Vol.2015-MUS-106 No.26
Vol.2015-EC-35 No.26
2015/3/3

# 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

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いる．我々は，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を以前提案した．これらに加え，ソース・フィルタモデルに代表される楽音の物理的な生成過程も，多重音解析の性能向上には重要である．そこで本報告では，離散時間信号領域で定義されたソース・フィルタモデルが調波時間因子分解の音源スペクトログラムモデルに導入できることを示し，そのパラメータ推論法について述べる．定量評価実験により，こ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示す．

## 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，パワーなど）を得る処理であり，音楽情報検索や自動採譜，音響信号の自動編集など様々なアプリケーションの基礎技術である．

多重音解析を行う際に，観測信号が多チャンネル信号であれば音源の空間的な手がかりを利用することができるが，モノラル信号であるときはこの手がかりに代わる何らかの仮定が必要となる．モノラル信号を入力とする多重音解析では，主に以下の 2 つのアプローチが取られてきた．1 つ目のアプローチは，計算論的聴覚情景分析のコンセプトに基づくものである（例，[1]）．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 [2] で示された聴覚機能を計算機で実現しようという試みである．このアプローチに則った手法である調波時間構造化クラスタリング（Harmonic-temporal clustering, HTC）[3,4] は，同一の音源に由来する時間周波数成分が音脈と呼ばれる知覚的な構造に群化されるプロセスを模倣したものである．HTC では，群化の要件（調波性，連続性，同時性，同期性など）が時間周波数領域の局所的な制約として記述されており，当該制約を満たすように観測信号の時間周波数成分を時間周波数平面上でクラスタリングする．

一方，2 つ目のアプローチの代表的な手法は，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）を適用するものである [5]．この手法は，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眼し，各フレームのスペクトルが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表現できるという仮定を用いている．スペクトルテンプレートと重みは非負値であるため観測信号は 2 つの非負値行列の積として記述される．したがって，観測振幅スペクトログラムをこの 2 つの行列の積に分解することによって，スペクトルテンプレートと重みを同時推定し，観測振幅スペクトログラムを音高毎のスペクトログラムに分離できる．

これら 2 つのアプローチは異なる手がかりを元に分離を行っている．前者のアプローチは時間周波数表現の局所的な性質，後者は音楽信号特有の大域的な性質を利用しており，これらは相容れない関係ではなくいずれも高精度な多重音解析を実現するためには有用である．この考えに基づき，我々はこれら 2 つの性質を同時に取り入れたスペクトログラムモデルと，それに基づく多重音解析手法である調波時間因子分解（Harmonic-temporal factor decomposition, HTFD）を提案した [6]．

HTFD の多重音解析性能をさらに向上させるためには，音源の物理的な生成過程に基づく手がかりが有用である．管楽器や弦楽器から発せられた各音高の楽音は，管楽器に

[1] 東京大学 大学院情報理工学研究科, 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

吹き込んだ呼気や弦の振動が楽器内部で共鳴して生成される．前者は主に音高に対応し，後者は音色に対応することが知られている．音高に対応する物理量である基本周波数（$F_0$）はビブラートやポルタメントなどで時間的に変動することも多いが，音色は比較的時不変とみなせる．したがって，これらの性質を導入した音源スペクトログラムモデルが適切に構築できれば，実際の楽器では起こりえないスペクトルの時間変動が抑制できるはずである．

本報告では，HTFD の多重音解析性能向上のため，上述の音源の物理的な生成過程に基づく手がかりを離散時間領域で定義されたソース・フィルタモデルを用いて記述し，ある特定のウェーブレット変換領域で定義された HTFD のスペクトログラムモデルに導入する．また，構築したスペクトログラムモデルの効率的なパラメータ推論法を提案する．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

### 2.1 楽音信号のウェーブレット変換

本節では，[6] と同様に [4] に倣って楽音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（特に調波楽器音）は，擬似周期信号（局所的に周期的とみなせ，周期や調波成分のパワーが時間的に滑らかに変化する信号）とみなせる．ここで，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 $a_{k,n}(u) \in \mathbb{C}$ の擬似周期信号の解析信号表現

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

として与える．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．この信号表現では，群化の要件のうち調波性と周波数変調の同期性が暗に満たされている．ここで，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を $\psi(u) \in \mathbb{C}$ と書くと，ウェーブレット基底関数 $\psi_{\alpha,t}(u)$ は

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

とかける．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である．このウェーブレット基底関数の複素共役 $\psi^*_{\alpha,t}(u)$ を用いて $f_k(u)$ のウェーブレット変換は，

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

と得られる．典型的に $\psi^*_{\alpha,t}(u)$ の優勢な部分が $t$ の周りに局在化しているので，式 (3) の積分結果は $t$ 周りの $\theta_k(u)$ と $a_{k,n}(u)$ の値のみに依存する．そのため，以下のように $\theta_k(t)$ と $a_{k,n}(t)$ を 1 次の Taylor 展開で近似する．

$$a_{k,n}(u) \simeq a_{k,n}(t), \quad \theta_k(u) \simeq \theta_k(t) + \dot{\theta}_k(t)(u-t). \tag{4}$$

ここで，$\dot{\theta}_k(u)$ は瞬時 $F_0$ である．上述の近似を用いつつ 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，式 (3) を

$$W_k(x,t) = \sum_{n=1}^{N} a_{k,n}(t) \Psi^*(ne^{-x+\Omega_k(t)}) e^{\mathrm{j}(n\theta_k(t)+\varphi_{k,n})}, \tag{5}$$

