市販楽曲中の歌声の分離と音高推定に基づく歌唱表現編集システム

池宮 由楽[1,a)]　糸山 克寿[1,b)]　吉井 和佳[1,c)]

概要：本稿では，市販 CD のような音楽音響信号に含まれる歌声に対して，伴奏音に影響を与えることなく歌唱表現（ビブラート・グリッサンド・こぶしなど）の編集を行うためのシステムについて述べる．近年，既存楽曲をユーザが自分好みに編集・加工することを可能にする能動的音楽鑑賞システムの研究が盛んである．なかでも，混合音中の歌声の編集は最も実現が難しい課題の一つであり，既存の歌声の**声質**を他の歌唱者の声質に直接変換する技術は提案されているが，歌声がもつ特徴的な**音高**軌跡，すなわち歌唱表現を編集する技術は実現されていなかった．我々は，歌声・伴奏音の分離と歌声の音高推定の相互依存性に着目し，独立に利用されていた Robust PCA と Subharmonic Summation を相補的に組み合わせることで，両タスクの精度を改善する手法を提案する（国際的な音楽認識コンテスト MIREX2014 の歌声分離トラックで世界最高性能を達成）．この技術を応用し，既存楽曲に含まれる歌声の任意の箇所に対して，任意の歌唱表現を付与するための GUI を実現した．実験により，提案システムの優れた歌声解析精度を定量評価し，実際に GUI を用いて市販楽曲を高品質かつインタラクティブに編集可能であることを確認した．

A Vocal Expression Editing System based on Singing Voice Separation and F0 Estimation for Music Recordings

Ikemiya Yukara[1,a)]　Itoyama Katsutoshi[1,b)]　Yoshii Kazuyoshi[1,c)]

Abstract: This paper presents a novel system that enables users to edit vocal expressions of singing voices (*e.g.*, vibrato, glissando, and kobushi) included in real-world music recordings while preserving the original accompanying sounds. Active music listening has recently gained a lot of attention for providing users with a way of modifing existing music signals as they like. In particular, editing accompanied singing voices is one of the most challenging problems. Although a promising method was proposed for directly converting the timbres of existing singing voices into those of another singer's voice, it had been infeasible to edit the characteristic patterns of singing F0 contours. In this paper we propose a method that significantly improves singing voice separation based on robust PCA (RPCA) and vocal F0 estimation based on subharmonic summation (SHS) by using the mutual dependency between these tasks (the proposed method took first place in the singing voice separation track of an international music recognition contest called MIREX 2014). We developed a GUI for adding arbitrary kinds of vocal expressions to users' specified regions of existing singing voices included in commercial CD recordings.

1. はじめに

近年，一般ユーザの音楽鑑賞体験をより豊かにするため，能動的音楽鑑賞インタフェース [1] の研究が盛んである．従来の音楽鑑賞は，市販 CD や MP3 などのオーディオメディアを再生するだけの受動的な体験であり，音楽に対する能動的な振る舞いとしては，自分好みのプレイリストを作成したり，イコライザを用いて周波数特性を調整するなど簡単な処理に限定されていた．昨今の音楽音響信号処理技術の発展は著しく，音楽的な専門知識を持たないユーザであっても，楽曲の音楽内容を反映したインタラクティブな音楽鑑賞を楽しむことが可能になってきている．特に，既存音楽音響信号の編集・加工に関する研究は，単純な音楽鑑賞支援にとどまらず，一種の創作支援と見ることもできる．例えば，ドラムパートの音量や音色，パターンを MIDI ファイルを扱うかのごとく編集する [2]，楽器パートの音量バランスを個別に調整する [3,4]，あるいは歌声と伴奏を分離するといったことが可能である [5].

我々は，歌声を編集・加工することができる能動的音楽鑑賞インタフェースを開発している．ポピュラー楽曲では

[1] 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University
[a)] ikemiya@kuis.kyoto-u.ac.jp
[b)] itoyama@kuis.kyoto-u.ac.jp
[c)] yoshii@kuis.kyoto-u.ac.jp

主旋律が歌声によって奏でられる場合が多く，歌声を操作することができれば，楽曲の印象を大きく変化させながら音楽鑑賞を楽しむことができる．このとき，音の三要素である音高・音量・音色に分けて歌唱者の個性を表現することが重要である [6]．例えば，音声分析合成システムである TANDEM-STRAIGHT [7] は，単独歌唱を F0（音高）とスペクトル包絡（音色），非周期性指標の三つのパラメータへ分解し，それらを独立に操作して高品質な音声を再合成することができる．大石ら [8] は，歌声の F0 軌跡を不連続な楽譜成分と微細な変動成分の重ね合わせとして表現する確率モデルを用いて，任意の楽譜から歌声の F0 軌跡を生成する手法を提案している．同様のモデルは，歌声の音量軌跡に対しても適用することができる [9]．ただし，これらの研究は無伴奏の歌声を対象としていた．一方，藤原ら [10] は，混合音を伴奏と歌声の重ね合わせとして表現する確率モデルを用いて，歌声信号を明示的に分離することなく，混合音中の歌声の声質変換を実現している．

