Vol.2015-MUS-107 No.50
2015/5/24

全極スペクトルモデルと擬似周期信号モデルのウェーブレット変換を用いた多重音スペクトログラムの調波時間因子分解

中村 友彦[1,a)]　亀岡 弘和[1,2,b)]

概要：多重音解析では，観測振幅スペクトログラムを非負値行列とみなし非負値行列因子分解を適用するアプローチと，計算論的聴覚情景分析に基づくアプローチが主に用いられてきた．本報告では，この２つのアプローチの利点を兼ね備えた新たなスペクトログラムモデルを導出し，それに基づく多重音解析手法である調波時間因子分解について述べる．音源の連続時間信号モデルとして擬似周期信号から出発し，各音源のウェーブレット変換を導出した後，それらを重畳したもので音楽音響信号のスペクトログラムを表現する．さらに，離散時間信号領域で定義された自己回帰モデルをスペクトログラムモデルに導入にできることを示す．このモデルに対し，ビブラートやポルタメントで起こる基本周波数の時間変動や，調による音高の出現頻度の偏りなどの音楽特有の性質を補助情報として組み込んだ，効率的なパラメータ推論アルゴリズムを導出する．

1. はじめに

多重音解析は，複数の音源信号が混合された観測信号から個々の音源の情報（基本周波数，発音開始時刻，音量など）を得る処理である．これは音楽情報処理の重要課題の一つであり，音楽情報検索や自動採譜，音楽音響信号加工など様々なアプリケーションの基礎技術である．

多チャンネル音響信号を入力とする場合には音源の空間的な手がかりを多重音解析に利用できるが，モノラル音響信号が入力である場合にはこれに代わる何らかの手がかりが必要である．モノラル音響信号を入力とする多重音解析のためのアプローチとして，計算論的聴覚情景分析のコンセプトに基づく手法が提案されている [1–3]．計算論的聴覚情景分析とは，Bregman によって提唱された聴覚情景分析 [4] で示された人間の聴覚機能を計算機で実現しようという試みである．調波時間構造化クラスタリング（Harmonic-Temporal Clustering, HTC）[2, 3] はこのアプローチに則り，人間が混在する複数の音の中で個々の音を聞き分けるために用いると考えられている要件（調波性，連続性，同時性，同期性）を時間周波数成分の局所的な制約として記述し，当該要件を満たすように観測信号の時間周波数成分をクラスタリングすることにより多重音解析を行う手法である．このアプローチでは，楽音のスペクトログラムの局所的な構造に着目し楽音に詳細な仮定を課している．

一方で，楽音に詳細な仮定を課さない手法として音楽の大局的な構造に着目したアプローチも提案されている．このアプローチでは，観測振幅スペクトログラムを非負値行列とみなして非負値行列因子分解（Non-negative Matrix Factorization, NMF）[5] を適用する．NMF は，限られた種類の音高の楽音がそれぞれ異なるタイミングで繰り返し生起するという音楽特有の性質に着目し，限られた種類のスペクトルテンプレートの適当な重み付き和で各時刻の観測スペクトルを表現できるという仮定を用いている．すなわち，観測振幅スペクトログラムは，2 つの非負値行列（各列がスペクトルテンプレートを表す基底行列と，各行が対応するスペクトルテンプレートの音量の時間発展を表すアクティベーション行列）の積として表現される．したがって，観測振幅スペクトログラムをこの 2 つの行列に分解することにより，スペクトルテンプレートと重みを同時推定し，各楽音の振幅スペクトログラムに分離できる．

前者のアプローチは音源のスペクトログラムの局所的な構造，後者のアプローチは音楽音響信号のスペクトログラムの大局的な構造を手がかりとしている．これらは互いに矛盾するわけではなく，いずれも高精度な多重音解析を実現するには有用な手がかりである．そこで，我々はこれら 2 つの手がかりを同時に取り入れたスペクトログラムモデ

[1] 東京大学大学院情報理工学系研究科，東京都文京区本郷 7-3-1，113-0033
[2] 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所，神奈川県厚木市森の里若宮 3-1，243-0198
[a)] Tomohiko_Nakamura@ipc.i.u-tokyo.ac.jp
[b)] kameoka@hil.t.u-tokyo.ac.jp/kameoka.hirokazu@lab.ntt.co.jp

ルを構築し，それに基づく多重音解析手法，調波時間因子分解（Harmonic-Temporal Factor Decomposition, HTFD）を提案する [6–8].

ところで，近年の音楽情報検索に関する国際会議や国際コンテスト [9] では，調推定，和音推定などの研究が急速に進展している．調，和音などの情報が高精度に得られるのであれば，多重音解析において有用な補助情報となりうる．そこで，HTFD では調推定，和音推定により得られた情報（音楽事前情報）をスペクトログラムモデルに組み込み，モデルパラメータの推定に活用するフレームワークも提案する．音源分離では，ユーザにより入力された情報を用いて分離精度を向上するユーザガイド付き音源分離手法 [10,11] や，楽譜を補助情報とする音源分離手法 [12,13] などの研究が進められており，提案するフレームワークも補助情報付き音源分離の一種として位置づけられる．

以下，実数集合と複素数集合，虚数単位をそれぞれ $\mathbb{R}, \mathbb{C}, \mathrm{j} := \sqrt{-1}$ と表記する．

2. 音楽音響信号のスペクトログラムの確率モデル化

2.1 楽音信号のウェーブレット変換

本節では，[3] に倣って楽音信号のウェーブレット変換領域でのモデルを導出する．多くの楽音は局所的に周期的とみなせるので，楽音の解析的なモデルとして周期や調波成分のパワーが時間的に滑らかに変化する擬似周期信号を用いることができる．音高のインデックスを k，調波成分のインデックスを $n = 1, 2, \ldots, N-1$ とする．音高 k の楽音の連続時間信号を，n 次調波成分の瞬時位相が $n\theta_k(u) \in \mathbb{R}$，瞬時振幅が $a_{k,n}(u) \in \mathbb{C}$ の擬似周期信号は

$$f_k(u) = \sum_{n=1}^{N} a_{k,n}(u) e^{\mathrm{j}(n\theta_k(u)+\varphi_{k,n})} \tag{1}$$

