4T-8

# 非負値調波時間構造因子分解法に基づく音楽音響信号の多重基本周波数解析

阪上 大地　　大塚 琢馬　　糸山 克寿　　奥乃 博

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

音楽音響信号は様々な音高の音が重ね合わさって構成され，時間方向の音の並びはメロディ，周波数方向の並びは和音と呼ばれる．音楽音響信号中のすべての音の音高・発音区間を同時に推定する多重基本周波数解析 [1–4] は，楽曲検索 [5]，和音・印象分析 [6]，音響信号加工 [7] など様々な応用研究を支える重要な技術である．

音楽はメロディ・対旋律・和音・リズム・楽器などの複合体として認識されるため，高精度な多重音解析を行うためには，スペクトログラムに対する直接のモデル化に加え，様々な潜在的特徴を統一的にモデル化する必要がある．従来，ベイズ推論にもとづく手法として非負値行列因子分解 (NMF) [8] と潜在的調波配分法 (LHA) [2] という 2 つの代表的手法が開発されているが，NMF は音量を，LHA は音高・音色のみをモデル化しており，音の三要素すべてを同時にモデル化していなかった．非負値調波因子分解 (NHF) [3] では NMF と LHA を統合し，これらの同時推定を実現したものの，楽器音の音量の時間的な推移に対する制約はなく，音量の時間的連続性が考慮されない問題が残っていた．本稿では，この問題を解決するため，NHF の各楽器音の音量推移を混合ガウス分布による滑らかな包絡線としてモデル化する手法を報告する．

我々は，音楽音響信号の観測スペクトログラムが高々 $K$ 種類の調波音の重ね合わせで成り立っていると仮定し，各調波音のエネルギー分布のスペクトル包絡と音量包絡をそれぞれ混合ガウス分布によってモデル化する．本手法を非負値調波時間構造因子分解法 (Nonnegative Harmonic-Temporal Factorization; NHTF) と呼ぶ．また，各調波音から生成されるスペクトログラムの時間周波数成分は，各点でのエネルギー密度をパラメータとするポアソン分布から生成されると仮定する．一般に，非共役な確率分布の組を使用するモデルでは，更新式の導出は困難である．一方，本手法では，エネルギー密度関数の積分幅 $\epsilon_{\mathrm{T}}, \epsilon_{\mathrm{F}} \to 0$ の極限で事後分布が共役形となり，容易に更新式が導出できることを示す．

## 2. 観測音のモデル化

観測されたウェーブレットスペクトログラムを以下のように確率的にモデル化する．時間フレーム，周波数ビン，基底，倍音のインデックスを $t, f, k, m$，ディリクレ分布，ガンマ分布，正規分布，ポアソン分布，ウィシャート分布の確率密度をそれぞれ $\mathcal{D}, \mathcal{G}, \mathcal{N}, \mathcal{P}, \mathcal{W}$ で表す．本手法では，各時間周波数点での振幅 $x_{tf}$ をある微小量の整数倍に量子化し，残差を切り捨てる．これを微小な音エネルギー粒子からなる二次元ヒストグラムと考える．このとき，各点の観測値はすべて整数となるため，ポアソン分布でモデル化できる．観測信号が $K$ 種類の調波音から生成されており，それぞれが独立に生成した音粒子のヒストグラム $s^k_{tf}$ の合計値が観測されていると仮定すると，$x_{tf} = \sum_k s^k_{tf}$ が成り立つ．$s^k_{tf}$ は直接観測できないため，潜在変数と呼ばれている．

