Vol.2014-MUS-104 No.23
2014/8/27

混合音中の歌声 F0 軌跡に対する歌唱表現転写システム

池宮 由楽[1,a)]　糸山 克寿[1,b)]　吉井 和佳[1,c)]　奥乃 博[2,d)]

概要：本稿では，音楽音響信号に含まれる歌声の基本周波数 (F0) 軌跡に対して歌唱表現（ビブラート・グリッサンド・こぶし）を転写することを可能とするシステムを提案する．能動的音楽鑑賞インタフェースは，エンドユーザのインタラクティブな音楽鑑賞を実現することを目的とした研究アプローチである．これには既存楽曲の加工支援も含まれ，歌声に関連するものでは，声質変換や歌声分離などの研究がなされている．本研究では，歌唱の歌い回しの加工を扱い，特に混合音中の歌声の F0 軌跡を任意に編集するインタフェースを実現する．ユーザは，歌声の任意の箇所を指定し，好みの歌唱表現を転写することで，歌い回しを自由に加工することができる．また，事前に市販楽曲からプロ歌手の歌唱表現を蓄積したデータベースを作成し，ユーザはそのデータベースから歌唱表現を参照することで直感的に転写を行うことが可能となる．歌唱表現の転写は，対数周波数軸において選択的に歌声のスペクトルのみをシフトさせ，伴奏音への影響を抑圧しながら歌声の音高を操作することで行われる．このとき，音韻性を保持するためスペクトル包絡を用いて音色の補正を行う．実際にユーザが表現の転写箇所を指定したり，F0 の存在範囲を提示するため，Graphical User Interface (GUI) の作成を行っている．実験では，音色補正の有効性やユーザ入力を用いた F0 推定の頑健性などを確認した．

1. はじめに

エンドユーザによる音楽の聴き方をより豊かにするため，能動的音楽鑑賞インターフェース [1] が提案されている．従来の音楽鑑賞は CD といったオーディオメディアを再生するだけの受動的な楽しみ方が一般的であり，能動的な楽しみ方としても，自分好みのプレイリストを作成したりイコライザにより周波数特性を調整するなどの簡単な処理によるものがほとんどであった．しかし，昨今の音楽音響信号処理技術の発展は著しく，一般のエンドユーザであっても，楽曲の内容に基づいた能動的でインタラクティブな鑑賞を楽しむことが可能となってきている．特に，従来は不可能であった既存楽曲の音楽的内容の編集への応用が多く研究されており，例えば，ドラムパートの音量や音色，パターンを MIDI シーケンサのように編集したり [2]，調波楽器の音量バランスを楽器別に調整する [3, 4]，歌声パートのみを分離するといったことが可能である [5]，

歌声はポピュラー楽曲における主旋律を司るパートであり，その音高・音量・音色に演奏者（歌唱者）の個性をより強く反映するため，多くの分析，編集技術が存在する．例えば，音声分析合成システムである TANDEM-STRAIGHT [6] は，歌声から基本周波数 (F0) とスペクトル包絡，非周期性指標の 3 パラメータを推定し，それぞれを独立に操作した後，高品質に再合成することが可能である．大石ら [7] は歌声 F0 軌跡を確率モデルで表現し，既存の無伴奏歌唱から楽譜のコンテキストと動的変動成分の関係を学習することで，任意楽譜から学習した歌唱者に対応する F0 軌跡を生成している．また，同様のモデルを歌声音量軌跡にも適用している [8]．これらは全て無伴奏歌唱を対象としている．藤原ら [9] は混合音中の伴奏と歌声のスペクトルを同時にモデル化することで，歌声のみの音響信号を明示的に分離することなく，歌声の声質変換を実現している．

本稿では，市販楽曲などの混合音中に含まれる歌声の F0 軌跡に対してユーザが任意に歌唱表現を転写することのできるシステムを提案する．歌声 F0 軌跡の変化パターンは歌唱者の個人性を表すとともに，歌唱の巧拙感に寄与している [10]．混合音中の歌声 F0 軌跡を自由に操作するインターフェースを作成することで，ユーザは任意の歌手の歌い方を自分好みに編集して楽しむことが可能となる．混合音中には歌声と同時に伴奏音が存在しているため，歌声の音高のみを変更する処理が必要となる．そこで本稿では，時間周波数領域において歌声のスペクトルのみを選択的にシフトすることで，これを実現する．

