辞書選択に基づく非負値行列因子分解による声質変換*

☆相原龍, 中鹿亘, 滝口哲也, 有木康雄 (神戸大)

1　はじめに

声質変換とは，入力された音声に含まれる話者性・音韻性・感情性などといった多くの情報の中から，特定の情報を維持しつつ他の情報を変換する技術である．音韻情報を維持しつつ話者情報を変換する “話者変換” [1] を目的として広く研究されてきたが，近年では，音声合成や音声認識における話者性の制御 [2] に用いられている他，感情情報を変換する “感情変換” [3, 4]，失われた話者情報を復元する “発話支援” [5] など多岐にわたって応用されている．本研究では，雑音環境下での声質変換など，これまでになかったタスクに対応可能な非負値行列因子分解（Non-negative Matrix Factorization : NMF）による声質変換 [6] を扱う．従来の NMF による声質変換に辞書選択を導入することで，変換精度の向上を目指した．

従来，声質変換においては統計的な手法が多く提案されてきた．なかでも混合正規分布モデル (Gaussian Mixture Model : GMM) を用いた手法 [1] はその精度のよさと汎用性から広く用いられており，多くの改良がされ続けられている．基本的には，変換関数を目標話者と入力話者のスペクトル包絡の期待値によって表現し，変数をパラレルな学習データから最小二乗法で推定する．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．Helander 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．またこれらの手法では，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．

我々はこれまで，従来の統計的手法とは異なる，スパース表現に基づく Exemplar-based な声質変換手法を提案してきた [6]．スパース表現に基づくアプローチは信号処理の分野において注目されており，音声信号処理の分野でも音声認識や音源分離，雑音抑圧などにおいて，その有効性が報告されている．このアプローチでは，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．例えば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 [12] 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．

本研究では，スパースコーディングの代表的な手法として NMF [13] を用いる．我々の提案している声質変換手法では，従来の声質変換手法でも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（入力話者辞書）と出力話者の音声辞書（出力話者辞書）からなる同一発話内容のパラレル辞書を構築する．変換時には，入力音声を NMF によって，入力辞書に含まれる少量の基底からなるスパース表現にする．得られた入力辞書の基底毎の重み係数（アクティビティ）に基づいて，入力話者辞書の基底を出力辞書内の基底と置き換え，線形結合することで，出力話者の音声へと変換する．従来の声質変換のように統計的モデルを用いない Exemplar-based な手法であるため，過学習がおこりにくく，自然性の高い音声へと変換可能であると考えられる．

さらに本手法では，従来の NMF を用いた雑音抑圧法を組み込み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が同時に可能である．この性質を利用して，これまで研究されてこなかった雑音環境下での声質変換を提案した [6]．また，NMF はクラスタリング手法でもあり，入力辞書行列の音素ラベルが既知であれば，音素識別が可能である．この性質を利用して，発話が不明瞭になりやすい脳性麻痺型の構音障害者を対象とした声質変換を提案した [14]．以上のように，NMF を用いた声質変換手法は従来手法にはなかった様々なタスクに応用可能な手法である．

本稿では，声質変換においてもっとも一般的な，音声スペクトルを特徴量とした話者変換をタスクとし，NMF を用いた声質変換手法の精度を向上させるため，辞書選択手法の導入を提案する．これまではパラレルデータの全フレームをそのまま辞書の基底として用いており，辞書のサイズが膨大となっていた．そのため，入力音声のフレームと，入力話者辞書から選ばれる基底の音素が必ずしも一致しないといった問題があった．そこで本稿では，入力・出力話者辞書を音素カテゴリに分けた副辞書を作成する．NMF を用いて音素カテゴリ認識を行い，選択した副辞書上でマッピングを行うことで声質変換を行う．

以下，第 2 章でこれまでの NMF による声質変換手法を述べ，第 3 章で本稿の提案手法を説明する．第 4 章で従来の GMM・NMF による声質変換手法と比較し，第 5 章で本稿をまとめる．