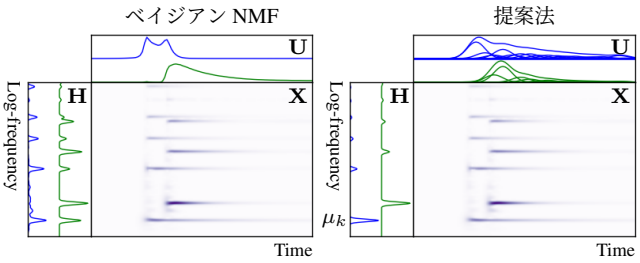


図 1: ベイジアン NMF と提案法の比較．提案法では調波構造・時間包絡が GMM により陽にモデル化されている．

NMF では，各潜在スペクトログラムは定常スペクトル $h^k_f$ と各時刻での音量 $u^k_t$ の積としてモデル化される．通常，これらの成分は各周波数ビン・時間フレームごとに独立と仮定されるが，実際には各基底の周波数成分・音量変化は多峰形の包絡線にしたがうため，独立とはいえない．そのため，NHF ではこのうち定常スペクトル部分を調波構造に対応する混合ガウス分布でモデル化した．数式で表すと，以下のようになる．

$$h^k_f = \sum_{m=1}^{M} \eta_{km} \int_{x_f-\epsilon_{\mathrm{F}}/2}^{x_f+\epsilon_{\mathrm{F}}/2} \mathcal{N}(x|\mu_k + o_m, \lambda_k^{-1}) dx$$
$$= \epsilon_{\mathrm{F}} \sum_{m=1}^{M} \eta_{km} \mathcal{N}(x_f|\mu_k + o_m, \lambda_k^{-1}) \quad (1)$$

ここで，$x_f$ は $f$ 番目の周波数ビンの対数周波数，$\eta_{km}$ は $m$ 番目の倍音の重み，$M$ は倍音数，$\mu_k$ は基本周波数，$\lambda_k$ は正規分布の精度である．$o_m = 1200 \log_2 m$ は各倍音の相対位置を表す．更新式の導出を簡単にするため，式中の正規分布が音エネルギーの分布密度であると想定し，$h^k_f$ は連続なエネルギー密度分布を各周波数ビンの周囲 $\pm\epsilon_{\mathrm{F}}/2$ の範囲で積分した値とする．この定式化により，各基底は基本周波数とその倍音の周囲でのみ大きな値をとるようになる．

本稿ではさらに，HTC [1] にならい，音量の時間構造についても混合ガウス分布を導入する（図 1）．これは以下の通りになる．

$$u^k_t = \sum_{\tau=1}^{N_\tau} u^k_\tau \int_{y_t-\epsilon_{\mathrm{T}}/2}^{y_t+\epsilon_{\mathrm{T}}/2} \mathcal{N}(y|\tau T, \lambda_{\mathrm{T}}^{-1}) dy$$
$$= \epsilon_{\mathrm{T}} \sum_{\tau=1}^{N_\tau} u^k_\tau \mathcal{N}(y_t|\tau T, \lambda_{\mathrm{T}}^{-1}) \quad (2)$$

ここで，$y_t$ は $t$ 番目の時間フレームの時刻，$\tau$ は正規分布のインデックス，$N_\tau$ は時間包絡を表現するガウス分布の個数，$u^k_\tau$ は $\tau$ 番目の時間コンポーネントの音量，$T$ は時間包絡を構成する正規分布の間隔，$\lambda_{\mathrm{T}}$ は精度，$\epsilon_{\mathrm{T}}$ は積分幅である．計算を簡単にするため，$\epsilon_{\mathrm{T}}, \epsilon_{\mathrm{F}} \ll 1$ を仮定する．モデル全体の尤度関数および事前分布は以下

の通りである．

$$x_{tf} = \sum_{\tau km} s_{tf}^{\tau km} \tag{3}$$

$$p(s_{tf}^{\tau km}|u,\eta,\mu,\lambda) = \mathcal{P}(s_{tf}^{\tau km}|\epsilon_{\mathrm{T}}\epsilon_{\mathrm{F}}u_\tau^k\eta_{km}\mathcal{N}(y_t|\tau T,\lambda_{\mathrm{T}}^{-1}) \times \mathcal{N}(x_f|\mu_k+o_m,\lambda_k^{-1})) \tag{4}$$

$$p(u_\tau^k) = \mathcal{G}(u_\tau^k|a_0,b_0) \qquad p(\eta_k) = \mathcal{D}(\eta_k|\alpha_m^0) \tag{5}$$

$$p(\mu_k,\lambda_k) = \mathcal{N}(\mu_k|m_0,(\beta_0\lambda_k)^{-1}))\mathcal{W}(\lambda_k|w_0,\nu_0) \tag{6}$$

