Vol.2012-MUS-96 No.5
2012/8/9

VocaListenerによる学習データ生成を利用した多対多固有声変換に基づく歌声声質変換

土井 啓成[1,a)]　戸田 智基[1,b)]　中野 倫靖[2,c)]　後藤 真孝[2,d)]　中村 哲[1,e)]

概要：歌声の声質には，歌手の個人性が反映されており，他者の声質に自在に切り替えて歌うことは難しい．そこで我々は，歌声の声質を他者の歌声の声質へと自動変換することで，任意の声質での歌唱を実現する手法を提案し，歌唱という音楽表現の可能性を広げることを目指す．従来，統計的声質変換に基づく歌声声質変換が実現されていたが，提案手法では様々な声質に少ない負担で変換可能にするため，多対多固有声変換を導入する．これにより変換時に数秒程度の少量の無伴奏歌声さえあれば，任意の歌手の歌声から別の任意の歌手の歌声への声質変換が実現できる．しかし，その声質変換モデルの事前学習データとして，ある参照歌手の歌声と多くの事前収録目標歌手の歌声とのペアから構成されるパラレルデータセットが必要で，その歌声収録は困難であった．そこで提案手法では，歌唱表現を模倣できる歌声合成システム VocaListener を用いて目標歌手の歌声から参照歌手の歌声を生成することで，その学習データ構築を容易にする．実験結果から提案手法の有効性を確認した．

1. はじめに

歌うことは多くの人々にとって容易だが，自分の歌声を自在に制御することは難しい．特に歌声の声質は，歌手の歌唱技術によりある程度制御可能なものの，歌手の身体的特徴に依るところが大きく，性別や体格が違う他者の声質を真似て歌唱することは難しい．楽器であれば，楽器の個体を選ぶだけでなく，曲調や好みに応じてエフェクタを使用したり，楽器の部品を交換したりして，その音色を変化させることができる．しかし歌声の場合，自分の声質に合った曲調や上手く歌える曲には限度があることが多く，それが自分自身の好みに合っているとは限らない．もし仮に，歌手が自身の声質に限らず，他者の声質で自在に歌唱することが可能になれば，歌唱の楽しさが増すだけでなく，より多様な表現が生まれる可能性がある．そこで本研究では，歌手自身の声質の限界を超え，多様な声質に変換しながら歌唱できる技術を開発することで，歌唱という音楽表現の可能性を広げることを目指す．

歌声合成システムは，声質を選択して合成できるだけでなく，再現性高く納得のいくまで制御しながら様々な歌唱表現を得ることが可能であり，歌唱付き楽曲の創作における可能性を広げてきた [1]．VOCALOID2 [2], [3] や Sinsy [4], [5] のような歌声合成システムでは，歌詞と楽譜情報から合成歌声を生成する方式が主流であり，text-to-singing システムと呼ばれる．この方式では，音高や音量といった合成パラメータを手動で操作できる場合もあるが，多様で自然な歌唱表現を得るのは容易でなかった．そこで，中野ら [6], [7] は，VOCALOID2 等の合成パラメータの音高と音量をユーザのお手本歌声から自動推定し，お手本歌声を模倣した表現力豊かな合成歌声を容易に生成できる歌声合成システム VocaListener を実現した．これを中野らは singing-to-singing システムと名付け，その後，音高と音量だけでなく声色変化もお手本から真似る歌声合成システム VocaListener2 [8], [9] も提案した．しかし，自然な歌唱表現でリアルタイムに歌声合成することはできなかった．

一方，歌声合成システムを用いずに，直接ユーザの歌声を信号処理して声質変換する歌声声質変換手法 [10] も提案されており，他にも混合音中の歌声の声質変換を対象とした研究 [11] や，二人の歌唱者による同一歌詞の歌声を用いた声質のモーフィング [12], [13] に関する研究もある．この従来の無伴奏歌唱に対する歌声声質変換手法 [10] は，高度化したボイスチェンジャに相当し，統計的声質変

[1] 奈良先端科学技術大学院大学
Graduate School of Information Science, 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)] hironori-d[at]is.naist.jp
[b)] tomoki[at]s.naist.jp
[c)] t.nakano[at]aist.go.jp
[d)] m.goto[at]aist.go.jp
[e)] s-nakamura[at]is.naist.jp

換 [14], [15], [16] に基づいて，特定の源歌手（ユーザ）の歌声と特定の目標歌手の歌声との一対一の対応関係を予め学習し，源歌手から目標歌手へ声質を変換する．この学習のためには，同一歌詞の同一楽曲を，源歌手と目標歌手がそれぞれ歌う必要がある．その一組の歌声を**パラレルデータ**と呼ぶ．これにより，両歌声の音響特徴量間の対応関係を，結合確率密度を表わす混合正規分布 (Gaussian Mixture Model: GMM) でモデル化できる．このパラレルデータが多いほど，違う楽曲の歌声が適切に声質変換でき，最低でも一曲程度（数分程度）の歌声が必要である．学習が終われば，その GMM を用いて最尤基準により，任意の歌詞の源歌手（ユーザ）の歌声の声質を，学習時の目標歌手の声質に変換できる．リアルタイムな変換も短遅延変換アルゴリズム [17] により可能である．このように歌声声質変換は，歌声合成で必要だった歌詞や楽譜の事前準備をすることなしに，即興の歌唱であっても，通常のボーカル用エフェクタと同様に用いることができる利点がある．しかし声質の観点からは，学習した源歌手と目標歌手のペアにしか適用できず，別の歌手の声質を扱うためには，パラレルデータを収録し直す必要がある問題があった．両者の歌唱力や性別，声質が異なると，同じ楽曲を歌うのがそもそも困難な場合もあった．

本研究では，多対多固有声変換 [18] を歌声声質変換に初めて導入することで，パラレルデータ収録の問題を解決する．多対多固有声変換では，変換前の声質となる源歌手（ユーザ）が歌ったワンフレーズ程度の短い歌声と，変換後の声質となる目標歌手が歌った同様に短い歌声さえあれば，パラレルデータを収録し直さずに，別の歌手の声質を扱うことが可能になる．これは，変換前後で同じフレーズである必要すらなく，それぞれ，数秒程度の任意の歌詞と音高の歌声でよい．ただし，そのための事前の準備として，できるだけ多くの**事前収録目標歌手**の歌声をそれぞれ数曲分収録し（最低一曲程度で，目標歌手ごとに違う楽曲でもよい），しかもそれら全部の楽曲を，ある特定の一人の**参照歌手**が歌った歌声を収録しておく必要がある．つまり，それぞれの事前収録目標歌手と参照歌手との一組のパラレルデータを，事前収録目標歌手の人数分用意した**パラレルデータセット**を用意しなければならない．それさえあれば，事前収録目標歌手全員の音響特徴量の存在する声質の空間を効率よく表現するような複数の「固有声」と，参照歌手との音響特徴量間の対応関係を，**固有声 GMM**(Eigenvoice GMM: EV-GMM) として学習できる．固有声 GMM は，それぞれの固有声の声質と参照歌手の声質とを相互に変換できる．固有声は代表的な声質を表現しているので，それらの重み付けで任意の声質を表現できる．この重みだけであれば，ワンフレーズ程度の短い歌声であっても推定できるので，上記の変換が実現できる．しかし，上記のパラレルデータセットの収録は，話声ならば可能でも歌声の場合には非常に困難である．依然として，事前収録目標歌手と参照歌手が同じ楽曲を歌える必要がある．

