4Q-03

音源スペクトログラムの低ランク性とスパース性を考慮したNMF-LDAに基づくマルチチャネル音源定位と音源分離

板倉 光佑　坂東 宜昭　中村 栄太　糸山 克寿　吉井 和佳
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

近年，マイクロホンアレイを用いたマルチチャネル音源分離において，時間・周波数クラスタリングに基づく手法が有望視されている [1–4]．このアプローチでは，各音源スペクトログラムが時間・周波数領域でスパースであると仮定することで，混合音のスペクトログラムにおける各時間・周波数ビンにおいては，いずれか一つの音源成分が直接観測されるとみなす．最近，音源分離と音源定位を一挙に行うため，時間周波数ビンを音源にクラスタリングすると同時に，音源を方向ごとにクラスタリングすることができる潜在的ディリクレ配分法 (LDA) の拡張モデルが提案されている [4]．

一方，単チャネル音源分離手法においては，空間的な情報を利用することができないため，非負値行列因子分解 (NMF) [5] のように音源スペクトログラムの低ランク性を仮定した手法が盛んに研究されている．この種のアプローチは，マイク数が一つであっても複数の音源を分離できること，すなわち劣決定条件でも利用できるという好ましい性質を持つ．このことは，音源スペクトログラムのスパース性と低ランク性の両方を考慮することにより，マルチチャネル環境下において高精度な音源分離が可能になることを示唆している．

本稿では，音源分離と音源定位の精度を向上させるため，音源スペクトログラムのスパース性に基づくLDAと低ランク性に基づくNMFとを包含する統一的なベイズモデルを提案する．本手法は，独立ベクトル分析 (IVA) とNMFを統合したランク1マルチチャネルNMF (MNMF) に着想を得ている [6]．提案法の特徴は，LDAによるクラスタリングで得られる欠損値を含む（その音源が優位なビンのみ値を持つ）音源スペクトログラムに対して，NMFを用いて行列補完を行うところにある．その結果，全時間・周波数領域の音源スペクトログラムを復元してからNMFを用いて低ランク近似するMNMFと比べて，安定した音源分離ができることが期待される．

2. NMF-LDA

NMF-LDAのベイズモデル（図1）の定式化について説明する．マイク数を M，音源数を K とし，ある音源 k $(1 \leq k \leq K)$ がどのように観測されるかを考える．

- $\boldsymbol{x}_{tfk} \in \mathbb{C}^M$：音源 k を M 個のマイクロホンアレイで観測したとき，時刻 t，周波数 f の成分
- $y_{tfk} \in \mathbb{C}$：音源 k の時刻 t，周波数 f の成分
- $\boldsymbol{b}_{fd} \in \mathbb{C}^M$：方向 d，周波数 f の伝達関数

このとき，周波数領域の瞬時混合過程を仮定すると，観測スペクトルと音源スペクトルの関係は次式となる．

$$\boldsymbol{x}_{tfk} = \boldsymbol{b}_{fd_k} y_{tfk} \tag{1}$$

ここで，d_k は音源 k の方向を示す．さらに，音源スペクトルが以下の等方的な複素ガウス分布に従うと仮定する．

$$y_{tfk_{tf}} \sim \mathcal{N}_c(y_{tfk_{tf}}|0, \lambda_{tfk_{tf}}) \tag{2}$$

Multi-channel Sound Source Localization and Separation based on NMF-LDA Considering Low-rankness and Sparseness of Source Spectrograms : *Kousuke Itakura, Yoshiaki Bando, Nakamura Eita, Katsutoshi Itoyama, Kazuyoshi Yoshii (Kyoto Univ.)*

