一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SP2015-59 (2015-10)

任意話者を対象としたExemplar-based声質変換

相原　　龍† 　滝口　哲也†† 　有木　康雄††

† 神戸大学システム情報学研究科　〒 657-8501 兵庫県神戸市灘区六甲台町 1-1
†† 神戸大学自然科学系先端融合研究環　〒 657-8501 兵庫県神戸市灘区六甲台町 1-1
E-mail: †aihara@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

あらまし　本報告では，Exemplar-based 声質変換の枠組みにおいて，任意話者の発話を異なる任意話者の発話へと変換する多対多声質変換を提案する．従来，声質変換は入力話者の声質を出力話者のものへ変換する話者変換を目的として広く研究されてきた．声質変換において最も一般的な手法は混合正規分布モデル (GMM) を用いた統計的手法であり，統計的声質変換の枠組みは複数の事前収録話者から構成されるパラレルデータセットを用いて，任意の話者から他の任意の話者への変換へと拡張されている．一方，統計的声質変換に代わる手法として Non-negative Matrix Factorization (NMF) を用いた Exemplar-based 声質変換がある．この手法は，NMF が有する雑音除去機能と，Exemplar-based 手法がもつ変換音声の自然性保持という利点から研究が進められている．しかしながら，NMF 声質変換においては入力話者と出力話者のパラレルデータの存在が前提であり，これまでは任意話者の声質変換は不可能であった．そこで本報告では，NMF を拡張した Multiple Non-negative Matrix Factorization (Multi-NMF) を用いた多対多声質変換を提案する．従来手法で必要であった入力話者と出力話者のパラレルデータは，事前に学習された複数話者パラレルデータの線形結合で置き換えられる．本手法は行列表現における話者情報と音韻情報の分解であり，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．

キーワード　声質変換，音声合成，非負値行列因子分解，Exemplar-based，多対多

Exemplar-based voice conversion for arbitrary speakers

Ryo AIHARA†, Tetsuya TAKIGUCHI††, and Yasuo ARIKI††

† Graduate School of System Informatics, Kobe University 1-1 Rokkodai, Nada, Kobe 657-8501, Japan
†† Organization of Advanced Science and Technology, Kobe University 1-1 Rokkodai, Nada, Kobe 657-8501, Japan
E-mail: †aihara@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

Abstract　Voice conversion (VC) is being widely researched in the field of speech processing because of increased interest in using such processing in applications such as personalized Text-To-Speech systems. Statistical approach using Gaussian Mixture Model (GMM) is widely researched in VC and eigen-voice GMM enables one-to-many and many-to-one VC from multiple training data sets. We present in this paper an exemplar-based Voice Conversion (VC) method using Non-negative Matrix Factorization (NMF), which is different from conventional statistical VC. NMF-based VC has advantages of noise robustness and naturalness of converted voice compared to Gaussian Mixture Model (GMM)-based VC. However, because NMF-based VC is based on parallel training data of source and target speakers, we cannot convert the voice of arbitrary speakers in this framework. In this paper, we propose a many-to-many VC method that makes use of Multiple Non-negative Matrix Factorization (Multi-NMF). By using Multi-NMF, an arbitrary speaker's voice is converted to another arbitrary speaker's voice without the need for any input or output speaker training data. We assume that this method is flexible because we can adopt it to voice quality control or noise robust VC.

Key words　Voice Conversion, Speech synthesis, Non-negative Matrix Factorization, Exemplar-based, Many-to–many

1. はじめに

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” [1] を目的として広く研究されてきたが，感情情報を変換する “感情変換” [2]，失われた話者情報を復元する “発話支援” [3] など多岐にわたって応用されている．特に近年は音声合成技術の発達に伴い，音声合成における話者性の制御 [4]，スペクトル復元 [5] や帯域幅拡張 [6] などに応用され注目を集めている．

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 [1] はその精度のよさと汎用性から広く用いられており，多くの改良が行われている．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．戸田ら [7] は従来の GMM を用いた声質変換法に動的特徴と Global Variance を導入することでより自然な音声として変換する手法を提案している．Helander ら [8] は従来手法における過適合の問題を回避するため，Partial Least Squares (PLS) 回帰分析を用いる手法を提案している．

