ディリクレ事前分布を用いた音声のスパース性に基づく音源数推定と音源分離*

○荒木章子, 中谷智広, 澤田宏
(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)

1 はじめに

ブラインド音源分離 (BSS) は、信号の独立性を仮定する独立成分分析に基づく手法 ([1] 他) や音声のスパース性を仮定する時間周波数マスクに基づく手法 ([2] 他) など、既に多数の提案があり、実環境においても比較的頑健かつ高精度に動作することが確認されている。ところが、従来の BSS 手法は、音源数が既知であることを仮定している手法がほとんどであり、音源数未知の場合の BSS 手法は確立されていない。実際に BSS を利用する場合、音源数が既知であることは稀であり、観測信号から音源数を推定しながら音源分離を達成する方法が望まれる。

そこで本稿では、音源数が未知の場合において、音源数と時間周波数マスクとを同時に推定する方法を提案する。提案法では、観測信号から推定した各時間周波数の音源方向 (DOA) 情報をクラスタリングすることで音源分離を行なう。音源がスパース [2] である場合、DOA の分布は図 1(a) に示すように音源個数分 (図 1(a) では 4 個) のクラスタを形成し、それぞれのクラスタが各音源に対応する。そこで提案法では、DOA の分布を混合正規分布 (GMM) でモデル化し、特に、各クラスタをそれぞれ 1 個の正規分布で表すことを考える。これが可能になれば、クラスタを表す正規分布を数えあげることで音源数推定ができる。またそれぞれの正規分布から時間周波数マスクを作成することも可能となる。

GMM を用いた DOA のクラスタリングは、これまでも [3, 4] などで提案されている。しかしこれらは音源数 N_s を必要とし、音源数 N_s が未知の場合には例えば 1 つのクラスタを複数の正規分布でモデル化してしまう。本稿では、この問題を回避し、各クラスタにそれぞれ 1 個の正規分布をあてはめるために、GMM の混合重みにディリクレ分布による事前分布を与える。この事前分布を用いると、始めに必要数以上の個数の正規分布を用意しても、DOA の分布をなるべく少ない個数の正規分布でモデル化しようとし、その他の不要な正規分布の混合重みはほぼゼロとなる。これにより、各クラスタをそれぞれ 1 個の正規分布で表すことができ、音源数推定と音源分離が可能となる。

本稿では、GMM の混合重みにディリクレ事前分布を用いた DOA 分布のモデルパラメタ推定アルゴリズムを具体的に示す。またモデルパラメタから、音源数と時間周波数マスクを推定する方法についても述べる。最後に、提案法が音源数を正確に推定できることを確認したので報告する。

2 問題設定

N_s 個の音声信号 $\mathrm{s}_1, \ldots, \mathrm{s}_N$ が、部屋の残響の影響を受け、N_m 個のマイクで観測されたとすると、観測信号は次のようにモデル化できる。

$$\mathrm{x}_j(t) = \sum_{i=1}^{N_s} \sum_l \mathrm{h}_{ji}(l)\, \mathrm{s}_i(t-l),\ j = 1, \ldots, N_m, \quad (1)$$

ここで、$\mathrm{h}_{jk}(l)$ は音源 k からマイク j へのインパルス応答である。また本稿では音源数 N_s は未知であるとする。

本稿の目的は、収録された観測信号 x_j のみから、音源数 N_s と、それぞれの音声信号の推定値である分離音 y_i を得ることである。

本稿では、時間領域での観測信号 $\mathrm{x}_j(t)$ $(j = 1, \ldots, M)$ に短時間フーリエ変換 (STFT) を適用し、時間周波数領域にて信号を

$$x_j(f,\tau) = \sum_{i=1}^{N_s} h_{ji}(f) s_i(f,\tau)$$

として扱う。ここで、f と τ はそれぞれ、周波数とフレーム番号を示す。

また本稿では、信号のスパース性 [2] を仮定する。即ち、各時間周波数において、高々1 つの信号 $s_i(f,\tau)$ のみが支配的であり、複数の信号が互いに重ならないことを仮定する：

$$x_j(f,\tau) \approx h_{ji}(f) s_i(f,\tau)$$

このスパース性の仮定は、音声信号などで、時間周波数領域で近似的に成り立つことが知られている [2, 5]。

2.1 音源分離の方法

音源分離の方法としては、各時間周波数で推定した音源方向 (DOA) を分類する方法を取る。

始めに、各時間周波数における DOA $d(f,\tau)$ を、各マイクペア j-j' 間の到来時間差 $q_{jj'}(f,\tau) = \frac{1}{2\pi f} \arg[x_j(f,\tau)\ x^*_{j'}(f,\tau)]$ とマイクの座標を用いて推定する [6]. 本稿では、マイク間隔が十分小さく、空間的エイリアジングが生じない場合について議論する。空間的エイリアジングが生じる場合の方法については [7] を参照されたい。

