# 複合自己回帰系による音声パワースペクトル密度モデルを用いたブラインド音源分離と残響除去*

☆濱村真理子 (NTT CS研/九大院・芸工), 亀岡弘和, 吉岡拓也
ルルージョナトン, 柏野邦夫 (NTT CS研)

## 1 はじめに

ブラインド音源分離 (BSS) やブラインド残響除去では，音源の成分と音源からマイクロホンまでの伝達特性がともに未知のもとで，マイクロホン入力信号から音源成分を復元することが目的であり，音源に関する何らかの仮定をもとに音源らしさの規準を定め，これにより立てられる目的関数を最小 (最大) 化する最適化問題として定式化されることが多い。音源らしさの規準は，どのような信号を生成しやすいか，という何らかの傾向を表した統計モデルを仮定することにより定義できる。どのような統計モデルを仮定すべきかは，対象とする音源信号の実際のふるまいや現象にいかに即しているか，数理的技法が馴染みやすい最適化問題に帰着するか，を考慮して判断することが重要である。

BSS は音源の混合過程に関する仮定に応じてさまざまなアプローチに大別でき，それぞれに一長一短がある。時間領域の畳み込み混合モデルに基づく BSS アルゴリズム [1] は，混合過程のインパルス応答が長くなると計算量が膨大になる問題があるのに対し，短時間 Fourier 変換 (STFT) 領域においてフレームが混合過程のインパルス応答より十分長い場合には時間領域の畳み込み混合は周波数領域の瞬時混合で近似することができ，効率の良いアルゴリズムが実現可能となる [2]。しかし，混合過程のインパルス応答が STFT のフレームの長さを超える残響環境下では，上記の近似の精度は必然的に低下してしまう。一方近年の残響除去の研究では残響音声が STFT のサブバンドごとの音声信号と室内インパルス応答の畳み込みにより良く表せることが示されており [3]，高残響環境下では STFT 領域における畳み込み混合モデルをベースにした BSS [4] が性能および効率性の面で有効であることが示されている。以下，このモデルをベースに BSS の問題を具体的に示す。まず，$M$ 個の音源の信号を $M$ 個のマイクロホンで収音する状況を考え，これらの収音信号を STFT により展開した時間周波数成分を観測データと見なすことにしよう。ここで，$l$ 番目のマイクロホンで観測される信号を $Y^l_{k,n}$(ただし，$k, n$ は周波数，時刻に対応するインデックス) とし，$\boldsymbol{Y}_{k,n} = (Y^1_{k,n}, \cdots, Y^M_{k,n})^{\mathrm{T}}$ とする。また，$M$ 個の音源成分を並べたベクトルを $\boldsymbol{S}_{k,n} = (S^1_{k,n}, \cdots, S^M_{k,n})^{\mathrm{T}}$ とする。本稿では [4] に倣い，多チャンネル有限インパルス応答によって表される分離システム

$$\boldsymbol{S}_{k,n} = \sum_{\tau=0}^{n_k} \boldsymbol{W}_{k,\tau} \boldsymbol{Y}_{k,n-\tau}, \tag{1}$$

を用いる。ただし，$\boldsymbol{W}_{k,\tau}$, $0 \le \tau \le n_k$ は $M \times M$ 行列の分離フィルタである。さて，ここで，$S^m_{k,n}$ を確率変数とし，$S^m_{k,n}$ と $S^{m'}_{k',n'}$ は $(k,n,m) \neq (k',n',m')$ のとき独立と仮定する。$S^m_{k,n}$ の確率密度関数 $f_{S^m_{k,n}}(s^m_{k,n};\theta^m)$ は音源信号の統計モデルに該当し，$\theta^m$ はその分布を特徴づけるモデルパラメータである。この統計モデルの具体形が決まれば $\boldsymbol{Y} := \{\boldsymbol{Y}_{k,n}\}_{k,n}$ の確率密度関数は $f_{S^m_{k,n}}(s^m_{k,n};\theta^m)$ を用いて自動的に

$$f_{\boldsymbol{Y}|\Theta}(\boldsymbol{Y}|\Theta) = \prod_k |\det \boldsymbol{W}_{k,0}| \prod_{t,m} f_{S^m_{k,n}|\theta^m} \left(\bar{S}^m_{k,n}|\theta^m\right)$$

