Multiple Non-negative Matrix Factorization に基づく多対一声質変換*

☆相原龍, 滝口哲也, 有木康雄 (神戸大)

1 はじめに

本研究では，非負値行列因子分解 (Non-negative Matrix Factorization : NMF) [1] を用いた Exemplar-based 声質変換の枠組みにおいて，入力話者音声の学習を必要としない多対一声質変換を提案する．

従来，声質変換においては統計的な手法が多く提案されてきた．なかでも混合正規分布モデル (Gaussian Mixture Model : GMM) を用いた手法 [2] はその精度のよさと汎用性から広く用いられており，多くの改良がされ続けられている．基本的には，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．

声質変換の基本的な枠組みは，入力話者と出力話者が同じテキストを発話して得られるパラレルデータを用いて，データ間のマッピング関数を求めるものであった．そのため，これまでの声質変換においては入力話者と出力話者の大量の同一発話を必要とするという制約があった．統計的声質変換においては，所望の話者間のマッピング関数を柔軟に構築するために他の話者の発話データを用いる手法がいくつか提案されている．戸田ら [3] は固有声に基づく声質変換 (Eigenvoice conversion: EVC) を提案し，特定話者の声質から任意話者の声質へと変換する一対多声質変換，あるいはその逆である多対一声質変換を実現した．大谷ら [4] は参照話者を用いた多対多 EVC を提案した．齋藤ら [5] は GMM でモデル化した音声に対してテンソル表現を用い，より柔軟な一対多声質変換を提案している．以上のように，統計的声質変換においては任意話者への変換が可能となりつつある．

我々はこれまで，従来の統計的手法とは異なる，NMF に基づく Exemplar-based 声質変換手法を提案してきた [6]．我々の提案している NMF 声質変換では，従来の声質変換手法でも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（入力話者辞書）と出力話者の音声辞書（出力話者辞書）からなる同一発話内容のパラレル辞書を構築する．変換時には，入力音声を NMF によって，入力辞書に含まれる少量の基底からなるスパース表現にする．得られた入力辞書の基底毎の重み係数（アクティビティ）に基づいて，入力話者辞書の基底を出力辞書内の基底と置き換え，線形結合することで，出力話者の音声へと変換する．従来の声質変換のように統計的モデルを用いない Exemplar-based 手法であるため，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．

さらに，NMF 声質変換は NMF に備わる雑音抑圧法を組み込み，入力辞書に雑音辞書を結合してアクティビティを求めることで，雑音抑圧と声質変換が同時に可能である [6]．また，NMF はクラスタリング手法でもあり，入力辞書行列の音素ラベルが既知であれば，音素識別が可能である．この性質を利用して我々は，発話が不明瞭になりやすい脳性麻痺による構音障害者を対象とした声質変換を提案した [7]．以上のように，NMF を用いた声質変換手法は従来手法にはなかった様々なタスクに応用可能な手法である．

しかしながら，NMF 声質変換もまた入力話者と出力話者のパラレルな発話データを必要とするため，声質変換の実用化面で大きな制約となっていた．本研究では，NMF 声質変換において任意話者からの変換を実現するため，Multiple Non-negative Matrix Factorization (Multi-NMF) を用いた多対一声質変換を提案する．これまで必要とされてきた入力話者辞書の代わりに，複数話者の辞書の線形和で近似した辞書を用いることで，任意の話者からの声質変換を可能にした．本手法は多対一声質変換のみならず，多対多声質変換や話者性を制御可能な声質変換へと応用可能であると考えられる．

以下，第 2 章でこれまでの NMF による一対一声質変換手法を述べ，第 3 章で本稿の提案手法を説明する．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．

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

$\mathbf{v}_l$ 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．

$$\mathbf{V} \approx \mathbf{W}\mathbf{H} \quad (2)$$

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

ここで L はフレーム数を表す．

---

*Many-to-one Voice Conversion Based on Multiple Non-negative Matrix Factorization by Ryo AIHARA, Tetsuya TAKIGUCHI, Yasuo ARIKI (Kobe University)

