1—P—5

時間周波数マスキングとICAの併用による音源数>マイク数の場合のブラインド音源分離 *

○荒木 章子, 向井 良, 澤田 宏, 牧野 昭二 (NTT 研究所)

1. はじめに

本稿では, 音源数 N >マイク数 M の場合のブラインド音源分離 (BSS: blind source separation) について報告する. これまで, 信号のスパース性を用いる手法 (e.g., [1]) が提案されてきた. これは, 1つの信号しか存在しない時刻にその信号を抽出し再構成する手法であるが, 信号に不連続に0が詰められるため Musical Noise が顕著であった. これに対し本稿では, 信号のスパース性を用いて観測信号から $N-M$ 個の信号を除去した後に, 残りの混合信号に対して通常の ICA (independet component analysis) による分離を行う手法を提案する. 提案法では, 分離信号に詰められる0を低減でき, 歪の少ない出力が得られる.

2. 問題設定

実環境では, 音源信号に残響が畳込まれ混合されて観測される. 音源 i からの信号 s_i に部屋のインパルス応答 h_{ji} が畳込まれた信号が混合するため, マイク j による観測信号 x_j は, $x_j(n)=\sum_{i=1}^{N}\sum_{k=1}h_{ji}(k)s_i(n-k+1)$ $(j=1,\cdots,M)$ とモデル化される (図1). ここで N は音源数, M はマイク数である. 本稿では $N=3, M=2$ とする. また, N 個の信号源は統計的に互いに独立であり, それぞれの信号はスパースであると仮定する. 信号がスパースであるとは, 信号が殆どの時刻 n において0であることを指す.

BSS の目的は, 観測信号 x_j やインパルス応答 h_{ji} を知らずに音源信号の推定 y_k を求めることである.

畳み込み混合の問題は扱いが繁雑であること, またスパース性の仮定は時間 - 周波数領域でよりよく成立することから [2], 信号を時間 - 周波数領域に変換した上で問題を扱うことが有効である. 観測信号は時間 - 周波数領域で $\boldsymbol{X}(\omega,m)=\boldsymbol{H}(\omega)\boldsymbol{S}(\omega,m)$ となる. ここで, ω は周波数, m はフレーム時刻を表す. また, $\boldsymbol{H}(\omega)$ は j-i 要素に信号源 i からセンサ j までの周波数応答 $H_{ji}(\omega)$ を持つ $(M\times N)$ 行列であり, $\boldsymbol{S}(\omega,m)=[S_1(\omega,m),\cdots,S_N(\omega,m)]^T$, $\boldsymbol{X}(\omega,m)=[X_1(\omega,m),\cdots,X_M(\omega,m)]^T$ は, それぞれ原信号と観測信号の短時間フーリエ変換 (STFT) の結果である.

3. 従来法

音源数 N >マイク数 M の場合の BSS 問題を解くために, 信号のスパース性を利用し, 時間 - 周波数領域でのバイナリマスク (= 時間周波数マスキング) を用いて信号を分離する方法が提案されている (e.g., [1]).

信号のスパース性を仮定することで, 複数の信号が同時に存在していても, 各サンプルレベルで見れば同時刻に互いに重なりあって観測される頻度は低いことを仮定できる (詳しくは [3] を参照). 従って, それぞれの時刻で観測された信号が, どの信号源から発せられた信号であるかを推定し, その時刻の信号のみを抽出することで信号を分離できる. 例えば各時刻における観測信号の位相差やレベル差を分類し, それぞれのクラスに属する時刻の信号を再構成することで各源信号を推定できる.

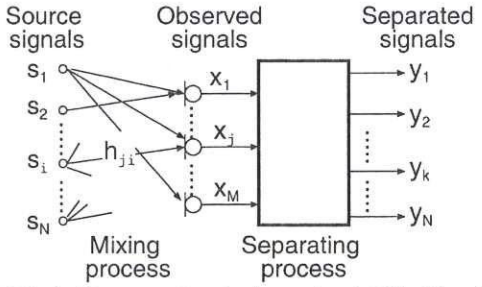

図 1 Block diagram of underdetermined BSS. $N>M$.