と変形できる．$\Psi$ は $\psi$ の Fourier 変換であり任意に選ぶことができるので，以下の $\omega = 1$ で最大値をとる対数正規分布型の実関数を用いる．

$$\Psi(\omega) = \begin{cases} e^{-\frac{(\ln \omega)^2}{4\sigma^2}} & (\omega > 0) \\ 0 & (\omega \le 0) \end{cases}. \tag{6}$$

$\sigma$ は $\Psi(\omega)$ を $\ln \omega$ 軸で見たときの標準偏差に対応する．これを用いて，式 (5) は

$$W_k(x,t) = \sum_{n=1}^{N} a_{k,n}(t) e^{-\frac{(x-\Omega_k(t)-\ln n)^2}{4\sigma^2}} e^{\mathrm{j}(n\theta_k(t)+\varphi_{k,n})}. \tag{7}$$

と変形できる．ここで，$n, n' (n \neq n')$ の指数項の重なりがほとんどない，すなわち調波性分が互いに重ならないと仮定できれば，$|W_k(x,t)|^2$ は近似的に

$$|W_k(x,t)|^2 \simeq \sum_{n=1}^{N} |a_{k,n}(t)|^2 e^{-\frac{(x-\Omega_k(t)-\ln n)^2}{2\sigma^2}} \tag{8}$$

と書ける．この仮定は調波成分のパワースペクトルが加法的であると近似していることに相当する．式 (8) で与えられるスペクトログラムモデルは HTC で用いられた調波時間構造化モデル [4] 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．さらに，パワースペクトルの加法性を仮定すれば，$K$ 個の楽音が重畳された音響信号のパワースペクトログラムは，式 (8) の $k$ について和をとったものとして与えられる．

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，我々が計算機によって実際に得ることのできる観測スペクトログラムは離散的な表現である．そのため本節以降は，等間隔に量子化された時間 $t_m$ $(m = 0, 1, \ldots, M-1)$ 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 $Y_{l,m} := Y(x_l, t_m)$ を表す．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

1 節で述べたように，楽音の物理的な生成過程を適切に記述できれば，より詳細なスペクトログラムモデルが得られ多重音解析性能が向上するはずである．楽音の物理的な

Vol.2015-MUS-106 No.26
Vol.2015-EC-35 No.26
2015/3/3

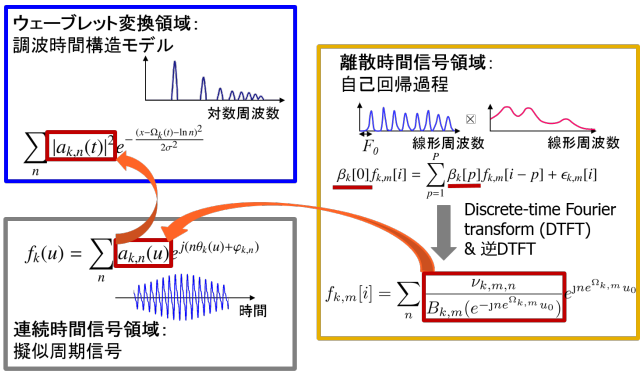


**図 1** 自己回帰過程のパラメータと調波時間構造化モデルのパラメータの対応関係.

生成過程はソース・フィルタモデルでよく表現でき，このモデルは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 (8) 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，[7] に従ってこれらのパラメータの関係を式 (1) で定義された擬似周期信号の解析的な表現を介して得ることを目指す（図 1）．

時刻 $t_m$ における式 (8) 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．$i$ は離散時間インデックスである．この $f_{k,m}[i]$ が $P$ 次の自己回帰過程によって

$$\beta_{k,m}[0] f_{k,m}[i] = \sum_{p=1}^{P} \beta_{k,m}[p] f_{k,m}[i-p] + \epsilon_{k,m}[i], \qquad (9)$$

と記述されると考えよう．$\beta_{k,m}[p]$ $(p = 0, 1, \ldots, P)$ 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 $\beta_{k,m}[p]$ をパラメータとして持つ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 $e^{\Omega_{k,m}}$ であるため，励起信号 $\epsilon_{k,m}[i]$ の $F_0$ も $e^{\Omega_{k,m}}$ である必要がある．したがって，$\epsilon_{k,m}[i]$ は

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

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（観測信号のサンプリング周期と一致）であり，$v_{k,n,m} \in \mathbb{C}$ は $n$ 番目の調波成分の複素振幅である．式 (9) に対して離散時間 Fourier 変換（discrete-time Fourier transform, DTFT）を適用すると，$f_{k,m}$ の DTFT は

$$F_{k,m}(\omega) = \frac{\sqrt{2\pi}}{B_{k,m}(e^{\mathrm{j}\omega})} \sum_{n=1}^{N} v_{k,n,m} \delta(\omega - n e^{\Omega_{k,m}} u_0), \qquad (11)$$

$$B_{k,m}(z) := \sum_{p=0}^{P} \beta_{k,m}[p] z^{-p} \qquad (12)$$

として得られる．ここで，$\omega$ は正規化角周波数，$\delta(\omega)$ は Dirac のデルタ関数である．式 (11) に逆 DTFT を適用すると，

$$f_{k,m}[i] = \sum_{n=1}^{N} \frac{v_{k,n,m}}{B_{k,m}(e^{\mathrm{j} n e^{\Omega_{k,m}} u_0})} e^{\mathrm{j} n e^{\Omega_{k,m}} i u_0}. \qquad (13)$$

