パワースペクトル上の無限オーダー畳込み型室内伝達モデルに基づく実時間セミブラインド音声強調*

○前澤　陽 (ヤマハ株式会社) 奥乃　博（京都大学）

1　はじめに

セミブラインド音声強調技術– ある機器から発せられる既知の音源（以下「自己発話」）と，分析の対象とする音源（「目的音声」）の混合音から，自己発話を抑圧すること–は，多くの信号処理技術の実用化に際して重要な技術である．たとえば，セミブラインド音声強調技術を音声認識に活用すると，ロボットへの組み込んだ際に，ロボットの動作音に対する頑健性を実現したり [1]，音声合成エンジンの発話途中に,「割り込んで」音声認識ができるようになる [2]．また，音楽情報処理（MIR）においても，ある機器から発せられる伴奏音に対して，ユーザが奏でる音響信号が上乗せされる場合，ユーザの演奏音を分析するためには，セミブラインド音声強調は有益だと考えられる．このようなユースは，カラオケをはじめとする自動伴奏技術には必要不可欠であろう．本研究は，音楽情報処理の実時間応用に際して，実用的なセミブラインド音声強調の実現を目的とする．

実時間 MIR タスクにおける，セミブラインド音声強調技術の実用化に際しては，三点の要件を満たすことが望ましい．一点目は，単一のマイクロフォンで動作することである．というのも，近年タブレット端末を始めとする携帯端末が普及し,「かさばらない」ことが，より重要になると想定されるためである．この場合，一つの観測音から，目的音と自己発話の二つを推定することは，劣決定問題となる．そのため，劣決定でないことを要求する，独立成分分析に基づく手法 [1, 3, 4] は使えない．そのため，モノラル音源の振幅スペクトログラムを，音源と観測スペクトルの非負値重みづけ和として表わすといったアプローチ [5] が有効になる．

二点目は，幅広い残響時間下で動作することである．なぜならば，音楽音響信号は，ドライなスタジオからコンサートホールといった，幅広い残響環境で再生されるためである．特に，残響時間に関連する次数を予め設定しないことが望ましい．次数を事前に決定することの障害は，音源分離などの枠組みで特に指摘されている [6]．特に，ノンパラメトリックベイズに基づくモデル化は，実質的な次数を，観測データに基づいて動的に変化させる方法として，信号の分解 [6] や残響抑圧 [7] の枠組みで提案されている．

三点目は，実時間で動作することである．二点目の要求を満たすためには，ベイズ推定をできることが望ましいが，逐次的にベイズ推定を行う方法は自明ではなかった．しかし，近年，ベイズ推定の方法論の一つである変分ベイズ法を，自然勾配法として見なすことにより，一部の問題で，逐次的な推論方法が確立 [8] されている．本稿では，この理論的枠組みをセミブラインド音源分離に活用する．

これら三点の要求を踏まえ，本稿では，単一の入力チャンネルで動作し，室内音響系に対する長さを明示的に与える必要のないことを特徴とした，実時間セミブラインドモノラル音声強調手法を提案する．

2　定式化

本手法は，単一のマイクロフォンで観測された，観測音の振幅スペクトログラムを，室内音響系に畳み込まれた自己発話と，目的音声の混合としてモデル化する．まず，周波数ビン f，時間インデックス t に対して，観測音の振幅スペクトログラムを $y_{f,t}$，ソースの振幅スペクトログラムを $s_{f,t}$，自己発話の振幅スペクトログラムを $x_{f,t}$ とする．また，自己発話の伝達系における，インパルス応答の振幅スペクトログラムを $h_{f,i}$ $(0 \leq i \leq I)$ とする．また，目的音のゲインを g_0，$h_{f,i}$ の各 i で共有されるゲインを g_{i+1} とする．本来このゲイン変数は s や h に吸収できるので冗長にも見えるが，後述する確率モデルで重要な役割を果たす．$h_{f,i}$ には，マイク，スピーカー，部屋の初期反射といった，スペクトルを f 軸に色付けするような要素が，部屋の長期残響によって i 方向に減衰するような形をしている．