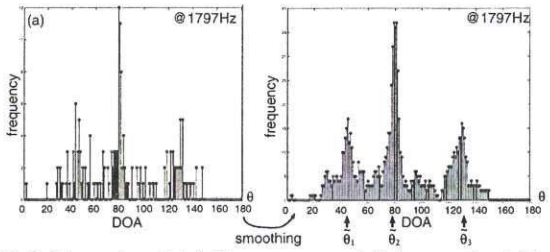

図 2 Example of (a) histogram and (b) smoothed histogram.

ここでは2つの無指向性マイクを用いることを想定し, 観測信号 $X_1(\omega,m)$ と $X_2(\omega,m)$ の間の位相差 $\varphi(\omega,m)=\angle\frac{X_1(\omega,m)}{X_2(\omega,m)}$ を用いる. そして, 信号到来方向 (DOA) を $\theta(\omega,m)=\cos^{-1}\frac{\varphi(\omega,m)c}{\omega d}$ によって推定する. ここで c は音速, d はマイク間隔である. $\theta(\omega,m)$ についての各周波数でのヒストグラムは3音源に対応する3つのピークを持つ. これを小さい方から $\tilde{\theta}_1,\tilde{\theta}_2,\tilde{\theta}_3$ とし (図2), また $\tilde{\theta}_k$ の近傍より到来する信号を $\tilde{S}_k(k=1,2,3)$ とする.

従来法ではこの $\tilde{\theta}_k$ を用いて作成したバイナリマスク

$$M_k(\omega,m)=\begin{cases}1 & \tilde{\theta}_k-\Delta\le\theta(\omega,m)\le\tilde{\theta}_k+\Delta\\ 0 & \text{otherwise}\end{cases}\qquad(1)$$

を観測信号に適用し, $Y_k(\omega,m)=M_k(\omega,m)X_j(\omega,m)$ $(j=1 \text{ or } 2)$ によって, k 番目の分離信号を得る.

ここで Δ を小さくすると高い分離性能が得られるが, バイナリマスクにより分離信号に多くの0が詰められるため歪が大きく, Musical Noise が発生する. 一方, この歪を回避するため Δ を大きくした場合には, 分離信号に詰められる0が減るため Musical Noise は減少するが, 分離性能が劣化する.

4. 提案法

前述の Musical Noise を回避するため, 時間周波数マスキングと ICA の併用による分離法を提案する.

提案法は2段から成る (図3). 1段目では, 信号のスパース性を用いて1つの信号を除去し, 2つの信号の混合音を抽出する. この2信号の混合音の抽出には, 従来法より1の多いバイナリマスクを用いるため信号への0詰めの影響が低減され, Musical Noise の軽減が期待できる. 次に2段目において, この2信号の混合音を ICA を用いて分離する.

* Blind Separation of more signals than sensors combining binary-masks and ICA, by S. Araki, R. Mukai, H. Sawada and S. Makino (NTT Corporation).

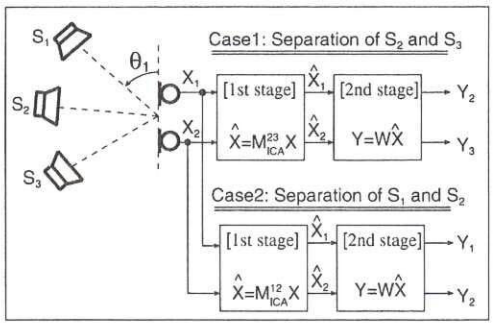

図 3　System setup

[1st stage] 1 信号除去： まず, 3 章と同様に $\tilde{\theta}_1, \tilde{\theta}_2, \tilde{\theta}_3$ を求める. 次にここでは, 1 つの信号を抽出するバイナリマスクではなく, 次のバイナリマスク

$$M_{\mathrm{ICA}}^{pq}(\omega, m) = \begin{cases} 1 & \theta_{min} \leq \theta(\omega, m) \leq \theta_{max} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

