EM アルゴリズムを用いた音声スパース性に基づく 2ch BSS *

◎和泉 洋介, 小野 順貴, 嵯峨山 茂樹 (東大情報理工)

1 はじめに

音声のスパース性を利用したブラインド音源分離は，観測信号数が音源信号数よりも小さい場合でも適用でき，特にマイクロホン数が少ない場合に有効な分離手法である．3 つ以上の音源信号を 2ch の観測から分離する問題に対し，従来では音源定位とその結果に基づく時間周波数マスクによる分離手法が用いられるが，雑音の影響で特徴量のクラスタリングが困難になることが課題であった．本稿では，この問題に対し EM アルゴリズムを適用することにより，雑音が加わった場合でも効率的に最尤法に基づく音源定位を行い，時間周波数マスクを設計する手法を提案する．最後にシミュレーションにより従来法との比較実験を行った結果を報告する．

2 スパース性に基づく 2chBSS の問題設定

信号はすべて時間周波数領域で扱い，n 番目の音源信号を $S_n(\tau,\omega)$，左右のマイクロフォンから得る観測信号を $M_L(\tau,\omega), M_R(\tau,\omega)$ とし，簡単のため以後 (τ,ω) は省略して表記する．音声のスパース性から，各時間周波数成分において観測信号に寄与する音源は 1 つだけであると仮定すると観測モデルは，

$$\begin{pmatrix} M_L \\ M_R \end{pmatrix} = S_i \begin{pmatrix} 1 \\ e^{j\omega\delta_n} \end{pmatrix} + \begin{pmatrix} N_L \\ N_R \end{pmatrix} \qquad (1)$$

と書ける．ただし，N_L, N_R は 2 つの観測信号それぞれについて生じる誤差項，δ_n は n 番目の音源に対応する時間差である．

各時間周波数成分について左右の観測信号の比 M_L/M_R から推定できる時間差 δ の散布図の一例を示すと Fig. 1 のようになり，各音源の真の δ_n を中心に分散して分布する．したがってこれをクラスタリングすることで，それぞれの音源が寄与する時間周波数成分を抜き出すことが可能になる．そこで多くの先行研究 [1–3] では，観測信号から抽出した時間差のクラスタリング，次に同じクラスタに属する時間周波数成分だけをマスキングにより抜き出して分離，という別個の 2 つのステップによって分離を行っていた．

3 本研究の着眼点

真の δ_n を推定する合理的な手法の 1 つは最尤法である．いま，ある音源信号 S_n が寄与している時間周波数成分の集合 Ω_n が既知の場合，この音源に対応する時間差 δ_n は，対数尤度

$$J_n = \sum_{(\tau,\omega)\in\Omega} \log p(\boldsymbol{M}|\ \delta_n) \qquad (2)$$

を最大化することによって求められる．ただし $\boldsymbol{M} = (M_L, M_R)^T$ とした．

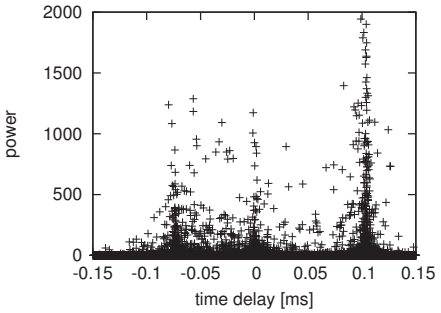

Fig. 1 時間差の散布図 (音源数 3)

ここで $p(\boldsymbol{M}|\ \delta_n)$ は，時間差 δ_n を生じるような方向に音源が存在する場合に $\boldsymbol{M}$ という信号が観測される確率であり，以後，これを方向尤度と呼ぶことにする．しかしながら，Ω_n を求めることは音源分離そのものであり，通常，直接この最尤法を実行することはできない．

我々はこの問題を，各時間周波数成分を個々の音源に帰属させるクラスタリングの問題ととらえ，混合ガウス分布の平均・分散推定問題との類似点に着目した．いずれの問題も，各データが帰属するクラスが未知の状態で，各クラスのパラメータを推定する問題であり，データと観測信号，各ガウス分布と各音源の方向尤度，ガウス分布の平均・分散と音源方向や誤差分散をそれぞれ対応させることで，EM アルゴリズムという効率的な解法を適用することができる．通常のクラスタリング手法では，データが各クラスタに属するかどうかを 0,1 で決定することが多い．EM アルゴリズムを適用すると帰属率を連続値の確率として扱い，分布同士が重なるような場合でもロバストに推定できる．このような枠組を導入することで，残響や背景雑音が存在し，時間差のクラスタリングが困難な場合に対して分離性能を向上させることが，本研究の目的の 1 つである．

4 2ch BSS への EM アルゴリズムの適用

EM アルゴリズムとは，隠れ変数 (観測できないデータ) が観測モデルに含まれている場合に，隠れ変数の期待値を求める E ステップと，対数尤度の条件付き期待値 (Q 関数) が最大となるようにパラメータを最適化する M ステップを反復することによって，モデルパラメータを局所最適解へと更新するアルゴリズムである．

本研究では，各時間周波数成分に寄与している音源のインデックスを隠れ変数として扱う．式 (1) の誤差 N_L, N_R に分散 σ^2 の独立な正規分布を仮定すると方向尤度の具体的な形は，

$$\begin{aligned} & p(\boldsymbol{M}|\ \delta_k) \\ = \ & -\log(2\pi) - \frac{1}{2}\log\sigma^2 - \frac{1}{4\sigma^2}|M_L - e^{-j\omega\delta_k} M_R|^2 \end{aligned}$$