と定義される。ただし，$\bar{S}^m_{k,n}$ は分離信号 $\sum_\tau \boldsymbol{W}_{k,\tau} \boldsymbol{Y}_{k,n-\tau}$ の $m$ 番目のベクトル要素である。この同時密度関数は，未知パラメータ $\Theta := \{\{\theta^m\}_m, \{\boldsymbol{W}_{k,\tau}\}_{k,\tau}\}$ の尤度関数に対応するものであり，[4] のように最尤推定の枠組でブラインド音源分離とブラインド残響除去を同時に行うための目的関数となる。

音源の統計モデルをどのように定めるかによっても BSS のアプローチは大別することができる。例えば音声の短時間信号のモデルとして自己回帰モデルが有名であるが，[5] や [4] では，自己回帰モデルに基づいて確率密度関数 $f_{S^m_{k,n}}(s^m_{k,n}|\theta^m)$ を記述した，最尤法に基づく BSS の解法が提案されている。本稿では，[4] の BSS システムの枠組に，自己回帰モデルに代わる新しい音声の統計モデルを組み込むことで BSS の性能向上を図ることが主目的である。

## 2 複合自己回帰系による音声の統計モデル

自己回帰モデルは，白色の Gauss 性雑音を入力とした全極システムからの出力と捉えることができ，入力とフィルタをそれぞれ声帯振動波形と声道伝達特性に対応させることで音声生成過程の一つの簡潔なモデルと見なすことができる。しかし，駆動信号を白色雑音で仮定す

---

*Blind source separation and dereverberation using speech power spectral density model derived from composite autoregressive system. by HAMAMURA Mariko (NTT Communication Science Laboratories/Kyushu University), KAMEOKA Hirokazu, YOSHIOKA Takuya, LE ROUX Jonathan and KASHINO Kunio (NTT Communication Science Laboratories)

ることは，駆動信号が一様なパワースペクトル密度 (以後，PSD) をもつことを仮定していることに相当し，高い基本周波数になるほど適合しなくなる。従って，駆動信号の PSD 自体も推定すべきパラメータとして扱いたいが，白色性に代わる何らかの新たな制約が必要である。そこで，発話全体を通して声の高さや音素の種類が限られているという音声の特徴に着目し，高々数種類の声道伝達特性と高々数種類の駆動信号 PSD の組み合せで音声を表現できるという仮定の下考案したのが以前提案した複合自己回帰系 [6] に基づく音声信号のモデルである。以下，このモデルについて具体的に説明する。

ここからは特定の音声信号について議論することとし，音源成分のインデックス $m$ は省略する。 ここで $n$ 番目のフレームにおける信号 $\{x_n[t]\}_{t=1}^{T}$ を，$P$ 次の自己回帰過程

$$x_n[t] = \sum_{p=1}^{P} a_p x_n[t-p] + \epsilon_n[t] \tag{2}$$

からの標本値系列と仮定する。ただし，$\epsilon[t]$ は声帯駆動信号に対応し，平均 0 の Gauss 性雑音としているが，必ずしも白色雑音とは限らない点を強調しておく。$\boldsymbol{x}_n = (x_n[1], \cdots, x_n[K])^{\mathrm{T}} \in \mathbb{R}^K$ と定義し，その離散 Fourier 変換 (以後，DFT) を $\boldsymbol{X}_n = (X_{1,n}, \cdots, X_{K,n})^{\mathrm{T}} \in \mathbb{C}^K$ とする。式 (2) より，DFT の線形性，および，$\epsilon_n[t]$ の定常性，Gauss 性から，$\boldsymbol{X}_n$ は，平均が 0，分散共分散行列が $\boldsymbol{\Lambda} = \mathrm{diag}\big(\lambda_1, \cdots, \lambda_K\big)$ の多次元複素正規分布 $\mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \boldsymbol{\Lambda})$ に従う。ただし，

$$\lambda_{k,n} = \frac{H_k U_n}{|A(e^{\mathrm{j}2\pi k/K})|^2}, \tag{3}$$

$$A(z) = 1 - a_1 z^{-1} \cdots - a_P z^{-P} \tag{4}$$

である。$H_1, \cdots, H_K$ は駆動信号 $\epsilon[t]$ の PSD であり，スペクトル微細構造に対応する。一方で，$1/|A(e^{\mathrm{j}2\pi 1/K})|^2$, $1/|A(e^{\mathrm{j}2\pi 2/K})|^2$, $\cdots$, $1/|A(e^{\mathrm{j}2\pi K/K})|^2$ は全極型伝達関数の PSD であり，スペクトル包絡に対応する。

