音楽音響信号の音源分離と能動的音楽鑑賞への応用

Sound source separation for music audio signals and its application to active music listening

糸山 克寿（Katsutoshi ITOYAMA, Ph. D.）
京都大学大学院 情報学研究科 助教
(Assistant Professor, Graduate School of Informatics, Kyoto University)
情報処理学会　日本音響学会　IEEE
受賞：平成 21 年度 IPSJ 論文船井若手奨励賞(2009)　第 24 回テレコムシステム技術学生賞 (2008)
研究専門分野：音楽情報処理　統計的信号処理・機械学習

あらまし　本研究では、音楽 CD などの市販楽曲を歌声・調波楽器音・打楽器音の 3 つに音源分離する手法の開発、および音源分離技術を応用した歌声編集システム、音楽演奏練習支援システムの開発に取り組んだ。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ロバスト主成分分析（Robust Principal Component Analysis : RPCA）と Subharmonic Summation を相補的に組み合わせることで、音源分離精度を改善した（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システムを実現した。さらに、伴奏音を調波楽器音（ギターやキーボードなど、メロディやコードを演奏できる楽器音）と打楽器音（ドラムなど、音の高さをもたない楽器音）へ分離する手法と組み合わせることで、歌唱や楽器演奏の練習を支援するシステムを実現した。

1．研究の目的・狙い

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[1]、楽器パートの音量バランスを個別に調整する[2]、あるいは歌声と伴奏を分離する[3]といったことが可能である。音楽 CD や MP3 を再生するだけの受動的な音楽鑑賞体験を超えた、このような豊かな音楽鑑賞体験は能動的音楽鑑賞[4]と呼ばれ、盛んに研究がなされている。本研究では、能動的な音楽鑑賞の実現に向けて、歌声・調波楽器音・打楽器音の音源分離技術の開発、および分離技術を応用した歌唱表現編集システム・音楽演奏練習システムの開発に取り組んだ。

ポピュラー楽曲では、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[5]。

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グラフィカルユーザーインタフェース（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）。

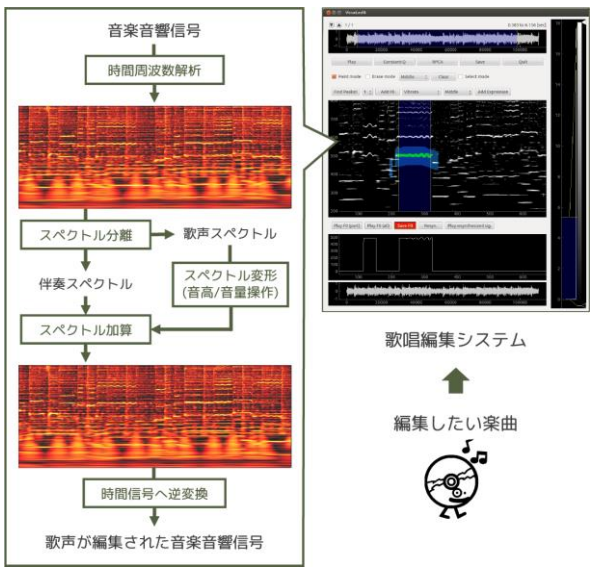

図 1　伴奏つきの歌声に含まれる歌唱表現を GUI 上で自由に編集できるシステム

音楽音響信号の音源分離と能動的音楽鑑賞への応用

Sound source separation for music audio signals and its application to active music listening

ここで、歌唱表現とは、ビブラートやグリッサンド、こぶしなどの音高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音高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を選択・付与することができる。

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音高推定が必要であり、筆者ら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、ロバスト主成分分析（Robust Principal Component Analysis; RPCA）[6]を用いてスペクトログラム*1 上で歌声・伴奏音分離を行うが、歌声の音高が分かっていれば、不要な伴奏音を抑制して分離性能を向上できる。一方、混合音に対する歌声の音高推定よりも、分離した歌声に対する推定の方がずっと容易である。

さらに、音源分離技術を応用し、任意の音楽音響信号（市販 CD に収録されているような歌唱を含む完全な楽曲）に対して、カラオケ音源やマイナスワン音源*2 を自動生成し、ユーザの歌唱や楽器の練習を支援するシステムを開発した。本システムは Android タブレット上に実装されており、どこでも手軽に音楽を楽しむことができる。ユーザが歌う際には、原曲中に含まれるコーラス歌唱を残しつつ、メインボーカルだけを抑制したカラオケ音源を再生することができる。このとき、原曲歌手の歌声の音高が画面上にガイドとして表示されており、ユーザはそれを参考にしながら歌うことができる。一方、ユーザの歌声の音高はリアルタイムで解析・可視化され、自分の歌声とプロ歌唱の音高を比較しながら練習できる。また、ユーザがギターやキーボードなどの調波楽器、あるいはドラムなどの打楽器を弾く際には、カラオケモードと同様に、調波楽器音あるいは打楽器音の音量を個別に抑制することができる。このとき、原曲のビート時刻やコード進行は音楽再生と同期してスクロール表示され、調波楽器を演奏する際には、そのコードがリアルタイムで解析・可視化される。