*Source number estimation and source separation of sparse signals with Dirichlet prior. by ARAKI, Shoko, NAKATANI, Tomohiro, SAWADA, Hiroshi (NTT Communication Science Laboratories, NTT Corporation)

次に DOA を分類し、同じ到来方向を持つ時間周波数成分を集める。ここで得られた各クラスタが、各音源に対応する。もし音源数 N_s が既知であれば、k-means 法などで分類できる [5]。本稿では、GMM をあてはめる方法をとる。方法の詳細は 3 章に述べる。

最後に、時間周波数マスク $M_i(f,\tau)$ を用いて i 番目のクラスタ成分を抽出し、分離音 $y_i(f,\tau)$ を得る：

$$y_i(f,\tau) = x_1(f,\tau)M_i(f,\tau). \tag{2}$$

2.2 DOA 分布の確率モデル

各時間周波数における DOA $d(f,\tau)$ を観測サンプル、$a(f,\tau) = |x(f,\tau)|^2$ をパワー重みとする。また今後、記載の簡単のため、$d(f,\tau) \to d_n$、$a(f,\tau) \to a_n$ と記載する。但し、$n = \tau F + f$ (F は周波数の数) である。すなわち、DOA 観測データ $d = \{d_1, d_2, \cdots, d_n, \cdots, d_N\}$ 、パワー重み $a = \{a_1, a_2, \cdots, a_n, \cdots, a_N\}$ とする。ここで $N = T \times F$(T はフレーム数) は観測の数である。

まず本稿では、音源数が 1 の場合、DOA 観測データが正規分布 (ラップドガウシアン (wrapped Gaussian)[8]、詳しくは後述)

$$\sum_{k=-\infty}^{\infty} \mathcal{N}(d_n + 2\pi k; \mu, \sigma^2) \tag{3}$$

に従うと仮定する。ここで、$\mathcal{N}(d_n;\mu,\sigma^2)$ は平均 μ、分散 σ^2 の正規分布、$-\pi \le d_n < \pi$ である。ここでは、位相のラッピングを考慮するため、ラップドガウシアン [8] を用いている。この理由を述べる。DOA 情報は $-\pi$ と π の値を取るため、例えば音源方向が $\pm\pi$ に近い場合には二峰性 (bimodal) の分布形状となる。これをここでは (3) のように、DOA 観測データに 2π の任意性を許すことでモデル化する。詳しくは [8] を参照されたい。

実際の観測信号には、充分数の音源が存在し、そのうちのいくつかのみがアクティブであると仮定する。この時、観測データが、1 つの正規分布が 1 つの音源方向に対応するような混合正規分布 (GMM) に従うと仮定する：

$$p(d_n;\mu,\boldsymbol{\sigma^2}) = \sum_{m=1}^{M} \alpha_m \sum_{k=-\infty}^{\infty} \mathcal{N}(d_n + 2\pi k; \mu_m, \sigma_m^2) \tag{4}$$

ここで α_m は混合重みであり、$\sum_m^M \alpha_m = 1$、$0 \le \alpha_m \le 1$ である。

さらに本稿では、パワー重み a_n を、観測サンプル d_n が観測された頻度を表す量と解釈する。すなわち、観測サンプル d_n が得られた時の信号パワーが a_n であるとき、この観測 d_n が得られる確率密度関数を

$$p(d_n;\mu,\boldsymbol{\sigma^2})^{f(a_n)} \tag{5}$$

と表現するものとする。ここで

$$f(a_n) = a_n / \sum_{n=1}^{N} a_n \tag{6}$$

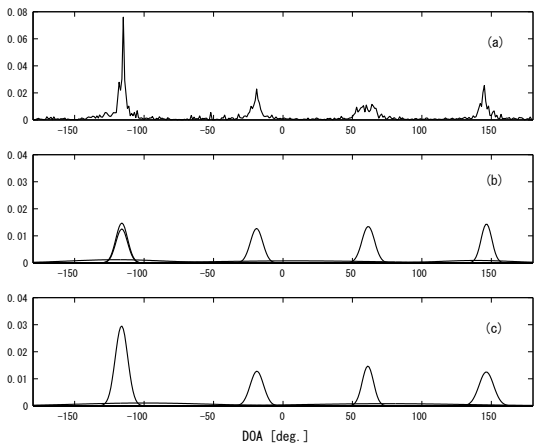

Fig. 1 An example GMM fitting result when $N_s = 4$. (a) DOA histogram, (b) GMM fitting result without prior. Two Gaussians are fit to the cluster around −115 degrees. (c) GMM fitting result with prior. One Gaussian fits to each cluster.