以上をもとに，複合自己回帰系を構築することができる。複合自己回帰系は，$I$ 種類の駆動信号 PSD と $J$ 種類の全極型フィルタにより構成される複合系であり，$i$ 番目の駆動信号 PSD と $j$ 番目の全極型伝達関数をそれぞれ $H_k^i, 1/A^j(e^{\mathrm{j}2\pi k/K})$ とする。この複合系は，すべての駆動信号 PSD と全極型フィルタを組み合わせた計 $I \times J$ 種類の音声要素を生成することができる。ここで例えば，各フレームで単一の音声要素だけがアクティブになると仮定した場合，この複合系の出力信号の統計モデルは [7] と同様の形態の混合正規分布により記述することができる。これに対し，複合自己回帰系では，これらの計 $I \times J$ 種類の音声要素の和を音声信号と考える。各音声要素のゲインがフレーム $n$ ごとに唯一異なる値を取り得るパラメータであり，この値を通して決まる $I \times J$ 種類の音声要素の混ざり具合により各フレームの音声スペクトルが特徴づけられる。$n$ 番目のフレームにおいて，$i$ 番目の駆動信号 PSD と $j$ 番目の全極型フィルタの組み合せから生成される音声要素のゲインを $U_n^{i,j}$ とすると，当該音声要素の DFT$\boldsymbol{X}_n^{i,j} = (X_{1,n}^{i,j}, \cdots, X_{K,n}^{i,j})^{\mathrm{T}}$ は多次元複素正規分布 $\mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \boldsymbol{\Lambda}_n^{i,j})$ に従う。$\boldsymbol{\Lambda}_n^{i,j}$ は $\boldsymbol{\Lambda}_n^{i,j} = \mathrm{diag}(\lambda_{1,n}^{i,j}, \cdots, \lambda_{K,n}^{i,j})$ の対角共分散行列で，対角要素は

$$\lambda_{k,n}^{i,j} = \frac{H_k^i U_n^{i,j}}{|A^j(e^{\mathrm{j}2\pi k/K})|^2} \tag{5}$$

$$A^j(z) = 1 - a_1^j z^{-1} \cdots - a_P^j z^{-P} \tag{6}$$

で与えられる。ここで，$\boldsymbol{X}_n^{i,j}$ と $\boldsymbol{X}_{n'}^{i',j'}$ は $(i,j) \neq (i',j')$ または $n \neq n'$ のとき独立と仮定すると，$\boldsymbol{X}_n^{1,1}, \cdots, \boldsymbol{X}_n^{I,J}$ の和，すなわち音声信号の STFT $\boldsymbol{S}_n \in \mathbb{C}^K$ はやはり正規分布に従い，

$$\boldsymbol{S}_n = \sum_i \sum_j \boldsymbol{X}_n^{i,j} \sim \mathcal{N}_{\mathbb{C}}\big(\boldsymbol{0}, \boldsymbol{\Phi}_n\big) \tag{7}$$

$$\boldsymbol{\Phi}_n = \sum_i \sum_j \boldsymbol{\Lambda}_n^{i,j} \tag{8}$$

が言える。式 (8) は，正規分布する独立な確率変数の和は各々の正規分布の分散の和を分散とする正規分布に従うことを意味する。以上より，$S_{k,n}$ の統計モデルは

$$f_{S_{k,n}}(s_{k,n}|\theta) = \frac{1}{\pi \phi_{k,n}} \exp\left(-\frac{|s_{k,n}|^2}{\phi_{k,n}}\right) \tag{9}$$

