一般社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
EA2012-114 (2012-12)

対数スペクトル事前分布を用いた MAPスペクトル推定に基づく劣決定音源分離

岩田 康明† 中谷 智広†† 藤本 雅清†† 吉岡 拓也†† 齋藤 洋典†

† 名古屋大学 大学院 情報科学研究科
〒 464-8601 名古屋市千種区不老町
†† 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
〒 619-0237 京都府相楽郡精華町光台 2-4
E-mail: †{iwata,saito}@cog.human.nagoya-u.ac.jp,
††{nakatani.tomohiro,fujimoto.masakiyo,yoshioka.takuya}@lab.ntt.co.jp

あらまし 音声信号を非定常ガウス過程でモデル化し，最尤法に基づきスペクトル推定を行うアプローチは，多くの音声強調手法で用いられている．我々はこれまでに，このアプローチの推定精度を改善するために，学習済みの対数スペクトル事前分布を導入し，事後確率最大化 (MAP) スペクトル推定に拡張する方法を提案してきた．本稿ではこの拡張を，Duong らが提案したマルチチャネルウィーナフィルタに基づく劣決定音源分離法 [1] に適用する．従来法では最尤法に基づき音源スペクトルと空間相関行列を推定することで音源分離を実現していたのに対し，提案法ではその音源スペクトルの推定の部分を MAP スペクトル推定に拡張する．これにより，高精度なスペクトル推定を行うことで，分離精度の改善を目指す．実験では，提案法により分離精度が改善することを示す．

キーワード ブラインド音源分離，対数パワースペクトル，事後確率最大化推定，混合ガウス分布

Under-Determined Audio Source Separation Based on MAP Spectral Estimation Using Log-Spectral Prior

Yasuaki IWATA†, Tomohiro NAKATANI††, Masakiyo FUJIMOTO††, Takuya YOSHIOKA††, and Hirofumi SAITO†

† Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8601 Japan
†† NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0237 Japan
E-mail: †{iwata,saito}@cog.human.nagoya-u.ac.jp,
††{nakatani.tomohiro,fujimoto.masakiyo,yoshioka.takuya}@lab.ntt.co.jp

Abstract Assuming speech to be non-stationary Gaussian process, maximum likelihood spectral estimation has been studied as an effective speech enhancement approach. Recently, to improve the estimation accuracy of this approach, we have proposed an extention of it, namely a maximum a posterior (MAP) estimation approach using pre-trained log-spectral priors, and showed its effectiveness. This paper newly applies this extention to a multi-channel Wiener filtering based undetermined blind source separation (BSS) technique proposed by Duong et al [1]. This conventional method adopts the likelihood maximization approach for estimating the source spectra and the spatial correlation matrices for the Wiener filtering. The proposed method extends it by introducing the MAP estimation approach for estimating the source spectra, and improves the accuracy of the Wiener filtering.

Key words blind source separation, log-power spectrum, maximum a posteriori estimation, Gaussian mixture model

1. はじめに

近年，未知の環境で収録された音響信号（雑音や残響などを含む）の中から，目的の音声のみを分離・抽出するブラインド信号処理の研究が盛んに行われている．その中でも特に，音声信号を非定常ガウス過程でモデル化し，最尤 (Maximum Likelihood, ML) 法によりそのパワースペクトルを推定するアプローチは，幅広い課題（音源分離／雑音除去／残響除去など）に対し有効な解決法として注目されている [1]〜[3].

我々はこれまでに，このアプローチに対し，その推定精度を改善するための汎用性の高い拡張方法を提案してきた．この方法では，音声信号の事前知識として，その対数パワースペクトルの分布を混合ガウス分布 (Gaussian Mixture Model, GMM) を用いて事前に学習しておき，収録音中の音源スペクトルを事後確率最大化 (Maximum A Posteriori, MAP) 推定により推定する．ここで重要なポイントは，スペクトルの事前分布として尤度関数と共役な分布ではなく，対数パワースペクトルの GMM を用いることが挙げられる．共役事前分布である逆ガンマ分布（あるいは逆ウィシャート分布など）を用いると，事後分布の関数の形が事前分布と同じになるため，その取り扱いが容易になるが [4]，それらの分布では音声の複雑なスペクトルパターンを高精度にモデル化することは困難である．そこで提案法では事前分布として，音声のスペクトルパターンを表現するのに適したモデルとされる，対数パワースペクトルの GMM を用いる．GMM は音声認識における音響モデルにも広く利用されている．この GMM により学習された音声のスペクトルパターンを考慮することで，より高精度なスペクトル推定が可能になると期待される．

なお，提案法では，MAP 推定で最大化すべき関数が非線形関数となるため，解析的な解を得ることが困難である．我々はこの問題に対し，ニュートン法に基づく計算量の少ない解決法を与えるとともに，マルチチャネル線形予測に基づく残響除去と，シングルチャネルウィーナフィルタに基づく雑音抑圧に適用し，その有効性を確認した [5], [6].

