ノンパラメトリックベイズに基づくパーミュテーションのない周波数領域でのブラインド音源分離

○柳楽浩平 大塚琢馬 奥乃博 (京都大学大学院)

1. はじめに

ロボット聴覚システム [1, 2] において音源分離システムは必要不可欠である [3, 4]．なぜなら，ロボットのマイクに入力される音は様々な音源からの混合音となるからである．例えば HARK[1] は音源分離，音源分離，分離音認識の機能を提供している．HARK を様々な環境で用いる場合，性能劣化の回避にはパラメータチューニングが必須である．

音源分離機能の実環境での利用のために満たすべき要件は以下の通りである．

1. 音声の混合過程が未知の状態での音源分離
2. 音源数未知状態での分離
3. 残響に対する頑健性

多くの音源分離手法は音源数や混合過程などの事前情報を必要としていた．これらの事前情報の入手は通常困難であるため，できる限り少ない事前情報での分離が求められるこのような分離手法をブラインド音源分離 (Blind source separation: BSS) と呼ぶ．

実環境での音源分離では，残響の混ざった音の分離が必要である．これはマイクには直接音の他に残響も同時に入力されるからである．残響の混ざった音の分離には周波数領域での処理が有効であり，数多くの周波数領域での BSS 手法が提案されている．周波数領域 BSS における問題の一つがパーミュテーション問題 [5] である．従来の周波数領域 BSS では各周波数帯域で独立に分離するため，各帯域での出力順序に曖昧性が生じる．周波数領域 BSS で分離信号を復元するためには，パーミュテーション問題の解決が必要不可欠である．

独立成分分析 (ICA) [6] は有名な BSS 手法である．周波数領域 ICA[7] は要件 1 と要件 3 を満たすが，ICA は音源数を事前に仮定している．つまり，要件 2 を満たさない．さらに，周波数領域 ICA ではパーミュテーション問題も生じてしまう．独立ベクトル分析（IVA）[8, 9] はパーミュテーション問題を回避した BSS 手法である．IVA は ICA に基づいており，音源数の仮定を置いている．つまり，要件 2 を満たさない．我々は以前，周波数領域 infinite sparse factor analysis (FD-ISFA) を提案した [10]．FD-ISFA は 3 つの要件すべてを満たすが，パーミュテーション問題の影響を受けてしまう．

本稿では以上の 3 要件を満たす分離とパーミュテーションの解決を同時に行う Permutation-free ISFA (PF-ISFA) を提案する．PF-ISFA はノンパラメトリックベイズに基づいており，音源数未知状況での BSS が可能である．本手法では分離とパーミュテーション解決を同時に行うために，全帯域統一の音源アクティビティを導入し，全周波数帯域を同時に処理する．

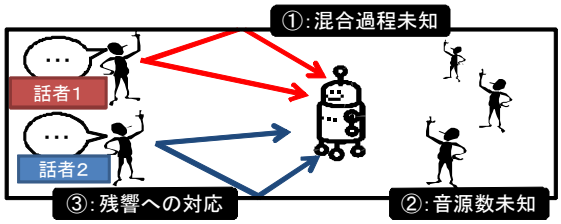

図 1 ロボット聴覚における音源分離の課題

2. 周波数領域での BSS

2·1 BSS の問題設定

BSS の問題設定は以下のようにまとめられる．

入力: D 本のマイクで観測される K 音源の混合音
出力: K 音源の分離信号とアクティビティ
仮定: 音源数はマイク数以上は存在しない

BSS は，D 本のマイクの観測信号から，K 個の音源の分離信号と各時間フレームでの音源のアクティビティを，音源やマイクの位置情報やマイクと音源の間の伝達過程などの事前知識なしで推定するという問題である．

2·2 周波数領域での処理

実環境での音声信号の混合仮定は畳み込み混合で表される．ロボットに搭載されているマイクで観測された信号は各音源からの信号の混合音であり，さらに各音源の反射音や残響，マイク間での信号の到来時間差などの影響を受ける．これらの時間遅れ信号をモデル化するために畳み込み混合モデルが用いられる．

$$\bar{\mathbf{x}}(t) = \sum_{j=0}^{J} \bar{\mathbf{A}}(j)\bar{\mathbf{s}}(t-j) \tag{1}$$

