一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SP2016-38(2016-08)

Discriminative Graph-embedded Non-negative Matrix Factorization を用いた声質変換のためのパラレル辞書学習

相原　龍†　滝口　哲也††　有木　康雄††

† 神戸大学システム情報学研究科　〒 657-8501 兵庫県神戸市灘区六甲台町 1-1
†† 神戸大学自然科学系先端融合研究環　〒 657-8501 兵庫県神戸市灘区六甲台町 1-1
E-mail: †aihara@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

あらまし　これまで一般的であった，混合正規分布モデル（GMM）に代表される統計的声質変換に代わる手法として，非負値行列因子分解（NMF）に基づく Exemplar-based 声質変換が研究されてきた．NMF 声質変換は，GMM 声質変換と比較して自然性の高い変換音声が期待されているが，一方で計算時間とメモリ使用量が多いという問題をかかえていた．NMF を用いた信号処理の手法には，Exemplar-based と辞書推定による手法があるが，NMF 声質変換では Exemplar-based がほとんどであり，辞書推定によるものは少なかった．本稿では，NMF に識別的制約を導入した Discriminative Graph-embedded Non-negative Matrix Factorization (DGNMF) を提案し，DGNMF を用いたパラレル辞書学習を NMF 声質変換に導入する．従来手法で用いられていたパラレル Exemplar から，与えられた音素ラベルに従った識別的でコンパクトな辞書を学習する．辞書学習によって基底数の少ない辞書を推定することで，さらなる計算コストの削減を図るとともに，従来の NMF 声質変換で指摘されていたアクティビティのアライメント問題を解決することができる．提案手法は従来の NMF 声質変換の問題点を解決するとともに，超解像など，NMF を用いた他の手法にも応用可能であると考えられる．実験結果より，提案手法は従来の NMF 声質変換の自然性を向上させることができると同時に，計算コストを削減させることができることがわかった．
キーワード　声質変換，音声合成，非負値行列因子分解，スパース表現

Parallel Dictionary Learning for Voice Conversion Using Discriminative Graph-embedded Non-negative Matrix Factorization

Ryo AIHARA†, Tetsuya TAKIGUCHI††, and Yasuo ARIKI††

† Graduate School of System Informatics, Kobe University 1-1 Rokkodai, Nada, Kobe 657-8501, Japan
†† Organization of Advanced Science and Technology, Kobe University 1-1 Rokkodai, Nada, Kobe 657-8501, Japan
E-mail: †aihara@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

Abstract　This paper proposes a discriminative learning method for Non-negative Matrix Factorization (NMF)-based Voice Conversion (VC). NMF-based VC has been researched because of the natural-sounding voice it produces compared with conventional Gaussian Mixture Model (GMM)-based VC. In conventional NMF-based VC, parallel exemplars are used as the dictionary; therefore, dictionary learning is not adopted. In order to enhance the conversion quality of NMF-based VC, we propose Discriminative Graph-embedded Non-negative Matrix Factorization (DGNMF). Parallel dictionaries of the source and target speakers are discriminatively estimated by using DGNMF based on the phoneme labels of the training data. Experimental results show that our proposed method can not only improve the conversion quality but also reduce the computational time.
Key words　voice conversion, speech synthesis, NMF, sparse representation

1. はじめに

非負値行列因子分解 (Non-negative Matrix Factorization : NMF) はスパース行列分解手法のひとつであり，ハイパースペクトルイメージング [1]，トピックモデル [2]，脳波解析 [3] など幅広い分野に応用されている．入力信号 $\mathbf{V}$ は，辞書行列を $\mathbf{W}$，係数行列（アクティビティ）を $\mathbf{H}$ とすると，以下のような式で表される．

$$\mathbf{V} \approx \mathbf{WH}. \tag{1}$$

音声信号処理においても，NMF はシングルチャンネル音声分離 [4], [5]，歌声分離 [6] などに応用されてきた．NMF のアプローチには，辞書とアクティビティを同時推定する教師なし NMF と，事前に与えられた Exemplar を辞書として固定し，アクティビティのみを推定する Exemplar-based 手法の 2 つがある．Gemmeke ら [7] は Exemplar-based NMF を用いたノイズロバストな音声認識手法を提案しており，注目を集めている．

