社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

［ポスター講演］マルチチャネル音源分離のための ネスト型基底・音源混合モデルに基づく時間周波数クラスタリング

板倉　光佑†　坂東　宜昭†　中村　栄太†　糸山　克寿†　吉井　和佳†
河原　達也†

† 京都大学大学院 情報学研究科
E-mail: †{itakura,bando,enakamura,itoyama,yoshii,kawahara}@sap.ist.i.kyoto-u.ac.jp

あらまし　本稿では，時間周波数クラスタリングによるマルチチャネル音源分離のためのネスト型混合モデルについて述べる．時間周波数クラスタリングによる音源分離では，混合音は各音源の観測モデルの混合モデルに基づいて生成されるとする．この混合モデルを推定するための特徴量として各マイクでの音の位相と各音源のパワースペクトログラムを用いることができる．提案法ではこのパワースペクトログラムを基底の混合モデルによりモデル化する．これにより，提案法では混合音が音源の混合モデルで，各音源が基底の混合モデルでモデル化される．これをネスト型基底・音源混合モデルと呼ぶ．評価実験により，音声などの混合音に対して提案法により SDR と SIR が向上することを確認した．

キーワード　マルチチャネル音源分離，時間周波数クラスタリング，潜在的ディリクレ配分法

1. はじめに

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，これまでに多くの手法が提案されてきた．広く用いられている手法のうちの一つに，独立成分分析 (ICA) [1] がある．ICA は各音源の統計的な独立性を仮定することにより分離行列を推定する．この ICA をもとに独立ベクトル分析 (IVA) [2] や FastICA [3] などのさまざまな手法が提案されているが，これらの手法は共通してマイク数が音源数より少ない劣決定条件では分離できないという問題点がある．

これに対し，劣決定条件でも分離が可能な手法として時間・周波数クラスタリングに基づく音源分離法が着目されている [4–9]．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，混合音スペクトログラムの各時間・周波数ビンにおける観測はそれぞれいずれか一つの音源成分が直接観測されたものであるとみなす．つまり，この仮定では，混合音の観測モデルは各音源の観測モデルの混合モデルとして扱われる．この混合モデルを推定するため，マイク間の音の位相差とパワー差を特徴量として用いた混合ワトソン分布のクラスタリング [4–7] や各マイクでの音の位相とパワーを特徴量として用いた混合ガウス分布のクラスタリング [8, 9] による分離法が提案されている．

特徴量にマイク間のパワー差ではなく各マイクでのパワーを用いる利点として，音源の混合過程のような空間モデルだけでなく各音源のパワースペクトログラムの性質に基づいた音源のモデル化も同時にできるという点が挙げられる．大塚ら [8] はパワースペクトログラムのスパース性に基づいたモデル化を行うため，このパワーに対しスパースとなる事前分布を用いたモデル化を行った．また，パワースペクトログラムの低ランク性を用いたモデル化も行われている [9]．この手法では，単チャネル音源分離でよく用いられる非負値行列因子分解 (NMF) [10] のようにパワースペクトログラムが基底スペクトルとアクティベーションの積の和を用いて表現される．マルチチャネル音源分離においても，空間モデルだけではなく音源モデルも考慮することにより分離性能の向上が期待される．

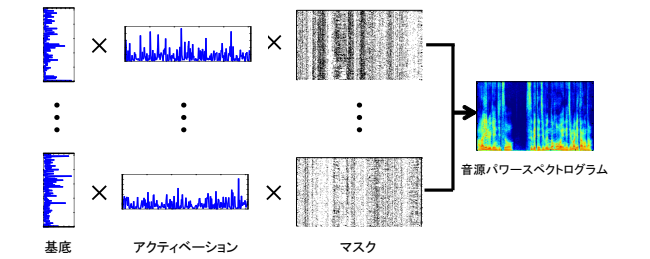
本稿では，このパワースペクトログラムをさらに混合モデルを用いてモデル化した２段階のネスト型混合モデルによる音源分離法を提案する．提案法では，NMF のようにパワースペクトログラムを基底スペクトルとアクティベーションに分解し，それに加えて基底スペクトルのスパース性を仮定する．つまり，パワースペクトログラムの各時間周波数ビンではいずれか一つの基底の成分のみが観測されるとする．これにより，提案法では空間モデルが音源の混合モデル，音源モデルが基底の混合モデルを用いてモデル化される．この２つの混合モデルをネスト型基底・音源混合モデルと呼ぶ．提案法ではネスト型基底・音源混合モデルに対して潜在的ディリクレ配分法 (LDA) の枠組みを用いてギブスサンプリングを行うことで二つの混合モデルを同時に推定する．図 1 に提案法における混合音の生成モデルを示す．

