Multiple Non-negative Matrix Factorization に基づく多対多声質変換*

◎相原龍, 滝口哲也, 有木康雄 (神戸大)

1 はじめに

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 [1] 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，あるいは最尤基準で推定する．

我々はこれまで，従来の統計的手法とは異なる，非負値行列因子分解 (Non-negative Matrix Factorization: NMF) に基づく Exemplar-based 声質変換手法を提案してきた [2]．NMF 声質変換では，GMM 声質変換手法で用いられていたパラレルデータを Exemplar（基底）として用いる．入力発話は基底の線形結合へと分解され，入力話者の基底を対応する出力話者の基底と置き換えることで出力話者の音声へと変換する．この手法は統計的モデルを用いない手法であるため，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．またこの手法は，ノイズロバスト声質変換 [2] や構音障害者を対象とした声質変換 [3] など統計的手法にはなかった様々なタスクに応用されてきた．

いずれの手法においても声質変換の基本的なアプローチは，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，データ間のマッピング関数を求める，ということに変わりはない．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，この制約を緩和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 [4] 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．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．能勢ら [6] はニューラルネットワークによる統計的な多対一声質変換を提案した．大谷ら [7] は EVC による多対一声質変換と一対多声質変換を参照話者で結びつけることで，多対多 EVC を提案した．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．

NMF を用いた Exemplar-based 声質変換では，これまで多対多声質変換は実現されてこなかった．本研究では，NMF 声質変換において任意話者間での変換を実現するため，Multiple Non-negative Matrix Factorization (Multi-NMF) を用いた多対多声質変換を提案する．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 による一対一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．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 $\mathbf{W}$ を “辞書”，基底の線形結合重みの集合 $\mathbf{H}$ 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 $\mathbf{V}$ 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．

$$\mathbf{V} \approx \mathbf{WH} \tag{1}$$

$$\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_L], \quad \mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_L]. \tag{2}$$

ここで L はフレーム数を表す．本手法において，$\mathbf{W}$ は学習データで固定され，NMF [8] 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. 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．D，J はそれぞれスペクトルの次元数，辞書の基底数である．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．

入力スペクトル $\mathbf{V}^s$ 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 $\hat{\mathbf{V}}^t$ は，$\mathbf{W}^t$ と推定した $\mathbf{H}^s$ の積によって得られる．

3 Multi-NMF を用いた多対一声質変換

3.1 概要

Fig. 2 に提案手法の概要を示す．$\mathbf{V}^s$ は変換前の入力話者スペクトル，$\mathbf{V}^t$ は適応データである変換話者スペクトル，$\hat{\mathbf{V}}^s$ は変換後のスペクトル，$\mathbf{a}^s$ は入力話者重みベクトル，$\mathbf{a}^t$ 出力話者重みベクトル，$\mathbf{H}^s$ は入

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

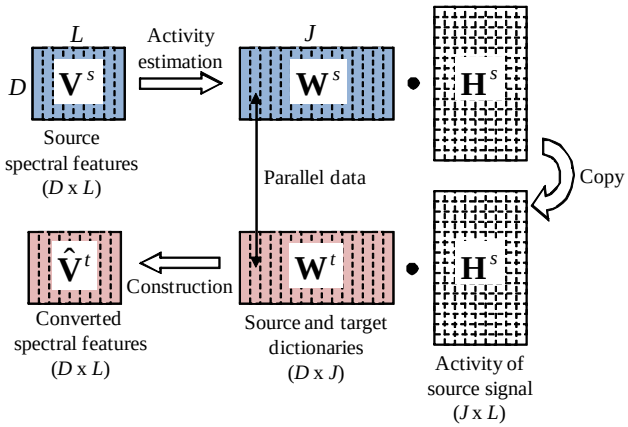

Fig. 1　Basic approach of NMF-based VC

力話者スペクトルのアクティビティ，$\mathbf{H}^t$ は変換話者スペクトルのアクティビティを表す．さらに，D, L, L', J はそれぞれスペクトルの次元数，入力話者スペクトルのフレーム数，出力話者スペクトルのフレーム数，辞書行列のフレーム数を表す．アクティビティ推定に用いる K 人のパラレルな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・出力話者は含まれない．

