非負値行列因子分解による構音障害者の話者性を維持した声質変換*

☆相原龍, 高島遼一, 滝口哲也, 有木康雄 (神戸大)

1 はじめに

本研究では，脳性麻痺の一種である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を提案する．アテトーゼ現象は意図的な動作に緊張状態を発生させるために，障害者の発話，特に子音が不安定になる．本稿では，非負値行列因子分解 (Non-negative Matrix Factorization:NMF) [1] を用いた Exemplar-based な声質変換を構音障害者の発話に適用し，不安定な発話音声をより聞き取りやすく変換することを目指す．従来の統計モデルを用いた声質変換技術は，主として話者変換を目的としていたため，入力話者の声質は完全に別の話者の声質に変換されてしまう．本研究では，障害者の母音と健常者の子音を組み合わせた Combined Dictionary を用いることで，入力障害者音声の話者性を維持しつつ，より聞き取りやすく変換することを可能にした．

現在，日本だけでも約 3 万 4 千人の言語・聴覚障害者がおり，言語障害の原因の一つとして脳性麻痺を挙げることができる．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．その原因としては様々なものが考えられ，脳性麻痺は脳の損傷部分によって主に 4 つに分類されている [2].

その中のひとつであるアテトーゼ型の脳性麻痺では，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）に損傷を受けたことにより，アテトーゼと呼ばれる筋肉が不随に動き正常に制御できない症状が現れる．この現象は意図的な動作を行う場合や緊張状態がある時に多く発生するため，発話時に筋肉の緊張がおこり正しく構音できない場合がある．アテトーゼ型脳性麻痺によ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，手話や音声合成システム [3] を使うことは困難である．そのため，構音障害者のための声質変換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提案する．アテトーゼ型脳性麻痺による構音障害者の発話の特徴として，音声の子音が不明確になることがある．アテトーゼにより子音を発音する際の筋肉の動きが制限されるためにおこる現象である．本稿では，声質変換技術を構音障害者音声に適用し，子音を明瞭化することで，障害者の話者性を維持しつつ，音声を聞き取りやすく自然に変換することを目指す．

声質変換の一般的な手法として，GMM に基づく手法 [4] を挙げることができる．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している．この手法は話者変換 [5]，感情音声変換 [6, 7]，無喉頭音声変換 [8] などに使われてきたが，アテトーゼ型の構音障害者への応用は研究が進んでいない．この手法を構音障害者に適用し，健常者の声質へと変換した場合，音声は聞き取りやすく変換されると考えられるが，構音障害者音声の話者性は完全に別の健常者の話者性へと置き換えられてしまう．

そこで，本研究では NMF を用いた Exemplar-based な声質変換を用いて構音障害者の話者性を維持した声質変換を行う．NMF は非負制約のある行列分解手法であり，雑音除去や音声強調において広く使われている手法である [9]．NMF において，観測信号は少量の基底の組み合わせで表現することができる．この基底の集合行列を辞書行列と呼ぶ．

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

ここで，$\mathbf{x}_l$ は観測信号の l 番目のフレーム，$\mathbf{a}_j$ と $h_{j,l}$ はそれぞれ 辞書行列の j 番目の基底とその重み，$\mathbf{A}$ は辞書行列，$\mathbf{h}_l$ は l 番目のフレームにおける基底の大きさを表し，ここではアクティビティと呼ぶ．このようにして，観測信号は辞書行列のスパース表現で書き換えることができる．Gemmeke らは，音声認識において，HMM の尤度の代わりにアクティビティを辞書の音素スコアとして使うことで認識率を向上させた [10].

本研究では，辞書行列を固定する教師あり NMF [11] を声質変換に用いる．入力された構音障害者のスペクトルは，障害者の基底と重みの線形和で表現できる．障害者の基底を健常者の基底に置き換えることで，構音障害者のスペクトルは健常者のスペクトルへと変換される．ここで，構音障害者の子音はアテトーゼ現象のために不安定になっている．したがって，構音障害者の子音に対応する基底のみを変換することにより，障害者の話者性を維持した変換が可能になる．

以下，第 2 章で提案手法を説明し，第 3 章で従来手法である混合正規分布モデル（Gaussian Mixture Model:GMM）に基づく手法と提案手法を比較する．第 4 章で結論を述べ，本稿をまとめる．

