非負値行列因子分解に基づく構音障害者音声の高域付加の検討 *

☆高島悠樹, 滝口哲也 (神戸大/JST さきがけ), 有木康雄 (神戸大)

1 はじめに

現在，我が国の障害者手帳を持つ 18 歳以上の人口は 350 万人を超えており，聴覚・言語障害者の数は 36 万人とされている [1]．文献 [2] では，構音障害者音声を対象とした音響モデル適応の検証を行っているが，言語障害者などの障害者を対象としている研究は非常に少ない．

言語障害には様々な種類の症状があるが，本研究では，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．アテトーゼ型の脳性麻痺では，意図的な動作を行う際に筋肉の不随意運動が発生するため，発話時に筋肉の緊張が起こり正しく構音できない場合がある．発話が困難な方でも，手話認識や音声合成システムを使用することでコミュニケーションをとることは可能であるが，脳性麻痺患者の多くは手足が不自由であり，音声に頼るしかない状況が考えられる．そのため，構音障害者のための音声による支援ツールには十分なニーズがあり，研究の必要性があるといえる．

構音障害者の発話スタイルは，筋肉の不随意運動により健常者と大きく異なり，安定した構音が難しく，特に子音は非常に不明瞭になる．そのため，周囲の人とのコミュニケーションに支障をきたす．構音障害者音声の特徴として，スペクトルにおける高周波成分のパワーが欠落するという点が挙げられる．Fig. 1 に健常者と構音障害者の発話 “あかちゃん” のスペクトログラムを示す．図に示すように，構音障害者のスペクトログラムの高周波成分は健常者のものに比べて弱くなっている．これが構音障害者音声を不明瞭にし，理解を難しくしている要因の 1 つだと考えられる．この現象は，摩擦音や破裂音などの高周波成分にパワーを持つ音素において特に顕著に現れる．そこで，本研究では，構音障害者音声の明瞭性を向上させるために，健常者スペクトルを用いて構音障害者スペクトルの高周波成分を生成する手法を提案する．

音声信号処理の分野の中でも，声質変換技術が様々なタスク [3] への応用が可能であることから近年盛んに研究されている．声質変換とは，入力話者音声の音韻情報を保存したまま，話者性に関する情報のみを出力話者のものへ変換させる技術である．これまでの声質変換法として，Gaussian mixture model を用いた手法 [4] が最も広く用いられており，様々な改良がなされてきた．その他の手法として，非負値行列因子分解 (non-negative matrix factorization; NMF) [5] や，ニューラルネットワークベースの手法 [6] が提案されてきた．

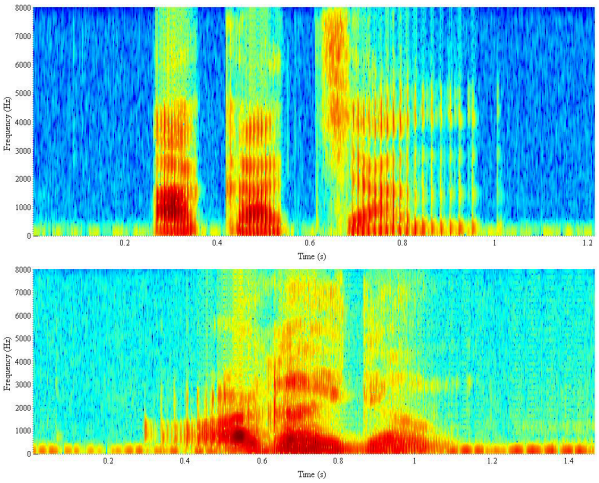


Fig. 1 Example of spectrogram uttered for /a k a ch a n/ of a physically unimpaired person (top) and a person with an articulation disorder (bottom)

