差分スペクトル補正による歌声声質変換のための F_0 変換に関する検討*

◎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

入力歌手の声質を目標歌手の声質へと変換する手法として，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]．差分 SVC は，入力歌声に対し，差分混合正規分布モデル（GMM: Gaussian Mixture Model）により推定された差分スペクトルを用いて，入力歌声波形に対する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．一方で，音源情報は保持されるため，変換可能な歌手の組み合わせは，同一性別内に制限されており，異性間の差分 SVC は実現していない．

本稿では，差分 SVC に対して，時間波形上での処理に基づく音源情報変換を導入することで，異性間の差分 SVC を実現する手法を提案する．実験結果より，音源情報変換を導入した差分 SVC は，従来の SVC と比較し，個人性変換精度を保持しつつ，変換歌声の音質を改善できることを示す．

2 差分 SVC

差分 SVC は，ボコーダによる波形合成処理を用いずに，入力歌手の声質を異なる歌手の声質へと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．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．また，差分スペクトル特徴量を $\boldsymbol{D}_t = [\boldsymbol{Y}_t - \boldsymbol{X}_t]$ とすると，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．

変換処理では，最尤系列変換法 [2] により，入力歌手のスペクトル特徴量系列を，差分スペクトル特徴量系列へと変換する．なお，その際には，スペクトル特徴量に対する系列内変動を考慮する [1]．得られた差分スペクトル特徴量系列に基づき，入力歌声波形に対してフィルタリング処理を施すことで，声質の変換を行う．ボコーダによる波形合成処理を回避することで，高い音質改善効果が得られる一方で，音源情報の変換が行われないため，例えば異性間の SVC のように，音高の変換も必要となる際には使用できない．

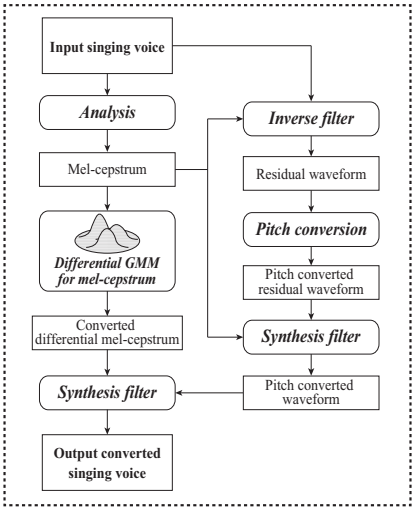

Fig. 1: Conversion process considering F_0 conversion in SVC based on spectral differential.

3 差分 SVC における F_0 変換

本稿では，変換対象となる音源情報として F_0 に着目し，差分 SVC の利点を保持しつつ，F_0 変換を実現する手法を提案する．提案法では，入力歌声の残差波形に対して時間軸上で処理を施すことで，ボコーダによる音源波形合成処理を用いずに F_0 変換を行う．提案法における処理の流れを図 1 に示す．

3.1 残差波形に対する F_0 変換

入力歌声に対して，スペクトル特徴量に基づく逆フィルタリングを行うことで，残差波形を抽出し，WSOLA[3] による継続長変換と，リサンプリングによる時間軸の伸長を実施することで，F_0 変換を行う．例えば，F_0 を 2 倍に変換する際には継続長を 2 倍に変換する処理と 0.5 倍のダウンサンプリング処理を行い，F_0 を 0.5 倍に変換する際には継続長を 0.5 倍に変換する処理と 2 倍のアップサンプリング処理を行う．上記の処理は，周波数軸を定数倍して，スペクトルを伸長させることに相当する．F_0 変換後の残差波形に対して，残差波形抽出時に用いたスペクトル特徴量に基づくフィルタリングを行うことで，F_0 変換が施された入力歌声が得られる．

3.2 高域周波数成分の復元

前節の処理で F_0 を下げる変換を行う際（主に女性から男性への歌声声質変換）に，F_0 変換後の入力歌声音声の高域周波数帯域の成分が欠落する．例えば，16 kHz サンプリングの入力歌声音声に対して，F_0 を 0.5 倍に変換すると，4 kHz 以上の周波数成分が欠落する．そのため，F_0 変換後の入力歌声の音質は著しく劣化する．

---

* An Investigation of F_0 Conversion in Statistical Singing Voice Conversion based on Spectral Differential, by KOBAYASHI, Kazuhiro, TODA, Tomoki, NEUBIG, Graham, SAKTI, Sakriani, NAKAMURA, Satoshi (NAIST)

