ソースフィルタ非負値行列因子分解によるボコーダを用いない声質変換の実験的検討*

☆ 須田仁志，齋藤大輔，峯松信明（東大）

1　はじめに

声質変換は一般に入力音声の特徴量を変換するタスクであり，その中でも話者変換は 2 話者間での特徴量の変換関数を推定することが主な目的となる．

特徴量の変換には混合ガウスモデル（GMM）のほかに非負値行列因子分解（NMF）を用いた手法も知られている．GMM を用いた手法では，2 話者の特徴量ベクトルを結合しそのベクトルの確率分布を推定することで変換関数を推定する [1]．一方 NMF を用いた手法では，特徴量が複数のパターン（exemplar）の加算で表されると仮定することで，パラレルデータのもつ時間的共通性を利用してパターンの辞書を推定している [2]．両者はまったく手法が異なる一方で同等の変換品質を実現している．

GMM を用いた声質変換では過平滑化やフレーム間の不自然な特徴量遷移が問題になることが知られており，それらの問題を解決する GV や動的特徴量を考慮する手法が検討されている [3]．NMF でも同様にパラレルデータの時間的不整合によって学習結果の品質が劣化することが指摘されている [4]．

最終的な音声の合成には，推定した変換後の特徴量に加えて対数領域で線形変換を行った基本周波数と組み合わせることが多い．分析・合成には STRAIGHT や WORLD といった分析合成システム（ボコーダ）が多く用いられており，これらは声質変換だけではなくテキスト音声合成などにも用いられている有効な手法である [5]．一方で合成された音声の品質はボコーダの性能が上限となる．そこで本研究では，ボコーダを用いない新しい枠組みについて検討する．

本稿ではまず NMF を用いた声質変換において，学習時の強い時間的同一制約を緩めることで時間的ミスマッチを改善する手法を述べる．また音声のスペクトルをそのまま扱えるように NMF を応用したソースフィルタ NMF（SF-NMF）を示す [6]．そして SF-NMF にもとづくボコーダを用いない新しい声質変換の枠組みを提案する．

SF-NMF を用いる場合には周波数軸が対数スケールのスペクトルが必要になる．そこでスペクトログラムの生成に連続ウェーブレット変換（CWT）を用い，またボコーダの代わりとなる音声の合成には CWT の擬似逆変換および位相推定を用いる [7]．

2　NMF による統計的声質変換

入出力話者のパラレルデータから非負の特徴量（スペクトル包絡）をそれぞれ抽出し，その系列に対して動的時間伸縮（DTW）により時間的対応付けを行う．得られた特徴量系列を $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T]$ および $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T]$ とする．

NMF による声質変換においては，

$$\boldsymbol{x}_t \approx \sum_{n=1}^{N} h_{n,t} \boldsymbol{a}_n^{(x)}, \quad \boldsymbol{y}_t \approx \sum_{n=1}^{N} h_{n,t} \boldsymbol{a}_n^{(y)} \tag{1}$$

が成立するように非負値 $\boldsymbol{a}_n^{(x)}$，$\boldsymbol{a}_n^{(y)}$，$h_{n,t}$ を推定する．ここで N は基底数，n は基底のインデックス，$\boldsymbol{a}_n^{(x)}$ および $\boldsymbol{a}_n^{(y)}$ は両話者の n 番目の基底ベクトル，$h_{n,t}$ は時刻 t における n 番目の基底の重みである．この重みは基底がどの程度生起しているかを表すため，NMF の文脈では生起状態と呼ぶ．上式は行列を用いて

$$\boldsymbol{X} \approx \boldsymbol{A}^{(x)} \boldsymbol{H}, \quad \boldsymbol{Y} \approx \boldsymbol{A}^{(y)} \boldsymbol{H} \tag{2}$$

$$\boldsymbol{A}^{(x)} = \left[\boldsymbol{a}_1^{(x)}, \boldsymbol{a}_2^{(x)}, \ldots, \boldsymbol{a}_N^{(x)}\right] \tag{3}$$

$$\boldsymbol{A}^{(y)} = \left[\boldsymbol{a}_1^{(y)}, \boldsymbol{a}_2^{(y)}, \ldots, \boldsymbol{a}_N^{(y)}\right] \tag{4}$$

$$\boldsymbol{H}_{n,t} = h_{n,t} \tag{5}$$

と表現でき，これは NMF のアルゴリズムを用いることで近似できる [8]．近似には目的関数を次のように定義し，補助関数法を用いて距離を単調に小さくするような反復を行う．

$$D = D_{\mathrm{KL}}\left(\boldsymbol{X} \,\middle|\, \boldsymbol{A}^{(x)} \boldsymbol{H}\right) \tag{6}$$