本稿では，市販 CD のように複雑な音楽音響信号を対象とし，**伴奏つきの歌声に含まれる歌唱表現を GUI 上で自由に編集**することができるシステムを提案する（図 1）．ここで，歌唱表現とはビブラートやグリッサンド，こぶしなどの F0 の特徴的な局所変動のことを意味する．GUI 上には自動推定された歌声の F0 軌跡が表示されており，ユーザはロングトーン部など任意の範囲を指定し，事前に用意した歌唱表現（別の歌声から抽出することも可能 [11]）を選択・付与することができる．ただし，F0 の自動推定には誤りが含まれることを前提として，ユーザの入力を援用することで高品質な歌声編集を可能にするインタフェースを実現した．ユーザはスペクトログラム上で歌声の F0 が含まれるであろう領域をラフに塗りつぶすだけの簡易な操作により，その領域中で歌声の F0 が自動的に再推定される．人間と機械の協調に基づく音源分離 [12] や音楽解析 [13] は近年非常に注目されており，本システムはインタフェースを介してユーザと計算機が密接に結びついてる．

混合音中の歌声に対する編集システムを実現するには，高精度な**歌声・伴奏音分離**と**歌声の F0 推定**が必要であり，我々は両タスクの相互依存性を考慮することで，精度を一挙に改善することができる手法を提案する．基本的には Robust Principal Component Analysis (RPCA) を用いてスペクトログラム上で歌声・伴奏音分離を行うが [14]，歌声の F0 情報を用いれば，不要な伴奏音を抑制することができる．一方，混合音に対して歌声の F0 推定を行うよりも，分離した歌声に対して F0 推定を行う方がずっと容易である．提案手法は，国際的な音楽認識コンテストである MIREX2014 [15] の歌声分離トラックにおいて，歌声と伴奏のいずれに対しても最も優れた分離精度を達成した．

本稿で提案するインタフェースや基盤技術は，混合音を対象に音の三要素である音高・音量・音色を個別に操作することができる究極の歌声編集システムを開発するうえで重要な一歩である．これが実現できれば，異なる楽曲間で歌唱者を入れ替えて歌わせるなどの新しい音楽鑑賞が可能になる．すでに故人となった歌手であっても，歌声の三要素を任意の楽曲中の歌声に一挙に転写することで，あたかもその歌手が歌っているかのような楽曲を制作できる．また，提案手法で分離された伴奏音の品質は，カラオケ音源として十分利用可能な水準に到達したことを記しておく[*1]．

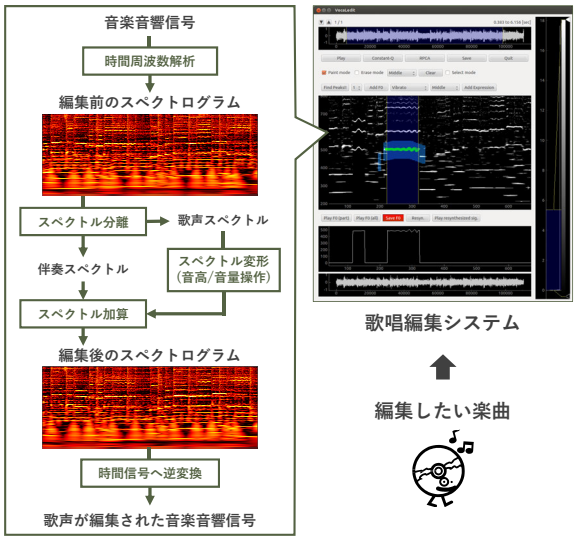

図 1　音楽音響信号中の歌声の歌唱表現編集システム

2.　歌唱表現編集インタフェース

本章では，混合音中に含まれる歌声の歌唱表現をインタラクティブに編集するための GUI について述べる．図 2 に，具体的な編集の流れを示す．

2.1　歌唱表現の選択と付与

本稿では，歌声の音高 (F0) 軌跡に現れる特徴的な局所変動を歌唱表現と定義し，特に，ポピュラー音楽や演歌などによく見られるビブラート・グリッサンド・こぶしの三種類を扱う．ここで，ビブラートはある程度の長さを持つ区間における F0 軌跡の 5Hz から 8Hz 程度の周期的な振動，グリッサンドはフレーズ開始での滑らかな音高上昇（グリスアップ）とフレーズ終端での滑らかな音高下降（グリスダウン），こぶしは比較的狭い区間における F0 の抑揚（非常に短いビブラートとみることもできる）を意味する（図 3）．実際にはそれぞれの歌唱表現について，音高のみでなく音量軌跡も変動していることに注意する．

各歌唱表現について典型的な F0 軌跡がテンプレートとして用意されており，ユーザはそれらのいずれかを選択して，歌声中の任意の箇所へ付与することができる（図 2，上から四番目）．文献 [11] を参考に，ビブラートは正弦波，グリッサンドは二次方程式（自由落下運動），こぶしは六次

[*1] デモページ：http://winnie.kuis.kyoto-u.ac.jp/members/ikemiya/demo/interaction2015/