ここで，s, x, h の独立性と，振幅の加法性を仮定し，更に，自己発話 x が，$h_{f,i}$ に時間方向に畳み込まれたものが，y で観測されたと仮定する．すると，次のような観測モデルを考えることができる：

$$y_{f,t} = g_0 s_{f,t} + \sum_{i=0}^{I} g_{i+1} h_{f,i} x_{f,t-i} \qquad (1)$$

このような観測モデルを元に，適当な統計モデルを構築し，統計モデルの事後分布を求めることを考える．事後分布が求まれば，事後分布の統計を取ることにより，あらゆる情報を得ることが出来る．たとえば，目的音を復元する際には，s の期待値を取ればよい．確率モデルの設計にあたっては，序文で説明した要件を満たすことが望ましい．特に，残響時間に対するロバストネスや，音楽音響信号を対象といった事柄は，適切な統計モデルの設計により，実現される．

---

*Semiblind Source Emphasis based on an infinite convolutive model of room acoustics.
by Akira MAEZAWA (Yamaha Corporation) and Hiroshi G. OKUNO (Kyoto University)

本手法における要求の一つに，残響時間に対するロバストネスを確保することがある．しかし，本手法の観測モデルにおいて，残響時間は，次数 I として陽に表わされ，その最適値は，残響時間によって適切な値が異なる．残響時間に対して，ロバストネスを確保するためには，次数が，観測音 y の特性に適応するような形が望ましい．

このような機能を実現するため，g には，縮退するような性質を持たせたい．つまり，観測音を，雑音成分と，過去の自己発話に h を重みづけた物の線形結合で説明する際に，必要最小限の変数で説明したいのである．そこで，g の事前分布として，ガンマ過程 [6,9] を考える．まず，L 個の確率変数 θ_l を，$\theta_l \sim \mathrm{Gamma}(\alpha/L, \alpha c)$ から生成することを考える．この過程は，L が増えるにつれ，形状パラメータ α と逆スケールパラメータ αc を持つガンマ過程に近づく．このようなガンマ過程から生成された点のうち，無視できない値 $\epsilon > 0$ を持つサンプルの数は，ほぼ確実に有限である．特に，L が α に対して十分大きければ，θ_l のうち，無視できない値を持つものは，L よりもはるかに少ない．よって，g のうち多くは無視できる値をとり，データを説明するのに効いてくる変数に対応する g_i の値のみが無視できない値になることが期待される．

y は，Poisson 分布に従って生成されるとする：

$$y|x,s,h \sim \mathrm{Pois}\left(\sum_i g_{i+1} x_{f,t-i} h_{f,i} + s_{f,t} g_0\right) \quad (2)$$

Poisson 分布は再生性を有しているため，観測モデルで仮定した，振幅の加法性と係数の独立性への相性がよい．再生性とは，$y_1 \sim \mathrm{Pois}(x_1)$ と，$y_2 \sim \mathrm{Pois}(x_2)$ となるような，独立な確率変数 y_1 と y_2 が与えられるとき，$y_1 + y_2 \sim \mathrm{Pois}(x_1 + x_2)$ となるような性質を指す．再生性により，補助変数 $M^{(s)}$ と $M^{(h)}$ を導入し，$M^{(s)}_{f,t} + \sum_i M^{(h)}_{f,t,i} = Y_{f,t}$ と制約することにより，次のような等価なモデルを考えることができる：

$$M^{(s)}_{f,t}|s \sim \mathrm{Pois}(s_{f,t} g_0) \quad (3)$$

$$M^{(h)}_{f,t,i}|x,h \sim \mathrm{Pois}(x_{f,t-i} h_{f,i} g_{i+1}) \quad (4)$$

推論の便宜上，h と s は，Poisson 分布の共役事前分布であるガンマ分布を設定する．ある密度関数 $p(x|\theta)$ に対する共役事前分布 $p(\theta)$ とは，事後分布 $p(\theta|x)$ が，事前分布 $p(\theta)$ と同じ関数形であるような分布のことである．g の事前分布は，残響時間に対するロバストネスを要求するため，ガンマ過程の近似に従うようにする．

$$g_i|\alpha^{(g)}, c^{(g)} \sim \mathrm{Gam}\left(\alpha^{(g)}/I, \alpha^{(g)} c^{(g)}\right) \quad (5)$$

$$h_{f,i}|a^{(h)}, b^{(h)} \sim \mathrm{Gam}\left(a^{(h)}, b^{(h)}\right) \quad (6)$$

$$s_{f,t}|a^{(s)}, b^{(s)} \sim \mathrm{Gam}\left(a^{(s)}, b^{(s)}\right) \quad (7)$$