と記述できる．ここで，$u \in \mathbb{R}$ は連続時間信号領域での時刻，$\varphi_{k,n} \in \mathbb{R}$ は初期位相である．$f_k(u)$ の連続ウェーブレット変換を得るための基底関数 $\psi_{\alpha,t}(u)$ を

$$\psi_{\alpha,t}(u) = \frac{1}{\sqrt{2\pi}\alpha} \psi\left(\frac{u-t}{\alpha}\right) \tag{2}$$

と定義する．ここで，$\alpha > 0$ はスケールパラメータ，$t \in \mathbb{R}$ は時間シフトパラメータ，$\psi(u)$ はアドミッシブル条件を満たす中心周波数 1 のアナライジングウェーブレットである．これを用いて $f_k(u)$ のウェーブレット変換は，

$$W_k(\ln \tfrac{1}{\alpha}, t) = \int_{-\infty}^{\infty} \sum_{n=1}^{N} a_{k,n}(u) e^{\mathrm{j}(n\theta_k(u)+\varphi_{k,n})} \psi^*_{\alpha,t}(u) \mathrm{d}u \tag{3}$$

と書ける．$\psi^*_{\alpha,t}(u)$ の優勢な部分は時刻 t の周りにのみ存在するので，$W_k(\ln \frac{1}{\alpha}, t)$ は時刻 t 周りの $\theta_k(u)$ と $a_{k,n}(u)$ に強く依存する．そこで，$\theta_k(u) \simeq \theta_k(t) + \dot{\theta}_k(t)(u-t), a_{k,n}(u) \simeq a_{k,n}(t)$ と近似し Parseval の定理を適用すれば，対数周波数（$x := \ln(1/\alpha)$）と対数瞬時 F_0（$\Omega_k(t) = \ln \dot{\theta}_k(t)$）を用いて，$W_k(x,t)$ は

$$\sum_{n=1}^{N} a_{k,n}(t) \Psi^*(ne^{-x+\Omega_k(t)}) e^{\mathrm{j}(n\theta_k(t)+\varphi_{k,n})}, \tag{4}$$

と表せる．ここで，$\dot{\theta}_k(u)$ は瞬時 F_0 である．ψ の Fourier 変換である Ψ は任意に選べるので，$\omega = 1$ で最大値をとる対数正規分布型の実関数 [3]

$$\Psi(\omega) = \begin{cases} e^{-\frac{(\ln \omega)^2}{4\sigma^2}} & (\omega > 0) \\ 0 & (\omega \le 0) \end{cases}. \tag{5}$$

を用いる．σ は $\Psi(\omega)$ を $\ln\omega$ 軸で見たときの標準偏差に対応する．式 (5) より，$W_k(x,t)$ は

$$W_k(x,t) = \sum_{n=1}^{N} a_{k,n}(t) e^{-\frac{(x-\Omega_k(t)-\ln n)^2}{4\sigma^2}} e^{\mathrm{j}(n\theta_k(t)+\varphi_{k,n})}. \tag{6}$$

と変形でき，さらに調波成分のパワースペクトルが加法的であると仮定すれば，$|W_k(x,t)|^2$ は

$$|W_k(x,t)|^2 \simeq \sum_{n=1}^{N} |a_{k,n}(t)|^2 e^{-\frac{(x-\Omega_k(t)-\ln n)^2}{2\sigma^2}} \tag{7}$$

と書ける．このモデルは，HTC で採用された調波時間構造化モデル [3] と同一であり，その時刻 t での断面は調波的に正規分布形の関数が並んだ混合正規分布モデルと同形の関数で表される．

ここまでスペクトログラムモデルを連続時間，連続対数周波数領域で定義してきたが，計算機で実際に得られる観測スペクトログラムは時刻 t と対数周波数 x に関して離散的である．そのため本節以降は，等間隔に量子化された時刻 t_m $(m = 0, 1, \ldots, M-1)$ と対数周波数 x_l $(l = 0, 1, \ldots, L-1)$ を用いて，観測スペクトログラムを $Y_{l,m} := Y(x_l, t_m)$ と表す．同様に $\Omega_{k,m} := \Omega_k(t_m)$，$a_{k,n,m} := a_{k,n}(t_m)$ とする．

2.2 ソース・フィルタモデルの導入

ここで，擬似周期信号について再考してみよう．擬似周期信号には局所的に周期的で滑らかに秋季や振幅が変化すること以外は仮定が課されていない．そのため，実際の楽音とは異なるスペクトル形状も許容してしまっている．そこで，楽音の物理的な生成過程に着目し適切に楽音のスペクトル形状に制約を与えることを考えよう．

楽音の物理的な生成過程はソース・フィルタモデルでよく表現でき，離散時間信号に対する自己回帰過程として記述できる．しかし，調波時間構造化モデルが定義されたのはウェーブレット変換領域であるため，式 (7) のパラメータと自己回帰過程のパラメータとの対応関係を直接得ることは難しい．そこで本節では，[14] に従ってこれらのパラメータの関係を式 (1) のパラメータを介して得ることを目指す．

離散時間インデックスを i とし，時刻 t_m での式 (7) の断

面に対応する連続時間信号モデルの離散時間表現を $f_{k,m}[i]$ とする．この $f_{k,m}[i]$ が P 次の自己回帰過程によって

$$\beta_{k,m}[0]f_{k,m}[i] = \sum_{p=1}^{P}\beta_{k,m}[p]f_{k,m}[i-p] + \epsilon_{k,m}[i], \qquad (8)$$

と記述できるとする．ここで，$\beta_{k,m}[p]\ (p = 0, 1, \ldots, P)$ は自己回帰過程のパラメータ（線形予測係数とも呼ばれる）である．この自己回帰過程は $\beta_{k,m}[p]$ をパラメータとする全極システムと等価であるため，$\epsilon_{k,m}[i]$ は全極システムの励起信号とみなせる．2.1 節で仮定したように，$f_{k,m}[i]$ の F_0 は $e^{\Omega_{k,m}}$ であるから，励起信号 $\epsilon_{k,m}[i]$ の F_0 も $e^{\Omega_{k,m}}$ でなくてはならない．したがって，$\epsilon_{k,m}[i]$ は

$$\epsilon_{k,m}[i] = \sum_{n=1}^{N} v_{k,n,m} e^{\mathrm{j}ne^{\Omega_{k,m}}iu_0}, \qquad (9)$$

