ノンパラメトリックベイズアプローチによる劣決定スパース BSS *

○亀岡弘和 [1,2], 佐藤美沙 [3], 小野拓磨 [1], 小野順貴 [4], 嵯峨山茂樹 [1]
([1] 東大院・情報理工, [2]NTT CS 研, [3] 東大・工, [4]NII)

1 序 論

ブラインド音源分離 (Blind Source Separation; BSS) とは，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技術である。音声信号を対象とした BSS は，ハンズフリーテレビ会議システムや会議録コンテンツの自動作成システムなど，多くの応用が期待されている。例えば会議の場面では参加人数が途中で変化したりドアの開閉音などが突発的に鳴ったりすることがあるように，実環境においてはあらかじめあらゆる音源の数を想定しておくことは難しい。従来の多くの BSS アルゴリズムは音源数を仮定して動作するものが多く，仮定した音源数が実際の音源数と異なる場合，高い性能を発揮できない場合がある。そこで本研究では，音声を対象とし，音源数を仮定することなく自律的に音源数を推論しながら動作する BSS システムの実現を目指している。

BSS では観測信号だけから音源信号とその混合過程を推定する必要があるため，通常は音源に関して何らかの仮定を置き，これにより立てられる規準をもとに両未知変数を推定する最適化問題として定式化される。例えば，観測信号数が音源数以上の場合には，独立成分分析が有効な手法として知られ，音源信号間の独立性を最大化するように分離フィルタを推定することが目的となる [1]。しかし，音源数を仮定しない BSS システムを実現するためには，マイクロホン数よりも音源数が多い劣決定な問題設定を想定しておく必要があり，独立成分成分分析をそのまま適用することはできない。劣決定の条件下では，たとえ混合過程が既知であったとしても解が一意に決められないため，音源に関して独立性よりさらに強い仮定が必要となる。

音声を対象とした劣決定条件での BSS では，音声の時間周波数成分のスパース性を利用したアプローチが有効である [2–6]。音声のスパース性とは，音声信号の時間周波数成分が多くの領域でほぼ 0 となる性質である。このため，複数の音声が同時に発話された状況でも，各時間周波数において音声の時間周波数成分は互いにほとんど重なり合わないと仮定できる場合が多い。この仮定をもとに，目的音声信号の時間周波数成分のみを通過させるような時間周波数マスクをいかにうまく設計するかがこのアプローチにおける問題の焦点となる。

各マイクロホンにおける観測信号は，通常，音源信号の時間遅れを含む畳み込み混合で表されるが，以上の音声のスパース性の仮定を組み込むには，観測モデルを時間周波数領域に展開して定式化する必要がある。音源からマイクロホンまでのインパルス応答長に対して十分に長い時間窓をもつ時間周波数分解 (短時間 Fourier 変換，ウェーブレット変換など) を用いると，畳み込み混合を近似的に瞬時混合で表すことができる。この観測モデルに基づく BSS は周波数領域 BSS と呼ばれ，時間領域の畳み込み混合モデルに基づく BSS に対し，演算量の少ないアルゴリズムを実現できる点や上述の音声のスパース性の仮定を組み込める点などの特長がある一方で，周波数ごとに分離した成分を音源ごとにまとめるためのパーミュテーション整合と呼ばれる問題を扱う必要がある。

以上の背景と要請のもと，本稿では，(1) 音源数の推論，(2) 音声のスパース性を仮定した劣決定周波数領域 BSS，(3) パーミュテーション整合，の問題を一挙に解決するアプローチを提案する。

