スパース基底空間上のマッピングに基づく声質変換*

◎高島遼一，滝口哲也，有木康雄 (神戸大)

1 はじめに

声質変換は，入力した音声を音韻情報などは保ったまま，話者性のような特定の情報のみを変換する技術であり，話者変換や感情変換 [1, 2]，発話支援 [3] など様々なタスクへの応用が期待されている．これまで声質変換のための統計的手法が多く提案されているが [4, 5, 6]，中でも Gaussian Mixture Model (GMM) を用いた手法 [6] が広く用いられており，多くの改良がされ続けている．

戸田ら [7] は従来の GMM を用いた声質変換法に動的特徴と Global Variance を導入することでより自然な音声として変換する手法を提案している．Helnder ら [8] は従来手法における過適合の問題を回避するため，Partial Least Squares (PLS) 回帰分析を用いる手法を提案している．また従来手法では，入力話者と出力話者が同じテキストを発話して得られるパラレルデータが必要であるが，このパラレルデータを使用せずに声質変換を行うために，GMM の話者適応を行う手法 [9] や Eigen-Voice GMM (EV-GMM)[10, 11] などが提案されている．

一方近年，信号処理の分野においてスパース表現に基づくアプローチが注目されており，音声信号処理の分野でも音声認識 [12, 13] や音源分離，雑音抑圧 [14, 15] などにおいて，その有効性が報告されている．スパース表現のアプローチでは，与えられた信号は少量の学習サンプルや基底の線形結合で表現される．音源分離に用いる場合，まず学習サンプルや基底を音源毎にグループ（辞書）化し，混合音声をそれらのスパース表現にする．その後，目的音声の辞書に対する重みベクトルのみを取り出して用いることで，目的音声のみを分離する．Gemmeke ら [13] は雑音の重畳した音声を，クリーン音声辞書とノイズ辞書のスパース表現にし，クリーン音声辞書に対する重みを音声認識における Hidden Markov Model (HMM) の尤度算出に用いることで，雑音にロバストな音声認識を行う手法を提案している．

我々はこれまで，音声のスパース表現に基づく声質変換手法を提案してきた [16]．この手法では，従来の声質変換手法で用いられていたパラレルデータから，入力話者の音声辞書（入力辞書）と出力話者の音声辞書（出力辞書）からなる同一発話内容のパラレル辞書を構築する．変換の際は，Non-negative Matrix Factorization (非負値行列因子分解，NMF)[17] を用いて，入力音声を入力辞書のスパース表現にする．そして得られた入力辞書内のサンプル毎の重み係数（アクティビティ）に基づいて，出力辞書内のサンプルを線形結合することで，出力話者の音声スペクトルへと変換する．また，この手法は入力音声に雑音が重畳している場合でも，従来の NMF を用いた雑音抑圧法と同様に，入力辞書に雑音辞書を結合してアクティビティを求めることで，雑音抑圧と声質変換を同時に行うことが可能である．

しかし，この手法はパラレルデータの各フレームをそのまま辞書のサンプルとして用いているため，辞書のサイズが膨大となり，変換のための計算時間がかかるという問題があった．そこで本稿では，入力音声と出力音声を同一のアクティビティで表現できるような部分空間を学習する NMF の枠組みを提案し，この空間上でマッピングを行うことで声質変換を行う手法を提案する．

2 スパース表現に基づく声質変換

本章では，これまで我々が提案してきた，音声のスパース表現に基づく声質変換法について述べる [16]．スパース表現のアプローチでは，与えられた信号は少量の学習サンプルや基底の線形結合で表現される．

$$\mathbf{x}_l \approx \sum_{j=1}^{J} \mathbf{a}_j h_{j,l} = \mathbf{A}\mathbf{h}_l \tag{1}$$

$\mathbf{x}_l$ は観測信号の l 番目のフレームにおける D 次元の特徴量ベクトルを表す．$\mathbf{a}_j$ は j 番目の学習サンプル，あるいは基底を表し，$h_{j,l}$ はその結合重みを表す．この手法では $\mathbf{a}_j$ は学習サンプルそのものを表す．学習サンプルを並べた行列 $\mathbf{A} = [\mathbf{a}_1 \dots \mathbf{a}_J]$ は “辞書” と呼び，重みを並べたベクトル $\mathbf{h}_l = [h_{1,l} \dots h_{J,l}]^T$ は “アクティビティ” と呼ばれる．このアクティビティベクトル $\mathbf{h}_l$ がスパースであるとき，観測信号は重みが非ゼロである少量の学習サンプルのみで表現されることになる．フレーム毎の特徴量ベクトルを並べて表現すると (1) 式は二つの行列の内積で表される．

$$\mathbf{X} \approx \mathbf{A}\mathbf{H} \tag{2}$$

$$\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_L], \quad \mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_L]. \tag{3}$$

