多チャンネルウィーナーフィルタを用いた音源分離における観測モデルの調査*

☆坂梨龍太郎, 宮部滋樹, 山田武志, 牧野昭二 (筑波大)

1 はじめに

音源分離はハンズフリー音声認識やコンピュータによる音環境理解のために不可欠な技術である．特に，音源の位置や話者の情報，また音源からの伝達関数など事前情報を必要としないブラインド音源分離（BSS: Blind Source Separation）はここ 10 年ほどで大きく進展している [1]．ブラインド音源分離は独立成分分析 [2] と呼ばれる手法が代表的であるが，この手法は音源数が観測信号数以下である場合にしか適用できない．音源の数が観測信号数よりも多い条件での音源分離を劣決定 BSS と呼び，これを実現する手法として一般的にスパース性の仮定を用いることが多い [3]．スパース性とは，対象とする信号のエネルギーが各時間周波数において一部の領域に集中し多くの領域でほぼ 0 であるような性質であり，音声の混合は時間周波数領域ではこの性質がよく成り立つため，劣決定 BSS で幅広く利用されている．このスパース性の仮定を利用した代表的な手法としては，時間周波数マスキングが挙げられる [3]．これは，分離対象とする信号が支配的である時間周波数スロットを推定し，その時間周波数成分をマスキング処理により分離する手法であり，これを用いた様々な手法が提案されている [4] [5] [6] [7]．

従来の劣決定音源分離ではスパース性を仮定するのが一般的であったのに対し，スパース性を入れないモデル化により高音質な分離を達成する多チャンネルウィーナーフィルタ [8] が Duong らにより提案されている．この手法は音源と伝達関数の積である音像が多変量正規分布に従うという仮定により，時不変の空間相関行列と時変の分散を EM アルゴリズムにより推定し，それらのパラメタを用いた多チャンネルウィーナーフィルタにより音源分離を行う．文献 [8] にはその最尤推定が明示的に定式化されていないが，Togami が [9] で指摘しているようにその EM アルゴリズムの Q 関数では隠れ変数として扱われる音像が多チャンネルの複素振幅のまま重ね合わせられて観測信号を生成するというモデル化となっていて，近似的にしか成り立たないスパース性の仮定を排している．

本稿では，Duong 法が高品質な音源分離を達成できる理由として考えられるいくつかの要因について詳しく分析する．具体的には，1）多変量正規分布による音像モデルを用いていること，2）多チャンネルウィーナーフィルタによる分離であること，3）スパース性を仮定せず導出した EM アルゴリズムであること，の 3 つの特徴に着目する．これらの有効性を確認するため，Duong 法と同様に音像を正規分布でモデル化しつつも，複数の音像の重ね合わせではなく単一の音像がスパースに出現して観測信号を形成する，というモデルの最尤推定を行う EM アルゴリズムを定式化し，推定したパラメタから，A）音源がアクティブである事後確率を用いた分離，B）最大事後確率（MAP）推定である最尤バイナリマスクによる分離，C）多チャンネルウィーナーフィルタによる分離の 3 手法を実装する．これを Duong 法及び一般的なバイナリマスク手法である MENUET と比較し，i）代表的なバイナリマスク手法である MENUET と定式化した最尤バイナリマスクの比較による多変量正規分布の音像モデル化の妥当性，ii）上記の A），B），C）3 手法の比較による多チャンネルウィーナーフィルタの有効性，iii）定式化した多チャンネルウィーナーフィルタと Duong 法の比較によるスパース性を仮定しない音像重ね合わせモデルによる推定の有効性，以上 3 点の検証を行った．その結果，上記 1）～3）全てが Duong 法の高品質な音源分離を達成できる要因と言えることが確認された．

2 章では本稿で扱う観測モデルについて述べる．3 章では Duong 法の概要について述べる．4 章では比較手法であるスパースモデルの定式化について述べる．5 章では Duong 法の有効性を確認するための音源分離の比較実験を行う．6 章では本稿の結論を述べる．

2 観測モデル

ここでは本稿で扱う Duong 法と新たに定式化したスパースモデルによる手法における観測モデルについての説明を行う．まず観測信号は，j 番目の原音源 $s_j(n,f)$，伝達関数ベクトル $\mathbf{h}_j(f)$ により，

$$\begin{aligned}\mathbf{x}(n,f) &= [x_1(n,f),...,x_I(n,f)]^T \\ &\approx \sum_{j=1}^{J} \mathbf{h}_j(f) s_j(n,f), \\ \mathbf{h}_j(f) &= [h_{1j}(f),...,h_{Ij}(f)]^T\end{aligned} \tag{1}$$