を得ることができ，これは $f_{k,m}[i]$ の別表現となっている（図 1 の灰色矢印）．式 (13) と式 (1) の離散時間表現

$$f_{k,m}[i] = \sum_{n=1}^{N} a_{k,n,m} e^{\mathrm{j}(n e^{\Omega_{k,m}} i u_0 + \varphi_{k,n})} \qquad (14)$$

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（図 1 の橙色矢印）を以下のように陽に記述することができる．

$$|a_{k,n,m}| = \left| \frac{v_{k,n,m}}{B_{k,m}(e^{\mathrm{j} n e^{\Omega_{k,m}} u_0})} \right|. \qquad (15)$$

### 2.3 モデルパラメータに対する拘束

前節では全極システムをスペクトログラムモデルに導入したが，そのパラメータの値によっては楽音の性質とはかけ離れたものも表しうる．そのため，モデルパラメータに適切な拘束を置く必要がある．NMF の重要な仮定は，各音高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したことにある．すなわち，どのスペクトルの構成要素を時変とみなすか時不変とみなすかがこの仮定の本質であり，ここではそれに従って考えてみよう．$F_0$ はビブラートやポルタメント中は時間的に変動すると仮定されるべきであるし，スペクトルのパワーについても同様である．（NMF もパワーは時変としている．）しかし，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．

これらの仮定は以下のようにスペクトログラムモデルに反映させられる．導出を簡単にするため，$|a_{k,n,m}|$ を以下のように 2 つの変数の積に分解する．

$$|a_{k,n,m}| = w_{k,n,m} \sqrt{U_{k,m}}. \qquad (16)$$

$w_{k,n,m}$ は音高 $k$ の楽音の時刻 $t_m$ における $n$ 番目の調波成分の相対的なパワー，$U_{k,m}$ 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 $\sum_{k,m} U_{k,m} = 1$ である．同様に，$v_{k,n,m}$ についても

$$v_{k,n,m} = \tilde{w}_{k,n,m} \sqrt{U_{k,m}}. \qquad (17)$$

と分解できる．全極スペクトル $1/|B_{k,m}(e^{\mathrm{j}\omega})|^2$ は音高 $k$ の楽音の音色に対応するので，時不変なものとして扱いたい．そのためには単純に $\beta_{k,m}[p]$ と $B_{k,m}(z)$ から時刻インデックス $m$ を削除すればよく，式 (15) は

$$w_{k,n,m} = \left| \frac{\tilde{w}_{k,n,m}}{B_k(e^{\mathrm{j} n e^{\Omega_{k,m}} u_0})} \right| \qquad (18)$$

と書き換えられる．HTFD では，物理的な生成過程を導入していなかったために全極スペクトルの時不変性を用いる

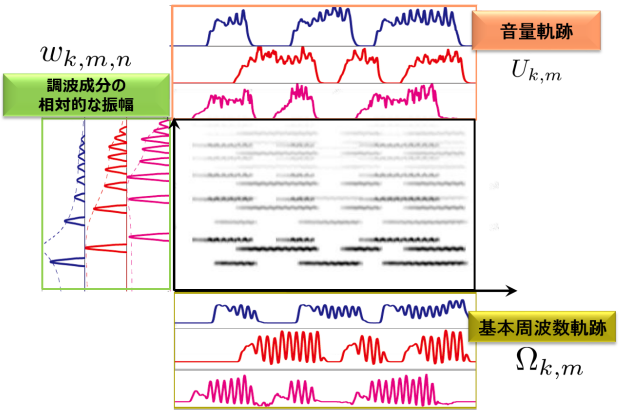


**図 2** 全極スペクトルモデルを導入したスペクトログラムモデル．緑枠で囲まれた部分の点線が各音高に対するスペクトル包絡を表す．

ことができず，解を制限するために $w_{k,n,m}$ を時不変なパラメータとして扱うことと等価な処理を行わざるをえなかった．一方で，提案モデルでは $w_{k,n,m}$ が時変なパラメータとして扱えるのが主要な違いであり，これにより励起信号の複素振幅の変化に伴うスペクトル形状の変化を記述できる．

まとめると，音高 $k$ のパワースペクトログラムモデル $C_{k,l,m}$ は音量軌跡 $U_{k,m}$，調波成分の相対振幅 $w_{k,m,n}$，$F_0$ 軌跡 $\Omega_{k,m}$ によって

$$C_{k,l,m} = \left( \sum_{n=1}^{N} w_{k,n,m}^2 e^{-\frac{(x_l - \Omega_{k,m} - \ln n)^2}{2\sigma^2}} \right) U_{k,m}, \tag{19}$$

