背景雑音抑圧を含めた時間周波数マスキングに基づく 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

複数の音声信号が混合した観測信号だけから個々の信号を分離するブラインド音源分離 (BSS) は様々なアプリケーションに応用できる技術として活発に研究されている [1]。特に劣決定の BSS 問題に対しては、対象とする音声信号のスパース性に基づく時間周波数マスキングが有効であることが知られている [4,5] が、目的音と同時に観測される拡散性の背景雑音は分離信号に含まれたままであった。しかし、実環境では多様な誤差要因により背景雑音が生じるので、方向性の音源信号の分離だけでなく拡散性の雑音も抑圧する分離手法が望ましく、ブラインド信号処理の枠組みにおいてもこうした試みが行われつつある。[6,8,9]。このような背景のもとに本稿では、劣決定の 2ch BSS 問題に対して我々がこれまで提案してきた最尤マスク設計法が背景雑音のパワースペクトルも同時に推定できることに着目し、時間周波数マスクによる分離と背景雑音の抑圧を行い分離信号を得る手法を提案する。

2 時間周波数マスキングによる 2ch BSS の問題設定

音声信号のエネルギーは時間周波数空間の疎な領域に集中しており (スパース性)、同時に K 個の音声を観測しても各時間周波数成分において個々の音声の成分が重なることはほとんど起こらない。したがって観測信号 $\boldsymbol{M} = (M_L(\tau,\omega), M_R(\tau,\omega))^T$ の各成分には $i(\tau,\omega)$ 番目の音声信号 $S_i(\tau,\omega)$ だけが寄与すると仮定できる。時間フレームと周波数のインデックス (τ,ω) を省いて表記を簡略化すると、観測モデルは、

$$\boldsymbol{M} = S_i \begin{pmatrix} 1 \\ e^{j\omega\delta_i} \end{pmatrix} + \boldsymbol{N} \qquad (1)$$

と表せる。ただし，δ_i は i 番目の音源に対応する時間差を、$\boldsymbol{N} = (N_L, N_R)^T$ は背景雑音を含む誤差項を表し、平均 $\boldsymbol{0}$、共分散行列 $V(\omega)$ の正規分布に従うとする。ただし、本稿ではあらゆる方向から確率的に等しく平面波が到来する拡散音場モデル [2] に基づいて以下のように $V(\omega)$ を仮定した。

$$V(\omega) = \sigma_N^2(\omega) \begin{pmatrix} 1 & \mathrm{sinc}(\omega D/c) \\ \mathrm{sinc}(\omega D/c) & 1 \end{pmatrix} \qquad (2)$$

ここで D はマイクロフォン間の距離、c は音速を表す。

3 2ch BSS への EM アルゴリズムの適用

時間差が δ_i となる方向から $i(\tau,\omega)$ 番目の音声信号 S_i が到来して $\boldsymbol{M}$ が観測される尤度は、

$$p\left(\boldsymbol{M}|\ \sigma_N^2(\omega), \delta_i, S_i\right) = \frac{1}{\pi\sqrt{|V|}} \exp\left(-\left(\boldsymbol{M} - \boldsymbol{b}_i S_i\right)^h V^{-1} \left(\boldsymbol{M} - \boldsymbol{b}_i S_i\right)\right)$$

と与えられる。ここで、各時間周波数成分に寄与する音源のインデックス i は観測できない隠れ変数であることに注意すると、EM アルゴリズムを適用することで効率的に未知パラメータ $\boldsymbol{\Theta} = \{\{S_i\}, \{\delta_i\}, \{\sigma_N(\omega)\}\}$ の最尤値を導出できる [3]。

具体的な更新として、t 回目の E ステップでは次式で表される Q 関数を計算する。

$$Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(t)}) = \sum_{\tau,\omega,i} m_{\tau,\omega,i}^{(t)} \log r_i\ p(\boldsymbol{M}|\ \sigma_N^2(\omega), \delta_i, S_i)$$

$$m_{\tau,\omega,i}^{(t)} = \frac{r_i^{(t)} p\left(\boldsymbol{M}|\ (\sigma_N^2(\omega))^{(t)}, \delta_i^{(t)}, S_i^{(t)}\right)}{\sum_{i'}^{(t)} r_{i'}^{(t)} p\left(\boldsymbol{M}|\ (\sigma_N^2(\omega))^{(t)}, \delta_{i'}^{(t)} S_{i'}^{(t)}\right)}$$

