歌声合成による学習データ生成を利用した歌声の声質評価値推定法*

☆山根 壮一, 小林 和弘, 戸田 智基（奈良先端大・情報）, 中野 倫靖，後藤 真孝（産総研），
ニュービッグ グラム, サクリアニ サクティ, 中村 哲（奈良先端大・情報）

1 はじめに

近年，主に楽曲製作において，VOCALOID[1] や UTAU[2] のような歌声合成システムが広く利用されている．歌声合成システムは，音高やリズムなどの楽譜情報と言語情報の入力により歌声を合成できる．さらに，歌声合成用の音源データ (歌手の声) の入れ替えにより容易に合成歌声の声質を変更することができる．一方で，利用可能な音源データの数は膨大であり，所望の声質を見つけることは容易ではない．

本稿では，主観的な声質評価値による音源データ検索の実現を目指し，音源データに対する声質評価値の自動推定法を提案する．多数の歌手に対する音源データを用いて学習された混合正規分布モデル (GMM : Gaussian Mixture Model) に基づき，各歌手の声質を捉える特徴量を抽出し，回帰分析により声質評価値を推定する．実験結果より，複数の声質表現語に対する声質評価値の推定精度を示す．

2 歌声合成による学習データ生成を利用した声質特徴量の抽出

歌声においては，音高や音韻継続長は楽曲に大きく依存するため，声質評価値に大きく影響を与える音響特徴量として，スペクトル包絡パラメータや非周期成分パラメータが有効であると考えられる．一方で，これらの音響特徴量は，声質のみでなく音韻の影響も大きく受ける．そのため，音韻の影響を取り除いた特徴量を抽出することが重要となる．また，音響特徴量から声質評価を推定する手法として，スペクトル包絡情報を Deep neural network でモデル化する手法が提案されている [3]．

本稿では，声質特徴量を抽出する手法として，声質変換処理においてその有効性が示されている参照歌手に基づく結合確率密度モデリング [4] を応用する．まず，参照歌手と多数の事前収録目標歌手の音源データを用いて，歌声合成により，楽譜情報を共有した歌声データ (パラレルデータ) を作成する．それらを用いて，参照歌手と個々の事前収録目標歌手の音響特徴量に対する結合確率密度関数を，次式に示す GMM によりモデル化する．

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

$$\boldsymbol{\mu}^{(s)} = \left[\boldsymbol{\mu}_1^{(s)\top}, \cdots, \boldsymbol{\mu}_M^{(s)\top}\right]^\top \quad (2)$$

ここで，$\boldsymbol{X}_t = \left[\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top\right]^\top$ と $\boldsymbol{Y}_t^{(s)} = \left[\boldsymbol{y}_t^{(s)\top}, \Delta\boldsymbol{y}_t^{(s)\top}\right]^\top$ は，参照歌手と s 番目の事前収録目標歌手の静的・動的結合特徴量ベクトルを表す．$\top$ は転置を表す．$\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ 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 m 番目の分布の混合重みである．平均ベクトル $\boldsymbol{\mu}_m^{(Y)}(s)$ は，s 番目の事前収録目標歌手に対する m 番目の分布における出力平均ベクトルを表す．それらを結合したスーパーベクトル $\boldsymbol{\mu}^{(s)}$ が，s 番目の事前収録目標歌手の声質特徴量となる．なお，$\boldsymbol{\lambda}$ は GMM のパラメータセットを表し，スーパーベクトル以外のパラメータを含む．

上記の GMM を学習する上で，まず，参照歌手と全事前収録目標歌手とのパラレルデータを用いて，次式により，目標歌手非依存 GMM を学習する．

$$\left\{\boldsymbol{\mu}^{(0)}, \boldsymbol{\lambda}^{(0)}\right\} = \arg\max_{\{\boldsymbol{\mu}, \boldsymbol{\lambda}\}} \prod_{s=1}^{S} \prod_{t=1}^{T_s} P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}|\boldsymbol{\lambda}\right) \quad (3)$$

ここで，s 番目の事前収録目標歌手に対するフレーム数は T_s であり，事前収録目標歌手の総数は S である．s 番目の事前収録目標歌手依存 GMM は，参照歌手と s 番目の事前収録目標歌手のパラレルデータを用いて，次式により目標歌手非依存 GMM のスーパーベクトル $\boldsymbol{\mu}^{(s)}$ を最尤基準により更新することで得られる．

$$\boldsymbol{\mu}^{(s)} = \arg\max_{\left\{\boldsymbol{\mu}^{(0)}, \boldsymbol{\lambda}^{(0)}\right\}} \prod_{t=1}^{T_s} P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}|\boldsymbol{\lambda}^{(0)}\right) \quad (4)$$