で与えられる。ただし，$\theta = \{H_k^i, a_p^j, U_n^{i,j}\}_{k,n,i,j}$ は以上の統計モデルにおける未知パラメータである。$\boldsymbol{\Phi}_n$ の対角要素の $\phi_{k,n}$ は複合系により生成される不規則信号の PSD であり，

$$\phi_{k,n} = \sum_i \sum_j \frac{H_k^i U_n^{i,j}}{|A^j(e^{\mathrm{j}2\pi k/K})|^2} \tag{10}$$

である。ところで，$J = 1$, $P = 0$ の特殊ケースにおいて，$\phi_{k,n}$ は $\phi_{k,n} = \sum_i H_k^i U_n^{i,1}$ のとおり行列積の形となり，この下での式 (9) は，非負値行列因子分解 (NMF) のある統計的解釈に基づいて導かれた尤度関数 [8] と同じ形となる。よって，複合自己回帰系による音源モデルを用いた全体の BSS システムは多チャンネル NMF[9] と構造的に深い関連があることが示唆される。

## 3 最適化アルゴリズム

観測信号 $\boldsymbol{Y}$ が与えられたとき，未知パラメータ $\boldsymbol{\Theta} = \{\{\theta^m\}_m, \{\boldsymbol{W}_{k,\tau}\}\}$ についての事後確率 $f_{\Theta|\boldsymbol{Y}}(\Theta; \boldsymbol{Y}) \propto f_{\boldsymbol{Y}|\Theta}(\boldsymbol{Y}; \Theta)\ f_{\Theta}(\Theta)$ の対数

$$L(\theta) := \log f_{\boldsymbol{Y}|\Theta}(\boldsymbol{Y}|\Theta) + \log f_{\Theta}(\Theta) \tag{11}$$

を最大化する $\boldsymbol{\theta}$ を探索するアルゴリズムについて述べる。この最適化問題は解析的に求めることはできないが，次の 3 つの最大化ステップ

$$\text{(S1)} \quad \theta^m \leftarrow \underset{\theta^m}{\operatorname{argmax}} L(\theta) \ \text{ for all } \ m,$$
$$\text{(S2)} \quad \boldsymbol{W}_{k,0} \leftarrow \underset{\boldsymbol{W}_{k,0}}{\operatorname{argmax}} L(\theta) \ \text{ for all } \ k$$
$$\text{(S3)} \quad \{\boldsymbol{W}_{k,\tau}\}_{\tau=1}^{n_k} \leftarrow \underset{\{\boldsymbol{W}_{k,\tau}\}_{\tau=1}^{n_k}}{\operatorname{argmax}} L(\theta) \ \text{ for all } \ k$$

を繰り返すことにより局所最適解を求めることができる。もし，各音源の PSD の推定値 $\phi_{k,n}$ が分かれば (S2) と (S3) の更新は [4] で述べられているプロセスで行うことができるため，(S1) の部分最適化解の探索アルゴリズムについてのみ以下に示す。(S1) は $\sum_\tau \boldsymbol{W}_{k,\tau}\boldsymbol{Y}_{k,n-\tau}$ の $m$ 番目のベクトル要素 $\bar{S}^m_{k,n}$ を $\theta_m$ に関して

$$\sum_{k,n} \log f_{S^m_{k,n}|\theta^m}(\bar{S}^m_{k,n}|\theta^m) + \log f_{\theta^m}(\theta^m) \qquad (12)$$

を $m$ ごとに最大化することと等しい。よってここより再び特定の $m$ についてのみ考えるので表記の簡単化のため $m$ は省略する。

式 (9) より，$\log f_{S_{k,n}}(\bar{S}_{k,n}|\theta)$ は $|\bar{S}_{k,n}|^2$ と $\phi_{k,n}$ との板倉斎藤距離と定数項を除いて等しい。第二項の事前確率は，音声要素のアクティベーションを抑制するためのスパース正則化項として利用する。以下では，指数分布 (片側 Laplace 分布)

$$f_\theta(\theta) = \prod_{i,j,n} \alpha \exp(-\alpha U^{i,j}_n) \qquad (13)$$

を仮定する。この事前確率のスパース化効果は $\alpha$ が大きいほど高くなる。このスパース正則化によって，各フレームの音声スペクトル $|\bar{S}_{k,n}|^2$ の中に高い頻度で生起するスペクトル微細構造ないしスペクトル包絡構造ほど駆動信号 PSD ないし全極型フィルタの推定結果となって表れやすい仕組となる。

