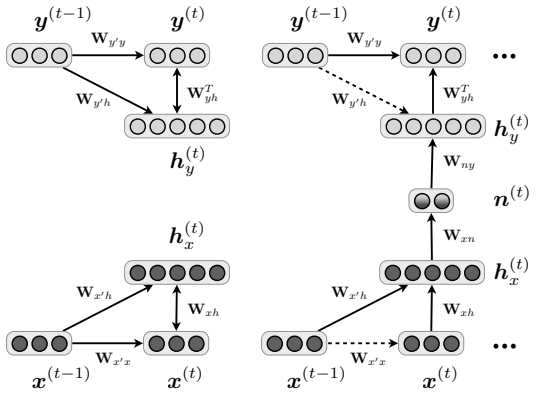

Fig. 1　(a) CRBMs for a source speaker (below) and a target speaker (above), (b) our proposed voice conversion architecture using two different CRBMs.

時刻 t における特徴ベクトル $\boldsymbol{x}^{(t)} \in \mathbb{R}^{D\times 1}$，時刻 $t-1$ における特徴ベクトル $\boldsymbol{x}^{(t-1)} \in \mathbb{R}^{D\times 1}$，潜在ベクトル $\boldsymbol{h}_x^{(t)} \in [0,1]^{H\times 1}$ の関係性を表している．通常の RBM (Restricted Boltzmann Machine) 同様，それぞれのベクトル内のノードは互いに条件付き独立であるが，ベクトル間では図のような依存関係がある．すなわち $\boldsymbol{x}^{(t-1)}$-$\boldsymbol{x}^{(t)}$，$\boldsymbol{x}^{(t-1)}$-$\boldsymbol{h}_x^{(t)}$ 間は一方向，$\boldsymbol{x}^{(t)}$-$\boldsymbol{h}_x^{(t)}$ 間は双方向の重み行列 (それぞれ $\mathbf{W}_{x'x}$，$\mathbf{W}_{x'h}$，$\mathbf{W}_{xh}$) が存在するようなモデルである．これらの重み行列は，以下の尤度を最小化するように求められる [7]．

$$p(\boldsymbol{x}^{(t)}|\boldsymbol{x}^{(t-1)}) = \frac{1}{Z}\sum_{\boldsymbol{h}_x^{(t)}} e^{-E(\boldsymbol{x}^{(t)},\boldsymbol{h}_x^{(t)}|\boldsymbol{x}^{(t-1)})} \quad (1)$$

ただし，Z は正規化項，E は以下で定義されるエネルギー関数である．

$$E(\boldsymbol{x}^{(t)},\boldsymbol{h}_x^{(t)}|\boldsymbol{x}^{(t-1)}) = -\boldsymbol{b}^T\boldsymbol{x}^{(t)} - \boldsymbol{c}^T\boldsymbol{h}^{(t)} - \boldsymbol{x}^{(t)^T}\mathbf{W}_{xh}\boldsymbol{x}^{(t)} \quad (2)$$

$$-\boldsymbol{x}^{(t-1)^T}\mathbf{W}_{x'x}\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)^T}\mathbf{W}_{x'h}\boldsymbol{h}^{(t)} \quad (3)$$

一度重み行列が求まれば，$\boldsymbol{x}^{(t)}$ と $\boldsymbol{x}^{(t-1)}$ が与えられたときの $\boldsymbol{h}^{(t)}$(前方推論)，$\boldsymbol{h}^{(t)}$ と $\boldsymbol{x}^{(t-1)}$ が与えられたときの $\boldsymbol{x}^{(t)}$(後方推論) はそれぞれ以下のように計算される．

$$p(\boldsymbol{h}^{(t)} = \mathbf{1}|\boldsymbol{x}^{(t)},\boldsymbol{x}^{(t-1)}) = \sigma(\boldsymbol{c} + \boldsymbol{x}^{(t)^T}\mathbf{W}_{xh} + \boldsymbol{x}^{(t-1)^T}\mathbf{W}_{x'h}) \quad (4)$$

$$p(\boldsymbol{x}^{(t)} = \mathbf{1}|\boldsymbol{h}^{(t)},\boldsymbol{x}^{(t-1)}) = \mathcal{N}(\boldsymbol{b} + \boldsymbol{h}^{(t)^T}\mathbf{W}_{xh}^T + \boldsymbol{x}^{(t-1)^T}\mathbf{W}_{x'x},\ 1) \quad (5)$$

*Voice conversion based on time-varying deep networks. by Toru NAKASHIKA, Tetsuya TAKIGUCHI, Yasuo ARIKI (Kobe University)

ただし，$\sigma(\cdot)$, $\mathcal{N}(\cdot)$ はそれぞれシグモイド関数，正規分布を表す．ターゲット話者に関しても同様に，ターゲット話者の音声のみを用いて CRBM の重み行列 (Fig. 1(a) 上）を学習させる．

音声特徴量の中には音韻情報と話者情報が存在する．話者性は変化せずに様々な音韻情報を豊富に含む学習データを用いて得られた潜在ベクトル $\boldsymbol{h}_x^{(t)}$ は，学習データの共通因子を捉えるように作用されているため，音韻情報よりも話者情報が多く含まれていると考えられる．また，重み行列 $\mathbf{W}_{x'x}$ と $\mathbf{W}_{x'h}$ が入力特徴・潜在特徴の時間的な変化を捉えているため，$\mathbf{W}_{xh}$ は特徴ベクトルの時間的変化に関係のない特徴抽出が可能であることを示唆している．