$\bar{\mathbf{x}}(t)$，$\bar{\mathbf{s}}(t)$，$\bar{\mathbf{A}}(j)$ はそれぞれ観測信号，音源信号，伝達関数の係数を表す．

畳み込み混合で表される信号の BSS 問題を解く時，短時間フーリエ変換（STFT）が用いられる．STFT を用いることで，時間領域での畳み込み混合問題は周波数領域での瞬時混合問題に変換される．

2·3 パーミュテーション問題

周波数領域での処理では各周波数帯域ごとに独立に処理を行う場合に分離信号の出力順序についてのパーミュテーションの曖昧性を解かなければならない．これをパーミュテーション問題と呼ぶ．パーミュテーション問題は周波数領域 BSS における有名な問題であり，周波数帯域間でのエンベロープの相関と到来方向推定を組み合わせた解法 [5] や，信号のパワー比を利用した解法 [11] など，様々な解法が提案されている．しかし，これまで画期的な解法はなく，パーミュテーション問題は今もなお活発に研究されている．

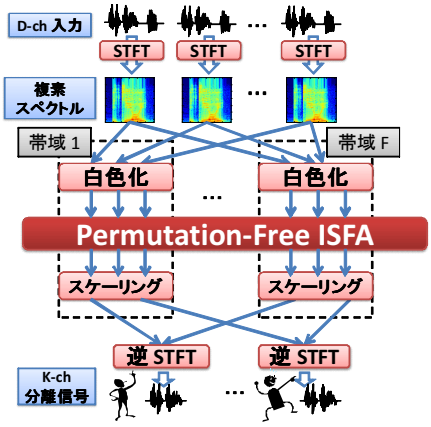

図 2 本手法の処理の流れ

3. Permutation-free ISFA

3·1 本手法の概要

図 2 に PF-ISFA の処理の流れを示す．入力信号に STFT を施した後，各周波数帯域ごとに複素スペクトルを白色化し，PF-ISFA を適用する．PF-ISFA の出力順序はすでに揃っているが，出力信号の振幅は元の音源の振幅と等しくない．これはスケーリング問題と呼ばれ，この問題も周波数領域の BSS の有名な問題の一つである．ここでは Projection back[12] を利用してスケーリング問題を解決する．その後，逆 STFT を施し，分離信号を出力する．

3·2 生成モデル・事前分布・尤度関数

K，D，F，T をそれぞれ音源数，マイク数，周波数帯域の数，信号の時間フレーム数とする．ある周波数帯域 f での瞬時混合モデルは以下のように表せる．

$$\mathbf{X}_f = \mathbf{A}_f(\mathbf{Z}_f \odot \mathbf{S}_f) + \mathbf{E}_f (f = 1, \cdots, F), \qquad (2)$$

ここで，$\mathbf{Z}_f = [\mathbf{z}_{f1}, \cdots \mathbf{z}_{fT}]$，$\mathbf{X}_f = [\mathbf{x}_{f1}, \cdots \mathbf{x}_{fT}]$，$\mathbf{S}_f = [\mathbf{s}_{f1}, \cdots \mathbf{s}_{fT}]$，$\mathbf{E}_f = [\varepsilon_{f1}, \cdots \varepsilon_{fT}]$ であり，$\mathbf{x}_{ft} = [x_{1ft}, x_{2ft}, \cdots, x_{Dft}]^{\mathrm{T}}$ は t フレーム目の混合信号ベクトル，$\mathbf{s}_{ft} = [s_{1ft}, s_{2ft}, \cdots, s_{Kft}]^{\mathrm{T}}$ は音源信号ベクトル，$\varepsilon_{ft} = [\varepsilon_{1ft}, \varepsilon_{2ft}, \cdots, \varepsilon_{Dft}]^{\mathrm{T}}$ はノイズ信号ベクトルを表す．また，$\mathbf{A}_f$ は $D \times K$ の混合行列，$\mathbf{z}_{ft} = [z_{1ft}, z_{2ft}, \cdots, z_{Kft}]^{\mathrm{T}}$ は f 番目の周波数帯域，時刻 t フレーム目での各音源のアクティビティを表す．音源アクティビティ z_{kft} は二値変数であり，時刻 t フレーム目，f 番目の帯域で音源 k が鳴っている場合には $z_{kft} = 1$，そうでない場合には $z_{kft} = 1$ となる．演算子 $\odot$ は要素ごとの積を表す．

