非負値タッカー分解によるNMF辞書学習に基づく非パラレル声質変換*

☆高島悠樹, 矢野肇 (神戸大), 中鹿亘 (電通大), 滝口哲也 (神戸大/JST さきがけ), 有木康雄 (神戸大)

1 はじめに

音声信号処理の分野の中でも，声質変換技術が様々なタスク [1] への応用が可能であることから近年盛んに研究されている．声質変換とは，入力話者音声の音韻情報を保存したまま，話者性に関する情報のみを出力話者のものへ変換させる技術である．これまでの声質変換法として，GMM (Gaussian mixture model) を用いた手法 [2] が最も広く用いられており，様々な改良がなされてきた [3]．その他の手法として，非負値行列因子分解 (non-negative matrix factorization; NMF) [4, 5] や，restricted Boltzmann machine (RBM) [6] に基づく手法が提案されてきた．

NMF [7] は，スパース行列分解手法の 1 つであり，入力信号を，基底行列と係数行列に分解する．NMF の目標は入力行列から，これら 2 つの行列を推定することである．本稿では，基底行列を辞書，係数行列をアクティビティと呼ぶ．NMF は 2 つの行列を同時に推定する辞書推定による手法 [5] と，辞書を Exemplar で固定しアクティビティのみを推定する Exemplar-based の手法 [4] に分けることができる．辞書推定による手法は，コンパクトな辞書を推定することができるため計算コストを削減できるが，アクティビティのみならず辞書基底もスパースになる傾向があるため，音声のフォルマント構造が壊れてしまい高い精度が得られないという問題点があった．Exemplar-based 手法ではそのような現象を防ぐことができるが，辞書推定による手法と比較して，計算コストが高く，分解精度誤差も大きくなるという問題点がある．

NMF 声質変換はこれまで Exemplar-based によるものがほとんどであった．しかしながら，Exemplar-based による手法は，モデルの学習時にパラレルデータを必要とする．パラレルデータとは，入力話者と出力話者の，同一発話内容による音声対であり，パラレルデータの作成には様々な制限が課せられる．第一に，発話データは同一の発話内容でないといけないという制限があるため，選択 (または作成) できる学習データセットの自由度は低い．第二に，フレーム単位で入出力音声の同期を取る必要があるため，動的計画法などを用いてアライメントを取るが，完全にフレームの同期が取れている保証がない，伸縮の際に音声に変換が加わっているなどの問題がある．Exemplar-based による手法において，辞書はパラレルデータから構成されるため，パラレルデータのアライメント誤差が声質変換性能に悪影響を及ぼす可能性がある．

入出力話者間のパラレルデータを必要としない，若しくは少量のパラレルデータを用いて，話者性を柔軟に制御するアプローチもいくつか提案されている [8]．例えば文献 [8] では，参照話者のパラレルデータを用いて二話者間の関係性を GMM でモデル化しておき，入力話者 (もしくは出力話者) を参照話者の特徴空間へ射影する行列を求めるため，入力話者-出力話者間のパラレルデータは必要としない (しかしながら，参照話者の間でパラレルデータを必要とする)．本稿では，従来の NMF に基づく声質変換をパラレルデータを使わない手法へ拡張する．

本稿では，パラレルデータを使用しない NMF 声質変換手法として，非負値タッカー分解 (non-negative Tucker decomposition; NTD) [9] に基づく辞書学習法を提案する．NTD はタッカー分解を非負拡張したものであり，入力信号を複数の行列と 1 つのコアテンソルに分解する．入力信号を 2 次元のスペクトル特徴量とした場合，NTD は，周波数と時間に対する 2 つのモード行列と 1 つのコア行列に分解する．我々はこれらの行列がそれぞれ，周波数基底行列，音韻情報，周波数基底と各音韻を対応づけるコードブックを表現すると仮定する．さらにこの仮定のもとで，従来の NMF におけるアクティビティがコードブックと音韻情報に分解されたと仮定する．辞書学習時に，従来の NMF ではパラレルデータを用いて話者間のアクティビティが共有されていたが，提案手法では，コードブックは話者間で共有し，音韻情報は話者依存項として学習される．これにより，時間変化のある非パラレルコーパスの音韻情報を話者ごとに扱うことが可能となる．変換時には，音韻情報の部分のみをアクティビティとして推定する．提案手法では，時間に依存する項を話者間で共有することなく，話者ごとに扱うためパラレルデータを必要としない学習が可能となる．