2. 提案法

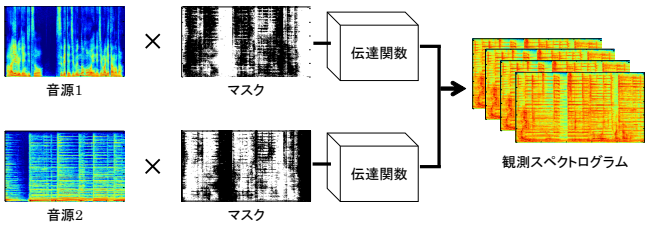
提案法では，ベイズモデルを用いてネスト型基底・音源混合モデルのモデル化を行い，LDA の枠組みを用いてそのモデルの推定を行う．本章ではそのモデル化と推定方法について述べる．

2.1 モデル化

ここでは提案法のモデルの定式化について述べる．提案法で



(a) 音源モデル



(b) 空間モデル

図 1　提案法の生成モデル．音源モデルでは各音源のパワースペクトログラムが基底・アクティベーション・マスクにより構成される．マスクにより一つの基底のみが選択される (黒色の部分が選択された部分)．空間モデルでは混合音スペクトログラムが音源スペクトログラム・伝達関数・マスクにより構成される．マスクにより一つの音源のみが選択される．

は時間領域の信号に対して短時間フーリエ変換 (STFT) を行うことにより得られる時間周波数領域の信号に対してモデル化を行う．まず K 個の音源を M 個のマイクを用いて録音するとし，時刻 t，周波数 f での観測 $\boldsymbol{x}_{tf}$ と音源信号 $\boldsymbol{y}_{tf}$ を以下のように定義する．

$$\boldsymbol{x}_{tf} = [x_{tf1}, \cdots, x_{tfM}]^T \in \mathbb{C}^M \tag{1}$$

$$\boldsymbol{y}_{tf} = [y_{tf1}, \cdots, y_{tfK}]^T \in \mathbb{C}^K \tag{2}$$

このとき周波数領域での瞬時混合を仮定すると，観測は以下のように表される．

$$\boldsymbol{x}_{tf} = \sum_{k=1}^{K} \boldsymbol{a}_{fk} \cdot y_{tfk} \tag{3}$$

ただし，$\boldsymbol{a}_{fk}$ は周波数 f での音源 k の伝達関数である．ここで，y_{tfk} が次のような複素ガウス分布に従うとする．

$$y_{tfk} \sim \mathcal{N}_{\mathbb{C}}(0, \lambda_{tfk}) \tag{4}$$

$\lambda_{tfk} = \mathbb{E}[y_{tfk}^2]$ は時刻 t，周波数 f での音源 k のパワーを表す．このとき音源 k のみを観測した時の観測 $\boldsymbol{x}_{tfk}$ は次のような複素ガウス分布に従う．

$$\boldsymbol{x}_{tfk} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \lambda_{tfk}\boldsymbol{G}_{fk}^{-1}) \tag{5}$$

ここで，$\boldsymbol{G}_{fk}^{-1}$ は周波数 f での音源 k の空間相関行列であり，$\boldsymbol{G}_{fk}^{-1} = \boldsymbol{a}_{fk}\boldsymbol{a}_{fk}^H$ である．ただし $*^H$ はエルミート共役を示す．

ここで，音源スペクトログラムがスパースである，すなわち，各時間周波数ビンにおいて観測される音は高々一つであるとし，そのときの混合音の観測モデルについて考える．まず各時間周波数ビンにおいて観測される音源を示すための変数を $\boldsymbol{z}_{tf} = [z_{tf1}, \cdots, z_{tfK}]^T$ とする．ただし，$\boldsymbol{z}_{tf}$ は 1 of K 表現のベクトルであり，音源 k が観測されるときは $z_{tfk} = 1$ となり，それ以外のときは 0 となる．このとき，観測 $\boldsymbol{x}_{tf}$ は次のような分布にしたがって生成される．

$$\boldsymbol{x}_{tf} \sim \prod_{k=1}^{K} \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \lambda_{tfk}\boldsymbol{G}_{fk}^{-1})^{z_{tfk}} \tag{6}$$