と記述され，観測信号のパワースペクトログラムモデル $X_{l,m}$ はそれらが重畳されたものとして以下のように表される（図 2）．

$$X_{l,m} = \sum_{k=1}^{K} C_{k,l,m}. \tag{20}$$

提案モデル $X_{l,m}$ は様々な仮定を置くことで従来のモデルと同一になる．$w_{k,n,m}$ を時不変にしたものを曲全体の音量を表す $V > 0$ と $n$ 番目の調波成分の相対的なパワーを正規化した $v_{k,n}$ ($\sum_n v_{k,n} = 1$) の積に分解すれば，HTFD のスペクトログラムモデルと同一となる．また，式 (20) の括弧内の項を $H_{k,l,m}$ とおけば $X_{l,m} = \sum_k H_{k,l,m} U_{k,m}$ と変形できるため，HTFD の場合と同様に NMF の時不変なスペクトルテンプレートが時変に拡張されたともみなせる．他にも HTFD と同じく様々な NMF の拡張や HTC との関連もあるが，既に [6] で記述しているため省略する．時間周波数表現として短時間フーリエ変換を用いた場合に，$H_{k,l,m}$ を時不変にした後励起信号のパワースペクトルを周波数点ごとにパラメータとみなせば複合自己回帰系 [8] と一致し，全極スペクトルモデル自体も周波数点ごとにパラメータとみなせば [9] と一致する．

### 2.4 確率モデルとしての定式化

本節では，HTFD と同様に式 (20) で定義された $X_{l,m}$ を確率モデルとして定式化する．これまでに用いた仮定や近似は現実には常に正確に成り立つとは限らないため，観測パワースペクトログラム $Y_{l,m}$ は最適なパラメータが得られたとしても $X_{l,m}$ と一致しないかもしれない．一つ一つの誤差要因を詳細にモデル化する代わりに，確率的生成モデルによってこの逸脱を確率的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，

$$Y_{l,m} \sim \mathrm{Pois}(Y_{l,m}; X_{l,m}) = \frac{X_{l,m}^{Y_{l,m}} e^{-X_{l,m}}}{\Gamma(Y_{l,m})} \tag{21}$$

と記述でき，この尤度関数は

$$\prod_{l,m} \mathrm{Pois}(Y_{l,m}; X_{l,m}), \tag{22}$$

と書ける．ただし，$\boldsymbol{Y} := (Y_{l,m})_{l,m}$ は観測パワースペクトログラムの行列表現である．式 (22) の $X_{l,m}$ に関する最大化は，I ダイバージェンス基準における $X_{l,m}$ の $Y_{l,m}$ への最適なフィッティングと等価である．

式 (18) から，条件付き確率分布 $p(\boldsymbol{w}|\tilde{\boldsymbol{w}}, \boldsymbol{\beta}, \boldsymbol{\Omega})$ は Dirac のデルタ関数を用いて

$$p(\boldsymbol{w}|\tilde{\boldsymbol{w}}, \boldsymbol{\beta}, \boldsymbol{\Omega}) = \prod_{k,n,m} \delta\left( w_{k,n,m} - \left| \frac{\tilde{w}_{k,n,m}}{B_k(e^{\mathrm{j}n e^{\Omega_{k,m}} u_0})} \right| \right) \tag{23}$$

と書ける．ただし，$\boldsymbol{w} := \{w_{k,n,m}\}_{k,n,m}$, $\tilde{\boldsymbol{w}} = \{\tilde{w}_{k,n,m}\}_{k,n,m}$ であり，$\boldsymbol{\beta}$ は $(k,p)$ 番目の要素が $\beta_k[p]$ の行列，$\boldsymbol{\Omega}$ は $(k,m)$ 番目の要素が $\Omega_{k,m}$ の行列を表す．これを元に条件付き確率分布 $p(\boldsymbol{w}|\boldsymbol{\beta}, \boldsymbol{\Omega})$ を導出するには，$p(\tilde{\boldsymbol{w}})$ を定義し $\tilde{\boldsymbol{w}}$ について周辺化すればよい．[7] のように，$\tilde{w}_{k,n,m}$ が等方的な複素正規分布

$$\tilde{w}_{k,n,m} \sim \mathcal{N}_{\mathbb{C}}(\tilde{w}_{k,n,m}; 0, \nu^2) = \frac{e^{-|\tilde{w}_{k,n,m}|^2/\nu^2}}{\pi \nu^2} \tag{24}$$

から生成されると仮定すると，$w_{k,n,m}$ は以下の Rayleigh 分布に従う．

$$\begin{aligned} w_{k,n,m} &\sim \mathrm{Rayleigh}\left( w_{k,n,m}; \frac{\nu}{|B_k(e^{\mathrm{j}n e^{\Omega_{k,m}} u_0})|} \right) \\ &= \frac{w_{k,n,m}}{(\nu/|B_k(e^{\mathrm{j}n e^{\Omega_{k,m}} u_0})|)^2} e^{-w_{k,n,m}^2/(2(\nu/|B_k(e^{\mathrm{j}n e^{\Omega_{k,m}} u_0})|)^2)}. \end{aligned} \tag{25}$$

これは全極スペクトルの効果が事前分布として確率モデルに導入できることを示している．

他のパラメータである $\boldsymbol{\Omega}$ や $U$ の事前分布については HTFD と同一のものを用いる．$F_0$ に関しては 2 つの物理的な性質を用いて事前分布を設計している．1 つ目の性質はビブラートやポルタメント中には $F_0$ が時間に関して連続的に変化しやすいことであり，2 つ目の性質は音高 $k$ の楽音はそれに対応する対数周波数 $\mu_k$ の周りに $F_0$ が分布しやすいことである．詳しい導出は省くが，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 $q_{\text{local}}(\boldsymbol{\Omega}_k), q_{\text{global}}(\boldsymbol{\Omega}_k)$ として以下のように記述できる．

$$q_{\text{global}}(\boldsymbol{\Omega}_k) = \mathcal{N}(\boldsymbol{\Omega}_k; \mu_k \mathbf{1}_M, \xi_k^2 \boldsymbol{I}_M), \tag{26}$$

$$q_{\text{local}}(\boldsymbol{\Omega}_k) = \mathcal{N}(\boldsymbol{\Omega}_k; \mathbf{0}_M, \tau_k^2 D^{-1}), \tag{27}$$

$$D = \begin{bmatrix} 1 & -1 & 0 & 0 & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & -1 & 2 & -1 \\ 0 & \dots & 0 & 0 & -1 & 1 \end{bmatrix}. \tag{28}$$