２．研究の背景

２．１ 歌声の分析・操作・合成

音声分析合成システムである TANDEM-STRAIGHT [7]は、単独歌唱を基本周波数（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、歌声の音高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音高軌跡を生成する手法を提案している。一方、藤原ら[9]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。

２．２ インタラクティブな音楽鑑賞

後藤ら[10]は、音楽内容を自動推定して可視化することでユーザが楽曲の要素をより深く理解できるようになる Web サービス Songle を開発している。安良岡ら[11]は、音響信号中の特定の楽器パートのフレーズを原曲の音色を保持しながら自由に編集できるシステムを提案している。深山ら[12]は、異なる音響信号中のコード進行の特徴を組み合わせて、新たなコード進行を生成することができるシステムを提案している。

３．研究の方法・研究の結果

３．１ 音源分離

音楽音響信号の歌声・調波楽器音・打楽器音への分離は、(1) 歌声・伴奏音分離、(2) 調波楽器音・打楽器音分離の 2 ステップからなる。

(1) 歌声・伴奏音分離

歌声・伴奏音分離には、ロバスト主成分分析（RPCA）と歌声音高推定の相補的実行に基づく手法[13]を用いる（図 2）。RPCA は、入力された行列を低ランク行列とスパース行列の 2 つに分解する教師なし機械学習法

音楽音響信号の音源分離と能動的音楽鑑賞への応用

Sound source separation for music audio signals and its application to active music listening

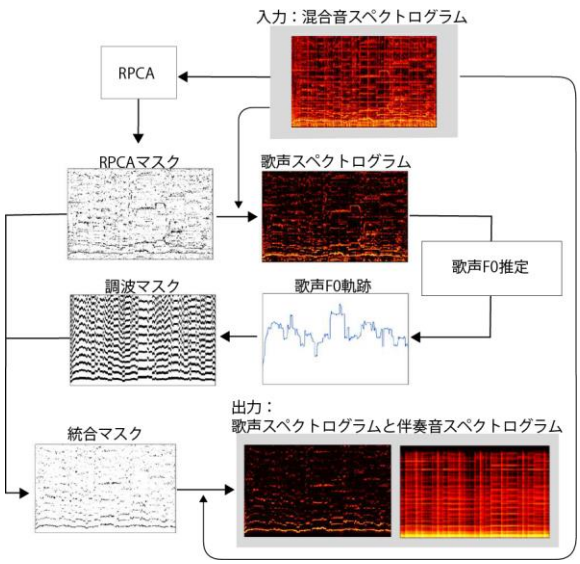

図２　ロバスト主成分分析と歌声音高推定の相補的実行に基づく歌声・伴奏音分離

の１つである。低ランク行列とは、少数のパターンが繰り返し出現する行列である。楽曲中で用いられる伴奏音のパターン（コードやリズムなど）は、音楽理論・用いる楽器・演奏者の個性などによって一定の範囲に制限されるため、伴奏音のスペクトログラムは低ランク性をもつ。一方でスパース行列は、同一パターンがめったに現れないような行列である。ビブラートなどの歌唱表現や不随意的な微少変動に起因して、歌声のスペクトログラムはスパース性をもつ。したがって、RPCA は楽曲スペクトログラムを歌声と伴奏音に教師なしで分解することができる。

まず、RPCA により、音楽音響信号のスペクトログラムを歌声にあたるスパース行列と伴奏にあたる低ランク行列の和に粗く分離する。次に、分離された歌声スペクトログラムから歌声の音高を推定する。最後に、推定された音高を用いて精密な歌声分離を行う。

(2) 調波楽器音・打楽器音分離

調波楽器音・打楽器音分離には、スペクトログラムの異方性に基づく手法[14]を用いる。この手法は、調波楽器音はスペクトログラム中で時間軸方向に滑らかであり、打楽器音はスペクトログラム中で周波数軸方向に滑らかであることに着目したものである。この異方性に基づき、調波楽器音抽出フィルタと打楽器音抽出フィルタを構築し、フィルタにより分離を行う。