ここで，空間相関行列は音源の種類ではなく音源の方向に依存するため，空間相関行列 $\boldsymbol{G}_{fk}$ を音源ごとに独立な変数ではなく，方向ごとに独立な変数 $\boldsymbol{G}_{fd}$ として考える．このとき，式 (6) は音源 k の方向を示すベクトル $\boldsymbol{s}_k = [s_{k1}, \cdots, s_{kD}]^T$ を用いると次のように表される．

$$\boldsymbol{x}_{tf} \sim \prod_{k,d=1}^{K,D} \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \lambda_{tfk}\boldsymbol{G}_{fd}^{-1})^{z_{tfk}s_{kd}} \tag{7}$$

ただし，$\boldsymbol{s}_k$ は 1 of D 表現のベクトルであり，音源 k が方向 d にあるときは $s_{kd} = 1$，それ以外のときは 0 となる．

次に音源 k のパワー λ_{tfk} のモデルについて考える．提案法では NMF のように基底スペクトル w_{klf} とアクティベーション h_{klt} を用いて λ_{tfk} を表現する．ただし，NMF では $\lambda_{tfk} = \sum_l w_{klf}h_{klt}$ と全ての基底の和でパワーを表現するのに対し，提案法では各時間周波数ビンごとに一つの基底 l' を選択して $\lambda_{tfk} = w_{kl'f}h_{kl't}$ と表す．その基底を選択するための変数を $\boldsymbol{u}_{tfk} = [u_{tfk1}, \cdots, u_{tfkL}]^T$ とすると式 (7) は次のように表される．

$$\boldsymbol{x}_{tf} \sim \prod_{k,d,l=1}^{K,D,L} \mathcal{N}_{\mathbb{C}}\left(\boldsymbol{0}, w_{klf}h_{klt}\boldsymbol{G}_{fd}^{-1}\right)^{z_{tfk}s_{kd}u_{tfkl}} \tag{8}$$

ただし，$\boldsymbol{u}_{tfk}$ は 1 of L 表現のベクトルであり，基底 l が用いられるときは $u_{tfkl} = 1$，それ以外のときは 0 となる．

2.2　事前分布の設計

提案法では式 (8) のパラメータに対し，それぞれ適切な事前分布を与えることで推論を行う．ここではその事前分布の与え方について述べる．まず，$\boldsymbol{z}_{tf}$，$\boldsymbol{s}_k$，$\boldsymbol{u}_{tfk}$ はクラスタリングによる推論を行うためにカテゴリカル分布から生成されるとする：

$$\boldsymbol{z}_{tf} \mid \boldsymbol{\pi}_t \sim \mathrm{Categorical}(\boldsymbol{\pi}_t) \tag{9}$$

$$\boldsymbol{s}_k \mid \boldsymbol{\phi} \sim \mathrm{Categorical}(\boldsymbol{\phi}) \tag{10}$$

$$\boldsymbol{u}_{tfk} \mid \boldsymbol{\psi}_{tk} \sim \mathrm{Categorical}(\boldsymbol{\psi}_{tk}) \tag{11}$$

ここで，ハイパーパラメータ $\boldsymbol{\pi}_t$，$\boldsymbol{\phi}$，$\boldsymbol{\psi}_{tk}$ は観測に依存して変動するため推論を必要とする．したがって $\boldsymbol{\pi}_t$，$\boldsymbol{\phi}$，$\boldsymbol{\psi}_{tk}$ はカテゴリカル分布と共役なディリクレ分布から生成されるとする：

$$\boldsymbol{\pi}_t \sim \mathrm{Dirichlet}(a_0^{\pi}\boldsymbol{1}_K) \tag{12}$$

$$\boldsymbol{\phi} \sim \mathrm{Dirichlet}(a_0^{\phi}\boldsymbol{1}_D) \tag{13}$$

$$\boldsymbol{\psi}_{tk} \sim \mathrm{Dirichlet}(a_0^{\psi}\boldsymbol{1}_L) \tag{14}$$

