差分スペクトル補正に基づく統計的歌声声質変換*

☆小林 和弘，戸田 智基，Graham Neubig，Sakriani Sakti，中村 哲（奈良先端大・情報）

1 はじめに

歌声は音楽を形成する上で重要な要素の 1 つであり，人は歌声の音高，リズム，強弱を巧みに操作する事で，多様な歌唱表現を生み出す．声質に関しても一定の範囲で操作可能であるが，個人が生成できる声質は身体的特徴により大きく制限されるため，身体的特徴を超えた声質での歌唱は困難である．これに対して，身体的制約を超えた多様な声質での歌唱の実現を目指し，統計的手法に基づく歌声声質変換（SVC: Singing Voice Conversion）が提案されている [1]．SVC は，混合正規分布モデル（GMM: Gaussian Mixture Model）に基づき入力歌手の音響特徴量を目標歌手の音響特徴量へと変換する事で，入力歌手の声質を目標歌手の声質へと変換する．一方で，ボコーダの使用に伴い，F_0 分析誤差やスペクトル包絡のモデリング誤差，さらには，GMM による変換誤差が発生するため，自然歌声に比べ音質劣化が生じる．

本稿では，主に同性歌手間における歌声においては音高変換が必要とならない点に着目し，高い自然性を持つ変換歌声を実現するために，スペクトル包絡の補正処理に基づく SVC を提案する．提案法は，音源特徴量の変換を行わないことで，ボコーダによる波形合成処理を回避する．実験結果から，従来の GMM に基づく SVC と比べ，提案法は高い自然性を持つ歌声変換が可能である事を示す．

2 GMM に基づく SVC

GMM に基づく SVC は，入力歌手の声質を異なる歌手の声質へと変換する技術であり，学習処理と変換処理から構成される．学習時には，入力歌手と目標歌手が同一曲を歌唱した歌声で構成されるパラレルデータを用い，両歌手の音響特徴量の結合確率密度関数を GMM でモデル化する．両歌手の静的・動的特徴量ベクトルをそれぞれ $\boldsymbol{X}_t = \left[\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top\right]^\top$ 及び $\boldsymbol{Y}_t = \left[\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top\right]^\top$ とすると，GMM は以下の式で表される．

$$P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda}\right) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix}\boldsymbol{X}_t \\ \boldsymbol{Y}_t\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)}\end{bmatrix}\right) \quad (1)$$

ここで $\mathcal{N}\left(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ は平均ベクトル $\boldsymbol{\mu}$ 及び共分散行列 $\boldsymbol{\Sigma}$ を持つ正規分布を表す．GMM の混合数は M であり，m は分布番号を示す．α_m は，各分布に対する混合重みを表す．

変換処理では，最尤系列変換法 [2] により，入力歌手の歌声から分析された音響特徴量を，目標歌手の音響特徴量へと変換する．入力歌手と目標歌手の特徴量系列ベクトルを，各々 $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_T^\top]^\top$ と $\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_T^\top]^\top$ とする．ここで，T はフレーム数である．変換される静的特徴量系列 $\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_T^\top]^\top$ は次式で示される．

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}) \text{ subject to } \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}, \quad (2)$$

ここで $\boldsymbol{W}$ は静的特徴量系列を静的・動的結合特徴量系列に拡張する行列である．なお，過剰な平滑化による変換歌声の音質劣化を緩和するため，系列内変動 (GV: Global Variance)[2] を考慮する．