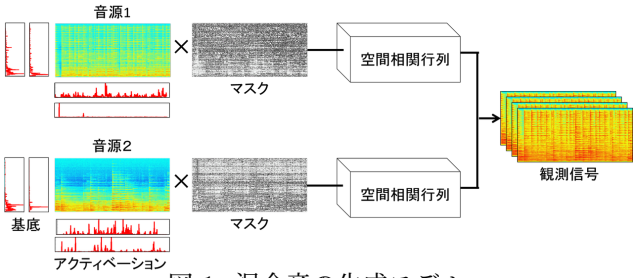

図 1: 混合音の生成モデル

このとき，式 (1) で与えられる線形関係から，観測スペクトルは以下の複素正規分布に従う [7]．

$$\boldsymbol{x}_{tfk} \sim \mathcal{N}_c(\boldsymbol{x}_{tfk}|\boldsymbol{0}, \lambda_{tfk_{tf}} \boldsymbol{G}^{-1}_{fd_{k_{tf}}}) \tag{3}$$

ここで，$\boldsymbol{G}^{-1}_{fd} = \boldsymbol{b}_{fd}\boldsymbol{b}^H_{fd}$ とした．このとき，観測される混合音スペクトル $\boldsymbol{x}_{tf}$ は次式で与えられる．

$$\boldsymbol{x}_{tf} = \sum_k \boldsymbol{b}_{fd_k} y_{tfk} \tag{4}$$

ここで，音源スペクトルがスパースである，すなわち，各時刻 t，周波数 f において，ただ一つの音源 k_{tf} が支配的である場合には，次式で近似できる．

$$\boldsymbol{x}_{tf} = \boldsymbol{b}_{fd_{k_{tf}}} y_{tfk_{tf}} \tag{5}$$

いま，時刻 t，周波数 f の観測に寄与する音源と音源 k の方向を示す変数 z_{tfk}，s_{kd} を新たに導入する．

$$z_{tfk} = \begin{cases} 1 & (k = k_{tf}) \\ 0 & (k \neq k_{tf}) \end{cases}, \quad s_{kd} = \begin{cases} 1 & (d = d_k) \\ 0 & (d \neq d_k) \end{cases}$$

このとき，式 (3) より観測スペクトルは次式に従う．

$$\boldsymbol{x}_{tf} \sim \prod_{k,d} \mathcal{N}_c(\boldsymbol{x}_{tf}|\boldsymbol{0}, \lambda_{tfk}\boldsymbol{G}^{-1}_{fd})^{z_{tfk}s_{kd}} \tag{6}$$

最終的に，尤度関数は次式で与えられる．

$$P(\boldsymbol{X}|\boldsymbol{\lambda}, \boldsymbol{G}, \boldsymbol{Z}, \boldsymbol{S}) = \prod_{t,f,k,d} \mathcal{N}_c(\boldsymbol{x}_{tf}|\boldsymbol{0}, \lambda_{tfk}\boldsymbol{G}^{-1}_{fd})^{z_{tfk}s_{kd}} \tag{7}$$

2.1 事前分布の設計

まず，LDA部分に関して定式化を行う．$\boldsymbol{G}$，$\boldsymbol{Z}$，$\boldsymbol{S}$ に関しては，従来法 [4] と同様の事前分布を設定した．

$$\boldsymbol{G}_{fd} \sim \mathrm{Wishart}_c(\boldsymbol{G}_{fd}|\nu, \boldsymbol{G}^0_{fd}) \tag{8}$$

$$\boldsymbol{z}_{tf} \sim \mathrm{Mult}(\boldsymbol{z}_{tf}|1, \boldsymbol{\pi}_t), \quad \boldsymbol{\pi}_t \sim \mathrm{Dir}(\boldsymbol{\pi}_t|a^\pi_0 \boldsymbol{1}_K) \tag{9}$$

$$\boldsymbol{s}_k \sim \mathrm{Mult}(\boldsymbol{s}_k|1, \boldsymbol{\phi}), \quad \boldsymbol{\phi} \sim \mathrm{Dir}(\boldsymbol{\phi}|a^\phi_0 \boldsymbol{1}_D) \tag{10}$$

