非負値行列因子分解による構音障害者の声質変換*

☆相原龍, 高島遼一, 滝口哲也, 有木康雄 (神戸大)

1 はじめに

近年，情報技術の福祉分野への応用が進んでいる．例えば，画像認識技術の応用による手話認識 [1]，文章読み上げシステム [2]，無喉頭音声変換 [3] など，その応用領域は幅広い．本研究では，脳性麻痺による構音障害者に焦点をあて，構音障害者の音声を健常者のものに変換することで，より聞き取りやすくすることを目指す．

現在，日本だけでも約 3 万 4 千人の言語・聴覚障害者がいる．言語障害の原因の一つとして脳性麻痺が考えられる．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．出生前・出生時・出生直後の脳への酸素供給，出生前の胎内感染，妊娠中毒症，分娩時の外傷，仮死状態，未熟出生，出生後の脳を覆う組織の炎症や外傷性損傷などが原因として考えられる．

脳性麻痺は，脳の損傷部分によって痙直型（大脳皮質），アテトーゼ型（中脳もしくは脳基底核），失調型（小脳），混合型（脳の広範囲）に分類される [4]．それぞれ，痙直型は正常な筋の伸張反射が過度になる，アテトーゼ型はアテトーゼと呼ばれる筋肉の不随意運動を伴う，失調型は協調運動の障害が現れ，混合型はそれぞれの症状が混合して現れる，というような症状が見られる．

本論文では，アテトーゼ型の脳性麻痺による構音障害者を対象としている．アテトーゼ型は，脳性麻痺患者の約 20% に発生する．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．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り，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．

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を挙げることができる．アテトーゼ現象により，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．本研究では，声質変換技術を構音障害者に適用し，音声の子音強調を行う．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．これまで，声質変換技術は話者変換，感情変換，あるい無喉頭音声変換に応用されてきたが，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．

声質変換の一般的な手法として，混合正規分布モデル (Gaussian Mixture Model:GMM) に基づく手法 [5] 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する．この手法の欠点として，“過学習”があり，モデルの自由度が学習データの量に対して大きすぎるときに発生することが知られている．本研究では，構音障害者の音声収録が困難であるため，GMM に基づく声質変換を行うのに十分な学習データが得られなかった．

そこで，本研究では非負値行列因子分解 (Non-negative Matrix Factorization:NMF) に基づく声質変換法を用いて構音障害者の音声を健常者の音声に変換し，子音強調を行う．NMF は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．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．

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

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行列と呼ぶ．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 [6]．

NMF には，教師あり [7] と教師なし [8] の 2 つの手法が存在し，本研究では教師あり NMF を用いる．学習に使う音声データからスペクトル包絡を抽出し，入力話者（構音障害者）の辞書と出力話者（健常者）の辞書から構成されるパラレルな辞書行列を作成する．入力音声のスペクトルは，入力話者の辞書のス

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

パース表現に変換できる．このとき，入力話者の辞書行列から選ばれた基底を，出力話者の辞書行列の同一アライメントの基底と交換することで，入力話者のスペクトルは出力話者のスペクトルと置き換えられる．NMF に基づく手法には，統計モデルが導入されていないため，過学習が起こりにくいと考えられる．

2　NMF による声質変換法

2.1　手法の流れ

Fig. 1 に本手法の概略を示す．D, L, J はそれぞれ，スペクトルの次元数，入力スペクトルのフレーム数，辞書行列のフレーム数である．まず，入力話者（構音障害者）スペクトルと出力話者（健常者）スペクトルから構成される，パラレルな辞書行列を用意する．辞書行列の構成法を Fig. 2 に示す．入力話者・出力話者の同一発話データから，音声分析合成 STRAIGHT [9] を用いてスペクトル包絡を抽出する．抽出されたスペクトルから，メルケプストラム係数を求め，DP マッチングを用いてアライメント情報を得る．こうして得られたアライメント情報をスペクトルに適用して各々のスペク トルの長さを調整し，学習データとして連結する．

変換する入力音声は，STRAIGHT でスペクトル包絡を抽出し，Fig. 1 に示すように，NMF を用いて入力話者の辞書行列とアクティビティ行列に分解する．アクティビティ行列には入力スペクトルが，辞書行列のどの基底で，どのくらいの大きさで構成されるかの情報が含まれる．