近年，Exemplar-based NMF は声質変換に応用されている [8], [9]．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” [10] を目的として広く研究されてきたが，感情情報を変換する “感情変換” [11]，失われた話者情報を復元する “発話支援” [12] など多岐にわたって応用されている．特に近年は音声合成技術の発達に伴い，音声合成における話者性の制御 [13]，スペクトル復元 [14] や帯域幅拡張 [15] などに応用され注目を集めている．

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 [10] はその精度のよさと汎用性から広く用いられており，多くの改良が行われている．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 [16] 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 [17] 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．

我々はこれまで，従来一般的であった統計的手法による声質変換 [16] とは異なる，スパース表現に基づく非負値行列因子分解 (Non-negative Matrix Factorization : NMF) [18] を用いた Exemplar-based 声質変換手法を提案してきた [8]．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” [10] を目的として混合正規分布モデル (Gaussian Mixture Model : GMM) を用いた手法を中心に広く研究されてきたが，NMF 声質変換は従来の声質変換のように統計的モデルを用いないため過学習がおこりにくいことに加え，高次元スペクトルを用いて変換するため，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，NMF 声質変換は，NMF によるノイズ除去手法と組み合わせることでノイズロバスト性を有する．

NMF は $\mathbf{W}$ と $\mathbf{H}$ を同時に推定する辞書推定による手法と，$\mathbf{W}$ を Exemplar で固定し $\mathbf{H}$ のみを推定する Exemplar-based の手法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．本稿では，NMF を声質変換における分解誤差削減による精度向上を目指し，識別的な Graph-embedded NMF (Discriminative Graph-embedded Non-negative Matrix Factorization: DGNMF) を提案し，この手法を用いたパラレル辞書学習を行う．類似した手法として Graph Regularized NMF (GRNMF) [19] が提案されているが，GRNMF は音素ラベルに基づく識別がないうえ，クラス間分散は考慮しておらず，真に識別的な学習が行われているとはいえない．DGNMF は音素ラベルを考慮し，クラス内クラス間分散制約に基づいて辞書を推定する識別的 NMF だといえる．

以下，第 2 章で従来の NMF は声質変換について説明し，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．

2. NMF 声質変換

2.1 概　　要

概要を図 1 に示す．$\mathbf{V}^s$ は入力話者スペクトル，$\mathbf{W}^s$ は入力話者辞書，$\mathbf{W}^t$ は出力話者辞書，$\hat{\mathbf{V}}^t$ は変換されたスペクトル，$\mathbf{H}^s$ は入力話者スペクトルから推定されるアクティビティを表す．D，J はそれぞれスペクトルの次元数，辞書の基底数である．この手法では，パラレル辞書と呼ばれる入力話者辞書 $\mathbf{W}^s$ と出力話者辞書 $\mathbf{W}^t$ 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 Dynamic Time Warping (DTW) 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たものである．

入力スペクトル $\mathbf{V}^s$ は NMF によって $\mathbf{W}^s$ と $\mathbf{H}^s$ の積に分解される．NMF のコスト関数を以下に示す．

$$d_{KL}(\mathbf{V}^s, \mathbf{W}^s\mathbf{H}^s) + \lambda||\mathbf{H}^s||_1 \quad s.t. \quad \mathbf{H}^s \geqq 0 \tag{2}$$

式 (2) において，第 1 項は $\mathbf{V}^s$ と $\mathbf{W}^s\mathbf{H}^s$ 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．辞書は固定し，アクティビティのみを推定する．コスト関数は，次のような更新式を用いて最小化される．

$$\mathbf{H}^s \leftarrow \mathbf{H}^s .* (\mathbf{W}^{s\mathsf{T}}(\mathbf{V}^s ./ (\mathbf{W}^s\mathbf{H}^s))) \\ ./ (\mathbf{W}^{s\mathsf{T}}\mathbf{1}^{(I\times J)} + \lambda\mathbf{1}^{(K\times J)}) \quad (3)$$

