声質変換のための音素識別的特徴量*

◎相原龍, 滝口哲也, 有木康雄 (神戸大)

1 はじめに

音声に含まれる音韻情報を維持しつつ話者情報を変換する手法として，声質変換がある．近年の深層学習に対する注目の高まりから Deep Neural Networks (DNN) に基づく声質変換 [1] や，Non-negative Matrix Factorization (NMF) を用いた Exemplar-based アプローチ [2] が提案されている．しかしながら，統計的声質変換は依然として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．

統計的声質変換の代表的例は混合正規分布モデル (Gaussian Mixture Model : GMM) を用いた手法 [3, 4] である．Helander ら [5] は GMM 声質変換における過適合の問題を回避するため，Dynamic Kernel Partial Least Squares (DKPLS) に基づく声質変換を提案しており，少量学習データ環境において GMM 声質変換を上回る結果を示している．

本稿では，PLS に基づく声質変換に対して適用可能な声質変換のための音素識別的特徴量を提案する．これまで，多くの声質変換手法において，特徴量の音素識別性は考慮されてこなかった．また，声質変換の応用例として障害者支援があり，特に構音障害者を対象とした声質変換の場合，入力話者の音韻空間が曖昧なため，音素識別的特徴量の必要性は極めて高いと考えられる．

提案手法では，Locality Sensitive Discriminant Analysis (LSDA) [6] を用いてスペクトル空間から音素識別的な特徴量空間を推定する．学習データの音素ラベルから，クラス間グラフとクラス内グラフを作成し，カーネルトリックを用いて非線形変換で音素識別的な高次元空間へと写像する．GMM 声質変換と異なり，PLS 声質変換は過学習を防ぐことができるため，提案手法のような局所性を考慮した特徴量に適していると考えられる．提案手法は，健常者話者変換と障害者話者変換の 2 つのタスクにおいて，客観評価と主観評価によって評価された．

以下，第 2 章で提案手法を述べる．第 3 章で評価実験を行い，第 4 章で本稿をまとめる．

2 音素識別的特徴量抽出

2.1 概略

提案手法は，学習と変換の 2 つ段階から構成される．Fig. 1 に学習段階の概要を示す．初めに，学習データである入力話者と出力話者のパラレル発話に対して STRIAGHT 分析 [7] を適用し，STRAIGHT スペクトルを得る．次に STRAIGHT スペクトルから低次元のメルケプストラムとその動的特徴量を計算し，それらを結合する．これらの特徴量をスペクトル特徴量と呼ぶ．入力話者のスペクトル特徴量に対して k-means クラスタリングを適用し，参照ベクトルを得る．さらに，得られた参照ベクトルを用いて，入力話者スペクトル特徴量にカーネル変換を適用し高次元空間に写像する．続いて，得られた入力話者の高次元特徴量に LSDA を適応することで音素識別的特徴量変換行列を推定する．高次元写像されたスペクトル特徴量は，推定された特徴量変換行列を用いることで音素識別的特徴量へと変換される．音声の動的特徴を捉えるため，変換された特徴量はセグメント化された後，出力話者のスペクトル特徴量と DTW を用いてアライメントをとる．最後に，PLS 回帰によって話者変換行列を得る．

Fig. 2 に変換段階の概要を示す．入力された STRAIGHT スペクトルから，メルケプストラムとその動的特徴量を計算し，スペクトル特徴量を得る．参照ベクトルによるカーネル変換の後，スペクトル特徴量は音素識別的特徴量へと変換される．最後に，PLS 回帰によって求められた話者変換行列によって，入力話者のスペクトル特徴量は出力話者のものへと変換される．

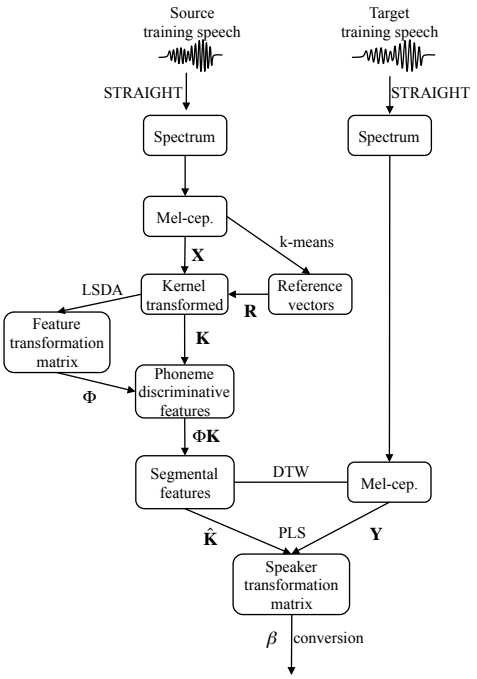

Fig. 1 Overview of the training phase of our proposed method.