である。これは、各 DOA 観測サンプルをパワーで重み付けして得られる重みつきヒストグラムに GMM をあてはめることに対応している。

3 提案法

この章では、まず 3.1 節で、DOA 分布を GMM でモデル化する時にディリクレ事前分布を用いる必要性とその方法を述べる。次に 3.2・3.3 節において、MAP 推定による DOA 分布への GMM フィッティングの方法を示す。最後に、3.4・3.5 節において、音源数推定と音源分離の方法について述べる。

3.1 混合重みへのディリクレ事前分布の適用

観測データに GMM をあてはめる場合には、充分数 M 個の正規分布を用意し、GMM のパラメタ θ(平均 μ_m、分散 σ_m^2、混合重み α_m) を求める。しかし、DOA 分布にそのまま GMM をあてはめると、1 つのクラスタを複数の正規分布でモデル化してしまう問題が生じる。図 1(b) は、図 1(a) の分布を 8 個の正規分布から成る GMM でモデル化した例である。−115° 付近のクラスタに複数の正規分布があてはまっていることが分かる。しかし本稿では、音源数推定を可能とするため、各クラスタをそれぞれ 1 個の正規分布で表したい。

そのため本稿では、DOA の分布をなるべく少ない正規分布でモデル化するために、分布の混合重み α_m の事前分布として、ディリクレ分布

$$p(\alpha) = \frac{1}{B(\phi)} \prod_{m=1}^{M} \alpha_m^{\phi-1} \tag{7}$$

を用いることを提案する。ここで、$\alpha = \{\alpha_1, \cdots, \alpha_m, \cdots, \alpha_M\}$ であり、$\sum_m^M \alpha_m = 1$、

$0 \leq \alpha_m \leq 1$ である。これは、GMM における混合重みの条件と同じであることに注意されたい。また、$B(\phi)$ はベータ分布である。ここでパラメタ ϕ を 1 より小さい正の値に設定すると、混合重み α_m のごく少数のみが十分に大きな値を持ち、残りは 0 に近い値を取るようになる [9]。この性質を用いることにより、なるべく少数の正規分布でのモデル化が可能となる。またディリクレ分布は、混合重みの共役分布としても広く用いられており [9]、GMM フィッティングへの適用が容易である。

3.2 MAP 推定のためのコスト関数

本節と次節にて、MAP 推定による DOA 分布への GMM フィッティングの方法を示す。モデルパラメタを $\theta = \{\alpha_m, \mu_m, \sigma_m, \cdots\}$、DOA 観測データを $d = \{d_1, d_2, \cdots, d_n, \cdots, d_N\}$、パワー重みを $a = \{a_1, a_2, \cdots, a_n, \cdots, a_N\}$ と書く (2.2 節参照)。正規分布のインデックス m とラップドガウシアンのパラメタ k は隠れ変数として扱う。

まず、MAP 推定のためのコスト関数を、パワー重みつきの平均対数尤度を用いて次のように与える：

$$\mathcal{L}(\theta) = \log p(d,\theta) = \log p(d|\theta) + \log p(\theta) \quad (8)$$

$$= \sum_{n=1}^{N} f(a_n) \log p(d_n|\theta) + \{\log p(\alpha) + \text{const.}\} \quad (9)$$

$$= \sum_{n=1}^{N} f(a_n) \log \left(\sum_{m=1}^{M} \sum_{k=-\infty}^{\infty} p(m,k,d_n|\theta) \right) + \log p(\alpha) + \text{const.} \quad (10)$$

ここで、$-\pi \leq d_n < \pi$ であり、

$$p(m,k,d_n|\theta) = \frac{\alpha_m}{\sqrt{2\pi\sigma_m^2}} \exp\left(\frac{-(d_n + 2\pi k - \mu_m)^2}{2\sigma_m^2} \right), \quad (11)$$

である。また、(10) において、$p(\alpha)$ にはディリクレ分布 (7) を用い、前述した通りそのハイパーパラメタを $\phi < 1$ とする。尚、ここで $\phi = 1$ とすると、事前分布を用いない GMM フィッティングに等しくなる。

