系列内変動を考慮した差分スペクトル補正に基づく短遅延歌声声質変換*

◎小林 和弘（奈良先端大），戸田 智基（名大・奈良先端大），中村 哲（奈良先端大）

1 はじめに

入力歌手の歌声から目標歌手の歌声へと高品質な声質変換を実現する手法として，系列内変動（GV: Global Variance）を考慮した差分スペクトル補正に基づく統計的歌声声質変換（SVC: Singing Voice Conversion）（以下，差分 SVC）が提案されている [1]．この手法では，入力歌声に対し，混合正規分布モデル（GMM: Gaussian Mixture Model）および GV に対する正規分布により推定される差分スペクトル系列を用いて，時変フィルタリングによる補正処理を行う事で，声質の変換を実現する．GV を考慮することで，変換歌声の品質を改善できるが，勾配法による繰り返し処理が必要となるため，短遅延声質変換 [2] によるリアルタイム変換処理を実現するのが困難となる．

本稿では，短遅延差分 SVC に適用可能な GV 補償ポストフィルタ処理を提案する．実験結果より，提案法は GV を補償しない差分 SVC と比べ，高い自然性と同等の個人性変換精度が得られる事を示す．

2 GV を考慮した差分 SVC

差分 SVC は，ボコーダによる音源波形の生成処理を用いずに，声質を変換する手法である．本処理は，学習処理と変換処理から構成される．

学習処理では，入力歌手と目標歌手のパラレルデータを用いて，入力歌手と目標歌手のスペクトル特徴量の結合確率密度関数を GMM によりモデル化する．得られた GMM に対して変数変換を施すことで，入力歌手のスペクトル特徴量と差分スペクトル特徴量の結合確率密度関数をモデル化する差分 GMM を求める．ここで，フレーム t における両歌手の静的・動的特徴量をそれぞれ $\boldsymbol{X}_t = \left[\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top\right]^\top$ 及び $\boldsymbol{Y}_t = \left[\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top\right]^\top$ とし，また，静的・動的差分特徴量を $\boldsymbol{D}_t = [\boldsymbol{Y}_t - \boldsymbol{X}_t]$ とする．差分 GMM による結合確率密度関数は以下の式で表される．

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

ここで $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ 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．$\boldsymbol{\lambda}$ は，GMM のパラメータセットを表す．また，目標歌手の静的特徴量系列に対する GV は，以下の式で表される．

$$\boldsymbol{v}(\boldsymbol{y}) = [v_1, v_2, \cdots, v_d, \cdots, v_D]^\top \quad (2)$$

$$v_d = \frac{1}{T}\sum_{t=1}^{T}\left(y_{t,d} - \bar{y}_d\right)^2 \quad (3)$$

$$\bar{y}_d = \frac{1}{T}\sum_{\tau=1}^{T} y_{\tau,d} \quad (4)$$

ここで，$y_{t,d}$ は，フレーム t における d 次元目の目標歌手の静的特徴量である．GV に対する確率密度関数は，正規分布によりモデル化される．

$$P\left(\boldsymbol{v}(\boldsymbol{y}) \mid \boldsymbol{\lambda}^{(v)}\right) = \mathcal{N}\left(\boldsymbol{v}(\boldsymbol{y}); \boldsymbol{\mu}^{(v)}, \boldsymbol{\Sigma}^{(vv)}\right) \quad (5)$$

ここで，$\boldsymbol{\lambda}^{(v)}$ は，平均ベクトル $\boldsymbol{\mu}^{(v)}$ 及び共分散行列 $\boldsymbol{\Sigma}^{(vv)}$ を持つ正規分布のパラメータセットである．

変換処理では，最尤系列変換法 [4] により，入力歌手の静的・動的特徴量系列を，静的差分特徴量系列へと変換する．入力歌声の静的・動的特徴量系列を $\boldsymbol{X}' = [\boldsymbol{X}_1'^\top, \cdots, \boldsymbol{X}_T'^\top]^\top$，変換歌声の静的特徴量系列を $\boldsymbol{y}' = [\boldsymbol{x}' + \hat{\boldsymbol{d}}]$ とすると，静的差分特徴量系列 $\boldsymbol{d}' = [\boldsymbol{d}_1'^\top, \cdots, \boldsymbol{d}_T'^\top]^\top$ は，次式の目的関数を最大化する系列として，勾配法を用いて推定される．

$$\hat{\boldsymbol{d}} = \underset{\boldsymbol{d}}{\operatorname{argmax}}\, P(\boldsymbol{D}|\boldsymbol{X}', \boldsymbol{\lambda})^\omega P(\boldsymbol{v}(\boldsymbol{y}')|\boldsymbol{\lambda}^{(v)}) \quad \text{s.t. } \boldsymbol{D} = \boldsymbol{W}\boldsymbol{d} \quad (6)$$

