# 音源信号のスパース基底モデルに基づくブラインド音源分離の検討*

◎北野佑 (東大・情報理工), 亀岡弘和 (NTT CS 研),
和泉洋介, 小野順貴, 嵯峨山茂樹 (東大・情報理工)

## 1 はじめに

音楽信号を対象とした音源分離は、音楽加工、音楽情報検索、自動採譜といった応用先の基礎的技術として広く研究されている。音源分離手法の代表的なものとして、独立成分分析 [1] があげられるが、音楽信号はステレオ録音で、音源数が観測数より多い劣決定な状況であることが多い。そういった劣決定ブラインド音源分離問題を解くためには、音源信号に何らかの仮定を設ける必要があり、最近では音源信号のスパース性に着目した手法が多く用いられている [4]。その一方で、モノラル信号を対象とした音源分離手法として、観測音響信号中に繰り返し生起するスペクトルパターンに着目して基底分解する非負値行列分解 (Nonnegarive Matrix Factorization: NMF) という手法が近年注目されている [2]。NMF のモデルにおいて、音響信号の時間周波数成分が限られた数の非負値スペクトルパターンの線形和により表現されているため、楽器の音色と基本周波数が同じなら、時間領域で一見異なるように見える信号でも一つのパターンで表せる点が特徴的であり、音楽信号をコンパクトに表現するにあたって合理的なモデルと言える。最近では NMF を多チャンネルへ拡張する研究もされている [3]。

本稿では音源の空間情報に基づく時間周波数マスクに NMF の基底分解モデルを組み合わせた楽音向き劣決定ブラインド音源分離手法を提案する。スパース性に基づく音源分離手法では各音源がスペクトログラム上でスパースに存在することを仮定して時間周波数マスクを設計するというものであるが、その際音源の空間的情報のみを用いるため、各音源位置が近い場合は上手く分離できないという問題点を抱えていた。そういった場合でも音源の持つ性質を有効に使うことにより分離性能を保ちたいという動機から、NMF の基底分解モデルを導入する。そしてスパース基底モデルを仮定することにより NMF とは異なった観点の定式化を以下で導入し、シミュレーション実験を行った結果を報告する。

## 2 提案手法

### 2.1 観測モデル

本稿において信号は全て短時間 Fourier 変換により時間周波数領域へ変換されたものとして扱う。$L$ 個の音源を 2 つの無指向性マイクロホンで録音する場合を考え、マイクロホン間距離に対して各音源が十分に遠方から到来すると仮定する。観測 $\mathbf{Y}_{\omega,t}$ は、

$$\mathbf{Y}_{\omega,t} = \sum_{l=1}^{L} \begin{pmatrix} 1 \\ e^{j\delta_l \omega} \end{pmatrix} S^{(l)}_{\omega,t} + \mathbf{N}_{\omega,t} \tag{1}$$

と近似的に表すことができる。$\omega, t$ はそれぞれ周波数と時間の index で、$S^{(l)}_{\omega,t}$ は $l$ 番目の音源の時間周波数成分、$\delta_l$ は $S^{(l)}_{\omega,t}$ に関してマイクロホン間で生じる時間差である。$\mathbf{N}_{\omega,t}$ はノイズで、両チャンネルが平均 0、分散 $\sigma^2$ の独立な複素正規分布に従うと仮定する。ここで音響信号が限られた数の振幅スペクトルパターンを持った信号で構成されるとすると、音源 $S^{(l)}_{\omega,t}$ は

$$S^{(l)}_{\omega,t} = \sum_{k \in K_l} H_{\omega,k} U_{t,k} e^{j\phi_{\omega,t,k}} \tag{2}$$