本学習処理において，参照歌手に関連する分布パラメータは，全事前収録目標歌手の間で共有される．また，参照歌手と各事前収録目標歌手のパラレルデータに基づき，スーパーベクトルが更新される．これらの処理により，個々の事前収録目標歌手依存 GMM において，各分布がモデル化する音韻成分の共有化が成される．その結果，個々の事前収録目標歌手に対するスーパーベクトル間の差は，主に声質の違いに起因するものとなる．

3 声質評価値推定

3.1 回帰分析による声質評価値推定

個々の事前収録目標歌手に対するスーパーベクトルと声質評価値に対して，回帰分析を行うことで，スーパーベクトルから声質評価値を推定するモデルを構築する．

3.1.1 重回帰分析に基づく手法

重回帰分析では，s 番目の事前収録目標歌手の声質評価値ベクトル $\boldsymbol{w}^{(s)} = \left[w_1^{(s)}, \cdots, w_J^{(s)}\right]^\top$ は，同歌手に対するスーパーベクトル $\boldsymbol{\mu}^{(s)}$ から，次式により推定される．

$$\boldsymbol{w}^{(s)} = \boldsymbol{A}\boldsymbol{\mu}^{(s)} + \boldsymbol{b} \quad (5)$$

ここで，声質表現語の数は J であり，j 番目の声質表現語に対する声質評価値は $w_j^{(s)}$ である．また，$\boldsymbol{A}$ および $\boldsymbol{b}$ は回帰パラメータであり，全事前収録目標歌手に対する声質評価値ベクトルおよびスーパーベクトルを用いて，最小平均二乗誤差推定により求める．

3.1.2 カーネル回帰分析に基づく手法

カーネル回帰分析では，s 番目の事前収録目標歌手の声質評価値ベクトル $\boldsymbol{w}^{(s)}$ は，同歌手に対するスーパーベクトル $\boldsymbol{\mu}^{(s)}$ から，次式により推定される．

---

*An estimation method of voice timbre evaluation values of singing voices using training data generated with singing voice synthesis, by YAMANE, Soichi, KOBAYASHI, Kazuhiro, TODA, Tomoki (NAIST), NAKANO, Tomoyasu, GOTO, Masataka (AIST), NEUBIG, Graham, SAKTI, Sakriani, NAKAMURA, Satoshi (NAIST)

Table 1　Voice timbre expression word

声質表現語	ラベル	詳細
年齢	AGE	幼い - 大人っぽい
綺麗さ	CLR	ノイジー - クリア
性別	GEN	女性的 - 男性的
滑舌	LSN	舌足らず - はきはき
力強さ	POW	優しい - 力強い
癖の強さ	UNQ	素直な - 癖がある

$$\boldsymbol{w}^{(s)} = \boldsymbol{V}\boldsymbol{\phi}(\boldsymbol{\mu}^{(s)}) \tag{6}$$

ここで，$\boldsymbol{\phi}(\cdot)$ はスーパーベクトルを高次元特徴量空間へ写像するための関数であり，$\boldsymbol{V}$ は高次元特徴量空間上における回帰パラメータである．上記の式は，カーネル関数 $k(\cdot,\cdot)$ を用いて，次式にて表される．

$$\boldsymbol{w}^{(s)} = \boldsymbol{Z}\boldsymbol{k}(\boldsymbol{\mu}^{(s)}) \tag{7}$$

$$\boldsymbol{k}\left(\boldsymbol{\mu}^{(s)}\right) = \left[k\left(\boldsymbol{\mu}^{(1)},\boldsymbol{\mu}^{(s)}\right),\cdots,k\left(\boldsymbol{\mu}^{(S)},\boldsymbol{\mu}^{(s)}\right)\right]^{\top} \tag{8}$$

ここで，パラメータ $\boldsymbol{Z}$ は，全事前収録目標歌手に対する声質評価値ベクトルおよびスーパーベクトルを用いて，正則化付き最小平均二乗誤差推定により求める．

3.2　任意の目標歌手に対する声質評価値推定

任意の目標歌手の音源データが与えられた際には，まず，歌声合成により，参照歌手とのパラレルデータを作成し，式 (4) に基づきスーパーベクトルを抽出する．得られたスーパーベクトルに対して，回帰分析に基づく手法を用いて，声質評価値ベクトルを推定する．