まず入力話者スペクトル $\mathbf{V}^s$ は，辞書行列 $\mathbf{W}^M$，入力話者重みベクトル $\mathbf{a}^s$，アクティビティ $\mathbf{H}^s$ の 3 要素に分解される．

$$\mathbf{V}^s \approx \left(\sum_{k=1}^{K} a_k^s \mathbf{W}_k^M\right)\mathbf{H}^s \tag{3}$$

a_k^s は $\mathbf{a}^s$ の k 番目の要素を表す．ここで，アクティビティ $\mathbf{H}^s$ が K 人の辞書行列に対して共通であることに注意されたい．

続いて，適応データである変換話者スペクトル $\mathbf{V}^t$ を用いて，出力話者重みベクトル $\mathbf{a}^t$，アクティビティ $\mathbf{H}^t$ を推定する．

$$\mathbf{V}^t \approx \left(\sum_{k=1}^{K} a_k^t \mathbf{W}_k^M\right)\mathbf{H}^t \tag{4}$$

最後に，変換スペクトル $\hat{\mathbf{V}}^s$ は推定された出力話者重みベクトルと入力発話のアクティビティで以下のように求められる．

$$\hat{\mathbf{V}}^t = \left(\sum_{k=1}^{K} a_k^t \mathbf{W}_k^M\right)\mathbf{H}^s \tag{5}$$

ここで，$\mathbf{W}^s \approx \sum_{k=1}^{K} a_k^s \mathbf{W}_k^M$ 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 $\mathbf{a}^s$ 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 $\mathbf{a}^s$ 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．

本手法においては，辞書行列は単一の性別の話者スペクトルから構成される．ここで，$\mathbf{W}_k^{M_s}$ と $\mathbf{W}_k^{M_t}$ をそれぞれ入力話者，出力話者と同一の性別の話者スペクトルで構成される辞書とすると，異性間の変換においては式 (3) の $\mathbf{W}_k^M$ は $\mathbf{W}_k^{M_s}$ で，式 (4) と式 (5) の $\mathbf{W}_k^M$ は $\mathbf{W}_k^{M_t}$ で置き換えられる．

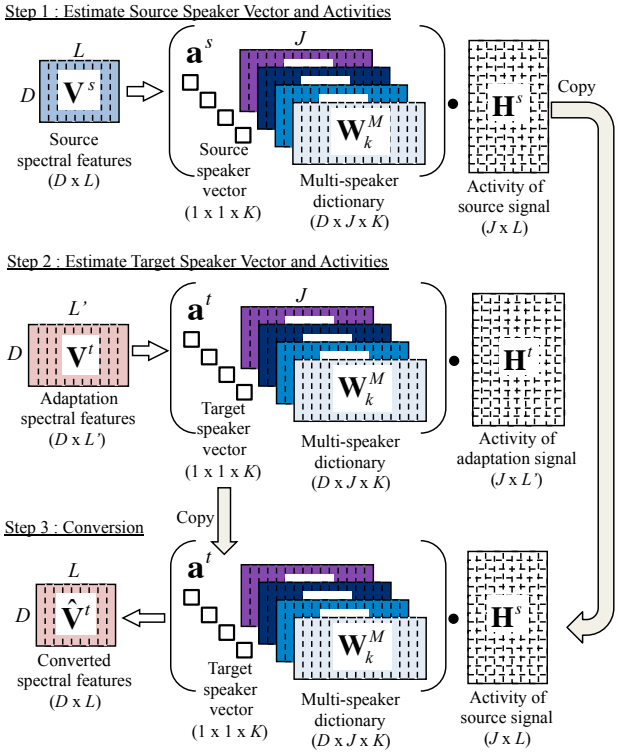

Fig. 2　Many-to-many VC using Multi-NMF

3.2　Multi-NMF

Multi-NMF は入力スペクトル $\mathbf{V} \in \mathbb{R}^{(D\times L)}$ と与えられた辞書行列 $\mathbf{W}^M \in \mathbb{R}^{(D\times J\times K)}$ から話者重みベクトル $\mathbf{a} \in \mathbb{R}^{(1\times 1\times K)}$ とアクティビティ行列 $\mathbf{H} \in \mathbb{R}^{(J\times L)}$ を推定する．Multi-NMF のコスト関数は下記のように定義できる．

$$d(\mathbf{V}, \sum_{k=1}^{K} a_k \mathbf{W}_k^M \mathbf{H}) + \lambda ||\mathbf{H}||_1 \tag{6}$$

ここで第 1 項は $\mathbf{V}$ と $\sum_{k=1}^{K} a_k \mathbf{W}_k^M \mathbf{H}$ の間の Kullback-Leibler(KL) ダイバージェンスであり，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．

