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
SP2015-72(2015-12)

Alternating Direction Method of Multipliers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．本稿では，Alternating Direction Method of Multipliers (ADMM) に基づいて最適化した Semi-NMF を用いてパラレル辞書の学習を行い，計算コストが少なく自然性の高い声質変換を提案する．Semi-NMF は，NMF における特徴量の非負制約をはずしたものであり，Mel-cepstrum などコンパクトな特徴量の利用により，使用メモリの削減が期待できる．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．従来提案されていた Semi-NMF は補助関数法を用いており収束に時間がかかるため，ADMM に基づく最適化を提案し，収束速度を向上させる．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 声質変換とほぼ同等の精度が，1/76 の計算時間で得られることがわかった．

キーワード　声質変換，音声合成，非負値行列因子分解，Exemplar-based

Parallel Dictionary Learning for Voice Conversion Using Alternating Direction Method of Multipliers

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　Voice conversion (VC) is being widely researched in the field of speech processing because of increased interest in using such processing in applications such as personalized Text-To-Speech systems. A VC method using Non-negative Matrix Factorization (NMF) has been researched because of its natural sounding voice, however, huge memory usage and high computational times have been reported as problems. We present in this paper a new VC method using Semi-Non-negative Matrix Factorization (Semi-NMF) using the Alternating Direction Method of Multipliers (ADMM) in order to tackle the problems associated with NMF-based VC. Dictionary learning using Semi-NMF can create a compact dictionary, and ADMM enables faster convergence than conventional Semi-NMF. Experimental results show that our proposed method is 76 times faster than conventional NMF, and its conversion quality is almost the same as that of the conventional method.

Key words　Voice Conversion, Speech synthesis, Non-negative Matrix Factorization, Exemplar-based

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，従来の統計的手法とは異なる，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]．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．

しかしながら，NMF 声質変換には，計算コストが高いという問題があった．その理由には，大きく分けて以下の 3 つが存在する．

（1） **高次元スペクトルの利用**：非負制約のため，NMF 声質変換では使用できる特徴量が限定され，Mel-cepstrum といった負値を含む低次元特徴量を使用できない．

（2） **膨大な基底数**：基本的に，NMF 声質変換では学習データの全てのスペクトルを基底として用いるため辞書の基底数が大きくなる．

（3） **最適化手法:** 従来の NMF 声質変換では，補助関数法を用いた最適化を行うため，コスト関数の収束に時間が掛かる．

本論文では，上記の問題点を解決するため，以下の 3 つの手法を採用する．

（1） **Semi-Non-negative Matrix Factorization (Semi-NMF)**：入力特徴量，辞書行列の非負制約を取り除き，負値を含むコンパクトな特徴量を利用する．

（2） **辞書学習**：Semi-NMF を用いた辞書学習を提案し，基底数の削減を行う．

（3） **Alternating Direction Method of Multipliers (ADMM)**：補助関数法ではなく，制約付き最適化手法のひとつである ADMM を用いることで，収束速度を向上させる．

補助関数法を用いた Semi-NMF は文献 [19] で提案されているが，これまで声質変換に Semi-NMF が用いられた例はない．ADMM を用いた NMF は文献 [20] において提案されており，本研究ではこの手法を Semi-NMF へ拡張する．

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．

2. 先行研究

2.1 NMF 声質変換

2.1.1 概　　要

スパース表現の考え方において，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．

$$\mathbf{v}_l \approx \sum_{j=1}^{J} \mathbf{w}_j h_{j,l} = \mathbf{W}\mathbf{h}_l \tag{2}$$

$\mathbf{v}_l$ 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．$\mathbf{w}_j$ 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 $\mathbf{w}_j$ とする．基底を並べた行列 $\mathbf{W} = [\mathbf{w}_1 \dots \mathbf{w}_J]$ 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 (2) は二つの行列の内積で表される．

$$\mathbf{V} \approx \mathbf{WH} \tag{3}$$

$$\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_L], \quad \mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_L]. \tag{4}$$

ここで L はフレーム数を表す．

本手法の概要を図 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，$\hat{\mathbf{V}}^t$ は変換音声，$\mathbf{H}^s$ 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．こ