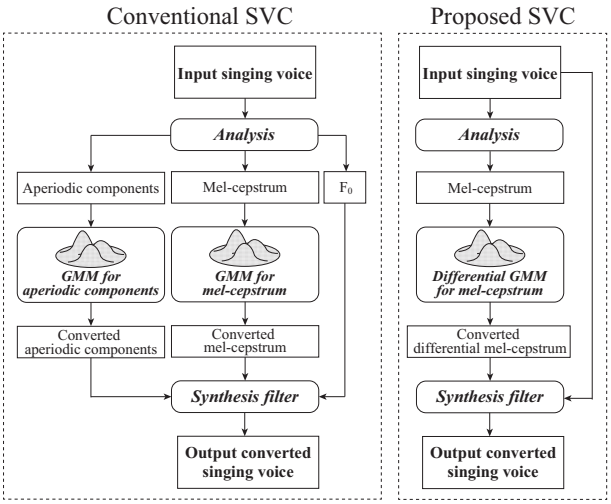

Fig. 1 従来法と提案法の変換処理

3 差分スペクトル補正に基づく SVC

主に同性間の歌声では，同一楽曲において顕著な音高の違いは発生しないため，音高の変換は必要とならない場合が多い．そこで，本稿では，音源特徴量の変換を行わずに，入力歌手と目標歌手のスペクトル特徴量の差分のみを補正する SVC を提案する．図 1 に，従来の SVC（左側）と提案法である差分スペクトル補正に基づく SVC（右側）の変換処理を示す．差分スペクトル補正に基づく SVC では，入力歌手のスペクトル特徴量から，入力歌手と目標歌手のスペクトル特徴量の差分を表す差分スペクトル特徴量を，GMM に基づき推定する．入力歌手の自然歌声波形に対して，差分スペクトル特徴量を合成フィルタにより畳み込むことで，入力歌手の声質を目標歌手の声質へと変換する．ボコーダによる波形合成処理を必要としないため，F_0 分析誤差やスペクトル特徴量の近似誤差を回避することができる．

本稿では，式 (1) の GMM に対して変数変換を行うことで，差分スペクトル特徴量の推定用の GMM を導出する．静的・動的差分特徴量ベクトルを $\boldsymbol{D}_t = \left[\boldsymbol{d}_t^\top, \Delta\boldsymbol{d}_t^\top\right]^\top$ とすると，入力特徴量ベクトルと差分特徴量ベクトルの結合特徴量ベクトルは以下のように表される．

$$\begin{bmatrix}\boldsymbol{X}_t \\ \boldsymbol{D}_t\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}_t \\ \boldsymbol{Y}_t - \boldsymbol{X}_t\end{bmatrix} = \boldsymbol{A}\begin{bmatrix}\boldsymbol{X}_t \\ \boldsymbol{Y}_t\end{bmatrix} \quad (3)$$

ここで $\boldsymbol{A}$ は，目標歌手の特徴量ベクトルを差分特徴

*Statistical Singing Voice Conversion based on Differential Spectral Compensation, by KOBAYASHI, Kazuhiro, TODA, Tomoki, NEUBIG, Graham, SAKTI, Sakriani, NAKAMURA, Satoshi（NAIST）

量ベクトルに変換する行列である．

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ -\boldsymbol{I} & \boldsymbol{I} \end{bmatrix} \tag{4}$$

この行列を式（1）に適用することで，入力特徴量ベクトルと差分特徴量ベクトルの結合確率密度をモデル化する以下の GMM が導出される．

$$P(\boldsymbol{X}_t, \boldsymbol{D}_t|\boldsymbol{\lambda}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix} \boldsymbol{X}_t \\ \boldsymbol{D}_t \end{bmatrix}; \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(D)} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XD)} \\ \boldsymbol{\Sigma}_m^{(DX)} & \boldsymbol{\Sigma}_m^{(DD)} \end{bmatrix} \right) \tag{5}$$

$$\boldsymbol{\mu}_m^{(D)} = \boldsymbol{\mu}_m^{(Y)} - \boldsymbol{\mu}_m^{(X)} \tag{6}$$

$$\boldsymbol{\Sigma}_m^{(XD)} = \boldsymbol{\Sigma}_m^{(DX)^\top} = \boldsymbol{\Sigma}_m^{(XY)} - \boldsymbol{\Sigma}_m^{(XX)} \tag{7}$$

$$\boldsymbol{\Sigma}_m^{(DD)} = \boldsymbol{\Sigma}_m^{(XX)} + \boldsymbol{\Sigma}_m^{(YY)} - \boldsymbol{\Sigma}_m^{(XY)} - \boldsymbol{\Sigma}_m^{(YX)} \tag{8}$$

この GMM に基づき，最尤系列変換法により静的差分特徴量ベクトルを推定する．なお，本稿では，差分スペクトル特徴量の GV については考慮しない．

4 実験的評価

4.1 実験条件

日本語民謡楽曲に対する歌唱データを用いる．楽曲数は 21 曲であり，計 152 フレーズ（各フレーズは 8 秒程度）から構成される．歌手は，男性 3 名，女性 3 名の計 6 名である．学習データとして，ランダムに選出した 80 フレーズを用い，残りをテストデータとして用いる．入力歌手と目標歌手の組み合わせは，同一性別内の総当りとする．被験者は，20 代の学生 8 名である．