そこでさらに，上記の我々の歌声合成システム VocaListener[6], [7] を利用した新たなパラレルデータセット構築手法を提案する．VocaListener は，任意の楽曲の歌声の歌唱表現を模倣して特定の歌手の歌声で歌声合成できる．多くの事前収録目標歌手による任意の楽曲の歌声（無伴奏独唱）さえ用意すれば，VocaListener によって，歌声合成音源の一人の歌手の声（例えば「初音ミク」）でそれらすべての楽曲の歌声を合成でき，それが所望のパラレルデータセット用の参照歌手の歌声となる．参照歌手の歌声が用意できない曲はなくなるので，事前収録目標歌手は任意の曲を歌ってもよくなる．しかもこの場合の参照歌手の声質変動は人間より少なく，人間と違って歌い回しまで真似た理想的なパラレルデータセットになる．

2. 従来の歌声合成システムと歌声声質変換

ユーザが自身以外の声質で歌唱表現するための代表的な関連研究として，まず，人間の歌声を収録したデータベースに基づいて歌声を合成する歌声合成システムを紹介する．次に，歌声の声質を他の歌手の声質へと変換する歌声声質変換において用いられる，統計的声質変換の枠組みについて説明する．

2.1 歌声合成システム

歌声合成システムを使用する条件の違いから，以下の三種類に分類できる．

(1) Text-to-singing (lyrics-to-singing)
VOCALOID2 [2], [3] や Sinsy [4], [5] のような，歌詞と楽譜情報から合成歌声を生成する方式である．この方式では，事前に，所望の声質を持つ歌手の歌声を歌声合成用コーパスとして収録し，それを元に，素片接続方式や隠れマルコフモデル (Hidden Markov Model: HMM) 合成方式といった合成手法により，歌声の合成を行う．歌声合成システムを声質変換技術によって拡張する研究 [19] もある．

(2) Speech-to-singing
コーパスを事前に用意することなく，合成対象の歌詞を朗読した話声からその声質を保ったまま歌声に変換する方式であり，齋藤らの SingBySpeaking[20], [21] の研究で名付けられた．SingBySpeaking では，話声の各音素の音高と音量，音長を，楽譜情報に応じて歌声らしく制御することで歌声に変換する．

(3) Singing-to-singing
お手本の歌声を入力として，その音高や音量等の歌唱表現を真似るように歌声合成する方式であり，中野らの VocaListener[6], [7] の研究で名付けられた．VocaListener では，歌詞は与える必要があるが，text-to-singing

のように楽譜は必要なく，お手本の入力歌声から自動推定する．歌声合成エンジンとしては，既存の歌声合成ソフトウェア VOCALOID あるいは VOCALOID2 を用い，その合成パラメータを，お手本歌声の音高と音量を真似るように反復推定して設定する．任意の歌声とその歌詞さえあれば，それを模倣して，歌声合成ソフトウェアとして市販されている様々な声質（歌声ライブラリ）で歌声合成できる特長があり，本研究の提案手法でも，後述するように学習データ生成において活用する．

2.2 統計的声質変換に基づく歌声声質変換

統計的声質変換に基づく歌声声質変換 [10] は，源歌手の歌声を目標歌手の歌声へと統計的手法で変換する技術であり，学習処理と変換処理から成る．学習時には，源歌手と目標歌手が同一曲を歌唱した歌声で構成されるパラレルデータから，音響特徴量を抽出し，両音響特徴量の結合確率密度関数を GMM でモデル化する．変換時には，新たに収録された源歌手の歌声から音響特徴量を抽出し，学習処理で得られた GMM に基づき，最尤系列変換法 [16] を用いて目標歌手の音響特徴量へと変換する．変換された音響特徴量から波形信号を合成することで，目標歌手の歌声が生成される．なお，統計的声質変換により変換する音響特徴量として，スペクトルパラメータや励振源パラメータが用いられるが，本稿では歌声の声質・個人性を最も強く捉える音響特徴量として，スペクトルパラメータの変換に着目する．図 1 に，この従来法の統計的声質変換に基づく歌声声質変換の学習処理及び変換処理を示す．

2.2.1 学習処理

時間フレーム t における源歌手と目標歌手の音響特徴量の静的特徴量ベクトルを，各々 $\boldsymbol{x}_t = [x_t(1), \cdots, x_t(D)]^\top$ 及び $\boldsymbol{y}_t = [y_t(1), \cdots, y_t(D)]^\top$ とする．ここで，$^\top$ は転置を表わす．各時間フレームにおいて動的特徴量ベクトル $\Delta\boldsymbol{x}_t$ 及び $\Delta\boldsymbol{y}_t$ を算出し，各々静的特徴量ベクトルと結合することで，$2D$ 次元の静的・動的結合特徴量ベクトル $\boldsymbol{X}_t = \left[\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top\right]^\top$ 及び $\boldsymbol{Y}_t = \left[\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top\right]^\top$ を構築する．各結合特徴量ベクトルの時系列データに対して，動的時間伸縮法によりフレーム間の対応付けを行うことで，各時間フレームにおける源歌手と目標歌手の静的・動的結合特徴量ベクトル対 $\{\boldsymbol{X}_t, \boldsymbol{Y}_t\}$ を求める．全時間フレームにおける静的・動的結合特徴量ベクトル対を学習データとして用いることで，次式に示す結合確率密度関数 $P(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda})$ を表す GMM を学習する．

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

ここで，$\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ は平均ベクトル $\boldsymbol{\mu}$ および共分散行列 $\boldsymbol{\Sigma}$ 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．GMM のパラメータセット $\boldsymbol{\lambda}$ は，個々の分布における混合重み α_m，源歌手の平均ベクトル $\boldsymbol{\mu}_m^{(X)}$，目標歌手の平均ベクトル $\boldsymbol{\mu}_m^{(Y)}$，源歌手の共分散行列 $\boldsymbol{\Sigma}_m^{(XX)}$，目標歌手の共分散行列 $\boldsymbol{\Sigma}_m^{(YY)}$，及び，源歌手と目標歌手の相互共分散行列 $\boldsymbol{\Sigma}_m^{(XY)} = \boldsymbol{\Sigma}_m^{(YX)^\top}$ から成る．