ここで距離規準 D_{KL} には KL ダイバージェンスを用いている．

NMF を用いた声質変換では，任意の時刻 t での特徴量が N 個の基底の重み付け和によって表すことができると仮定し，かつ 2 話者の同じ発話での生起状態系列が等しくなるように学習する．これにより 2 話者の基底の対応付けをとることができる．変換の際には，学習済みの基底にもとづいて入力音声の特徴量の生起状態を推定し，その生起状態と出力話者基底をかけあわせることで変換後特徴量を得ることができる．この概念図を Fig. 1 に示す．

---

*Experimental Study of Voice Conversion without Vocoders Based on Source-filter Non-negative Matrix Factorization, by Hitoshi SUDA, Daisuke SAITO, Nobuaki MINEMATSU (The University of Tokyo)

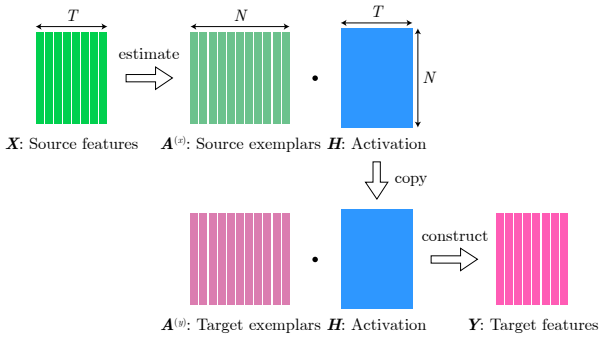


Fig. 1　NMF 声質変換の概念図

NMF を用いた声質変換では 2 話者の特徴量系列の生起状態が同じであるという仮定を行う．しかし生起状態が同じという制約条件が強すぎるために，基底の学習がうまく行えない場合がある．

本稿では出力話者の基底を学習する際に生起状態を固定しない学習法を用いる．ただし生起状態を完全に独立して学習してしまうと入出力話者間の基底の対応付けがとれないため，次式のように入力話者音声の生起状態と推定した生起状態との距離を距離規準に含めることで緩い対応付けをとる．

$$D = D\left(\boldsymbol{Y} \,\middle|\, \boldsymbol{A}^{(y)}\boldsymbol{H}^{(y)}\right) + \mu D\left(\boldsymbol{H}^{(x)} \,\middle|\, \boldsymbol{H}^{(y)}\right) \quad (7)$$

第 1 項は目標とする行列との距離，第 2 項は生起状態行列の距離を広げない制約である．μ は参照係数で正の定数である．

3　SF-NMF

ソースフィルタモデルでは，音声を音源波（ソース）とそれに畳み込まれるフィルタからなると仮定する．SF-NMF はソースフィルタモデルを NMF に適用したものである．ソースフィルタモデルを NMF に適用することで，特徴量としてスペクトル包絡を明示的に分離することなくスペクトルそのものをモデル化することができる．

時刻 t での音声のスペクトル $\boldsymbol{x}_t = [x_t(1), x_t(2), \ldots, x_t(K)]^\top$ を考える．ここで K は周波数ビンの数である．SF-NMF では次式のようにこのスペクトルをモデル化する．

$$x_t(k) \approx \sum_{n,m,s=1}^{N,M,S} h_{n,m,s,t} a_n^{(s)}(k-s) a_m^{(f)}(k) \quad (8)$$

ここで N はソース基底数，M はフィルタ基底数，S はソースの最大周波数ビンシフト数，n，m，s はそれぞれのインデックス，$a_n^{(s)}$ および $a_m^{(f)}$ は対応するインデックスのソースおよびフィルタの基底，$h_{n,m,s,t}$ はインデックスに対応する生起状態である．