これらを踏まえた上で，変分ベイズ法に基づく事後分布推定方法を導出する．まずは準備のため，オフライン推論手法について述べる．

変分ベイズ法とは，近似された事後分布 q を求める方法であり，目的関数 $J(\Theta) = \langle \log p(X,\Theta) \rangle_{q(\Theta)} + H_{q(\Theta)}$ を最大化させることに基づく．ここで，$\langle f(x,y) \rangle p(y)$ は，確率密度 $p(y)$ の下で $f(x,y)$ の期待値を評価することを意味し，また，$H_{q(\Theta)} = -\int q(\Theta) \log q(\Theta) d\Theta$ であり，エントロピーを指す．対数同時分布 $\log p(y,x,g,s,h)$ は，定数項を無視すると，次のように与えられる：

$$\begin{aligned}
\log p(y,x,g,s,h) &= \sum_{f,t} \left[a^{(s)}_{f,t} \log s_{f,t} - b^{(s)}_{f,t} h_{f,t} \right] \\
&+ \sum_{f,i} \left[a^{(h)} \log h_{f,i} - b^{(h)} h_{f,i} \right] + \sum_i \left[\alpha^{(g)}/I \log g_i - \alpha^{(g)} c^{(g)} g_i \right] \\
&+ \sum_{f,i,t} \left[M^{(h)}_{f,t,i} \log\left(g_{i+1} x_{f,t-i} h_{f,i}\right) - g_{i+1} X_{f,t-i} h_{f,i} - \Gamma(M^{(h)}_{f,t,i}) \right] \\
&+ \sum_{f,t} \left[M^{(s)}_{f,t} \log\left(g_0 s_{f,t}\right) - g_0 s_{f,t} - \Gamma(M^{(s)}_{f,t}) \right] \quad (8)
\end{aligned}$$

この目的関数は，最適化が困難であるが，q が $\prod_i q(g_i) \prod_{f,i} q(h_{f,i}) \prod_{f,t} q(s_{f,t})$ の形で近似できるとすると，解析的な最適化が可能となる．この様な近似は，言い換えると，事後分布と事前分布が，同じ関数形をしていると仮定している．すると，変分ベイズは，次の目的関数を最大化することに等しい：

$$\begin{aligned}
J(\Theta) = &\sum_i H_{q(g_i)} + \sum_{f,i} H_{q(h_{f,i})} + \sum_{f,t} H_{q(s_{f,t})} \\
&+ \langle \log p(y,x,g,s,h) \rangle_{q(g,s,h)} \quad (9)
\end{aligned}$$

このように，q における変数間の独立性を仮定すると，$J(\Theta)$ は，個々の変数に対して最適化を行うことができる．具体的には，ある変数 $\theta_i \in \Theta$ を最適化する際は，$\langle \log p(X,\Theta) \rangle_{q(\neg\theta_i)} + H_{q(\theta_i)}$ を最大化させればよい．ここで，$\neg\theta_i$ とは，Θ に含まれる θ_i 以外の変数という意味である．対数同時分布 $\log p(X,\Theta)$ の期待値は，期待値を取る分布が，指数族であり，観測尤度の共役事前分布である場合，解析的に求まる．本手法は，この要件を満たすため，期待値は簡単に求まる．なお，紙面の制約上，ここでは，この目的関数の最大化方法については省略する．

目的音を復元する際には，目的音の STFT して，振幅として目的音の期待値 $\langle s_{f,t} \rangle$，位相として観測音 STFT のものを用意し，信号を復元すればよい．

2.1 実時間推定

式 9 の最適化は実時間では行えない．というのも，g や h の事後分布は，全ての観測に依存し，s の事後分布は g と h の期待値に依存するため，逐次的な推論が不可能だからである．しかし，変分ベイズを，自然勾配法と捉えることにより [10]，このような場面で

も，逐次推定が可能な，変分ベイズ推定手法が確立されている [8]．そこで，このアイディアを，本手法の実時間変分ベイズ推論へ応用することを考える．

まず，時刻 0 からデータ長 T の間で一様に分布する確率変数 $\tau \sim \mathrm{Uniform}(0,T)$ を導入し，次のような目的関数 J_τ を定義する：

