Vol.2015-MUS-108 No.6
2015/9/1

歌声合成システムの音源データ検索のための声質評価値推定法

山根 壮一[1,a)]　小林 和弘[1,b)]　戸田 智基[1,c)]　中野 倫靖[2,d)]　後藤 真孝[2,e)]
ニュービッグ グラム[1,f)]　サクリアニ サクティ[1,g)]　中村 哲[1,h)]

概要：歌声合成システムは，歌声合成用の音源データ (歌手の歌声) の入れ替えにより，合成歌声の声質を容易に変更する事が出来る．一方で，利用可能な音源データ数は膨大であるため，所望の声質を持つ音源データを見つける事は困難である．この問題に対し，本研究では，主観的な声質評価値を用いた音源データの検索方法について検討し，音源データに対して，声質評価値を自動推定する方法を提案する．音源データ群から，参照歌手および多数の事前収録目標歌手を用意し，歌声合成システムを用いてパラレルデータを合成する．パラレルデータから音響特徴量を抽出し，結合確率密度関数を混合正規分布モデル（Gaussian Mixture Model : GMM）を用いてモデル化することで，個々の目標歌手の声質をよく捉える特徴量ベクトルの抽出を実現する．得られた特徴量ベクトルと各事前収録目標歌手に対する声質評価値を用いて，任意の音源データに対する声質評価値を推定するための回帰モデルを学習する．本稿では，声質評価値の自動推定に関する実験結果より，1）“年齢”，“性別” に関する声質評価値については高い推定精度が得られること，2）他の声質表現語に対して，重回帰分析よりもカーネル回帰分析を用いることで，推定精度が向上すること，3）パラレルデータの種類は，声質評価値の推定精度に大きな影響を与えないこと，を示す．

1. はじめに

歌声を含む楽曲の製作において， VOCALOID [1] や UTAU [2]，Sinsy [3] に代表される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，音高や発声のタイミング，継続長などの楽譜情報と歌詞などの言語情報の入力により，所望の旋律を歌う歌声を容易に合成できる．さらに，歌声合成用の音源データ（歌手の歌声）の種類や声質に関するパラメータなどの歌手情報を操作することで，合成歌声の声質を容易に変更することができる．歌声合成システムでは，歌手の声質を楽曲製作のどの段階においても自由に変更できるため，ユーザが製作した楽曲に対して，合致した声質を持つ歌声を選出することが可能である．このため，歌声合成システムにおける歌声の声質選択は，楽曲製作の自由度を大幅に広げているといえる．

歌声合成システムを用いて所望の声質を持つ歌声を合成する場合には，ユーザはまず，複数の音源データから音源データを一つ選択する必要がある．一方で，利用可能な音源データが膨大に存在する場合，それらの音源データから楽曲に適したものを一つ一つ探索することは非常に困難である．例えば，無料の音源データが豊富な UTAU 音声ライブラリ [2] では，存在する音源データの総数が 5000 を超えており [4]，この音源データを全て確認するには，多大な労力を要する．そのため，ユーザが容易に所望の音源データを検索できるシステムの構築が望まれる．

歌声を検索するシステムとして，様々な手法が提案されている．歌声の韻律的特徴をクエリとして用いた歌唱スタイルの検索手法として，F_0 軌跡の動的変動を相平面（$F_0 - \Delta F_0$ 平面）で表現し，この相平面上の確率分布に基づき，歌手の歌唱スタイルを同定する方法 [5] が提案されている．また，歌声の分節的特徴をクエリとして用いた楽曲の検索手法として，VocalFinder [6] が提案されている．この手法は，複数の歌声間の分節的特徴の類似度を用いて，似た声質を持つ歌声を含む楽曲を検索するシステムである．さらに，歌声を直感的な印象語によって分類し，

[1] 奈良先端科学技術大学院大学
Nara Institute of Science and Technology (NAIST)
[2] 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)
[a)] yamane.soichi.yj2@is.naist.jp
[b)] kazuhiro-k@is.naist.jp
[c)] tomoki@is.naist.jp
[d)] t.nakano@aist.go.jp
[e)] m.goto@aist.go.jp
[f)] neubig@is.naist.jp
[g)] ssakti@is.naist.jp
[h)] s-nakamura@is.naist.jp