ここで，$\boldsymbol{1}_N$ は要素が全て 1 の N 次元のベクトルとし，a_0^* はハイパーパラメータとする．また，空間相関行列 $\boldsymbol{G}_{fd}$，基底 w_{klf}，アクティベーション h_{klt} は式 (8) と共役になるように事前分布として次のような分布を与える．

$$\boldsymbol{G}_{fd} \sim \mathcal{W}_{\mathbb{C}}(\nu, \boldsymbol{G}_{fd}^0) \tag{15}$$

$$w_{klf} \sim \mathrm{Gamma}(a_0^w, b_0^w) \tag{16}$$

$$h_{klt} \sim \mathrm{Gamma}(a_0^h, b_0^h) \tag{17}$$

ここで，ν，a_0^*，b_0^* はハイパーパラメータであり，$\mathcal{W}_{\mathbb{C}}$ は複素ウィシャート分布 (付録参照) とする．

2.3 推　　論

提案法では，観測データ集合 $\boldsymbol{X}$ に対するすべてのパラメータの事後分布 $p(\boldsymbol{G},\boldsymbol{Z},\boldsymbol{S},\boldsymbol{U},\boldsymbol{\pi},\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{W},\boldsymbol{H}|\boldsymbol{X})$ を最大とするパラメータを求めることを目標とする．ただし，これらのパラメータを解析的に求めることは困難なので，提案法ではこれらのパラメータをギブスサンプリングにより求めることとする．ただし，事後分布に含まれるパラメータは $\boldsymbol{G}$，$\boldsymbol{Z}$，$\boldsymbol{S}$，$\boldsymbol{U}$，$\boldsymbol{W}$，$\boldsymbol{H}$，$\boldsymbol{\psi}$，$\boldsymbol{\pi}$，$\boldsymbol{\phi}$ の 9 つがあるが，提案法ではこのうちの $\boldsymbol{\pi}$，$\boldsymbol{\phi}$，$\boldsymbol{\psi}$ は積分消去を行い，残りのパラメータ $\boldsymbol{\Theta}=\{\boldsymbol{G},\boldsymbol{Z},\boldsymbol{S},\boldsymbol{U},\boldsymbol{W},\boldsymbol{H}\}$ を求めることとする．

ギブスサンプリングではそれぞれのパラメータの事後分布を求め，それらの事後分布からサンプリングを繰り返すことにより推定を行う．それぞれの事後分布は事前分布と尤度関数の積により求めることができ，以下のようになる．

$$\boldsymbol{G}_{fd} \mid \boldsymbol{X},\boldsymbol{\Theta}_{\neg \boldsymbol{G}_{fd}} \sim \mathcal{W}_{\mathbb{C}}(\nu'_{fd},\boldsymbol{G}'_{fd}) \tag{18}$$

$$\boldsymbol{z}_{tf} \mid \boldsymbol{X},\boldsymbol{\Theta}_{\neg \boldsymbol{z}_{tf}} \sim \mathrm{Categorical}(\boldsymbol{\pi}'_{tf}) \tag{19}$$

$$\boldsymbol{s}_{k} \mid \boldsymbol{X},\boldsymbol{\Theta}_{\neg \boldsymbol{s}_{k}} \sim \mathrm{Categorical}(\boldsymbol{\phi}'_{k}) \tag{20}$$

$$\boldsymbol{u}_{tfkl} \mid \boldsymbol{X},\boldsymbol{\Theta}_{\neg \boldsymbol{u}_{tfkl}} \sim \mathrm{Categorical}(\boldsymbol{\psi}'_{tfk}) \tag{21}$$

$$w_{klf} \mid \boldsymbol{X},\boldsymbol{\Theta}_{\neg w_{klf}} \sim \mathrm{GIG}(\gamma^w_{klf},\rho^w_{klf},\tau^w_{klf}) \tag{22}$$

$$h_{klt} \mid \boldsymbol{X},\boldsymbol{\Theta}_{\neg h_{klt}} \sim \mathrm{GIG}(\gamma^h_{klt},\rho^h_{klt},\tau^h_{klt}) \tag{23}$$

ここで，$\boldsymbol{\Theta}_{\neg *}$ は $\boldsymbol{\Theta}$ から $*$ の要素のみを除いた集合とする．また，GIG は一般化逆ガウス分布 (付録参照) を示す．ここで，ハイパーパラメータ ν'_{fd}，$\boldsymbol{G}'_{fd}$，$\boldsymbol{\pi}'_{tf}$，$\boldsymbol{\phi}'_{k}$，$\boldsymbol{\psi}$，$\boldsymbol{\gamma}^*$，$\boldsymbol{\rho}^*$，$\boldsymbol{\tau}^*$ は次のようになる．