---

*2ch BSS based on speech sparseness with EM algorithm by IZUMI Yosuke, ONO Nobutaka and SAGAYAMA Shigeki (Graduate School of Information Science and Technology, the Univ. of Tokyo)

のように求められる．ただし，スパース性の仮定から S_n は独立な変数ではなく EM アルゴリズムとは別に求める必要があるが，ひとつの手段として S_n に最尤値を用いた．

本稿での問題設定においてモデルパラメータの集合を $\Theta = \{\sigma, \delta_1, \delta_2, \dots\}$ とすると，EM アルゴリズムの t 回目の反復は，

- E ステップ：各時間周波数成分に対して期待値 $m^{(t)}_{\tau,\omega,k}$ を計算
- M ステップ：Q 関数を最大化するパラメータ $\Theta^{(t+1)}$ を求める

と書ける．ただし，

$$m^{(t)}_{\tau,\omega,k} \quad = \quad \frac{p(\boldsymbol{M}|\ \delta_k)}{\sum_{k'} p(\boldsymbol{M}|\ \delta_{k'})}$$

と定義され，これは (τ, ω) の成分が k 番目の音源に帰属する確率を表す．Q 関数の具体的な形は，

$$Q(\Theta|\Theta^{(t)}) = \sum_{\tau,\omega,k} m_{\tau,\omega,k} \log p(\boldsymbol{M}|\ \delta_k)$$

である．M ステップにおける σ の更新式は以下のように解析的に求まる．

$$\left(\sigma^2\right)^{(t+1)} = \frac{1}{2C} \sum_{\tau,\omega,k} m^{(t)}_{\tau,\omega,k} |M_L - e^{-j\omega\delta_k} M_R|^2$$

ただし C は全時間周波数成分の個数である．δ については更新式を解析的に求めることができないので，数値的に全探索を行って Q 関数が最大となる δ を採用することで更新した．

EM アルゴリズムによる分離手法を従来法と比較すると，E ステップでは連続値マスクによる音源分離を，M ステップが音源定位を行っていると捉えることができる．従来は別個の 2 つの処理によって分離を行っていたが，提案手法では，共通の目的関数である尤度を反復推定により最大化させることでこれらを統合した分離を行っている．

5 残響シミュレーション実験

EM アルゴリズムによる音源分離実験を，図 2 のように 3 つの音源および 2 つのマイクロフォンを配置し，鏡像法 [5] による残響シミュレーションを行った．分離性能の評価には分離の前後での元音声に対する S/N 比の改善値を用いた．音声データは研究用連続音声データベース (©板橋秀一 [日本音響学会／編]1991Vol. 1-3) を使用した．また，サンプリング周期 16kHz，フレーム長は 2^{10} 点, シフトは 2^9 点，窓関数を Hamming 窓として，観測信号を短時間 Fourier 変換して時間周波数表現を得た．比較対象とした従来法は Yilmaz らの手法に基づき，時間差のヒストグラムのピークを検出することで音源定位を行い，次にこれに基づいて最尤推定で音源分離マスクを設計するものである．残響時間 376ms の場合の音源分離・定位結果を Table 1 ,2 に, また，残響時間を変えたときの σ^2 の推定値を Table 3 に示す．これらの結果から，提案法は従来法よりも高い分離性能をもっていることが確認できる．さらに残響時間が増加するにしたがい雑音の分散 σ^2 の推定値も増加し，雑音環境の推定も同時に行えていることがわかる．

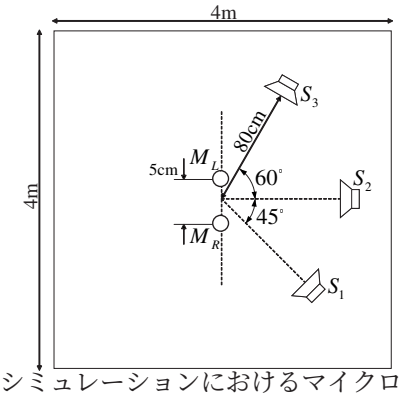

Fig. 2　シミュレーションにおけるマイクロフォンと音源の位置関係

Table 1　音源定位結果の比較 (時間差 [μs])

手法	s_1	s_2	s_3
従来手法	-103	14	22
提案手法	-118	0	108
真の位置	-127	0	104

Table 2　音源分離性能結果の比較 ([dB])

手法	s_1	s_2	s_3
従来手法	4.8	2.4	4.8
提案手法	5.9	4.1	7.1

Table 3　σ^2 の推定値と残響時間の関係

残響時間 [ms]	0	90	170	270	370
σ^2	0.12	0.14	0.17	0.21	0.25

6 結論

本稿では，スパース性を利用した 2chBSS に対して，EM アルゴリズムを適用することにより，最尤の音源方向と時間周波数マスクを設計し，その有効性を確認した．今後は，観測信号間の強度比や，雑音の物理特性を考慮した雑音の分布 [4] を導入し，分離性能の向上をはかるつもりである．

謝辞　本研究の一部は科学研究費補助金・若手研究 (B)(課題番号 18760303) の補助を受けて行なわれたので，ここに謝意を表する。

参考文献

[1] O. Yilmaz *et al.*, IEEE Trans. on Signal Processing, Vol. 52, No. 7, pp 1830-1847, 2004.
[2] S. Araki *et al.*, IWAENC2005, pp. 117-120, 2005.
[3] H. Sawada *et al.*, IEEE Trans. Audio, Speech and Language Processing, vol. 14, no. 6, pp. 2165-2173, 2006.
[4] 小野他, 音講論 (秋), 1-1-10 in CD-ROM, 2006.
[5] J. B. Allen et al., JASA, vol. 65, no. 4, pp. 943–950, Apr. 1979.