2 観測モデル

まず，K 個の信号源から到来する音源信号を M 個のマイクロホンで観測する場合を考え，m 番目のマイクロホンで観測される信号の時間周波数成分を $y_m(\omega,t)$，k 番目の音源信号の時間周波数成分を $s_k(\omega,t)$ とし，$\boldsymbol{y}(\omega,t)=(y_1(\omega,t),\ldots,y_M(\omega,t))^{\mathsf{T}}\in\mathbb{C}^M$，$\boldsymbol{s}(\omega,t)=(s_1(\omega,t),\ldots,s_K(\omega,t))^{\mathsf{T}}\in\mathbb{C}^K$ とする。ただし，$1\le\omega\le\Omega$, $1\le t\le T$ はそれぞれ周波数および時刻に対応するインデックスである。先に述べたとおり，時間周波数領域において観測信号 $\boldsymbol{y}(\omega,t)$ は近似的に

$$\boldsymbol{y}(\omega,t)=\boldsymbol{A}(\omega)\boldsymbol{s}(\omega,t)+\boldsymbol{n}(\omega,t) \qquad (1)$$

のように $\boldsymbol{s}(\omega,t)$ の瞬時混合の形で表すことができる。信号源 k からマイクロホン m までの伝達周波数特性 $a_{m,k}(\omega)$ を要素にした行列 $\boldsymbol{A}(\omega)=(a_{m,k}(\omega))_{M\times K}=(\boldsymbol{a}_1(\omega),\ldots,\boldsymbol{a}_K(\omega))\in\mathbb{C}^{M\times K}$ を混合行列と呼び，以下ではこれを時不変と仮定する。$\boldsymbol{n}(\omega,t)$ は，多数の方向から到来する背景雑音や，フレーム長を超える残響成分など，時不変な伝達特性として表現できない成分を表す。ここで，音声のスパース性が仮定できる場

* Bayesian nonparametric approach to underdetermined sparse BSS. by Hirokazu KAMEOKA, (University of Tokyo/NTT), Misa SATO, Takuma ONO, Nobutaka ONO, Shigeki SAGAYAMA (University of Tokyo)

合，各時間周波数ビン (ω, t) においてアクティブ (支配的) となる音源インデックスを $z_{\omega,t} \in \{1, \ldots, K\}$ で表すことにすると，式 (1) は

$$\boldsymbol{y}(\omega, t) = \boldsymbol{a}_{z_{\omega,t}}(\omega) s(\omega, t) + \boldsymbol{n}(\omega, t) \tag{2}$$

のように書き直せる。この観測モデルでは，$z_{\omega,t}$ 番目の音源以外の成分はすべて 0 であると仮定しているので，各時間周波数ビンにおいて音源成分を表す変数は一つだけで十分である。このため上式では $s_k(\omega, t)$ のインデックス k を省いている。すなわち，$s(\omega, t)$ は特定の音源の成分ではなく，各時間周波数ビンにおいてアクティブないずれかの音源の成分を表す変数である。紙面のスペースの節約のため，以後 ω と t を下付き添え字で表記することにする。

3 生成モデル

3.1 観測信号の生成プロセス

前節で立てた観測モデルをもとに，観測信号が生成されるプロセスを生成モデルにより記述する。

まず，雑音成分 $\boldsymbol{n}_{\omega,t}$ は，平均が $\boldsymbol{0}$，共分散が $\Sigma_\omega^{(n)}$ の複素正規分布に従うと仮定すると，もし $\boldsymbol{a}_{1:K,\omega} = \{\boldsymbol{a}_{1,\omega}, \ldots, \boldsymbol{a}_{K,\omega}\}$, $s_{\omega,t}$，および，各時間周波数ビンでどの音源がアクティブであるか，すなわち $z_{\omega,t}$ が既知であれば，式 (2) より，$\boldsymbol{y}_{\omega,t}$ は

$$\boldsymbol{y}_{\omega,t} | \boldsymbol{a}_{1:K,\omega}, s_{\omega,t}, z_{\omega,t} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{a}_{z_{\omega,t},\omega} s_{\omega,t}, \Sigma_\omega^{(n)}) \tag{3}$$