式 (12) を最大化する $\theta$ は解析的に求めることはできないが，以下に示すように $\boldsymbol{X}_{k,n} = (X^{1,1}_{k,n}, \cdots, X^{I,J}_{k,n})^{\mathrm{T}}$ を完全データと見なして Expectation-Maximization (EM) アルゴリズムにより効率良く局所解を探索することができる。なお，(S1)〜(S3) の各ステップで目的関数の単調増加さえ保証されていれば全体のアルゴリズムの収束性は保証される。$X^{i,j}_{k,n}$ と $X^{i',j'}_{k',n'}$ は $(k,n,i,j) \neq (k',n',i',j')$ のとき独立であるため，完全データ $\boldsymbol{X}_{k,n}$ の対数尤度は具体的に

$$\log f_{\boldsymbol{X}|\theta}(\boldsymbol{X}|\theta) = -\sum_{k,n}\Big[\log\det\pi\boldsymbol{\Lambda}_{k,n} + \mathrm{tr}\big(\boldsymbol{\Lambda}^{-1}_{k,n}\boldsymbol{X}_{k,n}\boldsymbol{X}^{\mathrm{H}}_{k,n}\big)\Big] \qquad (14)$$

と書ける。上式に対し，$\bar{S}_{k,n}$ および $\theta = \theta'$ の条件付期待値を取り，$\log f_\theta(\theta)$ を加えると，Q 関数

$$Q(\theta,\theta') = \log f_\theta(\theta) - \sum_{k,n}\Big[\log\det\pi\boldsymbol{\Lambda}_{k,n} + \mathrm{tr}\big(\boldsymbol{\Lambda}^{-1}_{k,n}\mathbb{E}\big[\boldsymbol{X}_{k,n}\boldsymbol{X}^{\mathrm{H}}_{k,n}|S_{k,n} = \bar{S}_{k,n}, \theta = \theta'\big]\big)\Big] \qquad (15)$$

を得る。ここで，不完全データ $\bar{S}_{k,n}$ と完全データ $\boldsymbol{X}_{k,n}$ に $\bar{S}_{k,n} = \mathbf{1}^{\mathrm{T}}\boldsymbol{X}_{k,n}$ (ただし，$\mathbf{1} = (1,\cdots,1)^{\mathrm{T}}$) なる多対一関係があるため，$\mathbb{E}\big[\boldsymbol{X}_{k,n}\boldsymbol{X}^{\mathrm{H}}_{k,n}|S_{k,n} = \bar{S}_{k,n}, \theta = \theta'\big]$ は具体的に

$$\begin{aligned}&\mathbb{E}\big[\boldsymbol{X}_{k,n}\boldsymbol{X}^{\mathrm{H}}_{k,n}|S_{k,n} = \bar{S}_{k,n}, \theta = \theta'\big] = \\ &\boldsymbol{\Lambda}'_{k,n} - \boldsymbol{\Lambda}'_{k,n}\mathbf{1}\big(\mathbf{1}^{\mathrm{T}}\boldsymbol{\Lambda}'_{k,n}\mathbf{1}\big)^{-1}\mathbf{1}^{\mathrm{T}}\boldsymbol{\Lambda}'_{k,n} + \\ &|\bar{S}_{k,n}|^2\boldsymbol{\Lambda}'_{k,n}\mathbf{1}\big(\mathbf{1}^{\mathrm{T}}\boldsymbol{\Lambda}'_{k,n}\mathbf{1}\big)^{-1}\big(\mathbf{1}^{\mathrm{T}}\boldsymbol{\Lambda}'_{k,n}\mathbf{1}\big)^{-1}\mathbf{1}^{\mathrm{T}}\boldsymbol{\Lambda}'_{k,n}\end{aligned} \qquad (16)$$

と書ける。ただし $\boldsymbol{\Lambda}'_{k,n}$ は $\boldsymbol{\Lambda}_{k,n}$ に $\theta = \theta'$ を代入したものを表す。$\theta$ に依らない項をまとめて $c$ とすると上式は

$$\begin{aligned}Q(\theta,\theta') = &-\sum_{k,n}\sum_{i,j}\left[\log H^i_k U^{i,j}_n + \frac{\Psi^{i,j}_{k,n}|A^j(e^{\mathrm{j}2\pi k/K})|^2}{H^i_k U^{i,j}_n}\right] \\ &-\alpha\sum_n\sum_{i,j}U^{i,j}_n + \mathrm{c}\end{aligned} \qquad (17)$$