$\mathcal{N}(\boldsymbol{\Omega}_k; \boldsymbol{\mu}, \Sigma)$ は平均 $\boldsymbol{\mu}$，分散 $\Sigma$ をもつ $K$ 次元正規分布である．また，$\mathbf{1}_M$ 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，$I_M$ は $M \times M$ の単位行列を表す．これら 2 つの性質を同時に満たすような事前分布を設計するためは，Product-of-experts [10] と呼ばれるコンセプトを用いて，$\boldsymbol{\Omega}_k$ の事前分布を

$$p(\boldsymbol{\Omega}_k) \propto q_{\text{local}}(\boldsymbol{\Omega}_k)^{\alpha_{\text{local}}} q_{\text{global}}(\boldsymbol{\Omega}_k)^{\alpha_{\text{glocal}}} \tag{29}$$

とすればよい．$\alpha_{\text{local}}, \alpha_{\text{global}}$ は，それぞれこの事前分布に対する $q_{\text{global}}(\boldsymbol{\Omega}_k), q_{\text{global}}(\boldsymbol{\Omega}_k)$ の寄与を調節するハイパーパラメータである．

$U$ に関しても 2 つの性質を用いて事前分布を設計できる．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に偏りがあり，この偏りは各音高の相対的なエネルギーに対する確率分布として記述できる．また，特定の音高のみが曲中演奏されるわけではなく，一つの音高に着目したときにはその音高をもつ音符の分布が時間的にスパースになりやすい．これも各音高のパワーの時間発展に対する確率分布として記述できる．これらの確率分布を簡単に導入するため，$U_{k,m}$ を音高方向に正規化された音量 $R_k = \sum_m U_{k,m}$ ($\sum_k R_k = 1$) と時刻方向に正規化された音量 $A_{k,m} = U_{k,m}/R_k$ ($\sum_m A_{k,m} = 1$) に分解する．すなわち，

$$U_{k,m} = R_k A_{k,m}. \tag{30}$$

とする．これによって上記の 2 つの確率分布が，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．

$$\boldsymbol{R} \sim \text{Dir}(\boldsymbol{R}; \boldsymbol{\gamma}^{(R)}) := \prod_k R_k^{\gamma_k^{(R)}-1}, \tag{31}$$

$$\boldsymbol{A}_k \sim \text{Dir}(\boldsymbol{A}_k; \boldsymbol{\gamma}_k^{(A)}) := \prod_m A_{k,m}^{\gamma_{k,m}^{(A)}-1} \tag{32}$$

$\boldsymbol{\gamma}^{(R)} := [\gamma_1^{(R)}, \dots, \gamma_K^{(R)}]^\top$ はは $\boldsymbol{R}$ の事前分布のハイパーパラメータであり，これに音高の出現頻度を反映できる．$\boldsymbol{\gamma}_k^{(A)} := [\gamma_{k,1}^{(A)}, \dots, \gamma_{k,M}^{(A)}]^\top$ 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．

## 3. パラメータ推論アルゴリズム

観測パワースペクトログラム $\boldsymbol{Y}$ が与えられたときに，2.4 節で構築した確率モデルに基づき事後確率 $p(\Theta|\boldsymbol{Y}) \propto p(\boldsymbol{Y}|\Theta)p(\Theta)$ を最大化するような $\boldsymbol{w}, \Theta := \{\boldsymbol{\Omega}, \boldsymbol{R}, \boldsymbol{A}\}$ を求めたい．したがって，目的関数

$$\mathcal{J}(\Theta) := \ln p(\boldsymbol{Y}|\Theta) + \ln p(\Theta) \tag{33}$$

を最大化する問題として定式化できる．生成モデルを表す項 $\ln p(\boldsymbol{Y}|\Theta)$，事前分布を表す項 $\ln p(\Theta)$ は $\boldsymbol{w}$ の定義域 $\mathcal{W}$ を用いて以下のように書ける．

$$\ln p(\boldsymbol{Y}|\Theta) = \ln \int_{\mathcal{W}} \prod_{l,m} \text{Pois}(Y_{l,m}; X_{l,m}) p(\boldsymbol{w}|\boldsymbol{\beta}, \boldsymbol{\Omega}) \text{d}\boldsymbol{w} \tag{34}$$

$$\ln p(\boldsymbol{\Theta}) = \sum_k \ln p(\boldsymbol{\Omega}_k) + \ln p(\boldsymbol{R}) + \sum_k \ln p(\boldsymbol{A}_k). \tag{35}$$

