社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B102-9

ブラインド音源分離のための Infinite Sparse Factor Analysis の複素拡張

Complex Extension of Infinite Sparse Factor Analysis for Blind Source Separation of Speech Signals

柳楽浩平 **Kohei NAGIRA**　**高橋徹** **Toru TAKAHASHI**　**尾形哲也** **Tetsuya OGATA**　**奥乃博** **Hiroshi G. OKUNO**

京都大学大学院 情報学研究科
Graduate School of Informatics, Kyoto University
{knagira, tall, ogata, okuno}@kuis.kyoto-u.ac.jp

Abstract

We present a method of blind source separation (BSS) for speech signals using a complex extension of infinite sparse factor analysis (ISFA) in the frequency domain. In real environment, microphone array embedded in robot captures sound mixture contaminated by delayed signals (i.e. reflections, short-time reverberations, and time lags of signals arriving at microphones). Our method achieves robust separation of sound mixture that contains such delayed signals. Our method uses complex normal distributions to estimate source signals and mixing matrix. Experimental results indicate that our method outperforms the conventional ISFA and in the average signal-to-distortion ratio (SDR).

1 はじめに

音声信号のブラインド音源分離は遠距離音声認識[Wölfel and McDonough, 2009; Seltzer *et al.*, 2004]やロボット聴覚システム[Nakadai *et al.*, 2010; Valin *et al.*, 2004] などの様々な領域で応用されており，それゆえに活発な研究トピックの一つとなっている．実環境においては，マイクからのシステムへの入力信号は複数話者の混合音声となり，さらに反射音や残響なども同時に入力される．このような混合信号からそれぞれの話者の音声を認識するために，混合音を分離する必要がある．

音声信号の音源分離に対する主な要求条件は以下の通りである．

要件 1. 事前情報を用いない分離

要件 2. アクティビティの同時推定

要件 3. 時間遅れ信号に対する頑健性

音源位置やマイク配置などの事前情報を用いない音源分離はブラインド音源分離[Belouchrani *et al.*, 1997]と呼ばれる．独立成分分析 (Independent component analysis: ICA) [Hyvärinen *et al.*, 2001] はブラインド音源分離によく利用される手法である．実環境下でのブラインド音源分離を達成する手法としてよく用いられるものに周波数領域の ICA[Sawada *et al.*, 2002]があるが，各音源のアクティビティの推定は行わないため，要件 2 を満たさない．

Infinite sparse factor analysis (ISFA) [Knowles and Ghahramani, 2007] はノンパラメトリックベイズに基づいたブラインド音源分離手法である．ISFA は音源分離と音源のアクティビティの同時推定を行うため，要件 1，2 を満たす．しかしながら従来の ISFA では反射音や残響，各信号のマイクへの到来時間差などの時間遅れ信号を含んだ混合音声をモデル化しておらず分離が困難であるため，要件 3 を満たさない．

我々の研究の目的は以上 3 つの要求を満たすブラインド音源分離システムの開発である．本稿では ISFA の複素拡張を用いて，これらの要求条件を満たす BSS 手法を提案する．

2 ISFA によるブラインド音源分離

本章では本稿で取り上げる問題を明らかにし，従来の ISFA について説明したのち，解決すべき問題点について述べる．

2.1 ブラインド音源分離の問題設定

本稿で扱うブラインド音源分離問題を要約すると以下のようになる．

入力: D 本のマイクに入力される K 音源の混合信号
出力: 元の K 個の音源信号とそれらのアクティビティ
仮定: $K \leq D$
残響時間は短時間フーリエ変換 (Short time Fourier transform: STFT) の窓幅より短い

D 個のマイクを用いてシステムに K 個の音源からの混合信号を入力し，音源方向やインパルス応答などの事前情報を用いずに元の K 個の音源信号を分離して出力する．

2.2 音声信号のブラインド音源分離

ここで，音声信号がマイクに入力される際の音声の混合過程について述べる．音源とマイクの間には距離があるので，音源から生じた音はすぐにマイクに届くのではなく，距離の分だけ遅延して入力される．また，壁などで反射した後にマイクに届く音などの間接音も同時に入力される．つまり，複数音源からの音声が混合される際，各音源からの音声信号それぞれの直接音及び間接音が同時に入力されることになる．このような混合過程は次の式のような時間領域での畳み込み混合モデルとして定式化できる．