と書き直せる。ただし，

$$\Psi^{i,j}_{k,n} = \frac{{\lambda'}^{i,j}_{k,n}}{{\phi'}_{k,n}}\big({\phi'}_{k,n} - {\lambda'}^{i,j}_{k,n} + \frac{{\lambda'}^{i,j}_{k,n}}{{\phi'}_{k,n}}|\bar{S}_{k,n}|^2\big) \qquad (18)$$

は各要素信号の PSD 推定値であり，E ステップでは $\theta'$ の更新に伴いこの値が更新されることになる。

以上より，各パラメータの M ステップ更新則が導ける。Q 関数を $H^i_k$ と $U^{i,j}_n$ に関して偏微分して 0 と置くと，

$$H^i_k = \frac{1}{NJ}\sum_n\sum_j \Psi^{i,j}_{k,n}|A^j(e^{\mathrm{j}2\pi k/K})|^2/U^{i,j}_n \qquad (19)$$

$$\begin{aligned}&\alpha U^{i,j\,2}_n + KU^{i,j}_n - \\ &\sum_k \Psi^{i,j}_{k,n}|A^j(e^{\mathrm{j}2\pi k/K})|^2/H^i_k = 0, \quad U^{i,j}_n \geq 0\end{aligned} \qquad (20)$$

が得られる。同様に，$a^j_1, \cdots, a^j_P$ に関して偏微分して 0 と置き，連立させると，Yule-Walker 方程式

$$r^j_p = \sum_{q=1}^{P} a^j_q r^j_{p-q} \ (p = 1, \cdots, P) \qquad (21)$$

を得る。ただし，$r^j_p$ は，インデックス $j$ の全ての音声要素のスペクトル包絡の平均を取ったものの逆 DFT

$$r^j_p = \sum_k\sum_{n,i}\frac{\Psi^{i,j}_{k,n}}{H^i_k U^{i,j}_n}e^{p\mathrm{j}2\pi k/K} \qquad (22)$$

である。以上より，$a^j_1, \cdots, a^j_P$ の更新値は，式 (21) を Levinson-Durbin アルゴリズムで解くことで得られる。

なお，駆動信号 PSD と全極型フィルタとの間にスケールの任意性があるため，各反復の $H$ の更新後に $H^i_k \leftarrow H^i_k/\sum_{k'} H^i_{k'}$ のような正規化ステップを設けることにする。このステップを含めることで厳密にはアルゴリズムの収束性は保証されなくなるが，実験を通してこのことが収束性能に大きく影響を及ぼすような場面は特に見受けられなかった。

表 1　入力 SIR(dB)

| Source | Channel #1 | #2 | #3 | #4 |
|---|---|---|---|---|
| #1 | -0.6 | -0.3 | -0.1 | 0.6 |
| #2 | 0.6 | 0.3 | 0.1 | -0.6 |

表 2　$I$ と $J$ の各設定条件における出力 SIR(dB)

| | $I=2$ | | $I=5$ | | $I=10$ | | Base- | Base- |
|---|---|---|---|---|---|---|---|---|
| Source | $J=12$ | $J=15$ | $J=8$ | $J=10$ | $J=8$ | $J=12$ | line1 | line2 |
| #1 | 18.0 | 16.8 | 19.9 | 18.5 | 17.7 | 17.6 | 11.6 | 17.2 |
| #2 | 11.7 | 10.9 | 14.1 | 13.9 | 13.6 | 14.5 | 11.0 | 13.9 |