として表される．ここで h_{ij} は j 番目の音源から i 番目のチャンネルへの伝達関数，J は音源数，I はチャンネル数である．また，残響などの空間特性を含んだ信号を，音像（source image）と呼び，j 番目の音像を以下のように定義する．

$$\mathbf{c}_j(n,f) = \mathbf{h}_j(f) s_j(n,f) \tag{2}$$

本稿では各音源の音像 $\mathbf{c}_j(n,f)$ を推定する問題を取り扱う．

3 Duong らの多チャンネルウィーナーフィルタ

ここでは Duong 法のパラメタ推定と音源分離について，文献 [8] で明示的に示されていなかった最尤推定の定式化を含めて説明する．

3.1 問題設定

Duong 法はモデルパラメタ推定を用いて音像 $\mathbf{c}_j(n,f)$ の推定を行う．本来音像は，式（2）のように表されるが，分析フレームの中に納まらない長い残響などの影響を考慮して伝達関数が時変であると仮定し，

$$\mathbf{c}_j(n,f) = \mathbf{h}_j(n,f) s_j(n,f) \tag{3}$$

と表し，伝達関数ベクトルも時変とする．このように変動する個々の音像 $\mathbf{c}_j(n,f)$ の事前確率密度がモデルパラメタ $\theta(f) = \{\nu_j(n,f), \mathbf{R}_j(f)\}$ による複素正規分布

$$p(\mathbf{c}_j(n,f)|\theta(f)) = \mathcal{N}_{\mathrm{c}}(\mathbf{c}_j(n,f); \mathbf{0}, \mathbf{R}_{\mathbf{c}_j}(n,f)) \tag{4}$$

---

*Research of the observation model in source separation using multichannel Wiener filter. by SAKANASHI, Ryutaro, MIYABE, Shigeki, YAMADA, Takeshi, MAKINO, Shoji (University of Tsukuba)

で表されると仮定する．ここで，$\mathcal{N}_{\mathrm{c}}(\cdot)$ は多変量複素正規分布の確率密度関数を表す．また，音像の共分散行列 $\mathbf{R}_{\mathbf{c}_j}(n,f)$ は，

$$\mathbf{R}_{\mathbf{c}_j}(n,f) = \nu_j(n,f)\mathbf{R}_j(f) \tag{5}$$

のように，$\mathbf{R}_{\mathbf{c}_j}(n,f)$ は時間依存の時変の分散 $\nu_j(n,f)$ と時不変の空間相関行列 $\mathbf{R}_j(f)$ として表せると仮定する。空間相関行列 $\mathbf{R}_j(f)$ は時不変の伝達関数 $\mathbf{h}_j(f)$ を仮定すると $\mathbf{R}_j(f) = \mathbf{h}_j(f)\mathbf{h}_j^H(f)$ と表せてランクが 1 となるが，分析フレームの中に納まらない長い残響などの影響を考慮し，伝達関数による制約は設けずフルランクであると仮定している．以上のように $\mathbf{c}_j(n,f)$ が生成されると仮定した場合の最尤推定を行う．

3.2 モデルパラメタ推定

式（1），(2），(4）を用いると観測信号が全ての音像の重ね合わせとして観測される確率は，

$$\begin{aligned} &p(\mathbf{C}(n,f),\mathbf{x}(n,f)|\theta(f)) \\ &= \prod_{j=1}^{J-1} \mathcal{N}_{\mathrm{c}}\left(\mathbf{c}_j(n,f);\mathbf{0},\mathbf{R}_{\mathbf{c}_j}(n,f)\right) \\ &\quad \cdot \mathcal{N}_{\mathrm{c}}\left(\mathbf{x}(n,f) - \sum_{j=1}^{J}\mathbf{c}_j(n,f);\mathbf{0},\mathbf{R}_{\mathbf{c}_J}(n,f)\right) \end{aligned} \tag{6}$$