３．２ 歌唱表現編集

分離された歌声を変形させることで、歌唱表現（ビブラート・グリッサンド・こぶし）を付与する（図 3）。

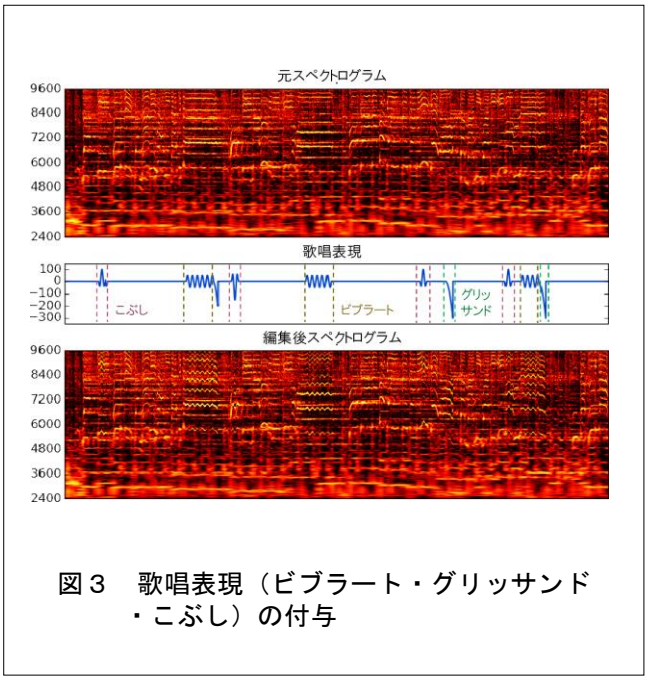

図３　歌唱表現（ビブラート・グリッサンド・こぶし）の付与

３．３ 歌唱表現編集インタフェース

各歌唱表現についての典型的な音高軌跡をテンプレートとして、それらの種類と大きさを選択し、歌声中の任意の箇所へ付与する。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）、こぶしは六次方程式で表現する。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。

３．４ 音色補正を用いた音高シフト

付与する歌唱表現に対応するように、分離された歌声の音高を変動させる。音高を変動させるためにはスペクトルを周波数方向へシフトすればよいが、単純なシフトではスペクトル包絡も合わせてシフトされるため、歌声の音色が不自然になってしまう。本研究では、

歌声スペクトルからスペクトル包絡を擬似的に推定し、音高シフト後に各倍音の強度を修正することで音色を補正し、音色の不自然さを軽減する。

３．５ 音楽演奏練習

音源分離技術を応用し、任意の音楽音響信号（市販 CD に収録されているような歌唱を含む完全な楽曲）に対して、カラオケ音源やマイナスワン音源を自動生成し、ユーザの歌唱や楽器の練習を支援するシステム（図 4）を制作した。本システムは、Android タブレット（HTC Nexus 9）上に実装されており、どこでも手軽に音楽を楽しむことができる。

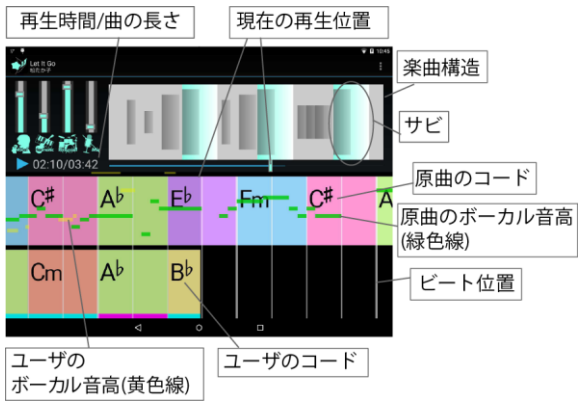

図４　カラオケ音源やマイナスワン音源を自動生成してユーザの歌唱や楽器の練習を支援するシステム

ユーザが歌う際には、原曲中に含まれるコーラス歌唱を残しつつ、メインボーカルだけを抑制したカラオケ音源を再生する。このとき、原曲歌手の歌声音高が画面上にガイドとして表示されており、ユーザはそれを参考にしながら歌うことができる。一方、ユーザの歌声音高はリアルタイムで解析・可視化され、自分の歌声とプロ歌唱の音高を比較しながら練習できる。また、ユーザがギターやキーボードなどの調波楽器、あるいはドラムなどの打楽器を弾く際には、カラオケモードと同様に、調波楽器音あるいは打楽器音の音量を個別に抑制することができる。このとき、原曲のビート時刻やコード進行は音楽再生と同期してスクロール表示され、調波楽器を演奏する際には、そのコードがリアルタイムで解析・可視化される。

本システムを実現するには、音源分離に加えて、音楽内容の可視化とリアルタイム演奏認識が必要となる。可視化すべき音楽内容は、クラウドソーシング型音楽鑑賞 Web サービス Songle [10]から取得する。Songle は、Web 上にある任意の音楽音響信号のビート時刻・コード進行・歌声の音高・楽曲構造などの要素を自動解析しブラウザ上に表示するサービスであり、Songle Widget と呼ばれる API を通じて解析結果を取得できる。また、ユーザ演奏のリアルタイム解析のため、テンプレートに基づくコード認識と、Subharmonic Summation [15]に基づく歌声の音高推定を用いる。これらの手法は簡便ではあるが、計算資源の限られた携帯端末上でも動作可能である。