$$
\begin{aligned}
J_\tau(\Theta) = &\sum_i \left(H_{q(g_i)} + \frac{\alpha^{(g)}}{I}\langle \log g_i \rangle - \alpha^{(g)} c^{(g)} \langle g_i \rangle \right) \\
&+ \sum_{f,i} \left(H_{q(h_{f,i})} + a^{(h)} \langle \log h_{f,i} \rangle - b^{(h)} \langle h_{f,i} \rangle \right) \\
+ T \max_{M^{(h,s)}_{:,\tau}, a^{(s)}_{:,\tau}, b^{(s)}_{:,\tau}} &\Bigg[H_{q(s_\tau)} + M^{(s)}_{f,\tau} \langle \log g_0 s_{f,\tau} \rangle - \langle g_0 s_{f,\tau} \rangle \\
&+ a^{(s)}_{f,\tau} \langle \log s_{f,\tau} \rangle - b^{(s)}_{f,\tau} \langle s_{f,\tau} \rangle \\
&+ \sum_i \left(M^{(h)}_{f,\tau,i} \langle \log(g_{i+1} h_{f,i} x_{f,\tau-i}) \rangle - g_{i+1} x_{\tau-i} \langle h_{f,i} \rangle \right) \Bigg] \\
&\text{subject to } Y_\tau = \sum_i M^{(h)}_{f,\tau,i} + M^{(s)}_{f,\tau} \quad (10)
\end{aligned}
$$

この目的関数は，$J(\Theta)$ に対して，(1) 全ての観測データが，時刻 τ の観測値であると仮定し，(2) $s_{f,t}$，$M^{(s)}$，$M^{(h)}$ を先に最適化する，という変更を加えているものである．ここで重要なのは，$\max \langle J_\tau(\Theta) \rangle_\tau = \max J(\Theta)$ となることである．つまり，$\langle J_\tau \rangle$ を最適化することは，変分ベイズ推定を行っていることになる．

そこで，$\langle J_\tau(\Theta) \rangle_\tau$ を最適化することを考える．τ を一様に分布させるためには，$t=0$ から $t=T$ までの全てのステップを評価すればよい．ここで，パラメータ Θ が，時刻 t のみに依存する $\Psi(t)$ と，時刻に依存しない Λ の二つで構成されていると考える．すると，各ステップにおいて $J_t(\Theta)$ を $\Psi(t)$ に対して最適化し，目的関数を自然勾配法を用いて Λ に対して最適化すれば，$\langle J_\tau(\Theta) \rangle_\tau$ を，オンラインで最適化できる．また，サンプルを増やせば増やすほど，目的関数の最適値は，オフライン推論のそれに近づいていく．

これらを踏まえると，実時間での変分ベイズ推定は次のような逐次的アルゴリズムとして与えられる：

1. 初期化：$\tilde{a}^{(h)}_{f,i}(0) = a^{(h)}$，$\tilde{b}^{(h)}_{f,i}(0) = b^{(h)}$，$\tilde{a}^{(g)}_i(0) = \alpha^{(g)}/I$，$\tilde{b}^{(g)}_i(0) = \alpha^{(g)} c^{(g)}$ とする．
2. $t = 1 \cdots T$ の間，次の処理を行う
   (a) 式 10 の max 演算を，時刻 t で評価する．具体的には，以下 2 つの処理を，Φ が収束するまで行う．まず，

$$
\Phi_{f,i} \propto \begin{cases} \exp\langle \log(g_0 s_{f,t}) \rangle & i = 0 \\ x_{f,t-i} \exp\langle \log(g_{i+1} h_{f,i}) \rangle & i > 0 \end{cases} \quad (11)
$$

とする．ただし，$\sum_i \Phi_i = y_{f,t}$ になるよう正規化する．次に，s_t の事後分布を次のように更新する：

$$
q(s_{f,t}) = \mathrm{Gam}\left(a^{(s)} + \Phi_{f,0}, b^{(s)} + \langle g_0 \rangle \right) \quad (12)
$$

   (b) 式 10 の期待値を，自然勾配法により最適化する．これは，次のように事後分布を更新することに相当する．

$$
\begin{pmatrix} \tilde{a}^{(h)}_{f,i} \\ \tilde{b}^{(h)}_{f,i} \\ \tilde{a}^{(g)}_i \\ \tilde{b}^{(g)}_i \end{pmatrix} := (1-\rho_t) \begin{pmatrix} \tilde{a}^{(h)}_{f,i} \\ \tilde{b}^{(h)}_{f,i} \\ \tilde{a}^{(g)}_i \\ \tilde{b}^{(g)}_i \end{pmatrix} + \rho_t \begin{pmatrix} \bar{a}^{(h)}_{f,i} \\ \bar{b}^{(h)}_{f,i} \\ \bar{a}^{(g)}_i \\ \bar{b}^{(g)}_i \end{pmatrix} \quad (13)
$$

$$
q(h_{f,i}) = \mathrm{Gam}\left(\tilde{a}^{(h)}_{f,i}(t), \tilde{b}^{(h)}_{f,i}(t) \right) \quad (14)
$$

$$
q(g_i) = \mathrm{Gam}\left(\tilde{a}^{(g)}_i(t), \tilde{b}^{(g)}_i(t) \right) \quad (15)
$$

