非負値行列因子分解による声質変換における INCA アルゴリズムをもとにした基底のノンパラレル学習法*

☆ 須田仁志，小谷岳，齋藤大輔（東大）

1 はじめに

声質変換，とくに話者変換とは，入力音声の言語的内容を維持したまま所望の話者が発話したかのような音声に変換する技術をいう．とくに入出力話者が同一言語内容を発話した音声を学習に要しない声質変換法をノンパラレル声質変換といい，同一内容を発話した音声を要するパラレル変換と比較して利便性が高い．種々のノンパラレル変換法が提案されているが，外部データを利用して話者情報と言語的情報に分解する手法と，外部データを利用せず入出力話者の音声のみからモデルを学習する手法の，大きく 2 手法に分類される．前者には固有声変換 [1]，音素事後確率を用いた手法 [2] などが挙げられ，緻密なモデルが得られやすい一方，言語依存性や外部データの入手の困難さが問題となる．後者には INCA [3] や CycleGAN-VC [4] が挙げられるが，入出力話者のデータ量が前者の手法と比較して多く必要な点，また学習が不安定になりうる点が問題となる．後者の手法の問題点は，入出力話者の音響特徴量の直接的な対応を学習することに終始しており，言語的な整合性を明示的に担保していないことによる．

そこで本稿では，外部知識なしに音響特徴量を話者情報と言語的情報に分解することで，ノンパラレルな入出力話者の発話のみから言語的整合性を維持しつつモデルを学習する手法を提案する．本稿では，非負値行列因子分解（non-negative matrix factorization; NMF）を用いた声質変換法 [5] において，音響特徴量に対する NMF が話者情報と言語的情報の分離を教師なしに実現する特徴を利用する．

2 ベースとなる声質変換技術

2.1 非負値行列因子分解（NMF）

NMF は，1 つの非負行列を 2 つの非負行列の積に分解する手法である [6]．$\boldsymbol{Y} \in \mathbb{R}^{\geq 0, K \times T}$ を分解する対象の行列とする．NMF では $\boldsymbol{Y} \approx \boldsymbol{H}\boldsymbol{U}$ が成立するように，行列 $\boldsymbol{H} \in \mathbb{R}^{\geq 0, K \times N}$ および $\boldsymbol{U} \in \mathbb{R}^{\geq 0, N \times T}$ を計算する．$\boldsymbol{H}$ は基底（base）あるいは辞書（dictionary, exemplar），$\boldsymbol{U}$ は生起状態（activation）と呼ばれる．

NMF は次式の最適化問題を解くことで実現される．

$$\mathcal{D}(\boldsymbol{Y} \,|\, \boldsymbol{H}\boldsymbol{U}) \to \min. \tag{1}$$

ここで $\mathcal{D}$ は分解の規準となるダイバージェンス関数であり，パワースペクトログラムを分解する場合には次式に示す一般化 KL ダイバージェンスを用いる．

$$\mathcal{D}_{\mathrm{KL}}(\boldsymbol{Y} \,|\, \boldsymbol{X}) = \sum_{k,t} \left(Y_{kt} \log \frac{Y_{kt}}{X_{kt}} - Y_{kt} + X_{kt} \right) \tag{2}$$

この最適化問題は制約付きの非線形最適化問題であるため一般には解析的に解けないが，補助関数法を用いて反復的にダイバージェンスを最小化するアルゴリズムが知られている．

$\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots \boldsymbol{y}_T]$ と $\boldsymbol{H} = [\boldsymbol{h}_1, \ldots \boldsymbol{h}_N]$ が縦ベクトルの列からなると考えると，NMF は各 $\boldsymbol{y}_t$ を N 個のベクトル $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_N$ の線形和で近似する．とくに $\boldsymbol{Y}$ がスペクトログラムのような時系列データと仮定すれば，各時刻 t における観測 $\boldsymbol{y}_t$ は，時間に依存しないテンプレート $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_N$ と各テンプレートの強さ $u_{n,t}$ に分解される．

