3—4—9

サブバンド処理によるブラインド音源分離に関する検討 *

◎荒木 章子[1], △ Robert Aichner[1,2], 牧野 昭二[1], 西川 剛樹[3] 猿渡 洋[3]

[1] 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,

[2]University of Applied Sciences Regensburg, [3] 奈良先端科学技術大学院大学

1. はじめに

ブラインド音源分離 (Blind Source Separation:BSS) は、観測された混合信号のみから音源信号を推定する手法であり、音源信号同士の独立性の仮定に基づく手法をはじめ近年多くの手法が提案されている [1]。

本稿では、はじめに、周波数領域 BSS において残響に対応するため十分な長さのフィルタを持つ分離系を推定する時に、各周波数で統計的性質が悪化するために分離性能が劣化するという問題を示す。そして、本問題を解決するため、各帯域で BSS に必要な信号の統計的性質を満たしたまま、十分な長さのフィルタを持つ分離系を推定できる「サブバンド BSS」を提案し、その効果を実験により確認したので報告する。

2. 混合信号のモデルと分離系

実環境での観測信号は、一般に音源からの信号に残響が畳込まれたうえで混合されて観測される。これより、実環境で観測される混合信号を次のようにモデル化する。

音源 i からの信号 s_i に部屋の応答 h_{ji} が畳み込まれた信号が混合し、マイク j で信号 x_j として観測されるとする。

$$x_j(n) = \sum_{i=1}^{N}\sum_{l=1}^{L} h_{ji}(l)s_i(n-l+1) \ (j = 1, \cdots, M). \quad (1)$$

h_{ji} は音源 i からマイク j への L タップのインパルス応答である。本稿では音源数 N= マイク数 M=2 とする (図 1)。

推定する分離系は K タップの分離フィルタ群 w_{ij} から成り、分離信号は次のように表される。

$$y_i(n) = \sum_{j=1}^{M}\sum_{k=1}^{K} w_{ij}(k)x_j(n-k+1) \ (i = 1, \cdots, N). \quad (2)$$

3. 周波数領域 BSS

3.1 周波数領域 BSS の方法

周波数領域 BSS[2, 3] では、式 (1) を T ポイントの DFT を用いて周波数領域の表現に変換する。

$$\boldsymbol{X}(\omega, m) = \boldsymbol{H}(\omega)\boldsymbol{S}(\omega, m). \quad (3)$$

これにより、式 (1) の畳込混合を各周波数で瞬時混合として表現でき、問題を簡単化できる。

次に、各周波数において出力信号 $Y_1(\omega, m)$、$Y_2(\omega, m)$ が互いに独立となるよう、(2×2) の分離行列 $\boldsymbol{W}(\omega)$ を推定する。

$$\boldsymbol{Y}(\omega, m) = \boldsymbol{W}(\omega)\boldsymbol{X}(\omega, m). \quad (4)$$

推定は例えば Kullback-Leibler divergence を最小にする手法を用いて行うことができる [3]。

3.2 周波数領域 BSS の問題点

周波数領域 BSS を用いて分離系 $\boldsymbol{W}$ を求める場合には、残響まで含めて妨害音を十分に除去するため、部屋のインパルス応答長 L より長いフレーム T を用いて DFT 分析を行い、周波数 bin の数を増やす必要がある。

しかし、我々は長いフレームを用いる場合に性能が劣化することを確認した [4]。図 2は、フレーム長 T を変化させた場合の分離性能である。実験で用いたセットアップはマイク間隔 4cm、音源方位 −30° 及び 40°、残響時間 T_R=300ms (2400 タップ) である。混合音 (= 学習データ) は音声に実測したインパルス応答を畳み込んで作成した。学習データの長さは 3 秒もしくは 8 秒であり、それぞれを用いて求めた分離行列で 8 秒全体を分離した。分離手法は文献 [4] を参照されたい。分離性能の評価尺度は、Signal to Interference Ratio (SIR= 出力 SNR− 入力 SNR) である。尚、時間周波数領域におけるデータ個数を揃えるためフレームシフトを $T/2$ に固定した。図 2より長いフレームを用いる際に性能が悪化することが分かる。