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

ここで，t は時刻を表し，$\overline{\mathbf{x}}(t)$，$\overline{\mathbf{s}}(t)$，$\overline{\mathbf{A}}(j)$ はそれぞれ観測信号，音源信号，伝達関数を表す．J は残響時間を意味しており，本稿では J が STFT の窓幅よりも短いことを仮定している．無響室などでの混合音声の場合この仮定が満たされるが，一般的な部屋での混合信号は必ずしもこれを満たさない．

畳み込み混合信号のブラインド音源分離問題を解く際には STFT がよく利用される．STFT により，式 (1) は以下のような式に変換される．

$$\mathbf{x}(f,t) = \mathbf{A}(f,t)\mathbf{s}(f,t) \tag{2}$$

f は周波数帯域のインデックスである．つまり，時間領域での畳み込み混合が周波数領域の瞬時混合に変換できる．この変換で，変換前は実数信号であったのに対し、変換後では複素信号を扱う必要が生じる．STFT を施したのちに，各周波数ごとに独立に分離処理を行い，分離結果に対して逆 STFT を施し元の音声信号を復元する．

2.3 従来の ISFA

Infinite sparse factor analysis [Knowles and Ghahramani, 2007]はノンパラメトリックベイズに基づくブラインド音源分離手法である．ここでは ISFA のモデルについて述べる．

はじめに，ISFA の混合モデルについて説明する．K，D，N をそれぞれ音源数，マイクの数，音源信号の長さとする．瞬時混合モデルは以下のように表される．

$$\mathbf{X} = \mathbf{A}(\mathbf{Z} \odot \mathbf{S}) + \mathbf{E}, \tag{3}$$

ここで，$\mathbf{Z} = [\mathbf{z}_1, \cdots \mathbf{z}_N]$，$\mathbf{X} = [\mathbf{x}_1, \cdots \mathbf{x}_N]$，$\mathbf{S} = [\mathbf{s}_1, \cdots \mathbf{s}_N]$, $\mathbf{E} = [\varepsilon_1, \cdots \varepsilon_N]$, $\mathbf{x}_t = [x_{1t}, x_{2t}, \cdots, x_{Dt}]^{\mathrm{T}}$ は時刻 t での混合信号ベクトル，$\mathbf{s}_t = [s_{1t}, s_{2t}, \cdots, s_{Kt}]^{\mathrm{T}}$ は音源信号ベクトル，$\varepsilon_t = [\varepsilon_{1t}, \varepsilon_{2t}, \cdots, \varepsilon_{Dt}]^{\mathrm{T}}$ はガウス性雑音のベクトルとする．また，$\mathbf{A}$ は $D \times K$ の混合行列，$\mathbf{z}_t = [z_{1t}, z_{2t}, \cdots, z_{Kt}]^{\mathrm{T}}$ は時刻 t での各音源のアクティビティを表す．z_{kt} 二値の変数であり，音源 k が時刻 t で音が鳴っている場合は $z_{kt} = 1$ となり，そうでない場合は $z_{kt} = 0$ となる．演算子 $\odot$ は要素ごとの積を表している．ISFA は観測信号 $\mathbf{X}$ のみを用いて音源信号 $\mathbf{S}$ とそれらのアクティビティ$\mathbf{Z}$，混合行列 $\mathbf{A}$，その他のパラメータを同時に推定する．

2.4 従来法の問題点

従来の ISFA[Knowles and Ghahramani, 2007]では複素数を扱えないため，STFT によって得られる混合音声の複素スペクトルに対して従来の ISFA を適用できず，畳み込み混合信号の分離ができない．これは音声信号のブラインド音源分離を行うにあたって解決すべき主要な問題の一つである．なぜなら，上記の通り音声信号の混合過程は伝達関数の畳み込みを用いて表されるからである．

3 ISFA の複素拡張

時間遅れ信号を含んだ混合信号を分離するために，周波数領域で ISFA を用いることを考える．我々の従来手法[柳楽ら, 2011]では，入力信号の実部と虚部を別々に実数 ISFA に入力していたが，実部と虚部の統合の際に別音源の実部と虚部が統合される可能性があり，推定精度が低下するという問題点があった．本稿では ISFA 自身を複素信号を扱えるように拡張することで周波数領域での ISFA を実現する．

