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
SP2014-53,WIT2014-8(2014-6)

スパース辞書学習による構音障害者の話者性を維持した声質変換

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　本研究では，アテトーゼ型脳性麻痺による構音障害者を対象とし，筋肉の不随意運動を原因とする障害者の不安定な発話を聞き取りやすく変換することを目指す.「自分らしい声で話したい」という障害者のニーズに応えるため，本研究では従来の統計的モデルによる声質変換とは異なる非負値行列因子分解（NMF）を用いた Exemplar-based 声質変換を用いて，話者性を維持しつつ聞き取りやすい音声に変換する．NMF 声質変換では，入力スペクトルは入力話者の exemplar の線形和で表現され，選ばれた exemplar を対応する出力話者のものと置き換えることで変換を行っていた．しかしこれまでの NMF 声質変換では，入力話者の exemplar から得られた重み行列をそのまま出力話者の重み行列として用いていた．実際には入力話者の重み行列と出力話者の重み行列は必ずしも一致するわけではなく，この問題が変換音声の劣化を引き起こしていると考えられていた．本研究ではこの問題を解決するため，NMF を用いたスパース辞書学習を行い，入力話者の線形重み行列を出力話者のものに変換するマッピング行列を導入する．提案手法の有効性を評価するため，従来の Gaussian Mixture Model に基づく声質変換，NMF 声質変換との比較実験を行った．
キーワード　声質変換，構音障害者，障害者支援，非負値行列因子分解

Individuality-preserving Voice Conversion for Articulation Disorders Using Sparse Dictionary Learning

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　We present in this paper a voice conversion (VC) method for a person with an articulation disorder resulting from athetoid cerebral palsy. The movement of such speakers is limited by their athetoid symptoms, and their consonants are often unstable or unclear, which makes it difficult for them to communicate. In our previous method, exemplar-based spectral conversion using Non-negative Matrix Factorization (NMF) was applied to a voice with an articulation disorder. To preserve the speaker's individuality, we used a combined dictionary that is constructed from the source speaker's vowels and target speaker's consonants. However, in this exemplar-based approach, source speaker's activity matrix which is estimated from input spectra and source speaker's exemplars are used as target speaker's. In this paper, we propose a sparse dictionary learning method for exemplar-based VC and estimate a mapping matrix between source speaker's activity and target speaker's activity. The effectiveness of this method was confirmed by comparing its effectiveness with that of a conventional Gaussian Mixture Model (GMM)-based method and a conventional NMF-based method.
Key words　Voice Conversion, Articulation Disorders, Asistive Technology, Non-negative Matrix Factorization

1. はじめに

これまで声質変換は，音声に含まれる話者性を変換しつつ音韻性・感情性などといった他の情報を維持する “話者変換” を目的として広く研究されてきた．しかし，近年では電気式人工喉頭を用いた発話の話者性の復元 [1] や無呼気音による非可聴つぶやき (Non-Audible Murmur: NAM）のささやき声への変換 [2] など声質変換を福祉分野へ応用する動きが進みつつある．

我々はこれまで，アテトーゼ型の脳性麻痺による構音障害者を対象とした音声支援技術を研究してきた [3], [4]．脳性麻痺とは，筋肉の動きをつかさどる脳の部分が受けた損傷が原因で筋肉の制御ができなくなり，けいれんや麻痺，そのほかの神経障害が起こる症状のことである [5]．脳性麻痺はその患部と症状によって様々に分類できるが，本研究では脳性麻痺患者の約 20% に発生するアテトーゼ型を対象とする．アテトーゼ型には，筋肉の随意運動や姿勢の調整を行っている大脳基底核（大脳皮質，視床や脳幹を結び付けている神経核の集まり）が損傷を受けたことによる筋肉が不随に動き正常に制御できない症状が現れる．とくに意図的な動作を行う場合や，緊張状態にある時に見られ，この運動障害の一つとして，正しく構音できない場合がある．症状は軽度から重度まで様々であるが，知能障害を合併していないケースや比較的知能障害の程度が軽いケースも多いのが特徴である．また，アテトーゼ型脳性麻痺による構音障害者の多くは，身体が不自由であるため，手話や文章読み上げ装置を用いたコミュニケーションは困難である．以上のことから，アテトーゼ型脳性麻痺による構音障害者のための音声支援には十分なニーズがあり，研究の必要性があるといえる．

