無限混合モデルを入れ子にしたmIBPに基づく音響イベント検出*

◎大石康智 (NTT), 持橋大地, 松井知子 (統数研), 中野允裕,
亀岡弘和, 泉谷知範, 柏野邦夫 (NTT)

1 はじめに

膨大な音や映像のメディアデータが身の回りにあふれる中，これらのデータを自在に検索して活用するためには，付随するテキストデータに頼るだけではなく，それぞれの中身を表す情報を，音や映像自体から自動的に引き出す技術が必要不可欠である。音響信号から，人に認識されうる音の事象，例えば，話声や歌声，笑い声やあいづちをはじめ，動物の鳴き声，楽器音，環境音，効果音などを自動的に書き起こす音響イベント検出もその一つの技術であり，現在盛んに研究が行われている [1–5]。

本稿では，音響イベント検出における2つの課題に取り組む。1つ目の課題は音響イベントの重なりがこれまで十分に考慮されなかった点である。多くの研究では，イベントの重ね合わせは無視し，各時刻で最も顕著なイベントだけを出力した。音響イベントがスパースに現れる音環境では十分な性能が得られるものの，音響イベントが豊富に含まれる環境を対象とすると検出が難しい [1]。文献 [7, 8] では，あらかじめ音響信号を複数のトラックに音源分離し，トラックごとに音響特徴量を抽出して，イベントを検出した。音響イベントの重なりを考慮できるが，音環境に合わせてトラック数を手動で調整する必要があった。

2つ目の課題は，数千時間の書き起こしデータを用いる音声認識に比べ，音響イベントの音響的特徴を学習するためのラベル付データベースが少ない点である [3, 9]。さらに，検出対象とする音響イベントを増やすには，新たなラベル付データが必要である。このようなデータスパースネスや未知の音響イベント検出問題に対処するために，教師なし，または半教師あり学習の枠組みを導入することは有望である。

本研究では，これら2つの課題に取り組むために，非負値行列因子分解（NMF）[10] をベースとした，音響イベント検出のための音響信号モデルを提案する。これは，音響イベントが複数の音響オブジェクト（音声における音素，楽器音における単音など）から構成されると想定し，まずはこのオブジェクトをすべて書き起こすモデルである。NMF を音響信号の振幅スペクトログラムに適用することで，音響イベントの重ね合わせを考慮した上で，各音響オブジェクトを表現する基底スペクトルが教師なしで学習される。

さらに，音響オブジェクトの音響的特徴や数を教師なしで学習するために，2つのノンパラメトリックベイズアプローチを導入する。各音響オブジェクトの音響的特徴を表現する基底スペクトルの状態数を自動的に推定するために，Chinese Restaurant Process (CRP)[11] を導入する。音響信号に含まれる音響オブジェクトの総数を自動的に推定するために，Markov Indian Buffet Process (mIBP)[12] を導入する。これらの確率過程のパラメータや NMF のパラメータは，スライスサンプリングを用いて，効率的に推論される。

評価実験では，3種類の音響信号を用いて，提案手法の基本動作と，推定される音響オブジェクトの発音区間を考察する。

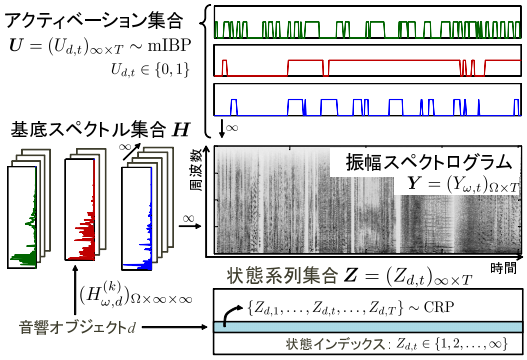

Fig. 1 可変基底型 NMF をベースとした音響信号モデル（D：音響オブジェクトの総数，K_d：音響オブジェクト d の基底スペクトルの状態数，Ω：ナイキスト周波数，T：分析フレーム数）

