# 声質変換における非周期性指標の影響とその評価*

☆伊藤大貴, 相原龍, 滝口哲也, 有木康雄 (神戸大)

## 1 はじめに

声質変換とは, 入力した音声の音韻情報などは保ったまま, 話者性や音韻性，感情性などの特定の情報を維持しながら他の情報を変換する技術である．音韻情報を保ちながら話者情報を変換する話者変換 [1], 感情情報を変換する感情変換 [2], 更には話者情報を復元する発話支援 [3] など様々なタスクへの応用が期待されている．

これまで声質変換では，STRAIGHT [4], WORLD [5] に代表される VOCODER を用いた特徴量分析が行われてきた．これらは音声信号を F0，スペクトル包絡，非周期性指標の 3 つの特徴量に分解する．従来の声質変換においては話者変換が主なタスクとして研究されてきたことから，スペクトル包絡を特徴量変換の対象とすることが多かった．

スペクトルの変換手法として統計的手法が多く提案され, 中でも混合正規分布モデル (Gaussian Mixture Model: GMM) を用いた手法 [1] が広く用いられている．戸田ら [6] は従来の GMM を用いた声質変換法に動的特徴と Global Variance を導入することによってより自然性の高い音声の変換を提案している．Helander [7] らは Partial Least Squares (PLS) 回帰分析を用いることにより，従来の手法における過学習の問題を回避するための手法を提案している．

また，従来の統計的手法とは異なる, スパース表現に基づく Non-negative Matrix Factorization(NMF) [8] を用いた Exemplar-based な声質変換手法が提案されている [9]．この手法では, 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. 変換時には構築された辞書を固定し, 入力音声を入力辞書の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手法である. スパース表現に基づく手法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 [10].

本稿では, 声質変換の精度向上を目指し, 非周期性指標 (Aperiodicity index: AP) に注目する. 大谷ら [11] は，GMM 声質変換手法を用いて AP を出力話者のものへと変換することで，変換音声の自然性を向上させられることを示した．しかしながら，GMM 特徴量変換では AP を圧縮した低次元特徴量を用いている．スペクトル変換における NMF 声質変換のように高次元特徴量での変換を行うことで，さらなる変換精度向上が期待できる．本稿では，Dynamic Kernel Partial Least Squares (DKPLS) [12] と NMF を用い，高次元 AP を変換する手法を提案する．

本手法では, 振幅スペクトルと AP は独立に学習され, 抽出された特徴量は WORLD を用いて合成する．評価実験から AP 変換のもたらす声質変換への影響について述べ, 比較検討を行う．

## 2 NMF を用いた声質変換

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．

$$\mathbf{v}_l \approx \sum_{j=1}^{J} \mathbf{w}_j h_{j,l} = \mathbf{W}\mathbf{h}_l \tag{1}$$

$\mathbf{v}_l$ 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．

$$\mathbf{V} \approx \mathbf{W}\mathbf{H} \tag{2}$$

$$\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_L], \quad \mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_L]. \tag{3}$$

ここで $L$ はフレーム数を表し, 本手法では，$\mathbf{W}$ は学習データで固定される．

本手法の概要を Fig. 1 に示す．$\mathbf{V}^s$ は入力話者音声特徴量，$\mathbf{W}^s$ は入力話者辞書，$\mathbf{W}^t$ は出力話者辞書，$\hat{\mathbf{V}}^t$ は変換された音声特徴量，$\mathbf{H}^s$ は入力特徴量から推定されるアクティビティを表す．$D$，$J$ はそれぞれ特徴量の次元数，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

*Influence of aperiodicity index on voice conversion and its evaluation, by Daiki Ito, Ryo Aihara, Tetsuya Takiguchi, Yasuo Ariki (Kobe univ.)

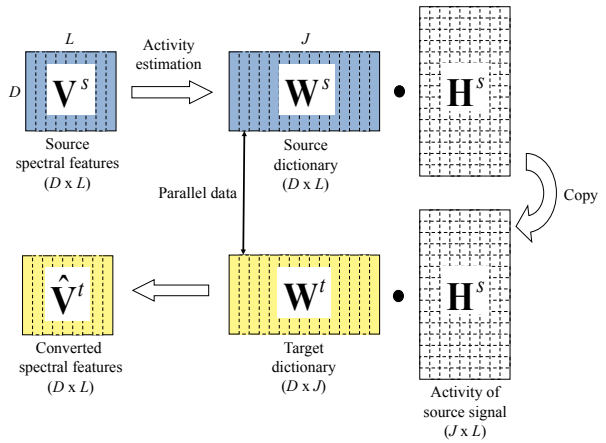


Fig. 1　Basic approach of NMF-based voice conversion

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．

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ングの代表的手法である非負値行列因子分解 (NMF) [13] を用いる．NMF のコスト関数は，$\mathbf{V}^s, \mathbf{W}^s, \mathbf{H}^s$ を用いて以下の式で表せる．

