

# Multiple Non-negative Matrix Factorization を用いた多対一声質変換

相原 龍<sup>†</sup> 滝口 哲也<sup>††</sup> 有木 康雄<sup>††</sup>

<sup>†</sup> 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  
<sup>††</sup> 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、非負値行列因子分解 (NMF) を拡張した Multiple Non-negative Matrix Factorization (Multi-NMF) を提案し、任意話者の発話を特定話者の発話へと変換する多対一声質変換を行う。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 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からの声質変換は不可能であった。そこで本報告では、Multi-NMF による、入力話者の発話データを学習せずとも変換できる多対一声質変換を提案する。入力話者の発話スペクトルは、事前に学習された複数の話者の発話スペクトルの線形和で表現され、その結合重み係数を用いて目標話者の発話スペクトルへと変換される。この手法は、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。

キーワード 声質変換, 音声合成, 非負値行列因子分解, Exemplar-based, 多対一

## Many-to-one Voice Conversion using Multiple Non-negative Matrix Factorization

Ryo AIHARA<sup>†</sup>, Tetsuya TAKIGUCHI<sup>††</sup>, and Yasuo ARIKI<sup>††</sup>

<sup>†</sup> Graduate School of System Informatics, Kobe University 1-1 Rokkodai, Nada, Kobe 657-8501, Japan  
<sup>††</sup> 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 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 GMM-based VC. However, because NMF-based VC is based on parallel training data of source and target speaker, we cannot convert voice of arbitrary speakers in this framework. In this paper, we propose a many-to-one VC using Multiple Non-negative Matrix Factorization (Multi-NMF). By using Multi-NMF, arbitrary speaker's voice is converted to target speaker's voice without any training data of input speaker's. We assume that this method is flexible because we can adopt it to many-to-many VC or voice quality control.

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

## 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], [3], 失われた話者情報を復元する“発話支援” [4] 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 [5], スペクトル復元 [6] や帯域幅拡張 [7] 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。戸田ら [8] 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 [9] 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 [10] 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 [11] 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 [12] は参照話者を用いた多対多 EVC を提案した。齋藤ら [13] 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) [14] を用いた Exemplar-based 声質変換手法を提案してきた [15]。スパース表現に基づくアプローチは信号処理の分野において注目されており、音声信号処理の分野でも音声認識や音源分離、雑音抑圧などにおいて、その有効性が報告されている。このアプローチでは、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。例えば音源分離に用いる場合、まず学習サンプルや基底を音源毎にグループ (辞書) 化し、混合音声をそれらのスパース表現にする。その後、目的音声の辞書に対する重みベクトルのみを取り出して用いることで、目的音声のみを分離する。Gemmeke ら [16] は雑

音の重畳した音声を、クリーン音声辞書とノイズ辞書のスパース表現にし、クリーン音声辞書に対する重みを音声認識における Hidden Markov Model (HMM) の尤度算出に用いることで、雑音にロバストな音声認識を行う手法を提案している。

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 [15]。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 [17]。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

本章では、従来の一対一声質変換における NMF 声質変換について、その手法と概要を述べる。

### 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

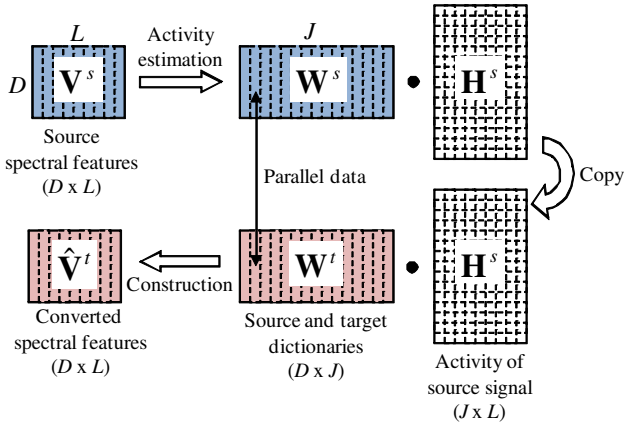


図1 NMFを用いた一対一声質変換の概要  
Fig.1 Basic approach of NMF-based VC

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。

本手法の概要を図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。図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。図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スパースコーディ