また，ユーザは転写する歌唱表現を事前に用意された

[1] 京都大学大学院情報学研究科
[2] 早稲田大学大学院創造理工学研究科
[a)] ikemiya@kuis.kyoto-u.ac.jp
[b)] itoyama@kuis.kyoto-u.ac.jp
[c)] yoshii@kuis.kyoto-u.ac.jp
[d)] okuno@aoni.waseda.jp

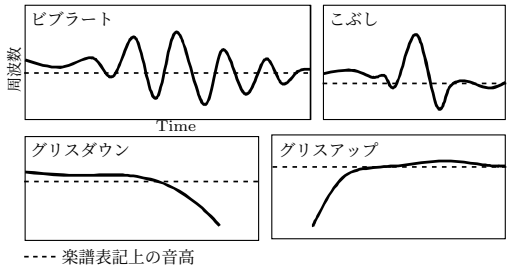


図 1 歌唱表現.

データベースの中から選択する．我々はこれまで，市販楽曲からプロ歌手の歌唱表現をデータベース化するための手法を開発してきた [11]．歌声 F0 軌跡の特徴的な変動であるビブラート・グリッサンド・こぶし (図 1) を歌唱表現として同定し，パラメータとして保存することでデータベースを構築する．転写時には，パラメータから歌唱表現が再合成される．ユーザはプロ歌手の歌唱表現を参照することで，最適なパラメータを手動で探索するなどの労力を必要としなくなる．もちろん，歌唱表現のパラメータを細かく設定したいユーザのため，任意のパラメータから合成を可能とするインタフェースを用意することは有用であると考えられる．実際に，ユーザが歌唱表現転写箇所や F0 の存在範囲の提示し，インタラクティブに操作を行うことを可能とする GUI を試作している．

2. 混合音中の F0 軌跡に対する歌唱表現転写

本章では，混合音中の歌声 F0 軌跡に対して歌唱表現を転写する手法について記述する (図 2)．入力歌唱はまず，定 Q 変換 [12] により時間周波数領域（スペクトログラム）へと変換される．時間周波数領域において歌声の F0 軌跡を操作した後，定 Q 逆変換により時間領域信号を得る．

2.1 定 Q 変換

時間領域信号 $x(n)$ に対する定 Q 変換 [12] は以下の式で定義される．

$$X(n,k) = \frac{1}{N_k} \sum_{j=n-\lfloor N_k/2 \rfloor}^{n+\lfloor N_k/2 \rfloor} x(j) a_k^*(j-n+N_k/2) \quad (1)$$

$$\begin{cases} a_k(n) = w(n/N_k)\exp(-i2\pi n f_k/f_s) \\ N_k = Q\frac{f_s}{f_k}, \quad Q = (2^{1/\mathrm{fratio}}-1)^{-1}\mathrm{qrate} \end{cases}$$

ここで，k は対数周波数インデクス，f_k は k に対応する線形周波数 [Hz]，f_s はサンプリング周波数を表す．$w(t)$ は区間 $[0,1]$ で正規化された窓関数である．また，fratio は対数周波数の離散化において，1 オクターブをいくつに分割するかを表し，qrate は時間周波数分解能のトレードオフを決定する．実用上は，全ての時間サンプル n における対数スペクトルを求めるのではなく，例えば 10 [msec] などのホップサイズで切り出し計算する．以後分かりやすさのため，定 Q 変換スペクトログラムの時間インデクス，周波数インデクスをそれぞれ t，f とし，（振幅）スペクトログラム自体を $X(t,f)$ と記述する．

