

複数楽器個体による事前分布を用いた 調波・非調波統合モデルのパラメータ推定

糸山 克寿[†]

後藤 真孝[‡]

駒谷 和範[†]

尾形 哲也[†]

奥乃 博[†]

[†] 京都大学 大学院情報学研究科 知能情報学専攻

[‡] 産業技術総合研究所

1. はじめに

音楽音響信号の音源分離は、高度な信号処理技術という学術的な面からだけでなく、楽曲の局所的な編集による新たな音楽鑑賞方法の創出 [1] など、応用の面からも注目される技術である。ポピュラー音楽などの音楽音響信号を分離するためには、調波構造を持つ楽器音（音高を感じる音、ピアノなど）と調波構造を持たない楽器音（音高を感じない音、ドラムスなど）の混合音を適切に扱う必要がある。

我々は調波・非調波統合モデルを用いることであらゆる楽器音を統一的に扱える音源分離手法、及びそのモデルパラメータ推定手法 [2] を設計し、実現した。モデルパラメータ推定においては、非調波構造モデルに対する周波数方向への平滑化という制約を用いることでモデルの過学習問題を解決した。また、MIDI 音源から生成した混合音を用いて、統合モデルを用いることで調波構造モデルや非調波構造モデルを単独で用いるよりも分離性能が向上することを実験で確認した。しかしながら、分離には MIDI 音源で生成したテンプレート音が必要という課題が残されていた。テンプレート音を統合モデルパラメータ推定の際のパラメータの規範として用いると、テンプレート音に用いた楽器個体・奏法への過剰な適応が問題となる。

本稿では、上記の課題を解決するためのモデルパラメータの事前分布を用いた統合モデルのパラメータ推定手法について述べる。事前分布は楽器単音に対して統合モデルのパラメータを推定したものから学習する。

2. 調波・非調波統合モデル

調波・非調波統合モデル $J(t, f)$ は、楽器音の単音を表現するモデルで、調波構造モデル $H(t, f)$ と非調波構造モデル $I(t, f)$ の重み付き和で、式 (1) で定義する。 t, f はそれぞれ時間、周波数を表す。 w_H と w_I はそれぞれ $H(t, f)$ と $I(t, f)$ の重みで、 $w_H + w_I = 1$ を満たす。

$$J(t, f) = w_H H(t, f) + w_I I(t, f) \quad (1)$$

統合モデルで用いられる他のパラメータを表 1 に示す。

調波構造モデル $H(t, f)$ は、パラメトリックな基底関数であるガウス分布関数の線形和として、パワーエンベロープを表現する $E(m, t)$ と各時刻での高調波構造を表現する $F(m, n, t, f)$ を用いて式 (2) で定義する。ただし、 M, N は定数で、評価実験では $M = 10, N = 30$ とした。

$$H(t, f) = \sum_{m=0}^{M-1} \sum_{n=1}^N E(m, t) F(m, n, t, f) \quad (2a)$$

Parameter Estimation for Harmonic and Inharmonic Models Using Prior Distributions from Multiple Instrument Bodies: Katsutoshi Itoyama (Kyoto Univ.), Masataka Goto (AIST), Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)

表 1: 統合モデルのパラメータ

記号	意味
w_H	調波構造モデルの重み
w_I	非調波構造モデルの重み
$u(m)$	パワーエンベロープの概形 ($\sum_m u(m) = 1$)
$v(m, n)$	n 次高調波成分の相対強度 ($\sum_n v(m, n) = 1$)
τ	発音時刻
$M\varphi$	音長 (M は定数)
$\omega(t)$	時刻 t における基本周波数
σ	周波数方向へのガウス分布関数の広がり
$I(t, f)$	非調波構造モデル ($\int \int I(t, f) dt df = 1$)

$$E(m, t) = \frac{u(m)}{\sqrt{2\pi\varphi}} \exp\left(-\frac{(t - \tau - m\varphi)^2}{2\varphi^2}\right) \quad (2b)$$
$$F(m, n, t, f) = \frac{v(m, n)}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(f - n\omega(t))^2}{2\sigma^2}\right)$$