で与えられる．ここで $\mathbf{C}(n,f) = \{\mathbf{c}_1(n,f),...,\mathbf{c}_j(n,f),...,\mathbf{c}_J(n,f)\}$ である．式（6）を全ての音像について周辺化することにより観測尤度

$$p(\mathbf{x}(n,f)|\theta(f)) = \mathcal{N}_{\mathrm{c}}(\mathbf{x}(n,f);\mathbf{0},\mathbf{R}_{\mathbf{x}}(n,f)) \tag{7}$$

が得られる．ここで，$\mathbf{R}_{\mathbf{x}}(n,f)$ は観測信号 $\mathbf{x}(n,f)$ の共分散行列であり，

$$\mathbf{R}_{\mathbf{x}}(n,f) = \sum_{j=1}^{J}\nu_j(n,f)\mathbf{R}_j(f) \tag{8}$$

で表される．この尤度関数により，Q 関数は，

$$\begin{aligned} &Q(\theta(f),\bar{\theta}(f)) \\ &= \sum_{n=1}^{N}\int p\left(\mathbf{C}(n,f)|\mathbf{x}(n,f),\theta(f)\right) \\ &\qquad \log p\left(\mathbf{C}(n,f),\mathbf{x}(n,f)|\bar{\theta}(f)\right) d\mathbf{C}_J(n,f) \\ &= \sum_{n=1}^{N}\int \frac{p(\mathbf{C}(n,f),\mathbf{x}(n,f)|\theta(f))}{p\left(\mathbf{x}(n,f)|\theta(f)\right)} \\ &\qquad \log p\left(\mathbf{C}(n,f),\mathbf{x}(n,f)|\bar{\theta}(f)\right) d\mathbf{C}_J(n,f) \\ &= \sum_{n=1}^{N}\left(-IJ\log\pi - I\sum_{j=1}^{J}\log\bar{\nu}_j(n,f)\right. \\ &\qquad - \sum_{j=1}^{J}\log\det(\bar{\mathbf{R}}_j(f)) \\ &\qquad \left. - \sum_{j=1}^{J}\frac{1}{\bar{\nu}_j(n,f)}\mathrm{Tr}\left(\mathbf{M}_j(n,f)\bar{\mathbf{R}}_j^{-1}(f)\right)\right), \end{aligned} \tag{9}$$

$$\begin{aligned} &\mathbf{M}_j(n,f) \\ &= \mathbf{R}_{\mathbf{c}_j}(n,f) - \mathbf{R}_{\mathbf{c}_j}(n,f)\mathbf{R}_{\mathbf{x}}^{-1}(n,f)\mathbf{R}_{\mathbf{c}_j}(n,f) \\ &\quad + \mathbf{R}_{\mathbf{c}_j}(n,f)\mathbf{R}_{\mathbf{x}}^{-1}(n,f)\mathbf{x}(n,f) \\ &\qquad \mathbf{x}^H(n,f)\mathbf{R}_{\mathbf{x}}^{-1}(n,f)\mathbf{R}_{\mathbf{c}_j}(n,f) \end{aligned} \tag{10}$$

となる．ここで $\mathbf{C}_J(n,f) = \{\mathbf{c}_1(n,f),...,\mathbf{c}_{J-1}(n,f)\}$ である．

3.3 音源分離

以上の Q 関数をパラメタで偏微分することにより，以下の EM アルゴリズムが与えられる．

E-step では以下の更新

$$\mathbf{W}_j(n,f) = \mathbf{R}_{\mathbf{c}_j}(n,f)\mathbf{R}_{\mathbf{x}}^{-1}(n,f) \tag{11}$$

$$\hat{\mathbf{c}}_j(n,f) = \mathbf{W}_j(n,f)\mathbf{x}(n,f) \tag{12}$$

$$\begin{aligned} \hat{\mathbf{R}}_{\mathbf{c}_j}(n,f) &= \hat{\mathbf{c}}_j(n,f)\hat{\mathbf{c}}_j^H(n,f) \\ &\quad + (\mathbf{I} - \mathbf{W}_j(n,f))\mathbf{R}_{\mathbf{c}_j}(n,f) \end{aligned} \tag{13}$$