2 音響イベント検出のための信号モデル

NMF に基づく音響信号解析では，振幅スペクトログラムもしくはパワースペクトログラム $\boldsymbol{Y} = (Y_{\omega,t})_{\Omega\times T} \in \mathbb{R}^{\geq 0,\Omega\times T}$ を基底スペクトル集合 $\boldsymbol{H} = (H_{\omega,d})_{\Omega\times D} \in \mathbb{R}^{\geq 0,\Omega\times D}$ とアクティベーション集合 $\boldsymbol{U} = (U_{d,t})_{D\times T} \in \mathbb{R}^{\geq 0,D\times T}$ の積で近似する。ここで，$\omega = 1,\ldots,\Omega$ は周波数インデックス，$t = 1,\ldots,T$ は分析フレームのインデックスとする。すなわち，$Y_{\omega,t} \approx \sum_d^D H_{\omega,d}U_{d,t}$ のように観測スペクトログラム $\boldsymbol{Y}$ を D 個の頻出の基底スペクトル $H_{:,d} = [H_{1,d},\ldots,H_{\Omega,d}]^{\mathrm{T}}$ とその音量変化を表すアクティベーション $U_{d,:} = [U_{d,1},\ldots,U_{d,T}]$ で近似することに相当する。$H_{:,d}$ と $U_{d,:}$ のペアを一般的にコンポーネントと呼ぶ [13]。このコンポーネントが音響オブジェクトに相当する。

本研究では，音源分離において性能が良いと報告される振幅スペクトログラムに対する一般化 Kullback-Leibler(KL) divergence 規準の NMF を利用する。こ

* Audio event detection based on Markov Indian buffet process nesting infinite mixture models. by OHISHI, Yasunori (NTT), MOCHIHASHI, Daichi, MATSUI, Tomoko (The Institute of Statistical Mathematics), NAKANO, Masahiro, KAMEOKA, Hirokazu, IZUMITANI, Tomonori, KASHINO, Kunio (NTT)

れは次のようなモデルの最尤推定問題と等価であることが知られている [14]。

$$Y_{\omega,t} = \sum_d C_{\omega,t,d},\ C_{\omega,t,d} \sim \text{Poisson}(H_{\omega,d}U_{d,t}) \quad (1)$$

さらに，Poisson 分布の共役事前分布である Gamma 分布を用いて，基底スペクトルとアクティベーションの事前分布 $H_{\omega,d} \sim \text{Gamma}(a_H, b_H), U_{d,t} \sim \text{Gamma}(a_U, b_U)$ を導入した Bayesian NMF も提案されている [15].

コンポーネントによって，一つの音響オブジェクトが表現されることが望ましいが，実際の音響オブジェクトのスペクトルは時間的に変化する（例えば，音素における定常状態や次の音素への “わたり”）。基底スペクトルが時間にともなって変化するように拡張した，可変基底型 NMF が提案されている [16]。これはフレーム t において，基底スペクトルが，ある一つの状態 $Z_{d,t} \in \mathbb{N}$ をとると見なし，式 (1) を以下のように拡張する。

$$Y_{\omega,t} = \sum_d C_{\omega,t,d},\ C_{\omega,t,d} \sim \text{Poisson}(H_{\omega,d}^{(Z_{d,t})}U_{d,t}) \quad (2)$$

基底スペクトルの詳細な時間構造をモデル化するために，状態遷移を考慮した NMF も提案されるが [17]，本研究では音響オブジェクトの検出を目的とするため（例えば，音素の時間構造まで表現する必要はないため），可変基底型の Bayesian NMF を土台する。ただし，音響オブジェクトの発音区間（ON/OFF）を表現するため，$\boldsymbol{U}$ は 0(OFF) もしくは 1(ON) の値からなる 2 値行列とする。これにより，音量に関する情報は基底スペクトルに含まれて表現される（Fig. 1）。

さらに 2 つのノンパラメトリックベイズアプローチを信号モデルに導入する。1 つ目のアプローチとして，各音響オブジェクトの基底スペクトルの状態系列 $\{Z_{d,1}, \ldots, Z_{d,T}\}$ の事前分布に CRP を導入する。これは中野らによって提案された無限状態スペクトルモデル [16] に相当し，従来は音楽音響信号解析のために用いられた。各音響オブジェクトの基底スペクトルの状態数がデータから決定される。2 つ目のアプローチとして，アクティベーション集合 $\boldsymbol{U}$ の事前分布に mIBP を導入する．これにより，$\boldsymbol{U}$ の各行成分にマルコフ性が仮定され，音響信号に含まれる音響オブジェクトの総数がデータから決定される。以下，2 つのアプローチの導入方法を述べる。