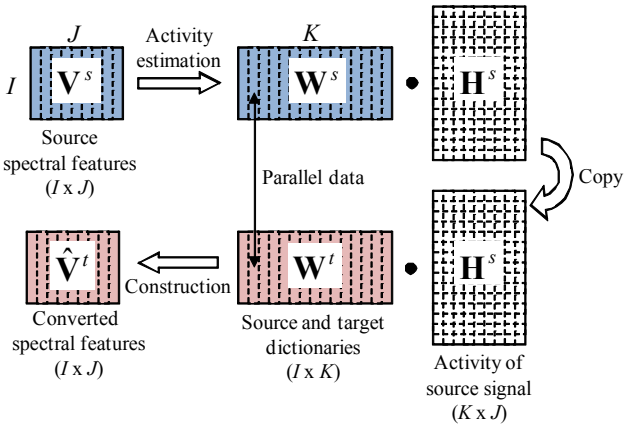

図 1　NMF 声質変換の概要

Fig. 1　Basic approach of NMF-based voice conversion

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て辞書化したものである．入力音声を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 [7]．NMF のコスト関数は，$\mathbf{V}^s$，$\mathbf{W}^s$，$\mathbf{H}^s$ を用いて以下のような式で表せる．

$$d(\mathbf{V}^s, \mathbf{W}^s\mathbf{H}^s) + \lambda||\mathbf{H}^s||_1 \tag{5}$$

ここで，第 1 項は $\mathbf{V}^s$ と $\mathbf{W}^s\mathbf{H}^s$ の間の Kullback-Leibler(KL) 距離であり，第 2 項はアクティビティ行列をスパースにするための L1 ノルム制約項である．λ はスパース重みを表す．このコスト関数は Jensen の不等式を用いることで，繰り返し適用を用いて最小化できる．

$$\begin{aligned}\mathbf{H}^s_{n+1} \;=\; & \mathbf{H}^s_n .* (\mathbf{W}^{s\mathsf{T}}(\mathbf{V}^s ./ (\mathbf{W}^s\mathbf{H}^s))) \\ & ./ (\mathbf{H}^{s\mathsf{T}}\mathbf{1}^{D\times L} + \lambda \mathbf{1}^{1\times L})\end{aligned} \tag{6}$$

変換音声 $\hat{\mathbf{V}}^t$ は出力話者辞書行列と推定されたアクティビティの内積をとることで得られる．

$$\hat{\mathbf{V}}^t = \mathbf{W}^t\mathbf{H}^s \tag{7}$$

2.1.2　問　題　点

NMF における非負制約のため，NMF 声質変換において用いられる特徴量は負値を含まない線形スペクトルなどに限定される．したがって，Δ や $\Delta\Delta$ 特徴量といった動的特徴量も使用できない．文献 [21] においては，513 次元の線形スペクトルとそのセグメント特徴量を使用しており，結果としてメモリ使用量が多くなるという問題が指摘されている．

さらに，NMF 声質変換においてはパラレル学習データ全てがパラレル辞書として用いられる．アクティビティは多くの基底をもつ辞書から推定されるため計算コストが高くなる．これらの点から，NMF 声質変換はリアルタイムでの実現が困難であるという問題がある．

また，NMF 声質変換においては，入力発話のアクティビティは入力辞書行列から推定され，パラレルな出力辞書行列と掛け合わせることで変換音声が得られる．図 2 は，性別の異なる話者から発話されたパラレルな発話のアクティビティを，それぞれの話者のパラレル辞書行列で推定したものである．入力発話は DTW によってアライメントが取られており，パラレル辞書行列の基底数は 250 である．図に示されたように，アクティビティ行列はそれぞれの入力発話がパラレルであるのにもかかわらず，異なった形状を示している．

この理由として，以下の 2 つが考えられる．一つ目はアライメントのミスマッチである．パラレル辞書は DTW によってアライメントが取られているが，アライメントのミスマッチは残っているの考えられる．2 つ目は，アクティビティ行列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である．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 [21]．文献 [22] 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．

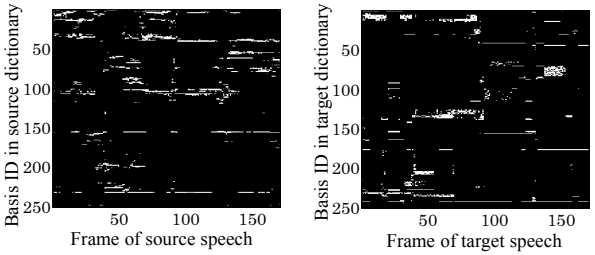

図 2　パラレル発話から推定したアクティビティの例

Fig. 2　Activity matrices for parallel utterances

2.2　補助関数法を用いた Semi-NMF

2.2.1　定　式　化

Semi-NMF のコスト関数は以下のように定義できる．

$$d_F(\mathbf{V}, \mathbf{WH}) + \lambda||\mathbf{H}||_1 \quad s.t. \;\; \mathbf{H} \geqq 0 \tag{8}$$