このモデルは、亀岡らの調波時間構造化クラスタリングで用いられる音源モデル [3] を参考に設計した。

非調波構造モデル $I(t, f)$ は、ノンパラメトリックな関数として、パワースペクトルの各時刻および周波数における周波数成分の強度を直接表現するように定義する。ただし、非調波構造モデルは任意のパワースペクトルを表現可能な自由度を持つため、パラメータ推定の際に過学習を起こす可能性がある。これを避けるために、非調波構造モデルに調波的なパワースペクトルを表現させなくする制約を与える。具体的には、非調波構造モデルをガウシアンフィルタで周波数方向に平滑化したものと非調波構造モデル自身との Kullback-Leibler Divergence (KLD) の最小化を行う。調波構造とは周波数方向に鋭いピークが等間隔で並んだ構造であるため、非調波構造モデルが周波数方向にピークを持たなければ調波構造を表現することも不可能になる。

3. 事前分布を用いたパラメータ推定

本章では、観測パワースペクトル $X(t, f)$ に対する統合モデルのパラメータの推定手法について述べる。

推定されたパラメータの「良し悪し」を計るために、 $X(t, f)$ と $J(t, f)$ との間の KLD をその尺度として導入し、これをモデルパラメータに関して最小化することで $X(t, f)$ に対する最適なパラメータを求める。これは、 $X(t, f)$ を 2 次元の観測確率密度関数とみなし、重み付きの混合分布として表現される統合モデルのパラメータの最尤推定に相当する。ただし、実際には最尤なモデルパラメータを解析的に求めることはできないため、Expectation-Maximization (EM) アルゴリズムを用いて反復的にパラメータを推定する。EM アルゴリズムの Q 関数、および具体的なパラメータ更新式の導出は紙面の制約上省略する。

3.1 事前分布

第1章で述べたように、パワースペクトルとの類似性のみに基づくパラメータ推定を行うと、特に混合音に大して複数モデルのパラメータを同時推定の際、モデルパラメータが過剰に混合音に適応してしまう可能性がある。この問題を避けるため、楽器ごとにパラメータの事前分布を作成し、モデルパラメータの「規範」とする。事前分布の学習は楽器ごとに行う。事前分布を用いない状態で楽器音に対して統合モデルを適応させたものを学習データとし、平均 μ 、共分散行列 Σ の正規分布を用いてモデルパラメータの分布を表す。ただし、パラメータによってはディリクレ分布なども使用可能である。

事前分布を用いた場合は最大事後確率推定でパラメータ推定を行う。その際、最小化されるコスト関数 (EM アルゴリズムでの負の Q 関数に相当) は次式で表される。

$$Q = \sum_{m=0}^{M-1} \sum_{n=1}^N \Delta_H(m, n, t, f) X(t, f) \times \log \frac{\Delta_H(m, n, t, f) X(t, f)}{w_H E(m, t) F(m, n, t, f)} + \Delta_I(t, f) X(t, f) \log \frac{\Delta_I(t, f) X(t, f)}{w_I w_I I(t, f)} + (\theta - \mu)^T \Sigma^{-1} (\theta - \mu) \quad (3)$$

式 (3) の最終項が事前分布によって追加された部分である。 $\Delta_H(m, n, t, f)$, $\Delta_I(t, f)$ は EM アルゴリズムでのパラメータ推定に用いる潜在変数の期待値で、 $X(t, f)$ を統合モデルの構成要素のパワー比に応じて各要素へと分配する。パラメータ更新式の導出は紙面の制約上省略する。

4. 評価実験

本手法の有効性を確認するため、評価実験を行った。本稿で述べた手法、および音楽音響信号に同期した Standard MIDI File (SMF) を用いて混合音中の各単音のモデルパラメータを推定し、そのモデルを用いて音楽音響信号を楽器パートごとに分離 [2] した。事前分布の学習データには RWC 研究用音楽データベース: 楽器音 [4] を、分離対象の楽曲には同データベース: ポピュラー音楽から選んだ 5 曲を用いた。各楽曲は、歌声とドラムス、および楽器音データベースに含まれない電子楽器によるパートをあらかじめ除き、残りの楽器パート (ピアノ、ギター、オルガンなど) を個別に MIDI 音源で生成した後に各パートをミックスしたもので、および対応する実楽曲のミックス前の音響信号のうち、MIDI 音源のものと同じパートを混合したものの [5] を用いた。