Table 1 は本手法の推論アルゴリズムである．本手法は Metropolis-Hastings アルゴリズムと Gibbs サンプリングに基づいている．ベイズの定理から，潜在変数の事後分布は事前分布と尤度関数の積から得られる．以下では，各パラメータの事前分布とこのモデルの尤度関数を示し，それぞれの事後分布について述べる．

3.1 事前分布

各変数の事前分布は以下の通りである．

$$\varepsilon_t \sim \mathcal{N}_C(0, \sigma_\varepsilon^2 \mathbf{I}) \qquad \sigma_\varepsilon^2 \sim \mathcal{IG}(p_1, p_2), \tag{4}$$

$$s_{kt} \sim \mathcal{N}_C(0, 1), \tag{5}$$

$$\mathbf{a}_k \sim \mathcal{N}_C(0, \sigma_{\mathbf{A}}^2 \mathbf{I}) \qquad \sigma_{\mathbf{A}}^2 \sim \mathcal{IG}(p_3, p_4), \tag{6}$$

$$\mathbf{Z} \sim \mathrm{IBP}(\alpha) \qquad \alpha \sim \mathcal{G}(p_5, p_6). \tag{7}$$

ここで，$\mathbf{a}_k$ は $\mathbf{A}$ の k 番目の列，p_1，p_2，p_3，p_4，p_5，p_6 はハイパーパラメータである．$\mathcal{N}_C(\mu, \sigma^2)$ は平均 μ，分散 σ^2 の一変量複素正規分布を表す．$\mathcal{G}(b, \theta)$ と $\mathcal{IG}(b, \theta)$ は形状母数 b，尺度母数 θ のガンマ分布と逆ガンマ分布を表す．それぞれの分布の確率密度関数は以下のようになって

Table 1: Algorithm for estimating model parameters of complex ISFA

1. 混合行列 $\mathbf{A}$，音源のアクティビティ $\mathbf{Z}$，音源信号 $\mathbf{S}$ を事前分布を元に初期化
2. 各時刻 t について以下を実行
   - 2-1 各音源 k ごとに式 (17) をもとに z_{kt} をサンプル
   - 2-2 $z_{kt}=1$ なら式 (13) から s_{kt} をサンプルそうでない場合は $s_{kt}=0$
   - 2-3 この時刻で初めて active になる音源の数 κ_t を決め，初期化
3. 各音源 k ごとに混合行列 $\mathbf{a}_k$ を式 (21) からサンプル
4. 全時刻通して inactive になっている音源があれば除去
5. σ_ε^2，$\sigma_{\mathbf{A}}^2$，α を式 (22)，(23)，(24) をもとに更新
6. 2 へ戻る

いる．

$$\mathcal{N}_C(x;\mu,\sigma^2) = \frac{1}{\pi\sigma^2}\exp\left(-\frac{|x-\mu|^2}{\sigma^2}\right), \quad (8)$$

$$\mathcal{G}(x;b,\theta) = \frac{x^{b-1}}{\Gamma(b)\,\theta^b}\exp\left(-\frac{x}{\theta}\right), \quad (9)$$

$$\mathcal{IG}(x;b,\theta) = \frac{x^{-(b-1)}}{\Gamma(b)\,\theta^b}\exp\left(-\frac{1}{\theta x}\right). \quad (10)$$

IBP(α) はパラメータ α の Indian buffet process (IBP) [Griffiths and Ghahramani, 2006] を表す．IBP は潜在的に無限個の音源を扱うことができる確率過程である．IBP の概要は以下のように表される．

1. 時刻 $t=1$ において
   初めから鳴っている音源の数を Poisson(α) からサンプリングする．
2. 時刻 $t=i$ において
   - 音源 k は確率 $\frac{m_k}{i}$ で active になる．ここで m_k は時刻 $t=1$ から $i-1$ までで音源 k が active になった時間の数を表す．
   - 既存の音源が active かどうかを決定した後，時刻 i で始めて active になる音源の数を Poisson($\frac{\alpha}{i}$) からサンプリングする．

