差分スペクトル補正に基づく歌声声質変換における パラメータ生成法に関する調査*

◎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，時間波形上で補正処理を行う事で，声質の変換を実現する．一方で，変換歌声のスペクトル特徴量は，差分GMMのモデリング誤差により平滑化されるため，入力歌声に比べて音質の劣化が生じる．

本稿では，差分SVCにおいて，変換歌声のスペクトル特徴量の平滑化を回避するパラメータ生成法に関して調査を行う．まず，変換歌声のスペクトル特徴量の系列内変動（GV: Global variance）を考慮した差分スペクトル特徴量のパラメータ生成法を提案する．また，静的特徴量系列空間における差分特徴量に基づくパラメータ生成法を提案する．実験結果より，両提案法による差分SVCは，従来の差分SVCに比べ，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，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り，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差分GMMによりモデル化する．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．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．

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

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．なお，差分GMMは，結合確率密度関数 $P(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda})$ に対するGMMから解析的に求める事が出来る [1]．

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を，差分スペクトル特徴量へと変換する．入力特徴量系列ベクトルと差分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 $\boldsymbol{D} = [\boldsymbol{D}_1^\top, \cdots, \boldsymbol{D}_T^\top]^\top$ とする．ここで，T はフレーム数である．静的差分特徴量系列ベクトル $\hat{\boldsymbol{d}} = [\hat{\boldsymbol{d}}_1^\top, \cdots, \hat{\boldsymbol{d}}_T^\top]^\top$ は，次式で示される．

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

ここで，各時刻における確率密度関数は

$$P(\boldsymbol{D}_t|m, \boldsymbol{X}_t, \boldsymbol{\lambda}) = \mathcal{N}\left(\boldsymbol{D}_t; \boldsymbol{E}_{m,t}^{(D)}, \boldsymbol{V}_m^{(D)}\right) \quad (3)$$

$$\boldsymbol{E}_{m,t}^{(D)} = \boldsymbol{\mu}_m^{(D)} + \boldsymbol{\Sigma}_m^{(DX)}\boldsymbol{\Sigma}_m^{(XX)^{-1}}\left(\boldsymbol{X}_t - \boldsymbol{\mu}_m^{(X)}\right) \quad (4)$$

$$\boldsymbol{V}_m^{(D)} = \boldsymbol{\Sigma}_m^{(DD)} - \boldsymbol{\Sigma}_m^{(DX)}\boldsymbol{\Sigma}_m^{(XX)^{-1}}\boldsymbol{\Sigma}_m^{(XD)} \quad (5)$$

で表される．また，$\boldsymbol{W}$ は静的特徴量ベクトルを静的･動的結合特徴量ベクトルに拡張する行列である．

3 差分SVCにおけるパラメータ生成法

3.1 GVを考慮したパラメータ生成法

差分SVCにおいて，変換歌声のスペクトル特徴量の平滑化を回避するために，GVを考慮したパラメータ生成法を提案する．目標歌手の静的特徴量系列ベクトルを $\boldsymbol{y} = [\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_T^\top]^\top$ とすると，目標歌手の静的特徴量に対するGVは，以下の式で表される．

$$\boldsymbol{v}(\boldsymbol{y}) = [v(1), v(2), \cdots, v(d), \cdots, v(D)]^\top \quad (6)$$

$$v(d) = \frac{1}{T}\sum_{t=1}^{T}\left(y_t(d) - \bar{y}(d)\right)^2 \quad (7)$$

$$\bar{y}(d) = \frac{1}{T}\sum_{\tau=1}^{T} y_\tau(d) \quad (8)$$

ここで $y_t(d)$ は，フレーム t における d 次元目の静的特徴量である．また，GVの確率密度関数を正規分布によりモデル化する．

$$P\left(\boldsymbol{v}(\boldsymbol{y}) \mid \boldsymbol{\lambda}^{(v)}\right) = \mathcal{N}\left(\boldsymbol{v}(\boldsymbol{y}); \boldsymbol{\mu}^{(v)}, \boldsymbol{\Sigma}^{(vv)}\right) \quad (9)$$

入力歌声の静的特徴量ベクトルを $\boldsymbol{x} = [\boldsymbol{x}_1^\top, \cdots, \boldsymbol{x}_T^\top]^\top$，変換歌声の静的特徴量ベクトルを $\hat{\boldsymbol{y}} = [\boldsymbol{x} + \hat{\boldsymbol{d}}]$ とすると，GVを考慮した静的差分特徴量系列のパラメータ生成処理は，次式で表される．

$$\hat{\boldsymbol{d}} = \underset{\boldsymbol{d}}{\operatorname{argmax}}\, P(\boldsymbol{D}|\boldsymbol{X}, \boldsymbol{\lambda}) P(\boldsymbol{v}(\hat{\boldsymbol{y}})|\boldsymbol{\lambda}^{(v)})^\omega \quad (10)$$

ここで ω は，尤度間の重みを調整するパラメータである．静的差分特徴量ベクトルは，勾配法により求める．

3.2 無声音に対する差分特徴量系列の平滑化

GVを考慮した差分SVCでは，無声音フレームにおいて，パラメータ系列の急須な変動に伴う変換歌声の音質劣化が生じる．一方で，差分SVCでは，個人性に与える影響が小さいと考えられる無声音フレームに対して，必ずしも高精度な変換処理を必要しない．そこで，差分特徴量系列に対する平滑化処理を導入することで，変換歌声の音質劣化を回避する．本稿では，無声フレームに対する確率密度関数のパラ