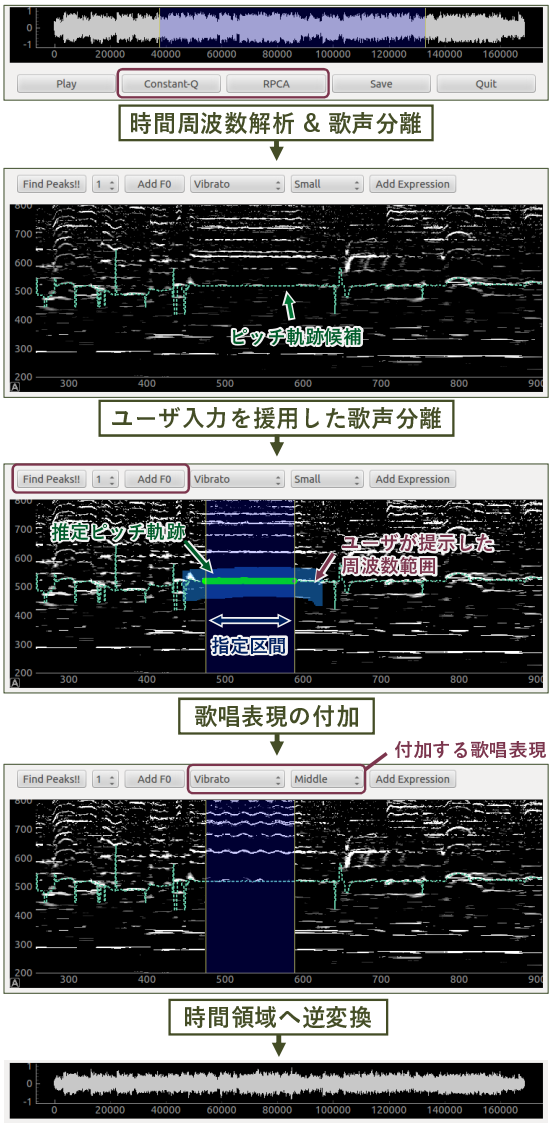

図 2　GUI 上での歌唱表現編集

方程式を用いて近似的に表現することにした．ただし，グリスダウンについては，F0 の下降に合わせて音量を減少させることで自然性を向上させている．また，各歌唱表現ついて三種の大きさ (small，middle，large) を用意することにより，ユーザの選択に柔軟性を持たせている．

ユーザは，マウスを用いて任意の時間区間を指定し，歌唱表現の種類と大きさを選択することで付与を行う．ただし，グリッサンドとこぶしについては，指定区間が選択された歌唱表現よりも短い場合には付与することはできない．ビブラートについては，指定区間前後での音高の時間的な不連続性を抑制するため，指定区間に収まり，周期の整数倍となる正弦波のうち最大長のものを生成する．

2.2　歌声の F0 再推定のための領域指定

3.2 節で述べる歌声・伴奏音分離手法では，歌声の F0 軌跡に基づいて混合音スペクトログラムから歌声の調波構造のみを抽出する．そのため，F0 推定に誤りがあると，歌声以外のスペクトルを含んだり（例：F0 推定結果が半ピッチ

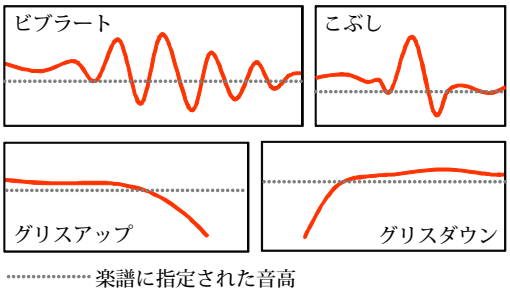

図 3　三種類の歌唱表現

誤りの場合），歌声のスペクトルを全く含まない分離結果となってしまい，音質を大きく損なう原因となる．

そこで，そもそも計算機による自動推定に誤りはつきものである（たとえ人間であっても自動採譜を完璧に行うことは非常に困難）という前提に立ち，ユーザとの協調により高精度に歌声の F0 を検出する（図 2，上から三番目）．ただし，F0 軌跡をペンツールで正確に描画することはユーザの労力の面で現実的ではないので，時間–周波数スペクトログラム上において，F0 の存在しそうな時間–周波数領域をペイントツールでラフに塗りつぶすことで範囲を指定する．指定された周波数の下限と上限が 3.2.2 項における $c_l^{(t)}$，$c_h^{(t)}$ となり，歌声の F0 推定が再度行われる．スペクトログラム上には当初推定された F0 軌跡がオーバーレイ表示されており，ユーザはスペクトルの調波構造を適宜参照しながら，F0 存在領域を塗りつぶすことができる．

2.3　各種信号処理の実行

本システムは，定 Q 変換による時間周波数解析，Robust PCA (RPCA) を用いた歌声分離 [16]，スペクトル位相復元などの複雑な信号処理は全てボタン一つで実行できる．ユーザが処理内容を意識する必要はほとんどなく，各部の計算時間も十分に高速である．また，歌声の F0 軌跡が正しく推定されているか確認するため，周波数が変調するサイン波を合成・再生する機能を実装した．入力音響信号と編集後の出力音響信号はどちらも GUI 上で再生可能である．