PF-ISFA は F 組の周波数帯域を同時に処理する．$\mathbf{Z}$，$\mathbf{X}$，$\mathbf{S}$，$\mathbf{E}$，$\mathbf{A}$ はそれぞれ $\mathbf{Z} = [\mathbf{Z}_1, \cdots \mathbf{Z}_F]$，$\mathbf{X} = [\mathbf{X}_1, \cdots \mathbf{X}_F]$，$\mathbf{S} = [\mathbf{S}_1, \cdots \mathbf{S}_F]$，$\mathbf{E} = [\mathbf{E}_1, \cdots \mathbf{E}_F]$，$\mathbf{A} = [\mathbf{A}_1, \cdots \mathbf{A}_F]$ とする．

全周波数帯域のアクティビティを束ねるため，以下のようなモデルを導入する．

$$z_{kft} = b_{kt}\phi, \quad \phi \sim \mathrm{Bern}(\psi_{kf}), \qquad (3)$$

ここで，$\mathrm{Bern}(x)$ はパラメータ x の Bernoulli 分布を表す．b_{kt} は音源 k の t フレーム目での全帯域で統一の音源アクティビティを，ψ_{kf} は音源 k の f 番目の帯域のアクティベーション確率を表す．$\mathbf{B}$ は b_{kt} を，Ψ は ψ_{kf} をそれぞれまとめた行列である．

表 1PF-ISFA のパラメータ推定アルゴリズム

入力: 観測信号 $\mathbf{X}$, 出力: 音源信号 $\mathbf{S}$.

1. 各変数を事前分布を使って初期化.
2. 各時間フレーム t で以下を実行.
   - 2-1 各音源 k について，b_{kt} を式 (14) からサンプル.
   - 2-2 $b_{kt} = 1$ の場合，各帯域 f について z_{kft} を式 (12) からサンプル．そうでない場合は $z_{kft} = 0$.
   - 2-3 $z_{kft} = 1$ の場合，s_{kft} を式 (10) からサンプル．そうでない場合は $s_{kft} = 0$.
   - 2-4 時刻 t で初めて現れる音源数 κ_t を決め初期化.
3. 各音源 k 帯域 f ごとに ψ_{kf} を式 (16) からサンプル.
4. 各音源 k 帯域 f ごとに $\mathbf{a}_{kf}$ を式 (17) からサンプル.
5. 常にアクティブでない音源があれば取り除く.
6. σ_ε^2，σ_{A}^2，α を式 (18)，(19)，(20) を用いて更新.
7. 2. へ戻る.

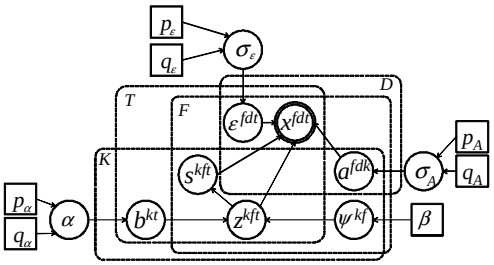

図 3PF-ISFA のグラフィカルモデル

各変数の事前分布は以下のように仮定する．

$$\varepsilon_{ft} \sim \mathcal{N}_C(0, \sigma_\varepsilon^2 \mathbf{I}),\ \sigma_\varepsilon^2 \sim \mathcal{IG}(p_\varepsilon, q_\varepsilon), \qquad (4)$$

$$s_{kft} \sim \mathcal{N}_C(0, 1), \qquad (5)$$

$$\mathbf{a}_{kf} \sim \mathcal{N}_C(0, \sigma_{\mathrm{A}}^2 \mathbf{I}),\ \sigma_{\mathrm{A}}^2 \sim \mathcal{IG}(p_A, q_A), \qquad (6)$$

$$\mathbf{B} \sim \mathrm{IBP}(\alpha),\ \alpha \sim \mathcal{G}(p_\alpha, q_\alpha), \qquad (7)$$

$$\Psi \sim \mathrm{Beta}(\beta/K, \beta(K-1)/K). \qquad (8)$$

$\mathbf{a}_{fk}$ は $\mathbf{A}_f$ の k 番目の行を表す．p_ε，q_ε，p_A，q_A，p_α，q_α，β はハイパーパラメータである．$\mathcal{N}_C$，$\mathcal{G}$，$\mathcal{IG}$ は一変数複素正規分布，ガンマ分布，逆ガンマ分布を，$\mathrm{IBP}(\alpha)$ は Indian buffet process (IBP)[13] を表す．