以下，第 2 章で先行研究について説明し，問題点を述べる．第 3 章で提案手法を説明する．第 4 章で評価実験を行い，第 5 章で本稿をまとめる．

---

*Parallel-Data-Free Dictionary Learning for Voice Conversion Using Non-negative Tucker Decomposition, by Yuki Takashima, Hajime Yano (Kobe University), Toru Nakashika (UEC), Tetsuya Takiguchi (Kobe University/JST PRESTO), Yasuo Ariki (Kobe University)

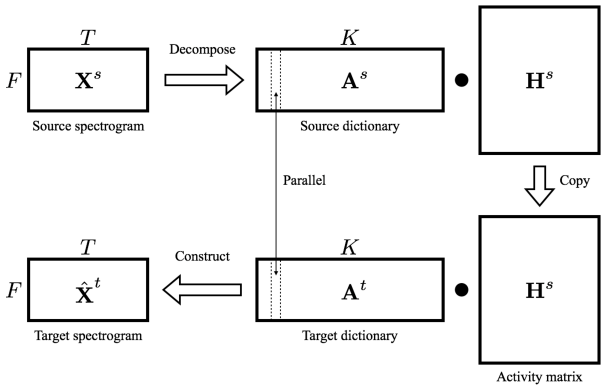

Fig. 1　Basic approach of NMF-based voice conversion

2　先行研究

2.1　NMF 声質変換

辞書学習による NMF 声質変換の概要を Fig. 1 に示す．$\mathbf{X}^s$ は入力話者スペクトル，$\mathbf{A}^s$ は入力話者辞書，$\mathbf{A}^t$ は出力話者辞書，$\mathbf{H}^s$ は入力話者スペクトルから推定されるアクティビティ，$\hat{\mathbf{X}}^t$ は変換されたスペクトルを表す．F，T，K はそれぞれ，スペクトルの次元数，フレーム数，辞書の基底数である．

この手法は，学習時に，入力話者スペクトル $\mathbf{X}^s$ と出力話者スペクトル $\mathbf{X}^t$ のパラレルデータを用いる．このパラレルデータは，入力話者と出力話者による同一発話内容の音声に dynamic time warping (DTW) を適用することでフレーム間の対応を取り作成される．

まず，入力スペクトル $\mathbf{X}^s$ に対して，NMF によって $\mathbf{A}^s$ と $\mathbf{H}^s$ が推定される．NMF のコスト関数を以下に示す．

$$d_{KL}(\mathbf{X}^s, \mathbf{A}^s\mathbf{H}^s) + \lambda||\mathbf{H}^s||_1 \ s.t.\ \mathbf{A}^s, \mathbf{H}^s \geq 0 \quad (1)$$

式 (1) において，第 1 項は $\mathbf{X}^s$ と $\mathbf{A}^s\mathbf{H}^s$ の間の Kullback-Leibler (KL) ダイバージェンスであり，第 2 項はアクティビティをスパースにするための L1 ノルム正則化項である．λ はスパース重みを表す．

次に，式 (1) より推定されたアクティビティ $\mathbf{H}^s$ を用いて，出力話者スペクトル $\mathbf{X}^t$ に対する出力話者辞書 $\mathbf{A}^t$ を推定する．$\mathbf{A}^t$ はアクティビティ $\mathbf{H}^s$ を固定して，以下のコスト関数で最適化される．

$$d_{KL}(\mathbf{X}^t, \mathbf{A}^t\mathbf{H}^s) \ s.t.\ \mathbf{A}^t \geq 0 \quad (2)$$

本手法では，「パラレル辞書で推定したパラレルな発話のアクティビティは置き換え可能である」と仮定している．従って，変換スペクトル $\hat{\mathbf{X}}^t$ は，推定された辞書 $\mathbf{A}^t$ とアクティビティ $\mathbf{H}^s$ の積によって得られる．

$$\hat{\mathbf{X}}^t = \mathbf{W}^t\mathbf{H}^s \quad (3)$$

2.2　問題点

NMF に基づく声質変換はいくつかの問題点がある．まず，入力話者と出力話者の発話はあらかじめ DTW によりアライメントを取るため，推定されたモデルパラメータはこのアライメントの精度に影響される．文献 [10] においては，アライメントのずれが引き起こす NMF 声質変換の精度劣化が指摘されている．次に，アクティビティ行列は音韻情報だけでなく，その他の情報も含むと考えられる．文献 [11] において，相原らはアクティビティ行列が音韻情報と話者情報を含むと仮定し，これらを扱うフレームワークを提案し，NMF 声質変換の性能を向上させた．本稿では，異なるアプローチとして，アクティビティ行列を，話者間で共有する行列と話者固有の行列に分解する．そして，話者固有の行列が音韻情報を表現すると仮定し，これをアクティビティとして変換を行う．これにより，入力スペクトルのフレーム長に依存する項を話者毎に持つため，学習時に入力話者と出力話者のパラレルデータを必要としない．