---

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

メータを以下のように修正することで，平滑化処理を実現する．

$$\boldsymbol{E}_{m,t}^{(D)} = \{\ \boldsymbol{0}\ \ (static\,\&\,delta) \tag{11}$$

$$\boldsymbol{V}_{m}^{(D)} = \begin{cases} \boldsymbol{\infty} & (static) \\ \Delta\boldsymbol{v}_{m}^{(D)} & (delta) \end{cases} \tag{12}$$

ここで $\boldsymbol{\infty}$ は無限大であり，$\Delta\boldsymbol{v}_{m}^{(D)}$ は，式 (5) の動的成分である．また，無声音フレームに対しては，勾配法によるパラメータ更新を行わない．

3.3 トラジェクトリ差分スペクトル特徴量に基づく差分 SVC

差分 GMM では，静的・動的差分特徴量空間における確率密度関数に基づき，静的差分特徴量系列が生成される．本稿では，さらに，静的差分特徴量系列空間における確率密度関数に基づくパラメータ生成法を提案する．静的特徴量系列空間における確率密度関数は，トラジェクトリモデル [3] として表現され，その平均ベクトルは式 (2) における $\hat{\boldsymbol{d}}$ により表される．そのため，静的差分特徴量系列空間における確率密度関数の平均ベクトルは，入力歌手の同一歌手 SVC[4] による変換特徴量系列ベクトルを $\boldsymbol{x}' = [\hat{\boldsymbol{x}}_1^\top, \cdots, \hat{\boldsymbol{x}}_T^\top]^\top$，SVC による入力歌手から目標歌手への変換特徴量系列ベクトルを $\boldsymbol{y}' = [\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_T^\top]^\top$ とすると，$\boldsymbol{d}' = [\boldsymbol{y}'_t - \boldsymbol{x}'_t]$ として表される．この静的差分特徴量系列 $\boldsymbol{d}'$ に基づき，入力歌声に対する補正を行う．

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，同一性別内の総当りとする．被験者は，20 代の学生 6 名である．

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．合成フィルタには，MLSA フィルタ [6] 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．スペクトル特徴量の GMM の混合数は 128 である．

実験では，以下の変換歌声に対して評価を行う．

- w/o GV: 従来の差分 SVC
- w/ GV: 3.1 節と 3.2 節の差分 SVC
- TrjDiff: 3.3 節の差分 SVC

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．なお，両実験共に 3 手法間の組み合わせに対し評価を行う．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 24 である．

4.2 実験結果

図 1 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．従来法と比べて，提案法はより音質の高い変換歌声を得られることが分かる．

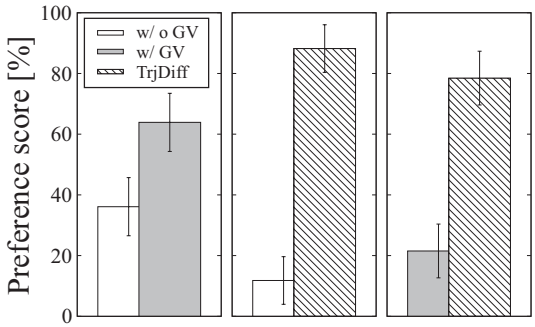

Fig. 1 Speech quality of converted singing voice

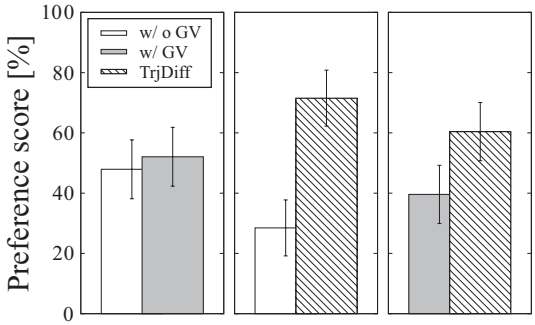

Fig. 2 Conversion accuracy of singer individuality

図 2 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．従来法と比べて，TrjDiff はより高いスコアが得られる．ただし，TrjDiff は他の手法に比べ音質が非常に高いことから，音質の面で自然歌声である参照歌声に最も似ていると判断された可能性が懸念される．そのため，個人性変換精度については，さらなる評価が必要である．

5 まとめ

差分 SVC において，パラメータ生成法に関する調査を行った．実験結果より，提案法である GV を考慮した差分 SVC とトラジェクトリ差分スペクトルに基づく差分 SVC は，従来の差分 SVC に比べ，より高音質な変換歌声が得られる事がわかった．今後の研究として，差分 SVC において異性間での声質変換に関する研究を行う．

謝辞 本研究の一部は，JSPS 科研費 26280060 および 24300073 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, pp. 2514-2518, 2014.

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

[3] H. Zen *et al.*, Computer Speech & Language, Vol. 21, No. 1, pp. 153–173, 2007.

[4] 小林和弘 他, 情報処理研報, Vol.2013–MUS–99 No.44, pp. 1–6, 2013.

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

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