Deep Belief Netsによる低次元空間表現を用いた声質変換の検討*

☆中鹿亘，高島遼一，滝口哲也，有木康雄（神戸大）

1 はじめに

近年，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である声質変換の研究が盛んに行われている．声質変換は，音声合成や音声認識における話者性の制御 [1]，雑音環境下音声の音声強調，感情変換 [2]，発話支援など [3]，多岐にわたる応用が期待されている [4, 5]．話者性を変換させる声質変換においては，特に音声スペクトル特徴が重要な役割を果たすため，多くの研究がスペクトル特徴の変換に主眼をおいており，本研究でもこれに準ずる．

これまでの声質変換に関する代表的な手法として，統計的手法が挙げられる [6, 7]．中でも Gaussian Mixture Model (GMM) を用いた手法 [8] が広く用いられており，その後多くの改良がされ続けている．戸田ら [9] 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．Helander ら [10] は従来手法における過適合の問題を回避するために，Partial Least Squares (PLS) 回帰分析を用いる手法を提案している．その他にも，学習時にパラレルデータ（同一発話内容において入力話者と出力話者間の時間的整合をとったデータ）を必要とせずに GMM の話者適応を行う手法 [11] や，Eigen-Voice GMM (EV-GMM) [13, 14]，話者モデルとの確率的統合による手法 [12] などが提案されている．

声質変換法に関する別の統計的アプローチとして，Neural Networks (NN) を用いた手法が提案されている [15, 16]．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．一方 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．このような識別的アプローチは，声質変換だけでなく，音声合成や音声認識の分野でもしばしば良い結果をもたらすことが示されている [17, 18]．また，入力・出力話者の声道形状は非線形的であり，変換関数を非線形とする NN との相性が良い．このような理由から，NN によるアプローチは声質変換においても高い精度を上げている [15, 16]．

また近年，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，音声認識や画像認識などの分野において高い精度を上げることで注目を浴びている [19, 20, 21, 22, 23]．Deep Learning には Convolutional Neural Networks [23] や Deep Belief Nets (DBN) [24]，Auto Encoder など，特徴抽出器や教師の有無などにより様々なタイプが存在するが，いずれも深い階層構造を持つことで，上層へいくにしたがって下層の情報を複雑に集約させるため，最上位の受容野において入力特徴をより表現することが可能となる．中でも DBN は 2006 年に Hinton ら [24] によって効率的な学習アルゴリズムが提案され，以降 DBN によるパターン認識が盛んに研究されている [19, 20, 21, 22]．

本研究では DBN を用いて入力・出力話者それぞれの特徴空間へ写像し，得られた低次元空間において変換を加えることで，最適な非線形関数で声質変換を行う手法を提案する．深い階層構造を持つ DBN において，各層の特徴空間は決められたノード数で入力特徴を表現するように自動形成されるため，最上位層の空間ではコンパクトに表現された基底集合となる．本稿では，階層の数を増やせば増やすほどこの傾向は顕著になり，無限まで階層の数を増やすと完全に話者性のない空間 (共通空間) が形成されると仮定する．現実的には階層数を有限に止めるが，最上位層における低次元空間では，元の特徴空間よりも話者性が薄くなり，入力・出力話者の共通空間に近い空間が形成されていると期待できる．本研究ではこの共通空間を NN を用いて近似する．すなわち NN を用いて低次元空間内における特徴変換を行う．共通空間を介して，入力話者から出力話者への変換を行い，DBN の逆プロセスを利用して出力話者の話者性を保持した高次元空間へ逆変換する．

評価実験では従来の GMM による声質変換法と比較し，提案手法の優位性を示す．

2 提案手法

2.1 DBN と NN による声質変換

本研究では Deep Belief Nets (DBN) と Neural Networks (NN) を組み合わせて，話者性の影響の少ない低次元空間において入力特徴・出力特徴を変換し，元の出力話者空間へ逆射影する (Fig. 1)．入力・出力話者の特徴変換器にはそれぞれ DBN_s, DBN_t という別々の DBN を用意するため，入力・出力話者それぞれの特徴空間をよく表現する空間を形成できる．DBN

*Study on voice conversion using Deep Belief Nets. by Toru NAKASHIKA, Ryoichi TAKASHIMA, Tetsuya TAKIGUCHI, Yasuo ARIKI (Kobe University)