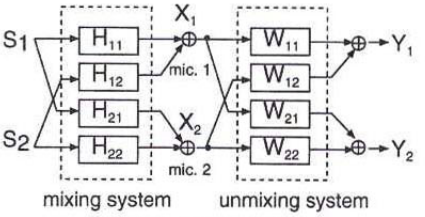

図 1 BSS の構成

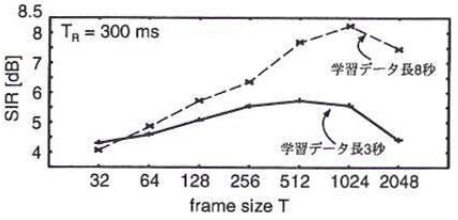

図 2 フレーム長と分離性能の関係。T_R=300 ms。

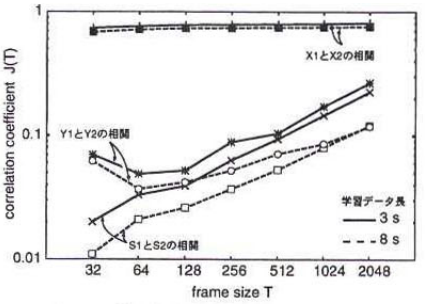

図 3 フレーム長と相関係数の関係。T_R = 0 ms。

この原因として、フレーム長が長いほど各周波数におけるデータ数が減るために各周波数での統計的性質が悪くなることが考えられる。そこで相関係数 r_ω を全ての周波数で求め平均した

$$J(T) = \frac{1}{T}\sum_{\omega}^{T} |r_\omega| \quad (5)$$

を用いて各フレーム長における独立性を評価した。相関係数は独立性を直接示すものではないが、他の評価量でも同様の傾向が得られている。結果を図 3に示す。長いフレームを用いる時に、信号が大きな相関を持ち、独立性の仮定が崩れていることが分かる。独立性の仮定が崩れると、分離行列の学習が阻害される [5]。

なお、我々は長い学習データを用いる場合には、より長いフレームでより良い性能が得られることを確認している (図 2破線)。しかし、実際は部屋のインパルス応答が長く変化し

*Blind source separation using SSB subband, by S. Araki, R. Aichner, S. Makino(NTT Communication Science Laboratories, NTT Corporation), T. Nishikawa, and H. Saruwatari (Nara Institute of Science and Technology).

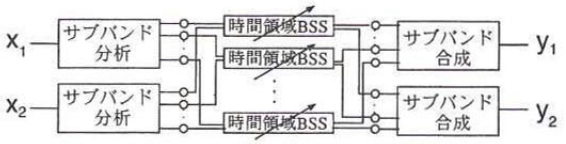

図 4　サブバンド BSS の構成

ないことはありえないため、長い学習データで良い分離性能を得ても意味がない。

4. サブバンド処理による BSS

本章では、できるだけ短い学習データを用いて残響に対応できる長い分離フィルタを推定するために、サブバンド処理を用いる BSS(サブバンド BSS) を提案する。

4.1 期待される利点

サブバンド BSS では分割するサブバンドの個数は自由に選べるため、各サブバンドで統計的性質を十分満たせる個数を選べばよい。また、周波数領域 BSS では各周波数で 1tap のフィルタしか推定できなかったため長いフレームを用いて多くの周波数に分割する必要があったが、サブバンド BSS では各帯域ごとに長いフィルタを持たせることができるため、分割数が少なくてもフルバンドで見たときに十分長いフィルタを推定できる。また、各帯域でのフィルタは短くて済むので、時間領域アルゴリズムの収束性の悪さを回避できる利点もある。

4.2 サブバンド BSS の方法

図 4にサブバンド BSS の構成を示す。まず入力された信号を、ポリフェーズフィルタバンクを用いてサブバンド分析する。このとき、各帯域で信号を実数で扱うために Single Side Band (SSB) サブバンド [6] を用いる。SSB サブバンドでは周波数領域でのエイリアジングを回避するために、M 個のサブバンドに分割する時、ダウンサンプリングレート R を $R = M/4$ とする必要がある。

次に各帯域において、時間領域 BSS を用いて分離を行う。最後に各帯域で分離された信号を合成し、出力を得る。

各帯域で用いる時間領域 BSS アルゴリズムは、信号の非定常性に基づく評価関数から導出されたもの [7] を用いる。出力信号の相互相関が全ての時間ブロックにおいて 0 になった時に最小値 0 をとる非負の評価関数 [8]

$$Q = \frac{1}{2}\sum_{b=1}^{B}\{\sum_{j=1}^{M}\log\langle y_j^2(n)\rangle_b - \log\det \boldsymbol{R}_y^b(0)\} \qquad (6)$$

を考える。ここで $\boldsymbol{y}(n) = [y_1(n), y_2(n)]^T$ は出力信号、$\boldsymbol{R}_y^b(\tau)$ は出力信号の共分散行列 $(=\langle \boldsymbol{y}(n)\boldsymbol{y}^T(n-\tau)\rangle_b)$、$\langle x\rangle_b$ はブロック b (b=1, $\cdots$, B) についての時間平均である。分離フィルタ w_{ij} の更新式は、この評価関数 Q を $\boldsymbol{w}(k)$ で微分し算出した natural gradient において、更に性能を高めるため $\boldsymbol{R}_y^b(0)$ のみでなく時間ずれの相関 $\boldsymbol{R}_y^b(\tau)$ も考慮することで次のように得られる。

$$\boldsymbol{w}_{i+1}(k) = \boldsymbol{w}_i(k) + \frac{\alpha}{B}\sum_{b=1}^{B}\{(\mathrm{diag}\boldsymbol{R}_y^b(0))^{-1}(\mathrm{diag}\boldsymbol{R}_y^b(k)) - (\mathrm{diag}\boldsymbol{R}_y^b(0))^{-1}\boldsymbol{R}_y^b(k)\}\boldsymbol{W}_i(z) \qquad (7)$$