を行い，M-step では以下の更新

$$\nu_j(n,f) = \frac{1}{I}\mathrm{tr}(\mathbf{R}_j^{-1}(f)\hat{\mathbf{R}}_{\mathbf{c}_j}(n,f)) \tag{14}$$

$$\mathbf{R}_j(f) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{\nu_j(n,f)}\hat{\mathbf{R}}_{\mathbf{c}_j}(n,f) \tag{15}$$

$$\mathbf{R}_{\mathbf{c}_j}(n,f) = \nu_j(n,f)\mathbf{R}_j(f) \tag{16}$$

$$\mathbf{R}_{\mathbf{x}}(n,f) = \sum_{j=1}^{J}\nu_j(n,f)\mathbf{R}_j(f) \tag{17}$$

を行う．これを繰り返すことによりパラメタを推定する．ここで $\mathbf{I}$ は単位行列である．EM アルゴリズムによるパラメタの推定後にそれらを用いて音像の推定を行う．$p(\mathbf{C}(n,f)|\mathbf{x}(n,f),\theta(f))$ を $\mathbf{c}_1(n,f),...,\mathbf{c}_{j-1}(n,f),\mathbf{c}_{j+1}(n,f),...,\mathbf{c}_J(n,f)$ について周辺化を行うと，

$$\begin{aligned} &p(\mathbf{c}_j(n,f)|\mathbf{x}(n,f),\theta(f)) \\ &= \mathcal{N}_{\mathrm{c}}\Big(\mathbf{c}_j(n,f);\mathbf{R}_{\mathbf{c}_j}(n,f)\mathbf{R}_{\mathbf{x}}^{-1}(n,f)\mathbf{x}(n,f), \\ &\qquad \left(\mathbf{R}_{\mathbf{c}_j}^{-1}(n,f) + \left(\mathbf{R}_{\mathbf{x}}(n,f) - \mathbf{R}_{\mathbf{c}_j}(n,f)\right)^{-1}\right)^{-1}\Big) \end{aligned} \tag{18}$$

となるため，求められたパラメタによる期待値または最大事後確率（MAP: Maximum A Posteriori）推定として各音源の音像が

$$\mathbf{c}_j(n,f) = \mathbf{R}_{\mathbf{c}_j}(n,f)\mathbf{R}_{\mathbf{x}}^{-1}(n,f)\mathbf{x}(n,f) \tag{19}$$

というマルチチャネルウィーナーフィルタの形で求められる．

4 スパース性を仮定した観測モデルと分離手法の定式化

スパース性を仮定しない Duong 法の有効性を検証するための比較手法として，ここでは Duong 法と同様のモデルの音像がスパースに生成されて観測されると仮定した音源分離を定式化する．

4.1 問題設定

各時間周波数スロット (n,f) において 1 つの音像 $\mathbf{c}_j(n,f)$ だけアクティブであると仮定したとき，$z(n,f)$ を各時間周波数スロット (n,f) においてアクティブである音源番号であると定義し，第 j 番目の音源がアクティブになる確率を

$$p(z(n,f)=j) = \mu_j(f),\ \sum_{j=1}^{J}\mu_j(f) = 1 \tag{20}$$

と表す．この仮定の下で，観測信号 $\mathbf{x}(n,f)$ が

$$\mathbf{x}(n,f) = \mathbf{c}_{z(n,f)}(n,f) \tag{21}$$

として与えられるというモデル化は，スパース性を仮定した観測モデルとなる．このモデルにおける音像 $\mathbf{c}_j(n,f)$ の生成は，Duong 法と同じく時変の分散 $\nu_j(n,f)$ と空間相関行列 $\mathbf{R}_j(f)$ の積を共分散行列とし，スパースの仮定を含むモデルパラメタ $\theta(f) = \{\mathbf{R}_j(f), \nu_j(n,f), \mu_j(f), \text{for } j = 1, ..., J\}$ の正規分布に従うと仮定すると，

$$\begin{aligned} &p\left(\mathbf{c}_j(n,f)|z(n,f)=j,\theta(f)\right) \\ &= \mathcal{N}_{\mathrm{c}}\left(\mathbf{c}_j\left(n,f\right);\mathbf{0},\nu_j\left(n,f\right)\mathbf{R}_j(f)\right) \end{aligned} \tag{22}$$