本稿では提案法の更なる有効性と汎用性を示すために，提案法をブラインド音源分離に応用する．具体的には，Duong らが提案したマルチチャネルウィーナフィルタに基づく劣決定音源分離法 (ML 法と呼ぶ) [1] に対し提案法を適用する．この分離法では従来，複数のマイクロホンから得られた観測ベクトルがゼロ平均の複素ガウス分布に従うと仮定し，その共分散行列を各音源のパワースペクトルと空間相関行列に分解し，それらを最尤法に基づき推定していた．この音源のパワースペクトルの推定の部分を，対数パワースペクトル GMM を用いた MAP 推定に拡張することで，スペクトル推定の高精度化すなわち分離精度の改善を目指す．

一方，この音源分離法は EM アルゴリズムによる繰り返し推定を用いるため，初期値依存性が高いという問題点があった．そこで礒らは，澤田らが提案した時間周波数マスクを用いたクラスタリング法 (Mask 法と呼ぶ) [7] で観測ベクトルをクラスタリングし，その結果を用いて空間相関行列の初期化を行う方法 (MaskML 法と呼ぶ) [8] を提案し，その有効性を示した．そこで本稿で提案する手法 (MAP 法と呼ぶ) でも同様に，空間相関行列の初期化には時間周波数マスクを用いたクラスタリングの結果を利用する．また，さらに高精度な初期値を用いるために，MaskML 法で得られた推定値を初期値として MAP 法を行う方法 (MLMAP 法と呼ぶ) についても検討する．

実験では，SiSEC [10]，Aurora-2 [12] の音声データを用い，MaskML 法（従来法），MAP 法，MLMAP 法の 3 手法の分離性能を比較した．実験結果より，提案法である MAP 法と MLMAP 法は，従来法の分離性能を改善することができることを示す．

2. 従来法

本節では提案する音源分離法の基礎となる ML 法と，実際の提案法適用対象である MaskML 法について述べる．

2.1 ML 法

ここでは Duong らが提案した ML 法について述べる．まず，I 本のマイクロホンで観測された観測ベクトル $\mathbf{x}(t)$ は以下のように表される．

$$\mathbf{x}(t) = \sum_{j=1}^{J} \mathbf{c}_j(t) \tag{1}$$

ここで J は音源数を表す．また，$\mathbf{c}_j(t) = [c_{1j}(t), \ldots, c_{I,j}(t)]^T$ は j 番目の音源の音像であり，以下のように表される．

$$\mathbf{c}_j(t) = \sum_{\tau} \mathbf{h}_j(\tau) s_j(t-\tau) \tag{2}$$

ここで $s_j(t)$ は j 番目の音源信号であり，$\mathbf{h}_j(\tau) = [h_{1j}(\tau), \ldots, h_{I,j}(\tau)]^T$ は j 番目の音源から各マイクロホンへの伝達特性を表すインパルス応答ベクトルである．この手法の目的は観測ベクトル $\mathbf{x}(t)$ のみから各音像 $\mathbf{c}_1(t), \ldots, \mathbf{c}_J(t)$ を推定することである．時間周波数領域では (2) は以下のように近似される．

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

ここで f，n はそれぞれ周波数ビン，時間フレームを表し，$\mathbf{c}_j(n,f)$，$\mathbf{h}_j(f)$，$s_j(n,f)$ はそれぞれ $\mathbf{c}_j(t)$，$\mathbf{h}_j(\tau)$，$s_j(t)$ の短時間フーリエ変換 (STFT) である．

$\mathbf{c}_j(n,f)$ がゼロ平均ガウス過程に従うと仮定し，その共分散行列 $\mathbf{R}_{\mathbf{c}_j}(n,f)$ は以下のように分解できると仮定する．

$$\mathbf{R}_{\mathbf{c}_j}(n,f) = v_j(n,f)\mathbf{R}_j(f) \tag{4}$$

ここで $v_j(n,f)$ は音源のパワースペクトルを表す時変のスカラー値である．また，$\mathbf{R}_j(f)$ は $I \times I$ の時不変の空間相関行列を表し，フルランク行列であると仮定する．各音源が無相関であると仮定すると，観測ベクトルもゼロ平均ガウス過程に従い，その共分散行列は以下のように表される．

$$\mathbf{R}_{\mathbf{x}}(n,f) = \sum_{j=1}^{J} v_j(n,f)\mathbf{R}_j(f) \tag{5}$$

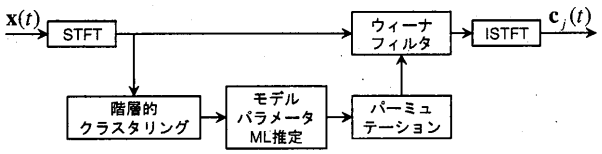

図 1　ML 法の処理の流れ
Fig. 1　Processing flow of ML method.