歌声の印象に基づいた検索 [7] が提案されている．これらの手法は，クエリとして歌声を入力し，音響特徴量の類似度に基づいて歌唱スタイルや楽曲を検索する手法である．そのため，その用途は，所望の歌声に類似した歌声が入手可能である場合や，自身でそのような歌声を生成可能である場合に限定される．一方で，本研究で取り扱う問題のように，大量に存在する音源データから，ユーザが所望の声質を持つ音源データを検索したい場合においては，類似した声質を持つ音源データが入手できるとは限らない．音源データの声質を特定の指標で表現し，クエリとして入力できる技術が必要となる．

本研究では，検索クエリに用いる指標として，声質表現語 [8] に対する主観的な声質評価値に着目する．声質表現語に対する声質評価値を用いることで，ユーザは所望の声質を直感的に表現することが可能となる．一方で，声質評価値による音源データ検索を実現するためには，全ての音源データに対して，声質評価値を付与する必要がある．主観的な声質評価値を全ての音源データに対して付与するためには，人手による膨大な作業が必要となる．そのため，各音源データに対して，声質評価値を自動推定する技術の構築が望まれる．

本稿では，声質評価値による音源データ検索の実現に向けて，歌声合成システムの音源データを対象とした声質評価値の自動推定技術の構築に取り組む．個々の音源データから，音韻の影響を取り除き，声質の影響を精度よく捉える特徴量を抽出するために，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 [9] を応用する．提案法では，歌声合成システムの特徴を活かし，異なる音源データ間で音韻情報が共有されたパラレルデータを合成し利用することで，音韻の違いにより生じる周波数特性の変動による影響を極力受けず，個々の音源データの声質を表す特徴量を抽出するための確率モデルの構築を可能とする．得られた特徴量と各音源データに対する声質評価値を用いて，声質評価値を推定するための回帰モデルを学習することで，任意の音源データに対する声質評価値の推定を実現する．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．また，パラレルデータの種類を変更した際の声質評価値の推定精度を比較することで，声質評価値の自動推定に適したパラレルデータの作成法についても検討する．

なお，本研究と同様に，既存の声質表現語とそれに対する評価値を用いて，スペクトル包絡パラメータを入力とする Deep Neural Network (DNN) を用いた声質評価法 [10] が提案されている．この手法は，音韻成分と声質成分の分離処理も同時に DNN に学習させる枠組みである．これに対し，提案法は，特徴量抽出に用いる確率モデルの内部構造で，音韻成分と声質成分を明示的に分離する枠組みであり，検索後の音源データに対する声質変換処理 [9] の適用とも親和性が高いという利点がある．

2. 声質特徴量の抽出

歌声においては，音高や音韻継続長などの韻律的特徴は楽曲に大きく依存するため，声質の評価を行うために使用す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などの分節的特徴が適切であると考えられる．一方で，スペクトル包絡パラメータや非周期成分パラメータは，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き，声質のみを表す特徴量を抽出することが重要となる．

本稿では，分節的特徴に基づく歌手依存の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関数に基づく手法 [9] 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，歌声合成により同一の楽譜情報を持つ歌声データ (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分節的特徴に関する音響特徴量を抽出し，それら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．

$$P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)} | \boldsymbol{\mu}^{(s)}, \boldsymbol{\lambda}\right)$$

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

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

ここで，$\boldsymbol{X}_t = \left[\boldsymbol{x}_t^\top, \boldsymbol{\Delta}\boldsymbol{x}_t^\top\right]^\top$ と $\boldsymbol{Y}_t^{(s)} = \left[\boldsymbol{y}_t^{(s)\top}, \boldsymbol{\Delta}\boldsymbol{y}_t^{(s)\top}\right]^\top$ 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 $\boldsymbol{\mu}_m^{(s)}$ 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．

式 (1) の GMM を学習するために，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．

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

ここで，s 番目の事前収録目標歌手に対するパラレルデータのフレーム数は 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に対する歌手依存 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，目標歌手非依存 GMM のスーパーベクトル $\boldsymbol{\mu}^{(0)}$ を次式の最尤基準に基づき更新することで得られる．