ここで，$.*$，$./$，$\mathbf{1}$ は要素積，要素商，全要素が 1 の配列を表す．

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 $\hat{\mathbf{V}}^t$ は，$\mathbf{W}^t$ と推定した $\mathbf{H}^s$ の積によって得られる．

$$\hat{\mathbf{V}}^t = \mathbf{W}^t\mathbf{H}^s \quad (4)$$

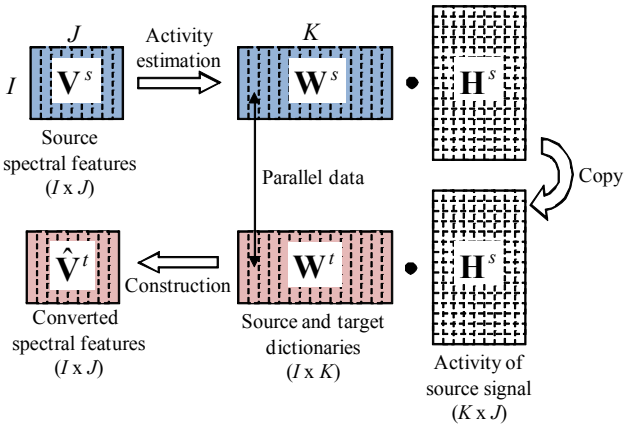

図 1　NMF 声質変換の概要

Fig. 1　Basic approach of NMF-based voice conversion

2.2 問 題 点

Exemplar-based による NMF 声質変換の場合，$\mathbf{V}^s$ と $\mathbf{W}^s\mathbf{H}^s$ の間の誤差値が大きくなる傾向がある．(4.2 節で示されている．) 誤差値を小さくするためには，exemplar からよりコンパクトな辞書を推定する必要がある．また，コンパクトな辞書によって NMF 声質変換の計算コストを削減することが可能になると考えられる．

さらに，NMF 声質変換の場合，入力スペクトルのアクティビティは入力話者辞書から，変換スペクトルは出力話者辞書から再構成される．図 2 に，パラレル発話から推定されたアクティビティの例を示す．簡単な例を示すため，アクティビティは男性 1 名，女性 1 名のパラレルな単語 1 語から推定されたものである．単語は DTW を用いてアライメントが取られている．推定に用いた辞書は 250 基底から構成されるパラレル辞書である．

図に示されているように，発話はパラレルであるにも関わらず，アクティビティは異なる形状を示している．これは，パラレル辞書発話のアライメントのずれによるものであると考えられる．パラレル辞書は DTW でアライメントを取られているが，ミスマッチは残っていると考えられ，このアライメントのずれが声質変換精度を劣化させていると考えられる [20].

さらに，アクティビティには，音韻情報のみならず話者情報も含まれており，話者依存性があるということが考えられる．これらの問題は NMF 声質変換の精度劣化を引き起こすと考えられる．文献 [21] では，この問題を解決するべくアクティビティマッピング手法が提案されているが，学習データに加えて適応データを必要とするため，実用性に乏しいという問題点があった．

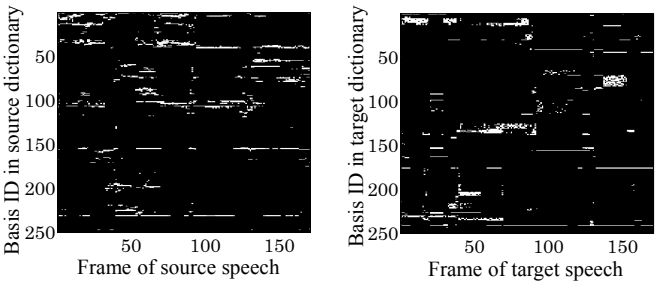

図 2　パラレル発話から推定したアクティビティ

Fig. 2　Activity matrices for parallel utterances

3. DGNMF を用いたパラレル辞書学習

前章で述べた問題点を解決するため，本章では DGNMF を用いた声質変換を提案する．DGNMF は NMF に識別学習を導入したものであり，識別能力を持ったコンパクトな辞書を推定することにより，アクティビティ推定時の推定誤差を減らすことができると考えられる．本章では，まず DGNMF のアルゴリズムを述べた後，パラレル制約を導入した DGNMF による辞書推定法を述べる．

3.1 Discriminative Graph-embedded Non-negative Matrix Factorization

学習データにおいて，クラス内分散グラフとクラス間分散グラフの隣接行列をそれぞれ次のように定義する．

$$\mathbf{A}^w_{ij} = \begin{cases} 1 & \left(\begin{array}{c} \mathbf{v}_i \in N_{k_w}(\mathbf{v}_i) \ or \ \mathbf{v}_j \in N_{k_w}(\mathbf{v}_j) \\ and \\ c_i = c_j \end{array} \right) \\ 0 & (otherwise) \end{cases} \quad (5)$$

$$\mathbf{A}^b_{ij} = \begin{cases} 1 & \left(\begin{array}{c} \mathbf{v}_i \in N_{k_b}(\mathbf{v}_i) \ or \ \mathbf{v}_j \in N_{k_b}(\mathbf{v}_j) \\ and \\ c_i \neq c_j \end{array} \right) \\ 0 & (otherwise) \end{cases} \quad (6)$$