α は母数 α のポアソン分布を表す．IBP にはサンプル順序の交換可能性があり，注目している時刻 t が最後にサンプルされると考えてよい．つまり，時刻 t 以外のアクティビティが与えられた状態で時刻 t のアクティビティを推定できる．これより，IBP に基づくアクティビティの事前分布は以下のようになる．

$$P(z_{kt}|\mathbf{z_{-kt}}) = \frac{m_{k,-t}}{N} \quad (11)$$

ただし，$m_{k,-t}=\sum_{s\neq t} z_{ks}$ を表し，$\mathbf{z}_{-kt}$ は $\mathbf{z}_t$ の要素のうち z_{kt} を取り除いたものを表す．

3.2 尤度関数

複素 ISFA の尤度関数は以下のように表される．

$$\begin{aligned} P(\mathbf{X}|\mathbf{A},\mathbf{S},\mathbf{Z}) &= \prod_{t=1}^{N} P(\mathbf{x}_t|\mathbf{A},\mathbf{s}_t,\mathbf{z}_t) \\ &= \prod_{t=1}^{N} \mathcal{N}_C(\mathbf{x}_t;\mathbf{A}(\mathbf{z}_t\odot\mathbf{s}_t),\sigma_\varepsilon^2\mathbf{I}) \\ &= \frac{1}{(\pi\sigma_\varepsilon^2)^{ND}}\exp\left(-\frac{\mathrm{tr}(\mathbf{E}^{\mathrm{H}}\mathbf{E})}{\sigma_\varepsilon^2}\right), \end{aligned} \quad (12)$$

ここで，

$$\mathbf{E}=\mathbf{X}-\mathbf{A}(\mathbf{Z}\odot\mathbf{S})$$

であり，各時刻でのデータは独立同分布であると仮定している．

3.3 事後分布

ここではこれまでに示した事前分布と尤度関数を用いて，ベイズの定理に基づいた事後分布の推論について述べる．ここで推論された事後分布からのサンプリングによって分離信号，各信号のアクティビティ，混合行列の推定を行う．

3.3.1 音源信号

z_{kt} が active であるとき，s_{kt} の事後分布は式 (5) と尤度関数から以下のようになる．

$$\begin{aligned} P(s_{kt}|\mathbf{A},\mathbf{s}_{-kt},\mathbf{x}_t\mathbf{z}_t) &\propto P(\mathbf{x}_t|\mathbf{A},\mathbf{s}_t,\mathbf{z}_t,\sigma_\varepsilon^2)P(s_{kt}) \\ &= \mathcal{N}_C\left(s_{kt};\mu_s,\sigma_s^2\right), \end{aligned} \quad (13)$$

ここで，

$$\sigma_s^2=\frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2+\mathbf{a}_k^{\mathrm{H}}\mathbf{a}_k},\quad \mu_s=\frac{\mathbf{a}_k^{\mathrm{H}}\varepsilon_{-kt}}{\sigma_\varepsilon^2+\mathbf{a}_k^{\mathrm{H}}\mathbf{a}_k}$$

である．$\mathbf{s}_{-kt}$ は $\mathbf{s}_t$ から s_{kt} を除いたもの，ε_{-kt} は $\varepsilon|_{z_{kt}=0}$ を意味する．

3.3.2 音源のアクティビティ

z_{kt} が active になる事後確率と inactive と事後確率の比は式 (14) によって計算される．この事後確率の比 r は二つの部分に分けられ，片方は事前確率の比 r_p，もう片方は尤度の比 r_l である．

$$\begin{aligned} r &= \frac{P(z_{kt}=1|\mathbf{A},\mathbf{s}_{-kt},\mathbf{x}_t,\mathbf{z}_{-kt})}{P(z_{kt}=0|\mathbf{A},\mathbf{s}_{-kt},\mathbf{x}_t,\mathbf{z}_{-kt})} \\ &= \frac{P(\mathbf{x}_t|\mathbf{A},\mathbf{s}_{-kt},\mathbf{x}_t,\mathbf{z}_{-kt},z_{kt}=1,\sigma_\varepsilon^2)}{P(\mathbf{x}_t|\mathbf{A},\mathbf{s}_{-kt},\mathbf{x}_t,\mathbf{z}_{-kt},z_{kt}=0,\sigma_\varepsilon^2)}\frac{P(z_{kt}=1|\mathbf{z}_{kt})}{P(z_{kt}=0|\mathbf{z}_{kt})} \\ &= r_l r_p. \end{aligned} \quad (14)$$