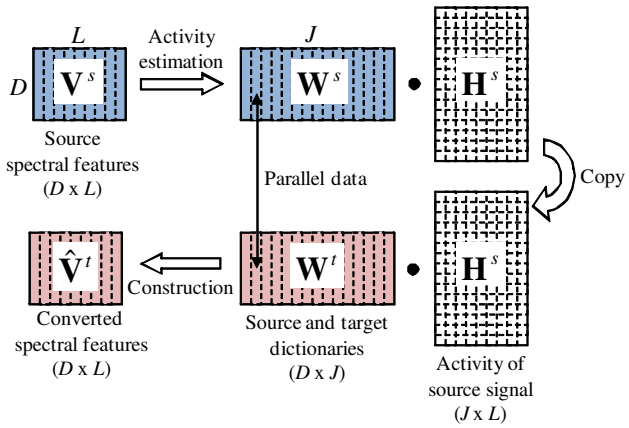

Fig. 1　Basic approach of NMF-based VC

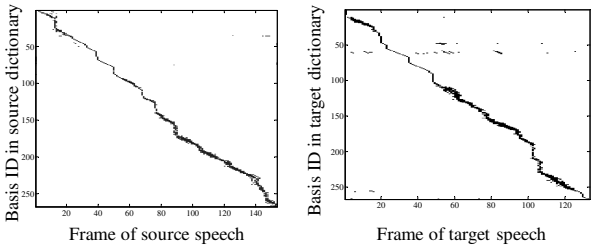

Fig. 2　Activity matrices for parallel utterances

本手法の概要を Fig. 1 に示す．$\mathbf{V}^s$ は入力話者スペクトル，$\mathbf{W}^s$ は入力話者辞書，$\mathbf{W}^t$ は出力話者辞書，$\hat{\mathbf{V}}^t$ は変換されたスペクトル，$\mathbf{H}^s$ は入力話者スペクトルから推定されるアクティビティを表す．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，「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の形状は類似する」と仮定している．Fig. 2 に男性 1 名，女性 1 名がそれぞれ発話した “ikioi” のスペクトルから，それぞれの話者のパラレル辞書を用いて推定されたアクティビティを示す．なお，簡単な例を示すために，辞書行列は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 より，アクティビティの形状は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．

2.2　NMF

本手法では，アクティビティ行列の推定に NMF を用いる [8]．NMF のコスト関数は，$\mathbf{V}^s$，$\mathbf{W}^s$，$\mathbf{H}^s$ を用いて以下のような式で表せる．

$$d(\mathbf{V}^s, \mathbf{W}^s\mathbf{H}^s) + \lambda||\mathbf{H}^s||_1 \tag{4}$$

ここで，第 1 項は $\mathbf{V}^s$ と $\mathbf{W}^s\mathbf{H}^s$ の間の Kullback-Leibler (KL) 距離であり，第 2 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．λ はスパース重みを表す．このコスト関数は Jensen の不等式を用いることで，繰り返し適用を用いて最小化できる．コスト関数を最小化するアクティビティは以下の更新式で求められる．

$$\begin{aligned}\mathbf{H}^s \quad \leftarrow \quad & \mathbf{H}^s .* (\mathbf{W}^{s\mathsf{T}}(\mathbf{V}^s ./ (\mathbf{W}^s\mathbf{H}^s))) \\ & ./ (\mathbf{W}^{s\mathsf{T}}\mathbf{1}^{D\times L} + \lambda\mathbf{1}^{1\times L})\end{aligned} \tag{5}$$

変換音声 $\hat{\mathbf{V}}^t$ は出力話者辞書行列と推定されたアクティビティの内積をとることで得られる．

$$\hat{\mathbf{V}}^t = \mathbf{W}^t\mathbf{H}^s \tag{6}$$

3　Multi-NMF を用いた多対一声質変換

3.1　概要

本手法は，第 2 章で用いたスパース表現の考え方を拡張したものであり，以下の 2 つの仮定に基づく．

1. 任意話者の発話スペクトルは，複数話者の発話スペクトルから成る，少量の基底の線形和で表現できる．
2. 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の形状は類似する．