ここで，第 1 項は $\mathbf{V}$ と $\mathbf{WH}$ の間のフロベニウスノルムであり，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．NMF と比較すると，Semi-NMF においては $\mathbf{W}$ の非負制約をはずしたため，コスト関数がフロベニウスノルムに限定される．

コスト関数は以下の更新式を繰り返し適用することで最小化される．

$$\begin{aligned}\mathbf{H} \leftarrow\; & (-\lambda\mathbf{H}^{\mathsf{T}} + (\mathbf{H}^{\mathsf{T}} .* \sqrt{\lambda^2 + 16(\mathbf{A} .* \mathbf{B})})) \\ & ./(4\mathbf{A})\end{aligned} \tag{9}$$

$$\mathbf{A} \;=\; (\mathbf{V}^{\mathsf{T}}\mathbf{W})^- + (\mathbf{H}^{\mathsf{T}}(\mathbf{W}^{\mathsf{T}}\mathbf{W})^+) \tag{10}$$

$$\mathbf{B} \;=\; (\mathbf{V}^{\mathsf{T}}\mathbf{W})^+ + (\mathbf{H}^{\mathsf{T}}(\mathbf{W}^{\mathsf{T}}\mathbf{W})^-) \tag{11}$$

ここで，正部と負部を分けるため，$\mathbf{X}^+ = (|\mathbf{X}| + \mathbf{X})/2, \mathbf{X}^- = (|\mathbf{X}| - \mathbf{X})/2$ のように定める．

2.2.2 問 題 点

Semi-NMF は負値を含む特徴量を分解することができるため，Mel-cepstrum や動的特徴量を用いることができ，NMF 比較して，メモリ使用量を削減することができる．しかしながら，式 (11) には $\mathbf{W}^\mathsf{T}\mathbf{W}$ の項がある．従来のパラレル辞書のような基底数の多い辞書行列を用いた場合，この項の計算に膨大なメモリを使用し，計算コストが高くなる問題がある．従って，Semi-NMF を用いた声質変換を考える場合，コンパクトなパラレル辞書の学習が必要となる．さらに，補助関数法による最適化は収束速度が遅いという問題もある．

3. Semi-Non-negative Matrix Factorization を用いた声質変換

3.1 Basic Idea

前章の問題点を解決するため，本研究では ADMM に基づく Semi-NMF を用いた声質変換法を提案する．まず，Semi-NMF を用いたパラレル制約付き辞書学習で入力と出力のパラレル辞書を学習する．パラレル制約は，従来の NMF 声質変換で問題となっていたアクティビティのずれを解消する手法である．さらに，コンパクトな辞書行列によって計算コストの削減が期待できる．

変換時には，ADMM ベースの Semi-NMF を用いて，入力スペクトルは学習された入力辞書基底の線形結合で表現される．ADMM を用いた Semi-NMF を用いることで，補助関数法を用いた Semi-NMF と比較して，よりスパースで精度の高いアクティビティが得られる．

3.2 辞 書 学 習

基底数の少ない，コンパクトな辞書を学習するため，入力話者と出力話者のパラレル辞書は，パラレル制約付き Semi-NMF によって推定される．

目的関数を下記のように定義する．

$$\begin{aligned} min \quad & d_F(\mathbf{V}^s, \mathbf{W}^s\mathbf{H}^s) + d_F(\mathbf{V}^t, \mathbf{W}^t\mathbf{H}^t) \\ & + \frac{\epsilon}{2}||\mathbf{H}^s - \mathbf{H}^t||_F^2 + \lambda||\mathbf{H}^s||_1 + \lambda||\mathbf{H}^t||_1 \\ sub\ to \quad & \mathbf{H}^s = \mathbf{H}^s_+, \mathbf{H}^s_+ \geqq 0, \mathbf{H}^t = \mathbf{H}^t_+, \mathbf{H}^t_+ \geq 0 \end{aligned} \tag{12}$$