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．Lee ら [9] は事後確率最大化を用いた教師なし学習による声質変換を提案した．戸田ら [10] 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 [11] は参照話者を用いた多対多 EVC を提案した．齋藤ら [12] 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，スパース表現に基づく非負値行列因子分解 (Non-negative Matrix Factorization : NMF) [13] を用いた Exemplar-based 声質変換手法を提案してきた [14]．NMF を用いた声質変換は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いことに加え，高次元スペクトルを用いて変換するため，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 によるノイズ除去手法と組み合わせることでノイズロバスト性を有する．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 章で本稿の提案手法を説明する．第 3 章で従来の GMM・NMF による声質変換手法と比較し，第 4 章で本稿をまとめる．

2. Multi-NMF を用いた多対多声質変換

2.1 概　　要

話者変換は，入力話者の音韻性を維持しつつ話者性を出力話者のものに変換する手法である．提案手法は，以下の 2 つの仮定をおく．

（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アクティビティは，話者非依存の音韻情報をもつ．

本論文でアクティビティとは，基底の線形重み係数行列のことをさす．本論文で提案する Multi-NMF は，入力スペクトルと辞書行列から話者重みベクトル，アクティビティを推定する．従来の一対一 NMF 声質変換で必要とされてきた，入力・出力話者のパラレル辞書は，複数話者のパラレル辞書の線形結合で表現され，その結合重み係数が話者重みベクトルとなる．上記の仮定により，Multi-NMF は話者重みベクトルが話者性を，アクティビティが音韻情報を推定していると考えられる．提案手法では，同性間変換・異性間変換で異なったフローを持つ．

2.1.1 同性間変換

Fig. 1 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 $\mathbf{W}^M_k \in \mathbb{R}^{(D\times J)}$ 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．

$$\mathbf{V}^s \approx \left(\sum_{k=1}^{K} a^s_k \mathbf{W}^M_k\right)\mathbf{H}^s \tag{1}$$

a^s_k 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，辞書行列は固定したまま話者重みベクトルとアクティビティを推定す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．

$$\mathbf{V}^t \approx \left(\sum_{k=1}^{K} a^t_k \mathbf{W}^M_k\right)\mathbf{H}^t \tag{2}$$

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．

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

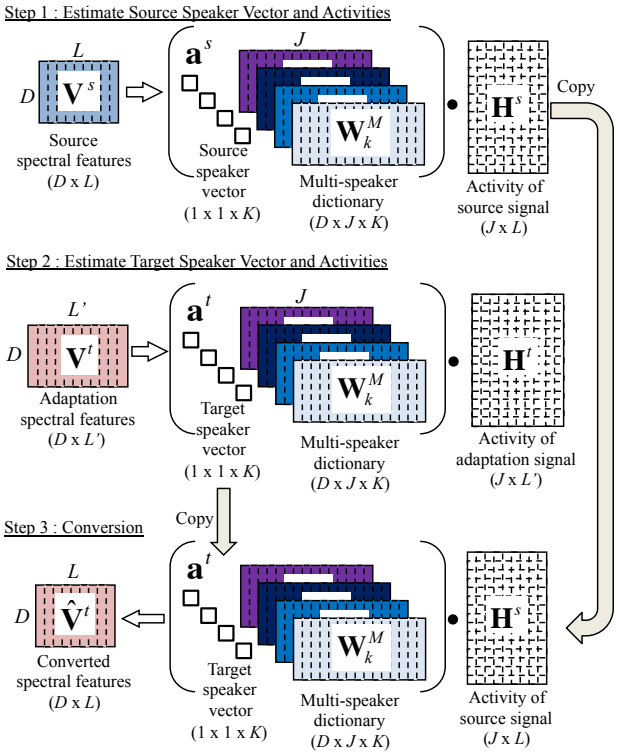


図 1　Multi-NMF を用いた多対多声質変換
Fig. 1　Many-to-many VC using Multi-NMF

2.1.2　異性間変換

計算コスト削減のため，異性間変換では入力辞書，出力辞書の 2 つの辞書行列を用いる．入力辞書行列 $\mathbf{W}^{M_s} \in \mathbb{R}^{(D \times J \times K_s)}$ は入力話者と同じ性別の話者スペクトルで，出力辞書行列 $\mathbf{W}^{M_t} \in \mathbb{R}^{(D \times J \times K_t)}$ は出力話者と同じ性別の話者スペクトルで構成される．K_s と K_t はそれぞれ，入力辞書行列と出力辞書行列に含まれる話者の数を表す．