2.2 音素識別的特徴量

音素識別的特徴量を LSDA [6] に基づいて推定する．まず，特徴量の非線形構造を表現するため，スペクトル特徴量をカーネル変換によって高次元空間へと写像

*Phoneme-Discriminative Features for Statistical Voice Conversion by Ryo AIHARA, Tetsuya TAKIGUCHI, Yasuo ARIKI (Kobe University)

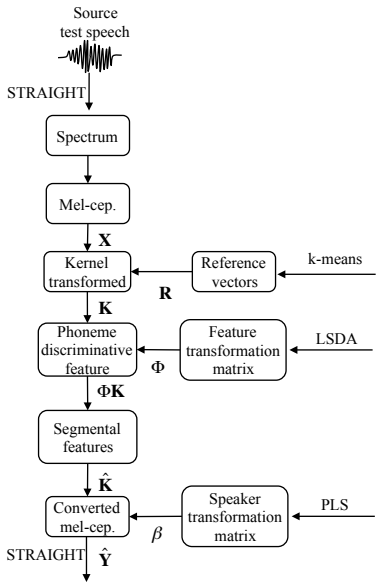

Fig. 2　Overview of the test phase of our proposed method.

する．カーネル関数には様々なものがあるが，本稿ではGaussian カーネルを用いる．入力スペクトル特徴量 $X_{d,i}, i = 1, ..., I$ は，参照ベクトル $R_{d,n}, n = 1, ..., N$ によって以下のように写像される．

$$k_{i,n} = exp(\frac{-||X_{d,i} - R_{d,n}||^2}{2\sigma^2}) \tag{1}$$

ここで，I と N はそれぞれ入力話者の学習データ数，参照ベクトル数を表す．このようにして入力話者スペクトルは非線形の高次元空間へと写像される．

カーネル写像された特徴量空間の音素的構造を推定するため，クラス内分散グラフとクラス間分散グラフの隣接行列をそれぞれ以下のように求める．

$$\mathbf{A}^w_{ij} = \begin{cases} 1 & \left(\begin{array}{c} \mathbf{k}_i \in N_{k_w}(\mathbf{k}_i) \ or \ \mathbf{k}_j \in N_{k_w}(\mathbf{k}_j) \\ and \\ c_i = c_j \end{array} \right) \\ 0 & (otherwise) \end{cases} \tag{2}$$

$$\mathbf{A}^b_{ij} = \begin{cases} 1 & \left(\begin{array}{c} \mathbf{k}_i \in N_{k_b}(\mathbf{k}_i) \ or \ \mathbf{k}_j \in N_{k_b}(\mathbf{k}_j) \\ and \\ c_i \neq c_j \end{array} \right) \\ 0 & (otherwise) \end{cases} \tag{3}$$

ここで，$N_{k_w}(\mathbf{v}_i)$，$N_{k_b}(\mathbf{v}_i)$ はそれぞれ，クラス内分散グラフにおける学習データ $\mathbf{v}_i$ の k_w 近傍集合，クラス間分散グラフにおける $\mathbf{v}_i$ の k_b 近傍集合を表す．c_i と c_j はそれぞれ $\mathbf{k}_i$ と $\mathbf{k}_j$ の音素ラベルである．隣接行列によってクラス間分散のグラフラプラシアンが以下のように定義される．

$$\mathbf{L}^b \quad = \quad \mathbf{D}^b - \mathbf{A}^b \tag{4}$$

ここで，$\mathbf{D}^b$ は対角成分に $\mathbf{A}^b$ の各行（あるいは列）の和をもつ重み行列である．

LSDA に基づいて，音素識別的特徴量変換行列 $\mathbf{\Phi}$ が下記のようにして求められる．

$$\begin{aligned} arg\ max\ & \mathbf{\Phi}^\mathsf{T}\mathbf{K}(\alpha\mathbf{L}^b + (1-\alpha)\mathbf{A}^w)\mathbf{K}^\mathsf{T}\mathbf{\Phi} \\ s.t.\ & \mathbf{\Phi}^\mathsf{T}\mathbf{K}\mathbf{D}^w\mathbf{K}^\mathsf{T}\mathbf{\Phi} = 1 \end{aligned} \tag{5}$$