この近似式はあるソース基底とあるフィルタ基底の掛け合わせによって得られたスペクトル成分がどの程度生起しているかを推定する式となっている．

上の近似ではソースとフィルタの組み合わせに対して生起状態を定めており，この場合 1 時刻あたりの生起状態の数が NMS となる．しかし発話者が 1 人の音声を分析する場合，ソースとフィルタがそれぞれのパターンにより生成され，それが掛け合わされた形になっていると考えても問題ない．したがって次式のようなモデルが新たに考えられる．

$$x_t(k) \approx \left(\sum_{n,s=1}^{N,S} h_{n,s,t}^{(s)} a_n^{(s)}(k-s)\right)\left(\sum_{m=1}^{M} h_{m,t}^{(f)} a_m^{(f)}(k)\right) \quad (9)$$

これにより 1 時刻あたりの生起状態の数は $NS+M$ となって大幅にパラメータを削減することができる．

単純化した SF-NMF は，ソースとフィルタをそれぞれ NMF で分解し，それに対して要素積をとることに当たる．ただしソースの NMF において基底を N 個のベースとなる基底からの複製と移動により SN 個準備する点が通常の NMF とは異なる．

学習時の更新式は，補助関数法より以下のように導ける．

$$h_{n,s,t}^{(s)} \leftarrow h_{n,s,t}^{(s)} \frac{\sum_k r_t(k) a_n^{(s)}(k-s)\beta_t(k)}{\sum_k a_n^{(s)}(k-s)\beta_t(k)} \quad (10)$$

$$a_n^{(s)}(k) \leftarrow a_n^{(s)}(k) \frac{\sum_{t,s} r_t(k+s) h_{n,s,t}^{(s)} \beta_t(k+s)}{\sum_{t,s} h_{n,s,t}^{(s)} \beta_t(k+s)} \quad (11)$$

$$h_{m,t}^{(f)} \leftarrow h_{m,t}^{(f)} \frac{\sum_k r_t(k) a_m^{(f)}(k)\alpha_t(k)}{\sum_k a_m^{(f)}(k)\alpha_t(k)} \quad (12)$$

$$a_m^{(f)}(k) \leftarrow a_m^{(f)}(k) \frac{\sum_k r_t(k) h_{m,t}^{(f)} \alpha_t(k)}{\sum_k h_{m,t}^{(f)} \alpha_t(k)} \quad (13)$$

ただしここで

$$\alpha_t(k) = \sum_{n,s=1}^{N,S} h_{n,s,t}^{(s)} a_n^{(s)}(k-s) \quad (14)$$

$$\beta_t(k) = \sum_{m=1}^{M} h_{m,t}^{(f)} a_n^{(f)}(k) \quad (15)$$

$$r_t(k) = \frac{x_t(k)}{\alpha_t(k)\beta_t(k)} \quad (16)$$

である．

変換の際には通常の NMF と同様に基底を交換してもう 1 度掛け合わせることで変換後のスペクトルを推定できる．

4　CWT および位相推定

音源波のスペクトルを周波数方向に移動させるとき，倍音を考慮すると周波数軸は対数スケールである必要がある．したがって SF-NMF を用いる場合には周波数軸が対数スケールのスペクトルが必要になる．分析のみの目的であれば，短時間フーリエ変換により得られた線形周波数軸上のスペクトルを補間しても問題ないが，最終的に合成することを考えると品質に疑問の余地が残る．そこで本稿では任意スケールでのスペクトルを分析することが可能な CWT を用いる．

CWT では目的の周波数をピークとするフィルタを畳み込むことで音声に含まれるその周波数の成分を抽出する．音声全体の CWT スペクトログラムのうち，周波数ビンが k の部分だけ取り出したものを $\boldsymbol{s}_k = [s_k(1), s_k(2), \ldots, s_k(T)]^\top$ とし，音声全体のスペクトログラムを $\boldsymbol{s} = \left[\boldsymbol{s}_1^\top, \boldsymbol{s}_2^\top, \ldots, \boldsymbol{s}_K^\top\right]^\top$ とする．分析対象の音声信号を $\boldsymbol{x} = [x(1), x(2), \ldots, x(T)]^\top$ とすれば，CWT は $\boldsymbol{s} = \boldsymbol{W}\boldsymbol{x}$ と表すことができる．ここで

$$\boldsymbol{W} = \left[\boldsymbol{W}_1^\top, \boldsymbol{W}_2^\top, \ldots, \boldsymbol{W}_K^\top\right]^\top \tag{17}$$

$$(\boldsymbol{W}_k)_{ij} = \psi_k(j - i) \tag{18}$$

であり，$\psi_k(t)$ は周波数ピークが k のウェーブレット関数を表す．

最終的に得られた音声は生成されたパワースペクトログラムから推定する必要がある．音声の推定は位相推定と擬似逆変換の 2 つのプロセスに分かれる．擬似逆変換は上で示した行列 $\boldsymbol{W}$ の擬似逆行列 $\boldsymbol{W}^+$ を用いて $\hat{\boldsymbol{x}} = \boldsymbol{W}^+\boldsymbol{s}$ とすれば得られる．