NMF [7] は，スパース行列分解手法の 1 つであり，入力信号を，基底行列と係数行列に分解する．NMF の目標は入力行列から，これら 2 つの行列を推定することである．本稿では，基底行列を辞書，係数行列をアクティビティと呼ぶ．相原ら [8] は NMF に基づく声質変換を用いて，構音障害者の話者性を維持したまま音声の明瞭性を向上させる手法を提案した．この方法では，辞書の構築に正確なアライメント情報を必要とし，複雑な処理を必要とする．しかし，構音障害者音声は不明瞭であるため，正確なアライメント情報を用意することは困難である．そこで本稿では，この問題点を解決するためのより単純な NMF 声質変換の枠組みを導入する．さらに，スペクトルの高周波成分を推定するため NMF を改良し，self-reconstructive NMF を提案する．提案モデルは，入力話者の辞書だけでなく，出力話者の辞書も用いてアクティビティを推定する．構音障害者スペクトルには本質的に高周波成分が存在しない．つまり，正解を用意することができないため，入力スペクトルと与えられた学習データから，いかに尤もらしい高周波成分を生成するか，ということが本手法の課題となる．評価実験により，提案手法はより尤もらしいスペクトルの高周波成分

*High-frequency Production Based on Non-negative Matrix Factorization for Articulation Disorders, by Yuki Takashima, Tetsuya Takiguchi (Kobe University/JST PRESTO), Yasuo Ariki (Kobe University)

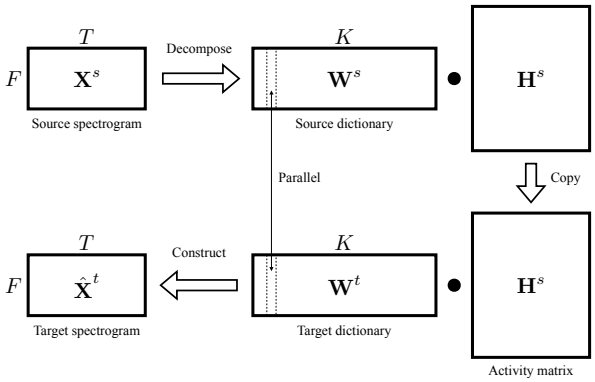


Fig. 2　Basic approach of NMF-based voice conversion

を生成できることを示した．

2　NMF による声質変換

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により表現される．NMF 声質変換では，基底は学習データのスペクトルであり，基底の集合 $\mathbf{W} \in \mathbb{R}^{F\times K}$ を辞書，基底の線形結合重みの集合 $\mathbf{H} \in \mathbb{R}^{K\times T}$ をアクティビティと呼ぶ．このアクティビティがスパースであるとき，観測信号 $\mathbf{X} \in \mathbb{R}^{F\times T}$ 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．

$$\mathbf{X} \approx \mathbf{WH}. \tag{1}$$

ここで，F，K，T はそれぞれ，観測信号の次元数，辞書の基底数，フレーム数を表す．本手法において，$\mathbf{W}$ は学習データで固定され，NMF のアルゴリズムを用いて入力スペクトルから $\mathbf{H}$ を推定する．

本手法の概要を Fig. 2 に示す．$\mathbf{X}^s \in \mathbb{R}^{F\times T}$ は入力話者スペクトル，$\mathbf{W}^s \in \mathbb{R}^{F\times K}$ は入力話者辞書，$\mathbf{W}^t \in \mathbb{R}^{F\times K}$ は出力話者辞書，$\mathbf{H}^s \in \mathbb{R}^{K\times T}$ は入力話者スペクトルから推定されるアクティビティ，$\hat{\mathbf{X}}^t$ は変換されたスペクトルを表す．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 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たものである．NMF のコスト関数は以下のように定義される．

$$d_{KL}(\mathbf{X}^s, \mathbf{W}^s\mathbf{H}^s) + \lambda||\mathbf{H}^s||_1 \quad s.t. \mathbf{H}^s \geq 0. \tag{2}$$

ここで，第 1 項は $\mathbf{X}^s$ と $\mathbf{W}^s\mathbf{H}^s$ の間の Kullback-Leibler ダイバージェンスであり，第 2 項はアクティビティをスパースにするための L1 ノルム正規化項である．λ はスパース重みを示す．辞書は固定し，アクティビティのみを推定する．