により生成される。

3.2 無限音源数混合モデル

前節では $z_{\omega,t}$ が既知の下での観測信号の生成プロセスを仮定したが，通常は各時間周波数ビンでどの音源がアクティブであるかに関する情報は観測することができない。そこで本節では，アクティブな音源インデックスを示すインジケータ $z_{\omega,t}$ を潜在変数と見なし，その生成プロセスをモデル化する。まず，音源数が K の場合，$z_{\omega,t}$ は，音源インデックスの集合 $\{1, \ldots, K\}$ からいずれかのインデックスがある離散分布に従って選ばれるプロセス

$$z_{\omega,t} | \boldsymbol{\pi} \sim \mathrm{Discrete}(\boldsymbol{\pi}) \tag{4}$$

によって生成されると仮定する。ただし，$\boldsymbol{\pi} = (\pi, \ldots, \pi_K)$ は離散分布における各インデックスの選ばれやすさを意味する確率値であり，$\sum_{k=1}^{K} \pi_k = 1$ とする。さらに，これらの確率値は Dirichlet 分布

$$\boldsymbol{\pi} \sim \mathrm{Dirichlet}(\alpha_0/K, \ldots, \alpha_0/K) \tag{5}$$

よって生成されると仮定する。

ここまでは有限個の音源数 K を想定していたが，ここで，$K \to \infty$ の極限をとると，式 (3), (4), (5) は Dirichlet 過程混合モデルとなり，

$$\boldsymbol{\pi} \sim \mathrm{GEM}(\alpha_0) \tag{6}$$

$$z_{\omega,t} | \boldsymbol{\pi} \sim \mathrm{Discrete}(\boldsymbol{\pi}) \tag{7}$$

$$\boldsymbol{a}_{k,\omega} | H \sim H \tag{8}$$

$$\boldsymbol{y}_{\omega,t} | \boldsymbol{a}_{1:\infty,\omega} s_{\omega,t}, z_{\omega,t} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{a}_{z_{\omega,t},\omega} s_{\omega,t}, \Sigma_\omega^{(n)}) \tag{9}$$

と表すことができる。$\mathrm{GEM}(\alpha_0)$ は Dirichlet 過程により生成される $\pi_1, \pi_2, \ldots$ の一つの具体的な構成方法 (棒折過程と呼ぶ。) であり，以下に従って与えられる。

$$v_k \sim \mathrm{Beta}(1, \alpha_0) \tag{10}$$

$$\pi_k = v_k \prod_{l=1}^{k-1} (1 - v_l) \tag{11}$$

以上のプロセスで生成される $\pi_1, \pi_2, \ldots$ は，平均的な意味で，k が大きいほど π_k が指数的に小さくなるという傾向を持つため，大きい k に対応した音源ほどアクティブになる確率が低くなることを意味する。よって，観測信号からパラメータを推論する際，必要最小限の音源インデックス数の混合モデルで観測信号を説明しようとする効果がもたらされる。以上の $\boldsymbol{y}_{\omega,t}$ の生成モデルを「無限音源数混合モデル」と呼ぶ。

3.3 混合 DOA モデル

以上のモデルにパーミュテーション整合機能を組み込むため，式 (9) の生成プロセスを以下でモデル化する。ここまで各音源の伝達周波数特性 $\boldsymbol{a}_{k,\omega}$ を周波数 ω ごとの独立な変数であるかのように扱っていたが，もし各音源が単一方向から平面波到来すると仮定できるならば，例えばマイクロホン数が 2 の場合，伝達周波数特性の各 ω 間の関係は，到来方向 (Direction-of-Arrival; DOA)θ の関数として

$$\boldsymbol{\ell}_{\theta,\omega} = \begin{bmatrix} 1 \\ e^{\jmath \omega d \cos\theta / c} \end{bmatrix} \tag{12}$$