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

本稿では，この歌手依存のスーパーベクトル $\boldsymbol{\mu}^{(s)}$ を声質特徴量として用いる．

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. 回帰分析を用いた声質評価値の自動推定

個々の事前収録目標歌手の声質特徴量とあらかじめ付与された声質評価値に対して回帰分析を行うことで，任意の目標歌手の声質特徴量から声質評価値を推定するモデルを構築する．

3.1 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，$\boldsymbol{A}$ 及び $\boldsymbol{b}$ 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及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．

3.2 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．

$$\boldsymbol{w}^{(s)} = \boldsymbol{V}^\top \boldsymbol{\phi}\left(\boldsymbol{\mu}^{(s)}\right) \quad (6)$$

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る．$\boldsymbol{V}$ は高次元特徴量空間上における回帰パラメータであり，スーパーベクトル $\boldsymbol{\mu}^{(s)}$ を用いて次式で表される．

$$\boldsymbol{V} = \sum_{s=1}^{S} \boldsymbol{\phi}\left(\boldsymbol{\mu}^{(s)}\right) \boldsymbol{Z}_s^\top \quad (7)$$

式 (6) に式 (7) を代入すると，

$$\boldsymbol{w}^{(s)} = \boldsymbol{Z}\boldsymbol{k}\left(\boldsymbol{\mu}^{(s)}\right) \quad (8)$$

$$\boldsymbol{k}\left(\boldsymbol{\mu}^{(s)}\right) = \left[k\left(\boldsymbol{\mu}^{(1)}, \boldsymbol{\mu}^{(s)}\right), \cdots, k\left(\boldsymbol{\mu}^{(S)}, \boldsymbol{\mu}^{(s)}\right)\right]^\top \quad (9)$$

が得られる．ここで，パラメータ $\boldsymbol{Z} = [\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_S]$ であり，$k(\cdot,\cdot)$ はカーネル関数を表す．本稿では，カーネル関数として，次式で示すガウスカーネルを用いる．

$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{||\boldsymbol{x} - \boldsymbol{x}'||^2}{\sigma^2}\right) \quad (10)$$

ここで，σ は任意の正値をとるパラメータである．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．

$$\boldsymbol{Z} = \boldsymbol{W}(\boldsymbol{K} + r\boldsymbol{I})^{-1} \quad (11)$$

ここで，$\boldsymbol{K} = \left[\boldsymbol{k}\left(\boldsymbol{\mu}^{(1)}\right), \cdots, \boldsymbol{k}\left(\boldsymbol{\mu}^{(S)}\right)\right]^\top$，$\boldsymbol{W} = \left[\boldsymbol{w}^{(1)}, \cdots, \boldsymbol{w}^{(S)}\right]$ であり，r は正則化に関するハイパーパラメータである．

3.3 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．

なお，本枠組みは合成歌声のみならず，自然歌声に対しても適用可能であり，大きく分けて 2 つの方法が考えられる．一つ目は，声質評価値推定の対象となる自然歌声と同じ音韻情報を持つように，参照歌手による歌声合成システムを用いて合成歌声を生成し，パラレルデータを作成する方法である．得られたパラレルデータから，合成歌声に対する場合と同様の手順により，自然歌声に対するスーパーベクトルを抽出することができる．

二つ目は，式 (4) で使用する目標歌手非依存 GMM を周辺化することで，目標歌手の音響特徴量に対する確率密度関数を導出し，自然歌声の音響特徴量のみを用いてスーパーベクトルを求める方法である．最大事後確率推定 [11]，最尤線形回帰 [12]，固有声 [13] などに代表される適応処理を用いることで，個々の分布がモデル化する音韻成分をできる限り保持しつつ，自然歌声の声質を捉えるスーパーベクトルが抽出されると考えられる．

4. 実験的評価

提案法による声質評価値推定の精度を評価するため，歌声合成用の音源データを用いて声質評価値を

表 1 声質表現語の種類

種類	ラベル	声質表現語
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

4.1 実験条件

声質評価値の推定に用いる音源データとして，40 個の UTAU 音声ライブラリ [2] を用意する．GMM の学習に用いる音声データとして，UTAU を用いて合成された “単音節発声”, “話声”, “歌声” の 3 種類の音声を用いる．“単音節発声” は，現代日本語の拍の音声記号の種類 [14] に基づき，100 種類の音節について，7 種類の音高の歌声を 1 音ずつ作成し，合計 700 種の合成された音声を用いる．1 音節の長さは約 2 秒である．“話声” は，音素バランスの取れた 50 文 [15] を，話声の韻律を手作業で再現し，歌声合成システムによって合成された音声を用いる．“歌声” は，RWC 研究用音楽データベース:ポピュラー音楽 (RWC-MDB-P-2001) [16] から日本語歌詞の楽曲 10 曲（A メロ・サビ等を 1 フレーズとし，1 曲平均 6 フレーズの合計 64 フレーズ）を使用し合成された歌声を用いる．

