# 辞書選択型NMFを用いた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)

## 1 はじめに

現在，日本だけでも約 4 万 2 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あげることができる．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．それらの原因は多様であり，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あげられている [1].

脳性麻痺は脳の損傷部分によって 4 つの種類に分類され，そのなかでも，脳性麻痺患者の約 20%に発生するアテトーゼ型は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．この症状は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，そ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．アテトーゼ症状は軽度から重度まで様々であり，さらに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ことから，アテトーゼ型脳性麻痺による構音障害者を対象とした発話支援システムが求められている．

手足が不自由なアテトーゼ型構音障害者の為の発話支援システムとして，声質変換が考えられる．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” [2] を目的として広く研究されてきたが，近年は感情変換 [3]，あるい無喉頭音声変換 [4] など様々なタスクに応用されてきた．

構音障害者の為の声質変換において，話者性の維持が大きな課題となる．従来，最も一般的な混合正規分布モデル (Gaussian Mixture Model : GMM) を用いた手法 [2] は話者変換を目的として研究されてきたため，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．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 [5] において，我々はアテトーゼ型脳性麻痺による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，従来の声質変換手法で用いられていたパラレルデータから，入力話者の音声辞書（入力辞書）と出力話者の音声辞書（出力辞書）からなる同一発話内容のパラレル辞書を構築する．変換の際は非負値行列因子分解 (Non-negative Matrix Factorization:NMF) [6] を用いて，入力音声を入力辞書のスパース表現にする．そして得られた入力辞書内のサンプル毎の重み係数（アクティビティ）に基づいて，出力辞書内のサンプルを線形結合することで，出力話者の音声スペクトルへと変換する．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．

しかしながら，この手法では同一フレーム内において，障害者の母音と健常者の子音の線形結合ができてしまうという問題点があった．さらに，パラレルデータの全フレームをそのまま辞書の基底として用いており，辞書のサイズが膨大となっていた．文献 [7] 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，辞書選択手法を導入し，健常者の話者変換においてその有効性が確認された．この手法では，入力・出力話者辞書を音素カテゴリに分けた副辞書を作成し，NMF を用いて音素カテゴリ認識を行った後，選択した副辞書上でマッピングを行うことで声質変換を行った．

本論文ではアテトーゼ型構音障害者を対象として，辞書選択を用いた NMF 声質変換による話者性を維持した声質変換を提案する．出力話者のカテゴリ辞書のうち，子音に関するカテゴリ辞書のみに健常者のスペクトルを用い，母音に関するカテゴリ辞書に障害者のスペクトルを用いることで，障害者の話者性を維持した声質変換を行う．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

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．

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

$\mathbf{x}_l$ 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．$\mathbf{a}_j$ 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 $\mathbf{a}_j$ とする．基底を並べた行列 $\mathbf{A} = [\mathbf{a}_1 \dots \mathbf{a}_J]$ 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．

$$\mathbf{X} \approx \mathbf{A}\mathbf{H} \tag{2}$$

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

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音声辞書と出力話者音声辞書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動的計画法 (DP)

*Individuality-preserving Voice Conversion for Articulation Disorders Based on Dictionary Selective NMF by Ryo AIHARA, Tetsuya TAKIGUCHI, Yasuo ARIKI (Kobe University)