まず，入力アクティビティ，入力話者重みベクトルを，入力スペクトルと入力辞書行列を用いて推定する．

$$\mathbf{V}^s \approx \left(\sum_{k=1}^{K_s} a_k^s \mathbf{W}_k^{M_s} \right) \mathbf{H}^s \tag{4}$$

つづいて，出力話者の適応データと出力辞書行列から，出力話者重みベクトルとそのアクティビティを推定する．

$$\mathbf{V}^t \approx \left(\sum_{k=1}^{K_t} a_k^t \mathbf{W}_k^{M_t} \right) \mathbf{H}^t \tag{5}$$

最後に，出力話者重みベクトル，出力辞書行列，入力アクティビティから変換スペクトルを合成する．

$$\hat{\mathbf{V}}^t = \left(\sum_{k=1}^{K_t} a_k^t \mathbf{W}_k^{M_t} \right) \mathbf{H}^s \tag{6}$$

2.2　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．

$$d(\mathbf{V}^s, \sum_{k=1}^{K} a_k \mathbf{W}_k^M \mathbf{H}^s) + \lambda ||\mathbf{H}^s||_1 \quad s.t. \quad \mathbf{H}^s \geqq 0, \mathbf{a} \geqq 0 \tag{7}$$

第 1 項は $\mathbf{V}^s$ と $\sum_{k=1}^{K} a_k \mathbf{W}_k^M \mathbf{H}^s$ 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．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．

$$a_k \leftarrow \frac{a_k}{\sum_{d,l} (\mathbf{W}\mathbf{H})_{dl}} \sum_{d,l} \left(\frac{v_{dl}^s (\mathbf{W}_k^M \mathbf{H})_{dl}}{\sum_k a_k (\mathbf{W}_k^M \mathbf{H})_{dl}} \right) \tag{8}$$

$$\mathbf{H}^s \leftarrow \mathbf{H}^s . * ((\sum_{k=1}^{K} a_k \mathbf{W}_k^M)^\mathsf{T} (\mathbf{V}^s ./ (\sum_{k=1}^{K} a_k \mathbf{W}_k^M \mathbf{H}^s))) \\ ./ ((\sum_{k=1}^{K} a_k \mathbf{W}_k^M)^\mathsf{T} \mathbf{1}^{D \times L} + \lambda \mathbf{1}^{1 \times L}) \tag{9}$$

v_{dl}^s は $\mathbf{V}^s$ の要素を表す．式 (8)，(9) の導出は付録に示す．

3.　評 価 実 験

3.1　実 験 条 件

ATR 研究用日本語音声データベース C セット [15] 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 [16] 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 (10) 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．

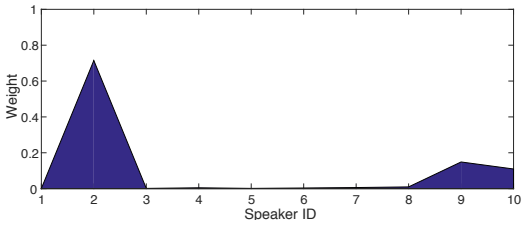
$$MCD = (10/\log 10)\sqrt{2\sum_{d}^{24}(mc_d^{conv} - mc_d^{tar})^2} \qquad (10)$$

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．

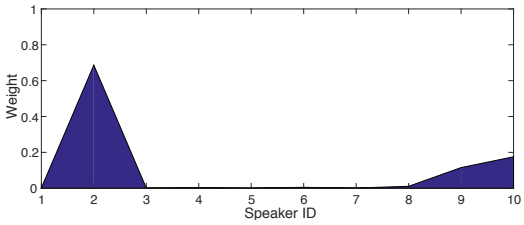
主観評価は成人男女 10 名に対して，音質と話者性の 2 項目について聴取実験を行った．音質の評価基準は MOS 評価基準に基づく主観評価（5:とてもよい，4:よい，3:ふつう，2:わるい，1:とてもわるい）とした．話者性の評価では，はじめに目標話者音声を聴かせた後，異なる手法によって変換した音声を試聴し，目標に話者に近い方を選ぶ XAB テストを行った．いずれの評価項目も，学習データに含まれない 25 文を静かな部屋においてヘッドホンを用いた両耳聴取で評価した．