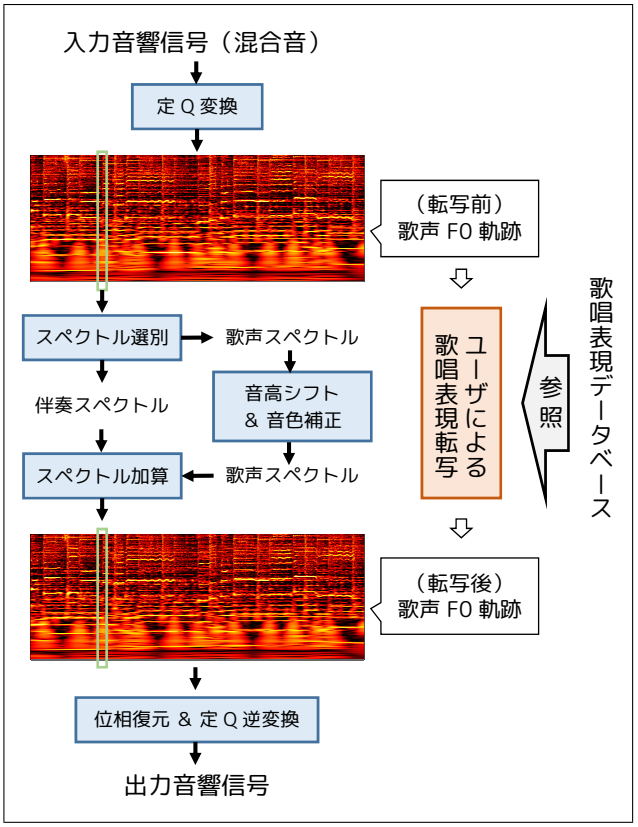


図 2 混合音中の F0 軌跡に対する歌唱表現転写.

2.2 Robust PCA を用いた歌声スペクトルの分離

Robust PCA (RPCA) [13] は行列（2 次元配列）を低ランク行列とスパース行列に分解する手法であり，以下で定式化される．

$$\text{minimize} \ \ \|L\|_* + \lambda\|S\|_1 \ \ (\text{subjectto} \ \ L+S=M) \quad (2)$$

ここで，M，L，S はそれぞれ入力行列，低ランク行列およびにスパース行列であり，$\|\cdot\|_*$，$\|\cdot\|_1$ はそれぞれ核ノルムと L1 ノルム，λ は低ランク性とスパース性のトレードオフパラメータを表す．一般に時間変化するデータ集合などを入力とし，頻出する成分が低ランク行列に，それ以外の成分がスパース行列に分解される．

RPCA を利用した歌声分離手法が提案されている [14]．つまり，混合音のスペクトログラムを入力とし，繰り返し演奏されるため多くの時間フレームで同じ形状をとる伴奏音（ドラムやギター）のスペクトルは低ランク行列へ，それ以外の歌声などのスペクトルはスパース行列へ分解される．分解した行列からバイナリマスクを作成し，元のスペクトログラムへ適用することで歌声が分離される．文献 [14] では短時間フーリエ変換によるスペクトログラムに対して処理を行っているが，本稿では対数周波数領域で処理を行うため，定 Q 変換スペクトログラムに対し同様の

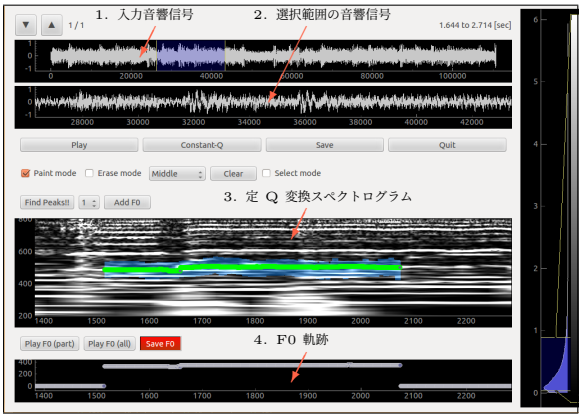


図 3 試作した GUI の画面．スペクトログラム上における，青い半透明の領域はユーザが提示した F0 存在範囲であり，緑の点は推定された F0 を示す．

アルゴリズムを適用する．

2.3 ユーザ入力を援用したスペクトルの選別

上記の手法により，ある程度の歌声分離は実現されるが完全ではない．特に，曲の一部でしか現れないベースやドラムのスペクトルは歌声とともにスパース行列へと分離されてしまう．そこで，ユーザの入力を援用し，歌声スペクトルを選別する処理を考える．ユーザがスペクトログラム上で歌声 F0 の存在範囲を提示し，その範囲を探索することで歌声 F0 推定を行い，得られた F0 からマスキング処理を施すことで歌声スペクトルを選別する．

2.3.1 ユーザ入力を用いた F0 推定

歌声 F0 は混合音スペクトルから入力無し（ブラインド）で推定することも考えられるが，歌声以外のスペクトル成分が多く存在するため，推定誤りを完全に抑制することは難しい．また，一部の誤りであっても大きな歪みを生む原因となり，転写の質を大きく下げることとなる．

推定誤りの少なく正確な F0 推定を実現するため，ユーザによる入力を援用することを考える．ユーザは表示されたスペクトログラム上において，領域を塗りつぶすことにより，F0 の存在する範囲を提示する．図 3 に試作した GUI 画面を示す．スペクトログラム上で，歌声の調波構造を視認し F0 付近を提示することは容易であると考えられる．提示された F0 存在範囲内を探索することにより歌声 F0 軌跡が推定される．