スペクトル包絡パラメータとして，STRAIGHT 分析 [17] 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 種類の声質表現語を用いる．各声質表現語に対する声質評価値は，19 名の評価者によって評価された値を用いる．ここで，声質評価値は，1-7 の 7 段階の値で評価されている．なお，各音源データに対する声質評価値は，声質表現語ごとに全評価者の平均値を用いる．

音源データに対し，声質特徴量ベクトルを抽出するために，全音源データの音声データを用いて，目標歌手非依存 GMM を学習する．学習された目標歌手非依存 GMM を初期モデルとして，各音源データの声質特徴量ベクトルを抽出する．回帰分析は，39 個の音源データに対する声質特徴量ベクトルと声質評価値を用いて回帰モデルを学習し，1 個の音源データに対する声質評価値を推定する交差検証により評価する．なお，本実験は，あらかじめ付与された各音源データの声質評価値を正解値として，推定値との相関係数によって評価を行う．

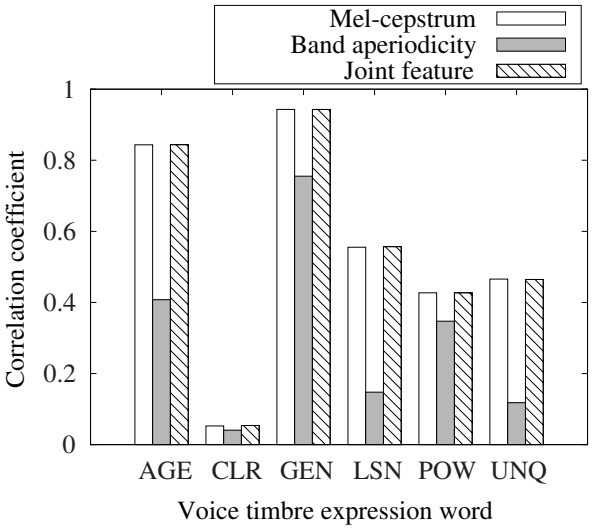

図 1 重回帰分析による推定値と正解値との相関係数

4.2 実験結果

図 1，2 に，学習データとして “単音節発声” を使用した場合の重回帰分析とカーネル回帰分析による声質評価値の推定結果を示す．各図には，音響特徴量として，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を用いた際の結果を示す．実験結果より，“年齢 (AGE)” と “性別 (GEN)” に対して，各々 0.8 以上，0.9 以上の相関係数が得られており，高い推定精度が得られていることが分かる．一方で，“綺麗さ (CLR)” に関しては，0.1 以下の相関係数しか得られておらず，推定精度が低いことがわかる．音響特徴量として，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．また，結合特徴量を用いても，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，非周期成分が声質表現語に与える影響は微小であると考えられる．なお，非線形回帰であるカーネル回帰分析を用いることで，“力強さ (POW)” 及び “癖の強さ (UNQ)” において，相関係数が 0.4 程度から 0.6 程度まで上昇しており，推定精度向上が得られることが分かる．

図 3 に，音声データとして “単音節発声”, “話声”, “歌声” をそれぞれ用いた場合の声質評価値の推定結果を示す．なお，用いた音響特徴量はメルケプストラム係数と非周期成分との結合特徴量であり，声質評価値推定に用いた回帰モデルはカーネル回帰分析である．実験結果より，“癖の強さ (UNQ)” に対してのみ，音声データに “単音節発声” を用いた際の推定精度が他の 2 種類の音声データを用いた場合と比べて高くなっているが，それ以外の声質表現語に対しては，推定精度に大きな変化が見られない．このことから，GMM の学習および声質特徴量抽出に用いるパラレルデータに関しては，その種類が推定精度に与える影響はさほど大きくないといえる．