事前確率の比 r_p は以下のように計算される．

$$r_p = \frac{P(z_{kt}=1|\mathbf{z}_{-kt})}{P(z_{kt}=0|\mathbf{z}_{-kt})} = \frac{m_{k,-t}}{N-m_{k,-t}}. \tag{15}$$

これは式 (11) の IBP に基づく音源のアクティビティの事前分布から導かれる[Griffiths and Ghahramani, 2006]．

尤度の比は式 (16) から計算される．

$$\begin{aligned} r_l &= \frac{P(\mathbf{x}_t|\mathbf{A},\mathbf{s}_{-kt},\mathbf{x}_t,\mathbf{z}_{-kt},z_{kt}=1,\sigma_\varepsilon^2)}{P(\mathbf{x}_t|\mathbf{A},\mathbf{s}_{-kt},\mathbf{x}_t,\mathbf{z}_{-kt},z_{kt}=0,\sigma_\varepsilon^2)} \\ &= \sigma^2 \exp\left(\frac{|\mu_s|^2}{\sigma_s^2}\right), \end{aligned} \tag{16}$$

これらを掛け合わせることで事後確率の比 r が得られ，$z_{kt}=1$ となる事後確率はこの比 r から計算される．

$$P(z_{kt}=1|\mathbf{A},\mathbf{s}_{-kt},\mathbf{x}_t,\mathbf{z}_{-kt}) = \frac{r}{1+r}. \tag{17}$$

z_{kt} が active かどうかを決定するために，一様分布 Uniform$(0,1)$ から u をサンプルし，それを $r/(1+r)$ と比較する．もし，$u \le r/(1+r)$ なら z_{kt} は active となり，そうでなければ active でないということになる．

3.3.3 新たに現れる音源の数

初めからは存在しておらず，時刻 t になって初めて出現する音源について考える．κ_t をそのような音源の数とすると，この κ_t は Metropolis-Hastings アルゴリズムによってサンプルされる．

まず，κ_t の事前分布は以下のようになる．

$$P(\kappa_t|\alpha) = \text{Poisson}\left(\frac{\alpha}{N}\right). \tag{18}$$

κ_t をサンプルしたのち，新しい音源とそのアク r ティビティを初期化する．

次に，この更新を受理するかどうかを決定する．現状態 ξ から，新しく κ_t 個の音源が加わった次状態 ξ^* への遷移確率 $J(\xi^*|\xi)$ は，Meeds ら[Meeds *et al.*, 2007]や Knowles ら[Knowles and Ghahramani, 2007]によると，次状態 ξ^* の事前分布と等しくなる．したがって，この遷移が採用される確率は $\min(1, r_{\xi\to\xi^*})$ となる．ただし，$r_{\xi\to\xi^*}$ は次式の通りである．

$$\begin{aligned} r_{\xi\to\xi^*} &= \frac{P(\xi^*|rest)J(\xi|\xi^*)}{P(\xi|rest)J(\xi^*|\xi)} \\ &= \frac{P(rest|\xi^*)P(\xi^*)P(\xi)}{P(rest|\xi)P(\xi)P(\xi^*)} \\ &= \frac{P(rest|\xi^*)}{P(rest|\xi)} \end{aligned} \tag{19}$$

$rest$ は ξ や ξ^* 以外のパラメータすべてをまとめたもの表す．つまり，$r_{\xi\to\xi^*}$ は更新前の状態と更新後の状態の尤度の比となる．この比を計算すると以下のようになる．

$$r_{\xi\to\xi^*} = (\det\Lambda_\xi)^{-1} \exp\left(\mu_\xi^{\mathrm{H}}\Lambda_\xi\mu_\xi\right), \tag{20}$$

ここで，

$$\Lambda_\xi = \mathbf{I} + \frac{\mathbf{A}^{*\mathrm{H}}\mathbf{A}^*}{\sigma_\varepsilon^2}, \quad \Lambda_\xi\mu_\xi = \frac{1}{\sigma_\varepsilon^2}\mathbf{A}^{*\mathrm{H}}\varepsilon_t.$$