スペクトル特徴量として，STRAIGHT 分析 [3] により得られるスペクトル包絡をモデル化したメルケプストラムを用いる．メルケプストラム次数は，1 次から 24 次，1 次から 32 次，1 次から 40 次と変化させる．合成フィルタには，MLSA フィルタ [4] を用いる．シフト長は 5 ms，サンプリング周波数は 16 kHz とする．従来の SVC で用いる音源特徴量として，F_0 と 5 周波数帯域における平均非周期成分を使用する．スペクトル特徴量と非周期成分に対する GMM の混合数はそれぞれ 128，64 である．本実験において F_0 の変換は行わない．

従来法と提案法による変換歌声の自然性を，AB テストにより評価する．従来法および提案法で変換された同一フレーズの歌声サンプルをそれぞれランダムな順序で再生する．どちらの変換歌声が高い自然性を持つかを評価する．また，従来法と提案法に個人性の変換精度を，XAB テストにより評価する．目標歌手の自然歌声を参照歌声とし，従来法と提案法の変換歌声をランダムな順序で再生する．どちらの変換歌声が目標歌手の自然歌声に似ているかという基準で評価する．なお，各被験者は，両実験共に 72 対のフレーズに対し，それぞれ評価を行う．

4.2 実験結果

図 2 に AB テストによる変換歌声の自然性に関する評価結果を示す．従来法と比べて，提案法はより自然性の高い変換歌声を得られることが分かる．これは，ボコーダ使用に伴う F_0 分析誤差やスペクトルモデリング誤差の影響を提案法では回避しており，入力歌声の情報を上手く活用できているためである．

図 3 に XAB テストによる変換歌声の個人性に関する評価結果を示す．メルケプストラムの次数に依存せず，提案法と従来法ではほぼ同等の個人性変換精度が得られることが分かる．以上の結果から，提案法は従来法よりも有効であることが分かる．なお，被験者からは，サンプルによっては，入力歌手と目標歌手の歌いまわしの違いが大きく，従来法および提案法の両手法とも，変換歌声が目標歌手にあまり似ていない場合があるという感想が得られている．この原因として，スペクトル特徴量や非周期成分などの分節的特徴と比べて，F_0 やパワーなどの韻律的特徴の方が，より個人性に大きな影響を与える点 [5] が考えられる．

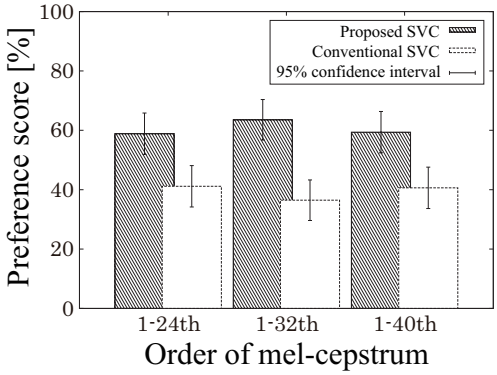

Fig. 2　自然性に関する評価結果

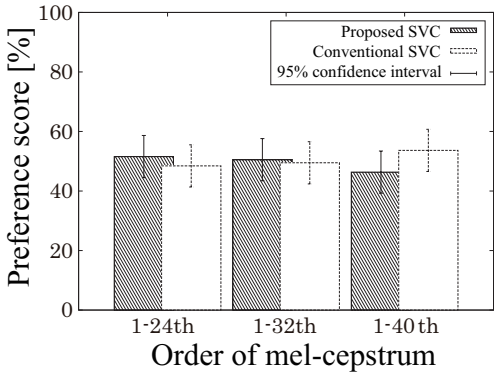

Fig. 3　個人性に関する評価結果

5 まとめ

統計的手法に基づく歌声声質変換において，差分スペクトル補正に基づく変換法を提案した．実験結果より，従来法に比べ提案法は，高い自然性を保ちつつ同等の個人性変換精度を達成できることを示した．今後の研究として，差分スペクトル特徴量に対する GV の検討や変換精度向上に取り組む．

謝辞　本研究の一部は，JSPS 科研費 22680016 および JST OngaCREST プロジェクトの助成を受け実施したものである．

参考文献

[1] H. Doi *et al.*, Proc. APSIPA ASC, 2012.
[2] T. Toda *et al.*, IEEE Trans. ASLP, Vol. 15, No. 8, pp. 2222–2235, 2007.
[3] H. Kawahara *et al.*, Speech Communication, Vol. 27, No. 3–4, pp. 187–207, 1999.
[4] 今井聖 他, 信学論 (A), Vol. J66-A, No. 2, pp. 122–129, 1983.
[5] 小林和弘 他, 情報処理研報, Vol.2013–MUS–99 No.44, pp. 1–6, 2013.