本研究では，アテトーゼ型脳性麻痺による構音障害者のための声質変換技術を研究する．構音障害者の聞き取りにくい発話を聞き取り易く変換することを目的とする．話者変換を用いて障害者の発話音声を健常者の発話へと変換することも考えられるが，構音障害者のなかには「自分らしい声で話したい」というニーズがあり，障害者の話者性を維持した声質変換が求められている．文献 [4] において，我々は非負値行列因子分解 (Non-negative Matrix Factorization:NMF) [6] を用いた構音障害者のための，話者性を維持した声質変換を提案した．この手法では，アテトーゼ型脳性麻痺による構音障害者の発話特徴である子音が不安定になりやすいという性質を利用し，入力辞書に障害者発話，出力辞書に障害者の母音と健常者の子音とを組み合わせた Combined-dictionary を用いることで，障害者の話者性を維持した変換を実現した．

NMF 声質変換は，これまで声質変換で一般的であった混合正規分布モデル (Gaussian Mixture Model : GMM) を用いた手法 [7], [8] とは異なり，スパース表現に基づく Exemplar-based なアプローチである [9]．入力音声を，入力話者の exemplar の線形和で表現し，選ばれた exemplar を対応する出力話者のものと置き換えることで変換を行う．この手法は，GMM を用いた手法で問題とされてきた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．しかしながら，これまでの手法では，入力話者の exemplar から得られた重み行列をそのまま出力話者の重み行列と置き換えており，この問題が変換音声の劣化を引き起こしていると考えられていた．文献 [10] において我々は，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したが，この問題を根本的には解決できずにいた．

そこで，本研究ではアテトーゼ型構音障害者を対象とした声質変換手法の精度向上を目的とし，スパース辞書学習を提案する．入力話者の線形重み行列を出力話者のものに変換する変換行列を，辞書行列と合わせて NMF で学習する．スパース辞書は複数のクラスタに分割され，話者性を維持するため子音クラスタのスペクトルのみを変換する．

以下，第 2 章で従来の NMF 声質変換手法を説明する．第 3 章で本稿の提案手法を述べた後，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

2.1 概　　要

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．

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

$\mathbf{v}_l$ 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 (1) は二つの行列の内積で表される．

$$\mathbf{V} \approx \mathbf{W}\mathbf{H} \tag{2}$$

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

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 Wrapping (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 [11]．

2.2 Combined Dictionary による話者性の維持

パラレルな辞書を構成するため，障害者と健常者による複数の同一内容発話を用意する．同一内容発話から抽出されたスペクトル包絡は，DTW よってアライメントをとる．アクティビ