である．また，$\mathbf{A}^*$ は $D\times\kappa_t$ の行列で，$\mathbf{A}$ の追加された部分を表している．

3.3.4 混合行列

混合行列は各列ごとに推定する．式 (6) で示した事前分布と尤度関数を用いると，事後分布は以下のようになる．

$$\begin{aligned} & P(\mathbf{a}_k|\mathbf{A}_{-k},\mathbf{S},\mathbf{X},\mathbf{Z},\sigma_\varepsilon^2,\sigma_{\mathbf{A}}^2) \\ \propto\ & P(\mathbf{X}|\mathbf{A},\mathbf{S},\mathbf{Z},\sigma_\varepsilon^2)P(\mathbf{a_k}|\sigma_{\mathbf{A}}^2) \\ =\ & \mathcal{N}_C(\mathbf{a}_k;\mu_{\mathbf{A}},\Lambda_{\mathbf{A}}^{-1}), \end{aligned} \tag{21}$$

ここで，

$$\begin{aligned} \Lambda_{\mathbf{A}} &= \left(\frac{\mathbf{s}_k^{\mathrm{H}}\mathbf{s}_k}{\sigma_\varepsilon^2} + \frac{1}{\sigma_{\mathbf{A}}^2}\right)\mathbf{I}_{D\times D}, \\ \mu_{\mathbf{A}} &= \frac{\sigma_{\mathbf{A}}^2}{\mathbf{s}_k^{\mathrm{H}}\mathbf{s}_k\sigma_{\mathbf{A}}^2+\sigma_\varepsilon^2}\mathbf{E}|_{\mathbf{a}_k=0}\mathbf{s}_k \end{aligned}$$

である．

3.3.5 雑音と混合行列の分散

雑音の分散は推定された信号の雑音のレベルに，混合行列の分散は推定された信号の振幅のスケールに対応している．それぞれの事後分布は以下のようになる．

$$\begin{aligned} P(\sigma_\varepsilon^2|\mathbf{E}) &\propto P(\mathbf{E}|\sigma_\varepsilon^2)P(\sigma_\varepsilon^2|p_1,p_2) \\ &= \mathcal{IG}\left(\sigma_\varepsilon^2; p_1+ND, \tfrac{p_2}{1+p_2\,\mathrm{tr}(\mathbf{E}^{\mathrm{H}}\mathbf{E})}\right). \end{aligned} \tag{22}$$

$$\begin{aligned} P(\sigma_{\mathbf{A}}^2|\mathbf{A}) &\propto P(\mathbf{A}|\sigma_{\mathbf{A}}^2)P(\sigma_{\mathbf{A}}^2|p_3,p_4) \\ &= \mathcal{IG}\left(\sigma_{\mathbf{A}}^2; p_3+DK, \tfrac{p_4}{1+p_4\,\mathrm{tr}(\mathbf{A}^{\mathrm{H}}\mathbf{A})}\right). \end{aligned} \tag{23}$$

3.3.6 IBP のパラメータ

IBP のパラメータ α の事後分布は以下のようになる．

$$\begin{aligned} p(\alpha|\mathbf{Z}) &\propto P(\mathbf{Z}|\alpha)P(\alpha|p_5,p_6) \\ &= \mathcal{G}\left(\alpha; K_+ + p_5, \frac{p_6}{1+p_6H_N}\right). \end{aligned} \tag{24}$$

ここで，K_+ は active となっている音源の数，$H_n = \sum_{j-1}^{N}\frac{1}{j}$ は N 番目の調和級数である．

3.4 後処理

周波数領域の ICA と同様に，本手法でもパーミュテーション問題とスケーリング問題について考えなければならない．これらの問題は，本手法では各周波数帯域で独立に分離を行うために，各帯域での出力信号の振幅および出力順序を揃える必要があるというものである．

ここで，スケーリング問題は projection back [Murata *et al.*, 2001]という方法で解決する．この方法は，分離信号に対して推定された混合行列の要素をかけあわせるこ

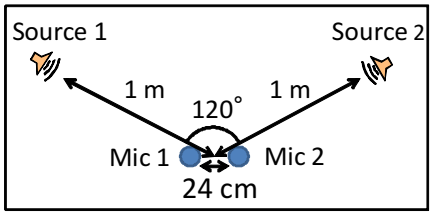

Figure 1: Locations of microphones and sources

Table 2: Experimental conditions

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 [msec]
STFT シフト幅	32 [msec]