声質変換では音韻情報を変えずに話者性のみを変換させるので，元の音響特徴 (スペクトルや MFCC など) に比べて，それぞれの話者性を強調させた潜在空間の方が正しく変換が変換できると期待される．本研究では，3 層の Neural Networks (NN) を用いて潜在ベクトル $\boldsymbol{h}_x^{(t)}$-$\boldsymbol{h}_y^{(t)}$ 間の変換を行う（この NN を中間 NN と呼ぶ）．このときに得られた重み行列を $\mathbf{W}_{xn}$, $\mathbf{W}_{ny}$ とする．結局，ソース話者の音響特徴ベクトル $\boldsymbol{x}^{(t)}$ は Fig. 1 (b) のように，ソース話者の CRBM の前方推論，中間 NN，ターゲット話者の CRBM の後方推論を経て，ターゲット話者の音響特徴ベクトル $\boldsymbol{y}^{(t)}$ へ変換される．このとき，最終層を除く全ての層の出力はシグモイド関数で表わされるため，CRBM や中間 NN で求めた重み行列を初期値として，全体のネットワークの Fine-tuning が可能である．

3 評価実験

3.1 実験条件

本実験では ATR 研究用日本語音声データベース [8] を用いた声質変換を行い，ネットワークベースの手法である DBN，NN による声質変換法と比較した．このデータベースから，男性話者 1 名 (MMY) の音声を入力話者音声に，女性話者 1 名 (FTK) の音声を出力話者音声として用いた．

学習・評価用音声データは STRAIGHT 分析を行い，得られた STRAIGHT パラメータから 0 次元を除く $D(=24)$ 次元の MFCC を抽出した．216 単語の音声から Dynamic Programming を用いて入力・出力話者のパラレルデータを作成し，本手法における中間 NN 及び Fine-tuning，従来法における DBN，NN の学習に用いた．潜在ベクトル $\boldsymbol{h}_x^{(t)}$，$\boldsymbol{h}_y^{(t)}$ の次元数はともに $H = 48$，中間 NN の隠れ層のノード数は 24 とした．また従来手法の DBN，NN に関しても，同様の構造を持つ (入力層から順に $24, 48, 24, 48, 24$ 個のノードを持つ 5 層の) ネットワークを用いた．

Methods	CRBM	CRBMi	DBN	NN	SRC
MCD(dB)	6.27	**5.93**	6.30	6.44	7.15

Table 1　Average MCDs obtained by each method.

提案手法の有効性を確かめるため，MCD(Mel-cepstrum distortion) を用いた客観的評価を行った．評価データには，ATR のデータベースからランダムに選択した 20 文の音声を用いた．

3.2 実験結果・考察

各手法による 20 文音声の平均 MCD を Table 1 に示す．提案法である “CRBM” と “CRBMi” の違いは，時刻 t における予測を行う際に，$\boldsymbol{y}^{(t-1)}$ の値として，時刻 $t-2$ からの予測値を用いるか，時刻 $t-1$ のターゲット話者の特徴ベクトルを用いるかである．また，参考までにソース話者とターゲット話者の特徴ベクトル間の距離を “SRC” として載せている．Table 1 より，提案手法の “CRBMi” から最も高い精度が得られたことが分かる．完全に未知の状態で推定を行う手法 (“CRBM”,“DBN”,“NN”) で比較しても，提案法の “CRBM” が従来手法よりも優れた結果を示している．これは，時間的な変化をネットワークに組み込むことで，より話者性が強調された空間が形成されたためであると考えられる．

4 おわりに

本稿では，時間的変化を捉える Deep Learning に着目し，話者ごとの CRBM によって得られた話者空間変換に基づく声質変換法を提案した．

参考文献

[1] Y. Stylianou, O. Cappe, and E. Moulines, “Continuous probabilistic transform for voice conversion,” IEEE Trans. Speech and Audio Processing, Vol. 6, No. 2, pp. 131—142, 1998.

[2] T. Toda, A. Black, and K. Tokuda, “Voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” IEEE Trans. Audio, Speech, Lang. Process., Vol. 15, No. 8, pp. 2222—2235, 2007.

[3] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, “Spectral mapping using artificial neural networks for voice conversion,” IEEE Trans. Audio, Speech, Lang. Process., Vol. 18, No. 5, pp. 954—964, 2010.

[4] G. Hinton, et al, “Deep neural networks for acoustic modeling in speech recognition,” IEEE Signal Processing Magazine, 2012.

[5] G. E. Hinton, S. Osindero, and Y. W. Teh, “A fast learning algorithm for deep belief nets,” Neural Computation, vol. 18, no. 7, pp. 1527-1554, 2006.

[6] T. Nakashika, R. Takashima, T. Takiguchi, and Y. Ariki, “Study on voice conversion using Deep Belief Nets,” ASJ 2013 Spring Meeting, pp.517-520, 2013.

[7] G.W. Taylor, G.E. Hinton, and S.T. Roweis, “Modeling human motion using binary latent variables,” Neural information processing, vol. 19, pp. 1345, 2007.

[8] 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, pp. 357—363, 1990.