Fig. 3 は障害者と健常者の単語 “ikioi”のアクティビティをそれぞれの辞書行列から推定したものである．アクティビティで高いエネルギーを示している要素が，障害者と健常者で類似していることがわかる．辞書行列がパラレルであるため，障害者で使われている基底と同じ基底が健常者で使われると考えられる．

Fig. 1 で推定された入力スペクトルのアクティビティ行列は，出力話者の辞書行列とかけあわされる．こうして，出力話者の辞書行列から入力話者の基底と同一アライメントの基底が選ばれ，入力話者のスペクトルが出力話者のスペクトルと変換される．

2.2　アクティビティ行列の推定

入力音声スペクトルの l 番目のフレームは，入力話者の辞書行列とアクティビティ行列によって以下のように表せる．

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

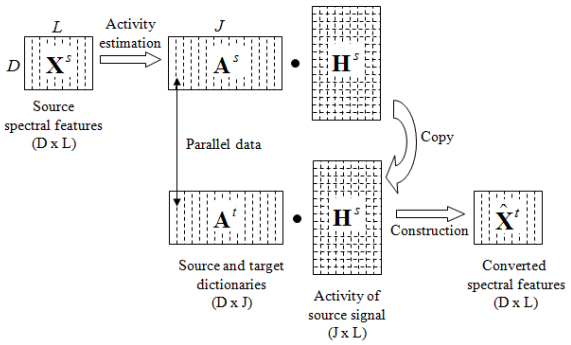

Fig. 1　Basic approach of exemplar-based voice conversion

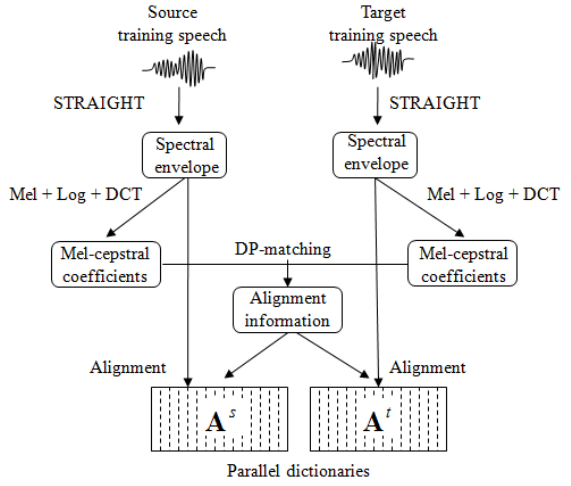

Fig. 2　Construction of source and target dictionaries

ここで，$\mathbf{x}_l^s$ は，入力音声の代表的なスペクトルである．入力音声の全スペクトルに対して，式 (2) は以下のように書き換えることができる．

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

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．

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

ここで，$\mathbf{1}$ は全要素が 1 の行列である．式 (4) の第 1 項は $\mathbf{X}$ 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．本研究では，スパース制約の重み λ は 0.1 とした．

3　評価実験

3.1　実験条件

実験データとして, 男性のアテトーゼ型構音障害者 1 名のデータを収録した．発話内容は，ATR 音素バ

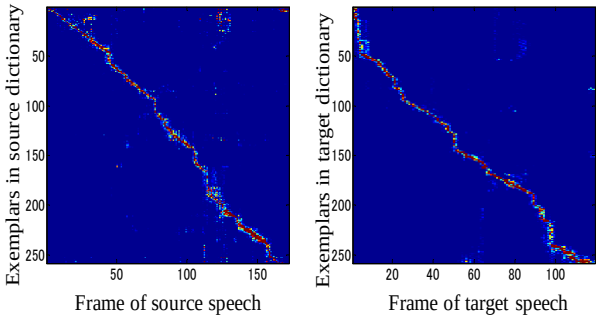

Fig. 3　Activity matrices of the source signal (left) and target signal (right)

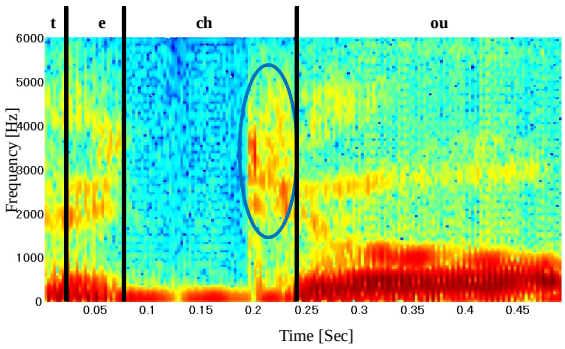

Fig. 4　Example of a spectrogram spoken by a person with an articulation disorder //t e ch ou