Jensen の不等式を用いて，コスト関数を最小化する $\mathbf{a}$ と $\mathbf{H}$ の更新式は，補助関数法で下記のように求められる．

$$a_k \leftarrow \frac{a_k}{\sum_{d,l}(\mathbf{W}_k^M\mathbf{H})_{dl}} \sum_{d,l}\left(\frac{v_{dl}(\mathbf{W}_k^M\mathbf{H})_{dl}}{\sum_k a_k(\mathbf{W}_k^M\mathbf{H})_{dl}}\right) \tag{7}$$

$$\mathbf{H} \leftarrow \mathbf{H}.*((\sum_{k=1}^{K} a_k \mathbf{W}_k^M)^{\mathsf{T}}(\mathbf{V}./(\sum_{k=1}^{K} a_k \mathbf{W}_k^M\mathbf{H}))) ./((\sum_{k=1}^{K} a_k \mathbf{W}_k^M)^{\mathsf{T}}\mathbf{1}^{D\times L} + \lambda \mathbf{1}^{J\times L}) \tag{8}$$

ここで，v_{dl} は $\mathbf{V}$ の要素であり，$.*$ と $./$ は要素単位の乗算，除算を表す．

4 評価実験

4.1 実験条件

ATR 研究用日本語音声データベース C セット [9] を用いて話者変換を行い，提案手法を従来の一対一 NMF 声質変換・一対一 GMM 声質変換と比較した．データベースに含まれる男性話者 20 名，女性話者 20 名のうち同性間変換の場合は，10 名のパラレルデータで辞書を構築し，残りの 10 名をテストデータとした．異性間変換の場合は，ソース話者と同性の話者 10 名のパラレルデータで入力辞書を，異性の話者のパラレルデータで 10 名で出力辞書を構築し，残りの 10 名をテストデータとした．いずれの場合も，辞書に含まれないターゲット話者の発話 2 文をデータベースからランダムに選択し，適応データとして用いた．

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 300，λ は 0.1 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 64 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．いずれの手法もサンプリング周波数は 12kHz である．

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 (9) で表されるメルケプストラム歪 (Melcepstrum distortion : MCD) [dB] によって各手法を比較した．

$$MCD = (10/\log 10)\sqrt{2\sum_{d}^{24}(mc_d^{conv} - mc_d^{tar})^2} \quad (9)$$

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．学習データに含まれない 50 文を評価に用いた．

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，目標に話者に近い方を選ぶ XAB テストを行った．いずれの評価項目も，学習データに含まれない 25 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

Table 1 から 4 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 は多対多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．提案手法は入力話者も出力話者も辞書に含まないにも関わらず，一対一声質変換との歪みの差が小さい．さらに，話者の組み合わせによっては，一対一声質変換と同等の変換精度を示しているものもある（F5→F10 や F2→M2 など）．これらの結果より，提案手法は一対一声質変換とほぼ同程度の変換精度を有することがわかる．

Table 1　MCD of male-to-male conversion [dB]

	Source	Multi	NMF	GMM
M1→M6	4.76	4.16	4.06	3.93
M2→M7	5.29	4.92	4.71	4.74
M3→M8	4.68	4.47	4.15	4.23
M4→M9	4.59	4.18	3.92	3.92
M5→M10	4.29	4.02	3.69	3.62
Mean	4.72	4.35	4.11	4.09

Table 2　MCD of female-to-female conversion [dB]

	Source	Multi	NMF	GMM
F1→F6	4.74	4.38	4.19	4.20
F2→F7	4.88	4.52	4.51	4.51
F3→F8	4.77	4.25	4.07	3.99
F4→F9	4.78	4.40	4.18	4.10
F5→F10	4.50	4.07	4.06	4.01
Mean	4.73	4.32	4.20	4.16

Table 3　MCD of male-to-female conversion [dB]

	Source	Multi	NMF	GMM
M1→F1	5.46	4.59	4.32	4.59
M2→F2	5.05	4.59	4.32	4.37
M3→F3	5.22	4.44	4.24	4.27
M4→F4	5.89	4.95	4.83	4.73
M5→F5	5.05	4.39	4.04	4.06
Mean	5.34	4.57	4.35	4.41