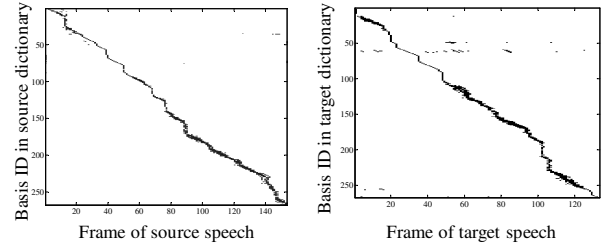


図2 パラレル発話から推定されたアクティビティの例  
Fig.2 Activity matrices for parallel utterances

ングの代表的手法である NMF を用いる [18]. 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。

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

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。 $\lambda$  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。

$$\mathbf{H}^s \leftarrow \mathbf{H}^s \cdot * (\mathbf{W}^{sT} (\mathbf{V}^s ./ (\mathbf{W}^s \mathbf{H}^s))) ./ (\mathbf{W}^{sT} \mathbf{1}^{D \times L} + \lambda \mathbf{1}^{1 \times L}) \quad (5)$$

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。

$$\hat{\mathbf{V}}^t = \mathbf{W}^t \mathbf{H}^s \quad (6)$$

## 3. Multi-NMF を用いた多対一声質変換

本章では、提案手法である Multi-NMF を用いた多対一声質変換を説明する。

### 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。

図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。

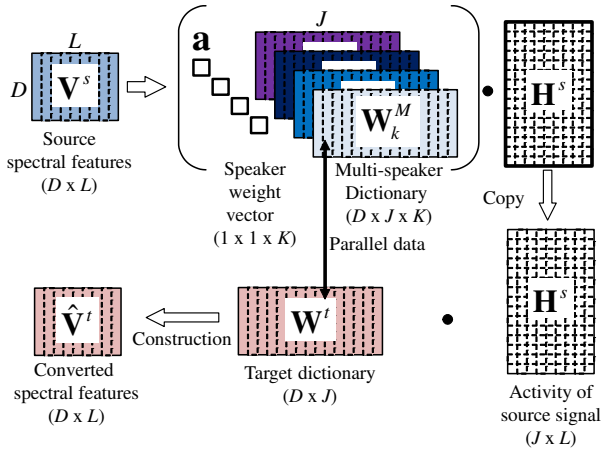


図3 Multi-NMFを用いた多対一声質変換の概要

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

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

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 $\mathbf{a}$ 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。

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

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。

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

$$\mathbf{H}^s \leftarrow \mathbf{H}^s \cdot \left( \left( \sum_{k=1}^K a_k \mathbf{W}_k^M \right)^\top (\mathbf{V}^s ./ \left( \sum_{k=1}^K a_k \mathbf{W}_k^M \mathbf{H}^s \right)) \right) ./ \left( \left( \sum_{k=1}^K a_k \mathbf{W}_k^M \right)^\top \mathbf{1}^{D \times L} + \lambda \mathbf{1}^{1 \times L} \right) \quad (10)$$

$v_{dl}^s$ は $\mathbf{V}^s$ の要素を表す。式(9),(10)の導出は付録に示す。

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。

$$\hat{\mathbf{V}}^t = \mathbf{W}^t \mathbf{H}^s \quad (11)$$

## 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[19]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。いずれの手法もサンプリング周波数は12kHzである。音素バランス50文を学習データとし、従来手法におけるGMMの学習ならびに従来のNMF、提案手法における辞書の構築にそれぞれ用いた。

一対一NMF及び提案手法ではSTRAIGHTスペクトル[20]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。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+ $\Delta$ MFCC+ $\Delta\Delta$ MFCCの60次元を特徴量とした。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。

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

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[21]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聴取しよい方を選ぶ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

図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。図4より、従来の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、ほとんどの話者においてこれら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。話者Cに関しては、若干ではあるが提案手法が一对一 NMF より歪が小さくなっており、この結果は話者によっては提案手法が一对一 声質変換より高い精度を得られる可能性を示している。