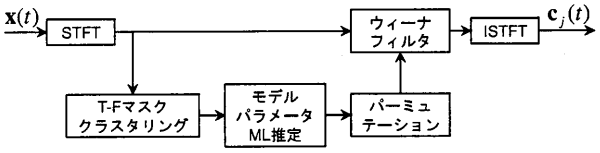

図 2　MaskML 法の処理の流れ
Fig. 2　Processing flow of MaskML method.

そして，音源のパワースペクトルの集合 $v=\{v_j(n,f)\}_{j,n,f}$ と空間相関行列の集合 $\mathbf{R}=\{\mathbf{R}_j(f)\}_{j,f}$ が与えられたときの観測ベクトルの集合 $\mathbf{x}=\{\mathbf{x}(n,f)\}_{n,f}$ の尤度関数は以下のように表される．

$$p(\mathbf{x}|v,\mathbf{R})=\prod_{n,f}\mathcal{N}_c(\mathbf{x}(n,f);\mathbf{0},\mathbf{R}_{\mathbf{x}}(n,f)) \tag{6}$$

ここで $\mathcal{N}_c(\mathbf{x}(n,f);\mathbf{0},\mathbf{R}_{\mathbf{x}}(n,f))$ は平均ベクトル $\mathbf{0}$，共分散行列 $\mathbf{R}_{\mathbf{x}}(n,f)$ の複素ガウス分布を表すものとする．

ML 法では，2 つのステップで音源分離を行う．第 1 ステップでは (6) が最大となるようなモデルパラメータ $\hat{v}$，$\hat{\mathbf{R}}$ を求め，第 2 ステップでは最小平均二乗誤差推定に基づき以下のようなマルチチャネルウィーナフィルタにより各音源の音像を求める．

$$\hat{\mathbf{c}}_j(n,f)=\hat{\mathbf{R}}_{\mathbf{c}_j}(n,f)\hat{\mathbf{R}}_{\mathbf{x}}^{-1}(n,f)\mathbf{x}(n,f) \tag{7}$$

ここで $\hat{}$ はそれが推定値であることを意味する．ML 法の処理の流れを図 1 に示す．

次にモデルパラメータ v，$\mathbf{R}$ の初期化について述べる．まず階層的クラスタリング [1] に基づき観測ベクトル $\mathbf{x}$ をクラスタ $C_1,\ldots,C_J$ に分類する．そして空間相関行列 $\mathbf{R}_j(f)$ を以下のように初期化する．

$$\mathbf{R}_j^{\mathrm{init}}(f)=\frac{1}{|C_j|}\sum_{\mathbf{x}(n,f)\in C_j}\mathbf{x}(n,f)\mathbf{x}(n,f)^H \tag{8}$$

ここで $|C_j|$ はクラスタ C_j の要素数を表し，周波数ビン f に依存する．また ^{H} は行列の共役転置を表す．一方，音源のパワースペクトルは以下のように初期化される．

$$v_j^{\mathrm{init}}(n,f)=1 \tag{9}$$

このようにして得られた初期値 $v_j^{\mathrm{init}}(n,f)$，$\mathbf{R}_j^{\mathrm{init}}(f)$ を用い，EM アルゴリズムによりモデルパラメータ v，$\mathbf{R}$ を推定する．$\theta=\{v,\mathbf{R}\}$ とし，$\{\mathbf{c}_j(n,f)\}_{n,f}$ を隠れ変数とすると，Q 関数は以下のように表される．

$$Q(\theta|\hat{\theta})=\int_{-\infty}^{\infty}p(\mathbf{c}_j|\mathbf{x},\hat{v},\hat{\mathbf{R}})\log p(\mathbf{x},\mathbf{c}_j|v,\mathbf{R})d\mathbf{c}_j \tag{10}$$

$$=\sum_{n,f}\left\{-\frac{\mathrm{tr}\left(\mathbf{R}_j^{-1}(f)\mathcal{R}_{\mathbf{c}_j}(n,f)\right)}{v_j(n,f)}\right.$$

$$\left.-I\log v_j(n,f)-\log|\mathbf{R}_j(f)|\right\}+\mathrm{const} \tag{11}$$

ここで $\mathrm{tr}(\cdot)$ は正方行列のトレースを表し，$\hat{\theta}=\{\hat{v},\hat{\mathbf{R}}\}$ は $\theta=\{v,\mathbf{R}\}$ の現在の推定値である．また $\mathcal{R}_{\mathbf{c}_j}(n,f)$ は以下のように計算される．

$$\mathcal{R}_{\mathbf{c}_j}(n,f)=\hat{\mathbf{c}}_j(n,f)\hat{\mathbf{c}}_j^H(n,f)$$