ここで，$N_{k_w}(\mathbf{v}_i)$，$N_{k_b}(\mathbf{v}_i)$ はそれぞれ，クラス内分散グラフにおける学習データ $\mathbf{v}_i$ の k_w 近傍集合，クラス間分散グラフにおける $\mathbf{v}_i$ の k_b 近傍集合を表す．c_i と c_j はそれぞれ $\mathbf{v}_i$，$\mathbf{v}_j$ 音素ラベルである．これらの識別行列を用いて，クラス内分散とクラス間分散のグラフラプラシアン行列を求める．

$$\mathbf{L}^w = \mathbf{D}^w - \mathbf{A}^w \quad (7)$$

$$\mathbf{L}^b = \mathbf{D}^b - \mathbf{A}^b \quad (8)$$

ここで，$\mathbf{D}^w$ と $\mathbf{D}^b$ はそれぞれ，対角成分に $\mathbf{A}^w$ と $\mathbf{A}^b$ の各行（あるいは列）の和をもつ重み行列である．

グラフラプラシアンを用いて，DGNMF のコスト関数を次のように定義する．

$$d_{KL}(\mathbf{V},\mathbf{WH})+\frac{\phi}{2}\mathbf{Tr}(\mathbf{H}^{\mathsf{T}}\mathbf{L}^{w}\mathbf{H})-\frac{\psi}{2}\mathbf{Tr}(\mathbf{H}^{\mathsf{T}}\mathbf{L}^{b}\mathbf{H}) \quad s.t.\ \ \mathbf{W}\geqq 0,\mathbf{H}\geqq 0 \tag{9}$$

ここで，$\mathbf{Tr}$，ϕ，ψ はそれぞれ行列のトレース，クラス内分散重み，クラス間分散重みを表す．式 (9) において，第 1 項は $\mathbf{V}$ と $\mathbf{WH}$ の間の KL ダイバージェンス，第 2 項はクラス内分散制約，第 3 項はクラス間分散制約を表す．

このコスト関数は次の更新式を用いて最小化できる．

$$\mathbf{W}\leftarrow\mathbf{W}.*((\mathbf{V}./(\mathbf{WH}))\mathbf{H}^{\mathsf{T}})./(\mathbf{1}^{(J\times I)}\mathbf{H}^{\mathsf{T}}) \tag{10}$$

$$\mathbf{H}\leftarrow\frac{-\beta+\sqrt{\beta^2+4\alpha\gamma}}{2\alpha} \tag{11}$$

$$\alpha=((\phi\mathbf{D}^{w}+\psi\mathbf{A}^{b})\mathbf{H})./\mathbf{H} \tag{12}$$

$$\beta=\mathbf{W}^{\mathsf{T}}\mathbf{1}^{(I\times J)}-\mathbf{H}(\phi\mathbf{A}^{w}+\psi\mathbf{D}^{b}) \tag{13}$$

$$\gamma=\mathbf{H}.*(\mathbf{W}^{\mathsf{T}}(\mathbf{V}./(\mathbf{WH}))) \tag{14}$$

これらの更新式は GNMF と同様にして導出できる [19]．しかしながら，GNMF には音素ラベルに基づく識別ならびにクラス間分散は考慮しておらず，DGNMF と GNMF は異なるアプローチである．