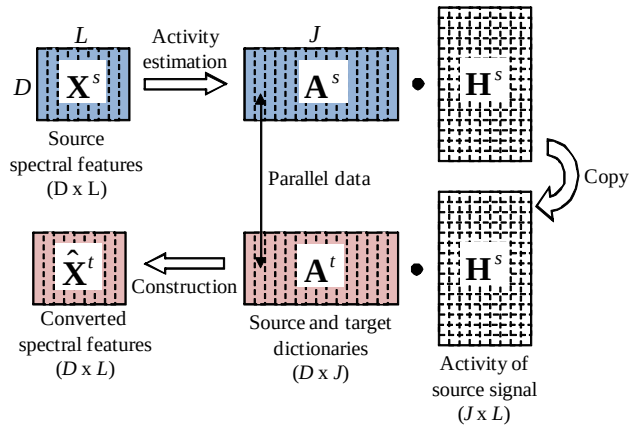


Fig. 1　Basic approach of NMF-based voice conversion

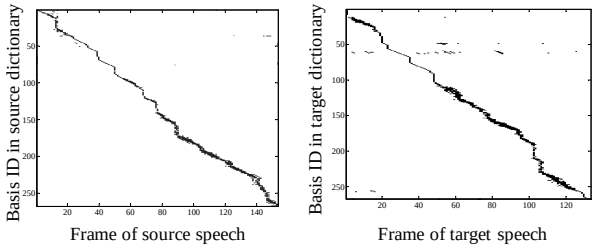


Fig. 2　Activity matrices for parallel utterances

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．

このとき，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仮定できる．Fig. 2 は 2 人の話者による発話単語 “ikioi” を入力信号としたときの，NMF で推定したアクティビティ行列である．ただし，簡易な例を示すため，辞書行列は 2 話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 を見ると，高い重みを持つ基底の位置が類似していることがわかる．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と考えられる．以上の仮定に基づき，入力音声は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．

## 3　辞書選択による NMF 声質変換

これまでの NMF による声質変換法では，学習サンプル全てを基底としてパラレルな辞書を構成していた．結果，辞書に含まれる基底数が膨大になり，入力信号が値の小さなアクティビティから成る多数の基底の線形結合で表現されてしまうという問題があった．線形結合の基底数が増加すると，パラレルな発話のアクティビティの形状が類似するという仮定が成り立たなくなり，これが変換音声の劣化につながっていると考えられる．そこで，入力信号を表現する基底数を限定するため，音素カテゴリごとに副辞書を作成し，限定された基底内でスパースコーディングを行う．

### 3.1　辞書構成法

Fig. 3 は副辞書の構成法を示したものである．従来の NMF による声質変換と同様にして障害者の発話スペクトルから構成される入力話者辞書 $\mathbf{A}^s$ と，健常者の発話スペクトルから構成される出力話者辞書行列 $\mathbf{A}^t$ を求める．次にパラレルな辞書行列は，Table 1 に示す音素カテゴリに従って，$K$ 個 の副辞書に分けられる．

$$\mathbf{\Phi}_k^s = [\mathbf{x}_1^{(k)}, \ldots, \mathbf{x}_{N_k}^{(k)}] \tag{4}$$

ここで，$\mathbf{\Phi}_k^s$，$N_k$ はそれぞれ $k$ 番目の副辞書，副辞書の基底数を表す．障害者の話者性を維持するため，音素カテゴリに分けられた副辞書のうち，母音の出力話者副辞書は，障害者の発話スペクトルから構成される入力話者副辞書と同じものを用いる．

さらに，それぞれの副辞書を表現する代表基底を集め，副辞書を選択するためのカテゴライズ辞書 $\mathbf{\Theta}$ を作成する．それぞれの代表基底は，副辞書において仮定した Gaussian Mixture Model (GMM) の平均ベクトルから構成される．

$$p(\mathbf{x}_n^{(k)}) = \sum_{m=1}^{M_k} \alpha_m^{(k)} N(\mathbf{x}_n^{(k)}, \boldsymbol{\mu}_m^{(k)}, \mathbf{\Sigma}_m^{(k)}) \tag{5}$$

ここで，$M_k$，$\alpha_m^{(k)}$，$\boldsymbol{\mu}_m^{(k)}$，$\mathbf{\Sigma}_m^{(k)}$ はそれぞれ $k$ 番目の副辞書における正規分布の混合数，$m$ 番目の正規分布の混合重み，平均，分散である．それぞれのパラメータは EM アルゴリズムを用いて推定する．正規分布の混合数は，対応する副辞書の基底数に応じて定めるため，副辞書を代表する基底数も副辞書ごとに異なる．