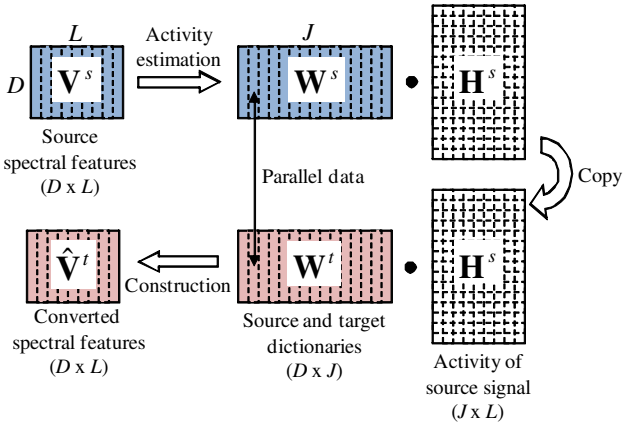


図 1　NMF を用いた声質変換の概要

Fig. 1　Basic approach of NMF-based voice conversion

ティを正確に推定するため，前後数フレームをまとめて 1 本のベクトルにしたセグメント特徴量を求めて入力特徴量とする．

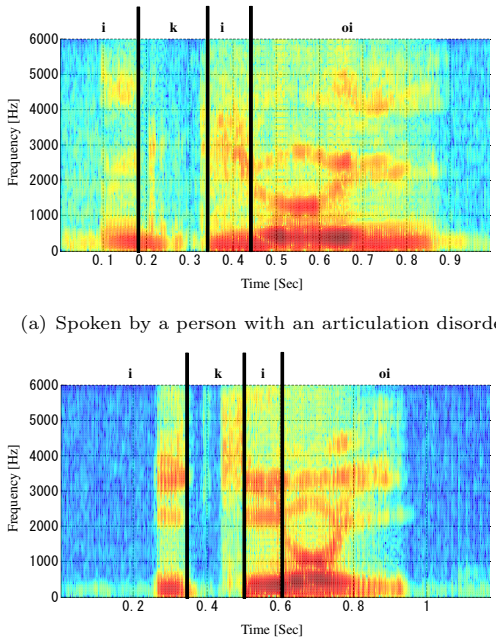
Fig. 2(a) に構音障害者による発話スペクトル，(b) に健常者による発話スペクトルを示す．障害者の発話例では，子音 ‘k’ に当たる部分のパワーが健常者と比較して弱くなっていることがわかる．その他の母音部分に関しては，障害者と健常者の間において，スペクトルの違いはあまり見られない．以上より，アテトーゼ型脳性麻痺による構音障害者の発話が聞き取りにくい原因は，子音にあると考えることができる．そこで，障害者の話者性を維持するため，アライメントをとったパラレルなスペクトル包絡から健常者の子音と障害者の母音を組み合わせて出力特徴量し，この出力特徴量を出力辞書に用いる．これらの処理を全ての同一内容発話について行い，抽出した特徴量を入力・出力それぞれについて水平に結合することで辞書行列とする．本論文ではこのような健常者の子音と障害者の母音から構成される出力辞書行列を Combined Dictionary と呼ぶ．

Fig. 1 において，出力話者辞書 $\mathbf{W}^t$ を Combined Dictionary におきかえる．入力された障害者スペクトルは，障害者の発話スペクトルから構成された入力話者辞書 $\mathbf{W}^s$ の基底の線形結合とその重みで表現される．重み行列 $\mathbf{H}^s$ は健常者の子音と障害者の母音から構成される出力辞書行列 $\mathbf{W}^t$ と掛け合わされる．このとき，入力された障害者スペクトルは障害者の母音基底と健常者の子音基底の線形結合で表現され，出力スペクトル $\hat{\mathbf{V}}^t$ は，入力スペクトル $\mathbf{V}^s$ のうち，子音フレームのみが健常者のものに変換されることになる．

3.　スパース辞書学習

3.1　概　　要

前章で述べた NMF を用いた声質変換法では，入力話者の辞書行列から推定したアクティビティをパラレルな出力話者の辞書行列と内積をとることで変換していた．これは，“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” という仮定に基づくものであった．



(a) Spoken by a person with an articulation disorder

(b) Spoken by a physically unimpaired person

図 2　障害者・健常者の発話スペクトル “ikioi”

Fig. 2　Examples of spectrogram //i k i oi

Fig. 3 に 2 人の日本語話者によって発話されたスペクトル “あこがれる”から推定したアクティビティを示す．発話スペクトルは DTW でアライメントがとられたもので，推定に用いた辞書は，パラレルな 1,000 基底から構成されているものを用いた．Fig. 3 からわかるように，同一発話スペクトルのアクティビティには差異があることがわかる．従って，本研究では辞書学習を行い，入力話者アクティビティを出力話者アクティビティへと変換する．

Fig. 1 で示されていた NMF 声質変換の概要は Fig. 4 のように変化する．第 1 段階では，入力話者スペクトル $\mathbf{V}^s$ が学習された入力話者辞書行列 $\hat{\mathbf{W}}^s$ の線形結合で表現される．このとき，基底の結合重みがアクティビティ $\mathbf{H}^s$ として NMF を用いて推定される．第 2 段階では，入力話者アクティビティ $\mathbf{H}^s$ が．学習された変換行列 $\mathbf{A}$ によって出力話者アクティビティ $\mathbf{H}^t$ へと変換される．第 3 段階では，変換されたアクティビティ $\mathbf{H}^t$ と出力話者辞書行列 $\mathbf{W}^t$ によって変換スペクトル $\hat{\mathbf{V}}^t$ が得られる．