で表される．

4.2 モデルパラメタ推定

モデルパラメタ $\theta(f)$ を最大化する観測信号 $\mathbf{x}(n,f)$ の尤度 $\prod_n p(\mathbf{x}(n,f)|\theta(f))$ は以下のように定義される．

$$\begin{aligned} &\prod_n p\left(\mathbf{x}(n,f)|\theta(f)\right) \\ &= \prod_n \sum_{j=1}^{J} \mu_j(f)\mathcal{N}_{\mathrm{c}}\left(\mathbf{x}(n,f);\mathbf{0},\nu_j(n,f)\mathbf{R}_j(f)\right) \end{aligned} \tag{23}$$

この最尤推定は、以下で与えられる $z(n,f)$ を隠れ変数とした Q 関数を最大化する EM アルゴリズムにより求めることができる。

$$\begin{aligned} &Q_f(\theta(f),\bar{\theta}(f)) \\ &= \sum_{n,j} m_j(n,f)\log\mu_j(f) \\ &\qquad \cdot \mathcal{N}_{\mathrm{c}}(\mathbf{x}(n,f);\mathbf{0},\nu_j(n,f)\mathbf{R}_j(f)) \\ &= \sum_{n,j} m_j(n,f)\Bigg(\log\mu_j(f) - I\log\pi - I\log\nu_j(n,f) \\ &\quad - \log\det\left(\bar{\mathbf{R}}_j(f)\right) - \frac{\mathbf{x}(n,f)^{\mathrm{H}}\bar{\mathbf{R}}_j^{-1}(f)\mathbf{x}(n,f)}{\nu_j(n,f)} \Bigg) \end{aligned} \tag{24}$$

ここで $m_j(n,f)$ は事後確率を与える分配関数（連続的な時間周波数マスク）

$$m_j(n,f) = p\left(z(n,f)=j|\mathbf{x}(n,f),\mathbf{R}_j(f),\nu_j(n,f)\right) \tag{25}$$

であり，

$$\sum_{j=1}^{J} m_j(n,f) = 1 \tag{26}$$

を満たす．この Q 関数の各パラメタでの偏微分を 0 とおくことにより EM アルゴリズムが導かれる．

M-step では以下の変数が更新される．

$$\nu_j(n,f) \leftarrow \frac{\mathbf{x}^H(n,f)\mathbf{R}_j^{-1}(f)\mathbf{x}(n,f)}{I} \tag{27}$$

$$\mathbf{R}_j(f) \leftarrow \frac{\sum_n \frac{m_j(n,f)}{\nu_j(n,f)}\mathbf{x}(n,f)\mathbf{x}^H(n,f)}{\sum_n m_j(n,f)} \tag{28}$$

$$\mu_j(f) \leftarrow \frac{\sum_n m_j(n,f)}{\sum_{n,j'} m_{j'}(n,f)} \tag{29}$$

E-step では以下の変数が更新される．

$$\begin{aligned} &m_j(n,f) \\ &\leftarrow \frac{\mu_j(f)\mathcal{N}_{\mathrm{c}}\left(\mathbf{x}(n,f);\mathbf{0},\nu_j(n,f)\mathbf{R}_j(f)\right)}{\sum_{j'}\mu_{j'}(f)\mathcal{N}_{\mathrm{c}}\left(\mathbf{x}(n,f);\mathbf{0},\nu_{j'}(n,f)\mathbf{R}_{j'}(f)\right)} \end{aligned} \tag{30}$$

4.3 音源分離

以上のパラメタ推定による音源分離として，以下の 3 手法を示す．

4.3.1 事後確率を用いたスパースソフトマスクによる分離

ソフトマスクによる分離は，j 番目の音源がアクティブになる事後確率を与える分配関数 $m_j(n,f)$ による分離であり，分離音は以下の式で表される．

$$\mathbf{c}_j(n,f) = m_j(n,f)\mathbf{x}(n,f) \tag{31}$$

4.3.2 スパースバイナリマスクによる分離

バイナリマスクによる分離は，分配関数 $m_j(n,f)$ における各時間周波数で最大の確率を持つ音源の確率を 1，それ以外を 0 にしたバイナリマスク $M_j(n,f)$ を作成し分離することであり，分離音は以下の式で表される．

$$M_j(n,f) = \begin{cases} 1 & \text{if } \ j^* = \underset{j}{\operatorname{argmax}}\, m_j(n,f) \\ 0 & \text{otherwise} \end{cases} \tag{32}$$

$$\mathbf{c}_j(n,f) = M_j(n,f)\mathbf{x}(n,f) \tag{33}$$

これは $p(\mathbf{c}_1(n,f),...,\mathbf{c}_J(n,f)|\mathbf{x}(n,f),\mathbf{R}_1(f),...,\mathbf{R}_J(f),\nu_1(n,f),...,\nu_J(n,f))$ を最大にする $\mathbf{c}_1(n,f),...,\mathbf{c}_J(n,f)$ の最大事後確率推定に相当する．