3　NTD を用いたパラレル辞書学習

3.1　非負値タッカー分解

N 階の非負テンソルが与えられた時，非負値タッカー分解 (NTD) は入力テンソルを，非負に制約された 1 つのコアテンソルと N 個のモード行列に分解する．本稿では，入力テンソルとして 2 次元のスペクトル特徴量を扱うため，コアテンソルは行列として表現され，モード行列の数は 2 となる．この条件下で，NTD は簡単に以下の式で表現される．

$$\mathbf{X} \approx \mathbf{U}\mathbf{G}\mathbf{V}^\top \ s.t.\ \mathbf{U} \geq 0, \mathbf{G} \geq 0, \mathbf{V} \geq 0 \quad (4)$$

ただし，$\mathbf{X} \in \mathbb{R}^{F\times T}$, $\mathbf{U} \in \mathbb{R}^{F\times M}$, $\mathbf{V} \in \mathbb{R}^{T\times L}$, $\mathbf{G} \in \mathbb{R}^{M\times L}$ はそれぞれ，入力スペクトル，周波数及び時間軸に対するモード行列，コア行列を表す．F, T, M, L はそれぞれ，周波数ビンの数，フレーム数，周波数基底及び時間基底の数である．NTD のコスト関数を以下に示す．

$$||\mathbf{X} - \mathbf{U}\mathbf{G}\mathbf{V}^\top||_F^2 \quad (5)$$

ただし，$||\cdot||_F$ はフロベニウスノルムを表す．NTD は，NMF を含む非負値テンソル分解の一般形であり，その更新則は文献 [12] で提案されている．この更新則は NMF の更新則に基づくものである．

3.2　NTD を用いたパラレル辞書学習

NTD を用いて，パラレルデータを用いないコンパクトな辞書を推定する．Fig. 2 に，提案する辞書学習

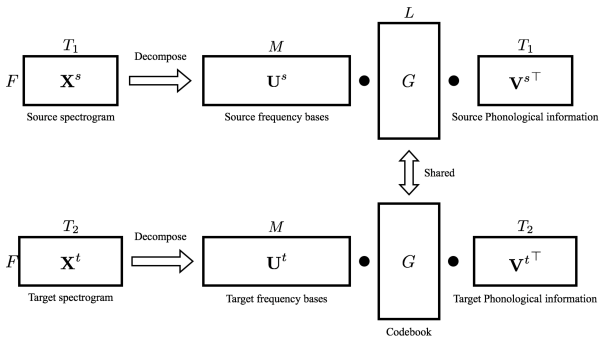

Fig. 2　Parallel dictionary learning using NTD

法の概要を示す．目的関数を次のように定義する．

$$||\mathbf{X}^s - \mathbf{U}^s\mathbf{G}\mathbf{V}^{s\top}||_F^2 + ||\mathbf{X}^t - \mathbf{U}^t\mathbf{G}\mathbf{V}^{t\top}||_F^2$$
$$s.t.\ \mathbf{U}^s \geq 0, \mathbf{U}^t \geq 0, \mathbf{G} \geq 0, \mathbf{V}^s \geq 0, \mathbf{V}^t \geq 0 \quad (6)$$

だたし，$\mathbf{X}^s \in \mathbb{R}^{F\times T_s}$, $\mathbf{X}^t \in \mathbb{R}^{F\times T_t}$, $\mathbf{U}^s \in \mathbb{R}^{F\times M}$, $\mathbf{U}^t \in \mathbb{R}^{F\times M}$, $\mathbf{V}^s \in \mathbb{R}^{T_s\times L}$, $\mathbf{V}^t \in \mathbb{R}^{T_t\times L}$, $\mathbf{G} \in \mathbb{R}^{M\times L}$ はそれぞれ，入力話者及び出力話者のスペクトル，周波数基底行列，時間基底行列，及び，コア行列を表す．F, T_s, T_t, M, L はそれぞれ，周波数ビンの数，入力話者及び出力話者スペクトルのフレーム数，周波数及び時間の基底数である．このコスト関数は，通常の NTD と同様の更新則により繰り返し最小化される．コア行列 G は以下の更新式により繰り返し更新される．