$$+(\mathbf{I}-\mathbf{W}_j(n,f))\hat{\mathbf{R}}_{\mathbf{c}_j}(n,f) \tag{12}$$

$$\mathbf{W}_j(n,f)=\hat{\mathbf{R}}_{\mathbf{c}_j}(n,f)\hat{\mathbf{R}}_{\mathbf{x}}^{-1}(n,f) \tag{13}$$

$$\hat{\mathbf{c}}_j(n,f)=\mathbf{W}_j(n,f)\mathbf{x}(n,f) \tag{14}$$

ここで $\hat{\mathbf{R}}_{\mathbf{c}_j}(n,f)$，$\hat{\mathbf{R}}_{\mathbf{x}}(n,f)$ はそれぞれ (4)，(5) で求められる．また $\mathbf{I}$ は $I\times I$ の単位行列である．

E ステップでは (13)〜(12) の計算を行い，M ステップでは Q 関数 (11) が最大になるようにモデルパラメータを以下のように更新する．

$$v_j(n,f)=\frac{1}{I}\mathrm{tr}\left(\hat{\mathbf{R}}_j^{-1}(f)\mathcal{R}_{\mathbf{c}_j}(n,f)\right) \tag{15}$$

$$\mathbf{R}_j(f)=\frac{1}{N}\sum_{n=1}^{N}\frac{1}{\hat{v}_j(n,f)}\mathcal{R}_{\mathbf{c}_j}(n,f) \tag{16}$$

以下 ML 法では，1) E ステップと M ステップを収束するまで繰り返し，2) 音源方向推定に基づきパーミュテーション問題を解き，3) マルチチャネルウィーナフィルタ (7) により各音源の音像を求め，4) 短時間逆フーリエ変換 (ISTFT) により時間領域に戻すことで，分離を実現する．

2.2　MaskML 法

ここでは礒らが提案した MaskML 法について述べる．この手法では，ML 法と比べ，モデルパラメータの初期化における観測ベクトルのクラスタリングの方法が異なる．具体的には，澤田らが提案した Mask 法 [7] により得られる事後確率マスクを用いる．そのマスク $M_j(n,f)$ の値は，観測ベクトル $\mathbf{x}(n,f)$ が与えられた時にそれがクラスタ C_j に属する確率 $p(C_j|\mathbf{x}(n,f))$ の推定値である．そこで，各周波数ビンにおいて観測ベクトルを以下のようにクラスタリングする．

$$\mathbf{x}(n,f)\in C_{j'}\qquad j'=\arg\max_j M_j(n,f) \tag{17}$$

その後，得られたクラスタを用い，(8) で $\mathbf{R}_j(f)$ を初期化する．また，パーミュテーション問題は音源間のパワー比の相関に基づく方法 [9] を用いて解く．他の部分は ML 法と同様である．MaskML 法の処理の流れを図 2 に示す．

3.　提　案　法

本節では我々の提案する MAP 法と MLMAP 法について述べる．

3.1　MAP　法

まず音源の対数パワースペクトル $\rho_j(n,f)$ を以下のように定義する．

$$\rho_j(n,f) = \log v_j(n,f) \tag{18}$$

そして $\boldsymbol{\rho}_j(n) = [\rho_j(n,1),\ldots,\rho_j(n,F)]^T$ が GMM に従うと仮定する．その集合 $\boldsymbol{\rho} = \{\boldsymbol{\rho}_j(n)\}_{j,n}$ の確率密度関数は以下のように表される．

$$p(\boldsymbol{\rho}) = \prod_{j,n}\sum_{k_j=1}^{K} w_{k_j}\mathcal{N}(\boldsymbol{\rho}_j(n);\boldsymbol{\mu}_{k_j},\boldsymbol{\Sigma}_{k_j}) \tag{19}$$

ここで $\mathcal{N}(\cdot)$ はガウス分布を表し，$\{w_1,\ldots,w_K\}$，$\{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K\}$，$\{\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_K\}$ はそれぞれ混合重み，平均ベクトル，共分散行列を表す．K は混合数である．MAP 法ではこの GMM を，学習用クリーン音声を用いて事前に学習しておき，スペクトル推定に利用する．

(6)，(19) より，観測ベクトルの集合 $\mathbf{x}$ が与えられたときの，音源の対数パワースペクトルの集合 $\boldsymbol{\rho}$，空間相関行列の集合 $\mathbf{R}$ の事後確率密度関数は以下のように表される．

$$\begin{aligned} p(\boldsymbol{\rho},\mathbf{R}|\mathbf{x}) &\propto p(\mathbf{x}|\boldsymbol{\rho},\mathbf{R})p(\boldsymbol{\rho}) \qquad (20)\\ &= \prod_{n,f}\mathcal{N}_c(\mathbf{x}(n,f);\mathbf{0},\mathbf{R}_{\mathbf{x}}(n,f)) \\ &\quad\times\prod_{j,n}\sum_{k_j=1}^{K} w_{k_j}\mathcal{N}(\boldsymbol{\rho}_j(n);\boldsymbol{\mu}_{k_j},\Sigma_{k_j}) \qquad (21)\end{aligned}$$