2.2 NMF によるパラレル声質変換

NMF の時系列データを分解する特徴を利用してパラレル声質変換を実現する手法が提案されている [5]．Fig. 1 にこの手法を概略を示す．

$\boldsymbol{Y}^{(s)} = [\boldsymbol{y}_1^{(s)}, \ldots, \boldsymbol{y}_T^{(s)}]$ および $\boldsymbol{Y}^{(t)} = [\boldsymbol{y}_1^{(t)}, \ldots, \boldsymbol{y}_T^{(t)}]$ をそれぞれ入力話者と出力話者が発話した音声の音響特徴量系列とする．ここで特徴量系列はパラレルである，すなわち発話内容は同一で，DTW によって時間的な対応付けがとられているものとする．NMF 声質変換では，パラレルな音響特徴量系列を，次式のように話者依存・時不変の基底 $\boldsymbol{H}$ と話者非依存・時変の生起状態 $\boldsymbol{U}$ に分解する．

$$\boldsymbol{Y}^{(s)} \approx \boldsymbol{H}^{(s)}\boldsymbol{U} \quad \text{and} \quad \boldsymbol{Y}^{(t)} \approx \boldsymbol{H}^{(t)}\boldsymbol{U} \tag{3}$$

式 (3) において，各 $\boldsymbol{h}_n$ は話者ごとの音響特徴量のテンプレートを表しており，生起状態 $\boldsymbol{U}$ は対応するテンプレートの発現の程度を表している．したがって，$\boldsymbol{H}$ は話者の情報を，$\boldsymbol{U}$ は言語的な情報を持つと解釈できる．ここで $\boldsymbol{h}_n^{(s)}$ および $\boldsymbol{h}_n^{(t)}$ は各基底インデックス n で言語的に対応している．

変換時には，入力音響特徴量 $\boldsymbol{Y}^{(s)}$ および入力話

---

*Nonparallel training of exemplars in voice conversion system based on non-negative matrix factorization, by SUDA, Hitoshi, KOTANI, Gaku, and SAITO, Daisuke (the University of Tokyo)

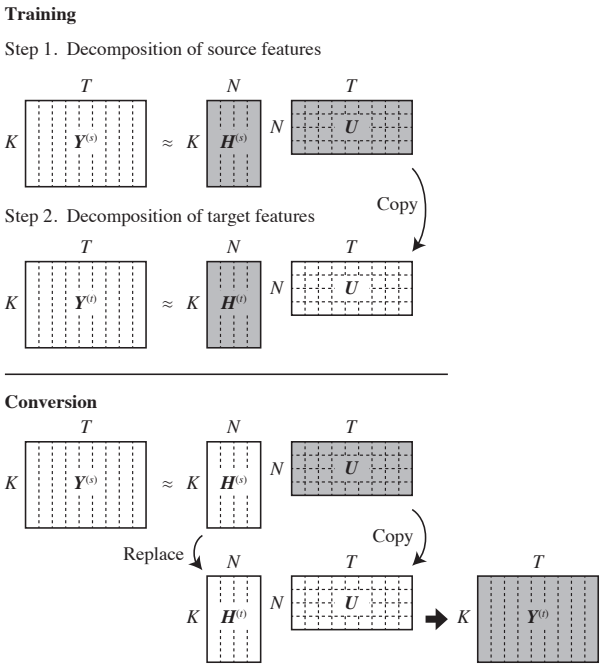


Fig. 1 Overview of the conventional parallel VC system based on NMF [5].

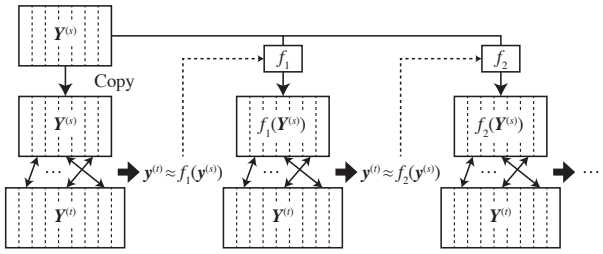


Fig. 2 Overview of an iteration process in INCA [3]. Through iterations, $f_i(\boldsymbol{Y}^{(s)})$ gets more likely to be of the target speaker, and alignment gets feasible.

者基底 $\boldsymbol{H}^{(s)}$ から NMF により生起状態 $\boldsymbol{U}$ を推定することで，変換後の音響特徴量 $\boldsymbol{Y}^{(t)} = \boldsymbol{H}^{(t)}\boldsymbol{U}$ が得られる．

2.3 INCA を利用したノンパラレル声質変換

INCA（an **i**terative combination of a **n**earest neighbor search step and a **c**onversion step **a**lignment method）は，ノンパラレルな特徴量系列対に対し制約のないアラインメントを得るアルゴリズムである [3]．INCA は，入力特徴量の変換，アラインメント，変換モデルの学習，の 3 ステップの繰り返しからなる．Fig. 2 に INCA の概略を示す．