$$\mathbf{G} \leftarrow \mathbf{G}.*(\mathbf{U}^{s\top}\mathbf{X}^s\mathbf{V}^s + \mathbf{U}^{t\top}\mathbf{X}^t\mathbf{V}^t)$$
$$./(\mathbf{U}^{s\top}\mathbf{U}^s\mathbf{G}\mathbf{V}^{s\top}\mathbf{V}^s + \mathbf{U}^{t\top}\mathbf{U}^t\mathbf{G}\mathbf{V}^{t\top}\mathbf{V}^t) \quad (7)$$

ただし，$.*$ と $./$ はそれぞれ，要素積と要素商を表す．

我々は，$\mathbf{U}^s$，$\mathbf{U}^t$ が周波数基底行列，$\mathbf{V}^s$,$\mathbf{V}^t$ が音韻情報を表すと仮定する．さらに，コア行列 $\mathbf{G}$ は入力スペクトルの次元数に依存しない行列であり，コア行列が周波数基底と音素のコードブックを表現すると仮定する．この仮定のもとで，コア行列は周波数基底群と各音素を対応づける行列だと考えられる．L 個の音素があり，それぞれが M 個の周波数基底の重み付け和で表現されていると考えることができる．アクティビティ行列が持つ情報は音韻情報だけでないと考えられるが，従来の NMF 声質変換においては，アクティビティ行列が音韻情報のみを持つとして推定されていた．それに対して，提案手法では，アクティビティ行列を，話者共有の行列 $\mathbf{G}$ と話者固有の行列 $\mathbf{V}^s$ (入力話者の場合) に分解しているとみなすことができる．この分解により，入力スペクトルのフレーム長の違いを話者固有の行列で表現できるため，パラレルデータを使用しない学習を行うことができる．

モデルパラメータが推定された後，入力話者のパラレル辞書は以下の式で定義される．

$$\mathbf{W}^s = \mathbf{U}^s\mathbf{G} \quad (8)$$

出力話者辞書も同様に計算される．これらの辞書を用いて，入力スペクトルは 2.1 節で示した手法と同様にして変換される．

4　評価実験

4.1　実験条件

提案手法は，クリーン環境下での話者変換をタスクとし，従来のパラレル手法である GMM 声質変換 [2]，従来の辞書学習による NMF 声質変換 [5]，パラレルデータを用いない adaptive restricted Boltzmann machine (ARBM) に基づく声質変換 [6] と比較した．

ATR 研究用日本語音声データベースに含まれる男性 1 名を入力話者，女性 1 名を出力話者とした．サンプリング周波数は 12kHz である．音素バランス文 50 文を学習データとし，学習データに含まない音素バランス文 10 文をテストデータとして用いた．GMM 及び NMF に基づく手法は，dynamic programming matching (DPM) を用いて作成されたパラレルデータを用いた．NTD の更新回数は学習時及び変換時，共に 300 とした．これらのパラメータは実験的に求められたものである．

提案手法では，STRAIGHT スペクトル 513 次元を入力スペクトルとして用いた．周波数基底数 $M = 1,000$，時間基底数 $L = 200$ とした．GMM 声質変換において，STRAIGHT スペクトルから計算された 24 次元のメルケプストラムを特徴量として用いた．GMM の混合数は 64 とした．従来の辞書学習による NMF 声質変換において，辞書基底数を 1,000 とした．ARBM に基づく声質変換において，入力特徴量として，STRAIGHT スペクトルから計算された 32 次元のメルケプストラムを用い，隠れ層数は 128 とした．

本稿では，F0 には平均，分散を考慮した線形変換を適用し，非周期成分は入力発話のものを用いた．

4.2　実験結果

主観評価実験として，10 人の日本語話者が 10 文のテストデータについて，それぞれの手法で変換した音声を評価した．本論文では，話者性と自然性，明瞭性の 3 つの観点において評価実験を行った．それぞれ 5 段階評価 (5: excellent, 4: good, 3: fair, 2: poor, 1: bad) で評価した．

Fig. 3 に主観評価実験結果を示す．エラーバーは，95%信頼区間を示す．GMM と NMF はパラレルデータを用いる手法であり，ARBM と NTD はパラレルデータを用いない手法である．図より，パラレルデー