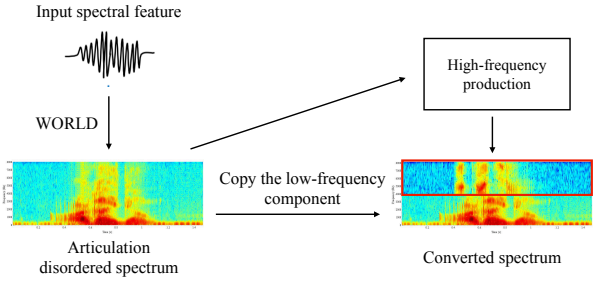


Fig. 3　Flow of our proposed framework

入力スペクトル $\mathbf{X}^s$ は NMF によって $\mathbf{W}^s$ と $\mathbf{H}^s$ の積に分解される．本手法では，「パラレル辞書で推定したパラレルな発話のアクティビティは置き換え可能である」と仮定している．従って，変換スペクトルは，$\mathbf{W}^t$ と推定した $\mathbf{H}^s$ の積によって得られる．

3　スペクトルの高周波成分生成

3.1　概要

Fig. 3 に提案する高周波成分生成法の概要を示す．まず，入力話者スペクトルから分析合成ツールの WORLD [9] を用いてスペクトル特徴量を抽出する．次に，入力話者辞書と出力話者辞書を用いて入力話者スペクトルに対応する高周波成分を生成する．入力話者の話者性を維持するため，低周波成分は入力話者スペクトルを使用し，高周波成分のみを生成したスペクトルで置換する．本研究では，高周波成分の生成方法として，従来の NMF と辞書適応型 NMF に基づく手法，そして提案モデルを用いた手法を用いる．

3.2　従来の NMF を用いた手法

与えられた構音障害者の入力スペクトルは，従来の NMF 声質変換の枠組みにより健常者スペクトルへ変換される．変換スペクトルは健常者の辞書基底を用いて構成されるため，高周波成分にパワーを持つと考えられる．しかし，文献 [10] において，入力スペクトルと辞書がパラレルであっても，推定されるアクティビティは話者毎に異なることが示唆されている．従って，このアクティビティの不一致により，適切な変換スペクトルが得られない可能性が考えられる．

3.3　辞書適応型 NMF を用いた手法

出力話者辞書を用いた入力話者スペクトルの表現について検討する．話者適応を用いた exemplar-based 声質変換が提案されている [11]．この方法では，少量のパラレルデータを用いて，NMF の枠組みにより入

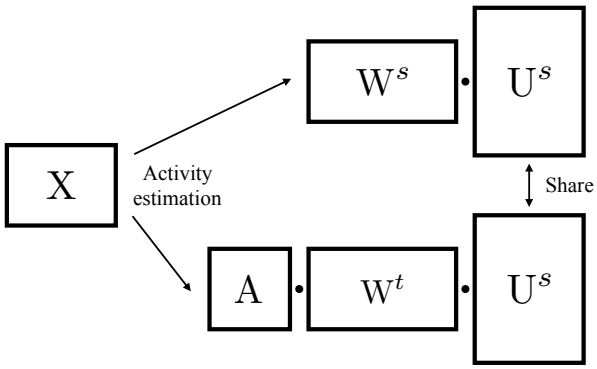


Fig. 4　Self-reconstructive NMF

力話者スペクトルから出力話者スペクトルへの線形変換を学習する．本稿では，これを高周波成分生成へ応用する．まず，入力話者辞書 $\mathbf{W}^s$ と出力話者辞書 $\mathbf{W}^t$ を用いて，下記のコスト関数により出力話者スペクトルから入力話者スペクトルへの線形変換を学習する．

$$d_{KL}(\mathbf{W}^s, \mathbf{A}\mathbf{W}^t) \quad s.t. \mathbf{A} \geq 0, \tag{3}$$

ここで，$\mathbf{A} \in \mathbb{R}^{F \times F}$ は適応行列を表す．適応行列のみを推定し，その他のパラメータは固定する．変換時には，与えられた入力スペクトルに対し，以下のコスト関数によりアクティビティ $\mathbf{U}^s \in \mathbb{R}^{K \times T}$ を推定する．

$$d_{KL}(\mathbf{X}^s, \tilde{\mathbf{W}}^t \mathbf{U}^s) \quad s.t. \mathbf{U}^s \geq 0, \tag{4}$$