2 NMF に基づく声質変換

2.1 Exemplar-based な声質変換

Fig. 1 に本手法の概要を示す．D, d, L, J はそれぞれ，入力特徴量の次元数，出力特徴量の次元数，辞書のフレーム数，辞書の基底の数を示す．ここで，辞書は入力話者，あるいは出力話者の基底の集合である．

提案手法では，主に障害者の基底から構成される入力辞書行列と，主に健常者の基底から構成される出力辞書行列の 2 つの辞書行列を用いる．これらは同一発話内容かつ同数の基底で構成されたパラレル

*Individuality-preserving Voice Conversion for Articulation Disorders Based on Non-negative Matrix Factorization by Ryo AIHARA, Ryoichi TAKASHIMA, Tetsuya TAKIGUCHI, Yasuo ARIKI (Kobe University)

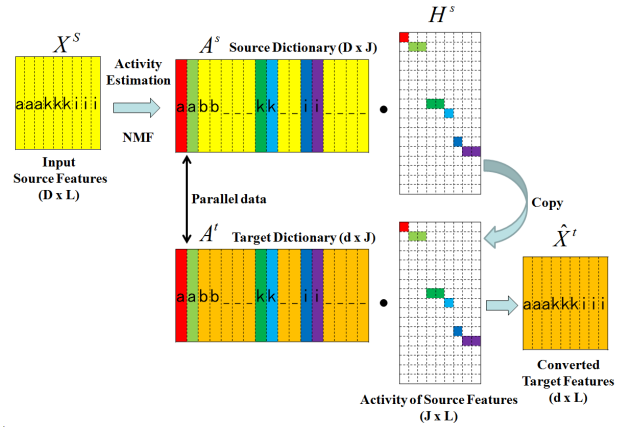

Fig. 1 Basic approach of NMF-based voice conversion

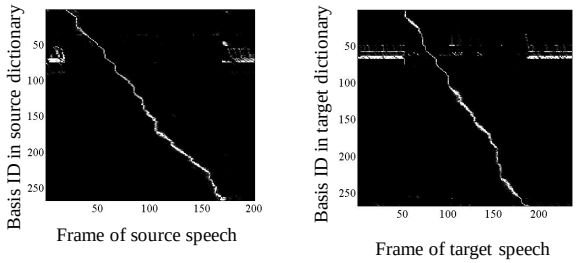

Fig. 2 Activity matrices for the articulation-disordered utterance (left) and well-ordered utterance (right)

データである．入力辞書行列の基底は入力特徴量で構成される．入力特徴量は障害者のスペクトル包絡とその周囲のフレームを含んだセグメント特徴量である．一方，出力辞書行列の基底は出力特徴量で構成される．出力特徴量は主に健常者のスペクトル包絡で構成されるが，セグメント特徴量は含まない．

入力特徴量 X^s は，NMF によって入力辞書行列 A^s に含まれる基底の線形和として表現される．このときの基底の重みがアクティビティ行列 H^s として推定される．つまり，アクティビティ行列には入力辞書行列の各基底に対する重みの情報が含まれることになる．次に，推定されたアクティビティ行列 H^s と出力辞書行列 H^t とかけあわせる．入力辞書行列と出力辞書行列は音韻的にパラレルであるため，入力辞書行列を用いて推定した重み行列と出力辞書行列によって，求めたい出力特徴量 $\hat{X}^t$ が出力辞書行列 A^t に含まれる基底の線形和として得られる．

Fig. 2 に単語 “ikioi”から推定したアクティビティ行列を示す．左側が構音障害者の発話を構音障害者の辞書行列で，右側が健常者の発話を健常者の辞書行列で推定したアクティビティである．簡単な例を示すために，辞書行列は障害者健常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 において，障害者と健常者間で高い重みを持つ基底の位置が類似している．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．したがって，Fig. 1 に示すように，求めたい出力特徴量は，出力辞書行列と，入力特徴量から推定されたアクティビティの線形和で表現することが可能であると考えられる．

スペクトル包絡は，音声分析合成 STRAIGHT [12] を用いて求められたものを用いる．STRAIGHT から抽出される F_0 や非周期成分といった他の特徴量は，変換せず，障害者のものをそのまま用いる．