2　NMF による声質変換

スパースコーディングの考え方において，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．

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

$\mathbf{x}_l$ 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．本手法では学習サンプルそのものを基底 $\mathbf{a}_j$ とする．基底を並べた行列 $\mathbf{A} = [\mathbf{a}_1 \dots \mathbf{a}_J]$ 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” と呼ぶ．このアクティビティベクトル $\mathbf{h}_l$ が

*Voice Conversion Based on Non-negative Matrix Factorization Using Phoneme Categorized Dictionary.
by Ryo AIHARA, Toru NAKASHIKA, Tetsuya TAKIGUCHI, Yasuo ARIKI (Kobe University)

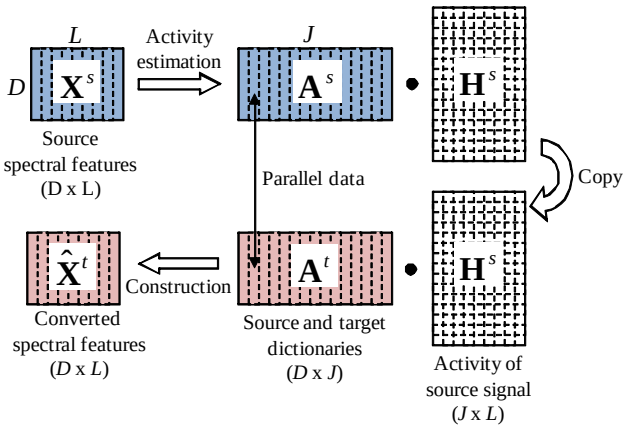

Fig. 1　Basic approach of NMF-based voice conversion

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基底ベクト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式 (1) は二つの行列の内積で表される．

$$\mathbf{X} \approx \mathbf{AH} \quad (2)$$

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_L], \quad \mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_L]. \quad (3)$$

ここで L はフレーム数を表す．

本手法の概要を Fig. 1 に示す．この手法では，パラレル辞書と呼ばれる入力話者音声辞書と出力話者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，入力話者と出力話者による同一発話内容の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．

このとき，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互いに類似していると仮定できる．Fig. 2 は 2 人の話者による発話単語 “ikioi”を入力信号としたときの，NMF で推定したアクティビティ行列である．ただし，簡易な例を示すため，辞書行列は 2 話者間でアライメントがとられた 1 単語のみから構成されている．Fig. 2 を見ると，高い重みを持つ基底の位置が類似していることがわかる．このことから，辞書行列がパラレルであれば，入力話者の辞書行列を用いて推定された入力特徴量のアクティビティは出力特徴量のアクティビティとして置き換え可能であると考えられる．以上の仮定に基づき，入力音声は入力話者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話者辞書の内積をとることで，出力話者の音声へと変換する．本手法では，アクティビティ行列の推定にスパースコーディングの代表的手法である NMF を用いる．

3　辞書選択による NMF 声質変換

これまでの NMF による声質変換法では，学習サンプル全てを基底としてパラレルな辞書を構成していた．結果，辞書に含まれる基底数が膨大になり，入力信号が値の小さなアクティビティから成る多数の基底の線形結合で表現されてしまうという問題があった．

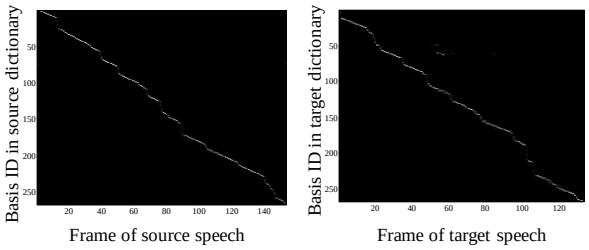

Fig. 2　Activity matrices for parallel utterances

線形結合の基底の数が増加すると，パラレルな発話のアクティビティの形状が類似するという仮定が成り立たなくなり，これが変換音声の劣化につながっていると考えた．そこで，入力信号を表現する基底数を限定するため，音素カテゴリごとに副辞書を作成し，限定された基底内でスパースコーディングを行う．