2.1 Chinese Restaurant Process の導入

音響イベントを表現するために必要な，基底スペクトルの状態数は音響イベントごとに異なる。例えば，ピアノの単音であれば，“attack”，“decay”，“sustain”，“release” と呼ばれる状態があるため，4 つの基底スペクトルを用いて表現されるだろう。ドアの開閉音や食器音は，楽器の単音に比べて突発音であるため，基底スペクトルの状態数は少ないかもしれない。交通騒音（ノイズ）は常に 1 つの基底スペクトルで表現されるかもしれない．一方で，音声は複数の音素で構成されるため，多くの基底スペクトルを必要とするだろう。このように，基底スペクトルの状態数は固定するのではなく，音響信号から自動的に決定されることが望ましい。中野らは可変基底型 NMF にディリクレ過程を導入した無限状態スペクトルモデルを提案し，音楽音響信号を楽器の単音ごとに分解できることを示した [16]。本研究でも同様の枠組みを音響イベント検出のために利用する。

音響オブジェクト d の状態系列 $Z_{d,1}, \ldots, Z_{d,T}$ はそれぞれ離散的な値 $1, \ldots, K_d$（状態インデックス）をとる。このとき，状態数を $K_d \to \infty$ として，$\{Z_{d,1}, \ldots, Z_{d,t-1}, Z_{d,t+1}, \ldots, Z_{d,T}\}$ (以降では $\boldsymbol{Z}_{d,\backslash t}$ と表す) が与えられたときの $Z_{d,t}$ の条件付き確率は次のように表せる。

$$p(Z_{d,t} = k | \boldsymbol{Z}_{d,\backslash t}, \beta_d) = \begin{cases} \dfrac{n_{d,\backslash t}^{(k)}}{(T-1+\beta_d)} & (n_{d,\backslash t}^{(k)} > 0) \\ \dfrac{\beta_d}{(T-1+\beta_d)} & (k = K_{\backslash t,+} + 1) \end{cases} \quad (3)$$

ここで，$n_{d,\backslash t}^{(k)}$ は $Z_{d,t'} = k$ $(Z_{d,t'} \in \boldsymbol{Z}_{d,\backslash t})$ を満たす t' の個数を表す。また，$K_{\backslash t,+}$ は $n_{d,\backslash t}^{(k)} > 0$ となるクラスの個数である。これが CRP[11] と呼ばれ，ディリクレ過程の一構成法を与える。この過程では，各時刻に用いられる状態が，他の時刻に多く用いられている状態ほど使われやすくなる性質がある。また，新しい状態が用いられやすくなるか否かは正のパラメータ β_d によって調整される。これらにより，K_d がデータから自動的に決定される。

2.2 Markov Indian Buffet Process の導入

Gael らは，Indian Buffet Process (IBP) と呼ばれるノンパラメトリックベイジアン因子モデルの一構成法を，時系列データのために拡張した mIBP を提案した [12]。この mIBP を提案モデルの 2 値行列 $\boldsymbol{U}$（D 行 T 列の行列）に適用すると，各音響オブジェクトの T フレームにわたる 2 値のアクティベーションが，一次マルコフ連鎖に従って生成される。一次マルコフ連鎖により，音響オブジェクトの時間的な持続性が表現されるため，2.1 節で述べた各音響オブジェクトの基底スペクトルの状態数がデータから効果的に決定されることも期待できる。

mIBP に基づいて，$\boldsymbol{U}$ の確率分布を導出する。いま音響オブジェクト d のアクティベーションの状態遷移を表現する遷移行列を

$$\boldsymbol{W}^{(d)} = \begin{bmatrix} 1-a_d & a_d \\ 1-b_d & b_d \end{bmatrix} \quad (4)$$