ここで，$\mathbf{D}^w$ は対角成分に $\mathbf{A}^w$ の各行（あるいは列）の和をもつ重み行列，α $(0 \leq \alpha \leq 1)$ はクラス間グラフに対する重み行列である．(5) を最大化する $\mathbf{\Phi}$ は下記の固有値最大化問題に基づいて求められる．

$$\mathbf{\Phi}^\mathsf{T}\mathbf{K}(\alpha\mathbf{L}^b + (1-\alpha)\mathbf{A}^w)\mathbf{K}^\mathsf{T}\mathbf{\Phi} = \lambda\mathbf{\Phi}^\mathsf{T}\mathbf{K}\mathbf{D}^w\mathbf{K}^\mathsf{T}\mathbf{\Phi} \tag{6}$$

ここで，λ は固有値を表す．

2.3　Dynamic Partial Least Square 回帰分析

特徴量間の動的特徴を捉えるため，入力話者の音素識別的特徴量からセグメント特徴量を構築する，

$$\hat{\mathbf{K}}_t = [\mathbf{\Phi}\mathbf{K}^\mathsf{T}_{t-1}\mathbf{\Phi}\mathbf{K}^\mathsf{T}_t\mathbf{\Phi}\mathbf{K}^\mathsf{T}_{t+1}]^\mathsf{T} \tag{7}$$

こうして得られた入力話者のセグメント特徴量 $\hat{\mathbf{K}}_t$ と出力話者のスペクトル特徴量 $\mathbf{Y}_t$ はそれぞれ，話者非依存の隠れベクトル $\mathbf{h}_t$ の線形結合で下記のように表される．

$$\hat{\mathbf{K}}_t = \mathbf{Q}\mathbf{h}_t + \mathbf{e}^x_t \tag{8}$$

$$\mathbf{Y}_t = \mathbf{P}\mathbf{h}_t + \mathbf{e}^y_t \tag{9}$$

ここで $\mathbf{Q}$ と $\mathbf{P}$ は話者依存の行列，$\mathbf{e}^x_t$ と $\mathbf{e}^y_t$ は残差項を表す．SIMPLS [8] に基づいて，話者変換行列 β が $\mathbf{Q}$ と $\mathbf{P}$ から求められる．

変換段階においては，セグメント特徴量 $\hat{\mathbf{K}}$ がテストデータから推定された特徴量変換行列に基づいて推定される．話者変換された特徴量 $\hat{\mathbf{y}}$ は下記のようにして得られる．

$$\hat{\mathbf{Y}}_t = \beta\hat{\mathbf{K}}_t \tag{10}$$

3　評価実験

3.1　実験条件

提案手法は，健常者話者変換と構音障害者を含む話者変換の 2 つをタスクとして評価した．健常者話者変換では，ATR 研究用日本語音声データベースに含まれる男性話者 2 人，女性話者 2 人を用いて，男性から男性 (M101→M102)，女性から女性 (F101→F102)，男性から女性 (M101→F101)，女性から男性 (F101→M101) の変換を行った．音素バランスを考慮した 50 文を学習に，学習データに含まない 50 文をテストに用いた．

障害者話者変換においては，アテトーゼ型脳性麻痺による日本人男性の構音障害者 1 名を入力話者とした．出力話者は ATR 研究用日本語音声データベースから男性話者 1 名を選択した．音素バランスを考慮した 50 文を学習に，学習データに含まない 50 文をテストに用いた．

音声のサンプリングレートは 16 kHz である．音声信号から音声分析合成手法 STRAIGHT [7] によって F0，スペクトル包絡，非周期成分が抽出される．STRAIGHT スペクトルからメルケプストラムと動的特徴量を計算し，これらを結合したものをスペクトル特徴量とする．この特徴量の次元数は，健常者話者変換ではパワーを含まない 48 次元，障害者話者変換ではパワーを含む 50 次元である．

本稿では，提案手法を含む以下の 4 つの手法を比較する．

- **ML-GMM-D**: 対角共分散行列を用いた最尤推定に基づく GMM 声質変換 [4].
- **ML-GMM-F**: 非対角共分散行列を用いた最尤推定に基づく GMM 声質変換 [4].
- **DKPLS**: カーネル変換された特徴量に PLS 声質変換を適用したもの [5].
- **PDKPLS** (提案手法): 音素識別的特徴量を PLS でモデル化したもの.