スペクトルの各周波数 c [cent] における「歌声の F0 らしさ」を表したコスト関数を $L_t(c)$ とし，ユーザの提示した F0 の存在する周波数範囲が $c_l \sim c_h$ [cent] とすると，F0 [cent] は以下の式で推定される．

$$F(t) = \arg\max_{c_l \le c \le c_h} L_t(c) \qquad (3)$$

対数周波数スペクトルから $L_t(c)$ を計算する手法はいくつか存在する [15, 16] が，本稿では計算の簡潔さ（計算コストの低さ）とノイズへの頑健性を兼ね備えた Subharmonic Summation (SHS) [17] を採用する．SHS を用いて $L_t(c)$ は，

$$L_t(c) = \sum_{n=1}^{N} \lambda^{n-1} S_t(c + 1200 \log_2 n) \qquad (4)$$

で与えられる．ここで，$S_t(c)$ は入力振幅スペクトル（連続表現）を表す．また，N は足し合わせる倍音数，λ は各倍音への重み係数を表し，本稿ではそれぞれ 15, 0.84 と設定する．

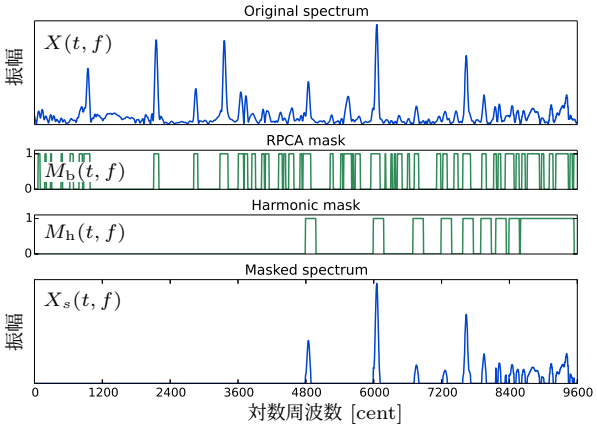


図 4 マスキングによる歌声スペクトルの選別．上から，元の混合音スペクトル，RPCA によるバイナリマスク，F0 を援用した倍音マスク，マスキングにより選別された歌声スペクトルを示している．

2.3.2 F0 軌跡を用いたマスキング処理

F0 軌跡を入力として，さらに歌声のスペクトルを選別する (図 4)．

$$M_h(t,f) = \begin{cases} 1 & \left[\begin{array}{l} H_t^h - \frac{w}{2} < C(f) < H_t^h + \frac{w}{2} \\ H_t^h = F_t + 1200 \log_2 h,\ 1 \le h \le H \end{array} \right. \\ 0 & otherwise \end{cases} \qquad (5)$$

ここで，F_t は時間フレーム t における F0 [cent]，$C(f)$ は周波数ビン f に対応する対数周波数 [cent]，H は倍音数，w は各倍音でマスクを取る幅 [cent] を示す．RPCA によるバイナリマスクを $M_b(t,f)$ と置くと，歌声と伴奏のスペクトログラム $X_s(t,f)$，$X_m(t,f)$ はそれぞれ以下のように得られる．

$$\begin{aligned} X_s(t,f) &= M_b(t,f) M_h(t,f) X(t,f), \\ X_m(t,f) &= (1 - M_b(t,f) M_h(t,f)) X(t,f) \end{aligned} \qquad (6)$$

2.4 音色補正を用いた音高シフト

線形周波数軸 [Hz] において音高を n 倍する処理はその周波数に関わらず，対数周波数軸 [cent] において $1200 \log_2 n$ [cent] 加算する処理に対応する．つまり，対数周波数領域でスペクトル全体をシフトすることで，音高シフトが可能

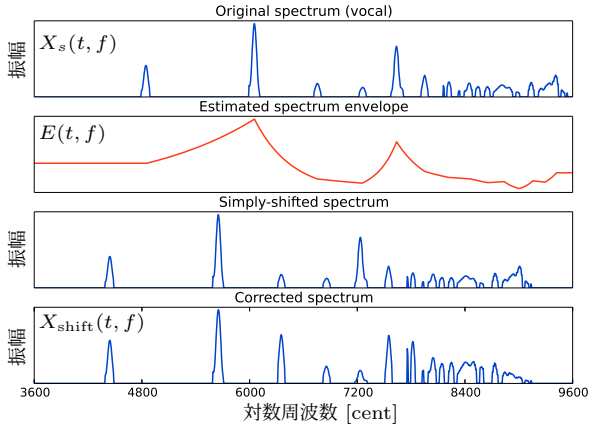


図 5 スペクトル包絡を用いた音色補正．上から，元の歌声スペクトル，推定されたスペクトル包絡，単純にシフトしたスペクトル，スペクトル包絡を用いて音色を補正したスペクトルを示している．