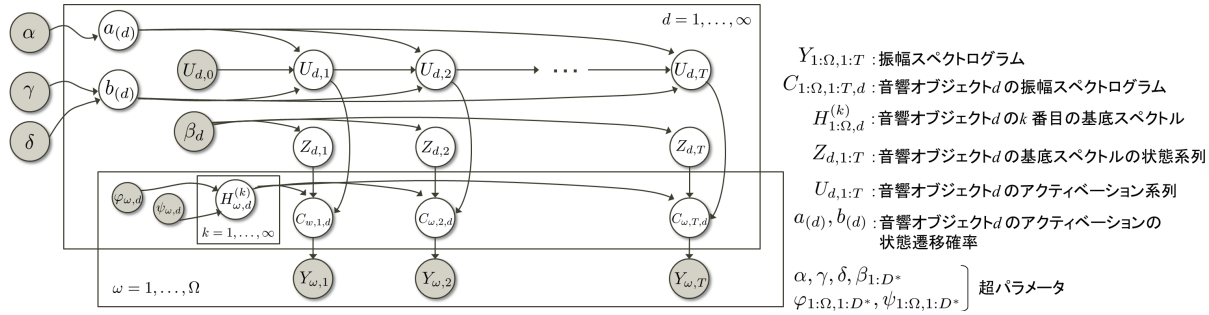

Fig. 2　提案する音響信号モデルのグラフィカル表現

とする．ここで，$W^{(d)}_{i,j} = p(U_{d,t+1} = j-1 | U_{d,t} = i-1)$，$i, j \in \{1, 2\}$ は遷移確率を表す．$U_{d,0} = 0$ とすると，$U_{d,t}$ は，$U_{d,t} \sim \text{Bernoulli}(a_d^{1-U_{d,t-1}} b_d^{U_{d,t-1}})$ のようなベルヌーイ分布から生成される確率変数として記述できる．ここで，$c_d^{00}, c_d^{01}, c_d^{10}, c_d^{11}$ をそれぞれ $0 \to 0$，$0 \to 1$，$1 \to 0$，$1 \to 1$ に遷移する回数とすると，$\boldsymbol{U}$ の確率分布は，

$$p(\boldsymbol{U}|\boldsymbol{a}, \boldsymbol{b}) = \prod_d (1-a_d)^{c_d^{00}} a_d^{c_d^{01}} (1-b_d)^{c_d^{10}} b_d^{c_d^{11}} \quad (5)$$

と記述できる．ここで，$\boldsymbol{a} = \{a_1, \ldots, a_D\}$，$\boldsymbol{b} = \{b_1, \ldots, b_D\}$ とした．a_d と b_d は，その共役性から $a_d \sim \text{Beta}(\alpha/D, 1), b_d \sim \text{Beta}(\gamma, \delta)$ とする．さらに，$D \to \infty$ とするために，Stick-breaking construction[18] を導入する．$\boldsymbol{a}$ を $a_{(1)} > a_{(2)} > \ldots > a_{(D)}$ のように順序付けて $D \to \infty$ とすると，$a_{(d)}$，$b_{(d)}$ の生成過程は

$$\nu_d \sim \text{Beta}(\alpha, 1),\ a_{(d)} = \nu_d a_{(d-1)} = \prod_{l=1}^{d} \nu_l \quad (6)$$

$$b_{(d)} \sim \text{Beta}(\gamma, \delta) \quad (7)$$

と記述できる。提案モデルの生成過程を表現するグラフィカルモデルを Fig. 2 に示す。ここで，基底スペクトルの事前分布はガンマ分布 $H^{(k)}_{\omega,d} \sim \text{Gamma}(\varphi_{\omega,d}, \psi_{\omega,d})$ を仮定する。また，超パラメータの値は本稿では固定する。Fig. 2 より，Gael らの提案する mIBP に無限混合モデルが入れ子になった音響信号モデルであると言える。

3　パラメータの推論

文献 [12] を参考に，スライスサンプリング [19] と動的計画法 [20] を組み合わせて，提案モデルのパラメータを推論する。これらにより，Stick-breaking construction の打ち切り数が自動的に調整される。

まず，スライス補助変数 s を導入する。

$$s|\boldsymbol{a}, \boldsymbol{U} \sim \text{Uniform}(0, \min_{d:\exists t,\ U_{d,t}=1} a_{(d)}) \quad (8)$$

s が与えられたとき，$\boldsymbol{U}$ の条件付き確率は，

$$p(\boldsymbol{U}|\boldsymbol{Y}, \boldsymbol{C}, \boldsymbol{Z}, \boldsymbol{H}, s, \boldsymbol{a}, \boldsymbol{b}) \quad (9)$$
$$\propto p(\boldsymbol{U}|\boldsymbol{Y}, \boldsymbol{C}, \boldsymbol{Z}, \boldsymbol{H}, \boldsymbol{a}, \boldsymbol{b}) \frac{\mathbb{I}(0 \le s \le \min_{d:\exists t,\ U_{d,t}=1} a_{(d)})}{\min_{d:\exists t,\ U_{d,t}=1} a_{(d)}}$$