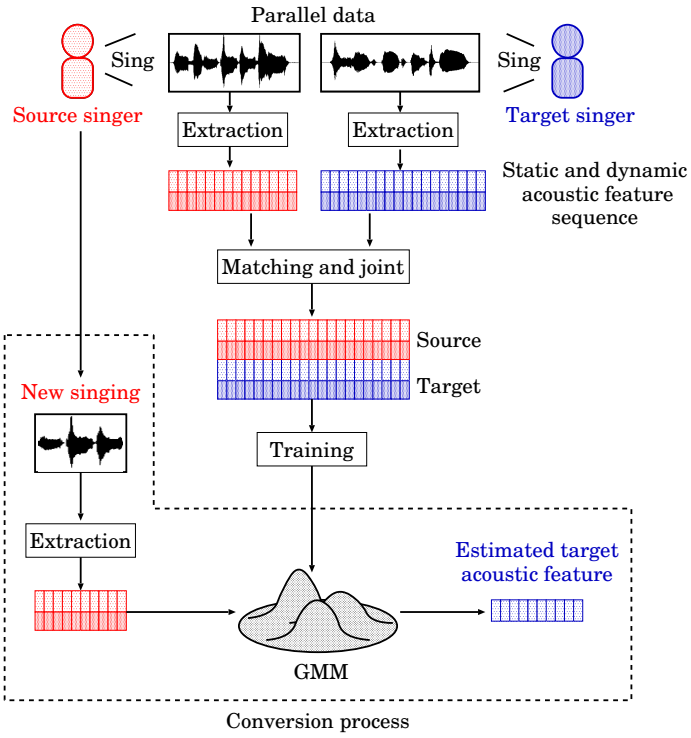

図 1 従来法の統計的声質変換に基づく歌声声質変換での学習過程

また，目標歌手の音響特徴量の時系列データにおいて，系列全体における静的特徴量の変動成分を表す系列内変動（global variance: GV）を求める．静的特徴量ベクトル系列 $\{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_T\}$ の GV ベクトル $\boldsymbol{v}_{\boldsymbol{y}} = [v_y(1), \cdots, v_y(D)]^\top$ は，次式で計算される．

$$v_y(d) = \frac{1}{T}\sum_{t=1}^{T}\left(y_t(d) - \frac{1}{T}\sum_{\tau=1}^{T} y_\tau(d)\right)^2 \quad (2)$$

本稿では，フレーズ単位で GV ベクトルを計算する．得られた GV ベクトルを学習データとして用いて，その確率密度関数を正規分布によりモデル化する．

$$P\left(\boldsymbol{v}_{\boldsymbol{y}}|\boldsymbol{\lambda}^{(v)}\right) = \mathcal{N}(\boldsymbol{v}_{\boldsymbol{y}}; \boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v) \quad (3)$$

ここで，$\boldsymbol{\lambda}^{(v)}$ は正規分布のパラメータセット（平均ベクトル $\boldsymbol{\mu}_v$ 及び共分散行列 $\boldsymbol{\Sigma}_v$）を表す．

2.2.2 変換処理

変換対象となる源歌手の歌声から抽出された静的・動的結合特徴量系列ベクトルを $\boldsymbol{X} = \left[\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_T^\top\right]^\top$ とする．また，これに対応する目標歌手の静的・動的結合特徴量系列ベクトルを $\boldsymbol{Y} = \left[\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_T^\top\right]^\top$ とし，静的特徴量系列ベクトルを $\boldsymbol{y} = \left[\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_T^\top\right]^\top$ とする．ここで，静的・動

的特徴量系列ベクトル $\boldsymbol{Y}$ と静的特徴量系列ベクトル $\boldsymbol{y}$ の間には，以下の関係が成り立つ．

$$\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y} \tag{4}$$

ここで，$\boldsymbol{W}$ は静的特徴量系列ベクトルから静的・動的特徴量系列ベクトルへの変換行列であり，動的特徴量を計算する際に用いる回帰係数を用いて決定される [22]．式 (2) 及び式 (4) を制約条件として，次式に示す目的関数を最大化する静的特徴量系列ベクトル $\boldsymbol{y}$ を求める．

$$\mathcal{L}(\boldsymbol{y}) = P(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\lambda})\, P\left(\boldsymbol{v}_{\boldsymbol{y}}|\boldsymbol{\lambda}^{(v)}\right)^{\omega} \tag{5}$$

ここで，条件付き確率密度関数 $P(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\lambda})$ は式 (1) で表される GMM により導出される．また，ω は GV の確率密度関数 $P\left(\boldsymbol{v}_{\boldsymbol{y}}|\boldsymbol{\lambda}^{(v)}\right)$ と $P(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\lambda})$ のバランスを調整する重みパラメータであり，本稿では両確率密度関数の次元数の比（$2T$）とする．静的・動的結合特徴量系列ベクトル $\boldsymbol{Y}$ 及び GV ベクトル $\boldsymbol{v}_{\boldsymbol{y}}$ は，共に静的特徴量系列ベクトル $\boldsymbol{y}$ から計算されるため，目的関数は $\boldsymbol{y}$ の関数となる．結果，静的特徴量，動的特徴量，及び，GV が適切となるような静的特徴量系列 $\boldsymbol{y}$ の推定が可能となり，動的特徴量により時間フレーム間相関を考慮した変換処理が実現され，GV により汎化処理に伴う音響特徴量の過剰な平滑化が効果的に抑えられる．

3. 話声に対する従来の多対多固有声変換

多対多固有声変換 [18] は，固有声変換技術 [23] の一つであり，任意の話者の音声を別の任意の話者の音声へと変換する手法である．任意の話者間の音響特徴量の対応関係は固有声 GMM（Eigenvoice GMM: EV-GMM）でモデル化する．本手法は，大量のパラレルデータセットを用いて事前に固有声 GMM を学習する事前学習処理，元話者と目標話者の音声に固有声 GMM を適応させる適応処理，また，適応固有声 GMM を使用して元話者の音声を目標話者の音声に変換する変換処理から構成される．