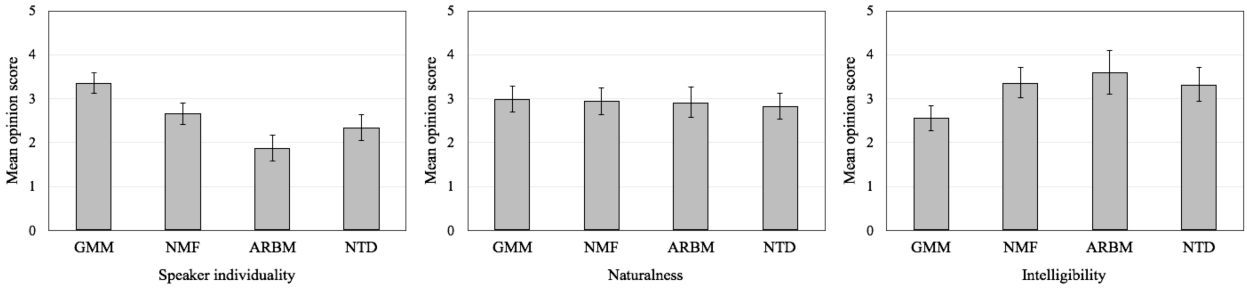

Fig. 3　Mean opinion scores (MOS) for each method

タを用いる手法は，用いない手法と比べて高い話者性が得られることが分かる．しかしながら，提案手法は ARBM と比較して有意に話者性を向上させている．さらに，明瞭性の評価において，提案手法は従来手法である GMM と比較して有意差が得られた．これらの結果は t 検定により有意であることが示されている．自然性の評価においては，全ての手法が同等の性能を示した．これらの結果より，提案手法は従来の非パラレル手法と比較して，効率的に話者性を変換することができる．しかしながら，従来の NMF 辞書学習法と比較して，話者性の劣化が見られる．この理由として，提案手法が非パラレル手法という点が挙げられる．また，他の理由として，提案手法のコスト関数が，NMF 辞書学習が持つスパース制約を含んでいないことが考えられる．NTD は NMF よりも複雑な分解であるため，何らかの制約を設けることで安定した変換が可能になると考えられる．また，提案法によるモデリングは，話者間で周波数基底行列がパラレルになっている保証がない．さらなる話者性向上のために，周波数基底行列をパラレルにするような制約が必要である．

5　おわりに

従来のパラレルデータを用いた NMF 声質変換の非パラレル拡張法として，NTD に基づくパラレルデータを用いない辞書学習法を提案した．実験結果により，NTD ベースの辞書学習は従来の NMF ベースの辞書学習と同等の声質変換性能を実現した．さらに，従来の非パラレル手法である ARBM よりも，話者性を向上させた．今後は，学習データの発話内容が話者間で異なる場合での評価を行う．

参考文献

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

[2] Y. Stylianou *et al.*, “Continuous probabilistic transform for voice conversion,” IEEE Trans. Speech and Audio Processing, vol. 6, no. 2, pp. 131–142, 1998.

[3] T. Toda *et al.*, “Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory,” IEEE Trans. Audio, Speech & Language Processing, vol. 15, no. 8, pp. 2222–2235, 2007.

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

[5] R. Takashima *et al.*, “Noise-robust voice conversion based on spectral mapping on sparse space,” in *SSW*, 2013, pp. 71–75.

[6] T. Nakashika *et al.*, “Non-parallel training in voice conversion using an adaptive restricted Boltzmann machine,” IEEE/ACM Trans. Audio, Speech & Language Processing, vol. 24, no. 11, pp. 2032–2045, 2016.

[7] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in *NIPS*, 2000, pp. 556–562.

[8] A. Mouchtaris *et al.*, “Nonparallel training for voice conversion based on a parameter adaptation approach,” IEEE Trans. Audio, Speech & Language Processing, vol. 14, no. 3, pp. 952–963, 2006.

[9] L. R. Tucker, “Some mathematical notes on three-mode factor analysis,” Psychometrika, vol. 31, pp. 279–311, 1966.

[10] R. Aihara *et al.*, “Voice conversion based on non-negative matrix factorization using phoneme-categorized dictionary,” in *ICASSP*, 2014, pp. 7894–7898.

[11] R. Aihara *et al.*, “Activity-mapping non-negative matrix factorization for exemplar-based voice conversion,” in *ICASSP*, 2015, pp. 4899–4903.

[12] Y.-D. Kim and S. Choi, “Nonnegative tucker decomposition,” in *CVPR*, 2007.