この問題を解決するために，高域周波数成分の復元処理を導入する．高域周波数成分は非周期成分が支配的である点に着目し，雑音源を利用する．具体的には， 雑音源に対して入力歌声のスペクトル特徴量に基づくフィルタリングを行うことで，雑音源に基づく入力歌声波形を生成し，高域通過フィルタを通すことで，高域周波数成分を生成する．それを，低域通過フィルタを通した F_0 変換後の入力歌声波形と足し合わせることで，高域周波数成分が復元された F_0 変換後の入力歌声波形を生成する．

3.3 低域周波数帯域における非周期成分の低減

3.2 節に述べた処理で F_0 を下げる変換を行うと，入力歌声の高域周波数成分が低域周波数成分へと移動する．結果として，低域周波数成分の非周期成分が大きくなり，F_0 変換後の入力歌声の音質が劣化する要因となる．

本稿では，非周期成分を低減させるために，フィードバック型くし型フィルタを用いて，調波成分間の周波数成分を抑圧する．時刻 t におけるくし形フィルタの伝達関数は，以下の式で表される．

$$H_t(z) = \frac{1-a}{1-az^{-(f_s/f_t)}} \tag{2}$$

ここで，a は，くし形フィルタの周波数応答を決める係数である．f_s は，入力歌声のサンプリング周波数であり，f_t は，時刻 t における F_0 変換後の基本周波数である．このくし形フィルタを用いることで，非周期成分を低減させる．なお，無声音のフレームに対しては，くし形フィルタによる非周期成分抑圧は行わない．

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する．入力歌手と目標歌手の組み合わせは，異性間の総当りとする．被験者は，20 代の学生 8 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 [4] 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．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．合成フィルタには，MLSA フィルタ [5] を用いる．スペクトル特徴量および非周期成分の GMM の混合数は，それぞれ 32, 16 である．F_0 変換は，従来法と提案法共に，男性から女性は 2 倍，女性から男性は 0.5 倍とする．低域通過フィルタおよび高域通過フィルタは，カットオフ周波数が 4 kHz，256 次の有限インパルス応答フィルタを用いる．くし形フィルタの係数 a は，0.6 である．

提案法である F_0 変換を導入した差分 SVC と，従来法であるボコーダを用いた SVC を比較する．まず，変換音声の音質を，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．被験者毎の各組み合わせに対する評価数は，両実験それぞれ 36 である．

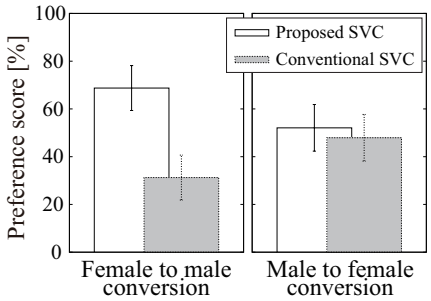

Fig. 2: Speech quality of converted singing voice.

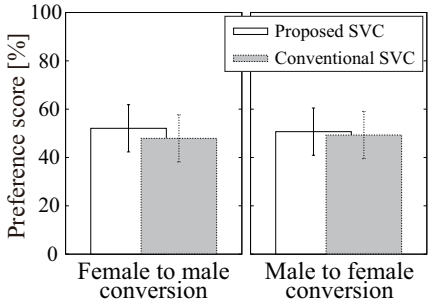

Fig. 3: Conversion accuracy of singer individuality.

4.2 実験結果

図 2 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．提案法は，女性から男性への変換において，従来法と比べてより音質の高い変換歌声が得られる．なお，男女間で音質改善効果が異なる傾向があるが，これは F_0 変換を行わない同性間における差分 SVC[1] においても同様の傾向がみられる．

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．提案法は，従来法とほぼ同等の個人性変換精度が得られる事がわかる．

5 まとめ

本稿では，残差波形に対する時間軸上での F_0 変換処理により，異性間の差分 SVC を実現する手法を提案した．実験結果より，F_0 変換を用いた異性間の差分 SVC は，従来の SVC に比べ，より高音質な変換歌声を得られる事がわかった．今後は，ボコーダによる音源生成を必要としない枠組みの中で，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.*, IEEE Trans. ASLP, Vol. 15, No. 8, pp. 2222–2235, 2007.

[3] Verhelst *et al.*, Proc. ICASSP, Vol. 2, pp. 554–557, 1993.

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

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