3.　混合音中の歌唱表現編集

本章では，音楽音響信号中の歌声に対して，音量・音高を編集する手法について説明する (図 1)．入力音響信号はまず，時間周波数領域（スペクトログラム）上において，操作対象の歌声スペクトルとそれ以外の伴奏音スペクトルへと分離される．次に，分離された歌声スペクトルの変形操作によって歌声の音量・音高操作を行う．

3.1　定 Q 変換を用いた時間周波数解析

まず，定 Q 変換 [17] を用いて入力音楽音響信号の時間周波数解析を行う．定 Q 変換はガボールマザーウェーブレットを用いたウェーブレット変換の変種とみなすことができ，時間領域信号 $x(n)$ に対して以下の式で定義される．

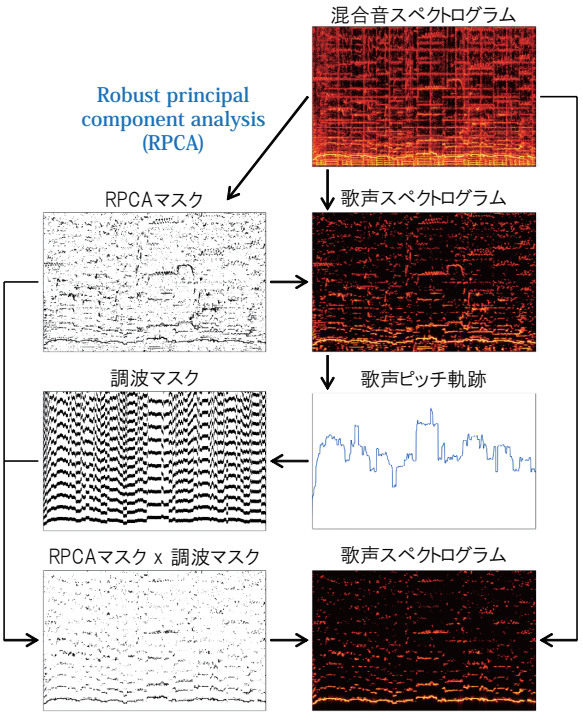

図 4　歌声・伴奏音分離と歌声の F0 推定

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

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

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 は対数周波数軸を離散化するための 1 オクターブの分割数を表し，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，全時間サンプル 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] などの時間幅で切り出す．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を $X(t,f)$ と記述する．

3.2　歌声・伴奏音分離と歌声 F0 推定の相補的実行

歌声・伴奏音分離と歌声の F0 推定は相互依存性をもっている．つまり，歌声の F0 軌跡が与えられていれば，歌声分離に利用することができる一方，歌声が分離されていれば，その F0 軌跡を推定することは比較的容易である．我々は，この相補的な関係を利用した歌声分離のための枠組みを提案する（図 4）．まず入力音響信号に対して， Robust PCA に基づくブラインド歌声分離を適用する．次に分離された歌声から F0 軌跡を推定し，それを用いてさらに精密な歌声分離を行う．

3.2.1　Robust PCA を用いた歌声分離

Robust PCA (RPCA) [14] は，与えられた行列（2 次元配列）を低ランク行列とスパース行列とに分解する手法であり，次式で定式化される．

$$\text{minimize}\ \ \|L\|_* + \lambda\|S\|_1\ \ (\text{subject to}\ \ L+S=X) \quad (2)$$

ここで，X，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，$\|\cdot\|_*$ と $\|\cdot\|_1$ 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．

混合音のスペクトログラムを入力行列 X と見なして RPCA を適用すると．繰り返し演奏されるため何度も出現する伴奏音（ドラムやギター）のスペクトルは低ランク行列 L へ，それ以外の歌声などの時間的な変動が大きいスペクトルはスパース行列 S へ分解される [16]．次に，分解結果からバイナリマスクを作成する．

$$M_r(t,f) = \begin{cases} 1 & |S(t,f)| > |L(t,f)| \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

バイナリマスクを混合音スペクトログラム $X(t,f)$ へ適用することで歌声スペクトログラムが分離できる．

3.2.2　Subharmonic Summation を用いた F0 推定

RPCA により分離された歌声スペクトル $X_s^{\mathrm{rpca}}(t,f)$ から，Subharmonic Summation (SHS) [18] を用いて歌声の F0 軌跡を推定する．SHS は計算コストの低さとノイズへの頑健性を兼ね備えた音高推定法であり，スペクトルの各周波数ビンについて，そのビンを F0 であると仮定したときの倍音に対応する周波数ビンのパワーを重みつきで足し合わせることで，当該ビンに F0 が存在する尤度を計算する．この音高尤度関数の計算は，対数周波数スケールでは以下で定式化される．

$$H(t,s) = \sum_{n=1}^{N} h_n P(t, s+1200\log_2 n), \quad (4)$$

ここで，t, s はそれぞれ時間インデクスと対数周波数 [cents] を表し，$P(t,s)$ は時間フレーム t，周波数 s [cents] における入力スペクトログラムの振幅である．N は足し合わせる倍音数，h_n は各倍音の重み関数であり，本稿ではそれぞれ 15 および 0.86^{n-1} とする．人間の聴覚特性の非線形性を考慮するため，SHS を適用する前に，入力スペクトルに対して A 特性補正 [*2] をかけるものとする．