と記述できる．ただし，$u_0 > 0$ は離散時間表現のサンプリング周期であり，$v_{k,n,m} \in \mathbb{C}$ は n 番目の調波成分の複素振幅を表す．ここで詳細は省略するが，$f_{k,m}[i]$ の離散時間 Fourier 変換（discrete-time Fourier transform, DTFT）に逆 DTFT を適用すると，

$$f_{k,m}[i] = \sum_{n=1}^{N} \frac{v_{k,n,m}}{B_{k,m}(e^{\mathrm{j}ne^{\Omega_{k,m}}u_0})} e^{\mathrm{j}ne^{\Omega_{k,m}}iu_0} \qquad (10)$$

$$B_{k,m}(z) := \sum_{p=0}^{P}\beta_{k,m}[p]z^{-p} \qquad (11)$$

として，$f_{k,m}[i]$ の異なる表現が得られる．式 (10) と式 (1) の離散時間表現を比較すると，全極システムのパラメータと 2.1 節で導入したパラメータの対応関係が，

$$|a_{k,n,m}| = \left|\frac{v_{k,n,m}}{B_{k,m}(e^{\mathrm{j}ne^{\Omega_{k,m}}u_0})}\right| \qquad (12)$$

と陽に得られる．

2.3 モデルパラメータに対する拘束

NMF における重要な仮定は，音源のスペクトルを時変な成分と時不変な成分の積として表現することであり，これによりモデルパラメータの空間を適切に制限できる．したがって，このモデルにおいてもスペクトルの構成要素がそれぞれ時変な成分または時不変な成分のどちらとみなされるべきかが重要となるはずである．F_0 はビブラートやポルタメント中は大きく時間変動し，スペクトルのスケールの時間変動も大きい．一方で，楽器の音色については曲全体を通して比較的一定とみなせることが多い．

簡単化のため，$|a_{k,n,m}|$ と $v_{k,n,m}$ を調波インデックスに依存する成分と依存しない成分に以下のように分解する．

$$|a_{k,n,m}| = w_{k,n,m}\sqrt{U_{k,m}}, v_{k,n,m} = \tilde{w}_{k,n,m}\sqrt{U_{k,m}}. \qquad (13)$$

$w_{k,n,m}$ は音高 k の楽音の調波成分の相対的な振幅，$\tilde{w}_{k,n,m}$ は音高 k の励起信号の調波成分の相対的な複素振幅，$U_{k,m}$ は音高 k の楽音の時刻 t_m における正規化振幅と解釈できる．ただし，$U_{k,m}$ の正規化条件は $\sum_{k,m} U_{k,m} = 1$ とする．全極スペクトルモデル $1/|B_{k,m}(e^{\mathrm{j}\omega})|^2$ は音色に対応するので，$\beta_{k,m}[p]$ と $B_{k,m}(z)$ から時刻インデックス m を削除する．これにより式 (12) は

$$w_{k,n,m} = \left|\frac{\tilde{w}_{k,n,m}}{B_k(e^{\mathrm{j}ne^{\Omega_{k,m}}u_0})}\right| \qquad (14)$$

と書き直せる．

2.4 確率モデルとしての定式化

以上をまとめると，音高 k のパワースペクトログラムモデル $C_{k,l,m}$ は

$$C_{k,l,m} = H_{k,l,m}U_{k,m}, \qquad (15)$$

$$H_{k,l,m} = \sum_{n=1}^{N} w_{k,n,m}^2 e^{-\frac{(x_l - \Omega_{k,m} - \ln n)^2}{2\sigma^2}} \qquad (16)$$

と記述され，HTC や NMF と同様にパワースペクトログラムの加法性を仮定すれば，観測パワースペクトログラムモデル $X_{l,m}$ は

$$X_{l,m} = \sum_{k=1}^{K} C_{k,l,m} \qquad (17)$$

と表せる（図 1）．$X_{l,m}$ には，実際の音源信号の擬似周期性の仮定からの逸脱による誤差，調波間干渉を無視したことによる誤差，パワーの加法性に起因する誤差，背景雑音や残響に起因する誤差など，様々な要因による誤差が存在する．提案法では，一つ一つの誤差要因を詳細にモデル化する代わりに，まとめて一挙に確率的な現象と捉えることにする．ここで，$Y_{l,m}$ の確率分布が平均 $X_{l,m}$ の Poisson 分布から生成されたとすると，

$$Y_{l,m} \sim \mathrm{Pois}(Y_{l,m}; X_{l,m}) = \frac{X_{l,m}^{Y_{l,m}} e^{-X_{l,m}}}{\Gamma(Y_{l,m})} \qquad (18)$$

と記述できる．この生成モデルの尤度関数を $X_{l,m}$ に関し最大化する問題は，I ダイバージェンス基準における $X_{l,m}$ の $Y_{l,m}$ への最適フィッティング問題と等価である．

詳細は省略するが，[14] のように $\tilde{w}_{k,n,m}$ が平均 0，分散 ν^2 の等方的な複素正規分布に従うとすれば，式 (14) より $w_{k,n,m}$ は以下の Rayleigh 分布に従う．

$$\begin{aligned} w_{k,n,m} \sim & \mathrm{Rayleigh}\left(w_{k,n,m}; \frac{\nu}{|B_k(e^{\mathrm{j}ne^{\Omega_{k,m}}u_0})|}\right) \\ = & \frac{w_{k,n,m}}{(\nu/|B_k(e^{\mathrm{j}ne^{\Omega_{k,m}}u_0})|)^2} e^{-w_{k,n,m}^2/(2(\nu/|B_k(e^{\mathrm{j}ne^{\Omega_{k,m}}u_0})|)^2)}. \end{aligned} \qquad (19)$$