入出力話者の音響特徴量系列をそれぞれ $\boldsymbol{Y}^{(s)} = [\boldsymbol{y}_1^{(s)}, \ldots, \boldsymbol{y}_N^{(s)}]$，$\boldsymbol{Y}^{(t)} = [\boldsymbol{y}_1^{(t)}, \ldots, \boldsymbol{y}_M^{(t)}]$ とする．変換ステップでは，入力特徴量を 1 つ前の反復で学習した変換モデルを用いて変換する．$i-1$ 回目の反復で得られた変換関数を f_i，変換後の入力特徴量を $\boldsymbol{y}_{i,n}^{(s)}$ とすれば，この変換は次のように表式される．

$$\boldsymbol{y}_{i,n}^{(s)} = f_i\Big(\boldsymbol{y}_n^{(s)}\Big) \tag{4}$$

最初の反復では恒等変換を変換関数 f_1 として扱う．アラインメントステップでは，変換後の入力特徴量 $\boldsymbol{y}_{i,n}^{(s)}$ と出力特徴量 $\boldsymbol{y}_m^{(t)}$ の対応を次式のように最近傍法によって得る．

$$p_i(n) = \arg\min_m d\Big(\boldsymbol{y}_{i,n}^{(s)}, \boldsymbol{y}_m^{(t)}\Big) \tag{5}$$

$$q_i(m) = \arg\min_n d\Big(\boldsymbol{y}_{i,n}^{(s)}, \boldsymbol{y}_m^{(t)}\Big) \tag{6}$$

ここで d は規準となる距離関数で，p_i と q_i は推定されたアラインメントを表す．モデル学習ステップでは，入力から出力への変換モデル f_{i+1} を推定されたアラインメントから作成したパラレルデータ $[[\boldsymbol{y}_n^{(s)}, \boldsymbol{y}_{q_i(m)}^{(s)}]^\top, [\boldsymbol{y}_{p_i(n)}^{(t)}, \boldsymbol{y}_m^{(t)}]^\top]^\top$ を用いて学習する．変換モデルには過学習を防ぐため混合数の少ない GMM などの粗い変換モデルを用いる．

3 提案する基底学習法

NMF 声質変換の枠組みにおいて，生起状態は音響特徴量と基底の緩い対応付け（soft alignment）と捉えられる．もし仮に入力話者の音響特徴量系列と出力話者の基底の対応付け（生起状態）が推定できれば，特徴量系列と得られた生起状態から入力話者の基底を推定できる．本稿で提案する基底の学習法は，この対応付けを INCA と同様に行うものである．

提案法は 3 ステップからなる．本学習法の概略を Fig. 3 に示す．

まず，出力話者の音響特徴量系列を分解し出力話者基底 $\boldsymbol{H}^{(t)}$ を得る．このステップは出力話者の音響モデルの構築にあたる．この分解は制約を受けないため，出力話者の特徴量を最も表現可能な基底を得ることができる．

次に，学習済みの出力話者基底 $\boldsymbol{H}^{(t)}$ を用いて，入力話者の音響特徴量系列 $\boldsymbol{Y}^{(s)}$ に対応する生起状態 $\boldsymbol{U}$ を得る．生起状態の推定は特徴量系列–基底間の対応付けであるから，このステップは INCA と同様の方法で実現できる．INCA と異なり対応付けは NMF により行い，変換モデルの学習に用いる出力特徴量は NMF の再構成 $\boldsymbol{X}_i = \boldsymbol{H}^{(t)}\boldsymbol{U}$ によって得る．この推定された特徴量系列 $\boldsymbol{X}_i$ は完全な変換後の特徴量ではないが，出力話者基底の作るベクトル空間に含まれることから $\boldsymbol{Y}^{(s)}$ と比較して出力話者の特徴量に近く，全体としての特徴量変換は徐々に達成される．反復の終了は NMF 前後のダイバージェンス $\mathcal{D}\big(f_i(\boldsymbol{Y}^{(s)}) \,\big|\, \boldsymbol{X}_i\big)$ により判定する．