Fig. 3 に本手法の概要を示す．$\mathbf{V}^s$ は入力話者スペクトル，$\hat{\mathbf{V}}^t$ は変換されたスペクトル，$\mathbf{H}^s$ は入力話者スペクトルから推定されるアクティビティを表す．アクティビティ推定に用いる K 人の話者スペクトルからなる辞書行列を $\mathbf{W}^M \in \mathbb{R}^{(D\times J\times K)}$ とし，k 人目の話者スペクトル辞書行列を $\mathbf{W}_k^M \in \mathbb{R}^{(D\times J)}$ で表す．K 人には入力話者は含まれない．$\mathbf{W}^t$ は出力話者辞書行列を表し，これら全ての辞書行列は複数の同一発話内容に対して DTW を適用したパラレルデータである．$\mathbf{a} \in \mathbb{R}^{(1\times 1\times K)}$ は話者重みベクトルとする．

入力話者スペクトル $\mathbf{V}^s$ は仮定（1）に基づいて以下のように表される．

$$\mathbf{V}^s \approx \quad \left(\sum_{k=1}^{K} a_k \mathbf{W}_k^M\right)\mathbf{H}^s \tag{7}$$

ここで，アクティビティ $\mathbf{H}^s$ が辞書行列に対して共通であることに注意されたい．$\mathbf{W}^s \approx \sum_{k=1}^{K} a_k\mathbf{W}_k^M$ とおけば，従来の一対一声質変換で必要であった入力話者辞書行列 $\mathbf{W}^s$ は，話者重みベクトル $\mathbf{a}$ によって複数話者辞書行列の線形和で近似されるとみることが

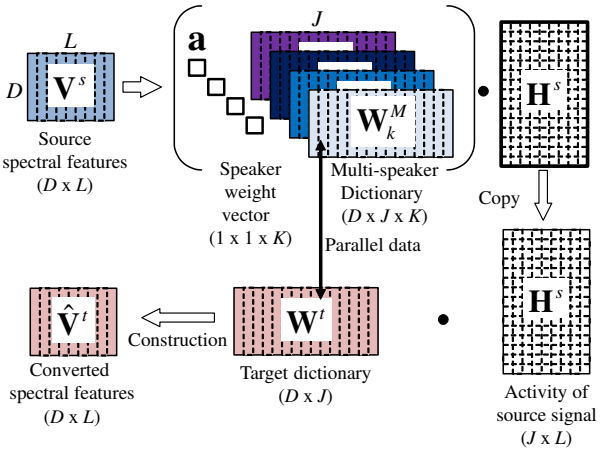

Fig. 3　Flow chart of Many-to-one VC using Multi-NMF

できる．アクティビティ$\mathbf{H}^s$が音韻性を表すと仮定すると，話者重みベクトル$\mathbf{a}$は話者性を表すと考えることができ，Multi-NMF は行列表現に基づいて，音声からの音韻性・話者性の分離を行っているといえる．本手法では，辞書を固定し，入力音声から音韻性と話者性の推定を同時に行うため，事前に入力話者の音声を学習しなくても変換が可能である．

変換スペクトルは仮定（2）に基づいて，出力話者辞書と，入力話者スペクトルから推定されたアクティビティの内積によって得られる．

3.2　Multi-NMF

複数話者辞書行列を用いて，話者重みベクトルとアクティビティ行列を推定する手法として Multi-NMF を提案する．Multi-NMF のコスト関数は$\mathbf{V}^s$, $\mathbf{a}$, $\mathbf{W}^M$，$\mathbf{H}^s$を用いて以下のような式で表せる．

$$d(\mathbf{V}^s, \sum_{k=1}^{K} a_k \mathbf{W}_k^M \mathbf{H}^s) + \lambda ||\mathbf{H}^s||_1 \tag{8}$$