カテゴライズ辞書の基底は，それぞれの副辞書において仮定した GMM の平均ベクトルから構成され，下のように表現される．

$$\boldsymbol{\theta}_k = [\boldsymbol{\mu}_1^{(k)}, \ldots, \boldsymbol{\mu}_{M_k}^{(k)}] \tag{6}$$
$$\mathbf{\Theta} = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K] \tag{7}$$

ここで，$\mathbf{\Theta}$，$M_k$ はそれぞれカテゴライズ辞書，$k$ 番目の副辞書の平均ベクトルの数を表す．

### 3.2　辞書選択・変換法

Fig. 4 は変換手法の流れを示したものである．変換する入力信号 $\mathbf{X}^s$ はカテゴライズ辞書 $\mathbf{\Theta}$ とアクティビティ $\mathbf{H}_{\mathbf{\Theta}}^s$ によって，カテゴライズ辞書の基底とその重みの線形和で表される．

$$\mathbf{X}^s \approx \mathbf{\Theta}\mathbf{H}_{\mathbf{\Theta}}^s \quad s.t. \quad \mathbf{H}_{\mathbf{\Theta}}^s \geq 0 \tag{8}$$
$$\mathbf{X}^s = [\mathbf{x}_1^s, \ldots, \mathbf{x}_L^s] \tag{9}$$
$$\mathbf{H}_{\mathbf{\Theta}}^s = [\mathbf{h}_{\mathbf{\Theta}1}^s, \ldots, \mathbf{h}_{\mathbf{\Theta}L}^s] \tag{10}$$
$$\mathbf{h}_{\mathbf{\Theta}l}^s = [\mathbf{h}_{\boldsymbol{\theta}_1 l}^s, \ldots, \mathbf{h}_{\boldsymbol{\theta}_K l}^s]^\mathsf{T} \tag{11}$$
$$\mathbf{h}_{\boldsymbol{\theta}_k l}^s = [h_{\boldsymbol{\theta}_1 l}^s, \ldots, h_{\boldsymbol{\theta}_M k l}^s]^\mathsf{T} \tag{12}$$

アクティビティは，スパース制約を持つ 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できる．コスト関数を最小化するように求められる．

$$d(\mathbf{x}^s, \mathbf{\Theta}\mathbf{h}_{\boldsymbol{\theta}_k l}^s) + ||\lambda . * \mathbf{h}_{\boldsymbol{\theta}_k l}^s||_1 \quad s.t. \quad \mathbf{h}_{\boldsymbol{\theta}_k l}^s \geq 0 \tag{13}$$