$$\nu'_{fd} = \nu + \sum_{t,k=1}^{T,K} z_{tfk}s_{kd} \tag{24}$$

$$\boldsymbol{G}'^{-1}_{fd} = (\boldsymbol{G}^0_{fd})^{-1} + \sum_{t,k,l=1}^{T,K,L} \frac{\boldsymbol{x}_{tf}\boldsymbol{x}_{tf}^H}{w_{klf}h_{klt}} z_{tfk}s_{kd}u_{tfkl} \tag{25}$$

$$\pi'_{tfk} = \prod_{d,l=1}^{D,L} \left\{ \left| \frac{\boldsymbol{G}_{fd}}{w_{klf}h_{klt}} \right| \exp\left(-\frac{\boldsymbol{x}_{tf}^H\boldsymbol{G}_{fd}\boldsymbol{x}_{tf}}{w_{klf}h_{klt}} \right) \right\}^{s_{kd}u_{tfkl}} \times (a_0^{\pi} + n_{tk}^{\neg tf}) \tag{26}$$

$$\phi'_{kd} = \prod_{t,f,l=1}^{T,F,L} \left\{ \left| \frac{\boldsymbol{G}_{fd}}{w_{klf}h_{klt}} \right| \exp\left(-\frac{\boldsymbol{x}_{tf}^H\boldsymbol{G}_{fd}\boldsymbol{x}_{tf}}{w_{klf}h_{klt}} \right) \right\}^{z_{tfk}u_{tfkl}} \times (a_0^{\phi} + c_{d}^{\neg k}) \tag{27}$$

$$\psi'_{tfkl} = \prod_{d=1}^{D} \left\{ \left| \frac{\boldsymbol{G}_{fd}}{w_{klf}h_{klt}} \right| \exp\left(-\frac{\boldsymbol{x}_{tf}^H\boldsymbol{G}_{fd}\boldsymbol{x}_{tf}}{w_{klf}h_{klt}} \right) \right\}^{z_{tfk}s_{kd}} \times (a_0^{\psi} + n_{tkl}^{\neg tfk}) \tag{28}$$

$$\gamma^w_{klf} = a_0^w - Mn_{fkl}, \quad \rho^w_{klf} = b_0^w \tag{29}$$

$$\tau^w_{klf} = \sum_{t,d=1}^{T,D} \frac{\boldsymbol{x}_{tf}\boldsymbol{G}_{fd}\boldsymbol{x}_{tf}}{h_{klt}} z_{tfk}s_{kd}u_{tfkl} \tag{30}$$

$$\gamma^h_{klt} = a_0^h - Mn_{tkl}, \quad \rho^h_{klt} = b_0^h \tag{31}$$

$$\tau^h_{klt} = \sum_{f,d=1}^{F,D} \frac{\boldsymbol{x}_{tf}\boldsymbol{G}_{fd}\boldsymbol{x}_{tf}}{w_{klf}} z_{tfk}s_{kd}u_{tfkl} \tag{32}$$

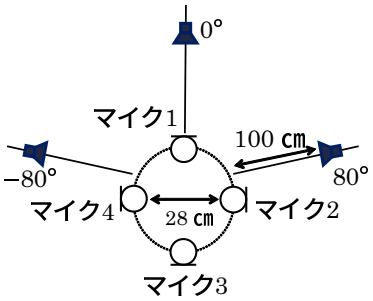


図 2　マイク配置

ここで，$n_{tk}^{\neg tf}$ は時刻 t 周波数 f でのサンプルを除いて，時刻 t において音源 k に割り当てられた時間周波数ビンの数を表し，$c_d^{\neg k}$ は音源 k を除いて方向 d に割り当てられた音源の数を表す．また，$n_{tkl}(n_{fkl})$ は時刻 t(周波数 f) において音源 k，基底 l に割り当てられた時間周波数ビンの数を示し，$n_{tkl}^{\neg tfk}$ は n_{tkl} から時刻 t 周波数 f で音源 k に割り当てられた要素を除いたものである．

2.4 分離音の生成

提案法では，時間周波数領域で音源方向ごとにマスクを推定することにより分離音を生成する．ギブスサンプリングで i 回目の試行により得られるサンプルを $\boldsymbol{W}^{(i)},\boldsymbol{H}^{(i)},\boldsymbol{G}^{(i)},\boldsymbol{Z}^{(i)},\boldsymbol{S}^{(i)},\boldsymbol{U}^{(i)}$ とすると，時刻 t 周波数 f での音源方向 d に対するマスク M^d_{tf} は次のようになる．

$$M^d_{tf} = \frac{1}{I}\sum_{i=1}^{I}\sum_{k=1}^{K} z^{(i)}_{tfk}s^{(i)}_{kd} \tag{33}$$