L はフレーム数を表す．

---

* Voice Conversion Based on Spectral Mapping on Sparse Space. by Ryoichi Takashima, Tetsuya Takiguchi, Yasuo Ariki (Kobe univ.)

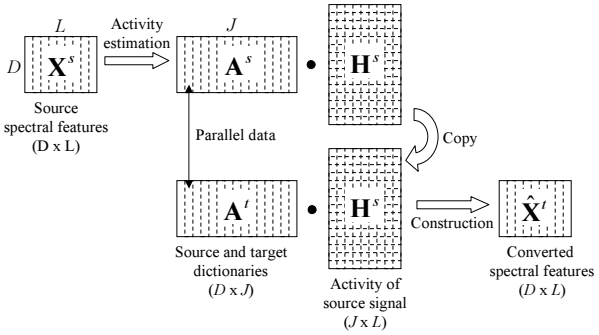

Fig. 1　スパース表現に基づく声質変換の概要

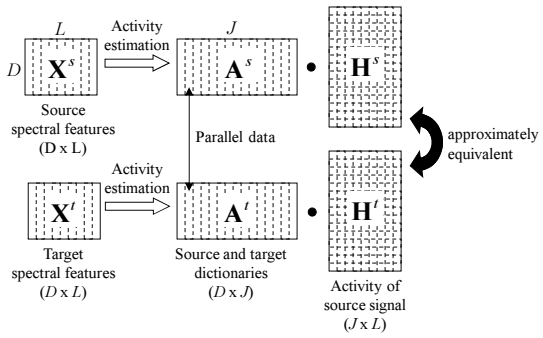

Fig. 2　入力辞書と出力辞書の並列性の仮定

スパース表現に基づく声質変換法の概要を Fig. 1 に示す．この手法では，パラレル辞書と呼ばれる入力音声辞書と出力音声辞書からなる辞書の対を用いる．この辞書の対は，従来の声質変換法と同様にパラレルデータに動的計画法 (DP) を適用することでフレーム間の対応を取った後，入力話者と出力話者の学習サンプルをそれぞれ並べて辞書化したものである．ここで，仮に入力話者の音声と，それと同一発話の出力話者の音声をそれぞれ入力辞書と出力辞書のスパース表現にした場合，それぞれから得られるアクティビティ行列は Fig. 2 のように互いに類似していると仮定する．スパース表現に基づく声質変換では，この仮定に基づき，入力話者の音声を入力辞書のスパース表現にし，得られたアクティビティ行列と出力辞書の内積を取ることで，出力話者の音声へと変換する．

この手法は入力音声に雑音が重畳している場合でも，従来の NMF を用いた雑音抑圧法と同様に，入力辞書 $\mathbf{X}^s$ に雑音辞書を結合してアクティビティ行列を求めることで，雑音抑圧と声質変換を同時に行うことが可能である．しかし，パラレルデータの各フレームをそのまま辞書のサンプルとして用いているため，辞書のサイズが膨大となり，変換のための計算時間がかかるという問題があった．

従来の NMF を用いた雑音抑圧法では，行列 $\mathbf{A}$ は学習データサンプルではなく，サンプル数よりもはるかに少量の基底ベクトルの集合で表現している．この基底行列はあらかじめ NMF により学習されている．しかし声質変換においては，辞書毎に NMF を行って独立に基底行列を学習すると，Fig. 2 のような同一のアクティビティ行列が得られるという辞書の並列性が失われてしまう．そこで本稿では，入力音声と出力音声を同一のアクティビティで表現できるような部分空間を学習する NMF の枠組みを提案し，この空間上でマッピングを行うことで声質変換を行う手法を提案する．