ただし $p(\mathbf{R})$ は一様分布であると仮定し略記した．

MAP 法では (21) を最大にするモデルパラメータ $\boldsymbol{\rho}$，$\mathbf{R}$ を EM アルゴリズムにより求める．$\theta = \{\boldsymbol{\rho},\mathbf{R}\}$ とし，$\{\mathbf{c}_j(n,f)\}_{n,f}$ と k_j を隠れ変数とすると，Q 関数は以下のように表される．

$$\begin{aligned} Q(\theta|\hat{\theta}) &= \sum_{k_j=1}^{K}\int p(\mathbf{c}_j|\mathbf{x},\hat{\boldsymbol{\rho}},\hat{\mathbf{R}})p(k_j|\{\hat{\boldsymbol{\rho}}_j(n)\}_n) \\ &\quad\times\log p(\mathbf{x},\boldsymbol{\rho},\mathbf{R},\mathbf{c}_j,k_j)d\mathbf{c}_j \qquad (22)\\ &= \sum_n\Bigg[\sum_f\Bigg\{-\frac{\mathrm{tr}\left(\mathbf{R}_j^{-1}(f)\mathcal{R}_{\mathbf{c}_j}(n,f)\right)}{e^{\rho_j(n,f)}} \\ &\quad -I\rho_j(n,f) - \log|\mathbf{R}_j(f)|\Bigg\} \\ &\quad -\frac{1}{2}\sum_{k_j=1}^{K} p(k_j|\hat{\boldsymbol{\rho}}_j(n))(\boldsymbol{\rho}_j(n)-\boldsymbol{\mu}_{k_j})^T\boldsymbol{\Sigma}_{k_j}^{-1} \\ &\quad (\boldsymbol{\rho}_j(n)-\boldsymbol{\mu}_{k_j})\Bigg] + \mathrm{const} \qquad (23)\end{aligned}$$

ここで $\hat{\theta} = \{\hat{\boldsymbol{\rho}},\hat{\mathbf{R}}\}$ は $\theta = \{\boldsymbol{\rho},\mathbf{R}\}$ の現在の推定値であり，

$$p(k_j|\hat{\boldsymbol{\rho}}_j(n)) = \frac{w_{k_j}\mathcal{N}(\hat{\boldsymbol{\rho}}_j(n);\boldsymbol{\mu}_{k_j},\boldsymbol{\Sigma}_{k_j})}{\sum_{k_j=1}^{K} w_{k_j}\mathcal{N}(\hat{\boldsymbol{\rho}}_j(n);\boldsymbol{\mu}_{k_j},\boldsymbol{\Sigma}_{k_j})} \tag{24}$$

また，(24) の計算には全周波数ビンにおける対数パワースペクトルを用いるため，MAP 法ではモデルパラメータ推定の前に MaskML 法と同様の方法でパーミュテーション問題を解く必要がある．MAP 法の処理の流れを図 3 に示す．

(23) の最後の 2 行は $\mathbf{R}$ に依存しないため，$\mathbf{R}$ は従来と同様，(16) で更新される．一方 $\boldsymbol{\rho}$ の更新では，(23) が ρ に関する非線形関数になるため，非線形最適化問題を解かなければならない．これには我々が [5] で提案したニュートン法に基づく計算効率の高い方法を用いることができる．

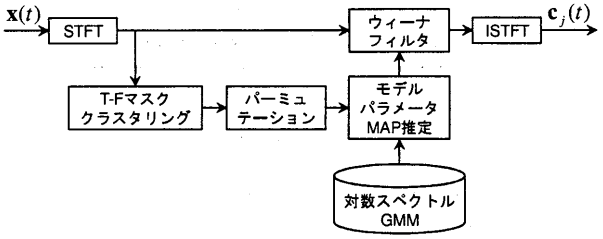

図 3　MAP 法の処理の流れ
Fig. 3　Processing flow of MAP method.

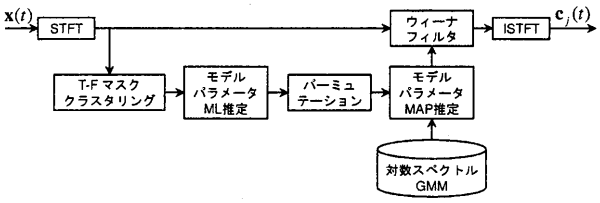

図 4　MLMAP 法の処理の流れ
Fig. 4　Processing flow of MLMAP method.

