EM アルゴリズムを用いた最尤時間周波数マスキングによる雑音環境下での 2ch BSS

2ch BSS under Noisy Environments by ML Time-Frequency Masking with EM Algorithm

小野 順貴 [1]　和泉 洋介 [1]　亀岡 弘和 [1]　嵯峨山 茂樹 [1]
Nobutaka Ono　Yosuke Izumi　Hirokazu Kameoka　Shigeki Sagayama

東京大学 大学院情報理工学系研究科 [1]
Graduate School of Information Science and Technology, The University of Tokyo

1　はじめに

複数の音源が存在する環境で目的音源信号のみを分離するための手法として，近年，独立成分分析をはじめとしたブラインド音源分離 (BSS) の手法が活発に進められ [1]，特に最近では，残響や背景音等が存在する実環境での適用が大きな関心の 1 つとなっている [2][3][4]。本稿では，目的信号として音声を想定し，未知の雑音環境下に対する BSS 手法として，1) 信号ドメインでのスパースな観測モデルと，2)EM アルゴリズムによる最尤時間周波数マスキング，に基づき，雑音レベル等も最尤推定の枠組みで同時推定しながら音源分離を行なう新たな手法を提案する。

2　従来の時間周波数マスキングによる音源分離

時間周波数マスキングとは，音声などの音源信号のエネルギー分布が時間周波数領域で疎らで互いの重なりが少ないことを前提に，目的音源成分のみを通過させ，それ以外を阻止する時間周波数領域でのマスキング処理により目的音源を分離する手法であり，マイクロフォン数より多くの音源を扱うことができる特長がある。従来の時間周波数マスクの設計手法としては，観測信号間の強度比，時間差，正規化ベクトルなどのクラスタリングが用いられてきた [5][6]。しかしながら，残響や背景音が存在する環境下では，時間差等の特徴量は大きなばらつきを含み，互いの分布が重なり合い，クラスタリングが困難になる問題が生じる (図 1)。

3　スパース信号の観測モデル

本研究の着眼点の 1 つは，時間差等の特徴量のばらつきを信号ドメインでモデル化することである。いま，複数存在する音源信号がスパースであり，時間周波数成分 (τ,ω) に寄与する音源は 1 つであると仮定すれば，観測モデルは，

$$\boldsymbol{M}(\tau,\omega) = S_k(\tau,\omega)\boldsymbol{b}_k(\omega) + \boldsymbol{N}(\tau,\omega) \qquad (1)$$

と表せる。ただし，$\boldsymbol{M}(\tau,\omega) = (M_L(\tau,\omega), M_R(\tau,\omega))^t$ は 2ch の観測信号，k は (τ,ω) 成分に寄与する音源番号を表すインデックス，$S_k(\tau,\omega)$ は音源信号，$\boldsymbol{b}_k(\omega) = (1, e^{j\omega\delta_k})^t$ はステアリングベクトル，δ_k は 2ch 間の時間差，$\boldsymbol{N}(\tau,\omega) = (N_L(\tau,\omega), N_R(\tau,\omega))^t$ は，残響，背景音，モデル化誤差を含む雑音項を表す。本研究では以下の定式化を容易にするため，$N_L(\tau,\omega), N_R(\tau,\omega)$ は $S_k(\tau,\omega)$ とは独立なガウス雑音であると仮定する。

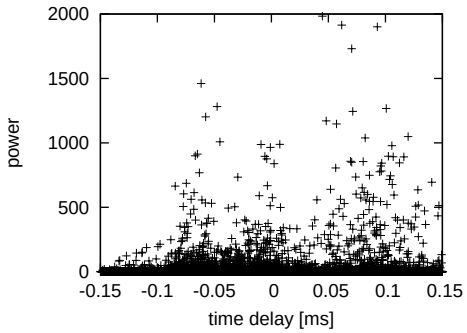

図 1　時間差の散布図 (音源数 3，残響時間 170ms)