3　提案手法

3.1　並列制約付き NMF

本章では，Fig. 2 のように，入力音声と出力音声をそれぞれ二つの基底行列 $\mathbf{A}^s$，$\mathbf{A}^t$ のスパース表現にしたとき，得られるアクティビティ行列が一致するような基底行列の学習方法について述べる．この手法は従来の NMF（スパース制約付き NMF）の枠組みを拡張したものであり，次式のコスト関数を最小とするような $\mathbf{A}^s$，$\mathbf{A}^t$ を推定する．

$$d(\mathbf{X}^s, \mathbf{A}^s\mathbf{H}) + d(\mathbf{X}^t, \mathbf{A}^t\mathbf{H}) + ||(\lambda \mathbf{1}^{(1\times L)}).*\mathbf{H}||_1 \quad (4)$$

$$s.t.\ \ \mathbf{A}^s, \mathbf{A}^t, \mathbf{H} \geq 0. \quad (5)$$

第一項及び第二項は $\mathbf{X}$ と $\mathbf{AH}$ の Kullback-Leibler (KL) divergence である．第三項は $\mathbf{H}$ をスパースにするための L1 ノルム正則化項 [13] である．(4) 式の KL divergence を展開すると以下のようになる．

$$\begin{aligned} &\textstyle\sum_{i,j}\left(\mathbf{X}^s_{i,j}\log\frac{\mathbf{X}^s_{i,j}}{(\mathbf{A}^s\mathbf{H})_{i,j}} - \mathbf{X}^s_{i,j} + (\mathbf{A}^s\mathbf{H})_{i,j}\right) \\ &+\textstyle\sum_{i,j}\left(\mathbf{X}^t_{i,j}\log\frac{\mathbf{X}^t_{i,j}}{(\mathbf{A}^t\mathbf{H})_{i,j}} - \mathbf{X}^t_{i,j} + (\mathbf{A}^t\mathbf{H})_{i,j}\right) \\ &+||(\lambda \mathbf{1}^{(1\times L)}).*\mathbf{H}||_1 \end{aligned} \quad (6)$$

$\mathbf{X}_{i,j}$，$(\mathbf{AH})_{i,j}$ はそれぞれ行列 $\mathbf{X}$，$\mathbf{AH}$ の i 行 j 列の要素を表す．

ここで，$\mathbf{Z}^T = [\mathbf{X}^{s^T}\mathbf{X}^{t^T}]$，$\mathbf{W}^T = [\mathbf{A}^{s^T}\mathbf{A}^{t^T}]$ と定義すると (T は転置を表す)，(6) 式は

$$\begin{aligned} &\textstyle\sum_{i,j}\left(\mathbf{Z}_{i,j}\log\frac{\mathbf{Z}_{i,j}}{(\mathbf{WH})_{i,j}} - \mathbf{Z}_{i,j} + (\mathbf{WH})_{i,j}\right) \\ &+||(\lambda \mathbf{1}^{(1\times L)}).*\mathbf{H}||_1 \end{aligned} \quad (7)$$

となり，最終的にコスト関数 (4),(5) 式は．

$$d(\mathbf{Z}, \mathbf{WH}) + ||(\lambda \mathbf{1}^{(1\times L)}).*\mathbf{H}||_1 \ \ s.t.\ \ \mathbf{W}, \mathbf{H} \geq 0. \quad (8)$$

として，通常のスパース制約付き NMF の目的関数と等価な関数で再定義できる．これは，従来の GMM に基づく声質変換法において，変換関数を学習するために，パラレルデータを連結させた super vector $\mathbf{Z}$ を用いて GMM を学習させるのと同様に，$\mathbf{Z}$ に NMF

を適用させることで，並列性を保ったまま二つの基底行列 $\mathbf{A}^s$，$\mathbf{A}^t$ を学習できることを示している．基底行列は連結させた行列 $\mathbf{W}$ の形のまま通常のスパース NMF を適用して求めても良いが，メモリ使用量削減のため以下の更新式により，$\mathbf{A}^s$ と $\mathbf{A}^t$ それぞれ分けて求められる．

$$\mathbf{A}^s \leftarrow \mathbf{A}^s .* (\mathbf{H}(\mathbf{X}^s ./ \mathbf{A}^s\mathbf{H})^T ./ \mathbf{H}\mathbf{1}^{(1\times D)})^T \quad (9)$$

$$\mathbf{A}^t \leftarrow \mathbf{A}^t .* (\mathbf{H}(\mathbf{X}^t ./ \mathbf{A}^t\mathbf{H})^T ./ \mathbf{H}\mathbf{1}^{(1\times D)})^T \quad (10)$$