本研究においては，障害者の子音のみを変換するため，スパース辞書学習は学習データの子音のみに適用される．

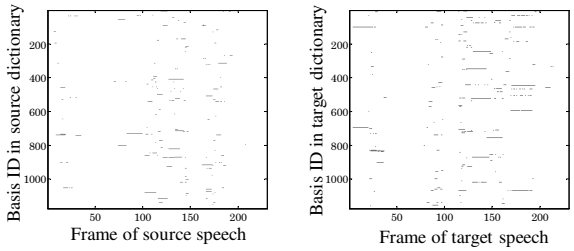


図 3　パラレル発話から得られたアクティビティ行列

Fig. 3　Activity matrices for parallel utterances

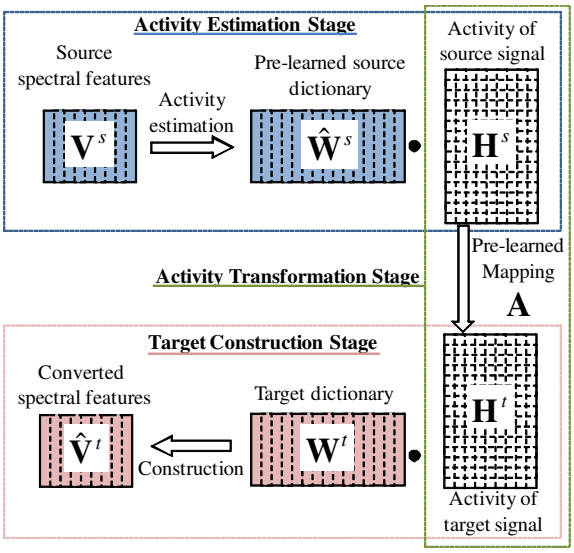


図 4　スパース辞書学習による声質変換の概要

Fig. 4　Flow chart of Voice Conversion Using Sparse Dictionary Learning

3.2　スパース辞書学習

辞書学習におけるコスト関数を以下のように定義する．

$$min_{(\mathbf{W}^s,\mathbf{A}^s)} d(\mathbf{V}^s,\mathbf{W}^s\mathbf{H}^s) + d(\mathbf{V}^t,\mathbf{W}^t\mathbf{H}^t) + \gamma d(\mathbf{H}^t,\mathbf{A}^s\mathbf{H}^s) + ||\lambda_s\mathbf{H}^s||_1 + ||\lambda_t\mathbf{H}^t||_1 \quad s.t.\ \ \mathbf{H}^s \geqq 0, \mathbf{H}^t \geqq 0 \quad (4)$$

第 1 項と第 2 項はそれぞれ $\mathbf{V}^s$ と $\mathbf{W}^s\mathbf{H}^s$，$\mathbf{V}^t$ と $\mathbf{W}^t\mathbf{H}^t$ の間の Kullback-Leibler(KL) 距離であり，第 3 項は $\mathbf{H}^t$ と $\mathbf{WH}^s$ の間の KL 距離である．$\mathbf{A}^s$ は $\mathbf{H}^t = \mathbf{A}^s\mathbf{H}^s$ を満たす変換行列であり，他の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．γ，λ_s，λ_t はそれぞれ第 3 項，入力話者アクティビティ行列のスパース制約，出力話者アクティビティ行列のスパース制約の重みである．

NMF に基づいて式 (4) を解くため，コスト関数を 3 つに分割する．まず，入力話者辞書行列 $\mathbf{W}^s$ とアクティビティ $\mathbf{H}^s$ は，入力話者の発話スペクトル $\mathbf{V}^s$ を用いて以下のコスト関数を最小化することで得られる．

$$d(\mathbf{V}^s,\mathbf{W}^s\mathbf{H}^s) + \gamma d(\mathbf{H}^t,\mathbf{A}^s\mathbf{H}^s) + ||\lambda_s\mathbf{H}^s||_1 \quad s.t. \quad \mathbf{H}^s \geqq 0 \quad (5)$$

式 (5) は $\mathbf{Z}^s = [\mathbf{V}^{s\mathsf{T}}, \gamma\mathbf{H}^{t\mathsf{T}}]^\mathsf{T}$，$\mathbf{D}^s = [\mathbf{W}^{s\mathsf{T}}, \gamma\mathbf{A}^{s\mathsf{T}}]^\mathsf{T}$ と置き換えることで，

$$d(\mathbf{Z}^s,\mathbf{D}^s\mathbf{H}^s) + ||\lambda_s\mathbf{H}^s||_1 \quad s.t. \quad \mathbf{H}^s \geqq 0 \quad (6)$$