2.5 従来のスペクトログラムモデルとの関連

提案モデル $X_{l,m}$ は，様々な仮定を置くことで従来のスペクトログラムモデルと同一になる．式 (16) のように $H_{k,l,m}$

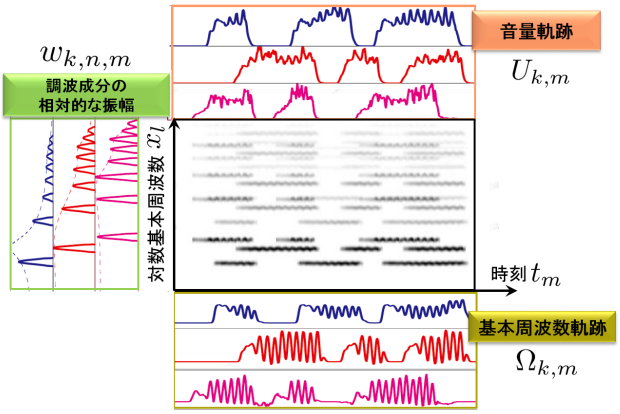


図 1 HTFD のスペクトログラムモデル．緑枠で囲まれた部分の点線が各音高に対するスペクトル包絡を表す．

をパラメトリックな関数で表さず，それ自体をパラメータとして扱えば，$X_{l,m}$ は可変基底 NMF [15] のスペクトログラムと一致する．さらに，$H_{k,l,m}$ を時不変として扱えば通常の NMF のスペクトログラムモデルと一致する．ここで，時不変にした $H_{k,l,m}$ が調波構造を持つことを仮定すれば調波 NMF [16, 17] のスペクトログラムモデルと同一である．一方で，式 (16) において $\Omega_{k,m}$ を時不変にすれば，[18] のスペクトログラムモデルと一致する．さらに，$U_{k,m}$ に拘束付きの混合正規分布モデルと同形の関数を仮定すれば，HTC における [2, 3] のスペクトログラムモデルと同一となる．

3. 音楽事前情報の組み込み

3.1 音楽事前情報の確率モデルへの導入方針

1 節で述べたように，前段処理で得られた音楽事前情報を，補助情報として活用できれば高精度な多重音解析を実現できるはずである．音楽事前情報には推定誤りを含むこともあるが，補助情報としての信頼度を確率と捉えれば，各パラメータの事前分布として推論に組み込める．

また，従来有効であることが知られている事前分布や複数の推定結果を同時に活用することもできる．例えば，調と和音の情報が得られたときには，その調で出現しやすくかつその和音で出現しやすい音高に，高い事前確率が割り当てられるべきである．この「かつ」に相当する演算は，両方の条件を表す確率分布の積で表現できる．そのため，複数の条件を同時に成立させるように事前分布を設計するには，各条件を表す確率分布の積をとり，それを正規化すればよい．この考え方は Products of Experts (PoE) [19] と呼ばれる．

3.2 音楽事前情報を反映した事前分布の設計例

具体例として，$\Omega_{k,m}$ と $U_{k,m}$ の事前分布を設計する．演奏中の各音符の F_0 を考えると，大域的にはその音符に対応する F_0 の近辺に存在する．一方で，特にバイオリンなどの弦楽器や管楽器による演奏では，ビブラートやポルタメントなどによって F_0 が局所的には時間に関して連続的に変化する傾向がある．この 2 つの性質は，$\boldsymbol{\Omega}$ の k 番目の行を転置した $\boldsymbol{\Omega}_k$ に対する確率分布 $q_{\mathrm{l}}(\boldsymbol{\Omega}_k), q_{\mathrm{g}}(\boldsymbol{\Omega}_k)$ として

$$q_{\mathrm{g}}(\boldsymbol{\Omega}_k) = \mathcal{N}(\boldsymbol{\Omega}_k; \mu_k \mathbf{1}_M, \xi_k^2 \boldsymbol{I}_M), \tag{20}$$

$$q_{\mathrm{l}}(\boldsymbol{\Omega}_k) = \mathcal{N}(\boldsymbol{\Omega}_k; \mathbf{0}_M, \tau_k^2 D^{-1}), \tag{21}$$

$$D = \begin{bmatrix} 1 & -1 & 0 & 0 & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & -1 & 2 & -1 \\ 0 & \dots & 0 & 0 & -1 & 1 \end{bmatrix}. \tag{22}$$

と記述できる．ここで，$\mathcal{N}(\boldsymbol{\Omega}_k; \boldsymbol{\mu}, \Sigma)$ は平均 $\boldsymbol{\mu}$，分散 Σ をもつ M 次元正規分布，$\mathbf{1}_M$ は全要素が 1 の M 次元のベクトル，$\mathbf{0}_M$ は M 次元の零ベクトル，$\boldsymbol{I}_M$ は $M \times M$ の単位行列を表す．この 2 つの確率分布を用いて $\boldsymbol{\Omega}_k$ の事前分布は

$$p(\boldsymbol{\Omega}_k) \propto q_{\mathrm{l}}(\boldsymbol{\Omega}_k)^{\alpha_{\mathrm{l}}} q_{\mathrm{g}}(\boldsymbol{\Omega}_k)^{\alpha_{\mathrm{gl}}} \tag{23}$$

と設計できる．ハイパーパラメータ $\alpha_{\mathrm{l}}, \alpha_{\mathrm{g}}$ により，事前分布に対する $q_{\mathrm{g}}(\boldsymbol{\Omega}_k), q_{\mathrm{g}}(\boldsymbol{\Omega}_k)$ の寄与を調節可能である．

ポピュラー音楽やクラシック音楽では調性があるため，曲中の調や和音によって音高の出現頻度に偏りがある．この偏りは，$U_{k,m}$ に対する確率分布として記述できる．また，時刻に関する $U_{k,m}$ のスパース制約が NMF において効果的であることが知られており，このスパース制約も確率分布として記述でき，音高の偏りを表す確率分布と統合的に扱える．これらの確率分布の導入を簡単にするため，$U_{k,m} = R_k A_{k,m}$ と分解する．$R_k := \sum_m U_{k,m}$ $(\sum_k R_k = 1)$ は音高方向に正規化された振幅であり，$A_{k,m} := U_{k,m}/R_k$ $(\sum_m A_{k,m} = 1)$ は時刻方向に正規化された振幅である．これらにより音高の出現頻度の偏りと時刻に関するスパース性を表す確率分布は，それぞれ $\boldsymbol{R} := [R_0, R_1, \dots, R_{K-1}]^\top, \boldsymbol{A}_k := [A_{k,0}, A_{k,1}, \dots, A_{k,M-1}]^\top$ の事前分布として以下のように記述できる．

$$\boldsymbol{R} \sim \mathrm{Dir}(\boldsymbol{R}; \boldsymbol{\gamma}^{(R)}), \quad \boldsymbol{A}_k \sim \mathrm{Dir}(\boldsymbol{A}_k; \boldsymbol{\gamma}_k^{(A)}) \tag{24}$$