４．将来展望

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。

歌声・伴奏音分離に用いた RPCA や、その基礎となる低ランク・スパース行列分解は、音楽音響信号に限らず様々な分野への応用が期待される。これらの行列演算の多くは GPGPU を用いた並列化が可能であるため、リアルタイムでの分離が行えるようになれば、さらに活用の範囲は広まる。また、コード進行やリズムパターンなどの音楽的知識を大量の音楽データから教師なしで獲得する試みが進められており、これらとの組み合わせで、より自然でインタラクティブな音楽鑑賞体験の実現が期待される。

音楽音響信号の音源分離と能動的音楽鑑賞への応用

Sound source separation for music audio signals and its application to active music listening

用語解説

*1 スペクトログラム

音響信号を窓関数に通して周波数スペクトルを計算した、時間・周波数・信号の強度の 3 次元グラフ。スペクトログラムの計算には短時間フーリエ変換（Short-time Fourier Transform : STFT）が一般的に用いられる。

*2 マイナスワン音源

様々な楽器パートの混合である楽曲から、特定の楽器パートだけを消し去ったもの。カラオケ音源は、ボーカルパートだけを消し去ったマイナスワン音源でもある。

参考文献

[1] K. Yoshii, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, Drumix: An Audio Player with Real-time Drum-part Rearrangement Functions for Active Music Listening, IPSJ Journal, Vol.48, No.3, pp.1229-1239, 2007.

[2] K. Itoyama, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, Parameter Estimation for Harmonic and Inharmonic Models by Using Timbre Feature Distributions, Journal of Information Processing, Vol.17, pp.191-201, 2009.

[3] Z. Rafii, F. G. Germain, D. L. Sun, and G. J. Mysore, Combining Modeling of Singing Voice and Background Music for Automatic Separation of Musical Mixtures, ISMIR 2013, pp.41-46, 2013.

[4] M. Goto, Active Music Listening Interfaces Based on Signal Processing, ICASSP 2007, Vol.IV, pp.1441-1444, 2007.

[5] T. Saito and M. Goto, Acoustic and Perceptual Effects of Vocal Training in Amateur Male Singing, INTERSPEECH 2009, pp.832-835, 2009.

[6] E. J. Candès, X. Li, Y. Ma, and J. Wright, Robust Principal Component Analysis?, Journal of the ACM, Vol.58, No.3, Article 11, 2011.

[7] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, Tandem-STRAIGHT: A Temporally Stable Power Spectral Representation for Periodic Signals and Applications to Interference-free Spectrum, F0, and Aperiodicity Estimation, ICASSP 2008, pp.3933, 2008.

[8] Y. Ohishi, D. Mochihashi, H. Kameoka, and K. Kashino, Mixture of Gaussian Process Experts for Predicting Sung Melodic Contour with Expressive Dynamic Fluctuations, ICASSP 2014, p.3714-3718, 2014.

[9] H. Fujihara and M. Goto, Concurrent Estimation of Singing Voice F0 and Phonemes by Using Spectral Envelopes Estimated from Polyphonic Music, ICASSP 2011, pp.365-368, 2011.

[10] 後藤真孝, 吉井和佳, 藤原弘将, M. Mauch, 中野倫靖, Songle: ユーザが誤り訂正により貢献可能な能動的音楽鑑賞サービス, インタラクション 2012, pp.1363–1372, 2012.

[11] N. Yasuraoka, T. Abe, K. Itoyama, T. Takahashi, T. Ogata, and H. G. Okuno, Changing Timbre and Phrase in Existing Musical Performances as You Like: Manipulations of Single Part Using Harmonic and Inharmonic Models, ACM Multimedia 2009, pp.203-212, 2009.

[12] S. Fukayama and M. Goto, HarmonyMixer: Mixing the Character of Chords among Polyphonic Audio, ICMC-SMC 2014, pp.1503-1510, 2014.

[13] Y. Ikemiya, K. Itoyama, K. Yoshii, Singing Voice Separation and Vocal F0 Estimation based on Mutual Combination of Robust Principal Component Analysis and Subharmonic Summation, IEEE/ACM Trans. ASLP, Vol.24, No.11, pp.2084-2095, 2016.

[14] D. Fitzgerald, Harmonic/percussive Separation Using Median Filtering, DAFx'10, pp. 246-253, 2010.

[15] D. J. Hermes, Measurement of Pitch by Subharmonic Summation, J. Acoust. Soc. Am, Vol.83, No.1, pp.257-264, 1988.

この研究は、平成２４年度ＳＣＡＴ研究助成の対象として採用され、平成２５～２７年度に実施されたものです。