PF-ISFA は観測信号 $\mathbf{X}$ のみから，音源信号 $\mathbf{X}$，時間周波数アクティビティ $\mathbf{Z}$，混合行列 $\mathbf{A}$，全帯域の統一アクティビティ $\mathbf{B}$，各帯域のアクティベーション確率 Ψ，のすべてを推定する．

PF-ISFA のモデルの尤度関数は以下の通りである．

$$P(\mathbf{X}|\mathbf{A}, \mathbf{S}, \mathbf{Z}) = \prod_{f=1}^{F} \frac{1}{(\pi\sigma_\varepsilon^2)^{TD}} \exp\left(-\frac{\mathrm{tr}(\mathbf{E}_f^{\mathrm{H}}\mathbf{E}_f)}{\sigma_\varepsilon^2}\right). \qquad (9)$$

各データ点は独立同分布に従うと仮定している。

3·3 潜在変数の推定による音源分離

表 1 はアルゴリズムの概要を，図 3 は PF-ISFA のグラフィカルモデルを表す．このアルゴリズムは Metropolis-Hastings アルゴリズムに基づいている．各潜在変数の事後分布は事前分布と尤度からベイズの定理によって導かれる．

3·3·1 音源信号

z_{kft} がアクティブの時，s_{kft} は以下の事後分布からサンプルされる．

$$P(s_{kft}|\mathbf{A}_f, \mathbf{s}_{-kft}, \mathbf{x}_{ft}\mathbf{z}_{ft}) \propto P(\mathbf{x}_{ft}|\mathbf{A}_f, \mathbf{s}_{ft}, \mathbf{z}_{ft}, \sigma_\varepsilon^2) P(s_{kft})$$
$$= \mathcal{N}_C\left(s_{kft}; \mu_{s,f}, \sigma_{s,f}^2\right), \qquad (10)$$

$$\sigma_{s,f}^2 = \sigma_\varepsilon^2 / \left(\sigma_\varepsilon^2 + \mathbf{a}_{kf}^{\mathrm{H}}\mathbf{a}_{kf}\right), \mu_{s,f} = \mathbf{a}_{kf}^{\mathrm{H}}\varepsilon_{-kft} / \left(\sigma_\varepsilon^2 + \mathbf{a}_{kf}^{\mathrm{H}}\mathbf{a}_{kf}\right).$$

$\mathbf{s}_{-kft}$ は $\mathbf{s}_{ft}$ の s_{kft} 以外の要素を表し，ε_{-kft} は $\varepsilon|_{z_{kft}=0}$ を意味する．

3·3.2 音源の時間周波数アクティビティ

$b_{kt}=1$ の時，z_{kft} の事後分布は以下で求められる．

$$P(z_{kft}|b_{kt},\psi_{kf},z_{-kft},\mathbf{x}_{ft},\mathbf{s}_{ft},\mathbf{A}_f)\propto P_pP_l \qquad (11)$$

ここで，

$$P_l=P(x_{ft}|\mathbf{A}_f,\mathbf{s}_{ft},\mathbf{z}_{ft},\sigma_\varepsilon^2),\ P_p=P(z_{kft}|b_{kt},\psi_{kf})$$

はそれぞれ尤度項と事前確率である．計算すると以下のようになる．

$$P(z_{kft}|b_{kt},\psi_{kf},z_{-kft},\mathbf{x}_{ft},\mathbf{s}_{ft},\mathbf{A}_f)=\ \mathrm{Bern}\left(p_1/(p_0+p_1)\right), \qquad (12)$$

$$\log(p_1)=\log(\psi_{kf})+\frac{2\mathrm{Re}(s_{kft}^*\mathbf{a}_{kf}^H\varepsilon_{-kft})+|s_{kft}|^2\mathbf{a}_{kf}^H\mathbf{a}_{kf}}{\sigma_\varepsilon^2}$$