この M^d_{tf} を用いて次の式により方向 d の分離音を生成することができる．

$$\boldsymbol{x}^d_{tf} = M^d_{tf}\boldsymbol{x}_{tf} \tag{34}$$

3. 評価実験

提案法の分離性能を評価するため，シミュレーションにより混合した音を用いた実験を行った．比較手法として，IVA [11]，マルチチャネル NMF (MNMF) [12]，音源モデルをスパースとし空間モデルに LDA を用いた分離法 (Sparse-LDA) [8]，音源モデルに NMF，空間モデルに LDA を用いた分離法 (NMF-LDA) [9] を用いた．これに対し，提案法では音源モデル・空間モデルともに LDA を用いているためここでは提案法を LDA-LDA と呼ぶ．また，Sparse-LDA では音源数の同時推定も行うが，条件を対等にするため音源数は既知とした．

3.1 実験条件

図 2 に音源とマイクの配置を示す．残響時間 400 ms のインパルス応答を用いて 3 音源を混合した音声を用いた．マイク数は 4 とした．混合音には，音声のみの混合音と音楽のみの混合音，音声と音楽の混合音をそれぞれ 10 個ずつ使用した．用いる音楽と音声は SISEC [13] と JNAS の音素バランス文 [14] から選択した．サンプリング周波数は 16 kHz とし，STFT では窓幅 512 のハミング窓をシフト幅 256 で使用した．基底数 $L=20$ とし，ハイパーパラメータは，$\nu=M+1$，$a_0^{\pi}=a_0^{\phi}=10$，$a_0^{\psi}=a_0^w=a_0^h=b_0^w=b_0^h=1$ とした．また，

表 1　音楽による評価

	SDR	SIR	SAR
IVA	0.3 dB	4.9 dB	5.7 dB
MNMF	1.0 dB	6.2 dB	**6.7** dB
Sparse-LDA	0.7 dB	7.4 dB	4.1 dB
NMF-LDA	**1.2** dB	**9.6** dB	3.5 dB
LDA-LDA	1.1 dB	8.8 dB	3.7 dB

表 2　音声による評価

	SDR	SIR	SAR
IVA	3.4 dB	7.5 dB	7.1 dB
MNMF	4.8 dB	10.0 dB	**7.7** dB
Sparse-LDA	5.5 dB	15.1 dB	6.3 dB
NMF-LDA	4.9 dB	15.0 dB	5.7 dB
LDA-LDA	**5.9** dB	**16.8** dB	6.4 dB

表 3　音楽 + 音声による評価

	SDR	SIR	SAR
IVA	0.1 dB	5.3 dB	5.3 dB
MNMF	1.8 dB	8.6 dB	**6.1** dB
Sparse-LDA	2.4 dB	11.5 dB	4.5 dB
NMF-LDA	1.3 dB	10.6 dB	3.9 dB
LDA-LDA	**2.6** dB	**13.2** dB	4.2 dB

$G^0_{fd} = (\boldsymbol{a}_{fd}\boldsymbol{a}^H_{fd} + 0.01 \times \boldsymbol{I})^{-1}$ とし，$\boldsymbol{a}_{fd}$ には無響室で 5° 間隔で録音したインパルス応答を用いた．ギブスサンプリングの試行回数は 50 回とし，はじめの 30 回は burn-in として棄却した．評価尺度として，Signal-to-distortion ratio (SDR)，Signal-to-inference ratio (SIR)，Signal-to-artificial ratio (SAR) を用いた．SDR は総合的な分離性能，SIR は目的音以外の音の除去性能，SAR は分離音の歪みの少なさを表す尺度である．