ここで，$\mathbf{V}^s$，$\mathbf{V}^t$，$\mathbf{W}^s$，$\mathbf{W}^t$，$\mathbf{H}^s$，$\mathbf{H}^t$ はそれぞれ，入力話者と出力話者のパラレル学習データ，推定する入力話者と出力話者のパラレル辞書行列，入力と出力のアクティビティを表す．パラレル学習データは DTW でアライメントを取られたものを用いる．ϵ と λ はそれぞれ，パラレル制約重みとスパース制約重みを表す．式 (12) のラグランジアンは下記のようになる．

$$\begin{aligned} & L_\rho(\mathbf{W}^s, \mathbf{H}^s, \mathbf{W}^t, \mathbf{H}^t, \mathbf{H}^s_+, \mathbf{H}^t_+) = \\ & d_F(\mathbf{V}^s, \mathbf{W}^s\mathbf{H}^s) + d_F(\mathbf{V}^t, \mathbf{W}^t\mathbf{H}^t) + \frac{\epsilon}{2}||\mathbf{H}^s - \mathbf{H}^t||_F^2 \\ & + \lambda||\mathbf{H}^s||_1 + \langle \alpha_{\mathbf{H}^s}, \mathbf{H}^s - \mathbf{H}^s_+ \rangle + \frac{\rho}{2}||\mathbf{H}^s - \mathbf{H}^s_+||_F^2 \\ & + \lambda||\mathbf{H}^t||_1 + \langle \alpha_{\mathbf{H}^t}, \mathbf{H}^t - \mathbf{H}^t_+ \rangle + \frac{\rho}{2}||\mathbf{H}^t - \mathbf{H}^t_+||_F^2 \end{aligned} \tag{13}$$

ここで，ρ は収束率を調性するチューニングパラメータである．式 (13) は，表 1 に示されたアルゴリズムを適用することで最適化される．

表 1　辞書学習アルゴリズム
Table 1　Algorithm of Dictionary Learning

Input $\mathbf{V}^s, \mathbf{V}^t$
Initialize $\mathbf{W}^s, \mathbf{H}^s, \mathbf{W}^t, \mathbf{H}^t, \mathbf{H}^s_+, \mathbf{H}^t_+, \alpha_{\mathbf{H}^s}, \alpha_{\mathbf{H}^t}$
Repeat
$\mathbf{W}^s \leftarrow (\mathbf{V}^s(\mathbf{H}^{s\mathsf{T}}))/(\mathbf{H}^s(\mathbf{H}^{s\mathsf{T}}))$
$\mathbf{W}^t \leftarrow (\mathbf{V}^t(\mathbf{H}^{t\mathsf{T}}))/(\mathbf{H}^t(\mathbf{H}^{t\mathsf{T}}))$
$\mathbf{H}^s \leftarrow (2\mathbf{W}^{s\mathsf{T}}\mathbf{W}^s + (\rho+\epsilon)\mathbf{I})$
$\backslash(2\mathbf{W}^{s\mathsf{T}}\mathbf{W}^s - \alpha_{\mathbf{H}_s} + \rho\mathbf{H}^s_+ + \epsilon\mathbf{H}^t - \lambda)$
$\mathbf{H}^t \leftarrow (2\mathbf{W}^{t\mathsf{T}}\mathbf{W}^t + (\rho+\epsilon)\mathbf{I})$
$\backslash(2\mathbf{W}^{t\mathsf{T}}\mathbf{W}^t - \alpha_{\mathbf{H}_t} + \rho\mathbf{H}^t_+ + \epsilon\mathbf{H}^s - \lambda)$
$\mathbf{H}^s_+ \leftarrow max(\mathbf{H}^s + \frac{1}{\rho}\alpha_{\mathbf{H}^s}, 0)$
$\mathbf{H}^t_+ \leftarrow max(\mathbf{H}^t + \frac{1}{\rho}\alpha_{\mathbf{H}^t}, 0)$
$\alpha_{\mathbf{H}^s} \leftarrow \alpha_{\mathbf{H}_s} + \rho(\mathbf{H}^s - \mathbf{H}^s_+)$
$\alpha_{\mathbf{H}^t} \leftarrow \alpha_{\mathbf{H}_t} + \rho(\mathbf{H}^t - \mathbf{H}^t_+)$
Until convergence **return** $\mathbf{W}^s, \mathbf{H}^s_+, \mathbf{W}^t, \mathbf{H}^t_+$

