5S-01

音楽音響信号に対する相補的な歌声分離と音高推定

池宮 由楽 ‡　糸山 克寿 ‡　吉井 和佳 ‡

‡ 京都大学 大学院情報学研究科 知能情報学専攻

1. はじめに

歌声はポピュラー楽曲におけるメロディーラインを担っており，その特徴の多くを含んでいる．それゆえ，楽曲からの伴奏/歌声分離と歌声音高推定は，音楽情報検索 [1] や能動的音楽鑑賞 [2] といった幅広い分野への応用が期待される．例えば，分離された歌声や音高は歌手名同定に利用でき，伴奏を用いればオリジナル音源でのカラオケを楽しむことができる．

歌声分離と音高推定は相補的な関係を持つ．つまり，歌声がある程度分離されていれば，その音高の推定は比較的容易になり，逆に，歌声音高が既知であれば，それにより歌声分離の精度を向上できる．従来，これらのタスクは個別に行われており，例えば歌声音高推定の代表的なアプローチは，音響信号中で最も優勢な調波構造を抽出する [3] というものである．このとき，歌声以外の楽器の音高を推定する誤りが頻繁に起きる．

本稿では，歌声分離と音高推定を相補的に実行する手法を提案する．まず，音楽音響信号（混合音）を入力とし，ロバスト主成分分析を用いた歌声分離 [4] を行う．次に，分離歌声スペクトルをから音高推定を行い，最後に推定された音高軌跡を用いてさらに精細な分離を行う．これにより，他楽器音による音高推定の誤りを抑制するとともに，メロディーを担う歌声のみに着目した分離を行うことが可能となる．

2. 提案手法

図 1 に提案手法の全体図を示す．モノラル音楽音響信号を入力とし，時間周波数領域において歌声/伴奏スペクトルに分解する．これらを時間領域へ逆変換することで分離信号を得る．

2.1 ロバスト主成分分析を用いた歌声分離

Huang ら [4] は混合音の短時間フーリエ変換 (STFT) スペクトログラムにロバスト主成分分析 (RPCA) を適用することで世界最高水準の歌声分離を実現した．RPCA は行列 M を，低ランク行列 L とスパース行列 S へ分解するアルゴリズムであり，以下の最適化問題で表される．

$$\text{minimize} \ \ \|L\|_* + \lambda_k \|S\|_1 \ \ (\text{s.t.} \ \ L+S=M)$$

ここで $\|\cdot\|_*$，$\|\cdot\|_1$ はそれぞれ，核ノルム，L1-ノルムを表す．k は L の低ランク性，S のスパース性を調節するトレードオフパラメータである．

ドラムやリズムギターといった強い繰り返し構造を持つスペクトルは L，歌声などの頻繁に変動するスペクトルは S へと分解される．L と S から時間周波数領域のバイナリマスク (RPCA マスク) B_R を作成する．

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

バイナリマスクを混合音のスペクトログラムに適用することで，歌声のスペクトルを分離できる．

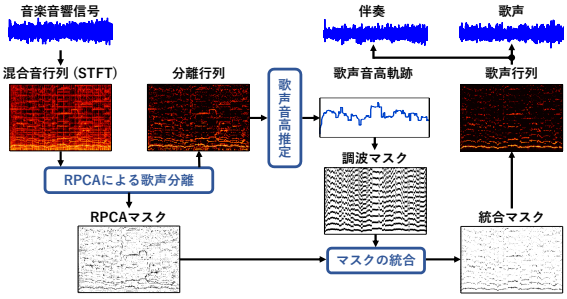

図 1: 提案手法の全体図

2.2 歌声音高推定

音高推定は，基本周波数 (F0) 軌跡を推定する問題である．STFT によるスペクトルはフーリエ変換の窓幅により周波数分解能が制限されている．そこで分解能を向上するため，振幅スペクトルを dB スケールでスプライン補間し，対数周波数軸上で一定間隔のスペクトルを得る．本稿では，分解能が 6 [cents] (200 bins per octave) となるように補間を行った．

まず人間の聴覚特性を考慮するため，スペクトルに対し A 特性関数を適用する．A 特性関数は等ラウドネス曲線の逆特性の近似関数であり，人間に聴こえにくい低周波のパワーを抑圧することができる．A 特性適用後のスペクトルから次式で Subharmonic Summation (SHS) を計算する．

$$S(t,c) = \sum_{n=1}^{N} h_n M_s^{\mathrm{A}} \left(t, c + \left\lfloor \frac{1200 \log_2 n}{6} \right\rfloor \right)$$