ここで，ω は 2 つの尤度関数の影響を調整するパラメータである．また，W は，静的特徴量系列を静的・動的特徴量系列へと変換する行列である．なお，GV を考慮せずに GMM の条件付き確率密度関数のみを最大化する際には，解析解が存在し，短遅延差分 SVC が可能となる．

入力歌声波形に対して推定された静的差分特徴量系列を補正する時変フィルタリング処理を施す事で，声質の変換を行う．

3 短遅延差分 SVC のための GV ポストフィルタ

3.1 同一歌手 SVC による変換スペクトル特徴量を用いたポストフィルタ

差分 SVC において GV を補償する際には，目標歌手のスペクトル特徴量系列の GV（すなわち変換スペクトル特徴量系列に相当する GV）を補償する差分スペクトル特徴量系列を推定する必要がある．そのため，変換スペクトル特徴量系列を差分スペクトル特徴量系列で表す必要がある．入力歌声に対して，STRAIGHT 分析 [5] などの高品質な分析系を用いてスペクトル特徴量の抽出を行う場合は，前節で述べた方法で，容易に変換スペクトル特徴量系列を求めることができる．一方で，リアルタイム変換 [2] では，計算量削減のため，固定の分析窓を用いた高速フーリエ変換とリフタリングによる単純な分析処理が用いられる．分析精度が低いため，得られるスペクトル特徴量系列は音源の周期構造の影響を受けやすく，変換スペクトル特徴量系列の推定精度の低下を招く．その結果，GV の補償効果の低下や変換音声の品質劣化が生じる傾向にある．

本稿では，この問題を緩和する手法として，同一歌手 SVC[3] による入力スペクトル特徴量系列の推定を用いた GV ポストフィルタ処理を提案する．ここで，同一歌手 SVC とは，入力歌手と目標歌手の結合確率密度関数に対し変数変換を施す事で，入力歌手から入力歌手への特徴量系列の変換を実現する枠組みである．本枠組みを拡張することで，単純な分析処理により得られる入力スペクトル特徴量系列から，STRAIGHT などの高度な分析処理により得られる入

*Low Delay Statistical Singing Voice Conversion based on Direct Waveform Modification Considering Global Variance, KOBAYASHI, Kazuhiro（NAIST），TODA, Tomoki（Nagoya University/NAIST），NAKAMURA, Satoshi（NAIST）

力スペクトル特徴量系列を近似的に推定することが可能となる．

フレーム t における d 次元目の GV を考慮しない差分 SVC による静的差分特徴量を $\hat{d}_{t,d}$，同一歌手 SVC による静的特徴量を $\hat{x}_{t,d}$ とすると，提案するポストフィルタ処理は以下の式で示される．

$$\hat{d}_{t,d}^{(GV)} = \mu_d^{(v)\frac{1}{2}} \bar{\mu}_d^{(v)-\frac{1}{2}} (\hat{x}_{t,d} + \hat{d}_{t,d} - \bar{y}_d) + \bar{y}_d - \hat{x}_{t,d} \quad (7)$$

ここで，$\mu_d^{(v)}$ は，d 次元目の目標歌手の静的特徴量系列の GV であり，$\bar{\mu}_d^{(v)}$ および $\bar{y}_d$ は，予め GV を考慮しない差分 SVC による変換歌声から分析された d 次元目の静的特徴量系列に対する GV と平均である．なお，提案法は，差分 SVC による静的差分特徴量の推定と同一歌手 SVC による入力歌手の静的特徴量を推定するために，2 つの変換を同時に必要があるが，どちらも短遅延変換処理を適用することが可能である．

3.2 無声音に対する変換処理の回避

無声音は，有声音に比べて，個人性知覚に対する寄与が小さい [6]．そのため，GV を考慮した差分 SVC では，無声音フレームに対する確率密度関数を修正する事で変換を抑圧する差分特徴量系列を推定し，無声音フレームの変換に伴う品質劣化を回避する．本稿では，類似の処理をポストフィルタ処理として導入する．無声音フレームに対しては，推定された差分特徴量の値を零とする事で，変換を抑圧する．