3.2 実験結果

表 1，2，3 に実験結果を示す．それぞれの条件において最も数値が大きくなったものを太字で示した．SDR と SIR は音楽においては NMF-LDA が最も大きく，音声と音楽+音声においては LDA-LDA が最も大きくなった．このことから，より低ランク性の強い音源である音楽に対しては低ランク近似を用いる NMF-LDA の方が分離精度は高く，比較的低ランク性の弱い音源である音声などが含まれる混合音に対しては LDA-LDA の方が分離精度が高いことがわかった．また，Sparse-LDA と比較すると SDR と SIR はすべての条件において提案法の方が大きくなった．このことから提案法の音源モデルの有効性が確認できた．ただし，SAR は全ての条件において MNMF が最も大きくなった．また，Sparse-LDA，NMF-LDA，LDA-LDA では SAR にはそれほど大きな差はなかった．これは，LDA を用いた分離法では完全に排他的な割り当てが行われるのに対し，MNMF ではソフトな割り当てを行えることから自然なモデル化が可能となり分離音の歪みが小さくなったのではないかと推測される．したがって，SAR の改善を行うためには空間モデルの改善が必要であると考えられる．

4. おわりに

本稿では，空間モデルを音源の混合モデルで記述し，音源モデルを基底の混合モデルで記述したネスト型基底・音源混合モデルを用いた音源分離法を提案した．ネスト型基底・音源混合モデルを用いることで，位相情報だけでなく，音源のパワースペクトログラムの構造も考慮した音源分離を可能とした．実験の結果，音声などの比較的低ランクでない音源が含まれる混合音に対しては提案法により従来法よりも SDR や SIR が向上することを確認した．今後は，提案法のオンライン化や音源数・基底数の推定を行う．

付　　録

複素ウィシャート分布と一般化逆ガウス分布の確率密度関数は以下のとおりである．

$$\mathcal{W}_{\mathbb{C}}(\boldsymbol{G}|\nu, \boldsymbol{G}^0) = \frac{|\boldsymbol{G}|^{\nu-M}\exp(-\mathrm{tr}(\boldsymbol{G}(\boldsymbol{G}^0)^{-1}))}{|\boldsymbol{G}^0|^{\nu}\pi^{M(M-1)/2}\prod_{m=0}^{M-1}\Gamma(\nu-m)} \quad \text{(A·1)}$$

$$\mathrm{GIG}(y|\gamma,\rho,\tau) = \frac{\exp\{(\gamma-1)\log y - \rho y - \tau/y\}}{2\tau^{\gamma/2}\mathcal{K}_{\gamma}(2\sqrt{\rho\tau})} \quad \text{(A·2)}$$

ただし $\mathcal{K}_{\gamma}$ は第 2 種変形ベッセル関数である．

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

文　　献

[1] A. Hyvärinen *et al.*, Independent component analysis, John Wiley & Sons, 2004.

[2] I. Lee *et al.*, “Fast fixed-point independent vector analysis algorithms for convolutive blind source separation,” Signal Processing, vol.87, no.8, pp.1859–1871, 2007.

[3] A. Hyvarinen, “Fast and robust fixed-point algorithms for independent component analysis,” IEEE Transactions on Neural Networks, vol.10, no.3, pp.626–634, 1999.

[4] I. Jafari *et al.*, “On the use of the watson mixture model for clustering-based under-determined blind source separation,” INTERSPEECH, pp.988–992, 2014.

[5] N. Ito *et al.*, “Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors,” ICASSP, pp.3238–3242, 2013.

[6] L. Drude *et al.*, “Blind speech separation based on complex spherical k-mode clustering,” ICASSP, pp.141–145, 2016.

[7] H. Sawada *et al.*, “Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment,” TASLP, vol.19, no.3, pp.516–527, 2011.

[8] T. Otsuka *et al.*, “Bayesian nonparametrics for microphone array processing,” TASLP, pp.493–504, 2014.

[9] K. Itakura *et al.*, “A unified bayesian model of time-frequency clustering and low-rank approximation for multi-channel source separation,” EUSIPCO, to appear, 2016.

[10] P. Smaragdis *et al.*, “Non-negative matrix factorization for polyphonic music transcription,” WASPAA, pp.177–180, 2003.

[11] N. Ono, “Stable and fast update rules for independent vector analysis based on auxiliary function technique,” WASPAA, pp.189–192, 2011.

[12] H. Sawada *et al.*, “Multichannel extensions of non-negative matrix factorization with complex-valued data,” TASLP, vol.21, no.5, pp.971–982, 2013.

[13] S. Araki *et al.*, “The 2011 signal separation evaluation campaign (sisec2011):-audio source separation,” Latent Variable Analysis and Signal Separation, pp.414–422, Springer, 2012.

[14] K. Itou *et al.*, “The design of the newspaper-based japanese large vocabulary continuous speech recognition corpus,” ICSLP, pp.3261–3263, 1998.