$$\log(p_0)=\log(1-\psi_{kf})$$

3·3.3 全帯域統一のアクティビティ

b_{kt} がアクティブとなる確率とそうでない確率の比は式 (13) で計算される．この比 r は事前分布の比 r_p と帯域 f での尤度の比 $r_{l,f}$ に分けられる．

$$r=\frac{P(b_{kt}=1|\mathbf{A},\mathbf{S}_{-kt},\mathbf{X}_t,\mathbf{S}_{-kt})}{P(b_{kt}=0|\mathbf{A},\mathbf{S}_{-kt},\mathbf{X}_t,\mathbf{Z}_{-kt})}=r_p\prod_{f=1}^{F}r_{l,f}. \qquad (13)$$

$$r_p=\frac{P(b_{kt}=1|\mathbf{b}_{kt})}{P(b_{kt}=0|\mathbf{b}_{kt})}=\frac{m_{k,-t}}{T-m_{k,-t}}$$

$$r_{l,f}=\frac{P(\mathbf{x}_{ft}|\mathbf{A}_f,\mathbf{s}_{-kft},\mathbf{x}_{ft},\mathbf{z}_{-kft},\mathbf{b}_{-kt},b_{kt}=1,\psi_{kf},\sigma_\varepsilon^2)}{P(\mathbf{x}_{ft}|\mathbf{A}_f,\mathbf{s}_{-kft},\mathbf{x}_{ft},\mathbf{z}_{-kft},\mathbf{b}_{-kt},b_{kt}=0,\psi_{kf},\sigma_\varepsilon^2)}$$

$$=\psi_{kf}\sigma_{s,f}^2\exp\left(|\mu_{s,f}|^2/\sigma_{s,f}^2\right)+(1-\psi_{kf}).$$

$\mathbf{X}_t$ は $\mathbf{x}_{1t},\cdots,\mathbf{x}_{Ft}$ を，$\mathbf{S}_{-kt}$ と $\mathbf{Z}_{-kt}$ はそれぞれ $\mathbf{S}$ と $\mathbf{Z}$ から $s_{k1t},\cdots,s_{kFt}$ と $z_{k1t},\cdots,z_{kFt}$ を取り除いたものである．ここで，$m_{k,-t}=\sum_{t'\neq t}b_{kt'}$ である．これは IBP に基づく音源アクティビティの事前分布から導出される [13]．

$b_{kt}=1$ となる事後確率は比 r を使って計算される．

$$P(b_{kt}=1|\mathbf{A},\mathbf{S}_{-kt},\mathbf{X}_t,\mathbf{Z}_{-kt},\mathbf{b}_{-kt})=r/(1+r) \qquad (14)$$

b_{kt} の値を決めるため，Uniform$(0,1)$ から u をサンプルし，$r/(1+r)$ と比較する．$u\leq r/(1+r)$ なら $b_{kt}=1$ に，そうでないなら $b_{kt}=0$ とする．

3·3.4 新しい音源の数

各音源は必ずしも初めから存在するわけではなく，時刻 t から現れる音源も存在する．このような音源の数を κ_t とする．κ_t の事前分布は $P(\kappa_t|\alpha)=\mathrm{Poisson}\left(\frac{\alpha}{T}\right)$ であり，κ_t をサンプルした後，新しい音源のアクティビティと音源信号の初期化を行う．次に，この更新を受理するかどうかを決定する．その受理確率は $\min(1,r_{\xi\to\xi^*})$ となる．Meeds ら [14] と Knowles ら [15] によると，$r_{\xi\to\xi^*}$ は現状態での尤度と次状態での尤度の比となり，$\mathbf{A}_f^*$ は $\mathbf{A}_f$ の追加部分を表す $D\times\kappa_t$ 行列とすると，比は以下のようになる．

$$r_{\xi\to\xi^*}=\prod_{f=1}^{F}(\det\Lambda_{\xi,f})^{-1}\exp\left(\mu_{\xi,f}^{\mathrm{H}}\Lambda_{\xi,f}\mu_{\xi,f}\right), \qquad (15)$$