位相推定は反復計算によって行うことができる．位相推定を行う対象となる絶対値をとったスペクトログラム $\boldsymbol{a}$ に対して位相 $\boldsymbol{\phi}$ を与え，$\boldsymbol{a}$ と $\boldsymbol{\phi}$ から得られる位相付きのスペクトログラム $\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi})$ を考えれば，位相推定は以下の目的関数を最小化する問題と考えられる．

$$D = \left\| \boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi}) - \boldsymbol{W}\boldsymbol{W}^+\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi}) \right\|^2 \tag{19}$$

これを単調減少させる更新式を補助関数法を用いて導くことができる．

$$\boldsymbol{\phi} \leftarrow \angle\left(\boldsymbol{W}\boldsymbol{W}^+\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi})\right) \tag{20}$$

CWT や位相推定を時間領域で行う場合には畳み込み演算が必要になるが，フーリエ変換領域で行うことで単純な要素積となり計算を高速化できる．

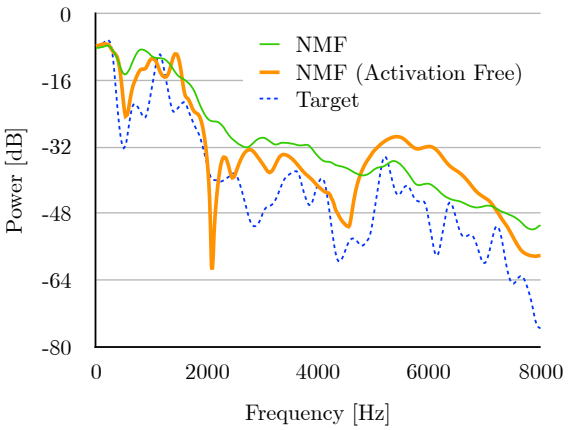


Fig. 2　NMF の学習手法による変換後のスペクトル包絡の変化

5　変換手法の評価

データセットは Voice Conversion Challenge 2016 [9] と同じものを用い，入力話者は SM1（男性），出力話者は TF1（女性）とした．サンプリング周波数は 16 kHz，学習音声は 162 文，変換音声は 54 文である．分析・合成システムには WORLD[*1] を用いた．SF-NMF の学習時には WORLD で得られた F_0 とスペクトル包絡を利用し，学習結果が実際のソース・フィルタに近くなるように距離規準を設けた．

NMF の基底数は 200 とした．SF-NMF の基底数はソースを 21（うち 20 は周波数移動なし），フィルタを 200 とした．また比較手法として GMM にもとづく変換も実装した．GMM の混合数は 128 とし，一次の動的特徴量および GV を考慮した．

5.1　生起状態フリー学習の NMF で生成されたスペクトル包絡

従来手法（出力話者生起状態学習時に生起状態を固定）で学習した NMF と，生起状態フリーで学習した NMF での変換により生成されるスペクトル包絡を比較する．ここでは反復回数を 250 で固定した．

ある時刻でのそれぞれのスペクトル包絡および目標音声のスペクトル包絡を Fig. 2 に示す．従来法はスペクトル包絡全体が滑らかになっているが，生起状態フリーの場合は生成された包絡の周波数方向の分散が大きくなっている．

変換後のスペクトル包絡から計算したメルケプストラム係数と目標音声とのメルケプストラム歪みを Fig. 3 に示す．メルケプストラム歪みの観点ではおおよそ固定が緩いほど目標音声との距離が大きい．