第 1 項は$\mathbf{V}^s$と$\sum_{k=1}^{K} a_k \mathbf{W}_k^M \mathbf{H}^s$の間の KL 距離であり，第 2 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．a_kは$\mathbf{a}$のk番目の要素を表す．

Jensen の不等式を用いて，コスト関数を最小化する$\mathbf{a}$と$\mathbf{H}^s$を補助関数法で求める．更新式は下記のようになる．

$$a_k \leftarrow \frac{a_k}{\sum_{d,l}(\mathbf{WH})_{dl}} \sum_{d,l} \left(\frac{v_{dl}^s (\mathbf{W}_k^M \mathbf{H})_{dl}}{\sum_k a_k (\mathbf{W}_k^M \mathbf{H})_{dl}} \right) \tag{9}$$

$$\mathbf{H}^s \leftarrow \mathbf{H}^s .* ((\sum_{k=1}^{K} a_k \mathbf{W}_k^M)^{\mathsf{T}} (\mathbf{V}^s ./ (\sum_{k=1}^{K} a_k \mathbf{W}_k^M \mathbf{H}^s))) \\ ./((\sum_{k=1}^{K} a_k \mathbf{W}_k^M)^{\mathsf{T}} \mathbf{1}^{D \times L} + \lambda \mathbf{1}^{1 \times L}) \tag{10}$$

v_{dl}^sは$\mathbf{V}^s$の要素を表す．

推定されたアクティビティによって，変換スペクトルは以下のように得られる．

$$\hat{\mathbf{V}}^t = \mathbf{W}^t \mathbf{H}^s \tag{11}$$

4　評価実験

4.1　実験条件

本実験では提案手法の有効性を示すため，従来のパラレルデータを用いた一対一 NMF 声質変換，一対一 GMM 声質変換と比較した．ATR 研究用日本語音声データベース [9] より，男性話者 6 名，女性話者 1 名の音声を用いた．提案手法においては，男性 6 名から入力話者 1 名を選び，残り 5 名で複数話者辞書行列を構築し，女性 1 名を出力話者とした．それぞれの辞書はパラレルな音素バランス 50 文から構成される．いずれの手法もサンプリング周波数は 12kHz である．

一対一 NMF 及び提案手法では，STRAIGHT スペクトル [10] と前後 2 フレームを含む 2,565 次元特徴量とした．それぞれの手法において NMF の更新回数は 400 とした．一対一 NMF の入力・出力辞書行列は，それぞれの話者のパラレルな 50 文から構成される．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 60 次元を特徴量とし，入力話者，出力話者のパラレルな 50 文で GMM を学習した．GMM の混合数は 112 である．本稿では，非周期成分は変換せず入力音声のものをそのまま用いている．F0 については，スペクトル変換における提案手法の有効性を示すため，提案手法においても従来手法と同様のパラレルデータを用いた単回帰分析によって変換している．

提案手法の有効性を確かめるため，客観評価と主観評価を行った．客観評価はメルケプストラム 24 次元を特徴量とし，式 (12) で表されるメルケプストラム歪 (Melcepstrum distortion : MelCD) [dB] によって各手法を比較した．

$$MelCD = (10/\log 10) \sqrt{2 \sum_{d}^{24} (mc_d^{conv} - mc_d^{tar})^2} \tag{12}$$

ここで，mc_d^{conv}，mc_d^{tar} は d 次元目の変換後のケプストラム，目標音声のケプストラムを表す．いずれの手法も 3 名の入力話者について，合計 75 文を評価した．

主観評価は成人男女 10 名に対して，音質と話者性の 2 項目について聴取実験を行った．音質の評価基準は 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．話者性の評価では，はじめに目標話者音声を聴かせた後，異なる手法によって変換した音声を試聴しよい方を選ぶ XAB テストを行った．いずれの評価項目も 3 名の入力話者について，合計 36 文を静かな部屋においてヘッドホンを用いた両耳聴取で評価した．