GMM 声質変換において，ガウス分布の混合数は $\{1, 2, 4, 8, 16, 32, 64, 128, 256\}$ の中から実験的に最適なものを選択した．PLS に基づく声質変換では隠れ素子数を同様の集合から実験的に選択した．提案手法において，α は $\{0.1, 0.5, 1.0\}$ のなかから実験的に最適なものを選択した．

3.2 客観評価

客観評価指標として，メルケプストラム歪（Mel-cepstral distortion: MelCD）[dB] を用いた．

$$MelCD = (10/\log 10)\sqrt{2\sum_{d}^{24}(mc_d^{conv} - mc_d^{tar})^2} \quad (11)$$

ここで，mc_d^{conv}，mc_d^{tar} は変換メルケプストラム，目標メルケプストラムにおける d 次元目の特徴量を表す．

Fig. 3 に健常者話者変換における音素識別的特徴量の次元数による変換精度の変化を示す．図より，特徴量の次元数が小さすぎる場合には，変換精度が劣化することがわかる．

Fig. 4 に健常者話者変換における DKPLS の隠れ素子数による変換精度の変化を示す．図より，提案手法である PDKPLS は，隠れ素子数が少ない場合でも精度良く変換できることがわかる．このことから，本稿で提案した音素識別的特徴量が音素空間を捉えていることがわかる．

Table 1 に各変換手法ごとの健常者話者変換の MelCD を示す．太字は変換対における最も良い結果を示す．表より，ML-GMM-D と ML-GMM-F の間にはほとんど差がないことがわかる．DKPLS は，女性間と男性から女性への変換においては GMM 声質変換より良い結果を示した．提案手法はすべての変換対について DKPLS を上回り，特に女性間と男性から女性への変換については他の手法を大きく上回った．

Fig. 5 に障害者話者変換における，各手法ごとの MelCD を示す．図より，提案手法が他の手法を上回ったことがわかる．

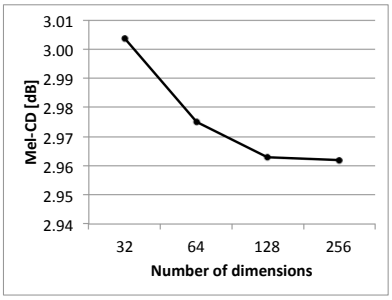

Fig. 3 MelCD [dB] of male-to-male conversion as a function of number of dimensions of phoneme discriminative feature.

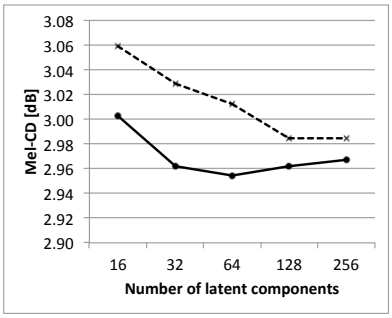

Fig. 4 MelCD [dB] of male-to-male conversion as a function of latent components of PLS. Dashed line denotes DKPLS and solid line denotes the proposed PDKPLS.

Table 1 Mel-CD of non-dysarthric speech conversion [dB]

	M-to-M	F-to-F	M-to-F	F-to-M
Source	3.96	3.69	4.17	4.17
ML-GMM-D	2.96	2.84	2.88	2.86
ML-GMM-F	**2.95**	2.84	2.88	**2.82**
DKPLS	2.98	2.84	2.87	2.84
PDKLPS	**2.95**	**2.81**	**2.84**	**2.82**

3.3 主観評価

主観評価実験は，10 人の日本語話者が 25 文のテストデータについてそれぞれの手法で変換した音声を評価した．本論文では，音質と話者性の 2 つの観点において評価実験を行った．音質に関しては 5 段階評価 (5: excellent, 4: good, 3: fair, 2: poor, 1: bad) で，話者性については，X を目標話者とし 2 つの変換手法のうちよい方を選ぶ XAB 法で評価した．考察は全て t 検定による有意差検定に基づく．

Fig. 6 に健常者話者変換における音質の評価結果を示す．エラーバーは，95% 信頼区間である．提案手法は，女性間と，男性から女性への変換について ML-GMM-D より良い結果を示した．他の変換対に関しては，有意差が得られなかった．Fig. 7 に健常者話者変換における話者性の評価結果を示す．提案手法と ML-GMM-D の間に有意差は見られなかった．以

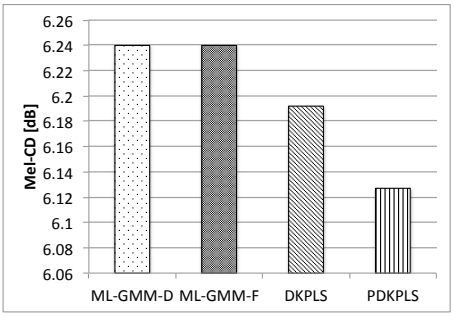

Fig. 5　MelCD [dB] of dysarthric speech conversion.