となる。ここで，$\mathbb{I}(A)$ は，A が真であるとき $\mathbb{I}(A) = 1$ となり，それ以外で 0 となる関数とする。ここで，D^* を $a_{(d)} > s$ となる最大の音響オブジェクトのインデックスとすると，上記の式は $d > D^*$ の音響オブジェクトは削除し，$d \le D^*$ の音響オブジェクトのパラメータ $\boldsymbol{C}, \boldsymbol{Z}, \boldsymbol{H}, \boldsymbol{U}, \boldsymbol{a}, \boldsymbol{b}$ を更新すれば良いことを意味する。D^* は打ち切り数に相当し，無限個の音響オブジェクトを対象とすることなく，計算コストを有限の量に制限できる。各パラメータの更新式は文献 [21] を参照されたい。

4　評価実験

提案法の基本動作とパラメータの推定結果を評価する。実験用に 3 種類の音響信号を用意した（サンプリング周波数 16 kHz，量子化ビット数 16 とする）。

音響信号 A　Free Sound Effects[22] から取得した合計 6 種類の効果音（救急車のサイレン音，拍手音，歓声，犬の鳴き声，飛行機のジェット音，雨音）および男女 1 名ずつによる 10 秒間の話声，男女 1 名ずつによる 10 秒間の歌声を 2 秒間の無音区間を挿入して繋ぎあわせた音響信号

音響信号 B　テレビ CM の音響信号（100 秒）

まず，音響信号をフレーム長 64 ms，フレームシフト長 250 ms で短時間フーリエ変換して振幅スペクトログラムを得る。そして，GaP-NMF[23] を利用して，$\boldsymbol{H}$ と $\boldsymbol{U}$ の初期値を求める。このとき，音響オブジェクトの初期値数は $D^* = 30$ とした。$\boldsymbol{H}$ は，各音響オブジェクトの基底スペクトルの初期値として利用する（状態数 K_d の初期値は 1 とする）。一方，$\boldsymbol{U}$ はその中央値を閾値として 2 値行列に変換し，初期値とする。超パラメータは，$\alpha = 4,\ \beta_d = 1,\ \gamma = 500,\ \delta = 1,\ \varphi_{\omega,d} = 1, \psi_{\omega,d} = 1\ (d = 1, \ldots, D^*,\ \omega = 1, \ldots, \Omega)$ と設定した。各パラメータの更新回数は 500 回とした。

Fig. 3，4 は，各音響信号の振幅スペクトログラム，音響イベントの発音区間のラベル，音響オブジェクトの発音区間推定結果 $\boldsymbol{U}$ を示す。ラベルと推定結果では，黒色部分が発音区間を示す。mIBP を導入するため，頻出する音響オブジェクトほど小さいインデックスとなる。推定された複数の音響オブジェクトを組み合せることによって，ラベル付された音響イベントを表現できる可能性が示唆される。また，30 個の音

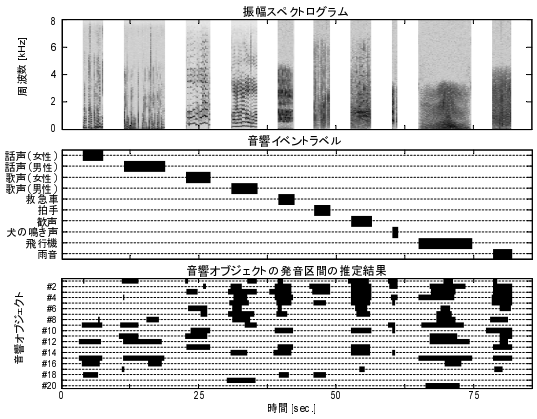

Fig. 3　音響信号 A に含まれる音響オブジェクトの発音区間の推定結果

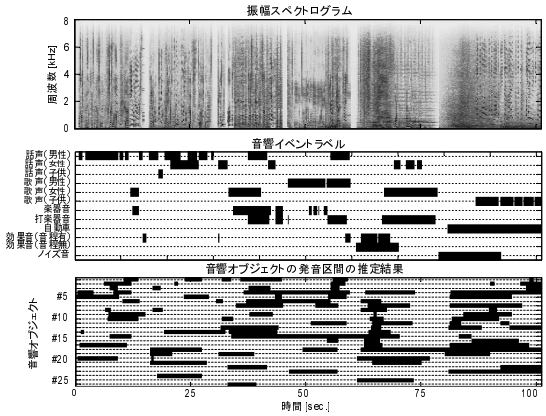

Fig. 4　音響信号 B に含まれる音響オブジェクトの発音区間の推定結果

響オブジェクトを初期値としたが，この個数が音響信号から自動的に学習される点も特筆すべき点である。

Fig. 3 は**音響信号 A** の推定結果であり，特に着目したいのは，飛行機のジェット音区間の推定結果である。音響オブジェクト#9 と#15 は終始アクティブであるが，#6，#7，#8 は部分的にアクティブとなる。これらは，時々刻々と変化するジェット音の高域の音響的特徴を表現する役割を担っていると言える。推定された音響オブジェトの役割の違いを確認できる。