3.3 変　　換

コンパクトなパラレル辞書行列 $\mathbf{W}^s$，$\mathbf{W}^t$ が推定された後，入力スペクトル $\mathbf{V}^s$ は，ADMM に基づく Semi-NMF を用いて変換スペクトル $\hat{\mathbf{V}}^t$ へ変換される．目的関数は下記のように定める．

$$\begin{aligned} min \quad & d_F(\mathbf{V}^s, \mathbf{W}^s\mathbf{H}^s) + \lambda||\mathbf{H}^s||_1 \\ sub\ to \quad & \mathbf{H}^s = \mathbf{H}^s_+, \mathbf{H}^s_+ \geqq 0. \end{aligned} \tag{14}$$

式 (14) のラグランジアンは次のようになる．

$$\begin{aligned} & L_\rho(\mathbf{W}^s, \mathbf{H}^s, \mathbf{H}^s_+) = \\ & d_F(\mathbf{V}^s, \mathbf{W}^s\mathbf{H}^s) + \lambda||\mathbf{H}^s||_1 \\ & + \langle \alpha_{\mathbf{H}^s}, \mathbf{H}^s - \mathbf{H}^s_+ \rangle + \frac{\rho}{2}||\mathbf{H}^s - \mathbf{H}^s_+||_F^2 \end{aligned} \tag{15}$$

ここで，$\mathbf{W}^s$ は固定され，$\mathbf{H}^s$ は表 2 に示されたアルゴリズムによって推定される．

表 2　変換アルゴリズム
Table 2　Algorithm of Conversion

Input $\mathbf{V}^s, \mathbf{W}^s$
Initialize $\mathbf{H}^s, \mathbf{H}^s_+, \alpha_{\mathbf{H}^s}$
Repeat
$\mathbf{H}^s \leftarrow (2\mathbf{W}^{s\mathsf{T}}\mathbf{W}^s + \rho\mathbf{I})$
$\backslash(2\mathbf{W}^{s\mathsf{T}}\mathbf{W}^s - \alpha_{\mathbf{H}_s} + \rho\mathbf{H}^s_+ - \lambda)$
$\mathbf{H}^s_+ \leftarrow max(\mathbf{H}^s + \frac{1}{\rho}\alpha_{\mathbf{H}^s}, 0)$
$\alpha_{\mathbf{H}^s} \leftarrow \alpha_{\mathbf{H}_s} + \rho(\mathbf{H}^s - \mathbf{H}^s_+)$
Until convergence **return** $\mathbf{H}^s_+$

推定されたアクティビティ $\mathbf{H}^s$ と辞書学習によって求められた出力辞書行列 $\mathbf{W}^t$ によって変換スペクトル $\hat{\mathbf{V}}^t$ は以下のように求められる．

$$\hat{\mathbf{V}}^t = \mathbf{W}^t\mathbf{H}^s \tag{16}$$

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 [23]，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 [24] 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 216 単語を学習データとし，音素バランス文 50 文をテストデータとして用いた．提案手法において，ρ, ϵ, λ はそれぞれ，1, 1, 0.1 とした．Semi-NMF の更新回数は辞書学習時には 50，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 GMM 声質変換において，音声分析合成手法 [25] を用いて推定されたスペクトルから計算した Mel-cepstrum と 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．一方，NMF 声質変換では，STRAIGHT 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 の混合数は 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 [16]，非周期成分は入力発話のものを用いた．

4.2 結果と考察

まず，ADMM ベースの Semi-NMF と，補助関数法を用いた Semi-NMF の間の収束速度を比較した．辞書の基底数は 1,000 とした．結果を図 3 に示す．横軸は更新回数，縦軸は log スケールにおける目的関数の値である．図より，ADMM を用いた場合，ρ が小さくなるにしたがって収束率が向上し，従来の補助関数法に基づく Semi-NMF を上回っていることがわかる．

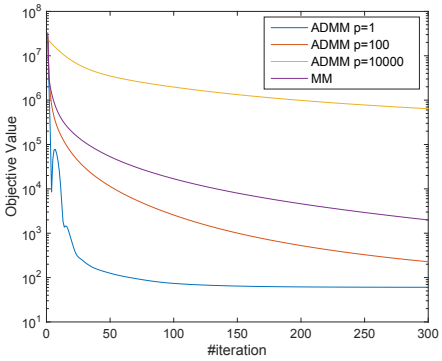

図 3　更新による目的関数値の変化
Fig. 3　Objective value as a function of iteration

客観評価指標として，Normalized Spectrum Distortion (NSD) [26] を用いた．

$$NSD = \sqrt{||\mathbf{X}^t - \hat{\mathbf{X}}^t||^2 / ||\mathbf{X}^t - \mathbf{X}^s||^2} \qquad (17)$$