手順をまとめると，E ステップでは (13)〜(12)，(24) を計算する．M ステップではまずニュートン法に基づき $\rho_j(n,f)$ を更新し，次に (16) で $\mathbf{R}_j(f)$ を更新する．

3.2　モデルパラメータの初期化

ここではモデルパラメータ v，$\mathbf{R}$ の初期化について述べる．まず注目すべき点は，ML 法において尤度関数 (6) を最大にする v，$\mathbf{R}$ は一意には決まらないということである．なぜなら，例えば $\{v(n,f)\}_n$ が定数 a 倍（すなわち $\{\rho(n,f)\}_n$ が $\log a$ だけ加算）されても，$\mathbf{R}(f)$ が $1/a$ 倍されれば尤度関数の値は変化しないためである．そのため従来法における v，$\mathbf{R}$ の初期化では，それらのスケールを考慮する必要はなかった．しかし，MAP 法における事後確率密度関数 (21) は ρ の事前分布を含むため，ρ，$\mathbf{R}$ は一意に決まる．そこで ρ，$\mathbf{R}$ の初期化は，それらのスケールを考慮し，以下のように行う．

$$\rho_j^{\mathrm{init}}(n,f) = \mu^{\mathrm{GMM}}(f) \tag{25}$$

$$\mathbf{R}_j^{\mathrm{init}}(f) = \frac{e^{-\mu^{\mathrm{GMM}}(f)}}{|C_j|}\sum_{\mathbf{x}(n,f)\in C_j}\mathbf{x}(n,f)\mathbf{x}(n,f)^H \tag{26}$$

ここで

$$\mu^{\mathrm{GMM}}(f) = \sum_{k_j=1}^{K} w_{k_j}\mu_{k_j}(f) \tag{27}$$

であり，$\mu_{k_j}(f)$ は $\boldsymbol{\mu}_{k_j}$ の f 番目の要素である．また，クラスタ C_j は MaskML と同様，Mask 法 [7] を用いて求める．

3.3　MLMAP 法

ここでは，より高精度な初期値を用いて MAP 推定することを目的とした MLMAP 法について述べる．MAP 法が最初から MAP 推定によりモデルパラメータを推定するのに対し，MLMAP 法は図 4 に示すように，まず ML 推定した後，さらに MAP 推定を行う．ML 推定により得られた $\rho_j(n,f)$，$\mathbf{R}_j(f)$

表 1　実験 1 の条件
Table 1　Condition of experiment 1

サンプリング周波数	8 kHz
マイクロホン数	2
マイクロホン間距離	4 cm
2 混合音声（テスト用）	30 発話
3 混合音声（テスト用）	40 発話
話者（テスト用）	男性 10 人，女性 10 人
発話時間	5 s
残響時間	110, 220, 400 ms
窓関数	ハニング窓
フレーム幅	128 ms（1024 点）
シフト幅	32 ms（256 点）
EM 繰り返し数	100
音声（学習用）	8440 発話
話者（学習用）	男性 55 人，女性 55 人
GMM 混合数	256

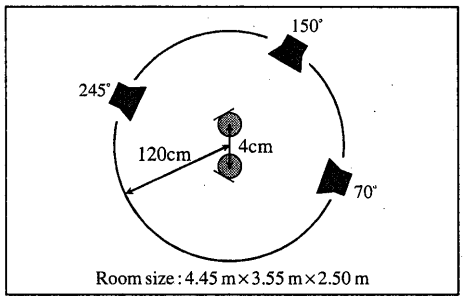

図 5　インパルス応答の収録環境
Fig. 5　Impulse response recording environment

の推定値のパーミュテーション問題を解いたものを $\rho_j^{\mathrm{ML}}(n,f)$, $\mathbf{R}_j^{\mathrm{ML}}(f)$ とすると，MAP 推定におけるモデルパラメータの初期化は以下のように行う．

$$\rho_j^{\mathrm{init}}(n,f) = \rho_j^{\mathrm{ML}}(n,f) - \mu^{\mathrm{ML}}(f) + \mu^{\mathrm{GMM}}(f) \tag{28}$$

$$\mathbf{R}_j^{\mathrm{init}}(f) = \frac{e^{\mu^{\mathrm{ML}}(f)-\mu^{\mathrm{GMM}}(f)}}{|C_j|} \sum_{\mathbf{x}(n,f)\in C_j} \mathbf{x}(n,f)\mathbf{x}(n,f)^H \tag{29}$$

ここで $\mu^{\mathrm{ML}}(f)$ は $\rho_j^{\mathrm{ML}}(n,f)$ の全時間フレームの平均である．この後の処理は MAP 法と同様である．