と陽に表される。ただし，$0 \leq \theta < 2\pi$，d をマイクロホンの間隔 (m)，c を音速 (m/s) とする。実際には残響や周波数領域の瞬時混合近似などの影響により，$\boldsymbol{a}_{k,\omega}$ は上記の理論式からは逸脱することが予想される。そこで，到来方向 θ_k が既知のとき，$\boldsymbol{a}_{k,\omega}$ は，$\boldsymbol{\ell}_{\theta_k,\omega}$ を中心とした複素正規分布より生成されたものと仮定する。しかし当然ながら到来方向 θ_k は実際には観測することができないため，これを潜在変数と見なすことにすると，$\boldsymbol{a}_{k,\omega}$ の生成モデルは，DOA を潜在変数とした混合モデルとなる。これを前節の無限音源数混合モデルに組み込み，観測信号が与えられた下で全体の生成モデルのパラメータ推論を行う

ことができれば，音源分離とパーミュテーション整合を同時解決できる可能性がある。

まず，$\Theta_1,\ldots,\Theta_M$(すべて定数) からなる M 個の DOA 候補値の集合を用意する。例えば，180 度を M 等分した角度 $\Theta_m=(m-1)\pi/M$, $(m=1,\ldots,M)$ の集合としよう。各音源の DOA がこの DOA 候補値の中から一つ選ばれて決定される，というプロセスを仮定するなら，θ_k が生成されるプロセスは以下のように記述できる。

$$x_k|\boldsymbol{\rho}\sim\mathrm{Discrete}(\rho_1,\ldots,\rho_M) \tag{13}$$

$$\theta_k=\Theta_{x_k} \tag{14}$$

ただし，$\boldsymbol{\rho}=(\rho_1,\ldots,\rho_K)$ である。$x_k\in\{1,\ldots,M\}$ は k 番目の音源にどの DOA 候補値が割り当てられるかを表すインジケータ変数であり，上式はこれが離散分布 (各確率値が $\rho_1,\ldots,\rho_M$) から生成されることを意味している。このプロセスにより各音源の DOA が決定され，伝達周波数特性 $\boldsymbol{a}_k(\omega)$ は，

$$\boldsymbol{a}_k(\omega)|x_k\sim\mathcal{N}_{\mathbb{C}}(\boldsymbol{\ell}_{\theta_k,\omega},\Sigma_\omega^{(a)}) \tag{15}$$

により生成される。また，ここで，$\boldsymbol{\rho}$ の事前分布として Dirichlet 分布

$$\boldsymbol{\rho}\sim\mathrm{Dirichlet}(\beta_0/M,\ldots,\beta_0/M) \tag{16}$$

を仮定する。以上の，混合モデルに基づく $\boldsymbol{a}_k(\omega)$ の生成モデルを「混合 DOA モデル」と呼ぶ。

4 変分推論アルゴリズム

観測信号 $\boldsymbol{y}=\boldsymbol{y}_{1:\Omega,1:T}$ が与えられたもとで，以上の生成モデルのパラメータ $\boldsymbol{a}=\boldsymbol{a}_{1:\infty,1:\Omega}$, $\boldsymbol{s}=s_{1:\Omega,1:T}$, $\boldsymbol{z}=z_{1:\Omega,1:T}$, $\boldsymbol{v}=v_{1:\infty}$, $\boldsymbol{x}=x_{1:\infty}$, $\boldsymbol{\rho}$ の事後分布 $p(\boldsymbol{a},\boldsymbol{s},\boldsymbol{z},\boldsymbol{v},\boldsymbol{x},\boldsymbol{\rho}|\boldsymbol{y})$ を求めたい。この事後分布を解析的に得ることは難しいが，変分推論法に基づき近似分布を反復計算により得ることができる。以下，簡単のため，$\Sigma_{1:\Omega}^{(n)}$, $\Sigma_{1:\Omega}^{(a)}$, α_0, β_0 は実験的に定める定数とする。