## 4　実験結果

ATR 音声データベースの音声データ (女性話者の音声信号で標本化周波数は 16kHz) を用いて提案法による残響環境下でのブラインド音源分離の実験を行った。混合過程のインパルス応答は，残響時間が 0.5s の可変残響室にて計測したた 2 入力 4 出力のものを用いた。各音源信号の入力信号対妨害比 (Signal-to-Interface ratio, SIR) は表 1 に示すとおりである。また，音源 1 と音源 2 の直接音対残響音比 (Direct-to-reverberation ratio, DDR) はそれぞれ 5.7dB，4.7dB であった。マイクロホン入力信号の時間周波数成分はポリフェーズフィルタバンクにより算出し，STFT で言うところのフレーム長 32ms 相当，フレームシフト 8ms 相当の分析条件とした。分離システムにおけるフィルタ長 $n_k$ は，$F_k < 0.8$ において $n_k = 25$，$0.8 \leq F_k < 1.5$ において $n_k = 20$，$1.5 \leq F_k < 3$ において $n_k = 15$，$F_k \geq 3$ において $n_k = 10$ とした。ただし，$F_k$ は $k$ 番目の周波数ビンの kHz 単位での周波数値を表す。複合自己回帰系における各全極モデルの次数 $P$ は 12，スパース正則化パラメータ $\alpha$ は $K/100$ とし，(S1) から (S3) の反復回数は 3 回，(S1) の内部で実行する EM アルゴリズムの反復回数は 100 回とした。

提案法を様々な $I$ と $J$ の組み合わせの設定条件で行った際の各々の分離性能の結果を，ベースライン手法による分離性能の結果と併せて表 2 に示す。Baseline1 および Baseline2 はそれぞれ澤田らの手法 [10] と吉岡らの手法 [4] をさす。Baseline1 については時間周波数解析をフレーム長 256ms，フレームシフト 64ms の条件で，Baseline1 についてはフレーム長 16ms，フレームシフト 8ms の条件でそれぞれ行った。実験結果より，提案法による分離精度は $I$ と $J$ が 5 と 8 のとき最も高く，両ベースライン手法を凌駕する結果を得た。

## 5　まとめ

本稿では，我々が以前提案した複合自己回帰系と呼ぶ音声信号の統計モデルを既存の BSS システムに組み込むことで BSS の性能向上を図ることを目的とし，実験によりその一有効性を示すことができた。また，[4] の手法と同様，残響環境下においても良い分離精度が得られることが分かった。

## 参考文献


[1] S. Douglas, H. Sawada and S. Makino, “Natural gradient multichannel blind deconvolution and speech separation using causal FIR filters,” *IEEE Trans. Speech, Audio Process.*, vol. 13, no. 1, pp. 92–104, 2005.

[2] P. Smaragdis, “Blind separation of convolved mixtures in the frequency domain,” *Neur. Comp.*, vol. 22, pp. 21–34, 1998.

[3] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi and B.-H. Juang, “Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation,” in *Proc. Int’l Conf. Acoust., Speech, Signal Process.*, 2008, pp. 85–88.

[4] T. Yoshioka, T. Nakatani, M. Miyoshi and H. G. Okuno, “Blind separation and dereverberation of speech mixtures by joint optimization,” accepted for publication in *IEEE Trans. Audio, Speech, Language Process.*, 2010.

[5] S. Dégerine and A. Zaïdi, “Separation of an instantaneous mixture of Gaussian autoregressive sources by the exact maximum likelihood approach,” *IEEE Trans. Signal Processing*, vol. 52, no. 6, pp. 1499–1512, 2004.

[6] H. Kameoka and K. Kashino, “Composite Autoregressive System for Sparse Source-Filter Representation of Speech,” In Proc. 2009 IEEE International Symposium on Circuits and Systems (ISCAS2009), pp. 2477-2480, 2009.

[7] L. Benaroya, F. Bimbot and R. Gribonval, “Audio source separation with a single sensor,” *IEEE Trans. Audio Speech Language Processing*, vol. 14, no. 1, pp. 191–199, 2006.

[8] C. Févotte, N. Bertin and J.-L. Durrieu, “Nonnegative matrix factorization,with the Itakura-Saito divergence. With application to music,analysis,” *Neural Comput.*, vol. 21, no. 3, pp. 793–830, Mar. 2009.

[9] A. Ozerov and C. Févotte, “Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 3, pp. 550–563, March 2010.

[10] H. Sawada, S. Araki and S. Makino, “Measuring dependence of binwise separated signals for permutation alignment in frequency-domain BSS,” in *Proc. Int’l Symp. Circ., Syst.*, 2007, pp. 3247–3250.