3.1 事前学習処理

学習にはまず，参照話者と多数の事前収録目標話者からそれぞれ同一内容の発話データを収録し，音響特徴量を抽出する．この時，時間フレーム t における参照話者と s 人目の事前収録目標話者の静的特徴量ベクトルをそれぞれ $\boldsymbol{x}_t = [x_t(1),\cdots,x_t(D)]^\top$, $\boldsymbol{y}_t^{(s)} = \left[y_t^{(s)}(1),\cdots,y_t^{(s)}(D)\right]^\top$ とし，静的・動的結合特徴量ベクトルをそれぞれ $\boldsymbol{X}_t = \left[\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top\right]^\top$, $\boldsymbol{Y}_t^{(s)} = \left[\boldsymbol{y}_t^{(s)\top}, \Delta\boldsymbol{y}_t^{(s)\top}\right]^\top$ とする．参照話者と各事前収録目標話者のペアに対して，時間フレームの対応付けを行うことで，静的・動的結合特徴量ベクトル対 $\left\{\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}\right\}$ を構築する．全ての事前収録目標話者に対する静的・動的結合特徴量ベクトル対を学習データとして用いて，話者適応学習（Speaker Adaptive Training: SAT）[24] を行うことで，固有声 GMM を学習する．参照話者と s 人目の事前収録目標話者の音響特徴量の結合確率密度関数をモデル化する固有声 GMM は，次式にて表される．

$$P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}^{(s)}\right) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix}\boldsymbol{X}_t \\ \boldsymbol{Y}_t\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y,s)}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)}\end{bmatrix}\right) \tag{6}$$

ここで，m 番目の分布における s 人目の事前収録目標話者に対する平均ベクトル $\boldsymbol{\mu}_m^{(Y,s)}$ は，次式で与えられる．

$$\boldsymbol{\mu}_m^{(Y,s)} = \boldsymbol{B}_m^{(Y)}\boldsymbol{w}^{(s)} + \boldsymbol{b}_{m,0}^{(Y)} \tag{7}$$

行列 $\boldsymbol{B}_m^{(Y)} = [\boldsymbol{b}_{m,1},\cdots,\boldsymbol{b}_{m,J}]$ 及びベクトル $\boldsymbol{b}_{m,0}$ は m 番目の分布の基底ベクトルセット（ベクトル数は J）及びバイアスベクトルであり，$\boldsymbol{w}^{(s)} = \left[w^{(s)}(1),\cdots,w^{(s)}(J)\right]^\top$ は s 人目の事前収録目標話者に対する J 次元の重みベクトルである．重みベクトルは個々の事前収録目標話者に依存するパラメータであり，全分布間で共有される．一方で，パラメータセット $\boldsymbol{\lambda}^{(EV)}$ は，個々の分布における混合重み α_m，参照話者の平均ベクトル $\boldsymbol{\mu}_m^{(X)}$，基底ベクトルセット $\boldsymbol{B}_m^{(Y)}$，バイアスベクトル $\boldsymbol{b}_{m,0}$，および，各共分散／相互共分散行列 $\boldsymbol{\Sigma}_m^{(XX)}$，$\boldsymbol{\Sigma}_m^{(XY)}$，$\boldsymbol{\Sigma}_m^{(YX)}$，$\boldsymbol{\Sigma}_m^{(YY)}$ から成り，全事前収録目標話者間で共有される．各分布の目標話者に対する平均ベクトル $\boldsymbol{\mu}_m^{(Y,s)}$ は，基底ベクトルで張られる部分空間上で表され，目標話者依存パラメータである重みベクトルを変化させることで，個々の分布の平均ベクトルがシフトし，参照話者と様々な話者間における結合確率密度関数が得られる．

3.2 適応処理及び変換処理

適応処理では，任意の元話者及び任意の目標話者の少量かつ任意の発話のみを用いて，それぞれ独立に固有声 GMM の話者依存重みベクトルを推定し，固有声 GMM を各話者に適応させる．任意の元話者 i の重みベクトル $\hat{\boldsymbol{w}}^{(i)}$ は次式により推定される．

$$\hat{\boldsymbol{w}}^{(i)} = \underset{\boldsymbol{w}}{\operatorname{argmax}} \prod_{t=1}^{T} \int P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(i)}|\boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}\right) \mathrm{d}\boldsymbol{X}_t \tag{8}$$

ここで，$\boldsymbol{Y}_t^{(i)}$ は，時間フレーム t における元話者 i の音響特徴量の静的・動的結合特徴量ベクトルである．同様に，任意の目標話者 o の重みベクトル $\hat{\boldsymbol{w}}^{(o)}$ も推定される．本推定処理では，各話者の音響特徴量のみしか用いておらず，言語情報などは一切必要としない．結果，元話者及び目標話者による任意の発話を使用することが可能となる．また，推定するパラメータ数（重みベクトルの次元数）は極めて少ないため，極少量の発話データのみでも頑健な推定処理が可能となる．

元話者 i と目標話者 o の音響特徴量の結合確率密度関数は，各話者に対して推定された重みベクトルを用いて適応

された結合確率密度関数に対して，次式の通り，参照話者の静的・動的結合特徴量ベクトル $\boldsymbol{X}_t$ の周辺化を行うことで導出される．

$$
\begin{aligned}
&P\left(\boldsymbol{Y}_t^{(i)}, \boldsymbol{Y}_t^{(o)} | \boldsymbol{\lambda}^{(EV)}, \hat{\boldsymbol{w}}^{(i)}, \hat{\boldsymbol{w}}^{(o)}\right) \\
&= \sum_{m=1}^{M} P\left(m|\boldsymbol{\lambda}^{(EV)}\right) \int P\left(\boldsymbol{Y}_t^{(i)}|\boldsymbol{X}_t, m, \boldsymbol{\lambda}^{(EV)}, \hat{\boldsymbol{w}}^{(i)}\right) \\
&\qquad P\left(\boldsymbol{Y}_t^{(o)}|\boldsymbol{X}_t, m, \boldsymbol{\lambda}^{(EV)}, \hat{\boldsymbol{w}}^{(o)}\right) P\left(\boldsymbol{X}_t|m, \boldsymbol{\lambda}^{(EV)}\right) \mathrm{d}\boldsymbol{X}_t \\
&= \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix}\boldsymbol{Y}_t^{(i)} \\ \boldsymbol{Y}_t^{(o)}\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}_m^{(Y,i)} \\ \boldsymbol{\mu}_m^{(Y,o)}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(YXY)} \\ \boldsymbol{\Sigma}_m^{(YXX)} & \boldsymbol{\Sigma}_m^{(YY)}\end{bmatrix}\right) \quad (9)
\end{aligned}
$$

ここで，

$$
\boldsymbol{\Sigma}_m^{(YXY)} = \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{\Sigma}_m^{(XX)^{-1}} \boldsymbol{\Sigma}_m^{(XY)} \quad (10)
$$