ここで r_i は i 番目の音源が active になる事前確率に相当する変数で $\sum_k r_k = 1$ を満たす。次に t 回目の M ステップの各パラメータの更新式は、

$$S_i^{(t+1)} = \left((\boldsymbol{b}_i^h)^{(t)} (V^{-1})^{(t)} \boldsymbol{M}\right) \Big/ \left((\boldsymbol{b}_i^h)^{(t)} (V^{-1})^{(t)} \boldsymbol{b}_i^{(t)}\right)$$

$$r_i^{(t+1)} = \sum_{\tau,\omega} m_{\tau,\omega,i}^{(t)} \Big/ \sum_{\tau,\omega,i'} m_{\tau,\omega,i'}^{(t)}$$

$$\left(\sigma_N^2(\omega)\right)^{(t+1)} = \frac{1}{C} \sum_{\tau,i} \frac{m_{\tau,\omega,i} \left(\boldsymbol{M} - \boldsymbol{b}_i^{(t)} S_i^{(t)}\right) (V^{-1})^{(t)} \left(\boldsymbol{M} - \boldsymbol{b}_i^{(t)} S_i^{(t)}\right)}{1 - \mathrm{sinc}^2(\omega D/c)}$$

と与えられる。C は全フレーム数を表す。ただし δ_i の更新は解析解が求まらないので適当に離散化し Q 関数を全探索することで更新した。各パラメータの推定値に基づき分離信号 $\hat{S}_i$ は、以下のようにソフトマスキングによる期待値として求められる。

$$\hat{S}_i = \mathrm{E}\left[S_i\right] = m_{\tau,\omega,i} \frac{\boldsymbol{b}_i^h V^{-1} \boldsymbol{M}}{\boldsymbol{b}_i^h V^{-1} \boldsymbol{b}_i}$$

4 後処理による背景雑音抑圧

前節のアルゴリズムはマスキング処理により方向性雑音は除去できるが拡散性の雑音成分 $\boldsymbol{N}$ は抑圧していない。ここで我々は、時間周波数マスクと同時に推定している雑音パワースペクトル $\sigma_N^2(\omega)$ に着目した。一般的な雑音抑圧手法は雑音成分のパワーなどの情報が既知であるか無音区間から推定する必要がある。しかし我々の手法では無音区間を利用するこ

* 2ch BSS based on time-frequency masking with background noise suppression by IZUMI Yosuke, ONO Nobutaka and SAGAYAMA Shigeki (Graduate School of Information Science and Technology, the Univ. of Tokyo)

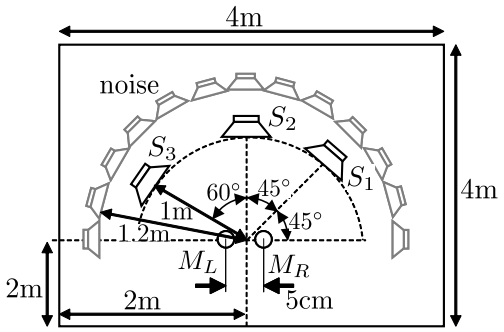

Fig. 1　マイクロフォンと音源の位置関係

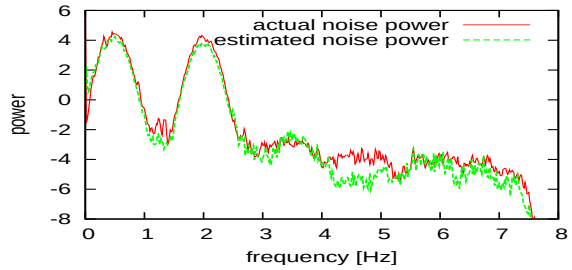

Fig. 2　雑音パワースペクトルの真値と推定値の比較

となく、むしろ目的音が寄与する観測成分の理想的な到来方向からの偏差からそのパワーを推定しており、我々の最尤マスク設計法を雑音抑圧に直結できる利点となっている。

拡散性雑音の抑圧の方法はいくつか考えられるが、本稿では基礎的手法としてスペクトルサブトラクション (SS) およびウィーナフィルタ (WF) を後処理として分離信号に施し、次節で述べるようにその性能改善を比較した。