である．本稿では，歌声のみ音高シフトを行うため，2.3 節において選別した歌声スペクトル $X_s(t,f)$ をシフトし，伴奏スペクトル $X_m(t,f)$ と合成する．

音声のスペクトル包絡には音韻性や音色の情報が含まれているため [6]，スペクトル全体を単純にシフトすると，本来保存されるべきスペクトル包絡もシフトされてしまい，不自然な音となってしまう．そこで，シフトする前の歌声スペクトルから擬似的にスペクトル包絡を推定し，倍音の比率を修正することで音色の補正を行う (図 5)．

倍音周波数のパワースペクトルの値を用いて，スペクトル包絡を推定する．離散的なスペクトルの値からスペクトル包絡を求めるため，離散全極型モデル (DAP) [18] などを利用することが考えられるが，本稿では簡単のため，対数スケールにおいて倍音周波数間を線形補間し，線形スケール（振幅スペクトル）へ戻すことでスペクトル包絡を得る．ただし，F0 以下の周波数の包絡は F0 部と同一にする．推定されたスペクトル包絡を対数周波数軸へスケーリングしたものを $E(t,f)$ とおく．ここでシフトする周波数ビン数を m とすると，音色の補正された歌声（振幅）スペクトルは以下の式で計算される．

$$X_{\mathrm{shift}}(t,f) = A_t\, X_s(t,f-m)\frac{E(t,f)}{E(t,f-m)} \qquad (7)$$

ただし，A_t は音高シフト前後の総スペクトルパワーを一定とするための正規化係数である．これより，最終的に得られる振幅スペクトルは以下である．

$$X_{\mathrm{new}}(t,f) = X_m(t,f-m) + X_{\mathrm{shift}}(t,f) \qquad (8)$$

2.5 時間領域への逆変換

定 Q 変換スペクトログラムから定 Q 逆変換 [12] により時間領域の信号を再構成することが可能である．ただし，2.4 節の音高シフトにより振幅スペクトルが変形されている．つまり，元の位相を用いて逆変換を行うことは適切ではなく，音の歪みの原因となる．そこで，定 Q 変換と逆変換を繰り返す位相復元法 [19] を変形後の振幅スペクトログラムに適用した後，逆変換を行うことで時間領域の信号を得る．

3. 歌唱表現データベース

本章では，ユーザが歌唱表現を参照するためのデータベースについて述べる [11]．本稿で扱う歌唱表現は, 歌声 F0 軌跡に含まれる特徴的な変動成分であるビブラート, こぶし, グリッサンドの 3 種である．ここで，ビブラートは F0 軌跡の周期的な振動，こぶしは短いメリスマ，グリッサンドはフレーズ初めの滑らかな音高上昇（グリスアップ）とフレーズ終りの滑らかな音高下降（グリスダウン）を表している．特に，こぶしは演歌や民謡といった日本の伝統的な歌曲で多用される．これらの歌唱表現を市販楽曲から抽出・蓄積することで，データベースを構築する．

入力歌唱音高列を援用して対象楽曲から歌声 F0 軌跡を推定し，次に，F0 軌跡上からパターンマッチングにより各表現を同定する．各歌唱表現の形状パターンは，簡潔に再合成が行えるようなパラメータとして定義されている．同定された各表現はパラメータとしてデータベースに保存され，転写時には保存パラメータから再合成が行われる．歌唱表現を保存する際，同定された音符の情報を同時に取得することで，楽譜を入力として与えられる場合の転写において，コンテキストの情報として利用することができる．しかし，本稿ではユーザが任意に指定した箇所へ転写を行うため，コンテキストの情報は利用できない．今後，蓄積されている歌唱表現をどのようにユーザへ提示するかを考える必要がある．

4. 実験

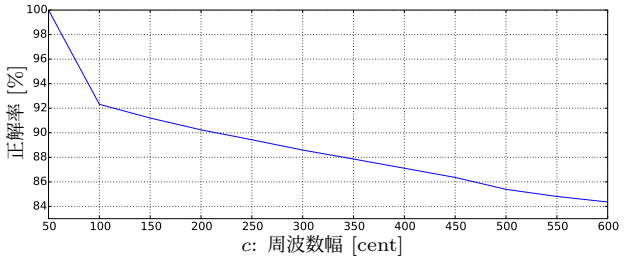
本章では，提案した音色補正の有効性，ユーザ入力を用いた F0 推定の頑健性を確認し，実際に音楽音響信号に対し歌唱表現を転写した結果を示す．