Fig. 4 は**音響信号 B** の分析結果である。音響信号を実際に聞きながら，12 種類の音響イベントを 200 ms ごとにラベル付けした。約 15 秒ごとに切り替わる CM の音楽（楽器音）を音響オブジェクトとして推定できた。一方，話声や効果音がどの音響オブジェクトによって表現されるか対応付けることが難しい。

実験結果より，音響オブジェクトを組み合わせて，それがどのラベルに相当するかを特定する枠組みが必要である。例えば，文献 [24] の sLDA を参考に，一部のラベル付データを利用した半教師あり学習への拡張が考えられる。各時刻の $U_{:,t}$ とラベルとの回帰問題を考えることになる。

しかしながら，複雑に重なり合った音響イベントを音響オブジェクトとして，2 値行列で書き起こすことのできる提案法の有効性を定性的ではあるが確認できた。情報圧縮された 2 値行列を利用すれば，例えばメディア検索への応用が期待できる.

5　まとめと今後の展開

本研究では，音の重ね合わせを考慮した上で，音響信号から様々な音響イベントを自動的に書き起こすことを目指す。本稿は，音響イベントがいくつかの音響オブジェクトによって構成されると想定し，教師なしの下，音響オブジェクトの音響的特徴や発音区間を周波数領域で推定する音響信号モデルを考案した。提案モデルは NMF をベースとし，2 つのノンパラメトリックベイズアプローチを導入することで，各音響オブジェクトの基底スペクトルの状態数や種類数を音響信号から自動的に学習する枠組みをもつ。2 種類の音響信号を用いて，提案モデルの基本動作を確認したところ，推定された音響オブジェクトの組み合わせることで，ラベルに対応付けられる可能性を示すことができた。

今後の課題は，半教師あり学習の下で音響イベントを特定すること，そして定量的な評価を行うことである。また，長時間の音響信号に適用できるよう，推論アルゴリズムを工夫し，高速化させることも検討中である。

参考文献

[1] A. Temko *et al.*, *Pattern Recogn. Lett.*, vol.30, no.14, pp. 1281–1288, 2009.
[2] T. Butko *et al.*, *Proc. ICASSP 2011*, pp. 357–360.
[3] 佐々木ほか, 情処研報, 2011–SLP87–6, 2011.
[4] A. Kumar *et al.*, *Proc. ICASSP 2012*, pp. 489–492.
[5] 井本ほか, 音講論集, 2-Q-32，pp. 975–976, 2012.
[6] A. Mesaros *et al.*, *Proc. EUSIPCO 2011*, pp. 1307–1311.
[7] T. Heittola *et al.*, *Proc. CHiME 2011*, pp. 36–40.
[8] M. Espi *et al.*, *Proc. ICASSP 2012*, pp. 4293–4296.
[9] Z. Zhang *et al.*, *Proc. ICASSP 2012*, pp. 333–336.
[10] D. D. Lee *et al.*, *Proc. NIPS 2000*.
[11] Y. W. Teh *et al.*, *Cambridge University Press*, 2010.
[12] J. V. Gael *et al.*, *Proc. NIPS 2008*.
[13] T. Virtanen, *IEEE TASLP*, 15, pp. 1066–1074, 2007.
[14] T. Virtanen *et al.*, *Proc. ICASSP 2008*, pp. 1825–1828.
[15] A. T. Cegmil, *Computational Intelligence and Neuroscience*, vol.2009, no.4, 2009.
[16] M. Nakano *et al.*, *Proc. ICASSP 2011*, pp. 22–27.
[17] M. Nakano *et al.*, *Proc. WASSPA 2011*, pp. 325–328.
[18] Y. W. Teh *et al.*, *Proc. NIPS 2007*.
[19] R. M. Neal, *Annals of Statistics*, vol.31, pp. 705–767, 2003.
[20] S. L. Scott, *JASA*, vol.97, pp. 337–351, 2002.
[21] 大石ほか, PRMU2012-29, pp. 37–42, 2012.
[22] http://www.partnersinrhyme.com
[23] M. Hoffman *et al.*, *Proc. ICML 2010*.
[24] D. M. Blei *et al.*, *Proc. NIPS 2007*.