変分推論は，事後分布 $p(\boldsymbol{a},\boldsymbol{s},\boldsymbol{z},\boldsymbol{v},\boldsymbol{x},\boldsymbol{\rho}|\boldsymbol{y})$ と，

$$\int\cdots\int q(\boldsymbol{a},\boldsymbol{s},\boldsymbol{z},\boldsymbol{v},\boldsymbol{x},\boldsymbol{\rho})\mathrm{d}\boldsymbol{a}\cdots\mathrm{d}\boldsymbol{\rho}=1 \tag{17}$$

を満たす非負の変関数 $q(\boldsymbol{a},\boldsymbol{s},\boldsymbol{z},\boldsymbol{v},\boldsymbol{x},\boldsymbol{\rho})$ との間の Kullback-Leibler ダイバージェンス

$$\mathcal{F}[q]=\left\langle\log\frac{p(\boldsymbol{a},\boldsymbol{s},\boldsymbol{z},\boldsymbol{v},\boldsymbol{x},\boldsymbol{\rho}|\boldsymbol{y})}{q(\boldsymbol{a},\boldsymbol{s},\boldsymbol{z},\boldsymbol{v},\boldsymbol{x},\boldsymbol{\rho})}\right\rangle_{q(\boldsymbol{a},\boldsymbol{s},\boldsymbol{z},\boldsymbol{v},\boldsymbol{x},\boldsymbol{\rho})} \tag{18}$$

を q に関して最小化することが目的となる。ただし $\langle f(x)\rangle_{q(x)}$ は $\int q(x)f(x)\mathrm{d}x$ を表す。そして q に関して

$$q(\boldsymbol{a},\boldsymbol{s},\boldsymbol{z},\boldsymbol{v},\boldsymbol{x},\boldsymbol{\rho})\simeq q(\boldsymbol{a})q(\boldsymbol{s})q(\boldsymbol{z})q(\boldsymbol{v})q(\boldsymbol{x})q(\boldsymbol{\rho}) \tag{19}$$

のように近似できると仮定し，$q(\boldsymbol{a})$, $q(\boldsymbol{s})$, $q(\boldsymbol{z})$, $q(\boldsymbol{v})$, $q(\boldsymbol{x})$, $q(\boldsymbol{\rho})$ について反復的に $\mathcal{F}[q]$ を最小化することで $p(\boldsymbol{a},\boldsymbol{s},\boldsymbol{z},\boldsymbol{v},\boldsymbol{x},\boldsymbol{\rho}|\boldsymbol{y})$ の近似分布を得ようというのが変分推論法の基本的な考え方である。また，$q(\boldsymbol{z})$ に関して，以下の打ち切り近似を行う。

$$q(z_{\omega,t}=K^*+1)=\cdots=q(z_{\omega,t}=\infty)=0 \tag{20}$$

この近似は，モデルの複雑度 (音源数) を固定した，ということではなく，q の関数空間をある領域に限定した，ということを意味している。よって，本来推定したい $p(\boldsymbol{a},\boldsymbol{s},\boldsymbol{z},\boldsymbol{v},\boldsymbol{x},\boldsymbol{\rho}|\boldsymbol{y})$ をできるだけ良く q でフィッティングしたければ，K^* は大きければ大きいほど良い，ということになる。

導出は省略するが，式 (18) を式 (17) の拘束の下で最小化する各 q は解析的に以下の形として求まる。

$$\hat{q}(\boldsymbol{a})=\prod_{k,\omega}\mathcal{N}_{\mathbb{C}}(\boldsymbol{a}_{k,\omega};\boldsymbol{m}_{k,\omega},\Gamma_{k,\omega}) \tag{21}$$

$$\hat{q}(\boldsymbol{s})=\prod_{\omega,t}\mathcal{N}_{\mathbb{C}}(s_{\omega,t};\mu_{\omega,t},\sigma_{\omega,t}^2) \tag{22}$$