4. 実　　験

本節では提案法の効果を検証するために行った 2 種類の実験について述べる．

4.1　実験 1：Aurora-2 を用いた評価

Aurora-2 の音声データを用いた評価実験を行った．実験条件を表 1 に示す．なお，表中にある EM 繰り返し数 100 について，MLMAP 法では $(\mathrm{ML},\mathrm{MAP}) = (10,90),(20,80),\ldots,(50,50)$ の 5 通りに分けて処理を行った．また，混合音声の合成に用いたインパルス応答の収録環境を図 5 に示す．2 音源の場合は 70°，150° の配置，3 音源の場合は 70°，150°，245° の配置を用いた．分離音声の評価には [11] で定義される以下の 4 つの評価尺度を用いた．

- SDR (Signal to Distortion Ratio)
- ISR (source Image to Spatial distortion Ratio)

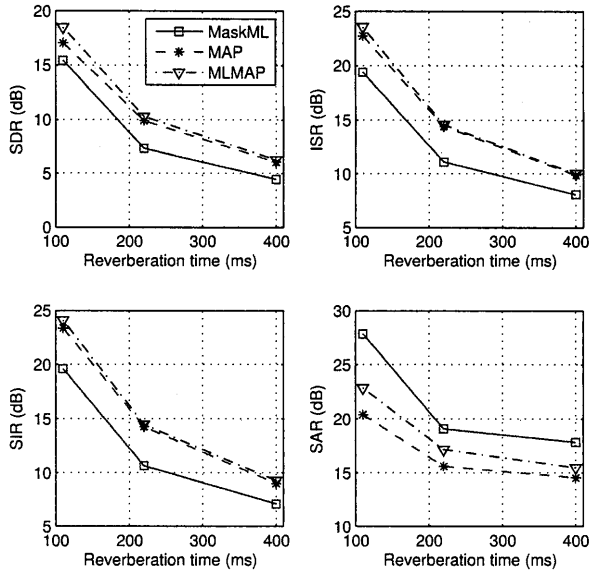

図 6　SDR, ISR, SIR, SAR（2 音源）
Fig. 6　SDR, ISR, SIR, SAR (2 sources)

- SIR (Source to Interference Ratio)
- SAR (Sources to Artifacts Ratio)

これらはすべて値が大きいほど性能が高いことを意味する．

2 音源の場合の評価結果を図 6 に示す．これらの図は，残響時間ごとに全音声データの評価値の平均をさらに全話者で平均したものをプロットしたものである．また MLMAP 法の EM 繰り返しの分け方は，$(\mathrm{ML},\mathrm{MAP}) = (10,90)$ が最も良い性能であったため，その結果を示している．結果を比較すると，SAR を除くすべての評価尺度において，2 つの提案法は従来法を大きく上回っている．これは，GMM が従来法の分離性能を改善させたことを示している．

続いて，3 音源の場合の評価結果を図 7 に示す．MLMAP 法の EM 繰り返しの分け方は，$(\mathrm{ML},\mathrm{MAP}) = (20,80)$ が最も良い性能であったため，その結果を示している．図より，2 音源の場合と同様，MLMAP 法は従来法を上回る性能を実現できている一方で，MAP 法は従来法と同程度の性能しか達成できていなかったことが確認できる．MLMAP 法と MAP 法は，初期値が異なる以外は同一の処理であるため，この結果は提案法の初期値依存性の高さを示していると思われる．

4.2　実験 2：SiSEC を用いた評価

音源分離の標準評価タスクである SiSEC の音声データを用いて評価を行った．実験条件を表 2 に示す．ただし MLMAP 法は実験 1 と同様の 5 通りに分けて処理を行った．分離音声の評価には SDR，ISR，SIR，SAR を用いた．

評価結果を表 3 に示す．MLMAP 法の EM 繰り返しの分け方は，男性の場合は (ML, MAP) $= (20,80)$，女声の場合は $(\mathrm{ML},\mathrm{MAP}) = (40,60)$ が最も良い性能であったため，それらの結果を示している．提案法と従来法を比較すると，MAP 法はほとんどの評価値で従来法を下回っているのに対し，MLMAP 法はほとんどの評価値で上回っている．この結果は，Aurora-2

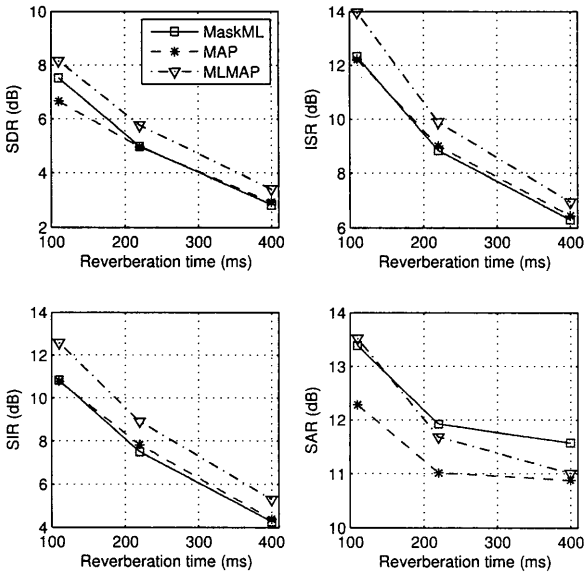

図 7 SDR, ISR, SIR, SAR (3 音源)
Fig. 7 SDR, ISR, SIR, SAR (3 sources)