ランス単語 216 語から選択した 50 語を用いた．対となる健常者の音声データは，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．音声のサンプリング周波数は 16kHz で，フレームシフトは 5ms である．例として，実験に用いた構音障害者の発話 “techou ”のスペクトルを Fig. 4 に，健常者の同一単語のスペクトルを Fig. 5 に示す．実験は，変換辞書に変換音声を含むクローズドな実験と，変換音声を含まずクロスバリデーションを用いたオープンな実験の 2 種類を行った．

F_0 の変換は，以下の式で表される平均と分散のみを考慮した線形変換を行った．

$$y_t = \frac{\sigma^{(y)}}{\sigma^{(x)}}(x_t - \mu^{(x)}) + \mu^{(y)} \tag{5}$$

ここで，x_t と y_t はそれぞれ障害者・健常者の t 番目のフレームにおける対数をとった F_0 である．変数 $\mu^{(x)}$ ，$\sigma^{(x)}$，$\mu^{(y)}$ ，$\sigma^{(y)}$ はそれぞれ，障害者・健常者の対数 F_0 の平均と分散であり，辞書を構成する学習データから求められる．

変換辞書となるパラレルデータの作成にあたって，まず障害者・健常者両方の音声に対して，HMM の強制アライメントを用いてラベル付けを行い，DP マッチングでデータ間の対応をとった．障害者の音声に関しては，子音の欠落や，健常者にはない無声区間の混入などのためラベル付けが不正確になる場合があり，このような場合はラベル付けを手動で訂正した．

3.2　実験結果

成人男女 5 名による聴取実験を行った．音質と子音の明瞭性の 2 つの項目について，MOS 評価基準に基づく 5 段階評価 (5:とてもよい，4:よい，3:ふつう，2:わるい，1:とてもわるい) の主観評価実験を行った．Fig. 4 の楕円で囲まれた部分は，構音障害者の子音 “ch ”を示している．Fig. 5 に示した，健常者の子音 “ch ”と比較して不明瞭になっていることがわかる．この子音の不明瞭さが，構音障害者の音声を聞き取

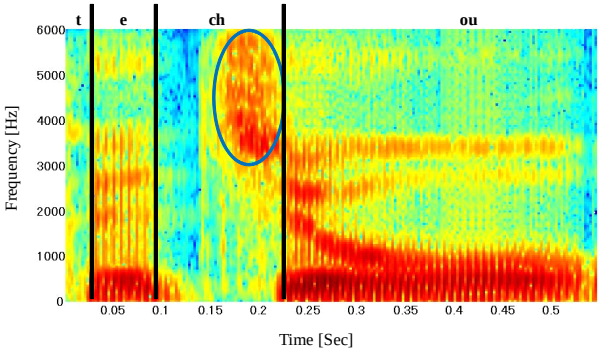

Fig. 5　Example of a spectrogram spoken by a physically unimpaired person //t e ch ou