$$\hat{q}(\boldsymbol{z})=\prod_{\omega,t}\hat{q}(z_{\omega,t}),\ \hat{q}(z_{\omega,t}=k)=\phi_{k,\omega,t} \tag{23}$$

$$\hat{q}(\boldsymbol{v})=\prod_{k}\mathrm{Beta}(v_k;\gamma_{k,0},\gamma_{k,1}) \tag{24}$$

$$\hat{q}(\boldsymbol{x})=\prod_{k}\hat{q}(x_k),\ \hat{q}(x_k=m)=\psi_{k,m} \tag{25}$$

$$\hat{q}(\boldsymbol{\rho})=\prod_{k}\mathrm{Dirichlet}(\zeta_{k,1},\ldots,\zeta_{k,M}) \tag{26}$$

5 実験

提案法の有効性を示すため，音源分離性能の検証を行った。3 人の話者 (女性 2 人，男性 1 人) の音声信号 [7] に，室内インパルス応答 (残響時間は 0 ms)[8] を畳み込み加算することで人工的に混合したものを観測信号とした。標本化周波数は 16 kHz とした。観測信号の時間周波数成分は，短時間 Fourier 変換 (フレーム長は 64 ms，フレームシフトは 16 ms) により算出した。$\Sigma_\omega^{(n)}$ と $\Sigma_\omega^{(a)}$ はそれぞれ $\boldsymbol{I}$, $10^{-1.5}\times\boldsymbol{I}$ とした。また，角度の分割数は $M=180$ とした。4 章の反復アルゴリズムの実行後，音源成分の推定値 $\mu_{\omega,t}$ に，音源 k が時間周波数ビンでどれだけアクティブらしいかを表す確率値 $\phi_{k,\omega,t}$ を乗じたものを，音源 k の推定時間周波数成分とした。また，今回の実験では，空間エイリアシングにより ζ が局所解に陥ってしまう可能性を考慮し，反復計算の初期段階では空間エイリアシングが起こらない低い帯域の観測情報のみを用いてアルゴリズムを実行し，反復回数の増加に従って徐々にその帯域を高帯域に広げていく方法をとった。音源分離性能の評価基準として，Signal-to-Distortion Ratio (SDR) と Signal-to-Interference

Table 1　打ち切りレベル $K^* = 3, 30$ の場合の提案法による分離性能 (単位はすべて dB)

K^*	音源番号	SDR	SIR	SAR
3	1	-2.55	1.03	2.48
	2	-4.50	2.48	-1.59
	3	2.44	10.67	3.50
	平均	-1.54	4.73	1.46
30	1	5.22	45.64	5.26
	2	5.32	26.44	5.42
	3	8.16	25.74	-6.80
	平均	6.23	32.61	5.34

Ratio (SIR) と Signal-to-Artifact Ratio (SAR) を採用した [9]。

まず，提案法の第一のポイントである，必要な音源数を適応的に推論する効果を検証した。また，打ち切りレベル K^* によって分離性能がどう影響するかを併せて確認した。$K^* = 3$(音源数と同数) の場合と $K^* = 30$ の場合における SDR，SIR，SAR を表 1 に示す。表 1 のとおり，$K^* = 30$ の場合の方がはるかに高い性能を得た。これは，先に述べたとおり，打ち切りレベル K^* が大きければ大きいほど q が真の事後分布の良い近似が得られる，ということを示した一つの例証である。また，小さいインデックス (k=1, 2, 3) に 3 音源の信号が自動的に集まる傾向は Dirichlet 過程による効果である。