4　実験的評価

4.1　実験条件

音源データとして 40 個の UTAU 音声ライブラリ [2] を用意し，1 ライブラリあたり 7 音高，100 種類の音節に対する歌声を合成する．1 音節の長さは 2 秒である．スペクトル包絡パラメータとして，STRAIGHT 分析 [5] 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．また，音源特徴量として STRAIGHT 分析によって得られる 0-1, 1-2, 2-4, 4-6, 6-8 kHz の 5 周波数帯域における平均非周期成分を使用する．シフト長は 5 ms，サンプリング周波数は 16 kHz とする．スペクトル包絡と非周期成分に対する GMM の混合数はそれぞれ 128, 16 である．

本稿では，表 1 に示す 6 種の声質表現語に対する声質評価値を用いる．19 名の評価者によって，各声質表現語，各ライブラリごとに 1-7 の範囲の 7 段階での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．

40 個の UTAU 音声ライブラリに対して，学習用データとして 39 個，声質評価値推定用データとして 1 個のライブラリに分ける一個抜き交差検証を行うことで，声質評価値推定精度を評価する．評価尺度として，正解値と推定値との相関係数を用いる．

4.2　実験結果

図 1 と図 2 に，重回帰分析による結果とカーネル回帰分析による結果を各々示す．各図には，メルケプストラム係数，非周期成分，及びそれらの結合特徴量ベクトルを用いた際の結果を示す．実験結果より，“年齢” と “性別” に対する声質評価値の推定精度が高いことがわかる．一方で，“綺麗さ” に対する声質評価値の推定精度が低く，本手法での声質評価値推定が困難である．また，特徴量ベクトルとして，非周期成分よりもメルケプストラム係数を使用する方が，高い推定精度を得られる．結合特徴量ベクトルを用いても，メルケプストラム係数を用いた際と同等の推定精度しか得られないことから，非周期成分が声質評価値に与える影響は微小であると考えられる．なお，カーネル回帰分析を用いることで，“力強さ” 及び “癖の強さ” において，10%程度の推定精度向上が得られる．

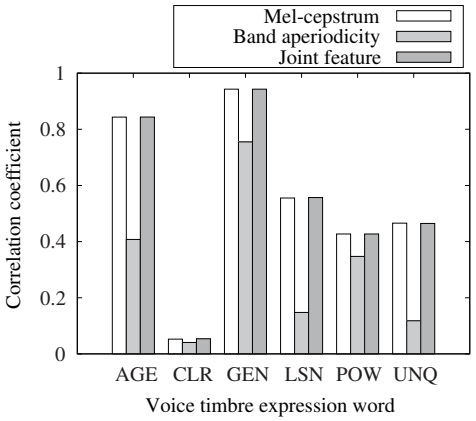

Fig. 1　Correlation coefficients between correct and estimated values by multiple regression

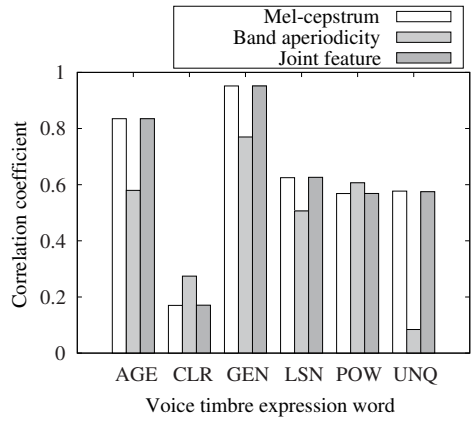

Fig. 2　Correlation coefficients between correct and estimated values by kernel regression

5　まとめ

音源データに対する声質評価値の推定法として，歌声合成による学習データ生成により得られた GMM を用いて声質特徴量を抽出し，回帰分析により声質評価値を推定する手法を提案した．実験結果より，メルケプストラム係数に基づく声質特徴量とカーネル回帰分析を用いることで，“年齢”, “性別” に関する声質評価値に対しては相関係数で 0.8 以上の推定精度が得られ，“滑舌”, “力強さ”, “癖の強さ” に対する声質評価値に対しては相関係数で 0.6 程度の推定精度が得られることが分かった．今後，本手法に基づく音源データの検索機能の実現に取り組む．

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] H. Kenmochi *et al.*, Proc. INTERSPEECH, pp.4011-4012, 2007.
[2] 歌声合成ツール UTAU, http://utau2008.web.fc2.com/, 2015-7-8.
[3] F. Yokomori *et al.*, IPSJ SIG, Vol. 2015-MUS-107, No. 61, 2015.
[4] H. Doi *et al.*, APSIPA ASC, 2012
[5] H. Kawahara *et al.*, Speech Communication, Vol. 27, No. 3-4, pp. 187-207, 1999.