SHS による音高尤度関数 $H(t,s)$ から歌声音高 $F(t)$ は以下の式で計算される．

$$F(t) = \underset{c_l^{(t)} \le s \le c_h^{(t)}}{\arg\max}\ H(t,s) \quad (5)$$

ここで，$c_l^{(t)}$，$c_h^{(t)}$ はそれぞれ，時間フレーム t における音高探索周波数範囲の下限と上限 ([cents]) である．

[*2] replaygain.hydrogenaud.io/proposal/equal_loudness.html

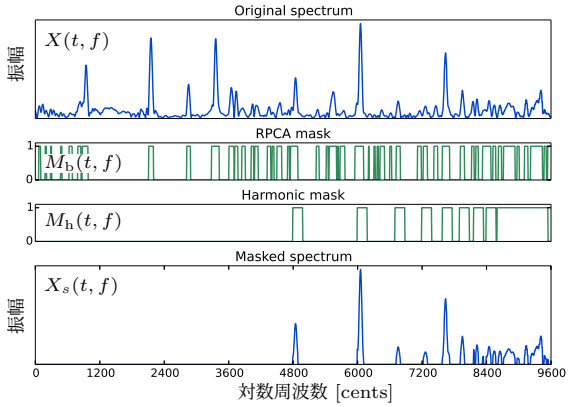

図 5 マスキングによる歌声スペクトルの選別: 上から，元の混合音スペクトル，RPCA によるバイナリマスク，歌声 F0 軌跡による調波マスク，選別された歌声スペクトルを示す．

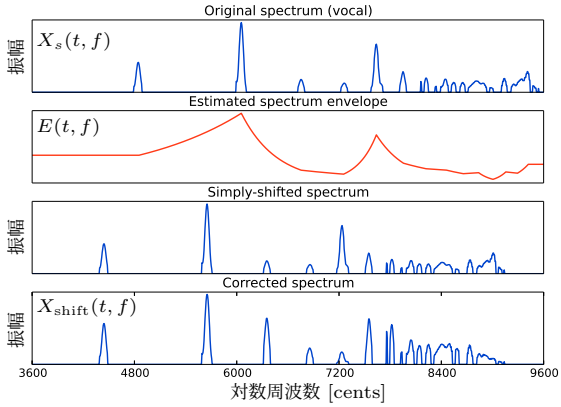

図 6 スペクトル包絡を用いた音色補正: 上から，元の歌声スペクトル，推定されたスペクトル包絡，単純にシフトしたスペクトル，スペクトル包絡を用いて音色を補正したスペクトルを示す．

3.2.3 F0 軌跡を用いた歌声分離

RPCA を用いた従来の歌声分離では，曲の一部しか現れないベースやドラム，メインボーカルと音高をずらして唱和するバックコーラスなども，歌声として分離されてしまう．そこで，F0 軌跡を利用して，さらに精度の高い歌声分離を行う（図 5）．具体的には，前項で得られた F0 軌跡から，基本周波数 (F0) と倍音周辺以外のパワーをマスキングする調波マスクを生成する．

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

ここで，F_t は時間フレーム t における F0 [cents]，$C(f)$ は周波数ビン f に対応する対数周波数 [cents]，H は倍音数，w は各倍音でマスクを取る幅 [cents] を示す．RPCA によるバイナリマスクと調波マスクを用いて，最終的な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．

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

3.3 音色補正を用いた音高・音量シフト

最後に，得られた歌声スペクトル $X_s(t,f)$ を変形し，伴奏スペクトル $X_m(t,f)$ と混合する．線形周波数軸 [Hz] において音高を a 倍する処理は，対数周波数軸 [cents] では $1200\log_2 a$ [cents] シフトする処理に対応する．つまり，対数周波数軸でスペクトル全体をシフトすれば，音高を変えることができる．また，音量を b 倍するには，スペクトル全体の大きさを b 倍する．しかし，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るため [7]，歌声のスペクトルを単純にシフトさせただけでは，スペクトル包絡も合わせてシフトされ，不自然な音となってしまう問題がある．

本研究では，音色の補正をするため，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，音高シフト後に各倍音の強度を修正する方法を提案する (図 6)．倍音のパワーからスペクトル包絡を推定するには，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）へ戻せばよい．ただし，F0 以下の周波数の包絡は一定にしておく．離散的な周波数におけるパワースペクトルからスペクトル包絡を精度よく求めるには離散全極モデル (DAP) [19] を利用する方法も考えられるが今後の課題とする．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次式で与えられる．

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

3.4 位相復元による時間領域信号の再合成

定 Q 変換スペクトログラムに対して定 Q 逆変換 [17] を適用することで，時間領域の信号を再合成することができる．しかし，音高シフトにより振幅スペクトルが変形されているため，元の位相をそのまま用いて逆変換を行うと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 [20] 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．

4. 評価実験

本章では，提案した歌声分離手法の定量的な評価を行う．この種の音楽信号編集システムを実用に耐えるものにするには，基礎技術の精度が何より重要である．また，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する歌唱表現の転写例を示す．