3.2 実験結果・考察

図 2 に男性話者 (M105) の発話文から推定した話者重みベクトル，図 3 に女性話者 (F105) の発話文から推定した話者重みベクトルを示す．どちらの図も，異なる発話から推定しているにも関わらず，話者間で類似したベクトルが得られている．さらに図より，話者重みベクトルにはスパース制約がないにも関わらず，スパースなアクティビティが推定できていることがわかる．



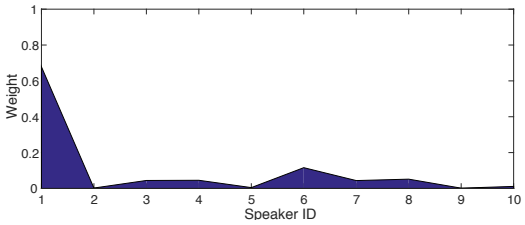
(a) Utterance #C01



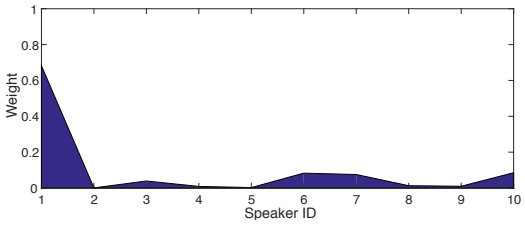
(b) Utterance #C06

図 2　話者重みベクトルの例 (M105)
Fig. 2　Example of speaker weight vectors (M105)

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は小さい．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．



(a) Utterance #C02



(b) Utterance #C11

図 3　話者重みベクトルの例 (F105)
Fig. 3　Examples of speaker weight vectors (F105)

表 1　男性 → 男性変換における MCD [dB]
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

表 2　女性 → 女性変換における MCD [dB]
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

表 3　男性 → 女性変換における MCD [dB]
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

表 4　女性 → 男性変換における MCD [dB]
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 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，入力辞書と出力辞書を分けて用いていることによると考えられる．同性間変換においては，入力話者重みベクトル，入力アクティビティ，出力話者ベクトルを同一の辞書行列から推定している．一方，異性間変換の場合は，入力辞書を用いて入力話者重みベクトルと入力アクティビティを推定し，出力辞書を用いて出力話者ベクトル推定し，入力アクティビティとの積により変換している．異性間変換の場合は入力辞書と出力辞書の間の空間のずれが変換精度に悪影響を及ぼしていると考えられる．

Fig. 5 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. 6 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．

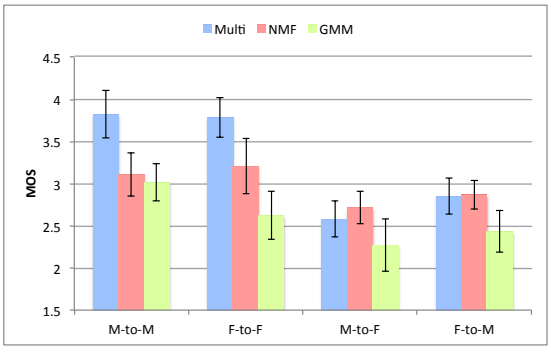


図 4　音質における主観評価実験
Fig. 4　MOS of speech quality

4. 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 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・出力話者の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り，話者によっては一対一 NMF 声質変換より高精度で変換できる可能性があることを示した．今後は，EVC など，他の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．