音源信号 $S_k(\tau,\omega)$ を観測信号から得られる最尤値

$$S_k(\tau,\omega) = \frac{\boldsymbol{b}_k(\omega)^h V(\omega)^{-1}\boldsymbol{M}(\tau,\omega)}{\boldsymbol{b}_k(\omega)^h V(\omega)^{-1}\boldsymbol{b}_k(\omega)} \qquad (2)$$

でおきかえると，時間差 δ_k に対応する音源が寄与する時間周波数 (τ,ω) で $\boldsymbol{M}(\tau,\omega)$ が観測される対数尤度のモデルとして

$$\begin{aligned}\log p(\boldsymbol{M}(\tau,\omega)|\ \delta_k) \ =\ & -\log(2\pi) - \frac{1}{2}\log|V(\omega)| \\ & -\frac{1}{2}(\boldsymbol{M}(\tau,\omega)^h V(\omega)^{-1}\boldsymbol{M}(\tau,\omega)) \\ & +\frac{1}{2}\frac{|\boldsymbol{b}_k(\omega)^h V(\omega)^{-1}\boldsymbol{M}(\tau,\omega)|^2}{\boldsymbol{b}_k(\omega)^h V(\omega)^{-1}\boldsymbol{b}_k(\omega)}\end{aligned} \qquad (3)$$

を得る。ここで $V(\omega)$ は雑音の共分散行列を表す。

従来のスパース性に基づく BSS の枠組みでは，特徴量（ここでは時間差 δ_k）のドメインで直接そのばらつきがモデル化されることが多かったが，このような観測信号領域でのモデル化は，拡散音場モデル [7] 等の物理モデルの導入を可能にし，また後述のように，共分散行列自体を観測データから推定することにより，周囲環境への自動的な適応を目的としている。

4　EM アルゴリズムによる最尤時間周波数マスキング

観測データからモデルパラメータを決定するための合理的な手法の 1 つは最尤推定である。いま，複数の音源が存在し，各音源の方向に対応して 2 個のマイクロフォン間に生じる時間差が $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_N)^h$ であるとき，$\boldsymbol{M}(\tau,\omega)$ が観測される尤度を $p(\boldsymbol{M}(\tau,\omega)|\ \boldsymbol{\delta})$ と表す。各時間周波数成分に寄与する音源が 1 個であるというス

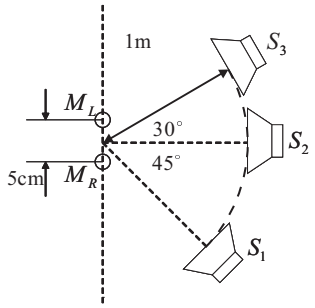

図 2　シミュレーションにおけるセットアップ

パースな観測モデルの下での $\boldsymbol{\delta}$ の最尤推定は,

$$J = \sum_{(\tau,\omega)} \log p(\boldsymbol{M}(\tau,\omega)|\ \boldsymbol{\delta}) = \sum_{k} p(\boldsymbol{M}(\tau,\omega)|\ \delta_{k(\tau,\omega)}) p(k(\tau,\omega)) \quad (4)$$

を最大化する $\boldsymbol{\delta}$ を求めることによって行なわれる。ここで $k(\tau,\omega)$ は，(τ,ω) 成分に寄与する音源のインデックスであり，実際には観測することができない。

本研究の 2 つ目の着眼点は，隠れ変数 $k(\tau,\omega)$ を含んだこの尤度最大化問題は混合ガウスモデル (GMM) と同型であり，EM アルゴリズムにより効率的に求めることができるという点にある。具体的な導出は省略するが，$V = \sigma^2 I$(I は単位行列) という共分散モデルの下では，パラメータの更新式は以下のように得られる。

$$m_{\tau,\omega,k}^{(t+1)} = \frac{p(\boldsymbol{M}(\tau,\omega)|\ \delta_k^{(t)})}{\sum_k p(\boldsymbol{M}(\tau,\omega)|\ \delta_k^{(t)})} \quad (5)$$

$$\delta_k^{(t+1)} = \arg\max_{\delta_k} \sum_{\tau,\omega} m_{\tau,\omega,k}^{(t)} \left| M_L - e^{-j\omega\delta_k} M_R \right|^2 \quad (6)$$

$$\left(\sigma^2\right)^{(t+1)} = \frac{1}{2C} \sum_{\tau,\omega,k} m_{\tau,\omega,k}^{(t+1)} |M_L - e^{-j\omega\delta_k^{(t+1)}} M_R|^2 \quad (7)$$