4.3.3 スパースウィーナーフィルタによる分離

ここでは分散 $\nu_j(n,f)$ の期待値を用いて設計した Duong 法と同様のウィーナーフィルタについて述べる．分散 $\nu_j(n,f)$ の事後期待値は $m_j(n,f)\nu_j(n.f)$ となるため，これを用いて式（19）と同様のマルチチャネルウィーナーフィルタを設計する．

$$\mathbf{R}_{\mathbf{x}}(n,f) = \sum_j m_j(n,f)\nu_j(n,f)\mathbf{R}_j(f) \tag{34}$$

$$\mathbf{R}_{\mathbf{c}_j}(n,f) = m_j(n,f)\nu_j(n,f)\mathbf{R}_j(f) \tag{35}$$

$$\mathbf{c}_j(n,f) = \mathbf{R}_{\mathbf{c}_j}(n,f)\mathbf{R}_{\mathbf{x}}^{-1}(n,f)\mathbf{x}(n,f) \tag{36}$$

5 実験

5.1 実験条件

SiSEC 2011 [10] の音声データ男女それぞれ 4 音声，計 8 音声から 3 音声を選び，異なる組み合わせで 6 パターンの音声混合を行い，一般的なバイナリマスク手法である MENUET と，スパースモデルによる 3 手法と Duong 法の性能比較を行う．マイク間距離は 2.15 cm，音源方向は $-40°$，$0°$，$30°$，FFT のサイズは 1024 点，サンプリング周波数は 16 kHz，EM アルゴリズムの反復回数は 50 回である．分離性能の客観評価値には文献 [11] で示された，SDR (Signal to Distortion Ratio)：総合的な歪み，ISR (Source Image to Spatial distortion Ratio)：線形歪み，SIR (Source to Interference Ratio)：他話者音声の消し残りによる歪み，SAR (Sources to Artifacts Ratio)：非線形歪み，以上 4 つの歪み尺度を使用する．単位は dB であり，この歪み尺度の数値が高いほど性能が良い．Duong 法と定式化したスパース性を導入した手法の初期値生成とパーミュテーション解決には，文献 [12] の手法を用いた．

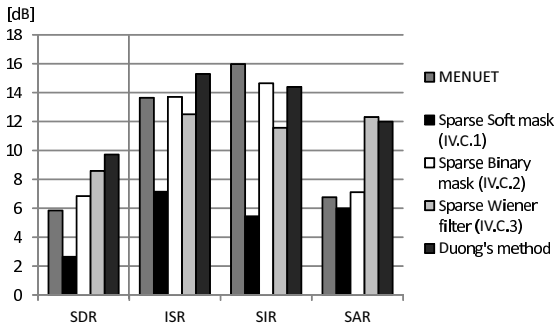

Fig. 1　実験結果

5.2　実験結果，考察

実験結果を図 1 に示す．スパースソフトマスクによる分離よりスパースバイナリマスクによる分離の方が性能は大幅に良く，スパースウィーナーフィルタによる分離はスパースバイナリマスクによる分離よりも SAR は大幅に良く SDR も高い．MENUET の性能はスパースバイナリマスクに劣るものの似たような傾向にあるが，SIR のみスパースバイナリマスクより高い値を持つ．Duong 法は SIR，SAR で他手法に劣るものの，ISR，また総合的な評価値である SDR の値は最も高い．