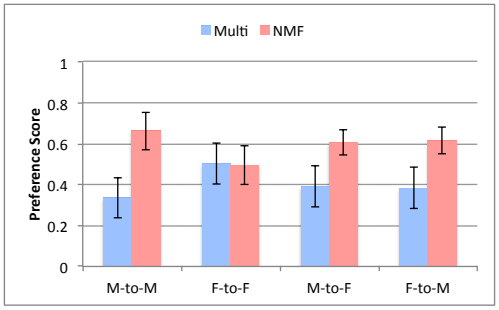


図 5　話者性における提案手法と一対一 NMF の比較
Fig. 5　XAB test between proposed method and NMF

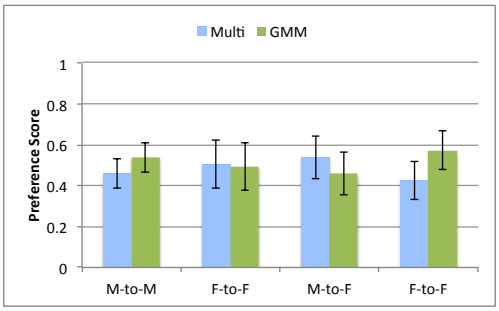


図 6　話者性における提案手法と一対一 GMM の比較
Fig. 6　XAB test between proposed method and GMM

提案手法は，Exemplar-based の枠組みにおける音声の話者性と音韻性の分離を可能にしたといえる．話者重みベクトルによって声質の制御が，アクティビティによって音韻の制御が可能である．話者重みベクトルを声質表現語と結びつけることで，統計的声質変換における重回帰 GMM による声質変換 [17] にあたる，話者性を制御可能な声質変換が実現可能であると考えられ，今後研究を進めていく．

提案手法の課題は，計算コストである．NMF を用いた声質変換は，GMM 声質変換と比較して計算コストが高い傾向にあるが，提案手法は多数話者のパラレルデータを辞書として用いるため，計算コストがさらに高くなっている．コスト削減のためには，よりコンパクトな辞書行列を求める必要があり，研究を進めていく．また，異性間変換においては入力・出力辞書を分けることで精度が下がる傾向があり，コンパクトな辞書表現で辞書を 1 つにすることで，精度向上も可能であると考えらえられる．

さらに，本手法は NMF に備わる雑音除去手法と組み合わせ

が可能であり今後は雑音環境下における任意話者を対象とした声質変換も実現したい．

付　　録

1. NMF 更新式

NMF のコスト関数は，下記のように定義される．

$$d(\mathbf{V}, \mathbf{WH}) + \lambda||\mathbf{H}||_1 \quad s.t. \quad \mathbf{H} \geqq 0 \tag{A·1}$$

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．

$$\mathbf{H} \leftarrow \mathbf{H}.*(\mathbf{W}^{\mathsf{T}}(\mathbf{V}./(\mathbf{WH})))$$
$$./(\mathbf{W}^{\mathsf{T}}\mathbf{1}^{D\times L} + \lambda\mathbf{1}^{1\times L}) \tag{A·2}$$

式 (A·2) の導出は，文献 [13] あるいは [18] を参照されたい．

2. Multi-NMF 更新式の導出

式 (9) は式 (A·2) における $\mathbf{W}$ を $\sum_{k=1}^{K} a_k \mathbf{W}_k^M$ と置き換えることで導出できる．以下，式 (8) の導出を示す．式 (7) のうち，$\mathbf{a}$ に関する第 1 項のみを文字を簡略化して取り出すと，Jensen の不等式を用いて以下のように近似できる．

$$d(\mathbf{V}, \sum_{k=1}^{K} a_k \mathbf{W}_k \mathbf{H}) \tag{A·3}$$
$$= \sum_{d,l} \left\{ v_{dl} \log v_{dl} - v_{dl} \log(\sum_k^K a_k \mathbf{W}_k \mathbf{H})_{dl} \right.$$
$$\left. - v_{dl} + \sum_k^K (a_k \mathbf{W}_k \mathbf{H})_{dl} \right\}$$
$$\leqq \sum_{d,l} \left\{ v_{dl} \log v_{dl} - v_{dl} \sum_k^K \alpha_k \log(\frac{a_k \mathbf{W}_k \mathbf{H}}{\alpha_k})_{dl} \right.$$
$$\left. - v_{dl} + \sum_k^K (a_k \mathbf{W}_k \mathbf{H})_{dl} \right\}$$
$$= Q(\mathbf{V}, \sum_{k=1}^{K} a_k \mathbf{W}_k \mathbf{H}, \alpha_k) \tag{A·4}$$
$$\leqq \sum_{d,l} \left\{ \log v_{dl} - v_{dl} \sum_k \alpha_k \sum_j \beta_j \log(\frac{a_k w_{dj}^k h_{jl}}{\beta_j}) \right.$$
$$\left. + v_{dl}\alpha_k \log \alpha_k - v_{dl} + \sum_k (a_k \sum_j w_{dj}^k h_{jl}) \right\}$$
$$= R(\mathbf{V}, \sum_{k=1}^{K} a_k \mathbf{W}_k \mathbf{H}, \alpha_k.\beta_j) \tag{A·5}$$