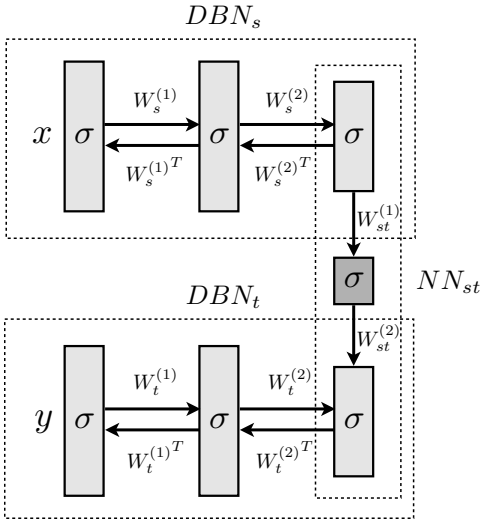

Fig. 1 Our proposed voice conversion architecture, combined with 2 different DBNs and concatenating NN. A source feature vector x is fed to DBN_x, NN_{xy}, and DBN_y in order, and then, converted to a target vector y. This figure shows an example of architectures with 2 hidden layers in the DBNs and with 1 hidden layer in the NN. σ indicates sigmoid function, i.e. $\sigma(x) = 1/(1 + exp(-ax))$.

の学習時には，入力話者の音声特徴ベクトル集合を用いて DBN_s を，出力話者のベクトル集合を用いて DBN_t を自己学習させる．DBN ではトップダウンとボトムアップの遷移重みを共有するという特徴がある．今，L 個の特徴抽出層を持つ DBN_s，DBN_t によるボトムアップ変換関数をそれぞれ $\zeta_s(\boldsymbol{x})$，$\zeta_t(\boldsymbol{y})$ とすると，重みパラメータ $\boldsymbol{W}_s^{(l)}$，$\boldsymbol{W}_t^{(l)}(l = 1, 2, \ldots, L)$ が与えられたとき，

$$\zeta_i(\boldsymbol{x}) = (\zeta_i^{(1)} \circ \zeta_i^{(2)} \circ \cdots \circ \zeta_i^{(L)})(\boldsymbol{x}) \quad (1)$$

$$= \bigodot_{l=1}^{L} \zeta_i^{(l)}(\boldsymbol{x}) \quad (2)$$

$$\zeta_i^{(l)}(\boldsymbol{x}) = \sigma(\boldsymbol{W}_i^{(l)}\boldsymbol{x}), \quad i \in \{s, t\} \quad (3)$$

と表せる．ただし，$\bigodot_{l=1}^{L}$ は L 個の合成関数を表す．例えば，$\bigodot_{l=1}^{2} \zeta_s^{(l)}(\boldsymbol{x}) = \sigma(\boldsymbol{W}_s^{(2)}\sigma(\boldsymbol{W}_s^{(1)}\boldsymbol{x}))$ となる．同様に，最上層における特徴ベクトル $\boldsymbol{x}'$ が与えられたとき，最下層 (可視層) のベクトルを求めるトップダウン変換関数 $\zeta_i^{-1}(\boldsymbol{x}')$ は以下のように表される．

$$\zeta_i^{-1}(\boldsymbol{x}') = \bigodot_{l=1}^{L} \sigma(\boldsymbol{W}_i^{(L-l+1)^T}\boldsymbol{x}') \quad (4)$$

また，式 (1) によって得られる入力音声の低次元特徴ベクトルは，$I + 1$ 層の多層パーセプトロンである NN_{st} によって出力音声の低次元特徴ベクトルへ変換される．NN_{st} のパラメータ $\boldsymbol{W}_{st}^{(l)}(l = 1, 2, \ldots, I)$ が予め推定されていれば，両者のベクトルは次式で変換することが可能である．

$$\eta_{st}(\boldsymbol{x}) = \bigodot_{l=1}^{I} \sigma(\boldsymbol{W}_{st}^{(i)}\boldsymbol{x}) \quad (5)$$

以上をまとめると，入力話者の特徴ベクトル $\boldsymbol{x}$ から出力話者の特徴ベクトル $\boldsymbol{y}$ へ変換する変換式は以下のようになる．

$$\boldsymbol{y} = \zeta_t^{-1}(\eta_{st}(\zeta_s(\boldsymbol{x}))) \quad (6)$$

$$= \bigodot_{l=1}^{2L+I} \sigma(\boldsymbol{W}^{(l)}\boldsymbol{x}) \quad (7)$$

ただし，パラメータ集合 $\boldsymbol{W} = \{\boldsymbol{W}^{(l)}\}_{l=1}^{2L+1} = \{\boldsymbol{W}_s^{(1)}, \cdots, \boldsymbol{W}_s^{(L)}, \boldsymbol{W}_{st}^{(1)}, \cdots, \boldsymbol{W}_{st}^{(I)}, \boldsymbol{W}_t^{(L)^T}, \cdots, \boldsymbol{W}_t^{(1)^T}\}$ を用いる．