のように書き換えられる．$\mathbf{W}^s$ と $\mathbf{H}^s$ は，次の更新式によって式 (6) を最小化できる．

$$\mathbf{D}^s_{n+1} = \mathbf{D}^s_n .* (\mathbf{H}^s_n(\mathbf{Z}^s./\mathbf{D}^s_n\mathbf{H}^s_n)^\mathsf{T}./(\mathbf{H}^s_n\mathbf{1}^{(L\times D)}))^\mathsf{T} \quad (7)$$

$$\mathbf{H}^s_{n+1} = \mathbf{H}^s_n .* (\mathbf{D}^{s^\mathsf{T}}_n(\mathbf{Z}^s./(\mathbf{D}^s_n\mathbf{H}^s_n))) ./(\mathbf{D}^{s^\mathsf{T}}_n\mathbf{1}^{(J\times L)} + \lambda_s\mathbf{1}^{(1\times L)}) \quad (8)$$

つづいて，出力話者のアクティビティ行列 $\mathbf{H}^t$ は，出力話者スペクトル $\mathbf{V}^t$ と，与えられた出力話者辞書行列 $\mathbf{W}^t$ によって以下のコスト関数を最小化することで得られる．

$$d(\mathbf{V}^t,\mathbf{W}^t\mathbf{H}^t) + \gamma d(\mathbf{H}^s,\mathbf{A}^t\mathbf{H}^t) + ||\lambda_t\mathbf{H}^t||_1 \quad s.t. \quad \mathbf{H}^t \geqq 0 \quad (9)$$

$\mathbf{Z}^t = [\mathbf{V}^{t\mathsf{T}}, \gamma\mathbf{H}^{s\mathsf{T}}]^\mathsf{T}$，$\mathbf{D}^t = [\mathbf{W}^{t\mathsf{T}}, \gamma\mathbf{A}^{t\mathsf{T}}]^\mathsf{T}$ と定義すれば，$\mathbf{H}^t$ は次の更新式によって，式 (9) を最小化できる．

$$\mathbf{H}^t_{n+1} = \mathbf{H}^t_n .* (\mathbf{D}^{t^\mathsf{T}}(\mathbf{Z}^t./(\mathbf{D}^t\mathbf{H}^t_n))) ./(\mathbf{D}^{t^\mathsf{T}}\mathbf{1}^{(J\times L)} + \lambda_t\mathbf{1}^{(1\times L)}) \quad (10)$$

最後に，変換行列 $\mathbf{A}^s$ は，次のコスト関数を最小化することで得られる．

$$d(\mathbf{H}^t,\mathbf{A}^s\mathbf{H}^s) \quad (11)$$

式 (11) を最小化する $\mathbf{A}^s$ は次に更新式によって得られる．

$$\mathbf{A}^s_{n+1} = \mathbf{A}^s_n .* (\mathbf{H}^s(\mathbf{H}^t./\mathbf{A}^s_n\mathbf{H}^s)^\mathsf{T}./(\mathbf{H}^s\mathbf{1}^{(L\times D)}))^\mathsf{T} \quad (12)$$

$\mathbf{A}^t$ は $\mathbf{A}^s$ と同様にして得られる．

3.3　辞書クラスタリング

第 2 章で述べた従来の NMF 声質変換では，学習データ全てをひとつの辞書行列のペアとして用いていた．しかしながら，発話スペクトルの多様性から，スパース辞書学習においてひとつの辞書行列ペアとひとつの変換行列で全ての発話を表現することは困難である．したがって，本論文ではパラレルデータをいくつかのクラスタに分類し，スパース辞書学習をクラスタ毎に適用する．