4.2 実験結果・考察

Fig. 4 左側に客観評価によるメルケプストラム歪を示す．Source は入力音声と目標音声間の歪，Multi は多対一声質変換における提案手法，NMF と GMM は一対一声質変換におけるそれぞれの手法による歪を表す．Fig. 4 左側より，従来のパラレルデータを用いた一対一声質変換においては，NMF と GMM はほぼ同程度の変換精度であることがわかる．提案手法は，これら 2 つの手法と比較して若干歪が大きくなっている程度であり，入力話者の発話スペクトルを学習していないにも関わらずほぼ同程度の変換が可能となっている．

Fig. 4 右側に音質における MOS 評価値を示す．誤差範囲は 95%信頼区間を示す．提案手法と一対一 NMF の間には有意な差は見られないが，これら 2 つの手法は一対一 GMM を上回っている．

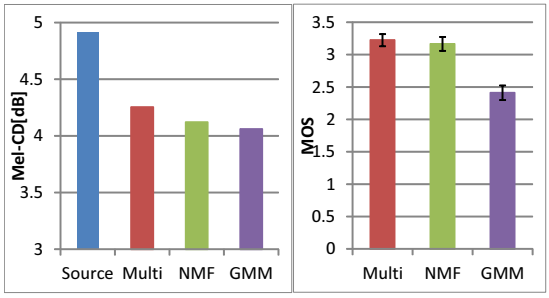

Fig. 4　Left : MelCD calculated from source speech and converted speech using each method. Right : MOS of speech quality.

Fig. 5 に話者性における XAB 評価結果を示す．左側は提案手法と一対一 NMF の間の比較であり，2 つの手法の間に有意差はない．右側は提案手法と一対一 GMM の間の比較を示しており，提案手法が一対一 GMM を上回っていることがわかる．

以上の結果より，提案手法は主観評価において一対一 GMM の精度を上回り，一対一 NMF とほぼ同等の精度が得られることが示された．

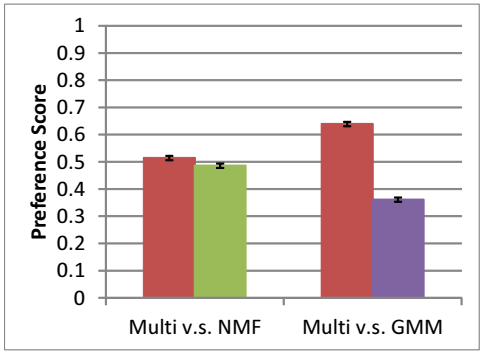

Fig. 5　XAB test between our proposed method and conventional methods

5　おわりに

本報告では，NMF を用いた Exemplar-based 声質変換の枠組みにおいて，入力話者の発話スペクトルを学習せずに変換が可能な多対一声質変換を提案した．従来の NMF を Multi-NMF へと拡張し，複数話者辞書行列と話者重みベクトルを導入することで，入力話者スペクトルを複数の話者スペクトルの線形結合で表すことを可能にした．客観評価実験・主観評価実験で，提案手法は従来のパラレルデータを用いた一対一 NMF 声質変換とほぼ同程度の精度で変換が可能であることを示した．今後は，EVC など，多対一声質変換を対象とした手法と本手法を比較する予定である．

参考文献

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

[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] T. Toda *et al.*, “Eigenvoice conversion based on Gaussian mixture model,” in Interspeech, pp. 2446–2449, 2006.

[4] Y. Ohtani *et al.*, “Many-to-many eigenvoice conversion with reference voice,” in Interspeech, pp. 1623–1626, 2009.

[5] D. Saito *et al.*, “One-to-many voice conversion based on tensor representation of speaker space,” in Interspeech, pp. 653–656, 2011.

[6] R. Takashima *et al.*, “Exemplar-based voice conversion in noisy environment,” in SLT, pp. 313–317, 2012.

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