と表すことができる。$H_{\omega,k}$ は時刻に依らずグローバルに決定される非負値振幅スペクトルパターン (以下、基底と呼ぶ)、$U_{t,k}$ は基底にかかる時変の重み係数 (以下、アクティベーションと呼ぶ)、$e^{j\phi_{\omega,t,k}}$ は位相スペクトルに当たり、この観測モデルは複素 NMF[5] と同様のモデルである。。$K_l$ は $S^{(l)}_{\omega,t}$ を構成する基底 $H_{\omega,k}$ の集合である。なお

$$\sum_{\omega} H_{\omega,k} = 1 \tag{3}$$

という規格化条件を設ける。

ここで各基底によって構成される信号がスペクトログラム上でスパースに存在していて、観測の各時間周波数 bin でただ一つの基底のみアクティベートすると仮定する。その際、観測の $(\omega, t)$ 成分においてアクティベートしている基底の index を $k'$ とすると、

$$\mathbf{Y}_{\omega,t} = \begin{pmatrix} 1 \\ e^{j\delta_{l'}\omega} \end{pmatrix} H_{\omega,k'} U_{t,k'} e^{j\phi_{\omega,t,k'}} + \mathbf{N}_{\omega,t} \tag{4}$$

と表すことができる。上記の式において $l'$ は $k' \in K_l$ を満たす $l$ とした。なお $k', l'$ は $\omega, t$ の変数だが、表記の簡略化のため、省略した。この基底のスパース性に基づく観測モデル (本稿ではこれをスパース基底モデルと呼ぶ) を仮定することにより、モデルの自由度を下げることができ、$\mathbf{H}$, $\mathbf{U}$ の解を一意に求めることが可能である。実際、各音源が単旋律を奏でている場合は、時間毎に音源内で一つの基底しかアクティベートしていないので、音源のスパース性が成立する限りは、この仮定は妥当である。

### 2.2 EM アルゴリズムによる対数事後確率最大化

表記の簡略化のために、$\mathbf{H} \equiv (H_{\omega,k})_{\Omega \times K}$, $\mathbf{U} \equiv (U_{t,k})_{T \times K}$, $\boldsymbol{\phi} \equiv (\phi_{\omega,t,k})_{\Omega \times T \times K}$ とし、さらに $\Psi \equiv \{\mathbf{H}, \mathbf{U}, \boldsymbol{\phi}\}$, $\boldsymbol{\delta} \equiv \{\delta_1, \cdots, \delta_L\}$ とする。$\Omega, T$ はそれぞれ周波数、時間の bin 数である。各パラメータは対数事後確率最大化問題に対し、[4] と同様に EM アルゴリズムを適用することにより推定可能である。今、観測 $\mathbf{Y}_{\omega,t}$ が得られた時に、事後確率は

$$p(\Psi, \boldsymbol{\delta}, \sigma^2 | \mathbf{Y}) \propto p(\mathbf{Y} | \Psi, \boldsymbol{\delta}, \sigma^2) p(\Psi, \boldsymbol{\delta}, \sigma^2) \tag{5}$$

と書くことができる。ここで $\mathbf{H}, \mathbf{U}, \boldsymbol{\phi}, \boldsymbol{\delta}, \sigma^2$ はそれぞれ独立とし、アクティベーション $U_{t,k}$ の事前確率 $p(U_{t,k})$ は一般化正規分布

$$p(U_{t,k}) = \frac{1}{2\Gamma(1+p^{-1})b} e^{-\frac{|U_{t,k}|^p}{bp}} \tag{6}$$

に従い、$\mathbf{U}$ 以外は一様分布に従うとする。但し、$0 < p \leq 2, b > 0$ とする。この時、解きたい問題は、

---

*Blind source separation based on a sparse basis activation model for source signals. by KITANO Yu(The University of Tokyo), KAMEOKA Hirokazu(NTT), IZUMI Yosuke, ONO Nobutaka, SAGAYAMA Shigeki(The University of Tokyo)

$$\Xi(\Psi, \boldsymbol{\delta}, \sigma^2) \equiv \log p(\mathbf{Y}|\Psi, \boldsymbol{\delta}, \sigma^2) + \log p(\mathbf{U}) \quad (7)$$