α はステップサイズ、$W_{ij}(z)=\sum_{k=0}^{K-1} w_{ij}(k)z^{-k}$、$z^{-1}$ はディレイオペレータである。導出の詳細については、文献 [7] を参照のこと。

4.3 実験結果および考察

サブバンド BSS の有効性を確かめるため、周波数領域 BSS とサブバンド BSS の分離性能を比較した。

用いたデータのセットアップは 3.2節と同様であり、残響時間 T_R = 150ms (1200 タップ) 及び 300ms (2400 タップ)、音声組合せは男男・男女・女女の 3 通りである。3 秒の混合音声を用いて分離を行い、その 3 秒を含む約 8 秒の信号を分離した。周波数領域 BSS は [4] の手法を用いた。

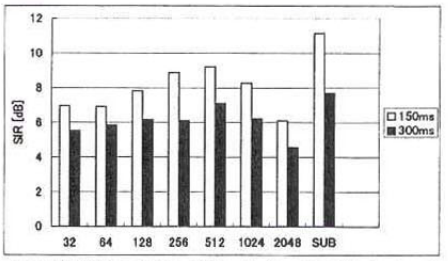

図 5 周波数領域 BSS 及びサブバンド BSS による性能比較。数字は周波数領域 BSS におけるフレーム長 (= フィルタ長)、SUB がサブバンド BSS による結果。

サブバンド BSS では、サブバンド数は 64、間引率は 16 とした。これは、間引率としては周波数領域 BSS の「32」に相当する。また各帯域において 64 タップの分離フィルタを用いた。これはフルバンドで 1024 タップの分離フィルタに相当する。

時間領域アルゴリズムで用いた α は 1.0×10^{-3}、B は 20、学習の反復回数は 500 回である。初期値には $\pm60°$ へ死角を向ける 1024 タップの死角型ビームフォーマをサブバンド分析して用いた。

結果を図 5に示す。周波数領域 BSS では長さ 1024 の長いフィルタ (= フレーム長) を用いた時に分離性能が劣化したが、サブバンド BSS では良い性能が得られることが分かる。なお、原信号をサブバンド分割したものの相関係数 (式 (5)) の値は、男男 0.028、男女 0.018、女女 0.020 であった。よって、独立性の仮定は十分保たれていると考えてよい。

5. まとめ

本稿では、周波数領域 BSS において残響に対応するため十分に長い分離フィルタを推定する時に生じる、各周波数での統計的性質の悪化により性能が劣化する問題を解決するため、(1) 各帯域で BSS に必要な信号の統計的性質を満たしたまま、(2) 十分に長い分離フィルタを推定できる、サブバンド BSS を提案し、その効果を示した。

謝辞　SSB サブバンドに関してご協力頂いた NTT サイバースペース研究所の羽田陽一博士、中川朗氏、阪内澄宇氏に感謝致します。

参考文献

[1] S. Haykin, *Unsupervised adaptive filtering.* John Wiley & Sons, 2000.

[2] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 3, pp. 320-327, May 2000.

[3] S. Ikeda and N. Murata, "A method of ICA in time-frequency domain," *Proc. ICA99*, pp. 365-370, Jan. 1999.

[4] S. Araki, S. Makino, T. Nishikawa, and H. Saruwatari, "Fundamental limitation of frequency domain blind source separation for convolutive mixture of speech," *Proc. ICASSP2001*, vol. 5, pp. 2737-2740, May 2001.

[5] S. Araki, S. Makino, R. Mukai, and H. Saruwatari, "Equivalence between frequency domain blind source separation and frequency domain adaptive null beamformers," *Proc. Eurospeech2001*, pp. 2595-2598, Sept. 2001.

[6] R. Crochiere and L. Rabiner, *Multirate Digital Signal Processing,* Englewood Cliffs, NJ: Prentice-Hall, 1983.

[7] T. Nishikawa, H. Saruwatari, K. Shikano, "Comparison of blind source separation methods based on time-domain ICA using nonstationarity and multistage ICA, " *IEICE Tech. Rep.*, Jan. 2002.

[8] M. Kawamoto, K. Matsuoka and N. Ohnishi, "A method of blind separation for convolved nonstationary signals," *Neurocomputing* vol. 22, pp. 157-171, 1998.