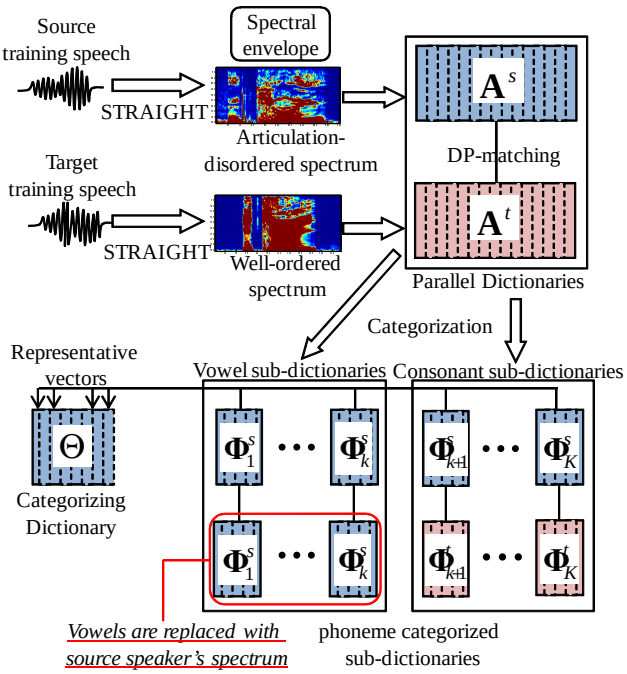


Fig. 3　Making sub dictionary

ここで，$\mathbf{1}$ は全要素が 1 の行列，$.*$ は行列の要素ごとの積である．式 (13) の第 1 項は $\mathbf{x}^s$ と $\mathbf{\Theta h}^s_{\boldsymbol{\theta}_k l}$ の間のカルバック・ライブラー情報量であり，第 2 項は $\mathbf{h}^s_{\boldsymbol{\theta}_k l}$ 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．それぞれの基底に対するスパース制約は $\lambda^{\mathsf{T}} = [\lambda_1 \dots \lambda_J]$ のようにして定められる．本研究では，スパース制約の重み $\lambda$ は 0.1 とした．ここで，$J$ は $\mathbf{\Theta}$ の基底数である．以下の更新式を繰り返し適用することで，式 (13) のコスト関数を最小化できる [6, 8]．

$$\begin{aligned}\mathbf{h}^s_{\boldsymbol{\theta}_k l} \quad \leftarrow \quad & \mathbf{h}^s_{\boldsymbol{\theta}_k l} .* (\mathbf{\Theta}^{\mathsf{T}}(\mathbf{x}^s ./ (\mathbf{\Theta h}^s_{\boldsymbol{\theta}_k l}))) \\ & ./ (\mathbf{\Theta}^{\mathsf{T}} \mathbf{1}^{D\times 1} + \lambda) \end{aligned} \tag{14}$$

ここで，$D$ は入力信号の次元数，$./$ は行列の要素ごとの商である．

つづいて入力信号のフレーム $\mathbf{x}^s_l$ は，対応する入力話者副辞書 $\mathbf{\Phi}^s_{\hat{k}}$ が以下のように選ばれ，その副辞書内の基底の線形和で表現される．

$$\begin{aligned}\hat{k} &= \arg\max_k \mathbf{1}^{1\times M_k} \mathbf{h}^s_{\boldsymbol{\theta}_k l} \\ &= \arg\max_k \sum_{m=1}^{M_k} h^s_{\boldsymbol{\theta}_m l} \end{aligned} \tag{15}$$

$$\mathbf{x}_l = \mathbf{\Phi}^s_{\hat{k}} \mathbf{h}_{\hat{k},l} \tag{16}$$

選ばれた入力話者副辞書 $\mathbf{\Phi}^s_{\hat{k}}$ が子音副辞書の場合，推定されたアクティビティと出力話者副辞書 $\mathbf{\Phi}^t_{\hat{k}}$ により，変換された特徴量は以下のように表現できる．

$$\hat{\mathbf{y}}_l = \mathbf{\Phi}^t_{\hat{k}} \mathbf{h}_{\hat{k},l} \tag{17}$$

アクティビティは式 (14) によって推定される．一方，選ばれた入力話者副辞書 $\mathbf{\Phi}^s_{\hat{k}}$ が母音副辞書の場合は，入力話者副辞書をそのまま用いる．

$$\hat{\mathbf{y}}_l = \mathbf{\Phi}^s_{\hat{k}} \mathbf{h}_{\hat{k},l} \tag{18}$$

Table 1　Sub-dictionary categories

| | Category | phoneme |
|---|---|---|
| vowels | a | a |
| | e | e |
| | i | i |
| | o | o |
| | u | u |
| consonants | plosives | p, t, k, b, d, g |
| | fricatives | s, h, j, z, |
| | nasals | m, n, N |
| | liquid | r |

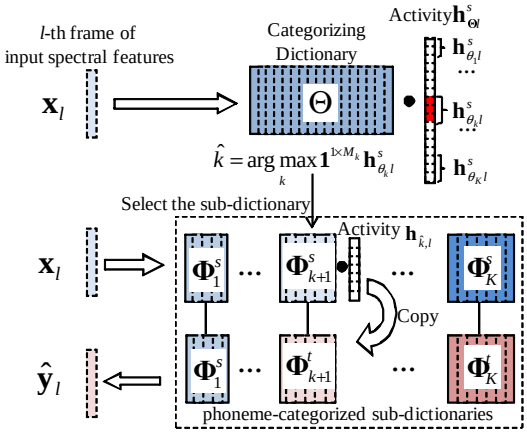


Fig. 4　NMF-based voice conversion using categorized dictionary

## 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 B セット [9] 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 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 [10] 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．