である．

変換処理では，元話者と目標話者に適応された固有声 GMM を用いて，新たに収録された元話者の音声を目標話者の音声へと変換する．具体的な変換処理は，2.2.2 節と同様である．ただし，GV の確率密度関数に関しては，予め全事前収録目標話者の GV ベクトルを用いて不特定モデルを構築しておき，その平均ベクトルのみを目標話者のものへと置き換える．この時，目標話者の GV の平均ベクトルは，適応処理に用いた目標話者の発話データから計算する．

4. 提案法

本提案法では，誰でも数秒の歌声を収録するだけで，様々な声質に自分の歌声の声質を変換できるようにするために，従来話声に対してのみ適用されてきた多対多固有声変換 [18] を歌声声質変換に導入する．これまで多対多固有声変換を歌声に適用困難だったのは，学習データ生成での歌声収録の難しさが原因であったが，その問題を解決する効率的な学習データ生成のために，歌声合成システム VocaListener [6], [7] を用いた新たな学習データ生成法も提案する．図 2 に提案法の学習及び適応過程を示す．

4.1 多対多固有声変換に基づく歌声声質変換

多対多固有声変換に基づく歌声声質変換では，多数のパラレルデータセット（図 2 左上）を用いて，予め固有声 GMM の学習（図 2 右上）を行う．それによりユーザが使用する際には，図 2 右下のように，任意の源歌手と任意の目標歌手の少量の歌声を使用して固有声 GMM を適応させるだけで，その源歌手から目標歌手への変換モデルが構築可能となる．

図 1 に示した従来法では，源歌手から目標歌手への変換モデルの構築のために，両歌手が歌唱した同一曲が数分程度必要であったのに対して，提案法では，両歌手の歌声が数秒程度あれば十分に変換できる特長を持つ．また従来法では，源歌手と目標歌手の二人が同一曲を歌唱することが必須であったのに対し，提案法では，それぞれが別の曲のワンフレーズを歌唱した歌声さえあればよい．これは提案法が，両歌手の歌声からそれぞれ独立に推定した重みベクトルを使用して固有声 GMM の適応を行うためである．このため提案法は，システム使用時におけるユーザ（源歌手）の事前準備の手間を大きく削減する長所と，様々な目標歌手の声質への変換が容易になる長所を併せ持っている．

例えば，ユーザがある演歌歌手の声質で歌いたい場合，従来法ではその演歌歌手の無伴奏歌唱を数分程度入手した上で，得意不得意にかかわらず，ユーザは同一の演歌を歌わなくてはならなかった．それに対して提案法では，ユーザはどんな曲を歌ってもよく，かつ，数秒程度の歌唱でいいため，従来法に比べてユーザの負担が格段に小さい上，ユーザの歌唱技術も問題にならない．

しかも提案法では，従来法と同じく，短遅延変換アルゴリズム [17] を用いることで，リアルタイム変換処理が可能である．これにより提案法も，楽曲制作時だけでなく，ライブやコンサート，カラオケなどのリアルタイム性が求められる様々な歌唱場面で使用できる．

4.2 VocaListener による学習データ生成

多対多固有声変換に基づく歌声声質変換では，固有声 GMM さえ学習しておけば，ユーザは上記のように容易にシステムを使用できる．しかし，固有声 GMM の学習には多数の歌声を含むパラレルデータセットの構築が困難であるという問題がある．

学習用のパラレルデータセットの構築では，まず，事前収録目標歌手の歌声として，様々な声質の歌手が，それぞれに何らかの楽曲を歌った無伴奏の歌声をできるだけ多く用意する必要がある．我々は RWC 研究用音楽データベース [25], [26] の楽曲を使用したが，Web サービス等では歌声のみが公開されている場合もあり，そうした歌声も使用できる可能性がある．それに対し，最も難しいのは，そうした事前収録目標歌手の歌声と対となる参照歌手の歌声の収録である．ある一人の参照歌手が，それらすべての楽曲を歌う必要があり，大きな負担となる．参照歌手の歌唱技術や発声可能な音域によっては，そもそも的確に歌唱できない場合もある．

そこで本稿では，VocaListener を利用して参照歌手の歌声を人工的に合成してパラレルデータセットを生成する手法を提案する．図 3 に，参照歌手の歌声として，従来の自然歌声を用いた場合と我々の VocaListener を用いた場合のパラレルデータ生成法を対比して示す．自然歌声を使用する場合，参照歌手は事前収録目標歌手の人数分だけ歌声を収録しなくてはならない．しかし必ずしもうまく歌唱できるとは限らず，曲毎に声質が大きく変動して歌声変換に悪影響を及ぼす可能性がある．一方，提案法では，事前収録目標歌手の歌声さえあれば，VocaListener で常に同一の