ここで，

$$\Lambda_{\xi,f}=\mathbf{I}+\mathbf{A}_f^{*\mathrm{H}}\mathbf{A}_f^*/\sigma_\varepsilon^2,\ \ \Lambda_{\xi,f}\mu_{\xi,f}=\mathbf{A}_f^{*\mathrm{H}}\varepsilon_{ft}/\sigma_\varepsilon^2.$$

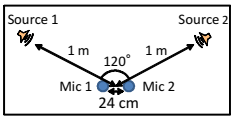

図 4 音源とマイクの位置関係

表 2 実験条件

音源数 K	2
マイク数 D	2
サンプリング周波数	16 kHz
STFT 窓幅	64 ms
STFT シフト幅	32 ms
反復回数	300 回

3·3.5 各帯域のアクティベーション確率

ψ_{kf} は以下の事後分布からサンプルされる．

$$P(\psi_{kf}|\mathbf{z}_{kf},\Psi_{-kf},\mathbf{B}_{-kt})\propto P(\psi_{kf}|\beta)\prod_{t=1}^{T}P(z_{kft}|\psi_{kf},b_{kt})$$

$$=\mathrm{Beta}\left(\beta/K+n_{kf},\beta(K-1)/K+m_k-n_{kf}\right), \qquad (16)$$

$n_{kf}=\sum_{t=1}^{T}z_{kft}$ f 番目の帯域で音源 k がアクティブになるフレーム数を，$m_k=\sum_{t=1}^{T}b_{kt}$ は音源 k がアクティブになるフレーム数を表す．

3·3.6 混合行列

混合行列は各音源ごとに推定する．

$$P(\mathbf{a}_{kf}|\mathbf{A}_{f,-k},\mathbf{S}_f,\mathbf{X}_f,\mathbf{Z}_f)\propto P(\mathbf{X}_f|\mathbf{A}_f,\mathbf{S}_f,\mathbf{Z}_f)P(\mathbf{a}_{kf}|\sigma_{\mathbf{A}}^2)$$

$$=\mathcal{N}_C(\mathbf{a}_{kf};\mu_{\mathbf{A}},\Lambda_{\mathbf{A}}^{-1}), \qquad (17)$$

ここで，

$$\Lambda_{\mathbf{A}}=\left(\frac{\mathbf{s}_{kf}^{\mathrm{H}}\mathbf{s}_{kf}}{\sigma_\varepsilon^2}+\frac{1}{\sigma_{\mathbf{A}}^2}\right)\mathbf{I}_D,\mu_{\mathbf{A}}=\frac{\sigma_{\mathbf{A}}^2}{\mathbf{s}_{kf}^{\mathrm{H}}\mathbf{s}_{kf}\sigma_{\mathbf{A}}^2+\sigma_\varepsilon^2}\mathbf{E}_f|_{\mathbf{a}_{kf}=0}\mathbf{s}_{kf}.$$

3·3.7 雑音と混合行列の分散

雑音の分散は推定信号のノイズレベルに，混合行列の分散は推定信号の振幅に影響する．

$$P(\sigma_\varepsilon^2|\mathbf{E})\propto P(\mathbf{E}|\sigma_\varepsilon^2)P(\sigma_\varepsilon^2|p_\varepsilon,q_\varepsilon)$$

$$=\mathcal{IG}\left(\sigma_\varepsilon^2;p_\varepsilon+FTD,\frac{q_\varepsilon}{(1+q_\varepsilon\sum_{f=1}^{F}\mathrm{tr}(\mathbf{E}_f^{\mathrm{H}}\mathbf{E}_f))}\right). \qquad (18)$$

$$P(\sigma_{\mathbf{A}}^2|\mathbf{A})\propto P(\mathbf{A}|\sigma_{\mathbf{A}}^2)P(\sigma_{\mathbf{A}}^2|p_{\mathbf{A}},q_{bfA})$$

$$=\mathcal{IG}\left(\sigma_{\mathbf{A}}^2;p_{\mathbf{A}}+FDK,\frac{q_{\mathbf{A}}}{1+q_{\mathbf{A}}\sum_{f=1}^{F}\mathrm{tr}(\mathbf{A}_f^{\mathrm{H}}\mathbf{A}_f)}\right). \qquad (19)$$