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．Fig. 3 の左側はパラレルな辞書行列の構成法を示している．同一内容発話から抽出されたスペクトル包絡は，DP マッチングによってアライメントをとる．アクティビティを正確に推定するため，ある程度の長さをもったセグメント特徴量を求めて入力特徴量とする．さらに，障害者の話者性を維持するため，出力特徴量は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とする．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. 3 は，障害者の話者性を維持した声質変換法を示している．一般的に，母音は子音と比較して話者性を強く含むと言われている．一方，構音障害者の子音は不安定になりやすく，聞き取りにくくする原因となっている．健常者の子音と障害者の母音から構成される Combined Dictionary を用いることで，構音障害者の子音のみを変換でき，障害者の話者性を維持することができる．

2.3 アクティビティの推定方法

入力信号の l 番目のフレームは入力辞書行列の基底とその重みの線形和で表現される．

$$\begin{aligned} \mathbf{x}_l &= \mathbf{x}^s_l \\ &\approx \textstyle\sum_{j=1}^{J} \mathbf{a}^s_j h^s_{j,l} \\ &= \mathbf{A}^s\mathbf{h}_l \quad s.t. \quad \mathbf{h}_l \geq 0 \end{aligned} \tag{2}$$

$\mathbf{x}^s_l \in \mathbf{X}^s$ は入力信号の代表的なスペクトル包絡である．全ての入力信号に対して，式 (2) は以下のように書き換えることができる．

$$\mathbf{X}^s \approx \mathbf{A}^s\mathbf{H}^s \quad s.t. \quad \mathbf{H}^s \geq 0 \tag{3}$$

アクティビティ行列 $\mathbf{H}^s$ はスパース制約をもつ NMF に基づいて，以下のコスト関数を最小化することで推定することができる．

$$d(\mathbf{X}^s, \mathbf{A}^s\mathbf{H}^s) + ||(\lambda \mathbf{1}^{1\times L}).*\mathbf{H}^s||_1 \quad s.t. \quad \mathbf{H}^s \geq 0 \tag{4}$$

ここで，$\mathbf{1}$ は全要素が 1 の行列である．式 (4) の第 1 項は $\mathbf{X}^s$ と $\mathbf{A}^s\mathbf{H}^s$ の間のカルバック・ライブラー情報量であり，第 2 項は $\mathbf{H}^s$ をスパースにするための L1 ノルム正規化を伴ったスパース制約項である．それぞれの exemplar に対するスパース制約は $\lambda^T = [\lambda_1 \ldots \lambda_J]$ のようにして定められる．本研究では，スパース制約