3.1　辞書構成法

Fig. 3 は副辞書の構成法を示したものである．従来の NMF による声質変換と同様にして入力話者辞書 $\mathbf{A}^s$ と出力話者辞書行列 $\mathbf{A}^t$ を求める．次にパラレルな辞書行列は，Table 1 に示す音素カテゴリに従って，K 個 の副辞書に分けられる．

さらに，それぞれの副辞書を表現する代表基底を集め，副辞書を選択するためのカテゴライズ辞書 $\boldsymbol{\Theta}$ を作成する．k 番目の入力話者副辞書を $\boldsymbol{\Phi}_k^s$ とするとき，その副辞書内の確率分布を仮定し，対応するカテゴライズ辞書の基底 $\boldsymbol{\theta}_k \in \boldsymbol{\Theta}$ を以下のように表す．

$$p(\boldsymbol{\Phi}_k^s) = \sum_{m=1}^{M} \alpha_m N(\boldsymbol{\Phi}_k^s, \boldsymbol{\mu}_m^{(k)}, \boldsymbol{\Sigma}_m^{(k)}) \quad (4)$$

$$\boldsymbol{\theta}_k = [\boldsymbol{\mu}_1^{(k)}, \ldots, \boldsymbol{\mu}_M^{(k)}] \quad (5)$$

$$\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K] \quad (6)$$

ここで，M，α，$\boldsymbol{\mu}_m^{(k)}$，$\boldsymbol{\Sigma}_m^{(k)}$ は正規分布の混合数，混合重み，平均，分散である．

Table 1　Category for sub dictionary

Category	phoneme
a	a
e	e
i	i
o	o
u	u
plosive	p, t, k, b, d, g, s
fricative	s, h, z
nasals	m, n, N
semi-vowel	j, w
liquid	r

3.2　辞書選択・変換法

Fig. 4 は変換手法の流れを示したものである．変換する入力信号 $\mathbf{X}^s$ はカテゴライズ辞書 $\boldsymbol{\Theta}$ とアクティビティ $\mathbf{H}_{\boldsymbol{\Theta}}^s$ によって，カテゴライズ辞書の基底とその

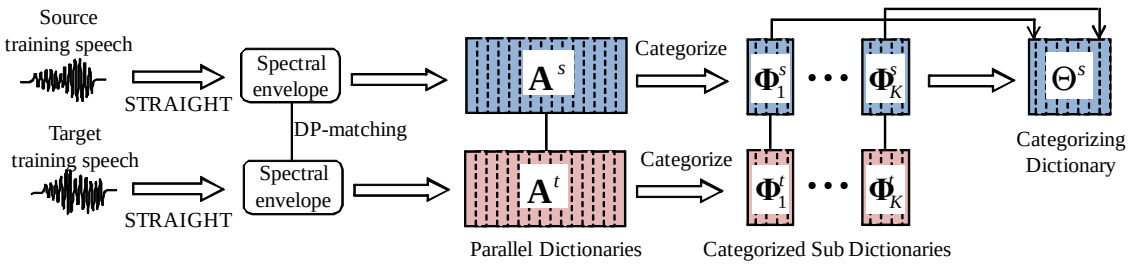

Fig. 3　Making sub dictionary

重みの線形和で表される．

$$\mathbf{X}^s \approx \mathbf{\Theta}\mathbf{H}^s_{\mathbf{\Theta}} \quad s.t. \quad \mathbf{H}^s_{\mathbf{\Theta}} \geq 0 \tag{7}$$

$$\mathbf{X}^s = [\mathbf{x}^s_1, \ldots, \mathbf{x}^s_L] \tag{8}$$

$$\mathbf{H}^s_{\mathbf{\Theta}} = [\mathbf{h}^s_{\mathbf{\Theta},1}, \ldots, \mathbf{h}^s_{\mathbf{\Theta},L}] \tag{9}$$

$$\mathbf{h}^s_{\mathbf{\Theta},l} = [h^s_{\mathbf{\Theta}_1,l}, \ldots, h^s_{\mathbf{\Theta}_K,l}]^T \tag{10}$$