音像の多変量正規分布によるモデル化の妥当性を検証するため，代表的なバイナリマスク手法である MENUET と多変量正規分布の最大事後確率推定であるスパースバイナリマスクを比較する。全ての評価尺度が同程度であるため、音像の多変量正規分布によるモデル化は副作用がなく妥当なものであるということが確認できる。次に，どのような分離手法が有効であるかの検証を行うため，定式化したスパースモデルによる 3 手法を比較した．スパースソフトマスクは全ての評価値において最も低い値を持つ．また，スパースバイナリマスクは，スパースウィーナーフィルタよりも SIR が高く分離度が高いが，SAR においてはスパースウィーナーフィルタの値が大幅に高く，歪みが少ない分離であることから，総合的な品質である SDR におけるスパースウィーナーフィルタの値が高くなっている．これらのことから，スパースウィーナーフィルタによる分離が定式化したスパースモデルによる手法の中で最も性能が高いと言える。最後に，スパース性を仮定しない音像の重ね合わせモデルの有効性の検証を行うため，定式化したスパースウィーナーフィルタと Duong 法を比較する．Duong 法の SAR の値は少々劣るものの，ISR，SIR の値は高い値を示しており，スパースウィーナーフィルタよりも Duong 法は分離度が高く歪みも少ないことがわかる．これらの値が，総合的な評価値である SDR の値にも作用していることがわかる．このことから，スパース性を仮定した観測モデルよりも，Duong 法のスパース性を仮定しない音像の重ね合わせによる観測モデルの方が有効であるということが確認された．

6　結論

本稿では，Duong 法が高品質な音源分離を達成できる要因として考えられる，多変量正規分布による音像モデル，多チャンネルウィーナーフィルタによる音源分離とスパース性を仮定しない重ね合わせによる観測モデルの 3 つの特徴の効果を検証した．そのために，Duong 法と同様の多変量正規分布による音像モデルと，複数の音像の重ね合わせではなく単一の音像がスパースに出現して観測信号を形成するという観測モデルの最尤推定から，音源がアクティブである事後確率を用いたスパースソフトマスクによる分離，MAP 推定であるスパースバイナリマスクによる分離，スパースウィーナーフィルタによる分離，の 3 つの音源分離手法を定式化し，Duong 法及び一般的な時間周波数バイナリマスク手法である MENUET と性能を比較した．その結果，Duong 法の上記 3 つの特徴全てが Duong 法の高品質な音源分離を達成できる要因と言えることが確認された．

謝辞　本研究は科研費（２３２４００２３）の助成を受けたものである．また，本研究を遂行するにあたって有益なご助言を頂いた日立製作所中央研究所の戸上真人氏に感謝いたします．

参考文献

[1] H. Sawada, S. Araki, and S. Makino, “Recent advances in audio source separation techniques,” *J. IEICE,* Vol. 91, No. 4, pp. 292–296, 2008.

[2] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent component analysis,* Wiley, New York, 2001.

[3] O. Yilmaz, S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE Trans. Signal Process.,* vol. 52, No. 7, pages 1830–1847, 2004.

[4] S. Araki, H. Sawada, R. Mukai, and S. Makino, “Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors,” *Signal Processing,* no. 87, pp. 1833–1847, 2007.

[5] Y. Izumi, N. Ono, and S. Sagayama, “Sparseness-based 2ch BSS using the EM algorithm in reverberant environment,” *Proc. WASPAA,* pp. 147–150, 2007.

[6] H. Sawada, S. Araki, and S. Makino, “Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment,” *IEEE Trans. Audio Speech Language Process.,* vol. 19, no. 3, pp. 516–527, 2011.

[7] J. Cermak, S. Araki, H. Sawada, S. Makino, “Blind speech separation by combining beamformers and a time frequency binary mask,” *Proc. IWAENC,* pp. 145–148, 2006.

[8] N. Q. K. Duong, E. Vincent, and R. Gribonval, “Under-determined reverberant audio source separation using a full-rank spatial covariance model,” *IEEE Trans. Audio Speech Language Process.,* vol. 18, no. 7, pp. 1830–1840, 2010.

[9] M. Togami, “Statistical estimation theory considering time-varying nature of systems and source-probability distributions,” *Ph. D. thesis,* the University of Tokyo, 2011.

[10] http://sisec.wiki.irisa.fr/tiki-index.php

[11] E. Vincent, H. Sawada, P. Boll, S. Makino, and J. P. Rosca, “First stereo audio source separation evaluation campaign: data, algorithms and results,” *Proc. ICA,* pp. 552–559, 2007.

[12] K. Iso, S. Araki, and S. Makino, T. Nakatani, H. Sawada, T. Yamada, and A. Nakamura, “Blind source separation of mixed speech in a high reverberation environment,” *Proc. HSCMA,* pp. 36–39, 2011.