式 (7) で示されるように，提案手法による声質変換法では，異なる非線形関数の合成関数によって出力特徴ベクトルがモデル化される．一方，従来の GMM(混合数 M) による声質変換では，

$$\boldsymbol{y} = \sum_{m=1}^{M} P(m|\boldsymbol{x})\boldsymbol{E}_m^{(y)} \quad (8)$$

$$P(m|\boldsymbol{x}) = \frac{w_m \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)}}{\sum_{m=1}^{M} w_m \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)}} \quad (9)$$

$$\boldsymbol{E}_m^{(y)} = \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yx)}\boldsymbol{\Sigma}_m^{(xx)-1}(\boldsymbol{x} - \boldsymbol{\mu}_m^{(x)}) \quad (10)$$

のように変換され，複数の非線形関数が加算されるようなモデルとなっている．したがって，複数の非線形関数の合成関数で表される，提案手法による変換関数の方が表現能力が増しているので，正しくパラメータが推定されれば GMM よりも高品質な変換結果が得られることが期待される．

2.2 モデルの学習

DBN は 2 層からなる Ristricted Boltzmann Machine (RBM) を下層から上層へ順に積み重ねたような構造となり，学習も RBM 単位で下層から順にパラメータが決定される．RBM は可視素子層と隠れ素子層の間の繋がりは許されるが，可視素子間または隠れ素子間の繋がりは存在しない無向グラフィカルモデルである．RBM では，二値可視素子集合 $\boldsymbol{v} = \{v_1, \cdots, v_i, \cdots\}, v_i \in \{0, 1\}$ と隠れ素子集合 $\boldsymbol{h} = \{h_1, \cdots, h_j, \cdots\}, h_j \in \{0, 1\}$ の同時確率 $p(\boldsymbol{v}, \boldsymbol{h})$ を以下のように定義する．

$$p(\boldsymbol{v}, \boldsymbol{h}) = \frac{1}{Z}\exp(-E(\boldsymbol{v}, \boldsymbol{h})) \quad (11)$$

$$E(\boldsymbol{v}, \boldsymbol{h}) = -\boldsymbol{b}^T\boldsymbol{v} - \boldsymbol{c}^T\boldsymbol{h} - \boldsymbol{v}^T\boldsymbol{W}\boldsymbol{h} \quad (12)$$

$$Z = \sum_{\boldsymbol{v},\boldsymbol{h}} \exp(-E(\boldsymbol{v}, \boldsymbol{h})) \quad (13)$$

ただし，$\boldsymbol{W}$，$\boldsymbol{b}$，$\boldsymbol{c}$ はそれぞれ可視素子と隠れ素子間の結合度合いを表す重みパラメータ，可視素子，隠れ素子のバイアスパラメータを表す．可視素子間，または隠れ素子間の繋がりは存在しないと仮定しているので，それぞれの条件付き確率を計算すると，次式のように非常に単純なものとなる．

$$p(\boldsymbol{h}=\boldsymbol{1}|\boldsymbol{v}) = \sigma(\boldsymbol{c}+\boldsymbol{v}^T\boldsymbol{W}) \quad (14)$$

$$p(\boldsymbol{v}=\boldsymbol{1}|\boldsymbol{h}) = \sigma(\boldsymbol{b}+\boldsymbol{h}^T\boldsymbol{W}^T) \quad (15)$$

これは，式 (3) や Fig. 1 の DBN において，データがある層から次の層へ遷移するときにシグモイド関数を通すことと一致する．

パラメータの推定には，可視素子の対数出現確率 $\log p(\boldsymbol{v})$ が評価関数として用いられる．これを各パラメータで偏微分すると，

$$\frac{\partial \log p(\boldsymbol{v})}{\partial w_{ij}} = \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model} \quad (16)$$

を得る．ここで，$\langle\cdot\rangle_{data}$，$\langle\cdot\rangle_{model}$ はそれぞれ入力データの期待値，内部モデルの期待値を表す．しかしながら，第二項の計算は困難である場合が多いため，式 (14)(15) を用いて入力データを再構築したデータの期待値 $\langle\cdot\rangle_{recon}$ を代わりに用いる [24]．式 (16) から，パラメータは確率的最急降下法などを使って更新することができる．