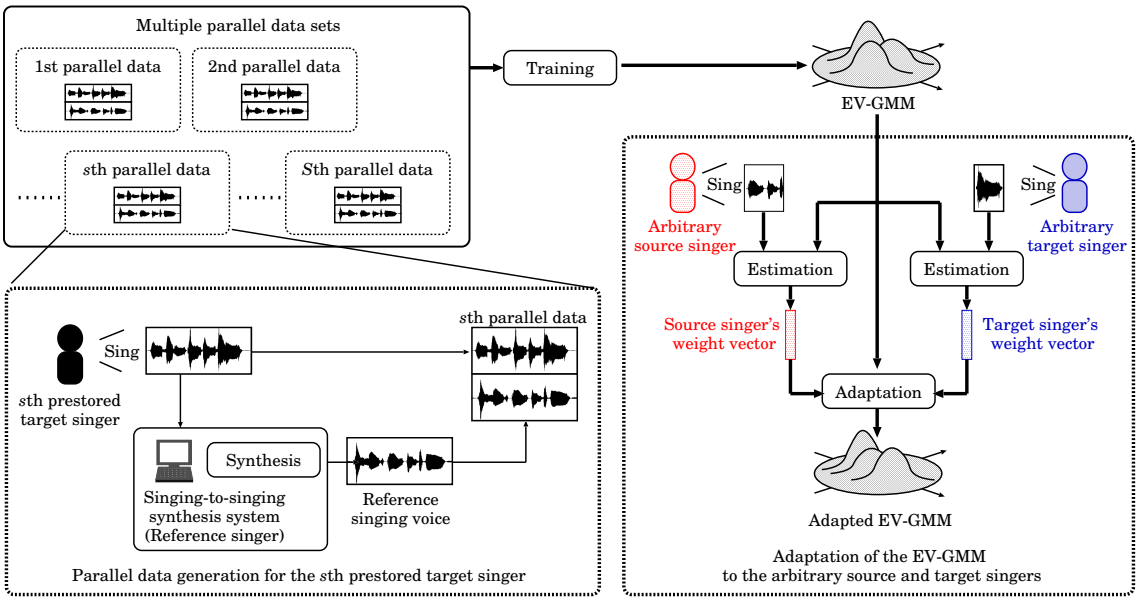

図 2 提案法の学習過程及び適応過程

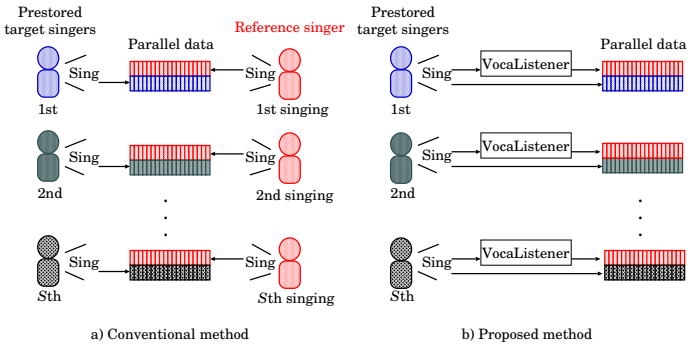

図 3 自然歌声と合成歌声を用いた場合のパラレルデータ生成法の違い

声質の歌声合成音源（歌声ライブラリ）を用いてその歌声を模倣するように合成することで，参照歌手の歌声を全ての楽曲において同じ声質で用意することが可能である．しかも曲毎に声質が大きく変動することが自然歌声よりも少ない．さらに人間と違って，単に同じ曲を歌っているというだけではない，歌い回しまでも真似たより高品質なパラレルデータセットとなる．従来法では，参照歌手と目標歌手の音響特徴量を結合する際に，動的時間伸縮法により時間フレーム間の対応付けで誤差を生じる可能性があるが，提案法で VocaLitener を用いると時間軸が一致しているので，その対応付けが不要で誤差を生じにくくなる．

以上のように本学習データ生成法では，まず，多数の事前収録目標歌手の歌声を準備し，次に，それぞれに対して VocaListener を用いて合成歌声を生成し，最後に，生成された合成歌声とその元となった事前収録目標歌手の歌声の音響特徴量を結合して学習データとする．

5. 実験による評価

提案法の有効性を客観的及び主観的に評価する．

5.1 実験条件

事前収録目標歌手の歌声として，RWC 研究用音楽データベース（ポピュラー音楽 RWC-MDB-P-2001） [25], [26] 中の 30 曲（男性歌唱 19 曲と女性歌唱 11 曲）の無伴奏歌唱を用いる．また，参照歌手の歌声として，歌声合成システム VOCALOID2（初音ミク [27]）を用いて，事前収録目標歌手の歌声を手本に VocaListener で自動推定された合成パラメータに基づいて生成される合成歌声を使用する．適応及び評価に用いる歌声として，RWC 研究用音楽データベースの中から，学習に使用されていない 2 曲（同一歌手による RWC-MDB-P-2001 No.35 及び No.71）の無伴奏歌唱と，これら 2 曲を新たに別の女性歌手 1 名が歌うのを収録した歌声を用いる．

スペクトル特徴量として，STRAIGHT 分析 [28] により抽出された 1 次から 24 次のメルケプストラム係数を用いる．シフト長は 5 ms，サンプリング周波数は 16 kHz とする．

提案法において，スペクトル変換用固有声 GMM は，30 人（上記 RWC 研究用音楽データベースの 30 曲）の事前収録目標歌手の歌声と，それらを VocaListener で変換した参照歌手の歌声から成るパラレルデータセットから学習される．固有声 GMM の重みベクトルの次元数は 29 とし，混合数は 128 とする．また，比較対象として，2.2 節で述べた従来法を用いる．従来法におけるスペクトル変換用 GMM の学習には，提案法における固有声 GMM の適応データとして用いる源歌手及び目標歌手の歌声と同一のものを用いる．ただし，提案法とは異なり，源歌手および目標歌手の歌声はパラレルデータとして取り扱う．また，従来法における GMM の混合数は，評価データに対する変換精度が最大になるように，事後的に最適化する．尚，本実験では，