辞書クラスタリングでは，k 近傍法のコスト関数をユークリッド距離から KL 距離に置き換えたものを用いる．入力ベクトル $\mathbf{v}_l = [\mathbf{v}^{s\mathsf{T}}_l, \mathbf{v}^{t\mathsf{T}}_l]^\mathsf{T}$ は次のようなコスト関数によってクラスタリングされる．

$$Dis = \sum_{l=1}^{L} d(\mathbf{v}_l, \mathbf{m}_{c_l}) \quad (13)$$

$\mathbf{v}_l$，$\mathbf{m}_{c_l}$，L はそれぞれ l 番目の入力ベクトル，c_l-番目のクラスタ，フレーム数を表す．c_l は l 番目のフレームが属するクラスタを表し，以下のように決定される．

$$c_l = \arg\min_k d(\mathbf{v}_l, \mathbf{m}_k) \quad (14)$$

ここで，K はクラスタ数を表す．

本論文においては，子音のみを変換する為，学習データを子音と母音に分割し，それぞれ別にクラスタリングを行う．

3.4　辞書選択と変換

変換時，入力スペクトルは NMF によって，次式を用いて辞書を選択する．

$$c_l = \arg\min_k d(\mathbf{v}^s_l, \mathbf{W}^s_k\mathbf{h}^s_{kl}) + ||\lambda_s\mathbf{h}^s_{kl}||_1 \quad s.t. \quad \mathbf{h}^s_{kl} \geqq 0 \quad (15)$$

ここで，$\mathbf{W}^s{}_k$ は学習前の辞書行列である．

式 (15) によって子音の辞書行列が選択された場合，入力スペクトル $\mathbf{v}^s$ は，学習された辞書行列 $\hat{\mathbf{W}}^s$ のスパース表現に置

き換えられる．アクティビティ行列 $\mathbf{H}^s$ は，次のコスト関数を最小化することで求められる．

$$d(\mathbf{v}^s, \hat{\mathbf{W}}^s\mathbf{h}^s) + ||\lambda_s\mathbf{h}^s||_1 \quad s.t. \quad \mathbf{h}^s \geqq 0 \tag{16}$$

式 (16) を最小化する $\mathbf{H}^s$ は次の更新式から得られる．

$$\mathbf{h}^s_{n+1} = \mathbf{h}^s_n . * (\hat{\mathbf{W}}^{s^\mathsf{T}}_n (\mathbf{v}^s ./ (\hat{\mathbf{W}}^s_n \mathbf{h}^s_n))) \\ ./ (\hat{\mathbf{W}}^{s^\mathsf{T}}_n \mathbf{1}^{(J\times L)} + \lambda_s \mathbf{1}^{(1\times L)}) \tag{17}$$

得られたアクティビティと学習された出力話者辞書行列，変換行列によって，変換スペクトルは次のように得られる．

$$\hat{\mathbf{v}}^t = \mathbf{W}^t\mathbf{A}^s\mathbf{h}^s \tag{18}$$

一方，式 (15) によって母音の辞書行列が選択された場合，入力スペクトル $\mathbf{v}^s$ は辞書行列 $\hat{\mathbf{W}}^s$ のスパース表現に置き換えられ，これを変換スペクトルとする．

$$\hat{\mathbf{v}}^t = \mathbf{W}^s\mathbf{h}^s \tag{19}$$

提案手法のアルゴリズムを表 1 に示す．

表 1　スパース辞書学習による学習・変換アルゴリズム
Table 1　Algorithm of Sparse Dictionary Learning

Initialize for Dictionary Learning
Set source and training exemplars to $\mathbf{V}^s$ and $\mathbf{V}^t$.
Set the other target exemplars to $\mathbf{W}^t$.
$\mathbf{W}^s$ is initialized with a random matrix.
Clustering
Jointed $\mathbf{V}^s$ and $\mathbf{V}^t$ are clustered by (13).
For each iteration
For each cluster of consonant
- Optimize $\mathbf{W}^s$ and $\mathbf{H}^s$ by (7) and (8)
- Optimize $\mathbf{H}^t$ by (10)
- Optimize $\mathbf{A}^s$ and $\mathbf{A}^t$ by (12)