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 (listening intelligibility)，話者性 (similarity)，自然性 (naturalness) の 3 項目とした．「聞き取りやすさ」と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 22 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50 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性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障害者の性質に似ているかを選択する 1 対 1 の対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

「聞き取りやすさ」の主観評価結果を Fig. 5 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．これは，変換ノイズによるものと考えられる．提案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．

「話者性」の主観評価結果を Fig. 5 に示す．提案手法は，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. 7 に示す．提案手法は，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．

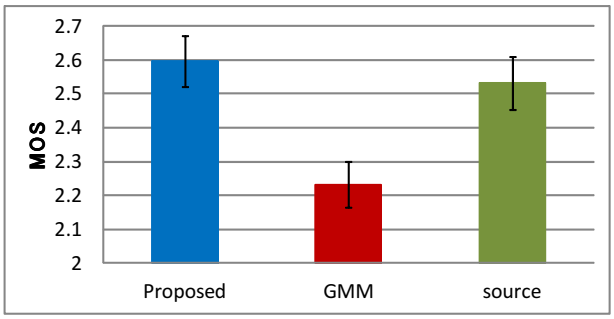


Fig. 5 Results of MOS test on listening intelligibility

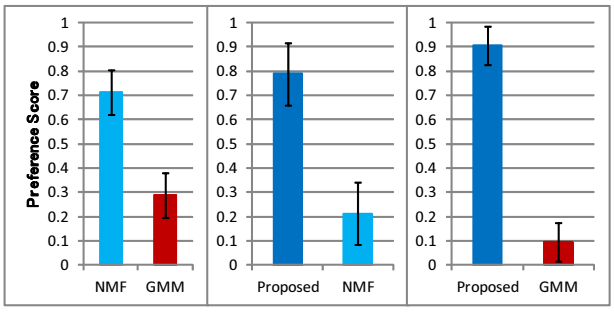


Fig. 6 Preference scores for the individuality

## 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辞書選択を導入し，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することを示した．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．

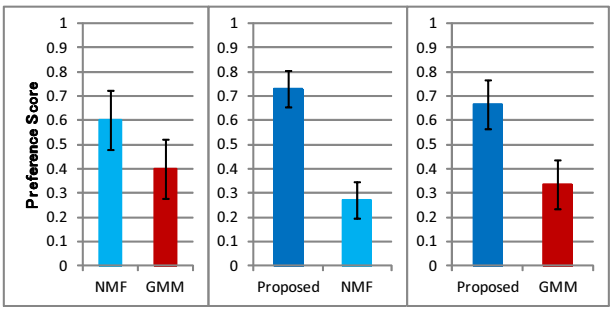


Fig. 7 Preference scores for the naturalness

## 参考文献


[1] S. T. Canale and W. C. Campbell, “Campbell’s operative orthopaedics,” Tech. Rep., Mosby-Year Book, 2002.

[2] Y. Stylianou *et al.*, “Continuous probabilistic transform for voice conversion,” IEEE Trans. Speech and Audio Processing, vol. 6, no. 2, pp. 131–142, 1998.

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

[4] 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.

[5] R. Aihara *et al.*, “Individuality-preserving voice conversion for articulation disorders based on Non-negative Matrix Factorization,” in ICASSP, pp. 8037–8040, 2013.

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

[7] 相原龍 *et al.*, “辞書選択に基づく非負値行列因子分解による声質変換,” 日本音響学会 2013 年秋季研究発表会, pp. 1473–1476, 2013.

[8] J. F. Gemmeke *et al.*, “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.

[9] A. Kurematsu *et al.*, “ATR Japanese speech database as a tool of speech recognition and synthesis,” Speech Communication, vol. 9, pp. 357–363, 1990.

[10] 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.