SS 法を施して得られる音源信号の推定値 $\hat{S}_i^{SS}$ のパワーは

$$|\hat{S}_i^{SS}|^2 = |\hat{S}_i|^2 - \sigma_N^2(\omega) \tag{3}$$

と与えられる。ただし右辺が負になる場合は 0 で置き換えた。これに $\hat{S}_i$ の位相を与えて $\hat{S}_i^{SS}$ を得た。

また、WF 法については以下のようなフィルタ処理を分離信号に施して雑音を抑圧した分離信号 $\hat{S}_i^{WF}$ を得た。

$$\hat{S}_i^{WF} = \frac{\sigma_S^2(\omega)}{\sigma_S^2(\omega) + \sigma_N^2(\omega)} \hat{S}_i \tag{4}$$

ここで音源信号のパワー $\sigma_S^2(\omega)$ はパワースペクトルの加法性を仮定し、式 (3) より近似的に求めたものを用いた。

5　シミュレーション実験

提案法によるシミュレーション音源分離実験を行った。図 1 のように 3 つの音源を取り囲む雑音源を配置して、雑音源には 500Hz と 2000Hz を中心周波数とするバンドノイズを使用し、拡散音場を模するためすべて同じパワーとした。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、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．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。

Table 1　音源定位結果の比較 (時間差 [μs])

手法	S_1	S_2	S_3
DUET	−99	−18	4
提案法	−104	0	129
真の位置	−104	0	127

Table 2　音源分離性能結果の比較 ([dB])

手法	S_1	S_2	S_3
DUET	3.2	−0.7	6.8
SS・WF 無し	3.6	7.6	12.2
提案法 (SS)	8.5	7.5	14.6
提案法 (WF)	10.7	6.8	13.2

雑音のパワースペクトル推定結果を Fig. 2 に、音源定位・分離結果を表 1,2 に示す。時間周波数マスキングの代表的な手法である DUET [10] と比較し、また、SS 法・WF 法を施さずに背景雑音を抑圧しない場合の分離結果も並記した。

これらの結果から従来法に比べて提案法は雑音環境下でも音源定位の精度が良く、また、背景雑音を抑圧した効果があることがわかる。

6　結論

本稿ではスパース性に基づく雑音環境下の 2ch BSS 問題に対し時間周波数マスクだけではなく雑音パワースペクトルも同時に推定することで、方向性雑音からの分離に加えて拡散性雑音も抑圧する手法を提案した。シミュレーション実験において、雑音抑圧の効果を確認し、従来法より優れた分離性能を示すことを確認した。

謝辞　本研究を進めるにあたり貴重な意見をいただいた NTT の荒木章子氏に謝意を表する。

参考文献

[1] J-F Cardoso, Proc. of the IEEE, vol. 90, no. 8, pp. 2009-2026, 1998
[2] R. K. Cook *et al.*, JASA, vol. 27, no. 6, pp. 1072-1077, 1955.
[3] 小野他, 電子情報通信学会技術研究報告 (応用音響), vol. 107, no. 240, pp. 25-30, 2007.
[4] S. Araki *et al.*, IWAENC, pp. 117-120, 2005.
[5] H. Sawada *et al.*, IEEE Trans. Audio, Speech and Language Proc., vol. 15, no. 5, pp. 1592-1604, 2007.
[6] Y. Takahashi *et al.*, Proc. HSMCA, pp. 164-167, 2008
[7] 小野他, 音講論 (秋), 1-1-10 in CD-ROM, 2006.
[8] H. Attias, Neural Comp., vol. 11, no. 4, pp. 803-851, 1999.
[9] E. Moulines *et al.*, Proc. ICASSP, pp. 3617-3620, 1997.
[10] O. Yilmaz *et al.*, IEEE Trans. Signal Proc., vol. 52, no. 7, pp. 1830-1847, 2004.