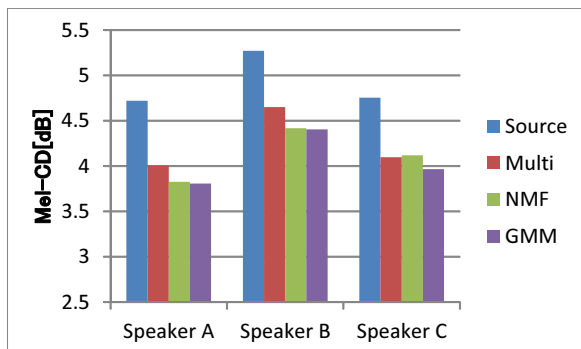


図4 メルケプストラム歪

Fig.4 MelCD calculated from source speech and converted speech using each method

図5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。いずれの話者についても、提案手法と一对一 NMFの間には有意な差は見られないが、これら2つの手法は一对一 GMMを上回っている。図6に話者性における提案手法と一对一 NMFの間のXAB評価結果を示す。話者A, Bに関しては2つの手法の間に有意差はなく、話者Cに関しては若干ではあるが、提案手法が一对一 NMFを上回っている。この結果は客観評価による結果を裏付けるものである。図7に提案手法と一对一 GMMの間のXAB評価結果を示す。全ての話者において、提案手法が一对一 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对一 GMMの精度を上回り、一对一 NMFとほぼ同等の精度が得られることが示された。

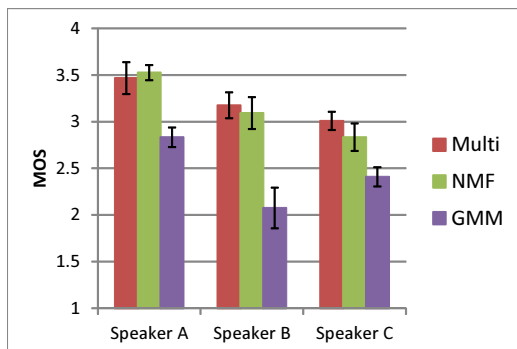


図5 音質のMOS評価値

Fig.5 MOS of speech quality

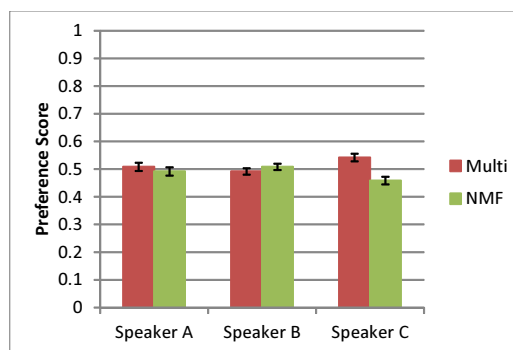


図6 提案手法とNMFにおける話者性の比較

Fig.6 XAB test between proposed method and NMF

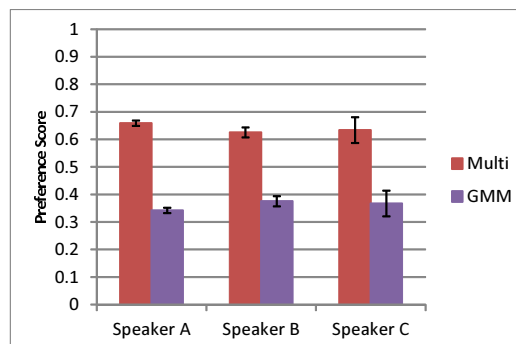


図7 提案手法とGMMにおける話者性の比較

Fig.7 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、提案手法は従来の一对一 NMF 声質変換とほぼ同程度の精度で変換が可能であり、話者によっては一对一 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、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。

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。本報告では任意話者から特定話者への変換である多対一声質変換を提案したが、特定話者から任意話者への変換である一對多声質変換あるいは一切の制約が存在しない多対多声質変換への拡張は容易に可能であると考えられる。また、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 [22] 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。

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. Multi-NMF 更新式の導出

式 (10) は式 (5) における  $\mathbf{W}^s$  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。式 (5) の導出は、文献[14]あるいは[18]を参照されたい。