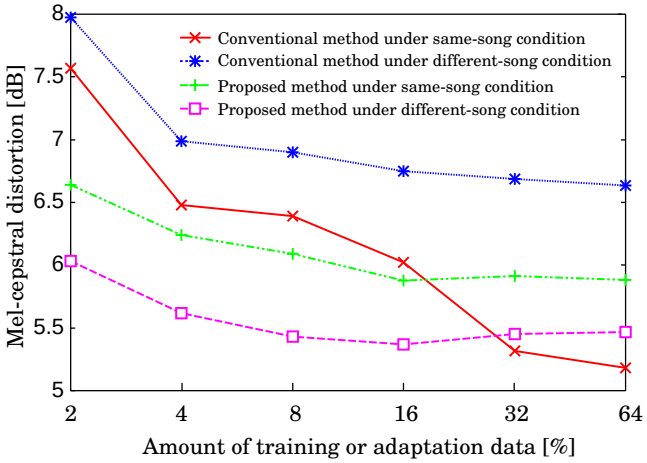

図 4 メルケプストラムひずみ．横軸は，従来法では学習に用いたデータ量，提案法では適応に用いたデータ量を示す．縦軸は，メルケプストラムひずみを示す

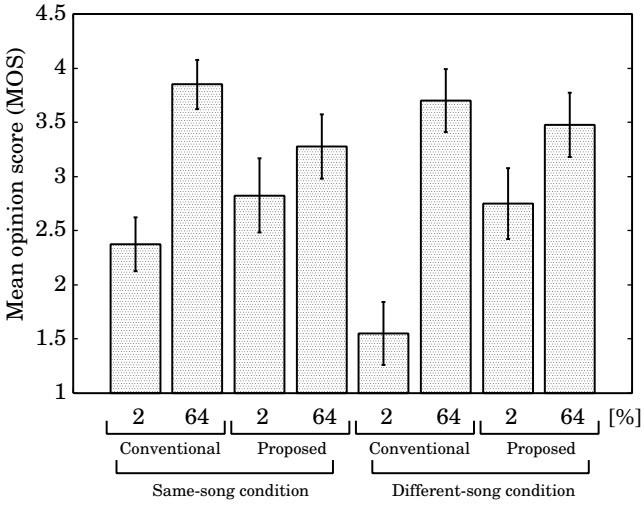

図 5 音質に関する主観評価結果

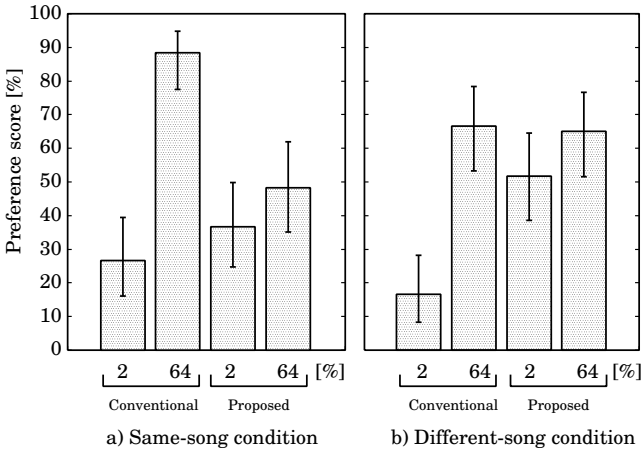

図 6 話者性に関する主観評価結果

従来法，提案法共に，短遅延変換アルゴリズムは用いず，バッチ処理に基づく変換アルゴリズム [16] により生成された変換歌声を用いる．

従来法の学習データ及び提案法の適応データとして，1 曲 (RWC-MDB-P-2001 No.35) に含まれる歌声中の 2, 4, 8, 16, 32, 64% を用い，残りの 36%を評価データとする．尚，楽曲の長さは 193 秒であり，その内，歌声の区間は 116 秒（100%に相当）である．

本稿では，客観評価及び主観評価を以下の 2 つの条件下で行う．

- 1) same-song condition: 学習・適応で用いた曲と同一の曲 (RWC-MDB-P-2001 No.35) を評価データとして使用する．
- 2) different-song condition: 学習・適応で用いた曲と同一歌手ではあるが異なる曲 (RWC-MDB-P-2001 No.71) を評価データとして使用する．

5.2 客観評価

従来法及び提案法の変換精度をメルケプストラムひずみにより評価する．図 4 に従来法及び提案法における 2 つの条件下での変換精度を示す．

same-song condition において，学習及び適応データが少ない場合 (16%以下の場合)，提案法は従来法よりも高い変換精度を示しており，提案法がデータ量に対して頑健であることが分かる．一方，学習及び適応データが多い場合 (64%の場合)，従来法が提案法よりも優れた変換精度を示している．これは，学習すべきパラメータが多い従来法においても，十分な学習データが得られたことにより，その楽曲における両歌手の声質をより効果的に表現できたためだと考えられる．尚，提案法は適応処理において，源歌手及び目標歌手の重みベクトルをそれぞれ独立に推定するため，パラレルデータを必要とせず，利便性においては，明らかに従来法に優っている．

different-song condition においては，データ量に関係なく提案法が従来法に優っている．これは，歌声の声質が，同一歌手による歌唱であっても，その曲調の違いに応じて変化するためであると考えられる．従来法では，学習曲における両歌手の声質に特化した GMM が学習されるため，別曲で評価した際に，性能が劣化する．一方で，提案法は，多数の事前収録目標歌手の声質をモデル化するように固有声 GMM が学習されるため，同一歌手の曲ごとの声質の変動に対しても頑健であると考えられ，別曲での評価においても性能劣化は小さい．ただし，提案法において different-song condition における変換精度が same-song condition に優っている点については，単に評価に用いる曲の違いに起因する特徴量分析精度や推定精度の差である可能性があり，必ずしも different-song condition の優位性を示すものであるとは限らない．

5.3 主観評価

主観評価では，各条件・手法における変換音声の音質及

び話者性を評価する．音質の評価は，5 段階の平均オピニオン評定による聴取実験で行う．評価する音声は，same-song condition 及び different-song condition において，2%または 64%の学習・適応データを用いた場合の従来法と提案法で生成された計 8 種類の変換歌声である．被験者は 5 名で，各被験者は，ランダムに提示される変換歌声サンプルを受聴し，その音質を 1(悪い)〜5(良い) の 5 段階で評価する．話者性の評価は，XAB 法による聴取実験で行う．ここで，話者性とは，声質における個人性を指す．評価対象は，音質評価と同じ 8 種類の変換歌声であり，被験者は 5 人である．被験者はまず，目標とする自然歌声を聴き，そののち 2 種類の変換歌声を聴く．そして，2 種類の内，より目標歌声に声質が近い変換歌声を選ぶ．尚，same-song condition と different-song condition は，目標とする曲が異なるため，それぞれ独立に評価する．

図 5 に音質の評価結果を示す．same-song condition において，提案法は 2%のデータを用いた際の音質が，従来法よりも高いこと，データ量の増加に伴い音質が改善することが分かる．しかしながら，データ量が多い場合，提案法の音質はパラレルデータで学習する従来法の音質に及ばない．これは，客観評価結果と同様である．different-song condition において，64%のデータを用いた際に，従来法は，same-song condition に近い音質を示しているにも関わらず，2%のデータを用いた際には，same-song condition よりも明らかに低い音質を示している．この事から，同一歌手においても，曲が異なる場合には，局所的にその声質が大きく変動することが窺える．一方，提案法は，客観評価と同じく，両条件下で同等の音質を示しており，曲ごとの声質の違いに対して頑健であることが確認できる．また，different-song condition において，提案法はパラレルデータで学習した従来法に匹敵する音質を示しており，当条件下における提案法の優位性は明らかである．

図 6 に話者性の評価結果を示す．話者性の評価においても，客観評価及び音質評価と同様の傾向を確認できる．すなわち，従来法は，源歌手と目標歌手のパラレルデータが大量に利用可能な場合においては，精度良く変換処理を実現することができるが，十分な量のパラレルデータが得られない際には，その変換精度は激しく劣化する．一方で，提案法は，データ量及び曲ごとの声質の差に非常に頑健であり，任意の少量のデータのみを用いて，源歌手と目標歌手の声質変換を比較的精度良く実現することができる．

以上の結果から，提案法は，利用可能な歌声が少量の場合であっても，高い変換性能を示すこと，適応に用いた曲と変換時の曲が異なる場合でも，頑健に変換可能であることが分かる．また，提案法が適応データとしてパラレルデータを必要としないことも，提案法の重要な利点の一つである．

6. まとめ

本稿では，混合音ではない無伴奏の独唱において，任意のユーザの歌声の声質を様々な歌手の声質に自動変換できる歌声声質変換手法を提案した．本手法では多対多固有声変換を導入したことで，変換前後の無伴奏歌唱データが少量あれば，事前に学習した固有声 GMM を適応させて変換に用いることを可能にした．これにより従来の歌声声質変換に比べ，より容易に幅広い場面で用いることができ，様々な声質への変換が実現できた．この優れた多対多固有声変換は，従来であれば歌声に適応することは現実的でなかったが，VocaListener を用いてパラレルデータセットを構築する斬新な学習データ生成により，固有声 GMM の学習を歌声でも可能にした．以上の提案法は，客観評価及び主観評価の結果から，高い変換精度を保ちつつ，かつ，利便性を大幅に向上できることが示された．

我々は既に，本手法に基づいて，マイクから入力されたユーザ（源歌手）の歌声の声質を，リアルタイムに他の声質に変換するプロトタイプシステムを試作した．しかし，歌声声質変換後の声質にはまだ改善の余地が大きく，今後もさらなる変換品質の向上とユーザの立場からの利便性の向上に取り組んでいくことを予定している．

謝辞 本研究の一部は，科研費補助金若手研究 (A) と科学技術振興機構 OngaCREST プロジェクトによる支援を受けた．STRAIGHT の使用を許可していただいた和歌山大学河原英紀教授に感謝いたします．

参考文献

[1] 後藤真孝, 奥乃博. 特集「CGM の現在と未来: 初音ミク，ニコニコ動画，ピアプロの切り拓いた世界」 編集にあたって. 情報処理 (情報処理学会誌), Vol. 53, No. 5, pp. 464–465, May 2012.