$$\begin{aligned}\mathbf{H} \leftarrow\ & \mathbf{H} \\ .* \ & (\mathbf{A}^{sT}(\mathbf{X}^s ./ (\mathbf{A}^s\mathbf{H})) + \mathbf{A}^{tT}(\mathbf{X}^t ./ (\mathbf{A}^t\mathbf{H}))) \\ & ./ (\mathbf{A}^{sT}\mathbf{1}^{(1\times L)} + \mathbf{A}^{tT}\mathbf{1}^{(1\times L)} + \lambda\mathbf{1}^{(1\times L)}). \end{aligned} \quad (11)$$

$\mathbf{1}$ は全ての要素が 1 のベクトルである．スパース制約の重み $\lambda^T = [\lambda_1 \ldots \lambda_J]$ は本実験では全て 0.2 とした．

3.2 並列制約付き NMF を用いた声質変換

本手法では，STRAIGHT スペクトル [18] を特徴量として声質変換を行う．以前に提案していた手法では，辞書構築用のパラレルデータの全フレームを並べて辞書行列 $\mathbf{A}^s$，$\mathbf{A}^t$ として，声質変換を行っていた．本稿で提案する手法では，パラレルデータを基底学習用の観測値行列 $\mathbf{X}^s_{train}$，$\mathbf{X}^t_{train}$ として，パラレルデータのフレーム数よりも少ない基底 $\mathbf{A}^s$，$\mathbf{A}^t$ を学習する．その際，$\mathbf{X}^s_{train}$，$\mathbf{X}^t_{train}$ はフレーム毎に，周波数上での総和が 1 になるように正規化した後，(9)，(10)，(11) 式により $\mathbf{A}^s$，$\mathbf{A}^t$ を学習する．

変換の際は，入力音声 $\mathbf{X}^s$ 及び入力話者の基底行列 $\mathbf{A}^s$ について，フレーム毎，あるいは基底毎に，周波数上での総和で正規化を行う．

$$\mathbf{M} = \mathbf{1}^{(D\times D)}\mathbf{X}^s \quad (12)$$

$$\mathbf{X}^s \leftarrow \mathbf{X}^s ./ \mathbf{M} \quad (13)$$

$$\mathbf{A}^s \leftarrow \mathbf{A}^s ./ (\mathbf{1}^{(D\times D)}\mathbf{A}^s) \quad (14)$$

次に，通常のスパース制約付き NMF を用いて，$\mathbf{X}^s$ を入力話者の基底 $\mathbf{A}^s$ のスパース表現にし，アクティビティ行列 $\mathbf{H}^s$ を得る．この際，NMF において $\mathbf{A}^s$ の更新は行わず，$\mathbf{H}^s$ のみ以下の更新式を用いて推定する [16].

$$\begin{aligned}\mathbf{H}^s \leftarrow\ & \mathbf{H}^s .* (\mathbf{A}^{sT}(\mathbf{X}^s ./ (\mathbf{A}^s\mathbf{H}^s))) \\ & ./ (\mathbf{1}^{(J\times L)} + \lambda\mathbf{1}^{(1\times L)}). \end{aligned} \quad (15)$$

J は基底数を表す．

次に，推定されたアクティビティ行列 $\mathbf{H}^s$ と出力話者の基底を用いて変換音声の生成を行う．このとき，出力話者の基底も入力話者の基底と同様に，周波数上での総和で正規化しておく．

$$\mathbf{A}^t \leftarrow \mathbf{A}^t ./ (\mathbf{1}^{(D\times D)}\mathbf{A}^t) \quad (16)$$

次に，正規化された出力話者の基底と $\mathbf{H}^s$ の内積を取り，(12) 式であらかじめ計算しておいた入力音声の振幅をかけることで，変換音声のスペクトルが生成される．

$$\hat{\mathbf{X}}^t = (\mathbf{A}^t\mathbf{H}^s) .* \mathbf{M} \quad (17)$$

最後に，変換された STRAIGHT スペクトルから変換音声を合成する．本稿では，このとき F0 情報は従来の単回帰分析により変換し，非周期成分は変換せず入力音声のものをそのまま用いている．

4 評価実験

4.1 実験条件

本実験では従来の GMM を用いた手法 [6] と，以前に提案したパラレルデータ全てを辞書に用いる手法 [16] (サンプルベースによる手法と呼ぶことにする) と比較を行った．ATR 研究用日本語音声データベースより，男性話者 1 名の音声を入力話者音声に，女性話者 1 名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．