3.2 パラレル辞書学習

パラレル制約付き DGNMF を用いて，コンパクトで識別的な辞書を推定する．図 3 に，パラレル辞書学習の概要を示す．パラレル制約付き DGNMF の目的関数を次のように定義する．

$$\begin{aligned}&d_{KL}(\mathbf{V}^{s},\mathbf{W}^{s}\mathbf{H}^{s})+d_{KL}(\mathbf{V}^{t},\mathbf{W}^{t}\mathbf{H}^{t})\\&+\frac{\phi}{2}\mathbf{Tr}(\mathbf{H}^{s\mathsf{T}}\mathbf{L}^{sw}\mathbf{H}^{s})+\frac{\phi}{2}\mathbf{Tr}(\mathbf{H}^{t\mathsf{T}}\mathbf{L}^{tw}\mathbf{H}^{t})\\&-\frac{\psi}{2}\mathbf{Tr}(\mathbf{H}^{s\mathsf{T}}\mathbf{L}^{sb}\mathbf{H}^{s})-\frac{\psi}{2}\mathbf{Tr}(\mathbf{H}^{t\mathsf{T}}\mathbf{L}^{tb}\mathbf{H}^{t})\\&+\lambda||\mathbf{H}^{s}||_1+\lambda||\mathbf{H}^{t}||_1+\frac{\epsilon}{2}||\mathbf{H}^{s}-\mathbf{H}^{t}||_F^2\\&s.t.\ \ \mathbf{W}^{s}\geqq 0,\mathbf{H}^{s}\geqq 0,\mathbf{W}^{t}\geqq 0,\mathbf{H}^{t}\geqq 0\end{aligned} \tag{15}$$

ここで，$\mathbf{V}^{s}$，$\mathbf{V}^{t}$，$\mathbf{W}^{s}$，$\mathbf{W}^{t}$，$\mathbf{H}^{s}$，$\mathbf{H}^{t}$ は入力話者学習データ，出力話者学習データ，入力話者辞書行列，出力話者辞書行列，入力アクティビティ，出力アクティビティを表す．さらに $\mathbf{L}^{sw}$，$\mathbf{L}^{sb}$ はそれぞれ，入力話者学習データのクラス内分散，クラス間分散のグラフラプラシアンを表し，$\mathbf{L}^{tw}$，$\mathbf{L}^{tb}$ は出力話者学習データのグラフラプラシアンを表す．入力話者学習データ，出力話者学習データはそれぞれ DTW で対応を取り，同一フレーム数となったものを用いる．ϵ，λ はそれぞれパラレル制約，スパース制約の重みである．式 (15) において，第 1 項から第 6 項は式 (9) を拡張したものである．第 7 項と第 8 項は $\mathbf{H}^{s}$ と $\mathbf{H}^{t}$ におけるスパース制約，最終項は $\mathbf{H}^{s}$ と $\mathbf{H}^{t}$ の間のパラレル制約を表す．

このコスト関数は次の更新式を用いて最小化できる．

$$\mathbf{W}^{s}\leftarrow\mathbf{W}^{s}.*((\mathbf{V}^{s}./(\mathbf{W}^{s}\mathbf{H}^{s}))\mathbf{H}^{s\mathsf{T}})./(\mathbf{1}^{(J\times I)}\mathbf{H}^{s\mathsf{T}}) \tag{16}$$

$$\mathbf{W}^{t}\leftarrow\mathbf{W}^{t}.*((\mathbf{V}^{t}./(\mathbf{W}^{t}\mathbf{H}^{t}))\mathbf{H}^{t\mathsf{T}})./(\mathbf{1}^{(J\times I)}\mathbf{H}^{t\mathsf{T}}) \tag{17}$$

$$\mathbf{H}^{s}\leftarrow\frac{-\beta^{s}+\sqrt{\beta^{s2}+4\alpha^{s}\gamma^{s}}}{2\alpha^{s}} \tag{18}$$

$$\alpha^{s}=((\phi\mathbf{D}^{sw}+\psi\mathbf{A}^{sb})\mathbf{H}^{s})./\mathbf{H}^{s}+\epsilon \tag{19}$$

$$\beta^{s}=\mathbf{W}^{s\mathsf{T}}\mathbf{1}^{(I\times J)}-\mathbf{H}^{s}(\phi\mathbf{A}^{sw}+\psi\mathbf{D}^{sb})-\epsilon\mathbf{H}^{t}+\lambda \tag{20}$$

$$\gamma^{s}=\mathbf{H}^{s}.*(\mathbf{W}^{s\mathsf{T}}(\mathbf{V}^{s}./(\mathbf{W}^{s}\mathbf{H}^{s}))) \tag{21}$$

$$\mathbf{H}^{t}\leftarrow\frac{-\beta^{t}+\sqrt{\beta^{t2}+4\alpha^{t}\gamma^{t}}}{2\alpha^{t}} \tag{22}$$