つづいて入力信号のフレーム $\mathbf{x}^s_l$ は，対応する入力話者副辞書 $\mathbf{\Phi}^s_{\hat{k}}$ が以下のように選ばれ，その副辞書内の基底の線形和で表現される．

$$\hat{k} = \arg\max_k \sum_n h^s_{\mathbf{\Theta}_k,l} / h^s_{\mathbf{\Theta}_n,\,1} \tag{11}$$

$$\mathbf{x}^s_l = \mathbf{\Phi}^s_{\hat{k}} \mathbf{h}_{l,\hat{k}} \tag{12}$$

推定されたアクティビティと出力話者副辞書 $\mathbf{\Phi}^t_{\hat{k}}$ により，変換された特徴量は以下のように表現できる．

$$\hat{y^s_l} = \mathbf{\Phi}^t_{\hat{k}} \mathbf{h}_{l,\hat{k}} \tag{13}$$

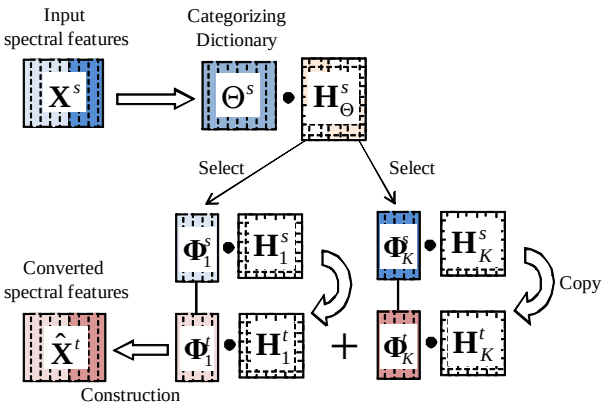

Fig. 4　NMF-based voice conversion using categorized dictionary

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 (サンプルベースによる手法と呼ぶ) と比較を行った．ATR 研究用日本語音声データベース [15] 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，提案手法における辞書の構築にそれぞれ用いた．評価には学習には含まれない 100 単語を用いた．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 [16] と前後 2 フレームを含む 2,565 次元特徴量とした．カテゴライズ辞書の基底を求める際の混合数は，副辞書の基底数の 1/500 とした．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 mfcc+Δmfcc+ΔΔmfcc の 64 次元を特徴量とした．GMM の混合数は 64 である．本稿では，提案手法・従来手法ともに 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．

提案手法の有効性を確かめるため，客観評価と主観評価を行った．客観評価は STRAIGHT スペクトル 516 次元を特徴量とし，式 (14) で表される NSD (Normalized Spectrum Distortion) [17] によって各手法を比較した．

$$NSD = \sqrt{\frac{||S^Y - S^{\hat{X}}||^2}{||S^Y - S^X||^2}} \tag{14}$$

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．主観評価は自然性と話者性の 2 つの項目について，10 人の主観評価実験を行った．自然性の評価基準は MOS 評価基準 [18] に基づく 5 段階評価（5:とてもよい，4:よい，3:ふつう，2:わるい，1:とてもわるい）とした．話者性については，出力話者の音声を聴いた後，各変換音声を聴き比べてどちらが出力話者の音声に似ているかを選択する XAB 法とした．

4.2　実験結果・考察

各手法における変換音声の NSD を Fig. 5 右側に示す．サンプルベースによる手法（NMF）は，従来の GMM による手法と比べてスペクトル歪みが大きくなっているが，提案手法は GMM よりも小さくなっている．このことから，従来の NMF による声質変換に辞書選択を導入することにより変換精度が向上したことがわかる．

自然性の主観評価結果を Fig. 5 左側に示す．従来の GMM による手法とサンプルベースの手法 (NMF) の間には有意な差は見られないが，提案手法はこれらを上回る結果となっている．

話者性の主観評価結果を Fig. 6 に示す．こちらも従来の GMM による手法とサンプルベースの手法 (NMF) の間には有意な差は見られないが，提案手法は従来のサンプルベースの手法 (NMF) よりも話者性が向上していることが分かる．