ここで，$\boldsymbol{\gamma}^{(R)} := [\gamma_1^{(R)}, \dots, \gamma_K^{(R)}]^\top$ は $\boldsymbol{R}$ の事前分布のハイパーパラメータである．出現頻度の高い音高に対応する $\gamma_k^{(R)}$ に大きな値を与えれば，その音高が出現しやすいように $\boldsymbol{R}$ の推定値を誘導できる．$\boldsymbol{\gamma}_k^{(A)} := [\gamma_{k,1}^{(A)}, \dots, \gamma_{k,M}^{(A)}]^\top$ は $\boldsymbol{A}_k$ の事前分布のハイパーパラメータであり，この値を小さくすればよりスパースな $\boldsymbol{A}_k$ の推定値に誘導できる．

4. パラメータ推論アルゴリズム

ここまでで構築した確率モデルのもとで，与えられた観測パワースペクトログラム $\boldsymbol{Y}$ に対し，事後確率 $p(\Theta|\boldsymbol{Y}) \propto p(\boldsymbol{Y}|\Theta)p(\Theta)$ を最大化するような $\boldsymbol{w}, \Theta := \{\boldsymbol{\Omega}, \boldsymbol{R}, \boldsymbol{A}\}$ を求めたい．このとき最大化したい目的関数は，

$$\mathcal{J}(\Theta) := \ln p(\boldsymbol{Y}|\Theta) + \ln p(\Theta) \tag{25}$$

である．ここで，$\boldsymbol{w} := \{w_{k,n,m}\}_{k,n,m}$ の定義域 $\mathcal{W}$ を用いて，

$$\ln p(\boldsymbol{Y}|\Theta) = \ln \int_{\mathcal{W}} \prod_{l,m} \mathrm{Pois}(Y_{l,m}; X_{l,m}) \times \prod_{k,n,m} \mathrm{Rayleigh}\left(w_{k,n,m}; \frac{\nu}{|B_k(e^{\mathrm{j}n e^{\Omega_{k,m}} u_0})|}\right) \mathrm{d}\boldsymbol{w} \tag{26}$$

$$\ln p(\boldsymbol{\Theta}) = \sum_k \ln p(\boldsymbol{\Omega}_k) + \ln p(\boldsymbol{R}) + \sum_k \ln p(\boldsymbol{A}_k). \tag{27}$$

と書ける．

式 (26) の右辺に $\boldsymbol{w}$ に対する周辺化が入った対数関数があるため解析的に大域最適解を導くことは難しいが，補助関数法を用いることによって閉形式の更新則を導出できる．補助関数法は，目的関数値 $\mathcal{J}(\Theta)$ の上界となる関数（補助関数）を設計し，補助関数を補助変数と呼ぶ変数とパラメータ Θ に対して交互に最大化することにより，$\mathcal{J}(\Theta)$ を単調増加させる手法である．

対数関数は凹関数であるため，式 (26) に Jensen の不等式を適用すると，

$$\ln p(\boldsymbol{Y}|\Theta) \geq \int_{\mathcal{W}} q(\boldsymbol{w}) \left(\sum_{l,m} Y_{l,m} \ln \frac{X_{l,m}}{Y_{l,m}} - \sum_{l,m} X_{l,m} + \sum_{l,m} Y_{l,m} + \sum_{k,n,m} \ln \mathrm{Rayleigh}(w_{k,n,m}; \nu / B_k(e^{\mathrm{j}n e^{\Omega_{k,m}} u_0})|) - \ln q(\boldsymbol{w}) \right) \mathrm{d}\boldsymbol{w} \tag{28}$$

補助変数 $q(\boldsymbol{w})$ は $\int_{\mathcal{W}} q(\boldsymbol{w})\mathrm{d}\boldsymbol{w} = 1, q(\boldsymbol{w}) \geq 0$ を満たす．等号成立条件は $q(\boldsymbol{w})$ が $\boldsymbol{w}$ の事後分布と一致するときである．以下では，$\mathrm{E}_{q(\boldsymbol{w})}[\boldsymbol{w}^2] := \int_{\mathcal{W}} q(\boldsymbol{w})\boldsymbol{w}^2\mathrm{d}\boldsymbol{w}$ と表記する．

式 (17) の $X_{l,m}$ は k, n に関する和を含むため，式 (28) の括弧内の第 1 項は対数関数の中に和を含むため，解析的に解くのが困難である．そこでさらに，式 (28) の括弧内の第 1 項に Jensen の不等式を適用すると，

$$Y_{l,m} \ln X_{l,m} \geq Y_{l,m} \sum_{k,n} \lambda_{k,n,l,m} \ln \frac{w_{k,n,m}^2 e^{-\frac{(x_l - \Omega_{k,m} - \ln n)^2}{2\sigma^2}} U_{k,m}}{\lambda_{k,n,l,m}}, \tag{29}$$

として下界が得られる．補助変数 $\boldsymbol{\lambda} := \{\lambda_{k,n,l,m}\}_{k,n,l,m}$ は $\lambda_{k,n,l,m} \geq 0, \sum_{k,n} \lambda_{k,n,l,m} = 1$ を満たし，等号成立条件は

$$\lambda_{k,n,l,m} = \frac{w_{k,n,m}^2 e^{-\frac{(x_l - \Omega_{k,m} - \ln n)^2}{2\sigma^2}} U_{k,m}}{X_{l,m}} \tag{30}$$