ここで，M_s^{A} は A 特性関数を適用した分離スペクトルを表し，t，c はそれぞれ，時間フレームと対数周波数ビンのインデクスである．また，N は足し合わせる倍音数，h_n は各倍音に対する重み係数であり本稿では 0.86^{n-1} とする．次に，SHS 関数から F0 軌跡を推定するため，最適経路探索問題を以下で定式化する．

$$\hat{C} = \underset{c_1,\ldots,c_T}{\operatorname{argmax}} \sum_{t=1}^{T-1} \left\{ \log \frac{S(t,c_t)}{\sum_{c_l}^{c_h} S(t,c)} + \log T(s_{t+1}|s_t) \right\}$$

ここで，c_l と c_h は歌声 F0 を探索する範囲の最低・最高周波数のインデクスである．$T(s_2|s_1)$ は現フレームの F0 s_1 から次フレームの F0 s_2 [cents] への遷移確率であり，標準偏差 150 [cents] のラプラス分布を用いる．ここで，時間フレーム間隔は 10 [msec] とする．この問題はビタビ探索により効率的に解くことができる．

2.3 歌声 F0 軌跡を用いた歌声分離

推定された歌声 F0 軌跡から，その倍音周辺をマスキングする調波マスク B_H を作成する．

$$B_H(t,f) = \begin{cases} 1 \ \text{if} \ nF_t - \frac{w}{2} < h(f) < nF_t + \frac{w}{2} \\ 0 \ \text{otherwise,} \end{cases}$$

ここで，F_t は時間フレーム t における推定 F0，w は各倍音におけるマスク幅 [Hz] である．$h(f)$ は周波数ビン f に対応する周波数 [Hz] を表す．RPCA マスクと調波

Mutually Dependent Vocal Separation and Melody Extraction for Music Audio Signals: *Ikemiya Yukara, Katsutoshi Itoyama, Kazuyoshi Yoshii (Kyoto Univ.)*

表 1: 歌声音高推定の結果 [%] (100 曲の平均)

RPCA	SHS-V	PreFEst-V	MELODIA-V	MELODIA
無し	71.50	70.07	67.79	69.97
有り	**77.41**	71.01	72.26	69.43

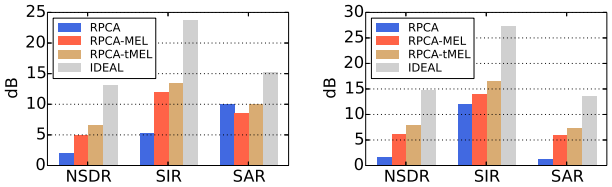

図 2: 歌声分離の結果 [dB]

マスクを統合し歌声/伴奏スペクトルを分離する．

$$M_v(t,f) = B_R(t,f)B_H(t,f)M(t,f)$$
$$M_a(t,f) = M(t,f) - M_v(t,f)$$

ここで M は混合音のスペクトルである．それぞれ混合音の位相を用いた逆 STFT により分離時間信号を得る．

3. 評価実験

提案手法の歌声音高推定・歌声分離に対する有効性を評価するため，以下の実験を行った．3.1，3.2 節におけるパラメータ設定は，RPCA のトレードオフパラメータ $k = 1.0$，調波マスクのマスク幅 $w = 80$ とする．

3.1 歌声音高推定

提案手法の歌声音高推定を，代表的なメロディー推定手法である PreFEst [3]，MELODIA[§] [5] と比較した．MELODIA は歌声音高軌跡の特徴を利用した手法であり，世界最高水準の精度を実現している．SHS の妥当性を評価するため，PreFEst，MELODIA で計算した尤度関数について 2.2 節で述べたビタビ探索を適用した．

SHS-V: A 特性関数+SHS+ビタビ探索 (提案手法)
PreFEst-V: PreFEst-core+ビタビ探索
MELODIA-V: MELODIA (salience)+ビタビ探索
MELODIA: The original MELODIA algorithm

また音高推定に対する歌声分離 (RPCA) の効果を調べるため，各手法について分離無し/有りの場合を比較した．推定精度は歌唱区間における F0 正解率とし，許容誤差は 50 [cents] とした．実験データとして，RWC 研究用音楽データベース [6] のポピュラー 100 曲を用いた．

表 1 に結果を示す．提案手法が最も高い精度を達成した．特に，RPCA により大幅に精度が向上しており，SHS と歌声分離の相性が良いことを示唆している．

3.2 歌声分離: 枠組みの評価

バイナリマスクの作成方法の違いにより，以下の四手法の歌声分離を比較した．

RPCA ：RPCA マスク [4]
RPCA-MEL ：RPCA+調波マスク（提案手法）
RPCA-tMEL ：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）