式 (34) 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的最適解を導くことは困難であるが，この場合には補助関数法 [4, 7, 11] を用いることができる．補助関数法は，目的関数値がちょうど上界になっているような関数（補助関数と呼ぶ）を設計し，補助関数を補助変数と呼ぶ変数とパラメータ $\Theta$ 化することによって，$\mathcal{J}(\Theta)$ 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 (34) に Jensen の不等式が適用でき，以下のように下界が得られる．

$$\ln p(\boldsymbol{Y}|\Theta) \geq \int_{\mathcal{W}} q(\boldsymbol{w}) \left( \sum_{l,m} \ln \text{Pois}(Y_{l,m}; X_{l,m}) + \ln \frac{p(\boldsymbol{w}|\boldsymbol{\beta}, \boldsymbol{\Omega})}{q(\boldsymbol{w})} \right) \text{d}\boldsymbol{w} \tag{36}$$

$$\underset{c}{=} \int_{\mathcal{W}} q(\boldsymbol{w}) \left( \sum_{l,m} Y_{l,m} \ln X_{l,m} - \sum_{l,m} X_{l,m} + \ln \frac{p(\boldsymbol{w}|\boldsymbol{\beta}, \boldsymbol{\Omega})}{q(\boldsymbol{w})} \right) \text{d}\boldsymbol{w} \tag{37}$$

ただし，$=_c$ は定数を除いて一致することを示している．ここで導入された補助変数 $q(\boldsymbol{w})$ は $\int_{\mathcal{W}} q(\boldsymbol{w}) \text{d}\boldsymbol{w} = 1, q(\boldsymbol{w}) > 0$ を満たす. 等号成立条件は $q(\boldsymbol{w})$ が $\boldsymbol{w}$ の事後分布一致するときである．すなわち，この補助関数は EM アルゴリズムの Q 関数であるとも言える．以下では，$\text{E}_{q(\boldsymbol{w})}[\boldsymbol{w}^2] := \int_{\mathcal{W}} q(\boldsymbol{w}) \boldsymbol{w}^2 \text{d}\boldsymbol{w}$ と表記する．

式 (20) を見ると $X_{l,m}$ は $k, n$ に関する和を含むため，式 (37) の括弧内の第 1 項は対数関数の中に和を含む関数型になっており，解析的に解くのが困難となる原因となっている．これは以下のように式 (37) の下界を設計することによって解決できる．式 (37) の括弧内の第 1 項に Jensen の不等式を適用すると，

$$Y_{l,m} \ln X_{l,m} \geq Y_{l,m} \sum_{k,n} \lambda_{k,n,l,m} \ln \frac{w_{k,n,m}^2 e^{-\frac{(x_l - \Omega_{k,m} - \ln n)^2}{2\sigma^2}} U_{k,m}}{\lambda_{k,n,l,m}}, \tag{38}$$

として下界が得られる．ここで，補助変数 $\boldsymbol{\lambda} := \{\lambda_{k,n,l,m}\}_{k,n,l,m}$ は正の実数であり $\sum_{k,n} \lambda_{k,n,l,m} = 1$ を満たす．不等式 (38) の等号成立条件は

$$\lambda_{k,n,l,m} = \frac{w_{k,n,m}^2 e^{-\frac{(x_l - \Omega_{k,m} - \ln n)^2}{2\sigma^2}} U_{k,m}}{X_{l,m}} \tag{39}$$

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

$$\begin{aligned}
&\mathcal{J}^+(\boldsymbol{\lambda}, q(\boldsymbol{w}), \Theta) \\
&\underset{c}{=} E_{q(\boldsymbol{w})}\left[\sum_{l,m} Y_{l,m} \sum_{k,n} \lambda_{k,n,l,m} \ln \frac{w_{k,n,m}^2 e^{-\frac{(x_l - \Omega_{k,m} - \ln n)^2}{2\sigma^2}} U_{k,m}}{\lambda_{k,n,l,m}}\right. \\
&\quad \left. - \sum_{l,m} X_{l,m} + \ln \frac{p(\boldsymbol{w}|\boldsymbol{\beta}, \boldsymbol{\Omega})}{q(\boldsymbol{w})}\right] + \ln p(\Theta).
\end{aligned} \tag{40}$$

と書ける．

式 (40) の括弧内の第 2 項は $\Omega_{k,m}$ に対して非線形ではあるが，$\sum_l X_{l,m}$ は $X(x_0, t_0), \ldots, X(x_{L-1}, t_{M-1})$ を等間隔 $\Delta_x$ でサンプルした点であるため，$x$ に関する積分を用いて $\sum_l X_{l,m}$ は以下のように精度良く近似することができる．

$$\begin{aligned}
\sum_l X_{l,m} &\simeq \frac{1}{\Delta_x} \int_{-\infty}^{\infty} X(x, t_m) \mathrm{d}x \\
&= \frac{1}{\Delta_x} \sum_{k,n} w_{k,n,m}^2 R_k A_{k,m} \int_{-\infty}^{\infty} e^{-\frac{(x - \Omega_{k,m} - \ln n)^2}{2\sigma^2}} \mathrm{d}x \\
&= \frac{\sqrt{2\pi}\sigma}{\Delta_x} \sum_k R_k A_{k,m} \sum_n w_{k,n,m}^2.
\end{aligned} \tag{41}$$