である．したがって，$\mathcal{J}(\Theta)$ の補助関数 $\mathcal{J}^+(\boldsymbol{\lambda}, q(\boldsymbol{w}), \Theta)$ は

$$\mathcal{J}^+(\boldsymbol{\lambda}, q(\boldsymbol{w}), \Theta) \underset{c}{=} E_{q(\boldsymbol{w})} \left[\sum_{l,m} Y_{l,m} \sum_{k,n} \lambda_{k,n,l,m} \ln \frac{w_{k,n,m}^2 e^{-\frac{(x_l - \Omega_{k,m} - \ln n)^2}{2\sigma^2}} U_{k,m}}{\lambda_{k,n,l,m}} - \sum_{l,m} X_{l,m} + \ln \frac{p(\boldsymbol{w}|\boldsymbol{\beta}, \boldsymbol{\Omega})}{q(\boldsymbol{w})} \right] + \ln p(\Theta). \tag{31}$$

と書ける．ただし，$=_c$ は定数を除いて一致することを示す．さらに，式 (31) の括弧内の第 2 項に対して，x に関する区分求積法による近似を用い，$x < x_0$ と $x_{L-1} < x$ の範囲の値が無視できるものとすると，

$$\sum_l X_{l,m} \simeq \frac{1}{\Delta_x} \int_{-\infty}^{\infty} X(x, t_m)\mathrm{d}x = \frac{\sqrt{2\pi}\sigma}{\Delta_x} \sum_k R_k A_{k,m} \sum_n w_{k,n,m}^2. \tag{32}$$

が得られる．この近似は，式 (31) の括弧内の第 2 項が $\boldsymbol{\Omega}_{k,m}$ にほとんど依存しないことを仮定することに相当する．

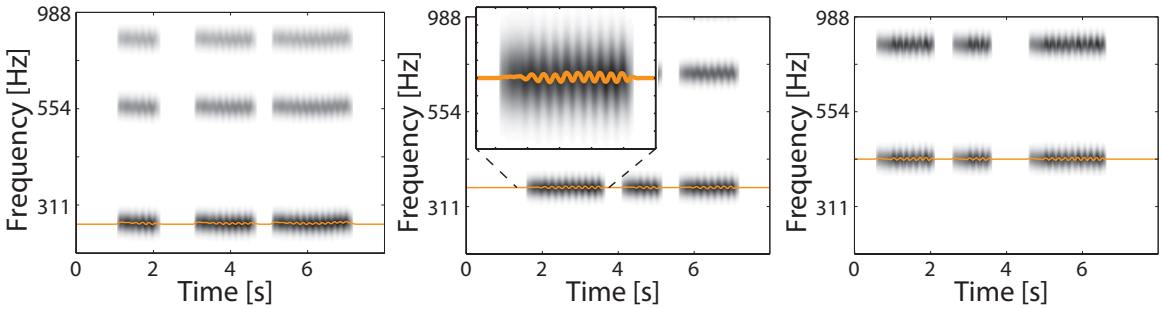
この近似を適用した $\mathcal{J}^+(\boldsymbol{\lambda}, q(\boldsymbol{w}), \Theta)$ を $\mathcal{J}^{++}(\boldsymbol{\lambda}, q(\boldsymbol{w}), \Theta)$ と置き，これを用いて補助変数とパラメータに関する更新式を導出する．補助変数の更新式は等号成立条件であり，モデルパラメータ Θ の更新式はそれぞれ $\mathcal{J}^{++}(\boldsymbol{\lambda}, q(\boldsymbol{w}), \Theta)$ に関する各パラメータでの偏微分が 0 となる値を求めることにより導出できる．詳しい導出や更新式は本稿では省略する（詳細は [7] 参照）．

5. 多重音解析の動作確認実験と定量評価実験

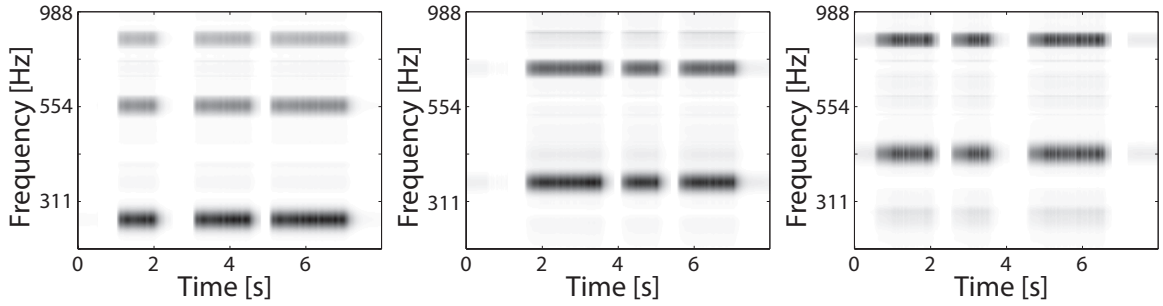
5.1 時間変化する F_0 の推定実験

HTFD が時間的に変化する F_0 を推定できるかを確認するため，RWC 楽器音データベース [20] から D♭4，F4，A♭4 のバイオリン音源のビブラート音を用いて，人工的に合成したサンプリング周波数 16 kHz の音響信号を分離する実験を行った．簡単のため，本節と次節の実験では全極スペクトルモデルを省いて $w_{k,m,n}$ の時不変性を仮定した HTFD [6] を用いた．スペクトログラムを求める際には，時間シフト間隔を 14.6 ms，解析周波数を 55 Hz から 7040 Hz まで 10 cent 間隔に設定し，高速近似ウェーブレット変換 [21] を行った．アナライジングウェーブレットとして，式 (5) で定義された対数正規分布型のウェーブレットを用いた（$\sigma = 0.02$）．調波成分の個数を $N = 8$，音高数を $K = 73$ とし，μ_k として A1（55 Hz）から A♯7 までの基本周波数にそれぞれ対応させた．他のパラメータは，全ての k に対し $\boldsymbol{\gamma}_k^{(A)} = (1 - 3.96 \times 10^{-6})\mathbf{1}_I$, $\tau_k = 0.83$, $\nu_k = 1.25$, $\alpha_g = \alpha_s = 1$, $\boldsymbol{\gamma}^{(R)} = (1 - 2.4 \times 10^{-3})\mathbf{1}_K$ とした．