音素バランス 216 単語を学習データとし，従来手法における GMM の学習，サンプルベースの手法における辞書の構築，提案手法における基底の学習にそれぞれ用いた．評価には 49 文を用いた．

サンプルベースによる手法及び提案手法では STRAIGHT スペクトル 512 次元を特徴量とした．各話者の学習データの総フレーム数は 28,517 であり，サンプルベースによる手法においてはこれが各話者の辞書のサンプル数に相当する．提案手法では，基底数を 1,000 として，各話者の基底行列を学習させた．それぞれの手法において NMF の更新回数は 300 とした．GMM を用いた従来手法では，STRAIGHT スペクトルから計算されたケプストラム 40 次元を特徴量とした．GMM の混合数は 64 である．

4.2 実験結果

各手法における変換音声それぞれのメルケプストラム歪みを Table 1 に示す．表より，提案手法はサンプルベースによる手法 (表中 Exemplar-based) と比べてメルケプストラム歪みが 0.3 dB 増加しているが，従来の GMM による手法よりも低い歪み値となっている．また，提案手法におけるアクティビティ行列の推定速度はサンプルベースの手法と比較して約 24 倍高速になっていた．このことから，提案手法は従来手法よりも高い変換性能を保ちつつ，以前に提案した手法よりも高速に変換が行えたと言える．

Table 1　各手法における変換音声のメルケプストラム歪み

	GMM	Exemplar-based	Proposed
Mel-CD [dB]	7.45	6.42	6.72

5　おわりに

本稿では，これまでに提案してきた音声のスパース表現に基づく声質変換法において，入力音声と出力音声を同一のアクティビティで表現できるような部分空間を学習する NMF の枠組みを提案し，この空間上でマッピングを行うことで声質変換を行う手法を提案した．部分空間の学習には，従来の GMM による変換と同様に，パラレルデータを結合した super vector の集合に対して通常の NMF を適用することで，並列性を保ったまま二つの基底行列を学習することができる．従来の GMM による手法と以前に提案したサンプルベースの手法と比較を行ったところ，サンプルベースの手法には変換精度は劣るものの，GMM よりも高い変換精度を示しており，またサンプルベースの手法よりも高速に変換を行うことができた．

今後は雑音辞書を結合することで雑音重畳音声の分離と声質変換を行う手法の性能を評価し，またセグメント特徴を導入することで動的変化を考慮した変換手法についてもを検討を行う．

参考文献

[1] Y. Iwami, T. Toda, H. Saruwatari, and K. Shikano, “GMM-based Voice Conversion Applied to Emotional Speech Synthesis,” IEEE Trans. Seech and Audio Proc., Vol. 7, pp. 2401–2404, 1999.

[2] C. Veaux and X. Robet, “Intonation conversion from neutral to expressive speech,” in Proc. INTERSPEECH, 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] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, “Vice conversion through vector quantization,” in Proc. ICASSP, pp. 655–658, 1988.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] T. N. Sainath, D. Nahamoo, B. Ramabhadran, and D. Kanevsky, “Enhancing Exemplar-Based Posteriors for Speech Recognition Tasks,” in Proc. INTERSPEECH, 4pages, 2012.

[13] J. F. Gemmeke, T. Viratnen, and A. Hurmalainen, “Exemplar-Based Sparse Representations for Noise Robust Automatic Speech Recognition,” IEEE Trans. Audio, Speech, Lang. Process., Vol. 19, Issue 7, pp. 2067–2080, 2011.

[14] T. Virtanen, “Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria,” IEEE Trans. Audio, Speech, Lang. Process., Vol. 15, No. 3, pp. 1066–1074, 2007.

[15] M. N. Schmidt and R. K. Olsson, “Single-channel speech separation using sparse non-negative matrix factorization,” in Proc. INTERSPEECH, pp. 2614-2617, 2006.

[16] R. Takashima, T. Takiguchi, and Y. Ariki, “Exemplar-Based Voice Conversion in Noisy Environment,” in Proc. SLT, pp. 313–317, 2012.

[17] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in Proc. Neural Information Processing System, pp. 556–562, 2001.

[18] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigne, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,” Speech Communication, Vol.27, pp. 187–207, 1999.