の最大化と等価である。

各時間周波数 bin でアクティベートする基底の index は直接観測できない隠れ変数として観測モデルに含まれているため、直接対数事後確率を最大化するのは非常に困難だが、EM アルゴリズムを適用することにより、反復的にパラメータ推定が可能である。EM アルゴリズムとは、観測できない隠れ変数が観測モデルに含まれている場合に、隠れ変数の期待値を求める E-step と、最大化したい対数事後確率の条件付期待値 (Q 関数) を最大化する M-step を反復することにより、対数事後確率を局所最大化することができるアルゴリズムである。今、基底 $k$ が単一方向より到来し観測の $(\omega, t)$ 成分でアクティベートする尤度は、

$$\log p(\mathbf{Y}_{\omega,t}|\Psi_k, \delta_l, \sigma^2) = -\log(\pi\sigma^4) - \frac{1}{\sigma^2}\left|\mathbf{Y}_{\omega,t} - \boldsymbol{a}_\omega^{(l)} H_{\omega,k} U_{t,k} e^{j\phi_{\omega,t,k}}\right|^2 \quad (8)$$

と表される。ここで $\boldsymbol{a}_\omega^{(l)} = \begin{pmatrix}1 & e^{j\delta_l\omega}\end{pmatrix}^T$ とし、$l$ は $k \in K_l$ を満たすものとした。これを用いて Q 関数は、

$$Q(\theta, \theta^{(n)}) = \sum_{k,\omega,t} m_{k,\omega,t} \log p(\mathbf{Y}_{\omega,t}|\Psi_k, \delta_l, \sigma^2) - \lambda \sum_{t,k} U_{t,k}^p \quad (9)$$

と設計できる。$\theta \equiv \{\Psi, \boldsymbol{\delta}, \sigma^2\}$ であり、$^{(n)}$ は $n$-step 目のパラメータである。$\lambda = b^{-p}$ とした。$m_{k,\omega,t}$ は、

$$m_{k,\omega,t} = p(k|\Psi^{(n)}, \delta^{(n)}, \sigma^{2(n)}, \mathbf{Y}_{\omega,t}) \quad (10)$$

$$= \frac{p(\mathbf{Y}_{\omega,t}|\Psi_k^{(n)}, \delta_l^{(n)}, \sigma^{2(n)})}{\sum_{k'} p(\mathbf{Y}_{\omega,t}|\Psi_{k'}^{(n)}, \delta_{l'}^{(n)}, \sigma^{2(n)})} \quad (11)$$

で定義され、本稿ではこれを基底 $k$ の帰属率と呼ぶ。E-step ではこの式により更新される。

M-step ではこの Q 関数を最大化するパラメータを求める。各パラメータで微分して 0 とおくと、

$$H_{\omega,k}^{(n+1)} = \frac{\sum_t m_{k,\omega,t}^{(n)} U_{t,k}^{(n)} \left|\mathbf{Y}_{\omega,t}^H (\boldsymbol{a}_\omega^{(l)})^{(n)}\right|}{2\sum_t m_{k,\omega,t}^{(n)} (U_{t,k}^{(n)})^2}$$

$$U_{t,k}^{(n+1)} = \frac{\sum_\omega m_{k,\omega,t}^{(n)} H_{\omega,k}^{(n)} \left|\mathbf{Y}_{\omega,t}^H (\boldsymbol{a}_\omega^{(l)})^{(n)}\right|}{2\sum_\omega m_{k,\omega,t}^{(n)} (H_{\omega,k}^{(n)})^2 + \lambda p \sigma^2 (U_{t,k}^{(n)})^{p-2}}$$

$$\phi_{\omega,t,k}^{(n+1)} = \arg\left(\mathbf{Y}_{\omega,t}^H (\boldsymbol{a}_\omega^{(l)})^{(n)}\right)^H$$

$$\left(\sigma^2\right)^{(n+1)} = \sum_{k,\omega,t} \frac{m_{k,\omega,t}^{(n)}}{2\Omega T} \left|\mathbf{Y}_{\omega,t} - (\boldsymbol{a}_\omega^{(l)})^{(n)} H_{\omega,k}^{(n)} U_{t,k}^{(n)} e^{j\phi_{\omega,t,k}^{(n)}}\right|^2$$

と導出される。ただし $\mathbf{H}$ の更新後は (3) を満たすように正規化する。また $\boldsymbol{\delta}$ の MAP 解は解析的に更新式を求められないため、離散値全探索により Q 関数を最大化するパラメータを求める。