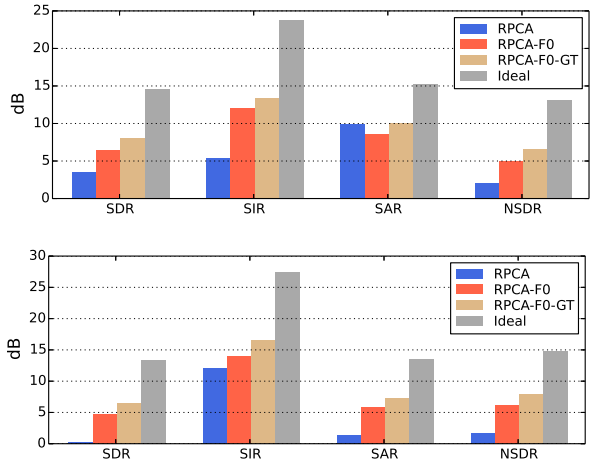

図 7 MIR-1K データセットを用いた歌声分離結果（全曲の平均値 11）：上図が歌声の分離精度，下図が伴奏の分離精度を示す．

4.1 実験条件

実験には 16kHz・16bits・モノラルの音響信号を用いた．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] とした．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 [16]，本稿では実験的に 0.1 と定めた．また，3.2.3 項での各倍音のマスク幅 w は 120 [cents] とした．

4.2 歌声分離枠組みの評価

3.2 節で述べた F0 推定に基づく歌声分離の有効性を評価するため，バイナリマスクの作成方法の違いにより，以下の四手法を比較した．

RPCA: RPCA マスク [16]
RPCA-F0: RPCA + 調波マスク（提案手法）
RPCA-F0-GT: RPCA + 調波マスク（正解 F0 を使用）
Ideal: 理想のバイナリマスク（精度の上限）

なお，歌唱表現編集システムでは音高操作の容易さから，時間周波数解析として定 Q 変換を用いているが，ここでは純粋な歌声分離精度を評価するため短時間フーリエ変換 (STFT) を用いた結果を示している．

実験データには MIR-1K データセット[*3]を使用した．データセットは 110 曲 (20-110 秒，サンプリング周波数 16 kHz）の歌唱入り楽曲からなる．各分離結果は BSS Eval Toolbox [21] を用いて，source-to-interference ratio (SIR)，sources-to-artifacts ratio (SAR)，source-to-distortion ratio (SDR)，及び Normalized SDR (NSDR) によって評価した．SIR，SAR，SDR と NSDR はそれぞれ，目的音源（歌声）以外の音源からの雑音，ミュージカルノイズなどの雑音，目的音源の歪み，分離前音源からの SDR の向上を表しており，単位はデシベル (dB) である．値が大きいほど分離精度が優れていることを示す．

[*3] sites.google.com/site/unvoicedsoundseparation/mir-1k

表 1 MIREX2014 の歌声分離タスクの結果．評価値は NSDR [dB].

手法	[22]	[23]	[24]	[25]
歌声	−1.40	−0.82	0.65	2.86
伴奏	0.35	−3.12	3.09	5.03

手法	GW1	RNA1	[26]	**提案手法**
歌声	2.89	3.69	4.17	**4.48**
伴奏	5.25	7.32	5.63	**7.87**

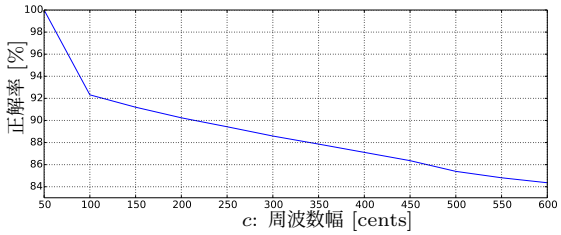

図 8 ユーザの提示する周波数幅と歌声音高推定精度の関係

図 7 に結果を示す．RPCA のみを用いた手法 (**RPCA**) と比較し，提案手法 (**RPCA-F0**) により歌声・伴奏双方について分離精度が有意に向上していることが分かる．また，表 1 は音楽認識の国際的なコンテストである MIREX 2014 [15] における歌声分離タスク（テストデータは非公開なのでフェアな評価が可能）の結果[*4]である．多くの最新手法 [22–26] が参加した中で，提案手法は歌声・伴奏音の双方について，最も高い分離精度を達成した．

4.3 指定する周波数幅の歌声 F0 推定への影響

2.2 節で述べた通り，スペクトログラム上でユーザが指定した周波数領域に基づいて歌声の F0 を推定する際に，指定された領域の大きさがどの程度推定精度へ影響するかを調べた．実験には，“RWC Music Database: Popular Music” (RWC-MDB-P-2001) [27] 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．正解音高から上下に $\pm c$ [cents] の幅を探索周波数範囲として F0 推定を行った．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た．

図 8 (a) に結果を示す．50 [cents] 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．$c = 100$ [cents] で既に 10 % 弱の誤りが存在しているが，実際にはこれらのほとんどが音符の遷移部など，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，これらの推定誤りは実用上それほど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，$c = 400$ [cents] においても 90%弱の F0 推定精度を保つことができた．これは，ユーザが正解 F0 から上下 4 半音の範囲を F0 存在領域として指定（ス