次に，NMF部分に関して定式化を行う．音源 k のパワースペクトログラム λ_{tfk} が低ランク性を持つと仮定すると，基底スペクトル $\boldsymbol{W}_k \in \mathbb{R}_+^{F\times L}$ とアクティベーション $\boldsymbol{H}_k \in \mathbb{R}_+^{L\times T}$ との積に分解できる．

$$\lambda_{tfk} = \sum_l w_{kfl} h_{klt} \tag{11}$$

ここで，基底数を L とした．さらに，$\boldsymbol{W}$ と $\boldsymbol{H}$ に対する事前分布としてガンマ分布を設定した．

$$w_{kfl} \sim \mathrm{Gamma}(w_{kfl}|a^w_0, b^w_0) \tag{12}$$

$$h_{klt} \sim \mathrm{Gamma}(h_{klt}|a^h_0, b^h_0) \tag{13}$$

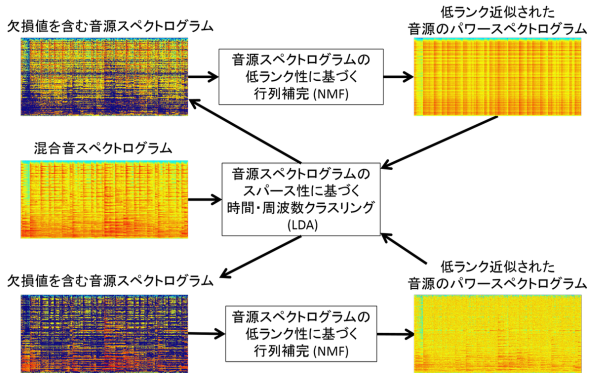

図 2: 音源数を 2 としたときの提案法の流れ．欠損値を含むスペクトログラムにおける青色の部分は欠損値を示す．

2.2 事後分布の推論

我々の目標は，観測データ $\boldsymbol{X}$ が与えられたときに，事後分布 $p(\boldsymbol{G}, \boldsymbol{Z}, \boldsymbol{S}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{W}, \boldsymbol{H}|\boldsymbol{X})$ を計算することである．しかし，解析的な計算は困難であるため，事前分布の共役性から $\boldsymbol{\pi}, \boldsymbol{\phi}$ を積分消去したうえで，ギブスサンプリングを用いることにした．具体的には，図 2 に示す通り，音源スペクトルの空間情報に関する変数 $\boldsymbol{G}$，スパース性に関する変数 $\boldsymbol{Z}, \boldsymbol{S}$ (LDA)，低ランク性に関する変数 $\boldsymbol{W}, \boldsymbol{H}$ (NMF) を交互に更新するのステップを収束するまで反復する．ただし，NMF においては，変分事後分布として一般化逆ガウス分布を導出し，それを提案分布とする Metropolis-Hastings (MH) 法 [8] を用いた．通常の NMF と異なり，LDA によるクラスタリングの結果，欠損値を含んだ音源スペクトログラムに対して NMF を行うことで，欠損している時間・周波数領域を復元することができる．最終的に，得られた変数と多チャネルウィナーフィルタを用いた分離音を生成する．

3. 評価実験

NMF によりパワーの推定を行う有用性を確認するため，パワーの事前分布にガンマ分布をおいた大塚らの手法 [4] と分離精度を比較した．ただし，[4] では音源数の推定も同時に行うが，対等な条件で評価するため，ここでは音源数は事前に与えるものとした．図 3 のように音源数 3，マイク数 4 とし，残響時間 400ms で録音したインパルス応答により作成したシミュレーション混合音を用いて評価した．提案法，大塚法ともにサンプリング回数は 100 回，$K = 3$，$D = 72$ とした．また提案法では NMF の基底数は音源ごとに 10 ずつとし，GIG のパラメータを定めるための更新回数は 50 回とした．音源には SiSEC [10] に収録されている bearlin-roads_snip_85_99 に含まれる guitar，piano，vocal，bass，hi hat のうちの 3 つを用いて得られる混合音 $10(={}_5C_3)$ 個を使用した．