3.3 EM アルゴリズム

次に、事前分布を含みながらパラメタ θ の更新を行なうための、EM アルゴリズムを導出する。

Q 関数は、

$$Q(\theta|\theta^t) = \mathrm{E}\left[\log p(d,\theta)|d_n;\theta^t\right] = \sum_n \sum_m \sum_k \left[p(m,k|d_n,\theta^t) f(a_n) \log p(m,k,d_n|\theta)\right] + \log p(\alpha) \quad (12)$$

である。ここで θ^t は、t 回目の反復更新で得られたパラメタを示し、

$$p(m,k|d_n,\theta^t) = \frac{p(m,k,d_n|\theta^t)}{\sum_m \sum_k p(m,k,d_n|\theta^t)} \quad (13)$$

である。パラメタ中の平均 μ_m と分散 σ_m の更新式は、$\frac{\partial Q(\theta|\theta^t)}{\partial \mu_m} = 0$ および $\frac{\partial Q(\theta|\theta^t)}{\partial \sigma_m^2} = 0$ とすることで、以下のように得られる。

$$\mu_m^{t+1} = \frac{\sum_n \sum_k p(m,k|d_n,\theta^t) f(a_n)(d_n + 2\pi k)}{\sum_n \sum_k p(m,k|d_n,\theta^t) f(a_n)} \quad (14)$$

$$(\sigma_m^2)^{t+1} = \frac{\sum_n \sum_k p(m,k|d_n,\theta^t) f(a_n)(d_n + 2\pi k)^2}{\sum_n \sum_k p(m,k|d_n,\theta^t) f(a_n)} - (\mu_m^{t+1})^2. \quad (15)$$

また、混合重み α_m の更新式は、$\frac{\partial Q(\theta|\theta^t)}{\partial \alpha_m} = 0$、$\sum_m^M \alpha_m = 1$ および (6) より、以下のように得られる。

$$\alpha_m^{t+1} = \frac{1}{1 + M(\phi - 1)} \left\{ \sum_n \sum_k p(m,k|d_n,\theta^t) f(a_n) + (\phi - 1) \right\} \quad (16)$$

以上の EM アルゴリズムをまとめる。E-step で (13) を計算し、M-step で (14)、(15)、(16) を用いてパラメタ θ を推定する。我々の実験において、混合重みが $\alpha_m < 0$ となってしまうことがあったので、その場合にはその重みに十分小さい値 ϵ を与え、パラメタ θ を推定した。

3.4 音源数推定

ディリクレ事前分布 (7) の利用により、音源方向に対応する正規分布以外の重み係数 α_m は十分に小さくなっている。よって音源数推定は、上記 EM アルゴリズムで推定された正規分布のうち、ある閾値以上の値を取る重み係数 α_m の個数を数えることで行なう。

実際には、重み係数 α_m の閾値処理だけでは充分ではない場合もあったため、本稿では、重み係数 α_m が十分に大きく分散 σ_m^2 が十分小さい正規分布の個数を数えることで、音源数推定を行なった：

$$\alpha_m \geq \epsilon \quad \& \quad \sigma_m \leq th$$

ここで ϵ は十分に小さな数、th は閾値であり、本稿では $\epsilon = 10^{-10}$ と $th = 20°$ を用いた。

3.5 音源分離

m 番目の信号を分離する時間周波数マスク $M_m(f,\tau)$ ((2) を参照) は、(13) で得られた分布を k で周辺化して以下のとおり推定する。

$$M_m(f,\tau) = p(m|d(f,\tau),\theta) = \sum_{k=-\infty}^{\infty} p(m,k|d(f,\tau),\theta) \quad (17)$$

分離信号は、このマスクを用いて、(2) により

$$y_m(f,\tau) = x_1(f,\tau) p(m|d(f,\tau),\theta) \quad (18)$$

で得る。

4 実験

4.1 実験条件

提案法の有効性を調べるため、実験を行なった。混合信号は、5 秒間の英語音声と、文献 [5] の図 4 の環

境 (残響時間 130 ms) で計測したインパルス応答とを畳み込んで作成した。実験では 3 つのマイクを用い、一辺 4 cm の正三角形の頂点にマイクを配置した。サンプリング周波数は 8 kHz、STFT のフレームサイズは 512 (64 ms)、フレームシフトは 128 (16 ms) とした。

EM アルゴリズムでは、提案法 ((7) のハイパーパラメタが $\phi = 0.98$)、従来法 ($\phi = 1.0$) とも $M = 8$ 個の正規分布を用いた。パラメタ θ の初期値としては、μ_m=[25°, 75°, 115°, 160°, 205°, 250°, 295°, 340°]、$\sigma_m = 40°$、$\alpha_m = 1/M = 0.125 (m = 1, \cdots, 8)$ を用いた。ラップドガウシアンとしては $-1 \leq k \leq 1$ について和をとった。更新回数は 40 とした。