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

ここで第 1 項は $\mathbf{V}^s$ と $\mathbf{W}^s\mathbf{H}^s$ の間の Kulback-Leibler(KL) 距離で，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．

$$\begin{aligned}\mathbf{H}^s \leftarrow \mathbf{H}^s .* (\mathbf{W}^{s\mathsf{T}}(\mathbf{V}^s ./ (\mathbf{W}^s\mathbf{H}^s))) \\ ./ (\mathbf{W}^{s\mathsf{T}}\mathbf{1}^{J\times L} + \lambda \mathbf{1}^{J\times L})\end{aligned} \tag{5}$$

変換特徴量 $\hat{\mathbf{V}}^t$ は，$\mathbf{W}^t$ と推定された $\mathbf{H}^s$ の内積をとることで得られる．

$$\hat{\mathbf{V}}^t = \mathbf{W}^t\mathbf{H}^s \tag{6}$$

## 3　DKPLS を用いた声質変換

本手法の概要を Fig. 2 に示す. NMF を用いた声質変換同様に，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る．入力話者の学習データに k-means クラスタリングを適用し，参照ベクトル $\mathbf{z}_j = [\mathbf{z}_1 \dots \mathbf{z}_D]^\mathsf{T}$ を生成する．生成された参照ベクトルと入力話者ベクトル $\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_N]^\mathsf{T}$ から，Gaussian Kernel を用いて，入力特徴量は高次元空間へと写像される．

$$k_{jn} = e^{\frac{-||\mathbf{x}_n - \mathbf{z}_j||^2}{2\sigma^2}} \tag{7}$$

$\sigma$ はパルツェン窓幅で適当な値 [14] を用いる．参照ベクトルの数を $C(j = 1, \dots, C)$ とすると，この結果から高次元特徴量が生成される．

$$\mathbf{K} = \begin{bmatrix} k_{11} & k_{12} & \dots & k_{1N} \\ k_{21} & k_{22} & \dots & k_{2N} \\ & & \dots & \\ k_{C1} & k_{C2} & \dots & k_{CN} \end{bmatrix} \tag{8}$$

以上のようなカーネルトリックにより，$D$ 次元の特徴量は $N$ 次元の高次元空間へと写像することが可能であり，入力特徴量は以下のような高次元特徴量 $\mathbf{k}_n$ で表される．

$$\mathbf{k}_n = \begin{bmatrix} k_{1n}, & k_{2n}, & \dots, k_{Cn} \end{bmatrix}^T \tag{9}$$

得られた入力話者の高次元特徴量 $\mathbf{k}_n$ と出力話者特徴量の間の回帰行列を推定する．

$$\mathbf{y}_n = \boldsymbol{\beta}\mathbf{k}_n + \mathbf{e}_n \tag{10}$$

$\mathbf{y}_n$ は出力話者ベクトル，$\mathbf{e}_n$ は剰余項，$\boldsymbol{\beta}$ は回帰行列を表す．この式 (10) の回帰行列 $\boldsymbol{\beta}$ を解くための手法として PLS を用いる．$\boldsymbol{\beta}$ は Table 1 に示されたアルゴリズム [15] によって求められる．