4.1 実験条件

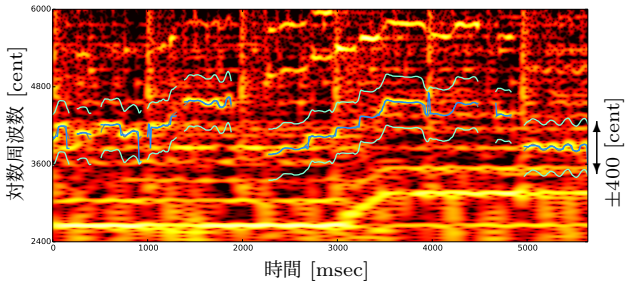
実験には 16kHz・16bit・モノラルの音響信号を用いる．また，定 Q 変換のパラメータは，fratio：0.005 (200 bins per octave)，qrate：0.2 とし，ホップサイズは 10 [msec] とする．RPCA を用いた歌声分離では，低ランク性とスパース性のトレードオフパラメータ k を決定する必要があるが [14]，本稿では実験的に 0.1 としている．また，2.3 節での各倍音のマスク幅 w は 120 [cent] とした．

4.2 提示する周波数幅の F0 推定への影響

2.3.1 節で述べた F0 推定において，ユーザは歌声 F0 の存在範囲をスペクトログラム上で提示する．そこで，提示する存在範囲幅がどの程度 F0 推定精度へ影響するかを調



(a) F0 推定精度



(b) F0 推定結果の例 ($c = 400$, RWC-MDB-P-2001: No.7)

図 6 ユーザの提示する周波数幅と F0 推定精度の関係．(b) において，灰色が正解 F0 軌跡，水色が推定 F0 軌跡，緑色が ±400 [cent] の幅を示している．

べた．実験には，“RWC Music Database: Popular Music” (RWC-MDB-P-2001) [20] から，ユニゾン歌唱や極端な音声加工（オートチューンなど）を含まないポピュラー楽曲 94 曲を用いた．正解 F0 から上下に $\pm c$ [cent] の幅を探索周波数範囲として F0 推定を行う．c の値を変化させることで F0 推定精度がどのように変化するかを調べる．

図 6 (a) に結果を示す．50 [cent] 以下の誤差を正解と判定し，推定精度は歌唱区間における正解率（全楽曲の平均）であるとする．結果を見ると，$c = 100$ [cent] で既に 10 % 弱の誤りが存在する．しかし実際には，これらはほとんど音符の遷移部など，極端に歌声の音量が小さくなる箇所で起こっていることを確認した．本稿で扱う歌唱表現はいずれも，一つの音符上で音を伸ばしている箇所に転写する性質のものであるため，これらの誤りは問題にならないと考えられる．また，c の変化に対して推定精度の下降は非常に緩やかであり，$c = 400$ [cent] においても 9 割弱の精度を保っている．これは，ユーザが正解 F0 から上下 4 半音の許容誤差をもって F0 存在範囲を提示（スペクトログラム上へのペイント）すればよいことを示しており，十分に実用的な値であると言える (図 6 (b))．

4.3 音色補正

2.4 節におけるスペクトル包絡を用いた音色補正の効果を調べる．ここでは純粋な音色補正の効果を調べるため，無伴奏歌唱を実験データとして用いる．女性 1 名が「い」「う」「お」という音韻を平坦に伸ばして発音している音声に対し，振幅が 100・200・300・400 [cent] のビブラート（周期は全て 6 [Hz]）を付加し，聴取実験によって音色の自然性を調べる．比較対象として，音色補正を行わず対数周波数軸で音高シフトした音声，TANDEM-STRAIGHT [6] による合成音声を用いる．TANDEM-STRAIGHT による音声は理想的に音色が補正されたものとして見ることができ，今回の評価値の上限を設定する役割を持つ．各音韻について，まず元音声を聴かせ，次に無作為な順番で 3 手法・3 振幅の音高シフト音声を聴かせ，「音色の自然性」について 5 段階で評価させる．被験者数は 7 名である．