とによって，各帯域で揃っていなかった振幅をマイクに入力される信号の振幅に合わせられるというものである．

パーミュテーション問題は本稿では混合前の原信号を用いて，分離信号との相関をとることで解決する．これは ISFA の複素拡張自身の分離性能を評価するためである．このパーミュテーション問題に対する解法は Sawada ら [Sawada *et al.*, 2004]などによって提案されているが，いまだ画期的な解法は開発されていないため，この問題の解法については今もなお活発に議論されている．

4 実験結果

本手法の分離性能の評価のために音声信号を用いた分離実験を行った．まず，本手法とベースラインである実数領域の ISFA とを比較する．実験は、瞬時混合・無響室録音のインパルス応答の畳み込み混合・会議室録音のインパルス応答の畳み込み混合の 3 種類の設定を用いた．Table 2 は実験状況をまとめたもので，Fig. 1 はマイクと音源の配置を示している．ATR 音素バランス単語データベース中の 32 単語の発話を用いた．反復回数は 150 回である．

Figures 2–5 のスペクトログラムはそれぞれ元音源，入力の混合信号，本手法による分離信号，ベースラインによる分離信号を表している．SDR (Signal to Distortion Ratio)，ISR (Image to Spatial distortion Ratio)，SIR (Source to Interference Ratio)，SAR (Source to Artifacts Ratio) [Vincent *et al.*, 2007]を用いた定量的な評価も行った．

結果は Table 3 の通りである．Baseline は時間領域の ISFA を表している．時間領域の ISFA については，瞬時混合の音声の分離実験では大変よい分離性能となっているが，畳み込み混合音声の分離実験では無響室程度のかなり短い残響時間の畳み込み混合でさえほぼ分離できていないのに対し，本手法は無響室，会議室などの畳み込み混合音声でも分離可能である．

無響室環境の場合，本手法はベースラインの手法と比較して SDR の平均で 2.91[dB] の改善がみられ，会議室環境においても本手法はベースラインに勝る分離性能となることが確認された．特に畳み込み混合の音声に対しては，SIR において本手法がベースライン手法に対し大きな改善が見られる．本手法の SAR の結果がベース手法と比較して悪化しているが，これは分離の際に用いるアクティビティ推定のミスにより生じるノイズによるものであると思われる．

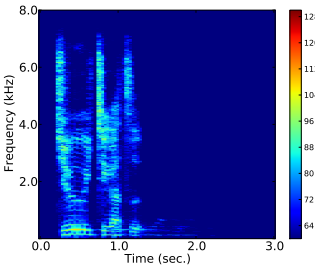

Figure 2: Spectrogram of source signal

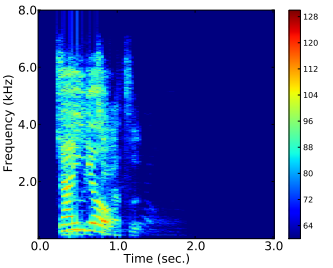

Figure 3: Spectrogram of mixed signal

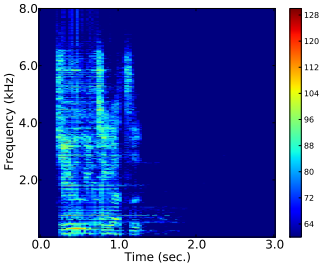

Figure 4: Separated signal with ours

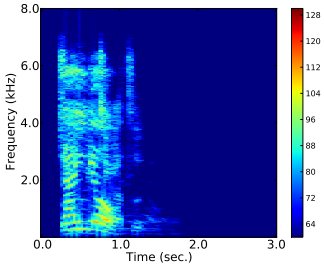

Figure 5: Baseline separated signal

また、我々の従来手法[柳楽ら, 2011]との性能比較も行った．実験条件は先ほどと同じであり，データは JNAS データベース中の 30 文を用いた．反復回数は 100 回である．その結果，無響室の場合，本手法が従来手法と比較して SDR で 0.27[dB]，SIR で 1.00[dB] の改善を確認した．