### 2.3 Wiener フィルタによる音源分離

前節に示したアルゴリズムにより推定されたパラメータより、各音源を推定することが可能である。どのようにして各音源を再現するかは様々な形があり得るが、ここでは推定されたパワースペクトルの情報を有効に用いるために、Wiener フィルタに基づく時間周波数マスクによる分離を考えると、$S_{\omega,t}^{(l)}$ は、

$$\hat{S}_{\omega,t}^{(l)} = \frac{\sum_{k\in K_l}(H_{\omega,k}U_{t,k})^2}{\sum_k (H_{\omega,k}U_{t,k})^2} \cdot \frac{(\boldsymbol{a}_\omega^{(l)})^H \mathbf{Y}_{\omega,t}}{2} \quad (12)$$

と表すことができる。

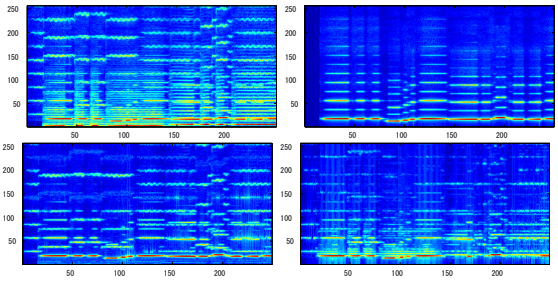

Fig. 1　観測信号と音源信号、分離信号のスペクトログラムの一例。左上：観測信号、右上：音源信号、左下：従来法による分離信号、右下：提案法による分離信号である。なお全て縦軸は周波数、横軸は時間の index である。

Table 1　分離性能 [dB]

| | $s_1$ | $s_2$ | $s_3$ |
|---|---|---|---|
| 混合信号 | -12.1 | 2.1 | 2.6 |
| 従来法 | -3.3 | 12.3 | 13.9 |
| 提案法 | 3.3 | 16.0 | 13.2 |

Table 2　音源定位結果 [ms]

| | $s_1$ | $s_2$ | $s_3$ |
|---|---|---|---|
| 真の値 | 0.20 | 0.15 | -0.12 |
| 従来法 | -0.24 | 0.17 | -0.09 |
| 提案法 | 0.19 | 0.12 | -0.12 |

## 3　シミュレーション実験

提案手法の有効性を検証するためにシミュレーション実験を行った。音源数 3、観測数 2、$\lambda = 0.01$, $p = 2$、雑音・残響なしの条件で混合シミュレーションを行った。各音源の楽器は $s_1$: violin, $s_2$: clarinet, $s_3$: contrabass、信号は MIDI 音源により作成した。今回は基底の形を予め別データから学習する Supervised なアプローチについて実験をした。学習データは分離データとは別の MIDI 音源を用い、各楽器が演奏可能な全ての音程について学習させた (Open data)。サンプリング周波数は 16kHz、フレーム長 1024 点、フレームシフト 512 点、窓関数には Hanning 窓を用いることにより、短時間 Fourier 変換をして時間周波数領域へ変換した。なお [4] を従来法とし、実験の比較対象とした。分離性能 (SN 比) と音源定位結果 (到来時間差) は Table. 1, 2 のようになった。また観測信号や音源信号、分離信号のスペクトログラムの一例を Fig. 1 に示した。これらからも分かるとおり、音源位置が近い場合でも提案法は従来法より音源定位の精度も分離性能も高いことが分かる。

## 4　おわりに

本稿ではスパース性に基づく音源分離手法に音源の持つパワースペクトルの情報を積極的に取り入れ、基底のスパースモデルを定式化した劣決定ブラインド音源分離手法を提案し、その有効性を確認した。今後は観測から基底を学習しつつ分離する Unsupervised なアプローチの実験を行い、また観測信号間の強度比やノイズに拡散音場モデルを導入することにより、分離性能の向上を狙ったり、残響にも対応した手法を確立していくつもりである。

## 参考文献


[1] A. Hyvärinen *et al.*, "Independent Component Analysis," John Wiley, New York, 2001.
[2] T. Virtanen, IEEE Trans. on ASLP, vol. 15, no. 3, pp.1066–1074, 2007
[3] A. Ozerov and C. Févotte, Proc. ICASSP'09, pp.3137–3140, 2009
[4] 和泉他, 音講論 (春), pp.555–556, 2007.
[5] 亀岡他, 音講論 (秋), pp.657–660, 2008.