$$\alpha^{t}=((\phi\mathbf{D}^{tw}+\psi\mathbf{A}^{tb})\mathbf{H}^{t})./\mathbf{H}^{t}+\epsilon \tag{23}$$

$$\beta^{t}=\mathbf{W}^{t\mathsf{T}}\mathbf{1}^{(I\times J)}-\mathbf{H}^{t}(\phi\mathbf{A}^{tw}+\psi\mathbf{D}^{tb})-\epsilon\mathbf{H}^{s}+\lambda \tag{24}$$

$$\gamma^{t}=\mathbf{H}^{t}.*(\mathbf{W}^{t\mathsf{T}}(\mathbf{V}^{t}./(\mathbf{W}^{t}\mathbf{H}^{t}))) \tag{25}$$

$\mathbf{A}^{sw}$ と $\mathbf{A}^{sb}$ は入力話者学習データにおけるクラス内分散グラフとクラス間分散グラフの隣接行列，$\mathbf{D}^{sw}$ と $\mathbf{D}^{sb}$ はそれらの重み行列である．$\mathbf{A}^{tw}$，$\mathbf{A}^{tb}$，$\mathbf{D}^{tw}$，$\mathbf{D}^{tb}$ はそれらのそれぞれ出力話者学習データである．

識別的なパラレル辞書が推定された後，入力スペクトルは第 2 章で示した手法と同様にして変換される．

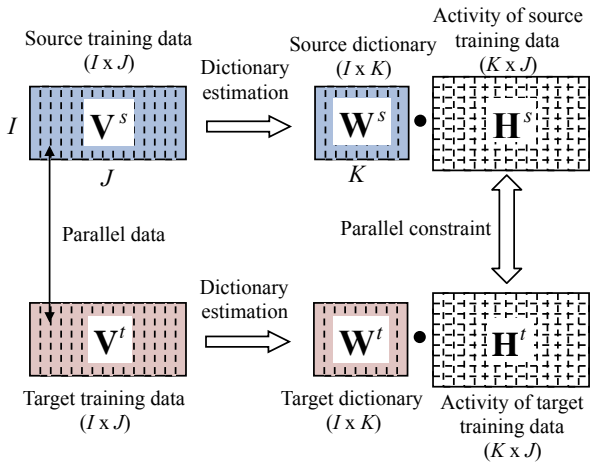

図 3　パラレル辞書学習の概要
Fig. 3　Parallel dictionary learning

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 NMF 声質変換 [8]，GMM 学習声質変換と比較した．

ATR 研究用日本語音声データベース [22] 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，学習データに含まれない音素バランス文 50 文をテストデータとして用いた．$\epsilon,\phi,\psi,k_w,k_b,\lambda$ はそれぞれ 40, 1, 10, 1024, 8192, 0.1 とした．NMF の更新回数は辞書学習時には 10，変換時には 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．

提案手法と NMF 声質変換では，音声分析合成手法 STRAIGHT [23] によって推定されたスペクトルと前後 1 フ

レームを含むセグメント特徴量を用いた．この次元数は 1,539 である．GMM 声質変換において，STRAIGHT を用いて推定されたスペクトルから計算した Mel-cepstrum と前後 1 フレームを考慮した Δ パラメータを特徴量として用いた．特徴量の次元数は 48 である．GMM の混合数は 64 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 [16]，非周期成分は入力発話のものを用いた．

4.2 実験結果

客観評価指標として，メルケプストラム歪（Mel-cepstral distortion: MelCD）[dB] を用いた．

$$MelCD = (10/\log 10)\sqrt{2\sum_{d}^{24}(mc_d^{conv} - mc_d^{tar})^2} \qquad (26)$$

ここで，mc_d^{conv}，mc_d^{tar} は変換メルケプストラム，目標メルケプストラムにおける d 次元目の特徴量を表す．

Table 1 にそれぞれの手法による MelCD と計算コストを表す．Conv. は目標メルケプストラムと変換メルケプストラムの間の歪，Recon. は入力メルケプストラムと分解再合成時のメルケプストラム（式 (2) における $\mathbf{W}^s\mathbf{H}^s$）を表し，これは行列分解における分解誤差にあたる．括弧内の数字はパラレル辞書の基底数である．PDL は，Graph Embedding を考慮していないパラレル辞書学習（DGNMF における $\phi = 0, \psi = 0$）を表す．