[*1] http://ml.cs.yamanashi.ac.jp/world/

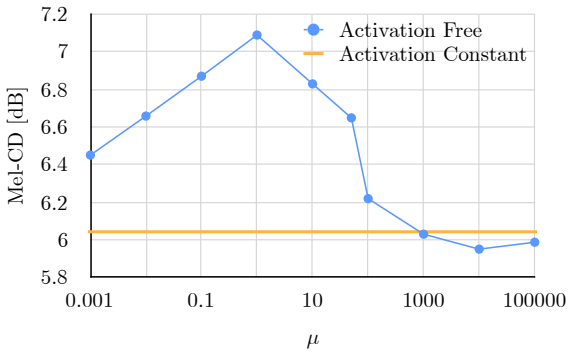
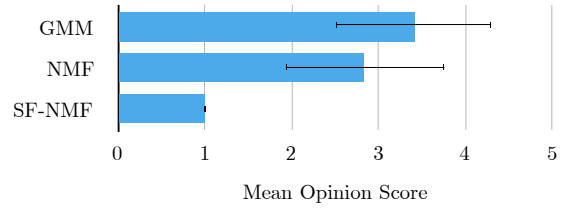
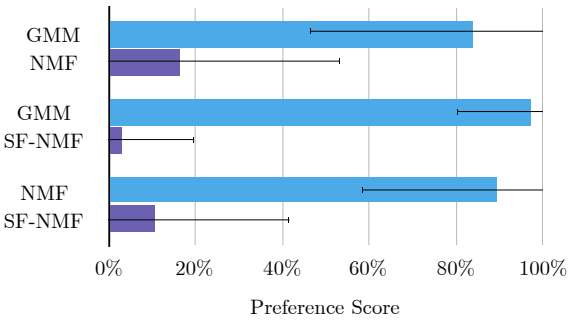


Fig. 3　生起状態フリー学習の参照パラメータ μ によるメルケプストラム歪みの比較



(a) 自然性



(b) 話者の類似性

Fig. 4　声質変換音声の主観評価結果
（エラーバーは 95% 信頼区間）

これは生起状態の変更を許すほど基底の対応付けが弱まるために話者の類似性が失われているためだと考えられる．

5.2　自然性および話者類似性の主観評価

変換音声の自然性および話者の類似性に対して主観評価を行った．自然性は「1: とても悪い」から「5: とても良い」までの 5 段階評価，話者の類似性は XAB テストにより評価した．評価者数は各 4 名である．

実験結果を Fig. 4 に示す．NMF は両評価基準ともに GMM と有意差はなかった．SF-NMF は自然性，話者の類似性ともに GMM および NMF に及ばず，十分な品質の音声が得られなかった．

5.3　SF-NMF による変換音声の考察

主観評価の結果から，今回の提案法では十分に自然な音声を合成することができなかった．より詳細に分析するため，いくつかの条件で音声を生成し，聴感上の印象を確認した．

入力音声を学習済みの基底を用いて分解しもう一度掛け合わせて位相推定した音声は，変換後音声よりも自然な音声であった．またフィルタの基底のみを変換した場合も，F_0 を変えずに変換を行ったような音声を合成することが可能であった．したがってソース基底の交換により大きく音声が劣化しており，入出力話者間のソース基底の対応を取ることが必要であると考えられる．

6　おわりに

本稿では SF-NMF と位相推定を用いた新たな枠組みでの声質変換法を提案した．音声として合成することができたものの，合成音声の自然性や話者の類似性には課題が残ることが実験により示された．今後は SF-NMF を用いたさらに高品質な音声のモデル化と入出力話者間の基底の対応を検討し，声質変換品質の向上を目指す．

参考文献

[1] Stylianou *et al.*, IEEE Speech Audio Process., 6 (2), 131–142, 1998.
[2] Takashima *et al.*, in *Proc. SLT*, 313–317, 2012.
[3] Toda *et al.*, IEEE Trans. Audio, Speech, Language Process., 15 (8), 2222–2235, 2007.
[4] Aihara *et al.*, in *Proc. ICASSP*, 4899–4903, 2015.
[5] Toda, Tokuda, in IEICE Trans. Inf. & Syst., E90-D (5), 816–824, 2007.
[6] Virtanen, Klapuri, in Workshop Advances in Models for Acoustic Processing at *Proc. NIPS*, 2006.
[7] Nakamura, Kameoka, in *Proc. DAFx*, 1–7, 2014.
[8] Lee, Seung, in *Proc. NIPS*, 556–562, 2000.
[9] Toda *et al.*, in *Proc. INTERSPEECH*, 1632–1636, 2016.