実験データには MIR-1K データセット[¶]を使用し，これは 110 曲 (20-110 秒, サンプリング周波数 16kHz) の歌唱入り楽曲からなる．評価尺度は，*BSS_EVAL toolbox*[‖] により得られる Normalized Source-to-Distortion Ratio (NSDR)，Source-to-Interference Ratio (SIR) と Sources-to-Artifacts Ratio (SAR) を用いた．NSDR, SIR と SAR はそれぞれ，目的音源以外の全ての雑音，目的音源以外の音源からの雑音，ミュージカルノイズなどの雑音からの比率 [dB] を表しており，値が大きいほど分離精度が優れていることを示す．

[§]http://mtg.upf.edu/technologies/melodia
[¶]sites.google.com/site/unvoicedsoundseparation/mir-1k
[‖]http://bass-db.gforge.inria.fr/bss_eval/

表 2: MIREX2014 の歌声分離タスクの結果 [dB]

手法	[7]	[8]	[9]	[10]	[11]	RNA1	[12]	**提案手法**
歌声	−1.40	−0.82	0.65	2.86	2.89	3.69	4.17	**4.48**
伴奏	0.35	−3.12	3.09	5.03	5.25	7.32	5.63	**7.87**

図 2 に結果を示す．**RPCA** と比較し，提案手法 (**RPCA-MEL**) により歌声・伴奏双方について分離精度が大幅に向上していることが分かる．歌声の SAR は減少しているが，これは **RPCA** が既に高い SAR を達成しているためであり，**RPCA-MEL** においても十分に高い値を維持している．

3.3 歌声分離: MIREX2014

The Music Information Retrieval Evaluation eXchange (MIREX) は音楽解析アルゴリズムの世界的なコンテストである．MIREX ではテストデータが未公開であるため，公平な評価が可能である．提案手法を評価するため，MIREX (2014 年) の歌声分離タスクへ参加した．参加手法 [7–12] は全て 2012–2014 年に提案された最新アルゴリズムである．表 2 に結果を示す．歌声・伴奏双方において，提案手法 ($k = 0.8$) が最高の分離性能を達成した．ここで評価尺度は NSDR [dB] である．

4. おわりに

本稿では，歌声分離と歌声音高推定を相補的に実行する手法を提案した．提案手法により，従来法と比較して歌声分離・音高推定双方で精度の向上を実現した．また，音楽解析の世界的なコンテストである MIREX2014 の歌声分離タスクにおいて，多くの最新手法が参加する中，最高の分離性能を達成した．

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

参考文献

[1] J. S. Downie: “Music Information Retrieval.”, *Annu. Rev. Inf. Sci. Technol.*, vol. 37, pp. 295–340, 2003.
[2] M. Goto: “Active Music Listening Interfaces Based on Signal Processing”, *Proc. ICASSP*, pp.1441–1444, 2007.
[3] M. Goto: “A Real-Time Music-Scene-Description System: Predominant-F0 Estimation for Detecting Melody and Bass Lines in Real-World Audio Signals”, *Speech Communication*, 2004.
[4] P. S. Huang *et al.*: “Singing-Voice Separation From Monaural Recordings Using Robust Principal Component Analysis”, *Proc. ICASSP*, 2012.
[5] J. Salamon*et al.*: “Melody Extraction from Polyphonic Music Signals using Pitch Contour Characteristics”, *IEEE TASLP*, 2012.
[6] M. Goto *et al.*: “RWC Music Database: Popular, Classical, and Jazz Music Databases”, *Proc. ISMIR*, 2002.
[7] P. S. Huang *et al.*: “Singing-Voice Separation from Monaural Recordings using Deep Recurrent Neural Networks”, *Proc. ISMIR*, 2014.
[8] F. Yen *et al.*: “Singing Voice Separation using Spectro-Temporal Modulation Features”, *Proc. ISMIR*, 2014.
[9] A. Liutkus *et al.*: “Kernel Additive Models for Source Separation”, *IEEE TSP*, 2014.
[10] Z. Rafii *et al.*: “Music/Voice Separation using the Similarity Matrix”, *Proc. ISMIR*, 2012.
[11] P. K. Yang *et al.*: “Bayesian Singing-Voice Separation”, *Proc. ISMIR*, 2014.
[12] I. Y. Jeong *et al.*: “Vocal Separation from Monaural Music Using Temporal/Spectral Continuity and Sparsity Constraints”, *Signal Processing Letters*, 2014.