まず，従来の NMF 声質変換 (Exemplar-based) において，基底数を削減していった場合を見ると，5,000 基底の場合と全基底を用いた場合の歪がほぼ一致するという結果になった．分解誤差は全基底を用いたときが最も小さく，次いで 5,000 基底を用いたときが小さい．

PDL でパラレル辞書を学習した場合は，NMF でランダムに基底を削減した場合と比較して歪が若干大きくなった．分解誤差も大きくなっていることがわかる．DGNMF でパラレル辞書を学習した場合，変換歪は従来の NMF とほぼ同等であり，分解誤差はこれらの手法のなかで最も小さくなった．この結果は，Graph Embedding を考慮することの効果を示している．

また，計算コストは辞書の基底数を削減するに従って小さくなっている．

主観評価実験は，10 人の日本語話者が 25 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，音質と話者性の 2 つの観点において評価実験を行った．音質に関しては MOS 評価基準による 5 段階評価 (5: excellent, 4: good, 3: fair, 2: poor, 1: bad) で，話者性については，X を目標話者とし 2 つの変換手法のうちよい方を選ぶ XAB 法で評価した．従来の NMF 声質変換について基底数は 5,000 とした．

図 4 に音質の評価結果を示す．エラーバーは，95% 信頼区間を示す．提案手法は，従来の 2 つの手法と比較して音質が向上していることがわかる．この結果は t 検定により有意であることが示されている．これは，提案手法が NMF の声質変換の分解誤差を削減しつつパラレル辞書のマッピング精度を向上させることができたことによるものだと考えられる．

図 5 に話者性による主観評価実験の結果を示す．提案手法と GMM 声質変換の間には有意差が見られないが，提案手法は NMF 声質変換と比較して有意に話者性を向上させていることがわかる．この結果は t 検定により有意であることが示されている．

表 1　各手法による MelCD [dB] と計算時間 [s]
Table 1　MelCD [dB] and computational time [s] of each method

	Conv.	Recon.	times
GMM	2.96	-	2
NMF (all)	3.05	1.37	890
NMF (10,000)	3.11	1.59	680
NMF (5,000)	3.05	1.56	310
NMF (1,000)	3.11	1.59	71
PDL (1,000)	3.11	1.85	71
DGNMF (1,000)	3.08	1.09	71

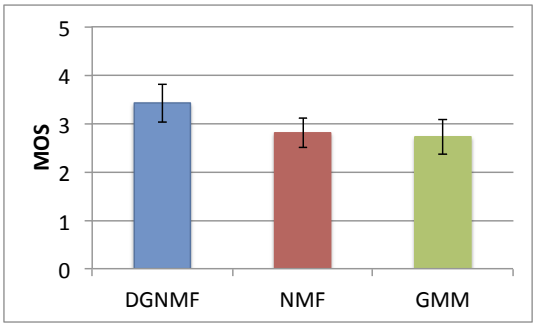

図 4　音質における MOS 値
Fig. 4　MOS test on speech quality

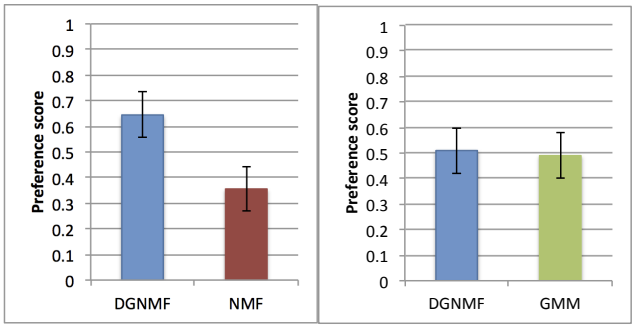

図 5　話者性における主観評価値
Fig. 5　Preference score on individuality

5. Conclusions

NMF 声質変換の変換精度を向上させるため，DGNMF を用いたパラレル辞書学習を提案した．提案手法である DGNMF はクラス間識別を導入した GNMF [19] に対して，音素ラベルによる識別とクラス内分散を導入したもので，NMF ベースの辞書学習における，辞書の過学習を防ぐことができる．実験結果により，DGNMF によるパラレル辞書学習は NMF 声質変換の精度を向上させたのみならず，基底数削減により計算コストの削減も可能にした．今後は，この手法を多対多声質変換 [24]