[*4] music-ir.org/mirex/wiki/2014:Singing_Voice_Separation_Results

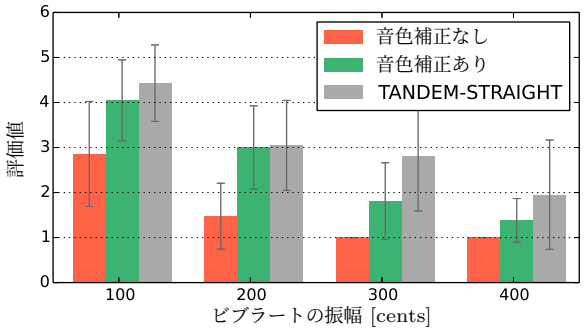

図 9 音色の自然性の評価: 赤は音色補正なし，緑は音色補正あり，灰は TANDEM-STRAIGHT の結果を示す．評価値は全被験者・音韻の平均であり，エラーバーは標準偏差を表す．

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，十分に実用的な許容誤差である．

4.4 音色補正の効果

3.3 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た．音色補正の効果を純粋に調べるため，無伴奏歌唱を実験データとして用いた．女性が「い」「う」「お」の音韻を平坦に伸ばして発音している音声に対し，周期 6 [Hz]，振幅 100・200・300・400 [cents] のビブラートを付与し，7 名の被験者数による聴取実験によって音色の自然性を調べた．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 [7] による合成音声を用いた．TANDEM-STRAIGHT は，単独音声の操作に関して非常に品質の高い合成音声を得ることができるため（混合音から分離された歪みのある歌声に対しては実用的ではない），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三手法・三種類の振幅の音高シフト音声を聴かせ，「音色の自然性」について五段階で評価させた．

図 9 に振幅毎の結果を示す．音色補正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 [cents] 以下では，音色補正を用いた音声は TANDEM-STRAIGHT と比較的近い評価を得た．多くのポピュラー歌手のビブラートが 200 [cents] 以下であることからも，音色補正は非常に有効であるといえる．振幅が大きくなるに従い，TANDEM-STRAIGHT を含めて全体的に音色自然性は低下する傾向にあった．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歌唱表現をする場合，音色（スペクトル包絡の形状）も合わせて大きく変化し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，倍音周波数のピークへ極端にフィッティングする現状の簡易なスペクトル包絡推定法が原因の一つと考えられる．今後は，離散全極モデル (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中の歌唱に対する音色・音質調査が必要であると考えている．

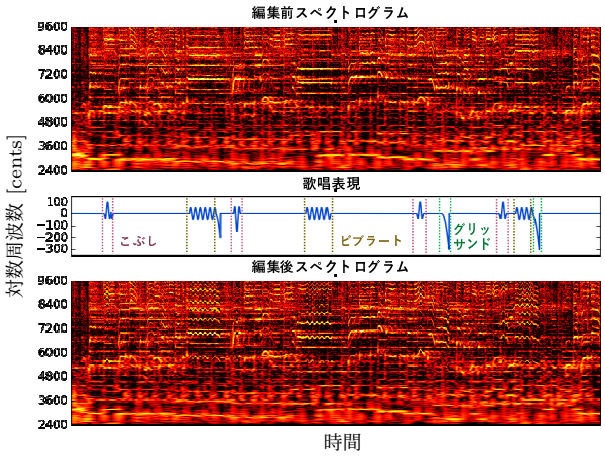

図 10 市販楽曲に対する歌唱編集の例（ラッキープール / JUDY AND MARY）：上から，入力音響信号のスペクトログラム，付与する歌唱表現，編集後のスペクトログラムを表す．

4.5 市販楽曲に対する歌唱編集

実装したシステムを用いて，市販楽曲中のプロ歌手の歌声に対して歌唱表現の編集を行った．図 10 はポピュラー楽曲（ラッキープール / JUDY AND MARY）に対する編集例である．伴奏音のスペクトルに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見て取れる．実際に編集後の音響信号を聴取したところ，歌声の歌い回しが違和感なく編集されていることを確認した．一部の楽曲では，編集後の音響信号に微小なノイズが感じられた．こ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．歌声のスペクトルにおいて，ある周波数以上の高周波成分はパワーも小さく，聴覚上も聴き取りづらい傾向があるため，歌声スペクトルの選別において高周波成分をある程度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．編集結果のサンプル音源を Web サイトで聴取することができる [28]．

5. インタフェースの課題

本稿で提案した歌唱表現編集インタフェースは，ユーザが調波構造や F0 といった音声の情報に対してある程度知識があることを前提としている．例えば，GUI 上で歌声 F0 の存在箇所を提示するには，表示された時間周波数スペクトログラムから，いずれが歌声のスペクトルであるかを判別する必要がある．実際，4.5 節で市販楽曲に対し編集処理を行ったのは，音声信号処理に精通した学生である．

そういった知識を持たない一般のユーザが直感的に操作を行えるインタフェースを実現するため，判別を助ける情報をシステムからフィードバックする機能を考える．具体的には，推定された F0 軌跡を音信号として再生することにより，ユーザはその F0 が正しいかを判断し，誤り箇所

を修正することが可能となる．この処理は，音楽解析サービス Songle [13] においても採用されている．また，ユーザに編集処理の流れを明示的に指示することで，円滑な操作が行えると考えている．今後はこれらの実装を行い，音声信号処理に精通していない被験者による主観評価を行う予定である．