表 2 実験 2 の条件
Table 2 Condition of experiment 2

サンプリング周波数	16 kHz
マイクロホン数	2
マイクロホン間距離	5 cm
3 混合音声 (テスト用)	男声 1 発話, 女声 1 発話
発話時間	10 s
残響時間	250 ms
窓関数	ハニング窓
フレーム幅	128 ms (2048 点)
シフト幅	32 ms (512 点)
EM 繰り返し数	100
音声 (学習用)	1000 発話
話者 (学習用)	男性 10 人, 女性 10 人
GMM 混合数	256

表 3 評価結果
Table 3 Evaluation result

		男声			女声		
		話者			話者		
		1	2	3	1	2	3
MaskML (従来法)	SDR (dB)	2.0	1.1	**5.6**	5.3	**4.6**	5.9
	ISR (dB)	5.6	3.5	**11.0**	6.9	10.0	11.4
	SIR (dB)	2.5	0.0	7.3	10.8	5.7	7.7
	SAR (dB)	8.3	**6.3**	**10.4**	10.6	**9.3**	**11.2**
MAP	SDR (dB)	2.8	1.4	4.5	4.8	3.2	4.5
	ISR (dB)	8.0	3.2	10.0	6.6	8.7	8.8
	SIR (dB)	3.8	2.6	6.5	9.9	4.4	6.6
	SAR (dB)	9.2	5.1	9.6	10.2	8.9	10.1
MLMAP	SDR (dB)	**3.4**	**2.0**	4.8	**5.5**	**4.6**	**6.1**
	ISR (dB)	**9.7**	**4.0**	8.7	**7.3**	**10.3**	**11.8**
	SIR (dB)	**4.2**	**3.5**	**8.0**	**11.3**	**6.3**	**8.3**
	SAR (dB)	**10.0**	5.3	8.8	**10.7**	9.2	11.0

の音声データを用いた 3 音源分離の結果と，ほぼ一致する．

5. まとめ

本稿では対数スペクトル GMM を用いた MAP スペクトル推定に基づく音源分離法を提案し，実験によりその有効性を示した．しかし，提案法は従来法と比べ初期値依存性が高く，分離が難しい条件 (2 マイク 3 音源など) では必ずしも性能が改善されるとは限らないということも示された．ただし，この問題は，従来法を用いて初期化を行った後に提案法を用いることにより解決できることを確認した．

文献

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

[2] M. Togami, Y. Kawaguchi, R. Takeda, Y. Obuchi, and N. Nukaga, "Multichannel speech dereverberation and separation with optimized combination of linear and non-linear filtering," *Proc. the 2012 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp.4057-4060, Mar. 2012.

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

[4] C.M. ビショップ, 元田浩, 栗田多喜夫, 樋口知之, 松本祐治, 松田昇 (監訳), "パターン認識と機械学習 上," 丸善出版, 2012.

[5] Y. Iwata and T. Nakatani, "Introduction of speech log-spectral priors into dereverberation based on itakura-saito distance minimization," *Proc. the 2012 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp.245-248, Mar. 2012.

[6] 岩田康明, 中谷智広, 藤本雅清, 吉岡拓也, 齋藤洋典, "対数スペクトル事前分布を用いた音声の MAP スペクトル推定と雑音抑圧による評価," 日本音響学会, 平成 24 年度秋季研究発表会, 1-P-30, Sep. 2012.

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

[8] K. Iso, S. Araki, S. Makino, T. Nakatani, H. Sawada, T. Yamada, and A. Nakamura, "Blind source separation of mixed speech in a high reverberation environment," *Proc. the 3rd Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, pp.36-39, May 2011.

[9] H. Sawada, S. Araki, S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS," *Proc. the 2007 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp.3247-3250, May 2007.

[10] E. Vincent, S. Araki, F.J. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, B.V. Gowreesunker, D. Lutter and N.Q.K. Duong, "The Signal Separation Evaluation Campaign (2007-2010): Achievements and remaining challenges," *Signal Processing*, vol.92, pp.1928-1936, Aug. 2012.

[11] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J.P. Rosca, "First Stereo Audio Source Separation Evaluation Campaign: Data, Algorithms and Results," *Proc. Independent Component Analysis (ICA) 2007*, pp.552-559, Sep. 2007.

[12] H.G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy condidions," *In Proc. ICSLP*, vol.4, pp.29-32, 2000.