図 4 に実験結果の SDR を示す．提案法では，従来法に比べて SDR の最小値が −17 dB から −11 dB に上昇した．その結果，従来法の SDR の平均値が 0.4dB であったのに対し，提案法では SDR の平均値は 1.1dB となり，提案法により SDR が 0.7 dB 向上した．また，分離結果の SDR を混合音ごとにまとめたものの一部を表 1，2 に示す．ただし，表 1 は，提案法のほうが従来法より優れていたもののうち上位 3 つの混合音，表 2 は，従来法のほうが提案法より優れていたもののうち上位 3 つの混合音を示す．ここで，表 1 の全ての混合音に hi hat が含まれないのに対し，表 2 の全ての混合音に hi hat が含まれている．hi hat は 5 つの音源の中で唯一調波構造を持たない音源である．したがって，提案法は調波構造をもつ音源の混合音に対してより有効であると言える．

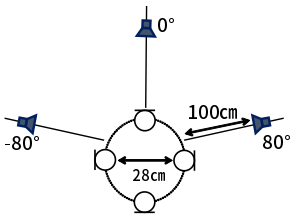

図 3: 実験条件

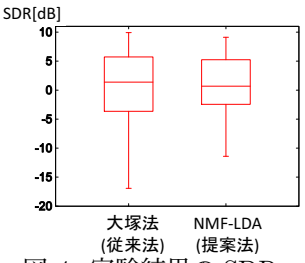

図 4: 実験結果の SDR

表 1: 提案法でより改善された混合音の分離結果．- はその混合音に含まれていない音源を示す．それぞれ，上段が提案法，下段が従来法による分離音の SDR[dB].

	guitar	piano	vocal	bass	hi hat
混合音 1	−2.4	−2.9	7.6	-	-
	−4.1	−5.0	6.8	-	-
混合音 2	0.96	-	7.4	−9.7	-
	0.73	-	7.2	−12	-
混合音 3	-	3.0	7.1	−11	-
	-	3.6	5.7	−14	-

表 2: 提案法で改善されなかった混合音の分離結果．

	guitar	piano	vocal	bass	hi hat
混合音 4	-	3.3	7.0	-	−1.6
	-	4.5	6.5	-	−0.83
混合音 5	−0.037	0.98	-	-	3.8
	2.0	2.7	-	-	3.2
混合音 6	4.3	-	-	−6.4	5.3
	5.1	-	-	−3.6	5.7

4. おわりに

本稿では，音源分離・定位を LDA により行い，パワーの推定を NMF を用いて行う NMF-LDA を提案した．実験では，パワーの推定において NMF を用いない手法と比べて提案法により SDR が平均で 0.7dB 向上することを確認した．今後は話し声や環境雑音など，音楽以外の音源による混合音に対しての分離性能の評価を行う．

謝辞 本研究の一部は、JSPS 科研費 24220006, 15K12063 の支援を受けた．

参考文献

[1] N. Ito *et al.* Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors. *ICASSP2013*, 3238–3242, May 2013.
[2] H. Sawada *et al.* Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. *TASLP*, 19(3):516–527, March 2011.
[3] H. Sawada *et al.* A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures. *WASPAA*, 139–142, Oct 2007.
[4] T. Otsuka *et al.* Bayesian nonparametrics for microphone array processing. *TASLP*, 493–504, 2014.
[5] 亀岡弘和. 非負値行列因子分解. **計測と制御**, 51(9):835–844, 2012.
[6] D. Kitamura *et al.* Relaxation of rank-1 spatial constraint in overdetermined blind source separation. *Proc. of EUSIPCO2015*, 1271–1275. [IEEE], 2015.
[7] N.Q.K. Duong *et al.* Under-determined reverberant audio source separation using a full-rank spatial covariance model. *TASLP*, 1830–1840, 2010.
[8] S. Chib *et al.* Understanding the metropolis-hastings algorithm. *The american statistician*, 49(4):327–335, 1995.
[9] B. Jørgensen. *Statistical properties of the generalized inverse Gaussian distribution*, volume 9. Springer Science & Business Media, 2012.
[10] S. Araki *et al.* The 2011 signal separation evaluation campaign (sisec2011):-audio source separation. *Latent Variable Analysis and Signal Separation*, 414–422. Springer, 2012.