を用いて, $\tilde{S}_p$ と $\tilde{S}_q$ の 2 信号の混合信号 $\hat{\boldsymbol{X}}(\omega, m) = M_{\mathrm{ICA}}^{pq}(\omega, m)\boldsymbol{X}(\omega, m)$ を得る. ここで θ_{min}, θ_{max} は, 例えば, 信号 $\tilde{S}_1$ の除去後 $\tilde{S}_2$ と $\tilde{S}_3$ を分離する場合, $\tilde{\theta}_1 < \theta_{min} < \tilde{\theta}_2, \theta_{max} = 180°$ を用いる. これを Case 1 と呼ぶ. 同様に信号 $\tilde{S}_3$ の除去後 $\tilde{S}_1$ と $\tilde{S}_2$ を分離する場合, $\theta_{min} = 0°, \tilde{\theta}_2 < \theta_{max} < \tilde{\theta}_3$ を用いる. これを Case 2 と呼ぶ (図 4).

[2nd stage] ICA による分離: バイナリマスク (2) により抽出した信号 $\hat{\boldsymbol{X}}$ は 2 信号の混合音であると考えて良いので, ここでは, 2 入力 2 出力の ICA を用いて $\hat{\boldsymbol{X}}$ を分離する. 具体的に分離信号 $\boldsymbol{Y}(\omega, m) = [Y_j(\omega, m), Y_k(\omega, m)]^T$ は $\boldsymbol{Y}(\omega, m) = \boldsymbol{W}(\omega)\hat{\boldsymbol{X}}(\omega, m)$ により計算される. ここで $\boldsymbol{W}(\omega)$ は 2×2 の分離行列であり, $Y_j(\omega, m)$ と $Y_k(\omega, m)$ が互いに独立となるように求める. 本稿では, 更新式 $\boldsymbol{W}_{i+1} = \boldsymbol{W}_i + \eta\{\mathrm{diag}\left(\langle\Phi(\boldsymbol{Y})\boldsymbol{Y}^{\mathrm{H}}\rangle\right) - \langle\Phi(\boldsymbol{Y})\boldsymbol{Y}^{\mathrm{H}}\rangle\}\boldsymbol{W}_i$ (但し $\Phi(\boldsymbol{x}) = \phi(|\boldsymbol{x}|) \cdot e^{j \cdot \angle(\boldsymbol{x})}$, $\phi(x) = \tanh(gx)$ (g=100) [4]) を用いた. また ICA の scaling 問題を解くため, MDP [5] を用いた.

なお, 1 系統では 2 出力しか得られないため, 3 出力を得るには図 3 に示すように 2 系統の分離処理を行う.

5.　実験とその結果

無響シミュレーション及び残響下での実験を行った. マイク間隔 4cm, 音源方位 45°(女声), 100°(男声), 135°(男声) とし, 無響シミュレーションでは原音声に該当する遅延を与えて混合した. 残響下での実験では, 残響時間 T_R = 130ms の室内で録音した各音声を足し合わせた. サンプリング周波数は 8kHz, DFT フレーム長は 512 である. またバイナリマスクのパラメタは, Case 1 では $\theta_{min} = \tilde{\theta}_2 - \Delta$, $\theta_{max} = 180°$, Case 2 では $\theta_{min} = 0°$, $\theta_{max} = \tilde{\theta}_2 + \Delta$, $\Delta = 10°$ とした.

分離性能の評価尺度には, 信号対妨害音比 (SIR) と信号対歪比 (SDR) を用いた.

表 1 は, バイナリマスクにより失われた信号のパワー比 $\frac{\sum_n s_i(n)^2 - \sum_n \hat{s}_i(n)^2}{\sum_n s_i(n)^2}$ を示している. ここで $\hat{s}_i(n) = \mathrm{IDFT}[M_i(\omega, m)S_i(\omega, m)]$ である. 従来法によるマスクでは約 20% のパワーが失われていたが, 提案法によるマスクでは数 % のパワーしか失われていないことが分かる. これより 2nd stage で分離ができれば, 零詰めによる歪みの少ない信号が得られることが期待できる.