事前分布は、楽器音データベース中の楽器個体をそれぞれの楽器ごとに 1 つだけ用いた場合、2 つ用いた場合、3 つすべて用いた場合の 3 通りを学習した。それぞれの事前分布を用いてパラメータ推定と音源分離を行った場合、および比較対象として事前分布を用いずに分離データとは異なる MIDI 音源で生成したテンプレートとのマッチングによって音源分離を行った場合に関して評価した。MIDI 音源での楽曲の分離信号の SNR を表 2 に、実楽曲の分離信号の SNR を表 3 に示す。

事前分布に用いる楽器個体数の増加に伴って、SNR も向上していることが分かる。表 2 では、個体数が 3 の場

表 2: MIDI 音源による音響信号を分離した場合の SNR [dB]

	個体数 1	個体数 2	個体数 3	テンプレート
P001	52.5	53.1	52.6	54.9
P002	53.2	53.2	53.5	53.3
P003	52.5	52.8	53.4	51.8
P008	49.3	49.0	49.5	46.2
P010	50.2	49.9	50.0	53.2
Ave.	51.5	51.6	51.8	51.8

表 3: 実楽曲の音響信号を分離した場合の SNR [dB]

	個体数 1	個体数 2	個体数 3	テンプレート
P001	50.3	50.4	50.4	50.0
P002	52.3	52.4	52.6	52.5
P003	45.2	45.3	45.4	44.4
P008	48.2	48.3	48.4	47.2
P010	49.4	49.5	49.5	49.0
Ave.	49.1	49.2	49.3	48.6

合は、テンプレートを用いて分離した場合と平均でほぼ等しい SNR を示している。テンプレート音を用いた場合は、楽曲によって SNR の上下が大きい (P001, P010 では事前分布を用いるよりも SNR が大きい、P003, P008 では逆に小さくなっている)。また、表 3 では、個体数が 1 の場合でもテンプレートを用いた場合よりも SNR が大きくなっている。テンプレートを用いると、テンプレート音と分離音の音色の違いが分離性能に影響するが、事前分布は複数の楽器個体・単音を用いるため、音色の違いに頑健なパラメータ推定が行われたと考えられる。

5. おわりに

本稿では、複数楽器個体による事前分布を用いた統合モデルのパラメータ推定の手法を述べ、事前分布に用いる楽器個体数を増加させることで推定されたモデルによる音源分離性能が向上することを示した。本手法では、単一のテンプレート音ではなく、多数の学習データを用いて事前分布を作成するため、楽器個体差だけでなく、奏法に起因する音色の変動にも頑健なパラメータ推定が可能になると期待される。今後は、ドラム音などの非調波成分のみからなる楽器音も追加した分離実験などを行う予定である。

謝辞 本研究の一部は、科研費、グローバル COE、Crest-Muse の支援をうけた。

参考文献

- [1] K. Yoshii *et al.*, “Drumix: An Audio Player with Real-time Drum-part Rearrangement Functions for Active Music Listening,” *IPSJ Journal*, vol. 48, No. 3, pp. 134–144, 2007.
- [2] 糸山他, “楽譜情報を援用した多重奏音楽音響信号の音源分離と調波・非調波統合モデルの制約付パラメータ推定の同時実現,” 情処論, Vol. 49, No. 3, 2008 (in Press).
- [3] H. Kameoka *et al.*, “A Multipitch Analyzer Based on Harmonic Temporal Structured Clustering,” *IEEE Trans. on ASLP*, Vol. 15, No. 3, 2007.
- [4] 後藤他, “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,” 情処論, Vol. 45, No. 3, pp. 728–738, 2004.
- [5] M. Goto, “AIST Annotation for the RWC Music Database,” *ISMIR 2006*, pp. 359–360.