これまでの結果から，“綺麗さ (CLR)” に関しては推定

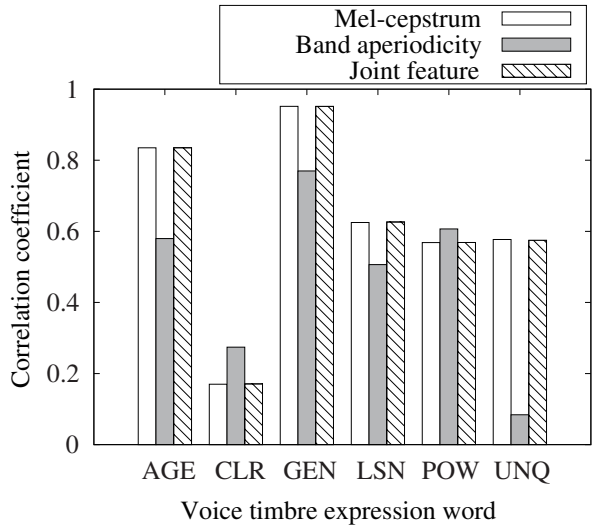

図 2 カーネル回帰分析による推定値と正解値との相関係数

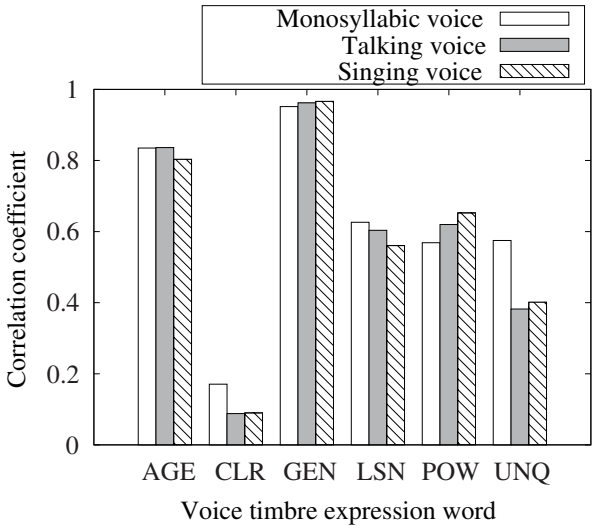

図 3 音源データを変えた場合におけるカーネル回帰分析による推定値と正解値との相関係数

精度が著しく低いことが分かる．この原因を調査するため，より詳細な調査を実施する．まず, “綺麗さ (CLR)” に関する声質表現語 “ノイジー - クリア” に対して，評価者による解釈の違いの有無を調査するために，個々の声質評価者の声質評価値に対して，ward 法によるクラスタ分析を行う．図 4 に結果を示す．この図において，高さは声質評価者間の類似度に関連する量を表しており，値が小さいほど類似度が高いことを示す．図から，大きく分けて二組のクラスタが存在することが分かる．ここで，評価者の人数が多い方を “Majority”，少ない方を “Minority” とする．“Majority” の人数は 14 人であり，“Minority” は 5 人である．なお，他の声質表現語に対する声質評価値に対しても同様のクラスタ分析を行った結果（一例として “年齢 (AGE)” に対する結果を図 5 に示す），“綺麗さ (CLR)” で見られたような類似度が低い複数のクラスタに分かれるという傾向は見られなかったことから, “綺麗さ (CLR)” に関