図 4 では, 各領域における各信号のパワー比を示している. Case 1 の領域 ($\tilde{\theta}_2 - \Delta \leq \theta(\omega, m) \leq 180°$) では

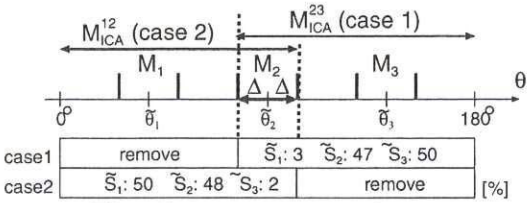

図 4　Source power in each area.

表 1　Power lost by binary masks (in %)

mask	M_1	M_2	M_3	M_{ICA}^{12}		M_{ICA}^{23}	
output	Y_1	Y_2	Y_3	Y_1	Y_2	Y_2	Y_3
[%]	17	14	23	2.5	5.7	8.1	0.7

表 2　Results. Case 1, 2 : by proposed method

(a) 無響　[dB]

	SIR1	SIR2	SIR3	SDR1	SDR2	SDR3
従来法	17.6	11.6	17.3	7.3	9.3	8.5
Case 1	-	8.8	13.6	-	12.5	**16.2**
Case 2	17.5	8.8	-	**20.8**	11.8	-

(b) 残響あり. T_R=130ms　[dB]

	SIR1	SIR2	SIR3	SDR1	SDR2	SDR3
従来法	9.9	4.7	8.3	4.0	8.3	4.8
Case 1	-	4.1	9.6	-	9.1	**6.9**
Case 2	9.0	3.3	-	**9.4**	10.3	-

$\tilde{S}_2$ と $\tilde{S}_3$ の 2 信号が, Case 2 の領域 ($0° \leq \theta(\omega, m) \leq \tilde{\theta}_2 + \Delta$) では $\tilde{S}_1$ と $\tilde{S}_2$ の 2 信号が主要である. これより 2nd stage において ICA を用いても良いと判断した.

表 2 に分離実験結果を示す. 従来法では歪みの指標である SDR が低いが, 提案法では分離性能 SIR をほとんど落とすことなく高い SDR を得ることができている. これは信号の歪みが少ない分離ができていることを示している. また残響下においても分離が可能であり, 従来法よりも歪が小さい出力が得られた. 尚, 他の音声組合せや音源配置についても同等の効果を確認している [6]. [7] にてデモ音声を公開している.

6.　まとめ

音源数>マイク数の場合の音源分離について検討した. 提案法では時間周波数マスキングと ICA を併用したため, 分離音に零詰めが起こりにくく, 従来法で問題であった Musical Noise を低減することができた.

参考文献

[1] S. Rickard and O. Yilmaz, "On the W-Disjoint orthogonality of speech," *Proc. ICASSP2002*, vol.1, pp. 529-532, 2002.

[2] P. Bofill and M. Zibulevsky, "Blind separation of more sources than mixtures using sparsity of their short-time Fourier transform," *Proc. ICA2000*, pp. 87-92, 2000.

[3] A. Blin, et al., "Blind source separation when speech signals outnumber sensors using a sparseness-mixing matrix combination," *Proc. IWAENC2003*, 2003.

[4] H. Sawada, et al., "Polar coordinate based nonlinear function for frequency domain blind source separation," *Proc. ICASSP2002*, pp. 1001-1004, May 2002.

[5] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," *Proc. ICA2001*, pp. 722-727, Dec. 2001.

[6] S. Araki et al., "Blind separation of more speech than sensors with less distortion by combining sparseness and ICA," *Proc. IWAENC2003*, 2003.

[7] http://www.kecl.ntt.co.jp/icl/signal/araki/underdeterminedBSSdemo.html