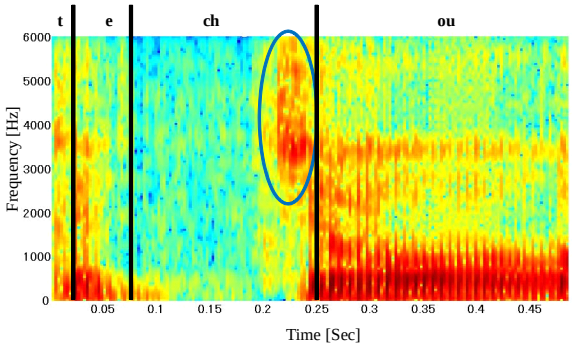

Fig. 6　Example of a spectrogram converted by a closed experiment //t e ch ou

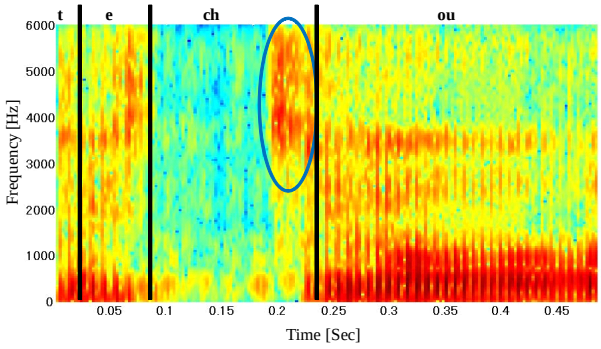

Fig. 7　Example of a spectrogram converted by an open experiment //t e ch ou

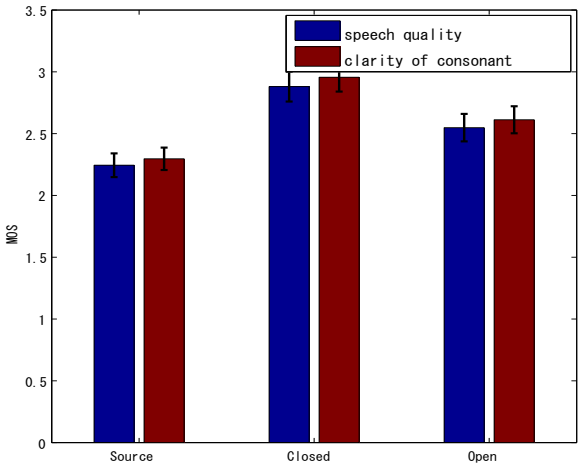

Fig. 8　Results of MOS test on speech quality and clearness of consonants

りにくくする原因の一つになっている．

Fig. 6 に，クローズドな実験で構音障害者音声 “techou ”を変換したスペクトルを示す．Fig. 4 の変換前の構音障害者音声の楕円部と比較して，Fig. 6 の楕円で囲まれた “ch ”の子音は明瞭になっていることがわかる．Fig. 7 は，同じ構音障害者音声 “techou ”を，オープンで変換したものである．こちらの楕円で囲まれた “ch ”の子音も，構音障害者のものと比較して明瞭化していることがわかる．

Fig. 8 に主観評価実験の結果を示す．提案手法によって，音質，子音の明瞭性ともに変換前のものより向上していることがわかる．構音障害者の音声は，音がこもりがちであるため音質がわるく，子音が不明瞭であるため聞き取りにくいが，変換によって音質が向上し，聞き取りやすくなっている．クローズドな変換では，音質，子音の明瞭性ともに大きく向上している．オープンな変換においては，クローズドな変換よりは落ちるものの，変換前のものと比較して音質，子音の明瞭性ともに向上していることがわかり，提案手法の有効性を示している．

4　おわりに

本論文では，子音が欠落し聞き取りにくい構音障害者の音声に対して NMF による声質変換を行うことにより，音質の改善と子音強調を行った．NMF に基づく声質変換は統計的手法を用いていないため，収録が困難な構音障害者の少量の学習データでも変換を行うことができた．

本研究では被験者は 1 名のみであったため，今後は学習する話者数を増やして手法の有効性を確認する必要がある．また GMM に基づく手法など，他の声質変換法と比較していく予定である．

参考文献

[1] J. Lin *et al.*, “Capturing human hand motion in image sequences,” IEEE Motion and Video Computing Workshop, pp. 99–104, 2002.

[2] M. K. Bashar *et al.*, “Unsupervised Texture Segmentation via Wavelet-based Locally Orderless Images (WLOIs) and SOM,” 6th IASTED International Conference COMPUTER GRAPHICS AND IMAGING, pp. 279–284, 2003.

[3] K. Nakamura *et al.*, “Speaking Aid System for Total Laryngectomees Using Voice Conversion of Body Transmitted Artificial Speech,” INTERSPEECH, pp. 1395–1398, 2006.

[4] S. T. Canale and W. C. Campbell, “Campbell’s Operative Orthopaedics,” Mosby-Year Book, 2002.

[5] Y. Stylianou *et al.*, “Statistical methods for voice quality transformation,” Eurospeech, pp. 447-450, 1995.

[6] J. F. Gemmeke, and T. Virtanen, “Noise robust exemplar-based connected digit recognition,” ICASSP, pp. 4546-4549, 2010.

[7] P. Smaragdis and J. Brown, “Non-negative matrix factorization for polyphonic music transcription, ” in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 177–180, 2003.

[8] M. N. Schmidt and R. K. Olsson, “Single-channel speech separation using sparse non-negative matrix factorization, ” INTERSPEECH, pp.2614–2617, 2006.

[9] H. Kawahara *et al.*, “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.