ここで，$\tilde{\mathbf{W}}^t = \mathbf{A}\mathbf{W}^t$ は適応された出力話者辞書を表す．適応された出力話者辞書は固定し，アクティビティのみを推定する．そして，推定されたアクティビティを用いて以下の式により変換スペクトル $\hat{\mathbf{X}}^t$ を構成する．

$$\hat{\mathbf{X}}^t \triangleq \mathbf{W}^t \mathbf{U}^s. \tag{5}$$

この手法は，出力話者スペクトル $\mathbf{W}^t\mathbf{U}^s$ を適応行列 $\mathbf{A}$ を用いて入力話者スペクトルへ変換しているとみなすことができる．従って，アクティビティを推定し，適応行列を用いないで再構成することで，出力話者の話者性を持つ変換スペクトル $\hat{\mathbf{X}}^t$ を得る．

3.4　Self-reconstructive NMF を用いた手法

辞書適応型 NMF において，式 (4) に示すように，アクティビティは適応された出力話者辞書を用いて推定される．しかしながら，変換スペクトルは式 (5) に示すように，適応されていない出力話者辞書を用いて計算される．このミスマッチは，変換スペクトルに対して悪影響を及ぼす可能性が考えられる．そこで，この問題を解決するために Fig. 4 に示す self-reconstructive NMF を提案する．

Self-reconstructive NMF のコスト関数は下記のように定義される．

$$d_{KL}(\mathbf{X}, \mathbf{W}^s\mathbf{U}^s) + d_{KL}(\mathbf{X}, \tilde{\mathbf{W}}^t\mathbf{U}^s) \quad s.t. \mathbf{U}^s \geq 0, \tag{6}$$

ここで，アクティビティ $\mathbf{U}^s$ のみを推定し，その他のパラメータは固定する．NMF 声質変換では，辞書とアクティビティによりスペクトルが構成されると仮定する．式 (6) は，入力話者スペクトル $\mathbf{W}^s\mathbf{U}^s$ と出力話者スペクトル $\mathbf{W}^t\mathbf{U}^s$ を含んでおり，これらは同一のアクティビティ $\mathbf{U}^s$ を用いたスペクトル変換を表現する．式 (6) 第 1 項は，パラレル辞書に推定されたアクティビティを掛けることでスペクトルが得られることを暗に保証している．式 (6) 第 2 項は，変換された出力話者スペクトルが，適応行列を用いることで入力話者スペクトルに対して一貫性が取れるようにアクティビティを推定することを保証する．従って，パラレル辞書は互いに対応関係にあるため，式 (5) により出力話者辞書と推定されたアクティビティを掛け合わせることでより尤もらしい変換スペクトルを得ることができる．

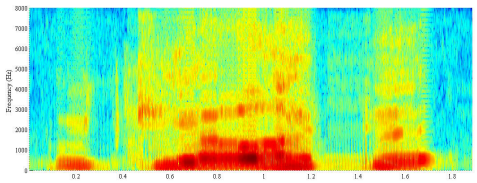
4　評価実験

4.1　実験条件

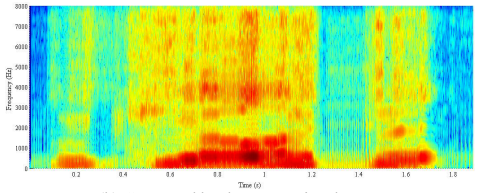
構音障害者男性 1 名と健常者男性 1 名の音声を使用した．構音障害者音声は，ATR 研究用日本語音声データベース [12] に含まれる 50 文と 10 単語を収録した．健常者音声は，ATR 研究用日本語音声データベースから 1 話者を選択し，構音障害者音声と同一内容の発話を使用した．50 文を学習データ (辞書構成用)，10 単語を評価に用いた．サンプリング周波数は 16kHz である．分析合成ツールの WORLD [9] によって得られたスペクトル 513 次元を前後 2 フレーム分を束ねた 2,565 次元を入力特徴量とした．パラレル辞書の基底数は 94,382 である．F0 と非周期成分は入力発話のものを用いた．本稿では，入力スペクトルの 211 次元より高域を置換した．