ただし，$\rho_t \in (0,1)$ は，勾配法のステップ係数である．$(t+t_0)^{-r}$, $r = [0.5, 1)$ という形であると，収束が保証される．また，

$$
\bar{a}^{(h)}_{f,i}(t) = a^{(h)} + T\Phi_{f,i} \quad (16)
$$

$$
\bar{b}^{(h)}_{f,i}(t) = b^{(h)} + T\langle g_{i+1} \rangle X_{f,t-i} \quad (17)
$$

$$
\bar{a}^{(g)}_i(t) = \alpha^{(g)}/I + T\sum_f \Phi_{f,i} \quad (18)
$$

$$
\bar{b}^{(g)}_i(t) = \alpha^{(g)} c^{(g)} + T\sum_f \langle h_{f,i} \rangle X_{f,t-i} \quad (19)
$$

である．

3 評価

本手法の有効性を確認するため，音源分離の評価項目 BSS_EVAL[11] のうち，Signal-to-Distortion Ratio (SDR) と Signal-to-Interference Ratio (SIR) の，セミブラインド音声強調適用前後での変化量 ΔSDR と ΔSIR を評価する．SDR は信号歪みを表し，目的音がどれだけ歪められているかを表す．また，SIR は，自己発話の除去度合いを表す．

目的音として，男性の文章朗読音 30 秒とギターのコード演奏音 30 秒を用意した．また，自己発話としては，楽曲自動伴奏用途を想定し，ギターコード演奏に加え，ソウル音楽とピアノ三重奏を 30 秒ずつ用意した．室内音響系として，RWCP 実環境データベース [12] に収録された，インパルス応答 3 種類 (E2A= “Dry,” JR1=“Tatami,” OFC= “Conf. Room”) と，Pori Concert hall のインパルス応答 [13](s1_p1_o= “Hall”) を用意した．なお，全ての波形は 44.1kHz でサンプリングされた．全ての波形に対して，フレーム長 4096 サンプル，ホップ長 1024 サンプル，ハミング窓で，振幅スペクトログラムを計算し，Y の平均値が 50 になるよう正規化し，X も同じ係数で正規化した．パラメータは，特に他に指定がない限り $a^{(s)} = b^{(s)} = a^{(h)} = b^{(h)} = \alpha(g) = c(g) = 1$，$I = 10$ とした．以下，目的音対自己発話の S/N 比が −5dB の結果を示す．ちなみに，S/N に応じて，評価値の差はあるものの，大まかな傾向は変わらない．