α_k と β_j は以下のように定義される．

$$\alpha_k = \frac{a_k \sum_j (w_{dj}^k h_{jl})}{\sum_m a_m \sum_j (w_{dj}^m h_{jl})} \tag{A·6}$$

$$\beta_j = \frac{w_{dj}^k h_{jl}}{\sum_n w_{dn}^k h_{nl}} \tag{A·7}$$

ここで，

$$\frac{\partial R}{\partial a_k} = \sum_{d,l} \left\{ -v_{dl} \sum_k \alpha_k \sum_j \beta_j \frac{1}{a_k} + \sum_k \sum_j w_{dj}^k h_{jl} \right\} \tag{A·8}$$

$\frac{\partial R}{\partial a_k} = 0$ とすると a_k の更新式は式 (8) のように求まる．

文　　献

[1] Y. Stylianou, O. Cappe, and E. Moilines, “Continuous probabilistic transform for voice conversion,” IEEE Trans. Speech and Audio Processing, vol.6, no.2, pp.131–142, 1998.

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

[3] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, “Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech,” Speech Communication, vol.54, no.1, pp.134–146, 2012.

[4] A. Kain and M.W. Macon, “Spectral voice conversion for text-to-speech synthesis,” in Proc. ICASSP, vol. 1, pp.285–288, 1998.

[5] K. Nakamura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, “A mel-cepstral analysis technique restoring high frequency components from low-sampling-rate speech,” in Proc. Interspeech, pp.2494–2498, 2014.

[6] Y. Ohtani, M. Tamura, M. Morita, and M. Akamine, “GMM-based bandwidth extension using sub-band basis spectrum model,” in Proc. Interspeech, pp.2489–2493, 2014.

[7] T. Toda, A. Black, and K. Tokuda, “Voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” IEEE Trans. Audio, Speech, Lang. Process., vol.15, no.8, pp.2222–2235, 2007.

[8] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, “Voice conversion using partial least squares regression,” IEEE Trans. Audio, Speech, Lang. Process., vol. 18, Issue:5,, pp.pp. 912–921, 2010.

[9] C.H. Lee and C.H. Wu, “MAP-based adaptation for speech conversion using adaptation data selection and non-parallel training,” in Proc. INTERSPEECH, pp.2254–2257, 2006.

[10] T. Toda, Y. Ohtani, and K. Shikano, “Eigenvoice conversion based on Gaussian mixture model,” in Proc. Interspeech, pp.2446–2449, 2006.

[11] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, “Many-to-many eigenvoice conversion with reference voice,” in Proc. Interspeech, pp.1623–1626, 2009.

[12] D. Saito, K. Yamamoto, N. Minematsu, and K. Hirose, “One-to-many voice conversion based on tensor representation of speaker space,” in Proc. INTERSPEECH, pp.653–656, 2011.

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

[14] R. Takashima, T. Takiguchi, and Y. Ariki, “Exemplar-based voice conversion in noisy environment,” in Proc. SLT, pp.313–317, 2012.

[15] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, “ATR Japanese speech database as a tool of speech recognition and synthesis,” Speech Communication, vol.9, pp.357–363, 1990.

[16] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigne, “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.

[17] K. Ohta, T. Toda, Y. Ohtani, H. Saruwatari, and K. Shikano, “Adaptive voice-quality control based on one-to-many eigenvoice conversion,” in Interspeech, pp.2158–2161, 2010.

[18] J.F. Gemmeke, T. Viratnen, and A. Hurmalainen, “Exemplar-based sparse representations for noise robust automatic speech recognition,” IEEE Trans. Audio, Speech, Lang. Process., vol.19, no.7, pp.2067–2080, 2011.