5 結論

本稿では実環境での反射音，残響，音源のマイクへの到達時間差などを考慮した畳み込み混合音声に対するブラインド音源分離と各音原のアクティビティの同時推定手法について述べた．本手法はノンパラメトリックベイズに基づいており，各周波数帯域ごとに ISFA の複素拡張を用いて複素混合信号を分離する．無響室環境での畳み込み混合音声の分離実験において，本手法によってベースラインの時間領域 ISFA と比較して平均 SDR で 2.91[dB] の改善がみられ，さらに会議室環境の畳み込み混合音声の分離実験でも分離性能の改善が見られた．また，我々の従来手法との比較においても改善が確認された．

今後の課題として，今回は音源のアクティビティについての評価を行い，これを発話区間検出（Voice Activity Detection）やパーミュテーション問題の解法に応用する事を考えている．そして，ロボット等への応用を考慮すると，リアルタイム処理を目指して本手法の処理速度の向

Table 3: Average separation performance from experimental results [dB]

	Instantaneous		
	Before	Baseline	Proposed
SDR	-1.19	**25.07**	2.27
ISR	2.35	**30.90**	4.06
SIR	1.17	**35.57**	10.45
SAR	75.77	**35.23**	2.83
	Anechoic chamber		
	Before	Baseline	Proposed
SDR	-1.01	-0.83	**2.08**
ISR	1.51	2.57	**3.86**
SIR	0.91	1.54	**8.91**
SAR	59.24	**36.25**	2.80
	Meeting room		
	Before	Baseline	Proposed
SDR	-1.96	-1.86	**0.60**
ISR	1.02	1.93	**2.98**
SIR	1.65	2.23	**4.90**
SAR	58.93	**36.08**	3.09

上について考える必要がある．

謝辞

本研究の一部は，科研費基盤 (S), JST-ANR BINAAHR, GCOE の支援を受けた．また，数多くの有益な助言をいただいた武田龍博士，平澤恭治氏に感謝の意を表する．

参考文献

[Belouchrani *et al.*, 1997] A. Belouchrani, K. Abed-Meraim, J.F. Cardoso, and E. Moulines. A blind source separation technique using second-order statistics. *Signal Processing, IEEE Transactions on*, 45(2):434–444, 1997.

[Griffiths and Ghahramani, 2006] T. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. *Advances in Neural Information Processing Systems*, 18:475–482, 2006.

[Hyvärinen *et al.*, 2001] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis*. Wiley-Interscience, 2001.

[Knowles and Ghahramani, 2007] D. Knowles and Z. Ghahramani. Infinite sparse factor analysis and infinite independent components analysis. *Independent Component Analysis and Signal Separation*, pages 381–388, 2007.

[Meeds *et al.*, 2007] E. Meeds, Z. Ghahramani, R.M. Neal, and S.T. Roweis. Modeling dyadic data with binary latent factors. *Advances in Neural Information Processing Systems*, 19:977–984, 2007.

[Murata *et al.*, 2001] N. Murata, S. Ikeda, and A. Ziehe. An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing*, 41(1-4):1–24, 2001.

[Nakadai *et al.*, 2010] K. Nakadai, T. Takahashi, H.G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino. 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.

[Sawada *et al.*, 2002] H. Sawada, R. Mukai, S. Araki, and S. Makino. 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.

[Sawada *et al.*, 2004] H. Sawada, R. Mukai, S. Araki, and S. Makino. 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.

[Seltzer *et al.*, 2004] M.L. Seltzer, B. Raj, and R.M. Stern. Likelihood-maximizing beamforming for robust hands-free speech recognition. *IEEE Trans. on Speech and Audio Processing*, 12(5):489–498, 2004.

[Valin *et al.*, 2004] J.M. Valin, J. Rouat, and F. Michaud. Enhanced robot audition based on microphone array source separation with post-filter. In *Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, volume 3, pages 2123–2128. IEEE, 2004.

[Vincent *et al.*, 2007] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. Rosca. First stereo audio source separation evaluation campaign: data, algorithms and results. *Independent Component Analysis and Signal Separation*, pages 552–559, 2007.

[Wölfel and McDonough, 2009] M. Wölfel and J. McDonough. *Distant Speech Recognition*. Wiley, 2009.

[柳楽ら, 2011] 柳楽 浩平, 高橋 徹, 尾形 哲也, 奥乃 博. ノンパラメトリックベイズによる時間周波数領域における音声信号のブラインド音源分離. **第 *29* 回日本ロボット学会学術講演会**, 3A2–5, 2011.