DBN における現在の層の RBM が更新されると，隠れ素子の確率値 (式 (14)) が一つ上の層の RBM の可視素子へコピーされる．

入力・出力話者 DBN の学習が終わったあと，入力音声と出力音声のパラレルデータ $\{\boldsymbol{x}_n, \boldsymbol{y}_n\}_{n=1}^N$ を用いて低次元空間変換 NN_{st} の学習を行う．NN の出力値 $\eta_{st}(\zeta_s(\boldsymbol{x}_n))$ と教師信号 $\zeta_t(\boldsymbol{y}_n)$ の誤差が最小となるように各重みパラメータを推定する．この際，入力音声と出力音声のパラレルデータを利用して誤差逆伝播を行い，DBN の各層の重みパラメータも同時に微調整する．

2.3 特徴ベクトルの前処理

式 (12) や式 (15) などで示される DBN(もしくは RBM) は，入力層の素子の値が二値であるという仮定のもとで定式化されている．そのため，音声特徴などの連続値を入力すると，うまく学習されないことがある．可視素子が単峰正規分布からサンプルされているとモデル化することにより，連続値の入力に対応する手法 [24] も存在するが，本研究では予め特徴ベクトルをソフト二値化させるアプローチをとっている．つまり，特徴ベクトルの各次元ごとに平均と偏差による正規化を行い，シグモイド関数を通すことで，可視素子が 0 もしくは 1 に近い連続値をとるようにしている．

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 [25] を用いた声質変換を行い，従来の GMM による手法と比較した．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．サンプリング周波数は 8kHz である．

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 次元のケプストラムを用いた．本研究では $D = 40$ である．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 DBN と NN の学習，従来法における GMM の学習に用いた．このときのデータ数（フレーム数）は 189,992 である．DBN の学習率を 0.05，更新回数を 50 とした．DBN の構造として，(DBN1) L=1,I=1，(DBN2) L=2,I=1 を与えた．入力から出力までの各層におけるノードの数はそれぞれ [40 80 80 40]，[40 120 30 30 120 40] とした．また GMM の混合数は 64 とし，分散共分散行列には対角行列を用いた．

提案手法の有効性を確かめるため，客観的指標と主観的指標による評価を行った．客観評価実験では ATR のデータベースから 15 文の音声をランダムに選択し，評価データとして用いた．提案法・従来法ともにスペクトル特徴量であるケプストラムのみ DBN，GMM を用いて変換し，パワー及び基本周波数については平均と標準偏差による線形変換によって変換した．ケプストラム変換後 STRAIGHT パラメータへ逆変換し，式 (17) で表される NSD (Normalized Spectrum Distortion) によって各手法を比較した．

$$NSD = \sqrt{\frac{\|S^Y - S^{\hat{X}}\|^2}{\|S^Y - S^X\|^2}} \quad (17)$$

ただし，S^X，S^Y，$S^{\hat{X}}$ はそれぞれ入力話者のスペクトル，出力話者のスペクトル，変換後のスペクトルを表す．

また，主観評価実験では XAB による評価を行った．主観評価実験で用いた 15 文の音声のうち，5 文を用いた．提案手法では，DBN1 の構造を採用した．この実験では，9 名の被験者に，各手法によって変換された音声を聴き比べて，どちらがより音声らしいかを選択してもらった．

3.2 実験結果・考察

各手法から得られた変換音声を元に，客観的尺度，主観的尺度を算出した値をそれぞれ Fig. 2 の左図，右

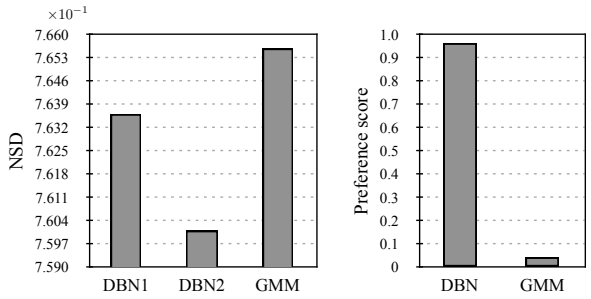

Fig. 2　Normalized spectrum distortion calculated from converted speech using each method (left), and preference score to see auditory measure (right)

図に示す．まず客観的尺度に注目すると，提案手法である DBN1,DBN2 がともに従来手法の GMM よりもスペクトル歪みが小さくなっていることが分かる．これは提案手法による変換が，複数の非線形関数の合成によって行われることで，GMM よりも深い次元で特徴空間をモデル化できたからだと考えられる．DBN1 と DBN2 を比べると，DBN2 の方がよい結果が得られた．これは，DBN の階層を増やすことで，話者性の薄れた空間が形成され，より変換のしやすい空間上で変換が行われたためであると考えられる．