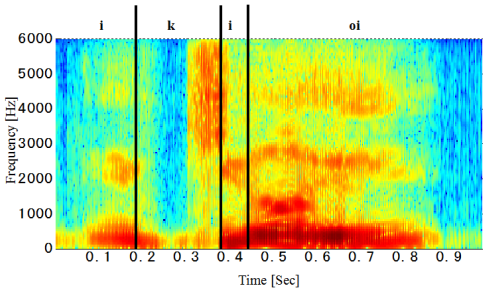
Initialize for Conversion
Set input spectra $\mathbf{V}^s$,
learned source dictionary $\hat{\mathbf{W}}^s$, target dictionary $\mathbf{W}^t$.
Clustering
Cluster the input spectrum $\mathbf{v}^s_l$ by (15).
For each iteration
For each cluster of consonant
- Optimize $\mathbf{H}^s$ by (17)

For each cluster
- Construct $\hat{\mathbf{V}}^t$ by (18) or (19)

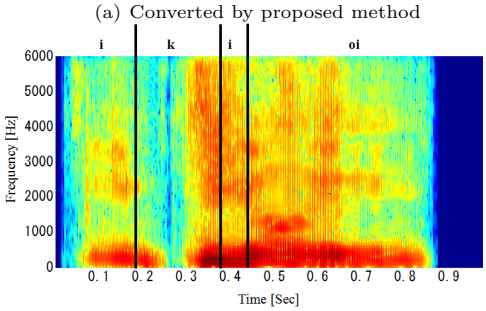
4. 評価実験

4.1 実験条件

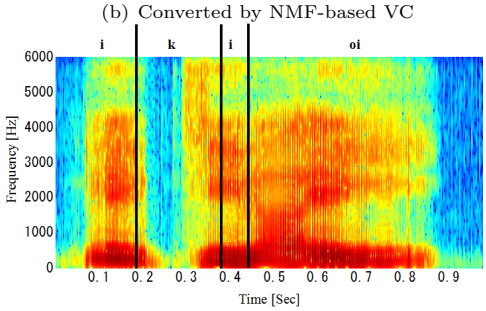
本実験では従来の GMM を用いた手法と，以前に提案したパラレルデータ全てを辞書に用いる手法 と比較を行った．入力話者として障害者の音声を使用するため，男性のアテトーゼ型構音障害者 1 名による 432 発話を収録した．発話内容は ATR 音素バランス単語 B セット [12] から 216 語を用いた．対となる健常者音声は，ATR 音声データベースに収録されている男性話者のものを使用した．それぞれの音声のサンプリング周波数



(a) Converted by proposed method



(b) Converted by NMF-based VC



(c) Converted by GMM-based VC

図 5　変換後のスペクトル “ikioi”
Fig. 5　Examples of converted spectrogram //i k i oi

は 12kHz，フレームシフトは 5ms である．対となったパラレルデータのうち，216 発話を学習に，残りの 216 発話をテストに用いた．入力特徴量，出力特徴量の次元数はそれぞれ 2565 次元と 513 次元である．パラレルデータ間の時間的なゆらぎを解消するため，STRAIGHT スペクトル [13] から求めた MFCC を用いて DTW を行った．

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．なお，本実験では F_0 は変換せず，障害者のものをそのまま用いた．

結果評価のために，成人男女 10 名による聴取実験を行った．評価項目は，聞き取りやすさ，話者性，自然性とした．聞き取りやすさの評価にはテストデータから構音障害者が発話しにくい 26 単語を選び，提案手法と従来手法それぞれで変換した音声と無変換の障害者音声を評価した．評価基準は MOS 評価基準に基づく主観評価（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，自然性の評価には，テストデータから 25 単語をランダムに選び，提案手法と従来手法それぞれで変換した．話者性の評価では，無変換の障害者の音声を聴いた後，各変換音声を聴き比べてどちらが障害者の声質に似ているかを選択する XAB 法とした．自然性の評価では，各変換音声を聴き比べてどちらが自然かを選択する 1 対 1 の対