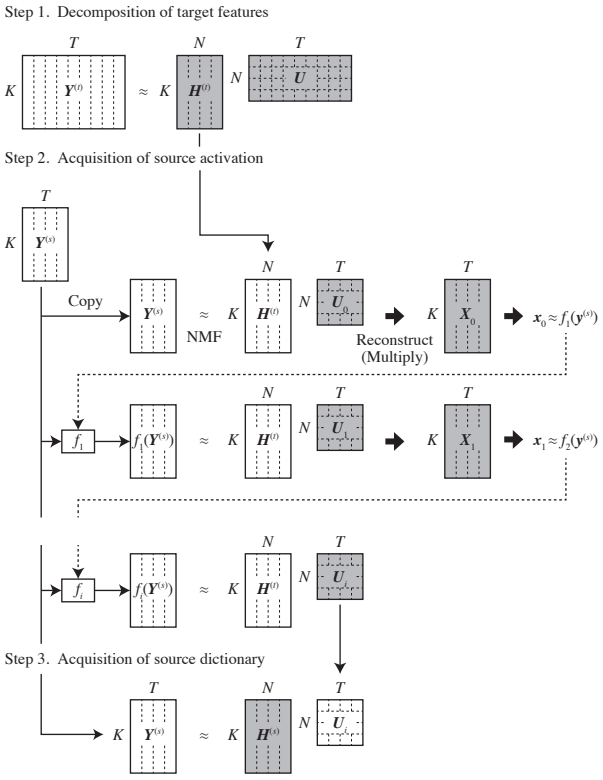


Fig. 3　Overview of the proposed training method of NMF-based VC.

最後に，推定された生起状態 $\boldsymbol{U}$ をもとに入力音響特徴量系列 $\boldsymbol{Y}^{(s)}$ を分解し，入力話者基底 $\boldsymbol{H}^{(s)}$ を得る．

4　実験

4.1　実験条件

本稿では，次の 5 条件下で変換した音声を評価した．

- NP-01，NP-10: 提案法．出力話者基底の学習に 60 文，入力話者基底の学習にそれぞれ 1 文もしくは 10 文を用いた．反復内で用いる変換関数は 25 次メルケプストラム係数のアフィン変換とした．
- CG-01，CG-10: CycleGAN-VC [4][*1]．学習は NP-01 および NP-10 と同様の条件とした．
- PR: NMF を用いたパラレル声質変換法．NP および CG において出力話者基底の学習に用いた文と同一の 60 文を用い学習した．

データセットには Japanese versatile speech（JVS）corpus [7] のうち女性話者 2 名が日本声優統計学会[*2]音素バランス文 100 文を発話した音声を用いた．音声のサンプリング周波数は 24 000 Hz である．入力話者に JVS066，出力話者に JVS010 を選んだ．音声の分析合成には音声分析合成システム WORLD [8]（D4C edition [9]）を用いた．分析時におけるフレームシフトは 1 ms とした．PR における DTW は 25 次メルケプストラム係数での affine-DTW を利用した．NMF の分解対象はパワースペクトログラムとし，距離規準にはメル重み付き一般化 KL ダイバージェンスを用いた．基底数は 200 とし，初期値に制約のない NMF においては対数スペクトルに対する k-means 法により基底の初期値を生成した．基本周波数は対数 F_0 の平均および分散にもとづき線形変換を行い，非周期性指標については変換を行わなかった．主観評価のための聴取実験として，自然性の評価について 2 択の強制選択によるプリファレンス・テスト，話者性の評価について ABX テストを行った．被験者はクラウドソーシングによって募集し，両実験ともに被験者数は 25 名であった．

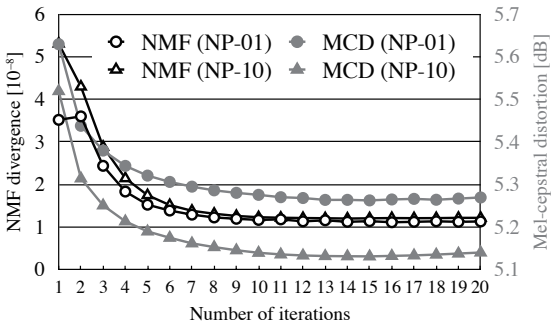


Fig. 4　Results of the transition of the NMF divergence and the mel-cepstral distortion with the number of iterations in the proposed systems.

4.2　提案法における NMF ダイバージェンスの収束

提案した基底学習法において，NMF 前後のダイバージェンス $\mathcal{D}\big(f_i\big(\boldsymbol{Y}^{(s)}\big)\,\big|\,\boldsymbol{X}_i\big)$ の収束について調査した．同時に，再構成後のスペクトル $\boldsymbol{X}_i$ と真の出力話者発話とのメルケプストラム歪みを計算した．出力話者発話はパラレルデータであり本来ノンパラレル変換の枠組み内では利用できないことに留意する．

Fig. 4 にこの結果を示す．反復を行うにつれ，NMF ダイバージェンスおよびメルケプストラム歪みは同様に減少した．反復にしたがって一時的な変換の性能が向上すること，また NMF ダイバージェ