[2] H. Kenmochi and H. Ohshita. VOCALOID – Commericial singing synthesizer based on sample concatenation. *Proc. INTERSPEECH*, pp. 4011–4012, Aug. 2007.

[3] 剣持秀紀, 大下隼人. 歌声合成システム VOCALOID — 現状と課題. 情報処理学会研究報告 音楽情報科学, Vol. 2008-MUS-74-9, pp. 51–58, 2008.

[4] K. Oura, A. Mase, T. Yamada, S. Muto, Y. Nankaku, and K. Tokuda. Recent development of the HMM-based singing voice synthesis system - Sinsy. *SSW7*, pp. 211–216, Sept. 2010.

[5] 徳田恵一, 大浦圭一郎. 自動学習により人間のように歌う音声合成システム —Sinsy—. 情報処理学会研究報告 音楽情報科学, Vol. 2012-MUS-94, No. 1, pp. 1–6, 2012.

[6] T. Nakano and M. Goto. VocaListener: A singing-to-singing synthesis system based on iterative parameter estimation. *Proc. SMC 2009*, pp. 343–348, July 2009.

[7] 中野倫靖, 後藤真孝. VocaListener: ユーザ歌唱の音高および音量を真似る歌声合成システム. 情報処理学会論文誌, Vol. 52, No. 12, pp. 3853–3867, Dec. 2011.

[8] T. Nakano and M. Goto. Vocalistener2: A singing synthesis sytem able to mimic a user's singing in terms of voice timbre changes as well as pitch and dynamics. *Proc.*

ICASSP, pp. 453–456, May 2011.
[9] 中野倫靖, 後藤真孝. VocaListener2: ユーザ歌唱の音高と音量だけでなく声色変化も真似る歌声合成システムの提案. 情報処理学会研究報告 音楽情報科学, Vol. 2010–MUS–86, No. 3, pp. 1–10, July 2010.
[10] 川上裕司, 坂野秀樹, 板倉文忠. 声道断面積関数を用いた GMM に基づく歌唱音声の声質変換. 信学技法, SP 110–297, pp. 71–76, Nov. 2010.
[11] 藤原弘将, 後藤真孝. 混合音中の歌声スペクトル包絡推定に基づく歌声の声質変換手法. 情報処理学会研究報告 音楽情報科学, Vol. 2010-MUS-86, No. 7, pp. 1–10, 2010.
[12] 河原英紀, 生駒太一, 森勢将雅, 高橋徹, 豊田健一, 片寄晴弘. モーフィングに基づく歌唱デザインインタフェースの提案と初期検討. 情報処理学会論文誌, Vol. 48, No. 12, pp. 3637–3648, 2007.
[13] H. Kawahara, R. Nisimura, T. Irino, M. Morise, T. Takahashi, and H. Banno. Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown. *Proc. ICASSP*, pp. 3905–3908, Apr. 2009.
[14] Y. Stylianou, O. Cappe, and E. Moulines. Continuous probabilistic transform for voice conversion. *IEEE Trans. SAP*, Vol. 6, No. 2, pp. 131–142, Mar. 1998.
[15] A. Kain and M. W. Macon. Spectral voice conversion for text-to-speech synthesis. *Proc. ICASSP*, pp. 285–288, May 1998.
[16] T. Toda, A. W. Black, and K. Tokuda. Voice conversion based on maximum likelihood estimation of spectral parameter trajectory. *IEEE Trans. ASLP*, Vol. 15, No. 8, pp. 2222–2235, Nov. 2007.
[17] T. Muramatsu, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano. Low-delay voice conversion based on maximum likelihood estimation of spectral parameter trajectory. *Proc. INTERSPEECH*, pp. 1076–1079, Sept. 2008.
[18] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano. Many-to-many eigenvoice conversion with reference voice. *Proc. INTERSPEECH*, pp. 1623–1626, Sept. 2009.
[19] F. Villavicencio and J. Bonada. Applying voice conversion to concatenative singing-voice synthesis. *Proc. INTERSPEECH*, pp. 2162–2165, Sept. 2010.
[20] T. Saitou, M. Goto, M. Unoki, and M. Akagi. Speech-to-singing synthesis: Converting speaking voices to singing voices by controlling acoustic features unique to singing voice. *Proc. WASPAA*, pp. 215–218, Oct. 2007.
[21] 齋藤毅, 後藤真孝, 鵜木祐史, 赤木正人. SingBySpeaking: 歌声知覚に重要な音響特徴を制御して話声を歌声に変換するシステム. 情報処理学会研究報告 音楽情報科学, Vol. 2008–MUS–74–5, No. 12, pp. 25–32, 2008.
[22] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. Speech parameter generation algorithms for HMM-based speech synthesis. *Proc. ICASSP*, pp. 1315–1318, June 2000.
[23] T. Toda, Y. Ohtani, and K. Shikano. One-to-many and many-to-one voice conversion based on eigenvoices. *Proc. ICASSP*, pp. 1249–1252, Apr. 2007.
[24] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano. Adaptive training for voice conversion based on eigenvoices. *IEICE Trans. Inf. and Syst.*, Vol. E93-D, No. 6, pp. 1589–1598, June 2010.
[25] 後藤真孝, 橋口博樹, 西村拓一, 岡隆一. RWC 研究用音楽データベース：研究目的で利用可能な著作権処理済み楽曲・楽器音データベース. 情報処理学会論文誌, Vol. 45, No. 3, pp. 728–738, Mar. 2004.
[26] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC Music Database: Popular, classical, and jazz music databases. *Proc. ISMIR*, pp. 287–288, Oct. 2002.
[27] 伊藤博之. 初音ミク as an interface. 情報処理学会誌, Vol. 53, No. 5, pp. 477–482, May 2012.
[28] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne. Restructuring 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, Apr. 1999.