$a_0, b_0, \alpha_m^0, m_0, \beta_0, w_0, \nu_0$ はモデルのハイパーパラメータである．

## 3. 更新式の導出

全潜在変数の同時分布を近似的に計算するため，これを $p(S,u,\eta,\mu,\lambda|X) \approx q(S)\prod_{\tau k} q(u_\tau^k)\prod_k\{q(\eta_k)\,q(\mu_k,\lambda_k)\}$ の形に因子分解し，各分布を変分ベイズ法によって交互に更新する．一般に，ポアソン分布のパラメータがディリクレ分布や正規ウィシャート分布など共役でない事前分布に従う手法では，更新式の導出は簡単ではない．しかし提案法の場合，$\epsilon_{\mathrm{T}}, \epsilon_{\mathrm{F}} \to 0$ の極限で変分事後分布が共役形になるので，更新式は容易に導出できる．たとえば，倍音比率 $\eta_{km}$ の最適な変分事後分布 $q^*(\eta_k)$ は

$$\begin{aligned}\ln q^*(\eta_k) &= \sum_{t\tau fm} \mathbb{E}[\ln p(S|u,\eta,\mu,\lambda)] + \ln p(\eta_k) \\ &= \sum_{t\tau fm} \mathbb{E}[s_{tf}^{\tau km}]\ln\eta_{km} + \ln p(\eta_k) \\ &\quad - \epsilon_{\mathrm{T}}\epsilon_{\mathrm{F}}\mathbb{E}[u_\tau^k\mathcal{N}(x_f|\cdots)\mathcal{N}(y_t|\cdots)]\eta_{km} \\ &\approx \ln\mathcal{D}(\eta_k|\alpha_{km}) \end{aligned} \tag{7}$$

$$\alpha_{km} = \sum_{t\tau f}\mathbb{E}[s_{tf}^{\tau km}] + \alpha_m^0 \tag{8}$$

の形になる．これは事前分布と同じくディリクレ分布になっているため，事後分布は共役形である．$u_\tau^k, \mu_k, \lambda_k$ についても同様に極限を取ることで事後分布の計算ができる．なお，$s_{tf}^{\tau km}$ の更新式は極限を用いずに導出できる．

## 4. 実験及び考察

従来法 (LHA, NHF) および本手法 (NHTF) に二種類の事前分布を設定して多重基本周波数推定を行い，結果を比較した．

### 4.1 観測スペクトログラムの作成

実験には，RWC 音楽データベース [9] の楽曲 40 曲，冒頭各 32 秒を使用した．内訳は，ピアノソロ 5 曲 (Jazz, No. 1–5)，ギターソロ 5 曲 (Jazz, No. 6–10)，ジャズデュオ 10 曲 (Jazz, No. 11–20)，ジャズバンド 10 曲 (Jazz, No. 21–30)，室内楽 10 曲 (Classic, No. 12–21) である．各楽曲は MIDI 音源 (YAMAHA MOTIF-XS) を用いて録音し，30 ∼ 3000 [Hz] の範囲でウェーブレットスペクトログラムに変換した．

### 4.2 実験条件

**事前分布とモデル次数**　実験では二種類の事前分布のもと性能評価を行った．ひとつは無情報事前分布であり，このとき各ハイパーパラメータを $a_0, b_0, \alpha_m^0, \beta_0, w_0, \nu_0 = 1, m_0 = 0$ に設定した．もうひとつの設定では，調波構造の各倍音の重みが徐々に小さくなるように，$\alpha_m^0 = 0.6547m^{-2}\sum_{tf}x_{tf}$ と設定した．この倍音比率は HTC [1] と同じ値である．なお，音源モデル数 $K = 73$, 倍音数 $M = 6$ とした．