この近似は，式 (40) の括弧内の第 2 項が $\Omega_{k,m}$ にほとんど依存しないことを意味している．

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．まず，補助変数の更新式は等号成立条件であるため，$\boldsymbol{\lambda}$ は式 (39) にしたがって更新し，$q(\boldsymbol{w})$ は $\boldsymbol{w}$ の事後分布を求めれば良い．$\boldsymbol{\lambda}$ 内の $\boldsymbol{w}$ に関する積分のため $\boldsymbol{\lambda}$ を直接計算することは難しいが，$w_{k,n,m}^2$ を $\mathrm{E}_{q(\boldsymbol{w})}[w_{k,n,m}^2]$ として近似的に計算する．詳しい計算は紙面の都合上省くが，$q(\boldsymbol{w})$ の更新則は

$$\begin{aligned}
q(w_{k,m,n}) = &\text{Nakagami}\left(w_{k,m,n}; \sum_l Y_{l,m} \lambda_{k,n,l,m} + 1,\right. \\
&\left. \frac{\sum_l Y_{l,m} \lambda_{k,n,l,m} + 1}{\sqrt{2\pi} R_k A_{k,m} \sigma / \Delta_x + |B_k(e^{jne^{\Omega_{k,m}} u_0})|^2 / (2v^2)}\right)
\end{aligned} \tag{42}$$

となる [12]．ただし $\text{Nakagami}(\zeta; a, b)$ は仲上分布

$$\text{Nakagami}(\zeta; a, b) = \frac{2a^a}{\Gamma(a) b^a} \zeta^{2a-1} e^{-a\zeta^2/b} \tag{43}$$

を表す．$\Theta$ に関する更新式は $\mathcal{J}^{++}(\boldsymbol{\lambda}, q(\boldsymbol{w}), \Theta)$ に対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って導出できる．$\boldsymbol{R}, \boldsymbol{A}$ に関する更新式は，

$$R_k \propto \frac{\sum_{l,m} Y_{l,m} \sum_n \lambda_{k,n,l,m} + \gamma_k^{(R)} - 1}{\sum_{m,n} A_{k,m} \mathrm{E}_{q(\boldsymbol{w})}[w_{k,m,n}^2]}, \tag{44}$$

$$A_{k,m} \propto \frac{\sum_l Y_{l,m} \sum_n \lambda_{k,n,l,m} + \gamma_{k,m}^{(A)} - 1}{R_k \sum_n \mathrm{E}_{q(\boldsymbol{w})}[w_{k,m,n}^2]}, \tag{45}$$

と導出できる．$\ln p(\boldsymbol{w}|\boldsymbol{\beta}, \boldsymbol{\Omega})$ に含まれる $\boldsymbol{\Omega}$ を無視すれば，$\boldsymbol{\Omega}$ の更新式は

$$\begin{aligned}
\boldsymbol{\Omega}_k \leftarrow &\left(\frac{\alpha_{\text{local}}}{\tau_k^2} D + \frac{\alpha_{\text{global}}}{\xi_k^2} \boldsymbol{I}_M + \sum_{n,l} \text{diag}(\boldsymbol{p}_{k,n,l})\right)^{-1} \\
&\times \left(\mu_k \frac{\alpha_{\text{global}}}{\xi_k^2} \boldsymbol{1}_M + \sum_{n,l} (x_l - \ln n) \boldsymbol{p}_{k,n,l}\right),
\end{aligned} \tag{46}$$

$$\boldsymbol{p}_{k,n,l} := \frac{1}{\sigma^2} \left[Y_{l,1} \lambda_{k,n,l,1}, Y_{l,2} \lambda_{k,n,l,2}, \ldots, Y_{l,M} \lambda_{k,n,l,M}\right]^\top \tag{47}$$

と得られる．$\text{diag}(\boldsymbol{p})$ は $\boldsymbol{p}$ の要素を対角成分に並べた対角行列である．この更新則では一部の項を無視しているため，$\mathcal{J}^{++}(\boldsymbol{\lambda}, q(\boldsymbol{w}), \Theta)$ に対する単調増加性が保証されないことに注意されたい．

全極システムのパラメータ $\boldsymbol{\beta}$ に関しては，[13] で定義された目的関数と $\boldsymbol{\beta}$ に依存する補助関数の項が同型であるため，[13] で提案された手法を修正して用いることができる．紙面の都合上詳細は省くが，以下の更新則を反復的に繰り返すことによって $\mathcal{J}^{++}(\boldsymbol{\lambda}, q(\boldsymbol{w}), \Theta)$ は単調増加する．

$$\boldsymbol{h}_k \leftarrow \hat{C}_k(\boldsymbol{\beta}_k) \boldsymbol{\beta}_k, \quad \boldsymbol{\beta}_k \leftarrow C_k^{-1} \boldsymbol{h}_k, \tag{48}$$

ただし，$C_k$ と $\hat{C}_k(\boldsymbol{\beta}_k)$ は $(P+1) \times (P+1)$ の Toeplitz 行列であり，それらの $(p, q)$ 番目の要素はそれぞれ

$$C_{k,p,q} = \frac{1}{MN} \sum_{m,n} \frac{\mathrm{E}_{q(\boldsymbol{w})}[w_{k,m,n}^2]}{2v^2} \cos[(p-q) n e^{\Omega_{k,m}} u_0], \tag{49}$$

$$\hat{C}_{k,p,q}(\boldsymbol{\beta}_k) = \frac{1}{MN} \sum_{m,n} \frac{1}{|B_k(e^{jne^{\Omega_{k,m}} u_0})|^2} \cos[(p-q) n e^{\Omega_{k,m}} u_0] \tag{50}$$

と書ける．

## 4. モノラル音源分離性能の定量評価

全極スペクトル導入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 HTFD と $F_0$ を時不変とした HTFD（スペクトルテンプレートを正規分布状に拘束した場合の調波 NMF [14, 15] に対応．以後，Harmonic NMF）を用いた．分離に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．各音高ごとに録音された演奏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 [16] の RM-C001 から RM-C005 の最初の 30 秒を MIDI シンセサイザー FluidSynth [17] で