ただし，t は反復の回数，C は全時間周波数成分の個数であり，また式中では M_L，M_R の引数 (τ,ω) は省略して表記している。$m_{\tau,\omega,k}$ は，観測信号 $\boldsymbol{M}(\tau,\omega)$ の音源 k への帰属度を表し，これが音源分離における連続値マスクの役割を果たすため，更新式はそれぞれ，音源分離，音源定位，雑音レベル推定に相当し，これらの反復により最尤解が得られる枠組みになっている。

5　シミュレーション実験による検証

図 2 のような 3 個の音源，2 個のマイクロフォン配置を想定し，球面波伝播と残響を鏡像法 [8] によりシミュレーションし，提案法の基本的な分離性能を検証した。分離性能の評価には，分離前後での原音声に対する S/N 比改善値を用いた。音声データは日本音響学会編集の研究用連続音声データベースを使用した。また，Yilmaz ら [5] の議論をもとに，実験条件は，サンプリング周期 16kHz，フレーム長 1024 点，フレームシフト 512 点，窓関数 Hamming 窓のように定め，短時間 Fourier 変換により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 [5] である。

表 1　従来／提案手法の音源定位結果 (時間差 [μs])

条件	手法	s_1	s_2	s_3
残響時間 0ms	従来手法	10.3	0.0	-6.7
	提案手法	9.8	0.0	-6.7
残響時間 370ms	従来手法	1.0	-4.2	-5.1
	提案手法	10.3	0.0	-8.8
	真値	10.4	0.0	-7.3

表 2　従来／提案手法の分離性能 (S/N 比改善値 [dB])

条件	手法	s_1	s_2	s_3
残響時間 0ms	従来手法	13.9	10.9	9.4
	提案手法	16.3	13.0	11.5
残響時間 370ms	従来手法	4.9	-8.3	2.9
	提案手法	7.8	3.9	8.1

表 3　雑音パワー (σ^2) の推定値と残響時間の比較

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

音源定位結果を表 1，分離結果を表 2 に示す。残響環境下においては従来手法の場合，残響による時間差のばらつきのためにクラスタリングがうまく働かず，分離性能は低い値に留まっているが，提案手法では音源定位，音源分離ともに，よい性能が確認できる。また，異なる残響環境における σ^2 の推定値と残響時間との関係を表 3 に示す。残響時間が長くなるにつれ σ^2 の推定値が大きくなっており，環境に応じて観測誤差の大きさを推定できていることがわかる。本稿では，雑音レベルの推定のみを示したが，同じ枠組みで周波数毎の雑音共分散行列自体の推定も可能である。これを利用した今後の課題の 1 つとして，最終段に最小分散ビームフォーマ等を組み合わせる手法を検討中である。

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] A. Hyv̈arinen et al., “Independent Component Analysis,” Wiley, 2001.

[2] H. Sawada et al, IEEE Trans, ASLP, vol. 14, no. 6, pp. 2165-2173, Nov. 2006.

[3] Y. Takahashi et al., Proc. IWAENC, Sep, 2006.

[4] Y. Izumi et al., JASA, vol. 120, p. 3047, 2006.

[5] O. Yilmaz et al., IEEE Trans. SP, vol. 52, no. 7, pp 1830-1847, 2004.

[6] S. Araki et al., Proc. IWAENC, pp. 117-120, Sep. 2005.

[7] R. K. Cook et al., JASA, vol. 27, no. 6, pp. 1072–1077, Nov. 1955.

[8] J. B. Allen et al., JASA, vol. 65, no. 4, pp. 943–950, Apr. 1979.