4.2　実験結果

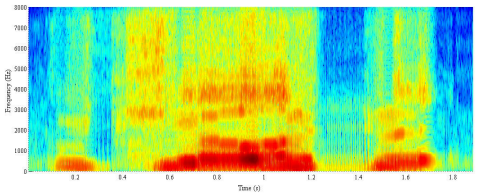
Fig. 5 は入力スペクトルと各手法により変換したスペクトルを示す．Figs. 5(a)，5(e) に示すように，構音障害者の高周波成分は，健常者のものと比べて弱くなっている．Fig. 5(b) は従来の NMF 声質変換に基づく手法を用いた変換結果である．高周波成分が全体的にぼやけていることが分かる．これは不正確なアクティビティを用いているためだと考えられる．Figs. 5(c)，5(d) は辞書適応型 NMF と self-reconstructive NMF を用いた手法による変換スペクトルを示す．Fig. 5(e)



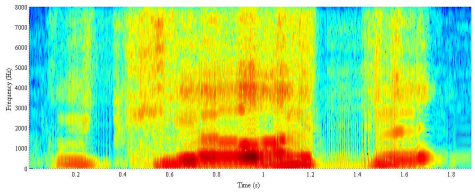
(a) Spoken by a person with an articulation disorder



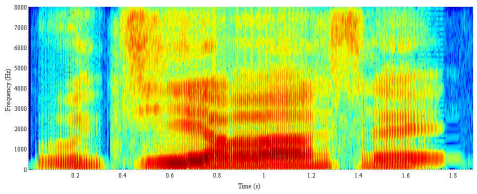
(b) Converted by the conventional NMF



(c) Converted by adaptation NMF



(d) Converted by self-reconstructive NMF



(e) Spoken by a physically unimpaired person

Fig. 5　Example spectra of spectrogram /u ch i a w a s e/

と比べて，尤もらしい高周波成分が生成できていることが分かる．この理由として，self-reconstructive NMF はパラレル辞書を掛けることでスペクトルに変換できることを保証しながらアクティビティを推定していることが考えられる．

明瞭性について，mean opinion score による主観評価実験を行なったが，有効性は確認できなかった．この理由として，NMF やボコーダによる分解/分析・再合成による誤差の影響が考えられる．

5　おわりに

構音障害者音声のための NMF に基づく高周波成分生成法として self-reconstructive NMF を提案し，変換スペクトルにおいてその有効性を確認した．今後は，より明瞭度の高い音声の生成を検討する．

謝辞　本研究の一部は，JSPS 科研費 JP17J04380，JST さきがけ JPMJPR15D2 の支援を受けたものである．

参考文献

[1] 内閣省，“平成 25 年版障害者白書，”．
[2] 中村圭吾 *et al.*，“発話障害者音声を対象にした健常者音響モデルの適応と検証，”日本音響学会講演論文集, pp. 109–110, 2015.
[3] A. Kain and M. W. Macon, “Spectral voice conversion for text-to-speech synthesis,” in *ICASSP*, 1998, pp. 285–288.
[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] R. Takashima *et al.*, “Exemplar-based voice conversion in noisy environment,” IEEE Workshop on Spoken Language Technology, pp. 313–317, 2012.
[6] S. Desai *et al.*, “Voice conversion using artificial neural networks,” in *ICASSP*, 2009, pp. 3893–3896.
[7] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” NIPS, pp. 556–562, 2000.
[8] R. Aihara *et al.*, “Individuality-preserving voice conversion for articulation disorders using phoneme-categorized exemplars,” TACCESS, vol. 6, no. 4, pp. 13, 2015.
[9] M. Morise *et al.*, “World: A vocoder-based high-quality speech synthesis system for real-time applications.,” IEICE Transactions, vol. 99-D, no. 7, pp. 1877–1884, 2016.
[10] R. Aihara *et al.*, “Parallel dictionary learning for voice conversion using discriminative graph-embedded non-negative matrix factorization,” INTERSPEECH, pp. 292–296, 2016.
[11] R. Aihara *et al.*, “Small-parallel exemplar-based voice conversion in noisy environments using affine non-negative matrix factorization,” EURASIP J. Audio, Speech and Music Processing, vol. 2015, pp. 32, 2015.
[12] A. Kurematsu *et al.*, “ATR Japanese speech database as a tool of speech recognition and synthesis,” Speech Communication, vol. 9, no. 4, pp. 357–363, 1990.