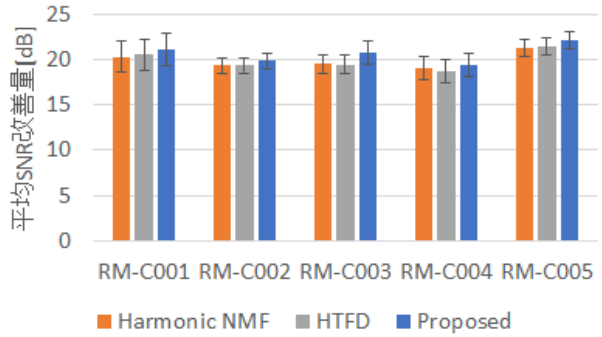


**図 3** $F_0$ を固定した HTFD（Harmonic NMF）と HTFD，提案法（Proposed）で得られた各楽曲に対する平均 SNR 改善量と標準誤差．

合成したの音響信号（サンプリング周波数 16 kHz）を入力として用いた．スペクトログラムを得る際には，時間シフト間隔を 14.6 ms とし解析周波数を 55 Hz から 7040 Hz まで 10 cent 間隔でとった高速近似ウェーブレット変換 [18] を用いた．アナライジングウェーブレットは式 (6) 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（$\sigma = 0.02$）．調波成分の個数は $N = 20$，音高数は $K = 73$ とし，$\mu_k$ は A1（55 Hz）から A♯7（3322 Hz）までの音高に半音単位で対応させた．$\boldsymbol{\Omega}_k$ に関するハイパーパラメータは $(\alpha_{\text{local}}, \alpha_{\text{global}}, \tau_k, \nu_k) = (1, 1, 1.0, 1.25)$ とし，$\boldsymbol{R}$ と $\boldsymbol{A}_k$ に関するハイパーパラメータは $\boldsymbol{\gamma}^{(R)} = 0.8\times\mathbf{1}_K$, $\boldsymbol{\gamma}_k^{(A)} = (1-1.0\times10^{-4})\mathbf{1}_I$ とした．全極システムの次数は $P = 20$ としハイパーパラメータは $\nu = 1$ とした．また，Harmonic NMF は 100 イテレーション，HTFD と提案法は 20 イテレーションとした．

図 3 に，分離音の signal-to-noise ratio（SNR）改善量の平均と標準誤差を楽曲毎に示した．Harmonic NMF と HTFD で得られた 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．これは，MIDI で音源を作成したため $F_0$ の変動が少ないからであると考えられる．提案法は，全曲で Harmonic NMF と HTFD よりも平均 SNR 改善量が高く，Harmonic NMF と HTFD との SNR の平均的な差はそれぞれ約 0.80 dB，約 0.78 dB であった．この結果から，全極スペクトルの導入によるモノラル音源分離性能の向上が確認できる．本実験では全極スペクトルのパラメータも教師なしで推定したが，事前にこのパラメータを学習して分離に用いることもでき，その場合にはより高精度な分離が期待できる．この効果を確かめるため，今後は教師あり学習を用いて実音源に対する分離実験を行う予定である．

## 5. 結論

本報告では，HTFD の多重音解析性能を向上させるために楽音の生成過程をソース・フィルタモデルを導入した多重音解析手法を提案した．ソース・フィルタモデルを全極システムとして導入しパワースペクトログラムモデルを構築したのち，そのパラメータ推論法を示した．定量評価実験により，提案法が HTFD よりも高いモノラル音源分離性能を持つことを確認した．今後は，調やオンセットに関する事前情報を事前分布に導入し，多重音解析に対する効果を定量的に評価する予定である．

**謝辞** 本研究の一部は，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] Bregman, A. S.: *Auditory scene analysis: The perceptual organization of sound*, MIT press (1994).
[3] 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).
[4] Kameoka, H.: Statistical Approach to Multipitch Analysis, PhD Thesis, The University of Tokyo (2007).
[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] 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).
[8] Kameoka, H. and Kashino, K.: Composite autoregressive system for sparse source-filter representation of speech, *Proc. 2009 IEEE International Symposium on Circuits and Systems*, IEEE, pp. 2477–2480 (2009).
[9] Virtanen, T. and Klapuri, A.: Analysis of polyphonic audio using source-filter model and non-negative matrix factorization, *Advances in Models for Acoust. Process., Neural Inf. Process. Syst. Workshop* (2006).
[10] Hinton, G. E.: Training products of experts by minimizing contrastive divergence, *Neural Comput.*, Vol. 14, No. 8, pp. 1771–1800 (2002).
[11] 亀岡弘和：音声音響信号処理のための確率モデルと学習アルゴリズム，第 17 回情報論的学習理論ワークショップ (2014).
[12] 亀岡弘和：音声生成過程の確率モデル，第 13 回情報論的学習理論ワークショップ (2010).
[13] El-Jaroudi, A. and Makhoul, J.: Discrete all-pole modeling, *IEEE Trans. Signal Process.*, Vol. 39, No. 2, pp. 411–423 (1991).
[14] 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).
[15] 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).
[16] Goto, M.: Development of the RWC Music Database, *Proc. Int. Congress Acoust.*, pp. I–553–556 (2004).
[17] : FluidSynth. `http://www.fluidsynth.org/`.
[18] 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).