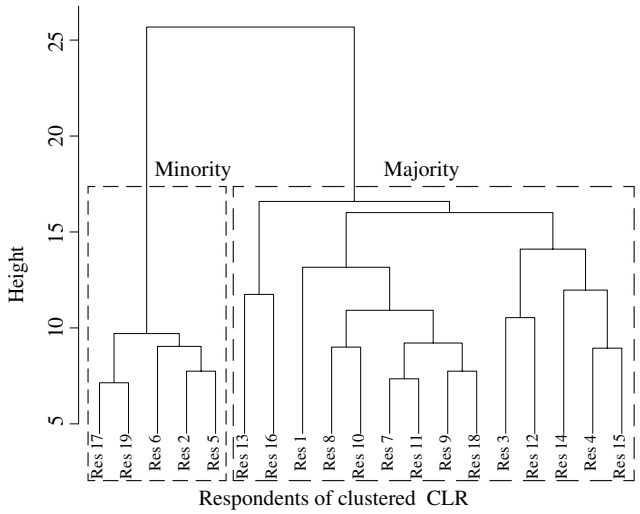

図 4 “CLR” の声質評価値を用いた 19 名の評価者による ward 法クラスタ分析の結果

しては評価者による解釈の違いが特に生じやすいということが分かる．

次に，“綺麗さ (CLR)” における “Majority” と “Minority” の各々のクラスタ内で，回帰分析を実施し，声質評価値を推定した際の結果を図 6 に示す．なお，用いた音響特徴量は結合特徴量であり，声質評価値推定に用いた回帰モデルはカーネル回帰分析である．“All” は全ての評価者の声質評価値を正解値としているため，図 3 の “CLR” の値と同じ結果を示している．実験結果より，“Minority” では相関係数が 0.5 程度まで改善するものの，“Majority” では，依然として相関係数が極めて低いことが分かる．このことから，“Minority” に関しては，他の表現語と同様に，“綺麗さ (CLR)” に関する声質はメルケプストラム係数である程度表現可能であると考えられるが，“Majority” に関しては，評価者によっては，メルケプストラム係数ではとらえられない声質成分を評価している可能性がある．

5. 結論

声質表現語に対する主観的な声質評価値を検索クエリとした音源データ検索の実現を目指し，本稿では，音源データに対する声質評価値の推定法を提案した．歌声合成システムを用いた参照歌手と目標歌手のパラレルデータ生成と結合確率密度モデリングを応用することで，音韻の影響を極力抑え，声質の影響を精度よく捉える特徴量の抽出を実現した．また，抽出された声質特徴量に対して，回帰分析により声質評価値を推定する処理を実現した．

“年齢”, “綺麗さ”, “性別”, “滑舌”, “力強さ”, “癖の強さ” といった 6 つの声質表現語に対する声質評価値を対象とした実験結果より，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 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