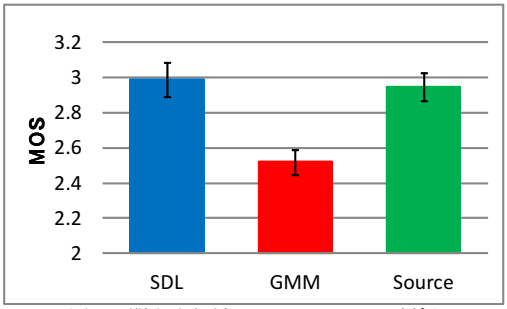


図 6　聞き取り易さにおける MOS 評価

Fig. 6　Results of MOS test on listening intelligibility

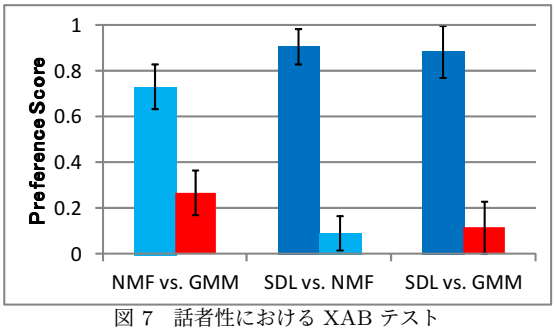


図 7　話者性における XAB テスト

Fig. 7　Preference scores for the individuality

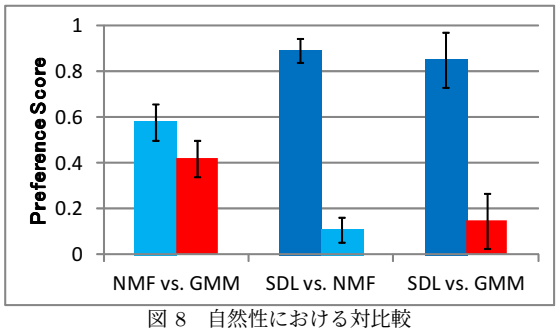


図 8　自然性における対比較

Fig. 8　Preference scores for the naturalness

比較法とした．いずれの評価項目も，静かな部屋においてヘッドホンを用いた両耳聴取を行った．

4.2　実験結果・考察

Fig. 5(a) に提案手法，(b) に従来の NMF 声質変換，(c) に GMM 声質変換で変換したスペクトルの例を示す．提案手法は他の手法と比較すると，母音部の誤変換がなくなっていることがわかる．聞き取りやすさの主観評価結果を Fig. 6 に示す．提案手法は，無変換の障害者音声と比較して，聞き取りやすさを向上させている一方，従来の GMM による声質変換は無変換音声と比較して劣化している．話者性の主観評価結果を Fig. 6 に示す．提案手法（SDL）は，従来の GMM，NMF による手法と比較して話者性が維持できている．従来の NMF による手法も GMM による手法と比較して話者性を維持している．自然性の主観評価結果を Fig. 8 に示す．提案手法（SDL）は，従来の GMM，NMF による手法と比較して，高い自然性を示している．

5.　おわりに

本論文では，アテトーゼ型構音障害者を対象とした話者性を維持した声質変換技術を提案した．これまで提案してきた NMF に基づく声質変換にスパース辞書学習を導入し，精度の向上を目指した．聴取実験によって，提案手法は構音障害者の話者性を維持しつつ聞き取りやすさを向上させられることを示した．さらに，従来手法と比較して，提案手法は自然性の高い音声で変換できることを示した．本実験では対象とした構音障害者は 1 名にとどまっているため，今後は話者数を増やして提案手法の有効性を確認する予定である．

文　献

[1] 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.

[2] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, “Speaking aid system for total laryngectomees using voice conversion of body transmitted artificial speech,” in Interspeech, pp. 148–151, 2006.

[3] T. Yoshioka, T. Takiguchi, and Y. Ariki, “Evaluation of random-projection-based feature combination on dysarthric speech recognition,” American Journal of Signal Processing, 2013 3 (3). doi:10.5923/j.ajsp.20130303.01, 2013.

[4] 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, 2014:5, doi:10.1186/1687-4722-2014-5, 2014.

[5] S.T. Canale and W.C. Campbell, “Campbell’s operative orthopaedics,” Technical report, Mosby-Year Book, 2002.

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

[7] Y. Stylianou, O. Cappe, and E. Moilines, “Continuous probabilistic transform for voice conversion,” IEEE Trans. Speech and Audio Processing, vol.6, no.2, pp. 131–142, 1998.

[8] 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.

[9] 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.

[10] 相原龍，滝口哲也，有木康雄，“辞書選択型非負値行列因子分解による構音障害者の声質変換,” 電子情報通信学会技術研究報告, vol. 113, no. 366, SP2013-87, pp. 71–76, 2013.

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

[12] 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.

[13] 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.