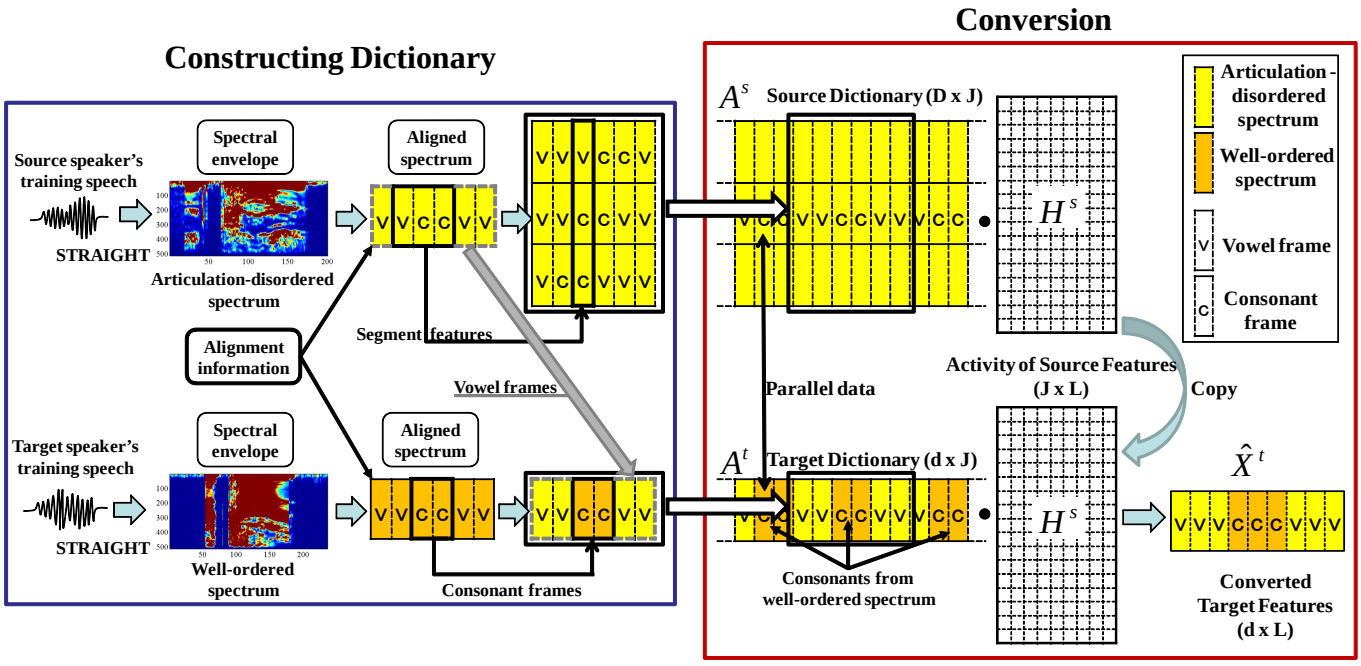

Fig. 3　Individuality-preserving voice conversion

の重み λ は 0.1 とした．以下の更新式を繰り返し適用することで，式 (4) のコスト関数を最小化できる [1].

$$\mathbf{H}^s_{n+1} = \mathbf{H}^s_n .* (\mathbf{A}^{sT}(\mathbf{X}^s ./ (\mathbf{A}^s\mathbf{H}^s))) ./ (\mathbf{A}^{sT}\mathbf{1}^{\mathbf{D}\times L} + \lambda \mathbf{1}^{1\times L}) \quad (5)$$

$\mathbf{H}^s$ のスパース性をさらに向上させるために，閾値以下の $\mathbf{H}^s$ の要素は 0 に丸められる．

推定されたアクティビティと出力辞書行列により，求めたい変換された特徴量は以下のように表現できる．

$$\hat{\mathbf{X}}^t = (\mathbf{A}^t\mathbf{H}^s) \quad (6)$$

3　実験結果

3.1　実験条件

提案手法を評価するため，構音障害者と健常者のパラレルなデータ対を用意した．障害者音声として使用するため，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 音素バランス単語セットから 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は 16kHz，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 スペクトルから求めたメルケプストラム係数を用いて DP マッチングを行った．

以下の実験結果では，提案手法である NMF に基づく声質変換と従来手法である GMM による声質変換を比較している．GMM に基づく声質変換では，入力特徴量・出力特徴量ともに障害者・健常者から STRAIGHT で抽出し求めた低次 16 次元のケプストラム係数を用いた．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．

3.2　主観評価実験

成人男女 5 名による，聴取実験を行った．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．評価項目は，聞き取りやすさ (listening intelligibility)，子音の明瞭性 (clarity of consonants)，話者性 (similarity)，自然性 (naturalness) の 4 項目とした．

「聞き取りやすさ」と「子音の明瞭性」の評価にはテストデータからランダムに 50 単語選び，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用いた．「聞き取りやすさ」の評価では，正解テキストを被験者に提示し音声がその単語にどの程度聞こえるかを評価した．「子音の明瞭性」の評価では音声中の子音がどの程度明瞭に聞こえるかを評価した．いずれの評価項目も，変換実験に用いた健常者音声を MOS 評価基準に基づく評価 5 の基準として被験者に与えた．

「話者性」「自然性」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 23 単語を選び，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，無変換の障害者音声に対し，変換音声が障害者の声質にどのくらい近いかを評価した．「自然性」の評価では，変換音声がどの程度音声として自然に聞こえるかを評価した．

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．

3.3　結果と考察

Fig. 4 に「聞き取りやすさ (intelligibility)」と「子音の明瞭性 (consonants)」の評価結果を示す．提案手法による NMF に基づく声質変換は，無変換の障害者音声と比較して聞き取りやすさと子音の明瞭性を向