ここで，$\mathbf{X}^s$, $\mathbf{X}^t$, $\hat{\mathbf{X}}^t$ はそれぞれ入力スペクトル，出力スペクトル，変換スペクトルを表す．図 3 にそれぞれの手法における NSD と計算時間を示す．提案手法において，辞書の基底数を 1,000 とした場合と 5,000 とした場合では，NSD にほとんど差がないことがわかる．提案手法は NMF 声質変換と比較して，わずかに NSD が大きくなっているが計算時間が大幅に削減できている．提案手法と GMM 声質変換の間の NSD はほとんど差がない．

表 3　各手法における NSD と計算時間
Table 3　NSD and computational times of each method

	NSD	times [s]
GMM	1.66	2
NMF	1.54	916
Proposed(1,000)	1.69	12
Proposed(5,000)	1.70	310

主観評価として，15 人の日本語話者によるヘッドホンを用いた聴取実験を行った．客観評価の結果から，提案手法における辞書の基底数は 1,000 とした．

図 4 の左側に，音質の評価結果を示す．評価基準として，Mean Opinion Score (MOS) による 5 段階評価 4 （5：とても良い，4：良い，3：普通，2：悪い，1：とても悪い）を用いた．図より，提案手法，NMF 声質変換はそれぞれ GMM 声質変換よりも音質が優れていることがわかる．この結果は t 検定により有意傾向が示されている．

図 4 の右側に話者性の評価結果を示す．評価基準は 5 段階評価（5：とても近い，4：近い，3：普通，2：遠い，1：とても遠い）を用いた．図より，3 つの手法の間には有意差がないことがわかる．

以上の結果から，提案手法は従来の NMF 声質変換法とほぼ同等の精度を得られ，計算コストをを削減することができたことがわかる．

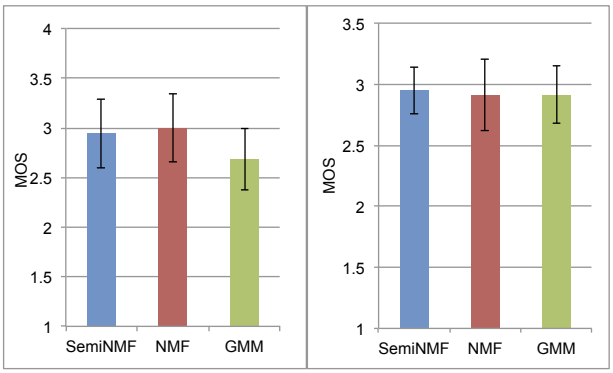

図 4　音質（左）と話者性（右）における MOS 試験
Fig. 4　MOS test on speech quality (left) and similarity (right)

5. おわりに

本稿では，ADMM に基づく最適化手法を採用した Semi-NMF による声質変換法を提案した．従来の NMF 声質変換の問題点であった計算コストとメモリ使用量を削減するため，NMF は Semi-NMF で置き換えられ，よりコンパクトなスペクトル特徴量を使用することが可能になった．さらに，従来の補助関数法に基づく Semi-NMF は収束速度に問題があったため，ADMM に基づく Semi-NMF を提案し，より少ない計算時間

でスパースなアクティビティが得られるようになった．また，基底数の少ないパラレル辞書行列を求めるため，パラレル辞書学習法を提案した．評価実験により，提案手法は従来の NMF 声質変換とほぼ同程度の精度で変換が可能でありながら，計算時間を大幅に削減することが可能になった．この手法はトピックモデルや超解像など，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．

また，今後は提案手法を構音障害者のための声質変換 [27] に応用していく予定である．

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.2614–2617, 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, 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, Issue: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] C. Ding, T. Li, and M.I. Jordan, “Convex and semi-nonnegative matrix factorization,” IEEE Trans. Pettern Analysis And Machine Intelligence, vol.32, no.1, pp.45–55, 2010.

[20] D.L. Sun and C. Févotte, “Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence,” in Proc. ICASSP, pp.6242–6246, 2014.

[21] 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.

[22] 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.

[23] R. Takashima, T. Takiguchi, and Y. Ariki, “Exemplar-based voice conversion using sparse representation in noisy environments,” IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol.E96-A, no.10, pp.1946–1953, 2013.

[24] 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.

[25] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigne, “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.

[26] T. En-Najjary, O. Roec, and T. Chonavel, “A voice conversion method based on joint pitch and spectral envelope transformation,” in Proc. ICSLP, pp.199–203, 2004.

[27] 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.