6. おわりに

本稿では，伴奏を含む市販楽曲中の歌声の歌唱表現を編集するシステムを提案した．入力音響信号は，歌声の F0 推定を相補的に用いる歌声・伴奏音分離手法により，歌声スペクトログラムと伴奏音スペクトログラムとに分解される．ユーザは GUI 上で，歌声の F0 軌跡の任意の箇所に，ビブラート，グリッサンド，こぶしなど任意の歌唱表現を選択的に付与することができる．

本研究で開発した高精度な歌声・伴奏音分離技術を生かして，混合音中の歌声の F0 以外にも，音色・音量を合わせて編集するような拡張を行う予定である．将来的に，我々が以前提案した歌唱表現ライブラリ [11] との統合を行い，ユーザが好みの歌手を指定しただけで，その歌手の特徴的な歌唱表現が，伴奏つき歌声の適切な位置に自動的に付与されるようシステムを洗練化していきたい．

謝辞 本研究の一部は，JSPS 科研費 26700020, 24220006，24700168, 26280089 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, *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, *ISMIR* (2008).
[4] Fritsch, J. and Plumbley, M. D.: Score Informed Audio Source Separation using Constrained Nonnegative Matrix Factorization and Score Synthesis, *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, *ISMIR* (2013).
[6] Saito, T. and Goto, M.: Acoustic and Perceptual Effects of Vocal Training in Amateur Male Singing, *INTERSPEECH* (2009).
[7] 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, *ICASSP* (2008).
[8] Ohishi, Y., Mochihashi, D., Kameoka, H. and Kashino, K.: Mixture of Gaussian Process Experts for Predicting Sung Melodic Contour with Expressive Dynamic Fluctuations, *ICASSP* (2014).
[9] 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，情処研報 (2013).
[10] Fujihara, H. and Goto, M.: Concurrent Estimation of Singing Voice F0 and Phonemes by Using Spectral Envelopes Estimated from Polyphonic Music, *ICASSP*, pp. 365–368 (2011).
[11] Ikemiya, Y., Itoyama, K. and Okuno, H. G.: Transcribing Vocal Expression from Polyphonic Music, *ICASSP* (2014).
[12] Bryan, N. J. and Mysore, G. J.: An Efficient Posterior Regularized Latent Variable Model for Interactive Source Separation, *ICML* (2013).
[13] 後藤真孝，吉井和佳，藤原弘将，Mauch, M.，中野倫靖：Songle: 音楽音響信号理解技術とユーザによる誤り訂正に基づく能動的音楽鑑賞サービス，情処論 (2013).
[14] Candes, E. J., Li, X., Ma, Y. and Wright, J.: Robust Principal Component Analysis?, *J. ACM* (2011).
[15] Downie, J. S.: The Music Information Retrieval Evaluation Exchange (2005—2007): A Window into Music Information Retrieval Research, *Acoustical Science and Technology*, Vol. 29, pp. 247–255 (2008).
[16] Huang, P.-S., Chen, S. D., Smaragdis, P. and Hasegawa-Johnson, M.: Singing-Voice Separation from Monaural Recordings Using Robust Principal Component Analysis, *ICASSP* (2012).
[17] Schorkhuber, C. and Klapuri, A.: Constant-Q Transform Toolbox for Music Processing, *SMC Conference* (2010).
[18] 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).
[19] El-Jaroudi, A. and Makhoul, J.: Discrete All-Pole Modeling, *IEEE Trans. on Signal Proc.* (1991).
[20] Irino, T. and Kawahara, H.: Signal Reconstruction from Modified Auditory Wavelet Transform, *IEEE Trans. on Signal Proc.* (1993).
[21] Vincent, E., Gribonval, R. and Févotte, C.: Performance Measurement in Blind Audio Source Separation, *IEEE Trans. on Audio, Speech and Language Processing*, Vol. 14, pp. 1462–1469 (2006).
[22] Huang, P.-S., Kim, M., Hasegawa-Johnson, M. and Smaragdis, P.: Singing-Voice Separation from Monaural Recordings using Deep Recurrent Neural Networks, *ISMIR* (2014).
[23] Yen, F., Luo, Y.-J. and Chi, T.-S.: Singing Voice Separation using Spectro-Temporal Modulation Features, *ISMIR* (2014).
[24] Liutkus, A., Fitzgerald, D., Rafii, Z., Pardo, B. and Daudet, L.: Kernel Additive Models for Source Separation, *IEEE TSP* (2014).
[25] Rafii, Z. and Pardo, B.: Music/Voice Separation using the Similarity Matrix, *ISMIR*, pp. 583–588 (2012).
[26] Jeong, I.-Y. and Lee, K.: Vocal Separation from Monaural Music Using Temporal/Spectral Continuity and Sparsity Constraints, *Signal Processing Letters*, Vol. 21, pp. 1197–1200 (2014).
[27] Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R.: RWC Music Database: Popular, Classical, and Jazz Music Databases, *ISMIR*, pp. 287–288 (2002).
[28] 池宮由楽：デモページ，`http://winnie.kuis.kyoto-u.ac.jp/members/ikemiya/demo/interaction2015/`.