3·3.8 IBP パラメータ

IBP のパラメータ α の事後分布は以下のようになる，

$$p(\alpha|\mathbf{B})\propto P(\mathbf{B}|\alpha)P(\alpha|p_\alpha,q_\alpha)$$

$$=\mathcal{G}\left(\alpha;K_++p_\alpha,q_\alpha/\left(1+q_\alpha H_T\right)\right). \qquad (20)$$

K_+ はアクティブな音源数を，$H_n=\sum_{j=1}^{n}\frac{1}{j}$ は n 番目の調和級数を表す．

4. 実験結果

本手法の分離性能を音声信号の分離実験によって確認する．ここではベース手法である FD-ISFA[10] と比較する．実験では無響室残響，会議室残響 ($\mathrm{RT}_{60}=460$ ms) の 2 種類の残響環境での信号を用いた．図 4 はマイクと音源の位置関係を表しており，実験条件は表 2 の通りである．実験では JNAS 音素バランス文から 200 発話分を用いた．

まず無響室残響での混合信号を用いた分離実験の結果の一例を示す．図 5–8 はそれぞれ音源信号，PF-ISFA による分離信号，FD-ISFA による分離信号，元音声を利用して FD-ISFA で分離した信号のパーミュテーションを解いた信号のスペクトログラムを表す．

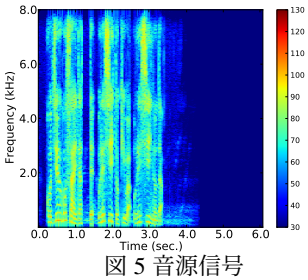

図 5 音源信号

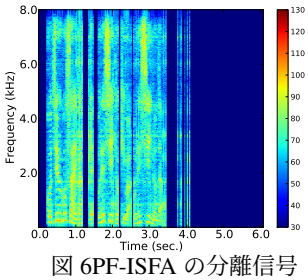

図 6PF-ISFA の分離信号

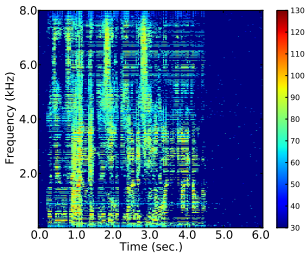

図 7 FD-ISFA のパーミュテーション解決前の分離信号

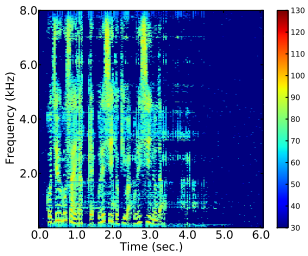

図 8 FD-ISFA の分離後にパーミュテーションを揃えた信号

図 7 の FD-ISFA の分離結果には数多くの横線が見られ，図 8 のパーミュテーション解決後ではその数は減少している．これらの横線は分離後の他の音源信号のスペクトログラムである．つまり，FD-ISFA では各帯域ごとの出力順序が揃っていないことが分かる．これに対し，図 6 の PF-ISFA の分離結果には横線は見られない．つまり，PF-ISFA の出力順序は全帯域で揃っており，PF-ISFA によってパーミュテーション問題が解決されたと言える．

また，Signal to Distortion Ratio (SDR)，Image to Spatial distortion Ratio (ISR)，Source to Interference Ratio (SIR)，Source to Artifacts Ratio (SAR) [16] を用いた評価も行った．結果は表 3 の通りである．“Non-Perm” は分離結果信号そのままを，“Perm” は元音声を用いて分離信号のパーミュテーションを解決した信号を表す．PF-ISFA は全条件，全指標で FD-ISFA を上回っている．特に SIR について，PF-ISFA は FD-ISFA と比較して無響室残響で 14.45 dB，会議室残響で 5.46 dB 改善した．

FD-ISFA の性能劣化の原因の一つがパーミュテーション問題であることは，分離結果に対するパーミュテーション解決後の信号の性能の改善からも分かる．これに対して，PF-ISFA ではその差は小さい．つまり PF-ISFA はパーミュテーション問題を解決している．

会議室残響での分離性能が無響室残響での分離性能に比べて低下している．これは会議室での残響 ($RT_{60} = 460$ ms) が STFT の窓幅 (64 ms) より長いことが原因である．残響時間が STFT の窓幅より長くなると，残響が複数の時間フレームに影響し分離性能が低下する．