次に，混合 DOA モデルの効果を見るため，音源の到来方向を正しく推定できているらしいかを確認した。Fig. 1 は，観測信号の合成に利用した室内インパルス応答のチャンネル間位相差と，推定した伝達周波数特性 $\boldsymbol{m}_{k,\omega}$ より算出されるチャンネル間位相差 $\arg([\boldsymbol{m}_{k,\omega}]_1/[\boldsymbol{m}_{k,\omega}]_2)$ を音源ごとに異なる色でプロットしたものである。ただし，$[\cdot]_i$ はベクトルの i 番目の要素を表す。空間エイリアシングがあっても各音源の到来方向が概ね正しく推定できていることが分かる。

6　おわりに

本稿では，音源数が変動したり事前に音源数に関する情報を知ることができない場合にも安定して動作する BSS アルゴリズムの実現を目指し，音源数を仮定することなく観測信号から適応的に音源数を推論しながら音源分離を行える手法を検討した。音声の時間周波数成分のスパース性に基づく周波数領域の劣決定 BSS モデルをベイズ的に記述し，Dirichlet 過程混合モデルにより観測信号の生成プロセスをモデル化したことで，音源数に合わせてモデルの複雑度を適応させながら音源分離を行える点，混合 DOA モデルと呼ぶ伝達周波数特性の生成モデルを導入したことで，周波数ごとの信号分離とパーミュテーション整合を同時に行える点，が提案法の主要な特徴である。

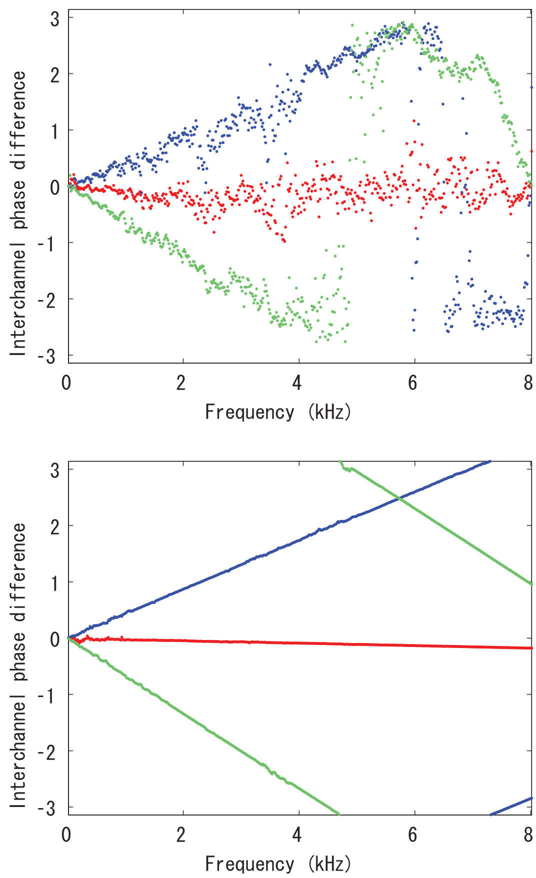

Fig. 1　正解チャンネル間位相差 (上) と推定チャンネル間位相差 $\arg([\boldsymbol{m}_{k,\omega}]_1/[\boldsymbol{m}_{k,\omega}]_2)$(下)

参考文献

[1] A. Hyvärinen *et al.*, “Independent component analysis,” John Wiley, New York, 2001.
[2] Yilmaz & Rickard, IEEE Trans. SP, **52**(7), pp. 1830–1847, 2004.
[3] Mandel et al., Adv. Neural Inf. Proc. Sys., 2006.
[4] Araki et al., Signal Process., Vol. 87, pp. 1833–1847, 2007.
[5] Mori et al., Proc. IWAENC’05, pp. 229–232, 2005.
[6] 和泉他, 音講論 (春)’07, 2-1-5, pp. 555–556, 2007.
[7] A. Kurematsu *et al.*, *Trans. Speech Communication*, pp. 357–363, 1990.
[8] S. Nakamura *et al.*, Proc. LREC, pp. 965–968, 2000.
[9] E. Vincent *et al.*, *Trans. ASLP*, pp.1462–1469, 2006.