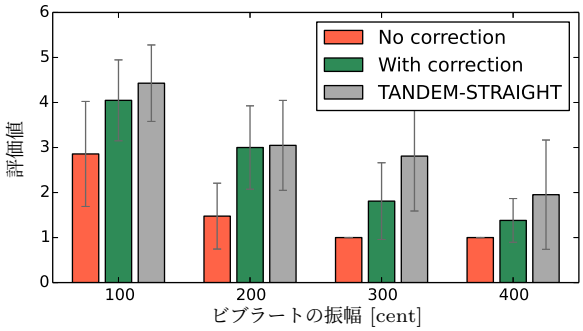


図 7 音色の自然性の評価．赤は音色補正なし，緑は音色補正あり，灰は TANDEM-STRAIGHT の結果を示す．評価値は全被験者・音韻の平均であり，エラーバーは標準偏差を表す．

図 7 は振幅毎の結果を示す．音色補正を行うことにより音色の自然性が大幅に向上していることが分かる．特に，200 [cent] 以下では，音色補正を行うことで TANDEM-STRAIGHT と比較的近い評価を得ている．多くのポピュラー歌手のビブラートが 200 [cent] 以下であることからも，音色補正は非常に有用であると言える．振幅が大きくなるに従い，TANDEM-STRAIGHT を含めた全体の音色自然性が下がっている．これは，同じ音韻・音符内であっても，音高を大きく変化させる表現をする場合，スペクトル包絡の形状を大きく変化させていることを示唆している．中でも音色補正を用いた音声は自然性が大きく下がっており，現在の倍音周波数のピークへ極端にフィッティングするスペクトル包絡推定法が原因の一つと考えられるため，今後 DAP などの推定法を採用することで自然性の向上が期待される．また，無伴奏歌唱だけでなく混合音に対する音色・音質調査が必要であると考えている．

4.4 伴奏音中の歌声 F0 軌跡に対する歌唱表現転写

実際に，伴奏音に対し，歌唱表現の転写を行った．図 8 に転写の一例を示す．伴奏音のスペクトルに対しほとんど影響を与えずに，歌声のスペクトルのみが変形されている様子がスペクトログラムから見て取れる．再合成された音響信号を聴取したところ，少しノイズ感があるものの，混合音中の歌声の歌い回しが変わっていることを確認した．ノイズ感の原因の一つとして，高周波数において歌声とその他のドラムなどのスペクトルの分離が不完全であることが考えられる．定 Q 変換では高周波数の周波数分解能が低くなり，スペクトルが密に重なりあうためである．歌声

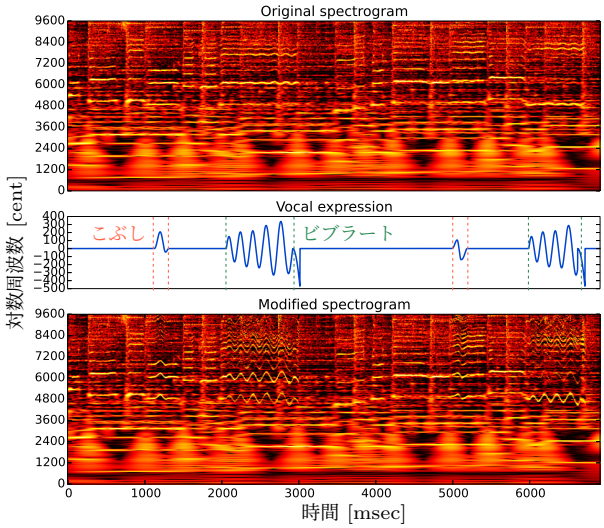


図 8 混合音に対する歌唱表現転写の例．上から，元のスペクトログラム，転写する歌唱表現，転写後のスペクトログラムを表す．

のある程度高周波の成分はパワーも小さく，聴覚上も聴き取り辛い傾向があるため，歌声スペクトルの選別において高周波数を切り捨てる処理を行うことで，聴覚的な自然さが増すのではないかと考えている．

5. おわりに

本稿では，市販楽曲などの混合音に含まれる歌声の F0 軌跡に対して，ユーザが自由に歌唱表現を転写することを可能とするシステムを提案した．ユーザは歌唱表現を事前に作成したプロ歌手のデータベースから参照することができる．混合音中の歌声のみを加工するため，時間（対数）周波数領域において F0 を用いたマスキング処理を行うことで歌声のスペクトルを選別する，選別された歌声スペクトルを対数周波数軸でシフトすることで音高を変更する．このとき，音韻性や声質を保存するため，スペクトル包絡を推定し音色の補正を行う．実験では，音色補正の有効性を確認した．また，ユーザが簡潔に歌唱表現転写箇所や F0 存在範囲を容易に提示することができる GUI を試作した．スペクトログラム上をなぞりペイントすることで F0 存在範囲を示すことができる．今後は GUI を洗練するとともに，実際に被験者実験を行い，使いやすさや手法の問題点を調べる予定である．また，歌唱表現をデータベース化する手法自体も改良が必要であると考えており，より大規模なデータベースを構築することを目標としている．

