歌声合成システムの音源データに対する声質評価値に基づく声質制御 *

☆山根 壮一, 小林 和弘 (奈良先端大・情報), 戸田 智基 (名大／奈良先端大),
中野 倫靖，後藤 真孝 (産総研), 中村 哲 (奈良先端大・情報)

1 はじめに

近年，楽曲製作において，VOCALOID [1] や UTAU [2] などの歌声合成システムが盛んに利用されている．歌声合成システムは，音源データの入れ替えにより，所望の歌手の歌声を合成する事が可能である．我々は，楽曲製作を支援する上で，膨大な種類の音源データから，所望の声質を持つ音源データを選択するために，音源データから主観的な声質評価値を推定する枠組みを提案し，その有効性を示した [3]．さらなる支援に向けて，既存の音源データを選択するのみでなく，さらに音源データの声質を自由に操作できる機能の実現が期待される．

本稿では，歌声合成システムの声質操作機能を拡張するために，音源データに対する声質制御法を提案する．統計的歌声声質変換技術を応用することで，主観的な声質評価値に基づく音源データの声質制御を実現する．歌声合成音声の品質および声質に関する実験的評価結果より，提案法の有効性を示す．

2 声質表現語に基づく音源データの声質制御

本稿では，歌声合成システムの各音源データの特徴を，“年齢” や “力強さ” などの声質表現語，および，それらに関する主観的な評価値である声質評価値を用いて表す．統計的歌声声質変換技術 [4] を応用し，声質評価値に基づき変換先の声質を設定可能な変換関数を構築する事で，歌声合成システムの声質制御を実現する．

2.1 修正重回帰混合正規分布モデルに基づく差分スペクトル補正による声質制御

声質表現語に基づく統計的歌声声質制御法 [5] は，学習処理と変換処理で構成される．

学習処理では，一人の参照歌手と複数の事前収録目標歌手が同一楽曲を歌唱した歌声で構成されるパラレルデータを用いて，重回帰混合正規分布モデル（MR-GMM: Multiple-Regression Gaussian Mixture Model）を学習する．次に，学習された MR-GMM に対して，変数変換を施す事で，差分スペクトル補正のための差分 MR-GMM を求める．ここで，事前収録目標歌手を S 人とし，s 番目の歌手の静的・動的特徴量ベクトルを $\boldsymbol{Y}_t(s) = \left[\boldsymbol{y}_t^\top(s), \Delta\boldsymbol{y}_t^\top(s)\right]^\top$，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，事前収録目標歌手 s に対する差分 MR-GMM は，以下の式で表される．

$$P\left(\boldsymbol{Y}_t(s), \boldsymbol{D}_t|\boldsymbol{\lambda}^{(DIFFMR)}, \Delta\boldsymbol{w}\right) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix}\boldsymbol{Y}_t(s)\\ \boldsymbol{D}_t\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}_m^{(Y)}(s)\\ \Delta\boldsymbol{\mu}_m\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_m^{(YY)} & \boldsymbol{\Sigma}_m^{(YD)}\\ \boldsymbol{\Sigma}_m^{(DY)} & \boldsymbol{\Sigma}_m^{(DD)}\end{bmatrix}\right) \quad (1)$$

$$\Delta\boldsymbol{\mu}_m = \boldsymbol{B}_m^{(Y)}\Delta\boldsymbol{w} \quad (2)$$

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 は分布番号を表す．α_m は m 番目の分布の混合重みである．$\boldsymbol{B}_m^{(Y)}$ は，S 人の事前収録歌手の平均ベクトルと声質評価値に対する重回帰分析によって得られた回帰ベクトルセットを表す．$\Delta\boldsymbol{w}$ は，声質を制御するための差分声質評価値ベクトルを表し，入力歌手に対する声質の変化量を定める．また，$\boldsymbol{\mu}_m^{(Y)}(s)$ は，s 番目の事前収録目標歌手依存平均ベクトルを表す．この差分 MR-GMM の入力平均ベクトルを，入力歌手依存平均ベクトル $\hat{\boldsymbol{\mu}}_m(k)$ に置き換えることで，修正差分 MR-GMM が得られる．なお，$\hat{\boldsymbol{\mu}}_m(k)$ は，参照歌手と入力歌手 k のパラレルデータを用いた最尤推定により求める．

変換処理では，修正差分 MR-GMM 及び手動で設定する差分声質評価値ベクトル $\Delta\boldsymbol{w}$ に基づき，与えられた入力歌手 k の入力スペクトル特徴量系列に対して，最尤系列変換法 [6] を用いて差分スペクトル特徴量系列を推定する [7]．入力歌手 k の静的・動的特徴量系列を $\boldsymbol{Y}(k) = \left[\boldsymbol{Y}_1^\top(k), \cdots, \boldsymbol{Y}_t^\top(k)\right]^\top$ とすると，差分静的特徴量系列 $\hat{\boldsymbol{d}} = \left[\hat{\boldsymbol{d}}_1^\top, \cdots, \hat{\boldsymbol{d}}_t^\top\right]^\top$ は，以下の式により推定される．

$$\hat{\boldsymbol{d}} = \arg\max_{\boldsymbol{d}} P\left(\boldsymbol{D}|\boldsymbol{Y}(k), \boldsymbol{\lambda}^{(DIFFMR)}, \Delta\boldsymbol{w}\right) \quad (3)$$