以下、式 (9) の導出を示す。式 (8) 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。

$$d(\mathbf{V}, \sum_{k=1}^K a_k \mathbf{W}_k \mathbf{H}) \quad (\text{A.1})$$

$$= \sum_{d,l} \left\{ v_{dl} \log v_{dl} - v_{dl} \log \left( \sum_k a_k \mathbf{W}_k \mathbf{H} \right)_{dl} - v_{dl} + \sum_k (a_k \mathbf{W}_k \mathbf{H})_{dl} \right\}$$

$$\leq \sum_{d,l} \left\{ v_{dl} \log v_{dl} - v_{dl} \sum_k \alpha_k \log \left( \frac{a_k \mathbf{W}_k \mathbf{H}}{\alpha_k} \right)_{dl} - v_{dl} + \sum_k (a_k \mathbf{W}_k \mathbf{H})_{dl} \right\}$$

$$= Q(\mathbf{V}, \sum_{k=1}^K a_k \mathbf{W}_k \mathbf{H}, \alpha_k) \quad (\text{A.2})$$

$$\leq \sum_{d,l} \left\{ \log v_{dl} - v_{dl} \sum_k \alpha_k \sum_j \beta_j \log \left( \frac{a_k w_{dj}^k h_{jl}}{\beta_j} \right) + v_{dl} \alpha_k \log \alpha_k - v_{dl} + \sum_k (a_k \sum_j w_{dj}^k h_{jl}) \right\}$$

$$= R(\mathbf{V}, \sum_{k=1}^K a_k \mathbf{W}_k \mathbf{H}, \alpha_k \cdot \beta_j) \quad (\text{A.3})$$

$\alpha_k$  と  $\beta_j$  は以下のように定義される。

$$\alpha_k = \frac{a_k \sum_j (w_{dj}^k h_{jl})}{\sum_m a_m \sum_j (w_{dj}^m h_{jl})} \quad (\text{A.4})$$

$$\beta_j = \frac{w_{dj}^k h_{jl}}{\sum_n w_{dn}^k h_{nl}} \quad (\text{A.5})$$

ここで、

$$\frac{\partial R}{\partial a_k} = \sum_{d,l} \left\{ -v_{dl} \sum_k \alpha_k \sum_j \beta_j \frac{1}{a_k} + \sum_k \sum_j w_{dj}^k h_{jl} \right\} \quad (\text{A.6})$$

$\frac{\partial R}{\partial a_k} = 0$  とすると  $a_k$  の更新式は式 (9) 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 *Interspeech*, pp.2765–2768, 2011.
- [3] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "GMM-based emotional voice conversion using spectrum and prosody features," *American Journal of Signal Processing*, vol.2, no.5, pp.134–138, 2012.
- [4] 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.
- [5] A. Kain and M.W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *ICASSP*, vol. 1, pp.285–288, 1998.
- [6] 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 *Interspeech*, pp.2494–2498, 2014.
- [7] Y. Ohtani, M. Tamura, M. Morita, and M. Akamine, "Gmm-based bandwidth extension using sub-band basis spectrum model," in *Interspeech*, pp.2489–2493, 2014.
- [8] 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.
- [9] 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.912–921, 2010.
- [10] C.H. Lee and C.H. Wu, "Map-based adaptation for speech conversion using adaptation data selection and non-parallel training," in *Interspeech*, pp.2254–2257, 2006.
- [11] T. Toda, Y. Ohtani, and K. Shikano, "Eigenvoice conversion based on Gaussian mixture model," in *Interspeech*, pp.2446–2449, 2006.
- [12] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Many-to-many eigenvoice conversion with reference voice," in *Interspeech*, pp.1623–1626, 2009.
- [13] D. Saito, K. Yamamoto, N. Minematsu, and K. Hirose, "One-to-many voice conversion based on tensor representation of speaker space," in *Interspeech*, pp.653–656, 2011.
- [14] D.D. Lee and H.S. Seung, "Algorithms for non-negative matrix factorization," *Neural Information Processing System*, pp.556–562, 2001.
- [15] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in *SLT*, pp.313–317, 2012.
- [16] J.F. Gemmeke and T. Virtanen, "Noise robust exemplar-based connected digit recognition," in *ICASSP*, pp.4546–4549, 2010.
- [17] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "Individuality-preserving voice conversion for articulation disorders based on Non-negative Matrix Factorization," in *ICASSP*, pp.8037–8040, 2013.
- [18] J.F. Gemmeke, T. Viratnen, and A. Hurmalainen, "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.
- [19] 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.
- [20] 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.
- [21] INTERNATIONAL TELECOMMUNICATION UNION, "Methods for objective and subjective assessment of quality," *ITU-T Recommendation P.800*, pp.800–899, 2003.
- [22] 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.