や構音障害者のための声質変換 [25] に応用させていく予定である．さらに，この手法はトピックモデル [2] や超解像 [1] など，他のタスクにも応用可能であると考えられる．

今後の課題として，依然として提案手法の計算コストが GMM 声質変換と比較して高いことがあげられる．Wu ら [9] は NMF 声質変換の計算コストを線形スペクトルとメルスペクトルのパラレルな特徴量を用いることで削減しており，この手法を導入することで，計算コストのさらなる削減を行うことができると考えられる．

文　献

[1] M.W. Berry, M. Browne, A.N. Langville, V.P. Pauca, and R.J. Plemmons, “Algorithms and applications for approximate nonnegative matrix factorization,” Computational Statistics and Data Analysis, vol.52, no.1, pp.155–173, 2007.

[2] T. Hofmann, “Probabilistic latent semantic indexing,” in Proc. SIGIR, pp.50–57, 1999.

[3] A. Cichocki, R. Zdnek, A.H. Phan, and S. Amari, Nonnegative Matrix and Tensor Factorization, WILKEY, 2009.

[4] M.N. Schmidt and R.K. Olsson, “Single-channel speech separation using sparse non-negative matrix factorization,” in Proc. Interspeech, pp.1652–1655, 2006.

[5] T. Virtanen, “Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria,” IEEE Trans. Audio, Speech, Lang. Process., vol.15, no.3, pp.1066–1074, 2007.

[6] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, “Efficient algorithms for multichannel extensions of Itakura-Saito nonnegative matrix factorization,” in Proc. ICASSP, pp.261–264, 2012.

[7] J.F. Gemmeke, T. Virtanen, and A. Hurmalainen, “Exemplar-based sparse representations for noise robust automatic speech recognition,” IEEE Trans. Audio, Speech, Lang. Process., vol.19, no.7, pp.2067–2080, 2011.

[8] R. Takashima, T. Takiguchi, and Y. Ariki, “Exemplar-based voice conversion in noisy environment,” in Proc. SLT, pp.313–317, 2012.

[9] Z. Wu, T. Virtanen, E.S. Chng, and H. Li, “Exemplar-based sparse representation with residual compensation for voice conversion,” IEEE Trans. Audio, Speech, Lang. Process., vol.22, no.10, pp.1506–1521, 2014.

[10] 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.

[11] C. Veaux and X. Robet, “Intonation conversion from neutral to expressive speech,” in Proc. Interspeech, pp.2765–2768, 2011.

[12] 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.

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

[14] K. Nakamura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, “A mel-cepstral analysis technique restoring high frequency components from low-sampling-rate speech,” in Proc. Interspeech, pp.2494–2498, 2014.

[15] Y. Ohtani, M. Tamura, M. Morita, and M. Akamine, “GMM-based bandwidth extension using sub-band basis spectrum model,” in Proc. Interspeech, pp.2489–2493, 2014.

[16] 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.

[17] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, “Voice conversion using partial least squares regression,” IEEE Trans. Audio, Speech, Lang. Process., vol.18, no.5, pp.912–921, 2010.

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

[19] D. Cai, X. He, and T.S.H. J. Han, “Graph regularized nonnegative matrix factorization for data representation,” IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.33, no.8, pp.1548–1560, 2010.

[20] R. Aihara, T. Nakashika, T. Takiguchi, and Y. Ariki, “Voice conversion based on non-negative matrix factorization using phoneme-categorized dictionary,” in Proc. ICASSP, pp.7944–7948, 2014.

[21] R. Aihara, T. Takiguchi, and Y. Ariki, “Activity-mapping non-negative matrix factorization for exemplar-based voice conversion,” in Proc. ICASSP, pp.4899–4903, 2015.

[22] 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.

[23] H. Kawahara, “STRAIGHT, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds,” Acoustical Science and Technology, pp.349–353, 2006.

[24] R. Aihara, T. Takiguchi, and Y. Ariki, “Multiple non-negative matrix factorization for many-to-many voice conversion,” IEEE/ACM Trans. on Audio, Speech, and Language Processing, vol.24, no.7, pp.1175–1184, 2016.

[25] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, “A preliminary demonstration of exemplar-based voice conversion for articulation disorders using an individuality-preserving dictionary,” EURASIP Journal on Audio, Speech, and Music Processing, vol.2014:5, doi:10.1186/1687-4722-2014-5, pp.1–10, 2014.