次に，主観的尺度を見ると，ほとんど全てのテストにおいて DBN による変換音声が GMM の変換音声に優っていたことが分かる．客観評価値である NSD ではそれほど顕著な差は見られなかったが，主観評価値により，DBN による声質変換は，GMM の変換に比べて聴覚的に優れた変換が行われたと言える．

4　おわりに

本稿では，DBN と NN を組み合わせて，話者性の取り除いた低次元空間で非線形変換を行う声質変換法を提案した．主観的・客観的に評価実験を行い，いずれの実験においても高い精度を示した．

参考文献

[1] A. Kain and M. W. Macon, “Spectral voice conversion for text-to-speech synthesis,” Proc. ICASSP, vol. 1, pp. 285—288, 1998.

[2] C. Veaux and X. Robet, “Intonation conversion from neutral to expressive speech,” in Proc. IN- TERSPEECH, pp. 2765—2768, 2011.

[3] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, “Speaking-aid systems using GMM- based voice conversion for electrolaryngeal speech,” Speech Communication, Vol. 54, No. 1, pp. 134— 146, 2012.

[4] L. Deng, A. Acero, L. Jiang, J. Droppo, and X. Huang, “High-performance robust speech recognition using stereo training data,” Proc. ICASSP, pp. 301—304, 2001.

[5] A. Kunikoshi, Y. Qiao, N. Minematsu, and K. Hirose, “Speech generation from hand gestures based on space mapping,” Proc. INTERSPEECH, pp. 308—311, 2009.

[6] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, “Vice conversion through vector quantization,” in Proc. ICASSP, pp. 655—658, 1988.

[7] H. Valbret, E. Moulines and J. P. Tubach, “Voice transformation using PSOLA technique,” Speech Communication, Vol. 11, No. 2-3, pp. 175—187, 1992.

[8] 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.

[9] 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.

[10] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, “Voice conversion using partial least squares regression,” IEEE Trans. Audo, Speech, Lang. Process., Vol. 18, No. 5, pp. 912—921, 2010.

[11] C. H. Lee and C. H. Wu, “Map-based adaptation for speech conversion using adaptation data selection and non-parallel training,” in Proc. INTERSPEECH, pp. 2254—2257, 2006.

[12] D. Saito, S. Watanabe, A. Nakamura, N. Minematsu, “Voice conversion based on probabilistic integration of joint density model and speaker model,” in Proc. Acoustic Society of Japan, pp. 335–338, 2010.

[13] T. Toda, Y. Ohtani, and K. Shikano, “Eigenvoice conversion based on Gaussian mixture model,” in Proc. INTERSPEECH, pp. 2446—2449, 2006.

[14] D. Saito, K. Yamamoto, N. Minematsu, and K. Hirose, “One-to-many voice conversion based on tensor representation of speaker space,” in Proc. INTERSPEECH, pp. 653—656, 2011.

[15] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, “Voice conversion using artificial neural networks,” in Proc. ICASSP, pp. 3893—3896, 2009.

[16] Z. Chen and L. H. Zhang, “A ANN Based High Quality Method for Voice Conversion,” in Proc. WiCOM, 2010.

[17] Y. J. Wu, H. Kawai, J. Ni, and R. H. Wang, “Minimum segmentation error based discriminative training for speech synthesis application,” in Proc. ICASSP 04, vol. 1, pp. 629-32, 2004.

[18] E. McDermott, T. Hazen, J. L. Roux, A. Nakamura, and S. Katagiri, “Discriminative training for large vocabulary speech recognition using minimum classification error,” IEEE Transactions on Speech and Audio Processing, vol. 15, no. 1, pp. 203-223, 2007.

[19] G. E. Dahl, D. Yu, L. Deng, and A. Acero, “Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition,”, IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, pp. 30-42, 2012.

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

[21] V. Nair and G. Hinton, “3-d object recognition with deep belief nets,” in Advances in Neural Information Processing Systems 22, pp. 1339—1347, 2009.

[22] R. Collobert and J. Weston, “A unified architecture for natural language processing: deep neural networks with multitask learning,” in Proc. 25th international conference on Machine learning, pp. 160—167, 2008.

[23] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, “Face recognition: A convolutional neural-network approach,”, IEEE Transactions on Neural Networks, vol. 8, pp. 98-113, 1997.

[24] 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.

[25] 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.