なお，差分特徴量系列を推定する際には，系列内変動を考慮する [7]．推定された差分特徴量系列（差分スペクトル包絡パラメータ系列）に基づき，時変フィルタリングにより入力歌声を補正する事で，入力歌声の声質を制御する．

2.2 カーネル回帰の導入

修正差分 MR-GMM に基づく声質制御は，事前収録目標歌手の平均ベクトルと声質評価値の対応関係に対して線形性を仮定して，声質評価値の変化に伴う声質変化をモデル化する．一方で，歌声音声データからの声質評価値推定の研究において，声質表現語の種類によっては，非線形性を考慮することで，推定精度の改善が得られることが報告されている [3]．同様の効果が歌声声質制御でも期待されるため，非線形回帰分析の導入を試みる．なお，本稿では，非線形回帰分析法として，カーネル回帰を用いる．

カーネル回帰では，s 番目の事前収録目標歌手に対する平均ベクトル $\boldsymbol{\mu}_m(s)$ は，次式で表される．

$$\boldsymbol{\mu}_m(s) = \boldsymbol{V}_m\phi(\boldsymbol{w}(s)) \quad (4)$$

ここで，$\phi(\cdot)$ は声質評価値を高次元特徴量空間へと写像するための非線形関数であり，$\boldsymbol{V}_m$ は m 番目の分布の高次元特徴量空間上における回帰パラメータである．入力歌手 k の声質評価値ベクトルを $\boldsymbol{w}(k)$ とすると，カーネル回帰による差分平均ベクトル $\Delta\boldsymbol{\mu}_m$ は以下の式で表される．

$$\Delta\boldsymbol{\mu}_m = \boldsymbol{V}_m\phi(\boldsymbol{w}(k) + \Delta\boldsymbol{w}) - \boldsymbol{V}_m\phi(\boldsymbol{w}(k)) \quad (5)$$

式 (4) 及び 式 (5) の演算は，全事前収録目標歌手の平均ベクトルの補間処理として表現される．その際に必要となる回帰パラメータは，正則化付き最小二乗誤差推定により求める．なお，カーネル回帰においては，差分声質評価値ベクトル $\Delta\boldsymbol{w}$ のみでなく，入力歌手の声質評価値ベクトル $\boldsymbol{w}(k)$ も必要となる．

---

* Voice timbre control based on voice timbre evaluation values applied to voice data of singing voice synthesis system, by YAMANE, Soichi, KOBAYASHI, Kazuhiro(NAIST), TODA, Tomoki (Nagoya Univ./NAIST), NAKANO, Tomoyasu, GOTO, Masataka (AIST), NAKAMURA, Satoshi (NAIST)

2.3 音源データに対する適用

本稿では，歌声合成システムとして UTAU [2] を用いる．UTAU では，音源データとして，歌声の素片データを集めた UTAU 音声ライブラリの入れ替えにより，任意の音源データによる歌声を合成できる．素片接続型合成方式であることから，本稿では声質制御を適用する手法として，合成歌声に対して変換処理を施す方法（U2C: UTAU synthesis to Conversion）と，音源データ中の素片データに対して変換処理を施す方法（C2U: Conversion to UTAU synthesis）の 2 つを検討する．U2C では，変換モデルのみを保持することで，様々な声質を持つ合成歌声を生成できるが，一方で，合成時に必要となる計算量は増加する．一方で，C2U では，事前に変換処理を施す事で合成時の計算量は増加しないが，その際には変換後の音源データを保持する必要がある．

3 実験的評価

3.1 実験条件

GMM の学習データとして，50 個の UTAU 音声ライブラリ [2] を用いる．学習に用いる音声データとして，UTAU [2] を用いて合成された，“ああいあうあえ” 等の 2 音素接続を可能な限り考慮した無意味詞で構成され，かつ，5 音階の音高遷移を含んだ約 10 分間の合成歌声データを使用する．スペクトル包絡パラメータとして，STRAIGHT 分析 [8] によって得られるスペクトル包絡から算出される 1 次から 24 次のメルケプストラム係数を使用する．シフト長は 5 ms，サンプリング周波数は 16 kHz とする．スペクトル包絡に対する GMM の混合数は 128 である．各 UTAU 音声ライブラリには，19 名の評価者による 1 - 7 の 7 段階の声質評価値が付与されており，全評価者の平均値を声質評価値として使用する．