Table 1　Algorithm of SIMPLS

| | |
|---|---|
| **Initialize** | $\mathbf{R}, \mathbf{V}, \mathbf{Q}$ and $\mathbf{T}$ |
| | $\mathbf{c} = \mathbf{K}\mathbf{Y}^\mathbf{T}$ , $\mathbf{q} = max(\mathbf{C}^\mathbf{T}\mathbf{C})$ |
| **Set** | $\mathbf{r} = \mathbf{C}\mathbf{q}$ , $\mathbf{K}^\mathbf{T}\mathbf{r}$ |
| | $\mathbf{t} = \mathbf{t} - mean(\mathbf{t})$ |
| **Normalize** | $\mathbf{r} = \mathbf{r}/\|\|\mathbf{r}\|\|$ and $\mathbf{t} = \mathbf{t}/\|\|\mathbf{t}\|\|$ |
| **Set** | $\mathbf{p} = \mathbf{K}\mathbf{t}$ , $\mathbf{q} = \mathbf{Y}\mathbf{t}$ and $\mathbf{u} = \mathbf{Y}^\mathbf{T}\mathbf{q}$ |
| **Set** | $\mathbf{v} = \mathbf{p}$ |
| | $\mathbf{v} = \mathbf{v} - \mathbf{V}\mathbf{V}^\mathbf{T}\mathbf{p}$ and $\mathbf{u} = \mathbf{u} - \mathbf{T}\mathbf{T}^\mathbf{T}\mathbf{u}$ |
| **Normalize** | $\mathbf{v} = \mathbf{v}/\|\|\mathbf{v}\|\|$ |
| **Set** | $\mathbf{C} = \mathbf{C} - \mathbf{v}\mathbf{v}^\mathbf{T}\mathbf{C}$ |
| Assign $\mathbf{r}, \mathbf{q}, \mathbf{v}$ and as the $i$th columns of matrices $\mathbf{R}, \mathbf{Q}, \mathbf{V}$ and $\mathbf{T}$, respectively and the regression matrix $\beta$ is obtained as $\boldsymbol{\beta} = \mathbf{R}\mathbf{Q}^\mathbf{T}$ | |

## 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セット [16] を用いて話者変換を行い，提案手法である NMF による AP の変換を用いた声質変換 (NMF-ap), DKPLS による AP の変換を用いた声質変換 (DKPLS-ap) を比較した． スペクトル包絡は NMF で話者変換を行ったものを用いた．入力話者は男性 A，B の 2 人, 出力

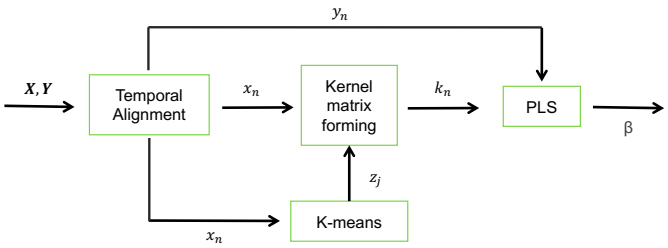


Fig. 2　Basic approach of DKPLS-based voice conversion

話者は女性 A，B の 2 人で男性 A → 女性 A，男性 A → 女性 B，男性 B → 女性 B の 3 通りの組み合わせを用いて実験し，サンプリング周波数は 16kHz とした．音素バランス 50 文を学習データとし，特徴量抽出手法として WORLD を用いた．

スペクトル包絡の変換には，WORLD スペクトル 513 次元と前後 2 フレームを含む 1539 次元特徴量とし，300 回の更新を行った．また辞書の基底数は約 80000 個ある辞書の中から，10000 個をランダムに取り出し，アクティビティの推定を行った．NMF の AP の変換には，スペクトルと同様に，WORLD 非周期性指標 ap513 次元と前後 2 フレームを含む 1539 次元特徴量とし，300 回の更新を行った．辞書基底は約 80000 個ある辞書の中から，1000 個をランダムに取り出し，アクティビティの推定を行った．DKPLS の AP の変換では，PLS のコンポーネント数は実験的に求められた RMSE 精度が最もよかった 128 とし，WORLD 非周期性指標 ap + Δ の 1026 次元を特徴量とした．

F0 については，AP 変換における提案手法の有効性を示すため，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．

テストデータとして客観評価には，パラレル辞書内に含まれない 50 文を用いた．客観評価指標として，AP の平均二乗誤差 (Root Mean Square Error，以下 RMSE とする．) を用いて各手法を比較した．RMSE の式は以下の式 (11) で表される．

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \qquad (11)$$

ここで，$y_i$ ，$\hat{y}_i$ はそれぞれ，入力話者の AP，出力話者の AP を表す．

主観評価は成人男女 10 名に対して，音質について聴取実験を行った．評価基準は MOS 評価基準 [17] 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）を用いた．最も客観評価指標のよかった 1 つの変換話者対の 20 文に対して，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

Fig.3 に RMSE による実験結果の比較を示す．A → A においては，どちらの手法も RMSE を減少できていることがわかる．しかし，A → B，B → B においては，DKPLS-ap は変換によって RMSE が減少しているが，NMF-ap は RMSE が増加するという結果になった．NMF-ap の RMSE が増加してしまったのは，実験では 80000 基底の中からランダムに選択した 1000 基底を用いているが，この選択が不適切だったためと考えられる．