4 実験的評価

4.1 実験条件

歌声データベースとして，日本語民謡楽曲を用いる．楽曲数は 21 曲，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する．入力歌手と目標歌手の組み合わせは，同性間の総当りとする．被験者は，20 代の学生 6 名である．

シフト長は 5 ms，サンプリング周波数は 16 kHz とする．スペクトル特徴量として，STRAIGHT 分析 [5] により得られるスペクトル包絡をモデル化した 1 次から 24 次のメルケプストラムを用いる．差分スペクトルを補正するための合成フィルタには，MLSA フィルタ [7] を用いる．スペクトル特徴量の GMM の混合数は，128 である．なお，短遅延変換 [2] と最尤系列変換 [4] のスペクトル特徴量の変換精度は同等である事より，本実験では，代替的な実験として最尤系列変換によるスペクトル特徴量の変換を行う．

提案法である GV を補償するポストフィルタを適用した差分 SVC（以下，“w/ GVPF”）と GV を考慮しない差分 SVC（以下，“w/o GVPF”）を比較する．まず，変換歌声の音質を，AB テストにより評価する．同一フレーズの変換歌声をそれぞれランダムな順序で再生し，どちらの変換歌声が高い音質を持つかを評価する．また，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，同一フレーズの 2 つ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．被験者毎の評価サンプル数は，両実験共に 32 である．

4.2 実験結果

図 1(a) に AB テストによる変換歌声の音質に関する評価結果を示す．“w/ GVPF” は，“w/o GVPF” と比べてより音質の高い変換歌声が得られる事がわかる．

図 1(b) 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．“w/ GVPF” は，“w/o GVPF” とほぼ同等の個人性変換精度が得られる事がわかる．

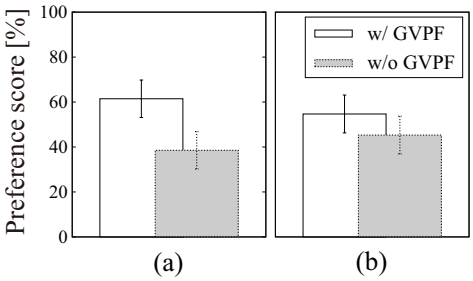

Fig. 1: Results of preference tests on (a) speech quality of converted singing voice and (b) conversion accuracy on singer individuality.

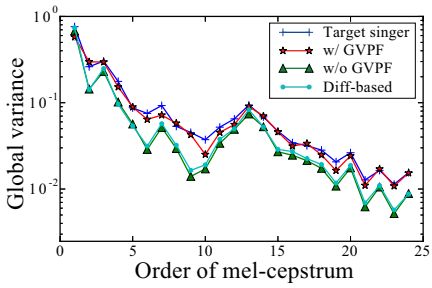

Fig. 2: GVs of mel-cepstral sequences of converted voices.

図 2 に変換歌声から分析されたメルケプストラム系列の GV を示す．“Diff-based” は，推定された静的差分特徴量に対して，静的差分特徴量の GV をポストフィルタ処理によって補償した変換歌声の GV である．静的差分特徴量に対する GV の補償では，変換歌声の GV は補償されていない事がわかる．一方で，“w/ GVPF” は，“w/o GVPF” に比べて，GV が補償されている事がわかる．

5 まとめ

本稿では，差分スペクトル補正に基づく短遅延 SVC の品質を改善するため，GV を補償するポストフィルタ処理を提案した．実験結果より，提案法は従来法に比べ，高い自然性と同等の個人性変換精度を実現する事がわかった．今後は，差分 SVC における非周期成分の変換処理に取り組む．

謝辞 本研究の一部は，JSPS 科研費 26280060 および OngaCREST の助成を受け実施したものである．

参考文献

[1] K. Kobayashi *et al.*, Proc. INTERSPEECH, 2015.

[2] T. Toda *et al.*, Proc, INTERSPEECH, pp. 94–97 .2012.

[3] K. Kobayashi *et al.* IEICE Trans. on Inf. and Syst., Vol. 97, No. 6, pp. 1419–1428, 2014.

[4] T. Toda *et al.*, IEEE Trans. ASLP, Vol. 15, No. 8, pp. 2222–2235, 2007.

[5] H. Kawahara *et al.*, Speech Communication, Vol. 27, No. 3–4, pp. 187–207, 1999.

[6] M. Sambur, IEEE Trans. ASSP, Vol. 23, No. 2, pp. 176–182, 1975.

[7] 今井聖 他, 信学論 (A), Vol. J66-A, No. 2, pp. 122–129, 1983.