図 2 に推定されたスペクトログラムモデルを示す．F_0 の時間変化を記述できない NMF（基底数 3，I ダイバージェンス基準）では，ビブラート時の F_0 の時間変化が平均化されてしまっていることが確認できる（図 2 (a)）．一方で，HTFD により得られたスペクトログラムモデルでは，図

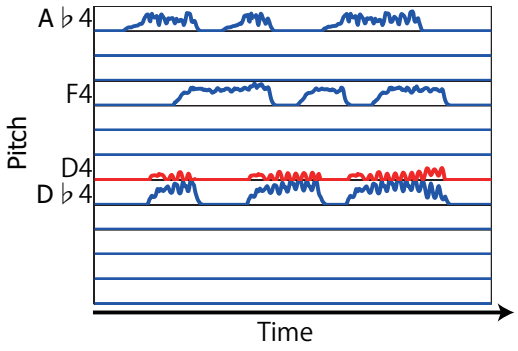


(a) HTFD により推定されたスペクトログラムモデルと F_0 軌跡（橙色の曲線）.

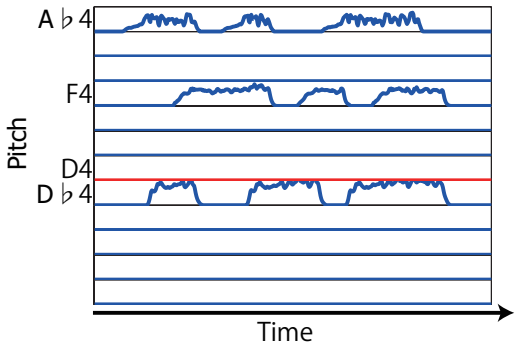


(b) NMF により推定されたスペクトログラムモデル

図 2 HTFD [6] と NMF により推定されたスペクトログラムモデル（ [8] より抜粋）. 左から順に，それぞれ D♭4，F4，A♭4 に対応.



(a) 調の情報なし



(b) 調の情報あり

図 3 調の情報を用いない場合 (a) と用いた場合 (b) の HTFD により推定された A3 から A♭4 に対応するアクティベーション（ [8] より抜粋）. 赤色の曲線は D4 のアクティベーションに対応する.

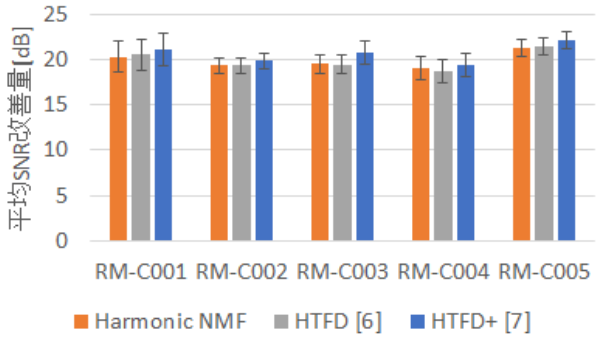


図 4 F_0 が時間変化しない拘束をおいた HTFD（Harmonic NMF）と，全極スペクトルモデルを導入していない HTFD（HTFD）[6]，全極スペクトルモデルを導入した HTFD（HTFD+）[7] で得られた各楽曲に対する平均 SNR 改善量と標準誤差.

2 (b) のように適切に F_0 軌跡が推定された.

5.2 音楽事前情報の効果確認実験

次に，調の情報を用いて事前分布を設計し音階外の音高のアクティベーションを抑制できるかを確認する．この音響信号の調を変ニ長調と仮定し，音階内の音高に対応する $\gamma_k^{(R)}$ は $1 - 2.4 \times 10^{-3}$，音階外の音高に対応する $\gamma_k^{(R)}$ は $1 - 3.0 \times 10^{-3}$ とした．他の条件は前節と同一である.

調の情報を用いない場合（図 3 (a)）には，実際に音響信号に含まれていないにも関わらず，D4 に対応するアクティベーションが高くなってしまっている．この場合に比べ，調の情報を用いた上述の事前分布を用いると，D4 は音階外であるため，この音高に対応するアクティベーションが抑制された．したがって，$\gamma_k^{(R)}$ に調の情報を反映し多重音解析性能を改善できることが確認できた.

5.3 モノラル音源分離性能の定量評価実験

最後に，全極スペクトルモデルの効果を確認するためモノラル音楽音響信号の各音高への分離実験を行った．比較手法として，全極スペクトルモデルを用いない HTFD [6] と，さらに F_0 を時不変とした HTFD （スペクトルテンプレートを，調波的に正規分布形の関数が並んだ混合正規分布モデルと同形の関数に拘束した場合の調波 NMF [16, 17] に対応．以後，Harmonic NMF）を用いた．分離では時間周波数マスクを $C_{k,l,m}/\sum_k C_{k,l,m}$ と設計して用いた．ただし，提案法では $\boldsymbol{w}$ の代わりに事後分布 $q(\boldsymbol{w})$ が推定されるため，$\boldsymbol{w}^2$ の推定値を $\mathrm{E}_{q(\boldsymbol{w})}[\boldsymbol{w}^2]$ として計算した．各音高ごとの演奏の録音を用意するのは困難であったため，RWC クラシック音楽データベース [20] の RM-C001 から RM-C005 の最初の 30 秒を MIDI シンセサイザー FluidSynth [22] で合成した音響信号（サンプリング周波数 16 kHz）を入力として用いた．スペクトログラムの計算には，時間シフトを 14.6 ms とした高速近似ウェーブレット変換を用いた．調波成分の個数は $N = 20$ とし，全ての k に対して $\tau_k = 1.0$, $\boldsymbol{\gamma}_k^{(A)} = (1 - 1.0 \times 10^{-4})\mathbf{1}_I$, $\boldsymbol{\gamma}^{(R)} = 0.8 \times \mathbf{1}_K$, $P = 20$, $\nu = 1$ とした．他のパラメータやアナライジングウェーブレットは 5.1 節と同一とした．また，Harmonic NMF は 100 イテレーション，HTFD と HTFD+は 20 イテレーション時点での推定値を用いた．