まずは，本手法における I を，有限としてモデル化した場合 ($p(g_i) = \delta(g_i - 1)$ とする) と，提案手法の

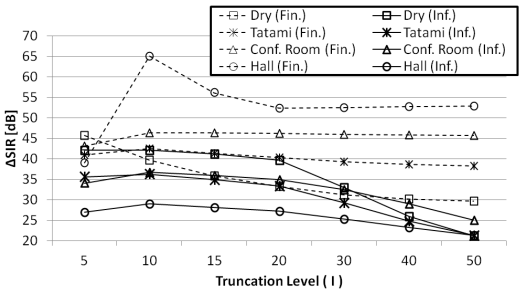

Fig. 1　無限と有限次数モデルにおける，次数 I に対する ΔSIR の，インパルス応答別の比較．“Fin.” は有限次数，“Inf.” は無限次数．

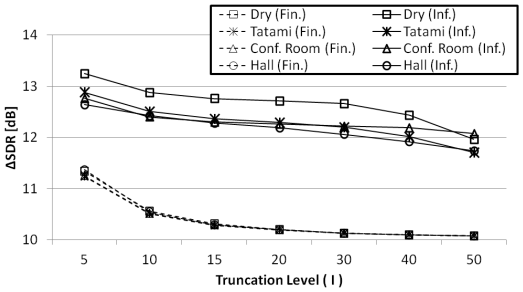

Fig. 2　無限と有限次数モデルにおける，次数 I に対する ΔSDR の，インパルス応答別の比較．“Fin.” は有限次数，“Inf.” は無限次数．

ように無限として扱った場合の近似としてモデル化した場合の評価値を比較する．ΔSIR の結果を図 1 に，ΔSDR の結果を図 2 に図示する．これらから，無限次数は，有限次数と比べ ΔSIR が低く，ΔSDR が高くなることが多いことが分かる．つまり，無限次数にすることにより，音を過剰に削られにくくなるということを表している．なお，ガンマ過程を導入しても，評価値に I に対する依存性があるのは，I を増やすことにより，g の縮退効果が，過剰に効きすぎてくるためだと考えられる．特に，$x_{f,t}$ と $x_{f,t-i}$ は，i が小さければ強い相関を持つ．そのため，信号モデルや，統計モデルで仮定している独立性が満たされず，モデルと現実の乖離が目立つために，I への依存性が起こると考えられる．

また，オンライン推論と，オフライン推論での，ΔSDR と ΔSIR を比較したものを，図 3 に示す．この図から，オンライン推論は，$a^{(s)}$ が小さいと，音を過剰に削りすぎる傾向があり，$a^{(s)} > 0.3$ 辺りからは，両者とも似た挙動を示すようになることが分かる．

なお，本手法のオンラインアルゴリズムは，Apple iPad2 上で実時間動作することが，確認された（Fs=8kHz）．

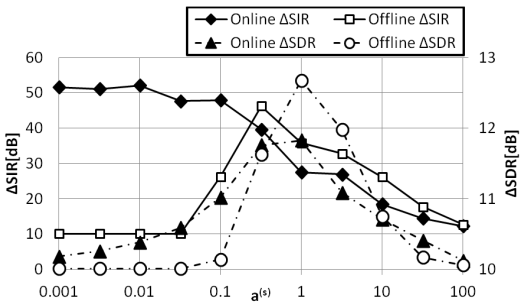

Fig. 3　$a^{(s)}$ に対する ΔSDR と ΔSIR.

4　まとめ

本稿では，変分ベイズ法に基づく実時間セミブラインド音源分離タスクのベイズ推定手法を提案した．筆者の知る限り，NMF や畳込 NMF のオンラインベイズ推定を定式化・評価したのは本稿が初めてである．残響のパラメータは，ガンマ過程を応用することにより，事前に設定する次数に対して，ロバストネスを実現した．今後は，本手法を，音楽情報処理タスクなどに応用したい．

参考文献

[1] J. Even, et al. Semi-blind suppression of internal noise for hands-free robot spoken dialog system. In *IROS*, pp. 658–663, October 2009.

[2] R. Takeda, et al. Barge-in-able robot audition based on ICA and missing feature theory under semi-blind situation. In *IROS*, pp. 1718–1723, 2008.

[3] 澤田紘志ほか. セミブラインド音源分離を用いたロボット音声対話システムのための内雑音抑圧手法. 音講論, 2-4-18, pp. 655–658, 2009.

[4] M. Babaie-Zadeh, et al. Semi-Blind Approaches for Source Separation and Independent component Analysis. In *ESANN*, pp. 301–312, 2006.

[5] 池澤浩気ほか. 非定常雑音・時変残響環境下でのパワースペクトログラム領域セミブラインド音声強調. 信学総大, A-4-5, p. 72.

[6] M. Hoffman, et al. Bayesian Nonparametric Matrix Factorization for Recorded Music. In *ICML*, pp. 439–446, 2010.

[7] 前澤陽. 無限オーダーの残響フィルタを持つ残響抑圧信号のベイズ推定手法. 音講論, March 2012.

[8] M. Hoffman, et al. Online learning for latent dirichlet allocation. In *NIPS*, pp. 856–864. 2010.

[9] J. F. C. Kingman. *Poisson Processes*. 1992.

[10] M. Sato. Online Model Selection Based on the Variational Bayes. *Neural Computation*, Vol. 13, No. 7, pp. 1649–1681, July 2001.

[11] E. Vincent, et al. Performance measurement in blind audio source separation. *TASLP*, Vol. 14, No. 4, pp. 1462–1469, 2006.

[12] 比屋根一雄ほか. RWCP 実環境音声・音響データベース. 信学総大, p. 257, 1999.

[13] J. Merimaa, et al. Concert Hall Impulse Responses - Pori, Finland: Reference. Retrieved from http://www.acoustics.hut.fi/projects/poririrs on 1/27/2013, 2005.