評価歌声として，AIST ハミングデータベース：ポピュラー音楽 (RWC-MDB-P2001) 日本語歌詞，サビパート [9] を用いる．評価楽曲は No.09 とする．被験者は 20 代の男性 8 名である．学習に用いた 50 個の UTAU 音声ライブラリの内，8 個を用いて評価歌声を生成する．各ライブラリが持つ声質評価値を，それぞれ-6, -3, +3, +6 と変化させることで，声質制御を行う．本実験では，“性別” という声質表現語に対応する声質評価値のみを操作する．

声質制御性能に関する評価として，声質制御の適用法である U2C 及び C2U を比較する．回帰手法は線形回帰とする．被験者は，各合成歌声に対して，“1-女性的”，“7-男性的” という 7 段階の声質評価値を付与することで評価を行う．一方，合成歌声の品質に関する評価として，5 段階オピニオン評定（“5-とても良い”，“4-良い”，“3-普通”，“2-悪い”，“1-とても悪い”）に基づく評価を行う．声質制御の適用法には C2U を用い，線形回帰とカーネル回帰の比較を行う．

3.2 実験結果

図 1 に声質制御性能に関する評価結果を示す．横軸は設定した差分声質評価値を示し，縦軸は被験者により評価された声質評価値の差分を示す．U2C および C2U ともに声質制御の効果が得られているが，C2U の方が若干精度が高い（平均 0.2 程度の精度向上）傾向が見られる．

図 2 に音質に関する評価結果を示す．提案する声質制御法は差分変換に基づいているため，差分声質評価値を 0 に設定した際（“Source voice”）には実質変換処理が行われず，変換前の UTAU による合成歌声と等価となる．結果から，差分評価値を 0 以外に

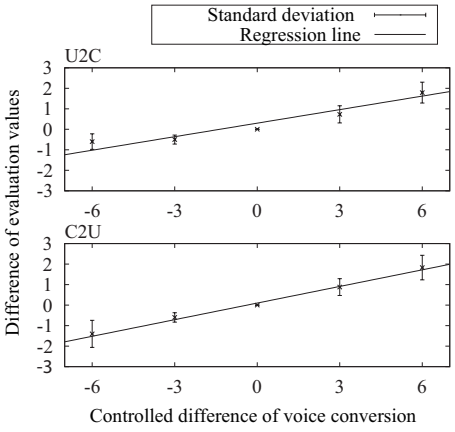

Fig. 1 “性別” に関する声質評価

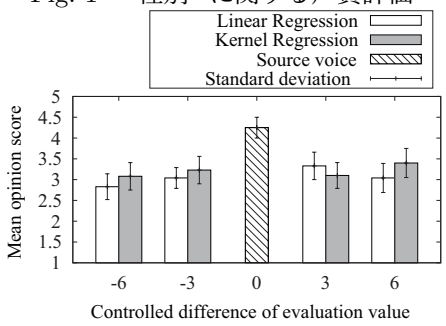

Fig. 2 音声品質に関する MOS 評価

設定した際には，変換処理による音質劣化が生じることが分かる．“性別” に関する声質制御においては，線形回帰とカーネル回帰の間には大きな差は認められない（0.3 程度の差）．そのため，両手法の比較については，高い非線形性が認められる声質表現語に対する実験も行う必要がある．

4 まとめ

本稿では，歌声合成システムの表現拡張を目指し，統計的歌声声質変換に基づく声質制御法を提案した．実験結果から，提案法により，一定の音声品質劣化を招くものの，主観値に沿った声質制御が可能である事が分かった．今後，他の声質表現語に関連する声質制御の評価や，声質制御性能の改善に取り組む．

謝辞 本実験で用いた UTAU 音声ライブラリを対象に，声質表現語に対する声質評価値をご提供頂いた鈴木せりふ氏に感謝する．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] H. Kenmochi *et al.*, Proc. INTERSPEECH, pp.4011-4012, 2007.
[2] 歌声合成ツール UTAU, http://utau2008.web.fc2.com/, 2015-11-24.
[3] S. Yamane *et al.*, IPSJ SIG, Vol. 2015-MUS-108 No. 6, 2015.
[4] H. Doi *et al.*, Proc. APSIPA ASC, 2012.
[5] K, Kobayashi *et al.*, Proc. IEICE Trans. Inf. Syst., Vol. E97-D, No. 6, pp. 1419-1428, 2014.
[6] T. Toda *et al.*, IEEE Trans. ASLP, Vol. 15, No. 8, pp. 2222-2235, 2007.
[7] K. Kobayashi *et al.*, Proc. INTERSPEECH, pp. 2754-2758 , 2015.
[8] H. Kawahara *et al.*, Speech Communication, Vol. 27, No. 3-4, pp. 187-207, 1999.
[9] M. Goto *et al.*, Proc.ISMIR, pp.229-230, 2003.