謝辞 本研究の一部は，JSPS 科研費 26700020, 24220006, 24700168 および JST CREST OngaCREST プロジェクトの支援を受けた．

参考文献

[1] Goto, M.: Active Music Listening Interfaces Based on Signal Processing, *Proc. ICASSP* (2007).
[2] Yoshii, K., Goto, M., Komatani, K., Ogata, T. and Okuno, H. G.: Drumix: An Audio Player with Real-time Drum-part Rearrangement Functions for Active Music Listening, *IPSJ Journal* (2007).
[3] Itoyama, K., Goto, M., Komatani, K., Ogata, T. and Okuno, H. G.: Instrument Equalizer for Query-by-Example Retrieval: Improving Sound Source Separation based on Integrated Harmonic and Inharmonic Models, *Proc. ISMIR* (2008).
[4] Fritsch, J. and Plumbley, M. D.: Score Informed Audio Source Separation using Constrained Nonnegative Matrix Factorization and Score Synthesis, *Proc. ICASSP* (2013).
[5] Rafii, Z., Germain, F. G., Sun, D. L. and Mysore, G. J.: Combining Modeling of Singing Voice and Background Music for Automatic Separation of Musical Mixtures, *Proc. ISMIR* (2013).
[6] Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T. and Banno, H.: Tandem-STRAIGHT: A Temporally Stable Power Spectral Representation for Periodic Signals and Applications to Interference-free Spectrum, F0, and Aperiodicity Estimation, *Proc. ICASSP* (2008).
[7] Ohishi, Y., Mochihashi, D., Kameoka, H. and Kashino, K.: Mixture of Gaussian Process Experts for Predicting Sung Melodic Contour with Expressive Dynamic Fluctuations, *Proc. ICASSP* (2014).
[8] 大石康智，持橋大地，亀岡弘和，柏野邦夫：混合ガウス過程に基づく歌声音量軌跡の生成過程モデル，情報処理学会研究報告 (2013).
[9] Fujihara, H. and Goto, M.: Concurrent Estimation of Singing Voice F0 and Phonemes by Using Spectral Envelopes Estimated from Polyphonic Music, *Proc. ICASSP*, pp. 365–368 (2011).
[10] Saito, T. and Goto, M.: Acoustic and Perceptual Effects of Vocal Training in Amateur Male Singing, *Proc. INTERSPEECH* (2009).
[11] Ikemiya, Y., Itoyama, K. and Okuno, H. G.: Transcribing Vocal Expression from Polyphonic Music, *Proc. ICASSP* (2014).
[12] Schorkhuber, C. and Klapuri, A.: Constant-Q Transform Toolbox for Music Processing, *SMC Conference* (2010).
[13] Candes, E. J., Li, X., Ma, Y. and Wright, J.: Robust Principal Component Analysis?, *J. ACM* (2011).
[14] Huang, P.-S., Chen, S. D., Smaragdis, P. and Hasegawa-Johnson, M.: Singing-Voice Separation from Monaural Recordings Using Robust Principal Component Analysis, *Proc. ICASSP* (2012).
[15] Goto, M.: PreFEst: A Predominant-F0 Estimation Method for Polyphonic Musical Audio Signals, *Proc. MIREX* (2005).
[16] Saito, S., Kameoka, H., Takahashi, K., Nishimoto, T. and Sagayama, S.: Specmurt Analysis of Polyphonic Music Signals, *IEEE Trans. on Audio, Speech, and Language Process* (2008).
[17] Hermes, D. J.: Measurement of pitch by subharmonic summation, *J. Acoust. Soc. Am.*, Vol. 83, No. 1, pp. 257–264 (online), DOI: 10.1121/1.396427 (1988).
[18] El-Jaroudi, A. and Makhoul, J.: Discrete All-Pole Modeling, *IEEE Trans. on Signal Proc.* (1991).
[19] Irino, T. and Kawahara, H.: Signal Reconstruction from Modified Auditory Wavelet Transform, *IEEE Trans. on Signal Proc.* (1993).
[20] Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R.: RWC Music Database: Popular, Classical, and Jazz Music Databases, *Proc. ISMIR*, pp. 287–288 (2002).