表 1: 多重基本周波数推定の F 値．太字は最大値を表す．LHA と NHF は更新式が同形のため，同一性能となる．

| | 無情報事前分布 | | | 事前分布あり | | |
|---|---|---|---|---|---|---|
| Genre | LHA | NHF | NHTF | LHA | NHF | NHTF |
| Piano | 0.558 | 0.558 | 0.590 | 0.584 | 0.584 | **0.590** |
| Guitar | 0.684 | 0.684 | 0.726 | 0.728 | 0.728 | **0.740** |
| Jazz (Duo) | 0.524 | 0.524 | 0.545 | 0.552 | 0.552 | **0.556** |
| Jazz (Trio∼) | 0.523 | 0.523 | **0.548** | 0.536 | 0.536 | 0.541 |
| Chamber | 0.481 | 0.481 | 0.508 | 0.503 | 0.503 | **0.512** |

**初期化**　初期化では，各音源モデルが C1 から C7 までの 6 オクターブをカバーするよう半音毎の基本周波数を設定した．各音源モデルの時刻 $t$ での音量は，各倍音に最も近い周波数ビンの振幅を $2^{-m}$ で重み付け，その総和に比例するよう設定した．倍音比率も同様に，$\eta_{km} \propto 2^{-m}$ に設定した．

**発音スレッショルドと正解率**　$N_{tk} = \sum_{\tau fm} s_{tf}^{\tau km}$ を各基底の時刻 $t$ での推定音量と考え，この値が一定のスレッショルドを超えているかどうかで発音・消音を判断した．各調波音に対する推定結果は基本周波数の推定値をもとに MIDI のノートナンバーに対応付け，$T \times 128$ の二値行列を出力とした．各手法の潜在的な推定性能を評価するため，スレッショルドは手法・楽曲ごとに最適化した．これを MIDI ファイルから直接作成した正解データと比較し，F 値を計算した．F 値は適合率と再現率の調和平均であり，値が大きいほど性能が高いことを示す．

### 4.3 実験結果

実験結果を表 1 に示す．いずれの音楽ジャンル・事前分布の下でも提案法が従来法を上回り，性能向上率は平均 1.8 ポイントであった．なお，実験に使用した楽曲のうち，4 ジャンルでは倍音比率の事前分布を設定することで性能が向上したが，残りの 1 ジャンルでは無情報事前分布がより高い性能を示した．結果として，各調波音の時間的連続性を陽にモデル化することで，多重基本周波数推定の性能が上がることが示せた．

## 5. おわりに

本稿では，調波クラスタリングと NMF を確率的に統合し，楽曲中の各音の時間的連続性を考慮した多重基本周波数推定法を報告した．今後は，楽器固有の音色や楽曲構造を反映したより高度な推論モデルにより推定精度の向上を目指したい．なお，本研究の一部は科研費 (S) No. 24220006，若手 (B) No. 24700168 の支援を受けた．

## 参考文献

[1] H. Kameoka *et al.*: A multipitch analyzer based on harmonic temporal structured clustering, *IEEE Trans. on ASLP*, vol.15, no.3 (2007), pp. 982–994.
[2] 吉井 他：多重音基本周波数解析のための無限潜在的調波配分法, 情処研報, 2010-MUS-86, 2010.
[3] 阪上 他：ベイジアン非負値調波因子分解と多重基本周波数推定への応用，情処研報，2012-MUS-96, 2012.
[4] D. Sakaue *et al.*: Initialization-Robust Multipitch Estimation based on Latent Harmonic Allocation using Overtone Corpus, *ICASSP 2012*, pp. 425–428.
[5] B. A. Casey *et al.*: Content-Based Music Information Retrieval: Current Directions and Future Challenges, *Proc. IEEE*, vol.96, no.4 (2008), pp. 668–696.
[6] Y. Ueda *et al.*: HMM-based approach for automatic chord detection using refined acoustic features, *ICASSP 2010*, pp. 5518–5521, 2010.
[7] K. Itoyama *et al.*: Parameter estimation for harmonic and inharmonic models by using timbre feature distributions, 情処論, vol.50, no.7 (2009), pp. 1757–1767.
[8] A. T. Cemgil *et al.*: Bayesian Inference for Nonnegative Matrix Factorization Models, *Tech. Rep. CUED/F-INFENG/TR.609*, 2008.
[9] 後藤 他:RWC 研究用音楽データベース，情処論, vol.45, no.3 (2004), pp. 728–738.