性能評価量としては、分離性能を示す信号対妨害音比 (SIR) と、音質を示す信号対歪み比 (SDR) を用いた [10]。性能評価の際には、3.4 節の方法で音源として数えあげられた正規分布にて分離した分離音声について SIR と SDR を評価し、音声組合せや音源位置の異なる 20 通りの混合について平均値を計算した。

4.2 実験結果

図 1 (a) の DOA 分布に対し、ディリクレ事前分布を用いた ($\phi = 0.98$) GMM フィッティング結果を図 1 (c) に示す。事前分布無しでは $-115°$ 方向に複数の正規分布があてはまっていたのに対し (図 1(b))、事前分布を用いることで、各音源クラスタに各 1 つの正規分布をあてはめることができ、全体で 4 個 (=音源数 $= N_s$) の正規分布で DOA の分布を表現することができていることがわかる。

表 1 は、音源数推定精度と音源分離性能を示す。表において、$\phi = 0.98$、$\phi = 1.0$ はそれぞれ、ディリクレ事前分布有りの場合と無しの場合の結果を示している。音源数推定精度は、それぞれ 20 通りの音源や音源位置組合せにおいて正しく音源数を推定した割合を%で示している。また、音源数が既知の場合に、k-means 法を用いて [5] 音源分離を行なった時の音源分離性能についても示している。

表 1 の結果より、ディリクレ事前分布を用いる ($\phi = 0.98$) ことで、音源数をかなり正確に推定できることがわかる。一方で、ディリクレ事前分布を用いない場合 ($\phi = 1.0$) には、音源数は過大評価され、正しい音源数推定はできなかった。

また表 1 より、音源分離性能についても、ディリクレ事前分布を使う ($\phi = 0.98$) 方が、使わない場合 ($\phi = 1.0$) に比べて優れていることが分かった。

5 おわりに

本稿では、DOA をクラスタリングする時に、GMM の混合重みにディリクレ分布による事前分布を用いることで、1 クラスタに 1 つの正規分布をあてはめることができ、その結果、音源数推定と音源分離を同時に実現できることを示した。

Table 1 Experimental results. $\phi = 0.98$: with prior; $\phi = 1.0$: without prior, K: with the k-means. InputSIR was 0.0 [dB] ($N_s = 2$), -3.1 [dB] ($N_s = 3$), and -4.9 [dB] ($N_s = 4$).

N_s	ϕ	Accuracy of $\hat{N}_s$ estimation [%]								Performance	
		$\hat{N}_s$:1	2	3	4	5	6	7	8	SIR	SDR
2	0.98		100							19.6	11.3
	1.0		0		10	25	35	20	10	11.6	4.2
	K	given								15.9	14.5
3	0.98		5	95						16.0	8.9
	1.0			35	15	30	20			14.5	7.5
	K	given								12.2	11.4
4	0.98				100					13.7	7.8
	1.0				20	75	5			12.8	6.7
	K	given								10.7	9.7

参考文献

[1] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[2] Ö. Yılmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[3] M. Mandel, D. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," in *Proc. Neural Info. Proc. Sys.* , 2006.

[4] P. O'Grady and B.Pearlmutter, "Soft-LOST: EM on a mixture of oriented lines," in *Proc. ICA 2004 (LNCS 3195)*, Sept. 2004, pp. 430–436.

[5] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 77, no. 8, pp. 1833–1847, aug 2007.

[6] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA estimation for multiple sparse sources with normalized observation vector clustering," in *Proc. ICASSP'06*, May 2006, vol. 5, pp. 33–36.

[7] S. Araki, T. Nakatani, H. Sawada, and S. Makino, "Stereo source separation and source counting with MAP estimation with Dirichlet prior considering spatial aliasing problem," in *Proc. ICA 2009 (LNCS 5441)*, Mar. 2009, pp. 742–750.

[8] P. Smaragdis and P. Boufounos, "Learning source trajectories using wrapped-phase hidden Markov models," in *Proc. of WASPAA'05*, oct 2005, pp. 114–117.

[9] C. M. Bishop, *Pattern recognition and machine learning*, Springer, 2008.

[10] S. Araki, T. Nakatani, H. Sawada, and S. Makino, "Blind sparse source separation for unknown number of sources using Gaussian mixture model fitting with Dirichlet prior," in *Proc. ICASSP'09*, Apr. 2009, pp. 33–36.

[11] K. V. Mardia, *Statistics of directional data*, Academic Press, 1972.