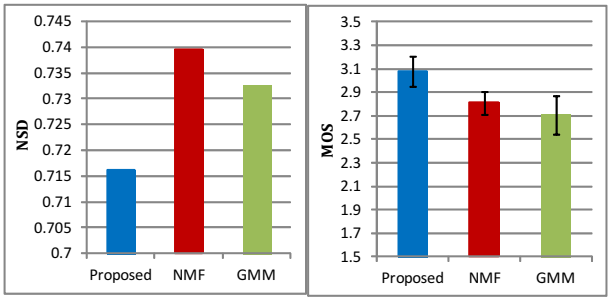

Fig. 5 Normalized spectrum distortion calculated from converted speech using each method (left), and results of MOS test (right)

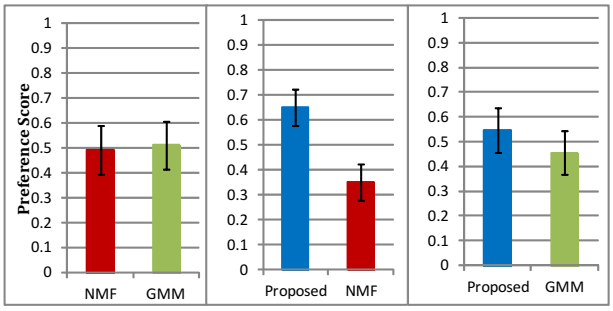

Fig. 6 Preference scores for the similarity

5 おわりに

本稿では，これまで提案してきたNMFに基づく声質変換法において，音素クラスタに分けた辞書選択を導入した．これにより，入力音声を音素的に近い基底のスパース性の高い線形和で表現することを可能にし変換精度を向上させた．主観・客観評価実験を行い，従来の統計的モデルを用いた声質変換法やNMFを用いたサンプルベースの手法よりも高い精度で変換できることを示した．

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する．また，構音障害者のための発話支援にも応用していく．

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] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in ICASSP, vol. 1, pp. 285-288, 1998.

[3] C. Veaux and X. Robet, "Intonation conversion from neutral to expressive speech," in Interspeech, pp. 2765–2768, 2011.

[4] R. Aihara *et al.*, "GMM-based emotional voice conversion using spectrum and prosody features," American Journal of Signal Processing, vol. 2, no. 5, 2012.

[5] K. Nakamura *et al.*, "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech," Speech Communication, vol. 54, no. 1, pp. 134–146, 2012.

[6] R. Takashima *et al.*, "Exemplar-based voice conversion in noisy environment," in SLT, pp. 313–317, 2012.

[7] T. Toda *et al.*, "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 *et al.*, "Voice conversion using partial least squares regression," IEEE Trans. Audio, Speech, Lang. Process., vol. 18, Issue: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 Interspeech, pp. 2254–2257, 2006.

[10] T. Toda *et al.*, "Eigenvoice conversion based on Gaussian mixture model," in Interspeech, pp. 2446–2449, 2006.

[11] D. Saito *et al.*, "One-to-many voice conversion based on tensor representation of speaker space," in Interspeech, pp. 653–656, 2011.

[12] J. F. Gemmeke and T. Virtanen, "Noise robust exemplar-based connected digit recognition," in ICASSP, pp. 4546–4549, 2010.

[13] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," Neural Information Processing System, pp. 556–562, 2001.

[14] R. Aihara *et al.*, "Individuality-preserving voice conversion for articulation disorders based on Non-negative Matrix Factorization," in ICASSP, pp. 8037–8040, 2013.

[15] A. Kurematsu *et al.*, "ATR japanese speech database as a tool of speech recognition and synthesis," Speech Communication, vol. 9, pp. 357–363, 1990.

[16] H. Kawahara *et al.*, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequencybased F0 extraction: possible role of a repetitive structure in sounds," Speech Communication, vol. 27, no. 3-4, pp. 187–207, 1999.

[17] T. En-Najjary *et al.*, "A voice conversion method based on joint pitch and spectral envelope transformation," in ICSLP, pp. 199–203, 2004.

[18] INTERNATIONAL TELECOMMUNICATION UNION, "Methods for objective and subjective assessment of quality," ITU-T Recommendation P.800, 2003.