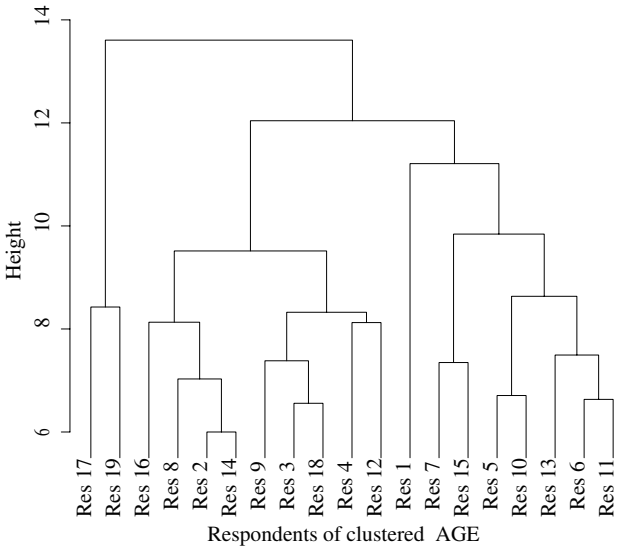

図 5 “AGE” の声質評価値を用いた 19 名の評価者による ward 法クラスタ分析の結果

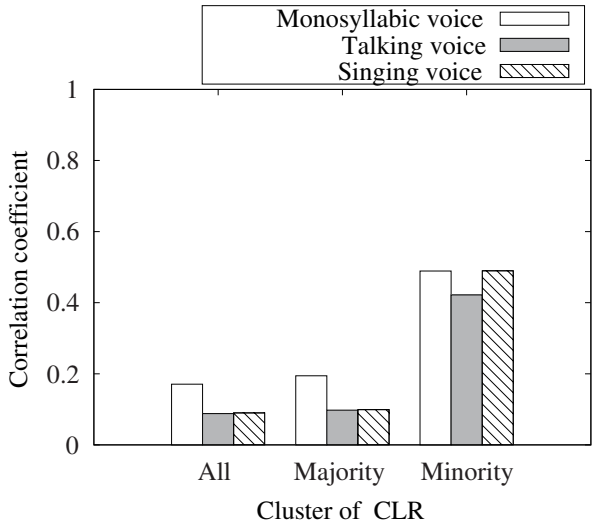

図 6 “CLR” における各クラスタのカーネル回帰分析による推定値と正解値との相関係数

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事が分かった．一方で，“綺麗さ” に対する声質評価値については，評価者間で解釈の相違が生じやすいこと，メルケプストラム係数や非周期成分といった音響特徴量では上手く捉えられない特徴が強く影響している可能性が高いこと，が分かった．また，声質特徴量抽出に使用するパラレルデータの種類として，“単音節発声”, “話声”, “歌声” の使用を比較した結果，推定精度に与える影響は小さいことが分かった．

今後，本手法を利用した音源データの検索機能の実現や，自然歌声を対象とした検索への拡張，声質変換処理との統合に取り組む．

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 and H. Ohshita : VOCALOID - Commercial singing synthesizer based on sample concatenation, *Proc. INTERSPEECH*, pp. 4011-4012 (2007).
[2] 歌声合成ツール UTAU : 〈http://utau2008.web.fc2.com〉(2015.07.08).
[3] 大浦圭一郎，間瀬絢美，山田知彦，徳田恵一，後藤真孝 : Sinsy :「あの人に歌ってほしい」をかなえる HMM 歌声合成システム, 情報処理学会研究報告, Vol. 2010-MUS-86 No. 1 (2010).
[4] UTAU ライブラリまとめ : 〈http://ruto.yu.to/〉(2015.08.04).
[5] T. Kako, Y. Ohishi, H. Kameoka, K. Kashino and K. Takeda : Automaic identification for singing style based on sung melodic contour characterized in phase plane, *Proc. ISMIR*, pp. 393-398 (2009).
[6] H. Fujiwara and M. Goto : A music information retrieval system based on singing voice timbre, *Proc. ISMIR*, pp. 467-470, (2007).
[7] A. Kanato, T. Nakano, M. Goto, H. Kikuchi : An automatic singing impression estimation method using factor analysis and multiple regression, *Proc. ICMC SMC*, pp. 1244-1251, (2014)
[8] 木戸博，粕谷英樹 : 通常発話の声質に関連した日常表現語 : 聴取評価による抽出, 日本音響学会誌, vol. 57, No. 5, pp. 337-344, (2001).
[9] H. Doi, T. Toda, T. Nakano, M. Goto and S. Nakamura : Singing voice conversion method based on many-to-many eigenvoice conversion and training data generation using a singing-to-singing synthesis system, *Proc. APSIPA ASC*, (2012).
[10] 横森文哉，大柴まりや，森勢将雅，小澤賢司 : スペクトル包絡情報を入力とした Deep Neural Network に基づく歌声のための声質評価，情報処理学会研究報告，Vol. 2015-MUS-107 No. 61 (2015).
[11] M. Tonomura, T. Kosaka and S. Matsunaga : Speaker adaptation based on transfer vector field smoothing using maximum a posteriori probablity estimation, *IEEE Trans. SAP*, vol. 1, pp. 688-691 (1995).
[12] C. J. Leggetter and P. C. Woodland : Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models, *Proc. CSL*, vol. 9, No. 2, pp. 171-185 (1995).
[13] R. Kuhn, J. C. Junqua, P. Nugyen and N. Niedzielski : Rapid speaker adaptation in eigenvoice space : *IEEE Trans. SAP*, vol. 8, no. 6, pp. 695-707 (2000).
[14] 大辞林 特別ページ 日本語の世界 日本語の音 : 〈http://daijirin.dual-d.net/extra/nihongoon.html〉(2015.07.29).
[15] 磯健一，渡辺隆夫，桑原尚夫 : 音声データベース用文セットの設計, 音響学会講演論文集，pp. 89-90, (1988).
[16] M. Goto, T. Nishimura, H. Hashiguchi and R. Oka : RWC Music Database : Music genre database and musical instrument sounddatabase, *Proc. ISMIR*, pp. 229-230, (2003).
[17] H. Kawahara, I. Masuda-Katsuse and A. Cheveigne : Rrestructing speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F_0 extraction : Possible role of a repetitive structure in sounds, *Speech Communication*, Vol. 27, No. 3-4, pp. 187-207 (1999).