[*1] 実装は https://github.com/leimao/Voice_Converter_CycleGAN による．

[*2] https://voice-statistics.github.io/

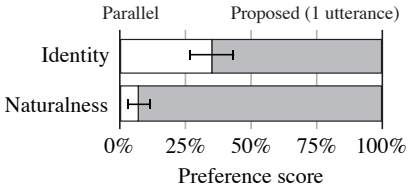
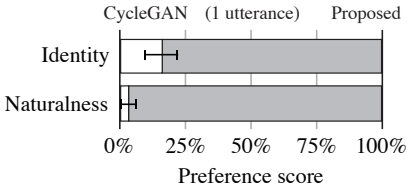
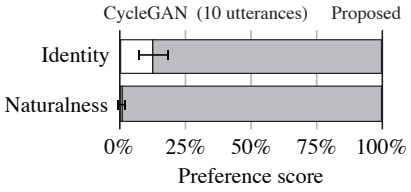


Fig. 5 Results of subjective evaluations about speaker identity and naturalness of converted utterances between the systems PR and NP-01. Error bars denote 95% confidential intervals (similarly in latter figures).



(a) Comparison of CG-01 and NP-01.



(b) Comparison of CG-10 and NP-10.

Fig. 6 Results of subjective evaluations about speaker identity and naturalness of converted utterances between the proposed systems and the CycleGAN-VC systems.

ンスにより収束が評価できることが示された．また，さらに反復を行うことにより NMF のダイバージェンスおよびメルケプストラム歪みがともに上昇することから，過学習も示唆された．

4.3 変換品質に関する主観評価

各条件下で変換した音声について，聴取実験によりその品質を評価した．

PR と NP-01 を比較した結果を Fig. 5 に示す．提案法 NP-01 は，パラレル NMF 変換 NP と比較して高い自然性および話者類似性を示した．DTW の誤りによる過平滑化の影響を受けたためパラレル手法の自然性は低下したと考えられる．

次に，ともにノンパラレル手法である CG と NP を比較した結果を Fig. 6 に示す．本実験の条件下では CycleGAN-VC と比較して提案法は高い品質を示した．提案法は NMF の強い制約のもとで学習するため，自由度の高い CycleGAN に比べ少量の学習

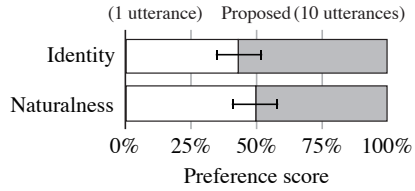


Fig. 7 Results of subjective evaluations about speaker identity and naturalness of converted utterances by the proposed method with the different number of training sentences of the source speaker, that is, the systems NP-01 and NP-10.

データにおいても自然な特徴量を生成できたと推察される．

最後に，学習文数の異なる NP-01 と NP-10 を比較した結果を Fig. 7 に示す．学習文数を増やしたことによる話者性の向上が見られた．学習文数が異なるのは入力話者の発話であるが，文数を増やすことで反復内の一時変換の品質が高くなり，話者変換がより正確に実現できたと考えられる．

5 結論

本稿では，NMF を用いたパラレル声質変換に対して INCA の枠組みを学習に取り入れることで，外部データの存在しないノンパラレル条件下でも学習可能な NMF 声質変換法を提案した．主観評価実験により，既存のパラレル手法や CycleGAN-VC と比較して品質が高く，また入力話者の学習データが少ない条件下でも声質変換を実現できることが確認された．さらなる実験として，異性間変換・異言語間変換においても同様の性能が得られるかどうかを評価する必要がある．

謝辞 本研究開発は総務省 SCOPE（受付番号 182103104）の委託を受けたものです．

参考文献

[1] Toda *et al.*, *in Proc.* INTERSPEECH 2006, 2446–2449, 2006.
[2] Sun *et al.*, *in Proc.* ICME 2016, 1–6, 2016.
[3] Erro *et al.*, IEEE Trans. Audio, Speech, Language Process., 18 (5), 944–953, 2010.
[4] Kaneko *et al.*, *in Proc.* EUSIPCO 2018, 2100–2104, 2018.
[5] Takashima *et al.*, *in Proc.* SLT 2012, 313–317, 2012.
[6] Lee *et al.*, *in Proc.* NIPS 13, 556–562, 2001.
[7] Takamichi *et al.*, arXiv:1908.06248, 2019.
[8] Morise *et al.*, IEICE Trans. Inf. Syst., E99-D (7), 1877–1884, 2016.
[9] Morise, Speech Communication, 84, 57–65, 2016.