楽曲毎の分離音の signal-to-noise ratio（SNR）改善量の平均と標準誤差を図 4 に示す．Harmonic NMF と HTFD を比べると SNR の差の平均値は約 0.02 dB であり，分離性能に大きな差は見られなかったが，MIDI で音源を作成したため F_0 の変動が少なかったからと考えられる．一方，HTFD+は全曲で Harmonic NMF と HTFD よりも SNR が約 0.80 dB，約 0.78 dB 平均的に改善した．この結果より，全極スペクトルモデルの導入によってモノラル音源分離性能が向上することを確認できた．

6. 結論

本報告では，NMF と HTC の利点を兼ね備えた新たなスペクトログラムモデルを提案し，そのモデルに基づく多重音解析手法 HTFD について述べた．各音高の音響信号の連続時間におけるモデルとして擬似周期信号から出発し，そのウェーブレット変換を近似的に導出した．また，離散時間信号領域で定義された全極システムをウェーブレット変換で定義されたスペクトルモデルに組み込んだ．また，音楽事前情報を事前分布として導入できることを示し，複数の音楽事前情報を用いる場合や従来用いられてきた事前分布と併用するときにも，PoE を用いて事前分布を適切に設計する方法を示した．提案した確率モデルに対して，閉形式の更新式からなるパラメータ推論アルゴリズムを導いた．モノラル音源分離実験により提案法の有効性を確認した．

謝辞 本稿は，四方紘太郎氏，高宗典玄氏らと共に著者らが行った研究の成果をまとめたものであり，実装や実験，議論を通して本研究に貢献していただいた両氏に感謝する．本研究は JSPS 科研費 26730100 の助成を受けたものである．

参考文献

[1] Hu, G. and Wang, D. L.: An auditory scene analysis approach to monaural speech segregation, *Topics in Acoust. Echo and Noise Contr.*, pp. 485–515 (2006).
[2] Kameoka, H., Nishimoto, T. and Sagayama, S.: A Multipitch Analyzer Based on Harmonic Temporal Structured Clustering, *IEEE Trans. Acoust., Speech, and Language Process.*, Vol. 15, No. 3, pp. 982–994 (2007).
[3] Kameoka, H.: Statistical Approach to Multipitch Analysis, PhD Thesis, The University of Tokyo (2007).
[4] Bregman, A. S.: *Auditory scene analysis: The perceptual organization of sound*, MIT press (1994).
[5] Smaragdis, P. and Brown, J. C.: Non-negative matrix factorization for polyphonic music transcription, *Proc. IEEE Workshop Applications Signal Process. Audio Acoust.*, IEEE, pp. 177–180 (2003).
[6] 四方紘太郎，高宗典弘，中村友彦，亀岡弘和：調波時間因子分解法に基づく事前情報付き多重音解析，情処研報，No. 39 (2014).
[7] 中村友彦，亀岡弘和：全極スペクトルモデルを用いた調波時間因子分解による多重音解析，情処研報，No. 26 (2015).
[8] Nakamura, T., Shikata, K., Takamune, N. and Kameoka, H.: Harmonic-Temporal Factor Decomposition Incorporating Music Prior Information for Informed Monaural Source Separation, *Proc. Int. Symposium Music Info. Retrieval*, pp. 623–628 (2014).
[9] [Online: 18, Apr. 2015], `http://www.music-ir.org/mirex/wiki/MIREX_HOME`.
[10] Smaragdis, P. and Mysore, G. J.: Separation by ”humming”: User-guided sound extraction from monophonic mixtures, *Proc. IEEE Workshop Applications Signal Process. Audio Acoust.*, IEEE, pp. 69–72 (2009).
[11] Ozerov, A., Févotte, C., Blouet, R. and Durrieu, J. L.: Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation, *Proc. Int. Conf. Acoust. Speech Signal Process.*, IEEE, pp. 257–260 (2011).
[12] Hennequin, R., David, B. and Badeau, R.: Score informed audio source separation using a parametric model of nonnegative spectrogram, *Proc. Int. Conf. Acoust. Speech Signal Process.*, pp. 45–48 (2011).
[13] Simsekli, U. and Cemgil, A. T.: Score guided musical source separation using generalized coupled tensor factorization, *Proc. Eur. Signal Process. Conf.*, IEEE, pp. 2639–2643 (2012).
[14] 亀岡弘和：全極型声道モデルと F_0 パターン生成過程モデルを内部にもつ統計的音声スペクトルモデル，信学技報，Vol. SP2010-74, pp. 29–34 (2010).
[15] Nakano, M., Le Roux, J., Kameoka, H., Ono, N. and Sagayama, S.: Infinite-state spectrum model for music signal analysis, *Proc. Int. Conf. Acoust. Speech Signal Process.*, pp. 1972–1975 (2011).
[16] Raczyński, S. A., Ono, N. and Sagayama, S.: Multipitch analysis with harmonic nonnegative matrix approximation, *Proc. Int. Conf. Music Info. Retrieval*, pp. 381–386 (2007).
[17] Vincent, E., Bertin, N. and Badeau, R.: Harmonic and inharmonic Nonnegative Matrix Factorization for Polyphonic Pitch transcription, *Proc. Int. Conf. Acoust. Speech Signal*

Process., pp. 109–112 (2008).
[18] Yoshii, K. and Goto, M.: Infinite Latent Harmonic Allocation: A Nonparametric Bayesian Approach to Multipitch Analysis, *Proc. Int. Soc. Music Info. Retrieval*, pp. 309–314 (2010).
[19] Hinton, G. E.: Training products of experts by minimizing contrastive divergence, *Neural Comput.*, Vol. 14, No. 8, pp. 1771–1800 (2002).
[20] Goto, M.: Development of the RWC Music Database, *Proc. Int. Congress Acoust.*, pp. l–553–556 (2004).
[21] 亀岡弘和，田原鉄也，西本卓也，嵯峨山茂樹：信号処理方法及び装置. 特開 2008-281898, (20. Nov. 2008).
[22] [Online: 21, Apr. 2015], `http://www.fluidsynth.org/`.