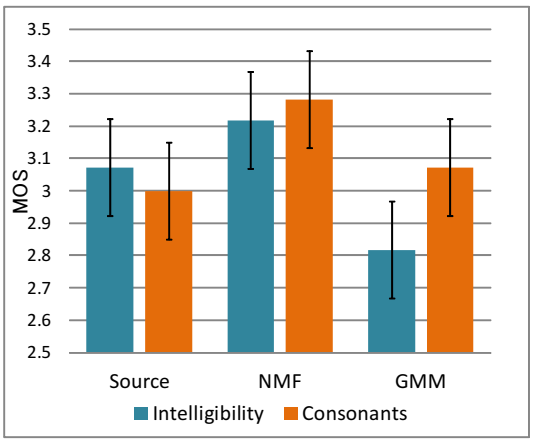

Fig. 4　Results of MOS test on listening intelligibility and clarity of consonants

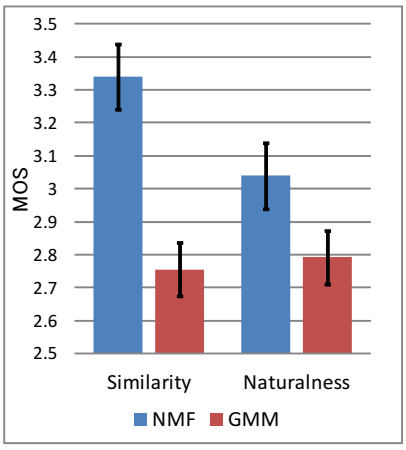

Fig. 5　Results of MOS test on the similarity to the source speaker and naturalness

上させていることがわかる．一方，従来手法であるGMM に基づく声質変換では子音の明瞭性は向上しているものの，聞き取りやすさは無変換音声と比較して劣化している．これは，変換ノイズによるものと考えられる．NMF に基づく変換手法も変換ノイズを発生させるものの，GMM に基づくものよりは少ない変換ノイズとなっている．

Fig. 5 に「話者性 (similarity)」と「自然性 (naturalness)」の評価結果を示す．提案手法である NMF に基づく声質変換は，従来手法と比較して話者性を維持できていることがわかる．これは，従来手法は全ての入力スペクトル包絡を健常者のものに変換しているのに対し，提案手法では Combined dictionary を用いて子音のみの変換を行っているためである．自然性についても，提案手法は異なる話者の子音と母音を組み合わせているのにも関わらず，従来手法よりも高いスコアを得ている．これは，従来手法の変換ノイズの問題に加え，提案手法は母音部分をほぼ無変換で合成できるためと考えられる．

4　結論

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．聴取実験によって，NMF に基づく提案手法は構音障害者の音声の子音を明瞭にし，聞き取りやすく変換できることを示した．さらに，GMM に基づく従来手法と比べて，提案手法は Combined dictionary を使うことでより自然で障害者の話者性を維持した音声に変換できることも示した．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．

参考文献

[1] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in Proc. Neural Information Processing System, pp. 556-562, 2001.
[2] S. T. Canale and W. C. Campbell, “Campbell’s operative orthopaedics,” Tech. Rep., Mosby-Year Book, 2002.
[3] C. Veaux *et al.*, “Using HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders,” Proc. Interspeech, 2012.
[4] Y. Stylianou *et al.*, “Continuous probabilistic transform for voice conversion,” IEEE Trans. Speech and Audio Processing, Vol. 6, No. 2, pp. 131-142, 1998.
[5] 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.
[6] Y. Iwami *et al.*, “GMM-based voice conversion applied to emotional speech synthesis,” IEEE Trans. Seech and Audio Proc., Vol. 7, pp. 2401-2404, 1999.
[7] R. Aihara *et al.*, “GMM-based emotional voice conversion using spectrum and prosody features,” American Journal of Signal Processing, Vol. 2 No.5, 2012.
[8] 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.
[9] T. Virtanen, “Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria,” IEEE Trans. Audio, Speech, Lang. Process., Vol. 15, No. 3, pp. 1066-1074, 2007.
[10] J. F. Gemmeke and T. Virtanen, “Noise robust exemplar-based connected digit recognition,” ICASSP, pp. 4546-4549, 2010.
[11] M. N. Schmidt and R. K. Olsson, “Single-channel speech separation using sparse non-negative matrix factorization,” INTERSPEECH, 2006.
[12] H. Kawahara *et al.*, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds,” Speech Communication, Vol. 27, No. 3-4, 1999.