5. 結論と今後の課題

本稿ではパーミュテーション問題を解決する周波数領域での新しい BSS 手法 PF-ISFA を提案した．本手法はノンパラメトリックベイズの手法に基づいている．全周波数帯域での出力順序を自動的に揃えるため，全帯域で統一な音源アクティビティを導入した．SIR による評価では，無響室残響，会議室残響 ($RT_{60} = 460$ ms) ともに PF-ISFA は FD-ISFA を上回る結果が得られた．

表 3 平均分離性能 [dB]

	無響室残響			
	FD-ISFA		**PF-ISFA**	
	Non-Perm	Perm	Non-Perm	Perm
SDR	0.38	11.96	**10.26**	**12.59**
ISR	4.98	18.23	**15.96**	**18.75**
SIR	1.38	18.58	**15.83**	**19.20**
SAR	5.22	14.39	**13.91**	**15.16**

	会議室残響 ($RT_{60} = 460$ ms)			
	FD-ISFA		**PF-ISFA**	
	Non-Perm	Perm	Non-Perm	Perm
SDR	0.35	5.85	**3.56**	**5.31**
ISR	4.73	10.41	**8.08**	**9.88**
SIR	1.12	9.86	**6.58**	**9.22**
SAR	5.72	10.36	**9.30**	**10.36**

今後の課題として，音源アクティビティの評価と音区間検出 (VAD) への応用があげられる．また，残響時間が長い状況でも分離が可能な手法の開発も求められる．さらに，ロボットへの応用を考慮すると，実時間処理を行えるように処理速度の改善が不可欠である．

謝辞本研究の一部は科研費 (S)，HRI-JP の支援を受けた．

参考文献

[1] K. Nakadai et al. Design and Implementation of Robot Audition System ”HARK” Open Source Software for Listening to Three Simultaneous Speakers. *Advanced Robotics*, 24(5–6):739–761, 2010.
[2] H. Sawada et al. Improvement of speech recognition performance for spoken-oriented robot dialog system using end-fire array. In *IROS 2010*, pages 970–975. IEEE, 2010.
[3] H. Saruwatari et al. Two-stage blind source separation based on ica and binary masking for real-time robot audition system. In *IROS 2005*, pages 2303–2308. IEEE, 2005.
[4] T. Mizumoto et al. Design and implementation of selectable sound separation on the texai telepresence system using HARK. In *ICRA 2011*, pages 2130–2137. IEEE, 2011.
[5] H. Sawada et al. A robust and precise method for solving the permutation problem of frequency-domain blind source separation. *IEEE Trans. on Speech and Audio Processing*, 12(5):530–538, 2004.
[6] A. Hyvärinen et al. *Independent component analysis*. Wiley-Interscience, 2001.
[7] H. Sawada et al. Polar coordinate based nonlinear function for frequency-domain blind source separation. *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP'02)*, pages 1001–1004, 2002.
[8] I. Lee et al. Fast fixed-point independent vector analysis algorithms for convolutive blind source separation. *Signal Processing*, 87(8):1859–1871, 2007.
[9] A. Hiroe. Solution of permutation problem in frequency domain ICA, using multivariate probability density functions. *Independent Component Analysis and Blind Signal Separation*, pages 601–608, 2006.
[10] K. Nagira et al. Complex extension of infinite sparse factor analysis for blind speech separation. *Latent Variable Analysis and Signal Separation*, pages 388–396, 2012.
[11] H. Sawada et al. Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS. In *Circuits and Systems, 2007. ISCAS 2007. IEEE International Symposium on*, pages 3247–3250. IEEE, 2007.
[12] N. Murata et al. An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing*, 41(1-4):1–24, 2001.
[13] T. Griffiths et al. Infinite latent feature models and the Indian buffet process. *Advances in Neural Information Processing Systems*, 18:475–482, 2006.
[14] E. Meeds et al. Modeling dyadic data with binary latent factors. *Advances in Neural Information Processing Systems*, 19:977–984, 2007.
[15] D. Knowles et al. Infinite sparse factor analysis and infinite independent components analysis. *Independent Component Analysis and Signal Separation*, pages 381–388, 2007.
[16] E. Vincent et al. First stereo audio source separation evaluation campaign: data, algorithms and results. *Independent Component Analysis and Signal Separation*, pages 552–559, 2007.