Fig. 3 に音質における主観評価結果を示す．誤差範囲は 95%信頼区間を示す．M-to-M，F-to-F，M-to-F，F-to-M はそれぞれ，男性間，女性間，男性から女性，女性から男性への変換であることを示す．同性間の変換においては，提案手法は従来手法を上回っている．異性間の変換においては，提案手法と一対一 NMF 声質変換との結果の差は有意でないものの，提案手法は一対一 GMM 声質変換を上回る結果となっている．以上の結果は 5% 水準の有意検定で確認されている．同性間変換，異性間変換における結果の違いは，異性間変換の場合，入力音声アクティビティの推定に用いる辞書行列と変換に用いる辞書行列が異なるために発生したと考えられる．

Table 4　MCD of female-to-male conversion [dB]

	Source	Multi	NMF	GMM
F1→M1	5.46	4.69	4.48	4.67
F2→M2	5.05	4.42	4.24	4.42
F3→M3	5.22	4.37	4.11	4.24
F4→M4	5.89	4.99	4.75	4.75
F5→M5	5.05	4.34	4.07	4.10
Mean	5.34	4.56	4.33	4.43

Fig. 4 は提案手法と一対一 NMF 声質変換の間の話者性の XAB テスト結果を示す．女性間の変換を除いて，提案手法は一対一 NMF 声質変換をわずかに下回る結果となった．Fig. 5 は提案手法と一対一 GMM 声質変換の間の話者性の XAB テスト結果を示す．提案手法と一対一 GMM 声質変換の間に有意な差がないことがわかる．以上の結果より，提案手法は入力話者の声質を出力話者の声質へと変換できていることが示された．

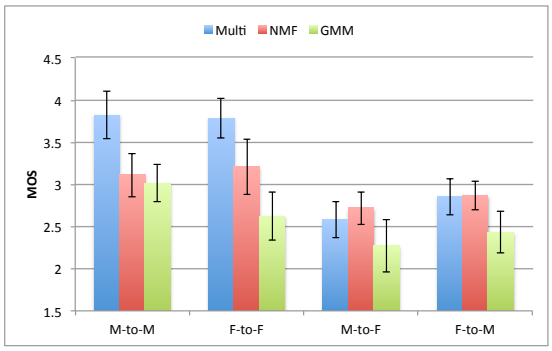

Fig. 3　MOS of speech quality

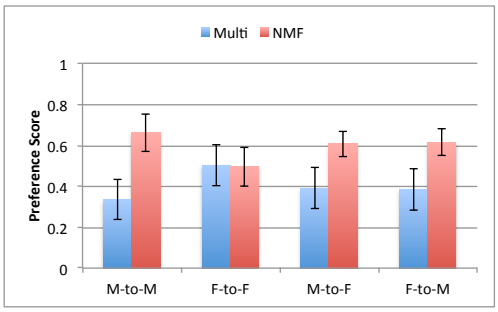

Fig. 4　XAB test between proposed method and NMF

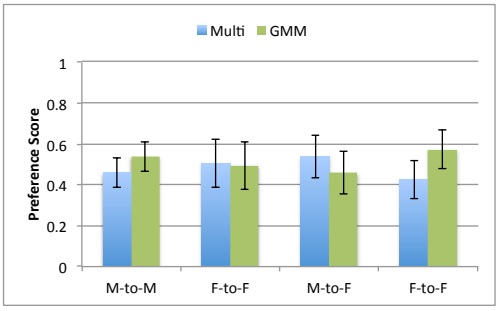

Fig. 5　XAB test between proposed method and GMM

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，入力話者の学習データを必要とせず，出力話者の少量の任意発話データのみで変換できる多対多声質変換を提案した．従来の一対一 NMF 声質変換で必要とされた入力話者と出力話者のパラレル辞書は，Multi-NMF を導入することで，事前に収録された複数の話者のパラレル辞書線形結合で置き換えられた．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，多対多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] Y. Stylianou *et al.*, "Continuous probabilistic transform for voice conversion," IEEE Trans. Speech and Audio Processing, vol. 6, no. 2, pp. 131–142, 1998.

[2] R. Takashima *et al.*, "Exemplar-based voice conversion in noisy environment," in Proc. SLT, pp. 313–317, 2012.

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

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

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

[6] 能勢隆 *et al.*, "ニューラルネットワークに基づくユーザ音声を必要としない多対一声質変換の検討," 日本音響学会 2015 年春季研究発表会講演論文集, pp. 271–274, 2015.

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

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

[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.