上の結果から，提案手法は音質において利点があると考えられる．

Fig. 8 に障害者話者変換における評価結果を示す．話者性において有意差は見られなかったものの，提案手法は音質において ML-GMM-D より有意に優れていることを示した．

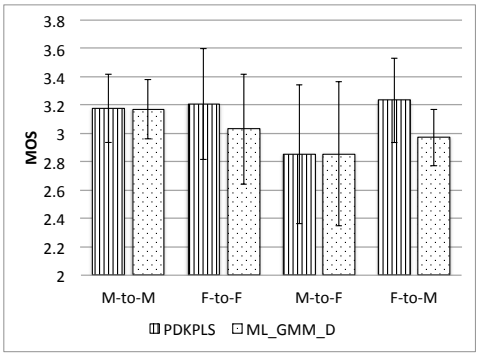

Fig. 6　Results of MOS test on speech quality for the task of non-dysarthric speaker conversion.

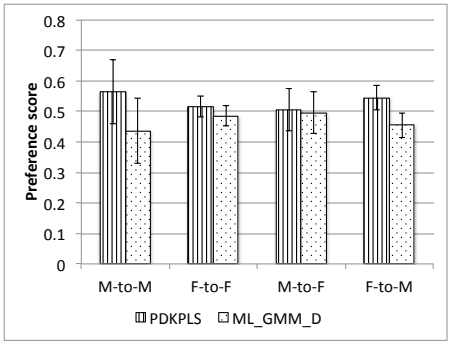

Fig. 7　Results of XAB test on similarity for the task of non-dysarthric speaker conversion.

4　終わりに

本稿では，PLS に基づく声質変換に対応した音素識別的特徴量を提案した．音素スペクトルの局所性と幾何学的な性質をとらえることを目的とし，音素ラベルに基づいて LSDA を適用し，音素識別的特徴量を推定した．音素識別的な空間は過学習しにくいことで知られる PLS 回帰によってモデル化される．評価実験によって，提案手法はこれまで一般的であった GMM 声質変換と比較して自然性の高い変換音声が得られることがわかり，特に構音障害者声質変換において有効であることが示された．今後の課題として，

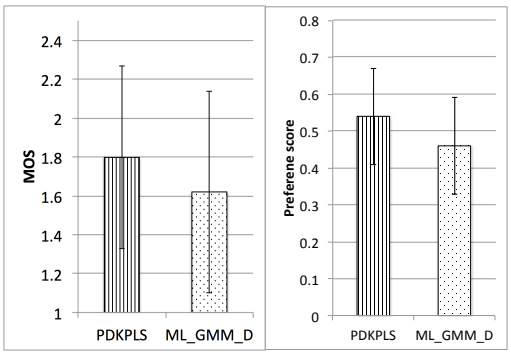

Fig. 8　Results of subjective evaluation for the task of dysarthric speaker conversion. Left: MOS test on speech quality. Right: XAB test on similarity.

提案手法における話者性の精度向上を挙げることができる．

参考文献

[1] C. Ling-Hui *et al.*, “Joint spectral distribution modeling using restricted Boltzmann machines for voice conversion,” in Proc. Interspeech, pp. 3052—3056, 2013.

[2] R. Aihara *et al.*, “Multiple non-negative matrix factorization for many-to-many voice conversion,” IEEE/ACM Trans. Audio, Speech, and Lang. Process., vol. 24, no. 7, pp. 1175–1184, 2016.

[3] Y. Stylianou *et al.*, “Continuous probabilistic transform for voice conversion,” IEEE Trans. Speech and Audio Processing, vol. 6, no. 2, pp. 131–142, 1998.

[4] 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.

[5] E. Helander *et al.*, “Voice conversion using dynamic kernel partial least squares regression,” IEEE Trans. Audio, Speech, Lang. Process., vol. 20, no. 3, pp. 806–817, 2012.

[6] D. Cai *et al.*, “Locality sensitive discriminant analysis,” in Proc the 20th International Joint Conference on Artificial Intelligence,, pp. 708–713, 2007.

[7] H. Kawahara, “STRAIGHT, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds,” Acoustical Science and Technology, pp. 349–353, 2006.

[8] S. de Jong, “SIMPLS: An alternative approach to partial least squares regression,” Chemometrics Intell. Lab. Syst, vol. 18, no. 3, pp. 251–263, 1993.