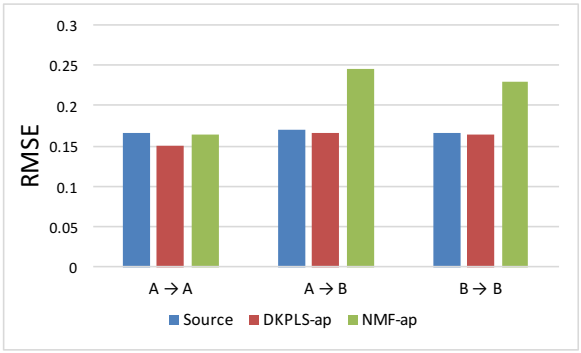


Fig. 3　RMSE of each method

次に主観評価の実験結果を Fig.4 に示す．エラーバーは，95%信頼区間を示す．図より，提案手法の 2 つは AP を変換しなかった場合と比較して音質が優れていることがわかる．

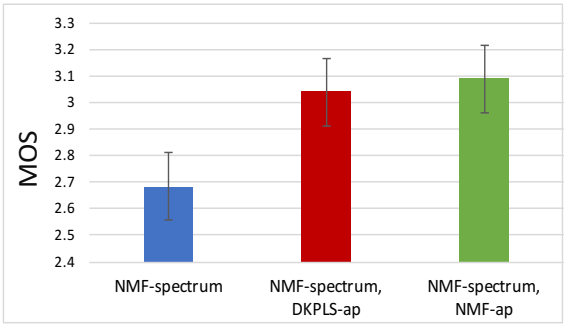


Fig. 4　MOS experiment of each method

## 5 おわりに

本稿では AP の変換を含む声質変換手法についての検討を行った．従来の NMF 声質変換では特徴量変換に用いていなかった AP を高次元特徴量の学習が可能な DKPLS，NMF を用いて AP 変換することにより，自然性の精度向上が可能であることを示した．しかし，客観評価においては，NMF-ap は RMSE の減少ができなかった話者のペアも存在した．

今後の課題として，AP を変換するための最適な辞書基底の選択手法の実現が挙げられる．また，非周期性指標は雑音混入の影響や，声帯振動の揺れを観測している成分であるため，AP 変換により雑音環境下における任意話者を対象とした声質変換の実現も考えられる．さらには，スペクトルと AP の同時モデリングによる自然性向上が考えられ，今後研究を進めていく．

## 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] C. Veaux and X. Robet, "Intonation conversion from neutral to expressive speech," in Proc. Interspeech, pp. 2765–2768, 2011.

[3] K. Nakamura *et al.*, "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech," Speech Communication, vol. 54, no. 1, pp. 134–146, 2012.

[4] H. Kawahara et al., "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneousfrequencybased F0 extraction: possible role ofa repetitive structure in sounds," Speech Communication, vol. 27, no. 3–4, pp. 187–207, 1999.

[5] M. Morise *et al.*, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," IEICE transactions on information and systems, vol. E99-D, no. 7, pp. 1877–1884, 2016.

[6] T. Toda *et al.*, "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.

[7] E. Helander *et al.*, "Voice conversion using partial least squares regression," IEEE Trans. Audio, Speech, Lang. Process., vol. 18, pp. 912–921, 2010.

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

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

[10] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 3, pp. 1066–1074, 2007.

[11] Y. Ohtani *et al.*, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed exciation," in Proc. Interspeech, pp. 2266–2269, 2006.

[12] E. Helander *et al.*, "Voice conversion using dynamic kernel partial least squares regression," IEEE Trans. Audio, Speech, Lang. Process., vol. 20, pp. 806–817, 2011.

[13] J. Gemmeke *et al.*, "Exemplar-based sparse representations for noise robust automatic speech recognition," IEEE Trans. Audio, Speech, Lang. Process., pp. 2067–2080, 2011.

[14] M. J. Embrechts and B. Szymanski, "Introduction to Scientific Data Mining: Direct Kernel Methods & Applications," Computationally Intelligent Hybrid Systems. New York: Wiley, ch., pp. 317–365, 2005.

[15] "SIMPLS: An alternative approach to partial least squares regression," Chemometrics Intell. Lab. Syst., vol. 18, no. 3, pp. 251–263, 1993.

[16] A. Kurematsu *et al.*, "ATR Japanese speech database as a tool of speech recognition and synthesis," Speech Communication, vol. 9, pp. 357–363, 1990.

[17] INTERNATIONAL TELECOMMUNICATION UNION, "Methods for objective and subjective assessment of quality," ITU-T Recommendation P.800, pp. 800–899, 2003.