

死角型ビームフォーマを初期値に用いる時間領域ブラインド音源分離*

◎荒木 章子¹, △Robert Aichner^{1,2}, 牧野 昭二¹, 西川 剛樹³ 猿渡 洋³¹ 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]。

実環境では、残響を抑圧するために十分な長い分離フィルタを推定する必要がある。しかし時間領域ブラインド音源分離の学習は初期値に非常に敏感で、長い分離フィルタの推定は難しかった。そこで本稿では、時間領域ブラインド音源分離のための分離フィルタの初期値設定法を提案する。我々は先に、周波数領域 BSS の分離のメカニズムが妨害音方向への死角形成であることを示した [2]。この事実を時間領域 BSS へ利用し、分離フィルタの初期値に死角型ビームフォーマを実現するフィルタを用いることを試みる。これにより、実環境での分離に必要な 1000 タップを越える長さの分離フィルタの推定に成功した。

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み込まれたうえで混合されて観測される。この観測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。

$$x_j(n) = \sum_{i=1}^N \sum_{k=0}^{P-1} h_{ji}(k) s_i(n-k) \quad (j=1, \dots, M). \quad (1)$$

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

推定する分離系は Q 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。

$$y_i(n) = \sum_{j=1}^M \sum_{k=0}^{Q-1} w_{ij}(k) x_j(n-k) \quad (i=1, \dots, N). \quad (2)$$

3. 時間領域 BSS

ここでは、信号の非定常性に基づく時間領域 BSS アルゴリズム [3] を用いる。

分離フィルタ $w(k) = [w_{ij}(k)]$ (i, j 要素が $w_{ij}(k)$ である行列) は、出力信号 $y_i(n)$ が、複数の分析区間全てにおいて、時間ずれの相関まで含めて無相関となるよう学習される。

$$w_{i+1}(k) = w_i(k) + \frac{\alpha}{B} \sum_{b=1}^B \{ (\text{diag} R_y^b(0))^{-1} (\text{diag} R_y^b(k)) - (\text{diag} R_y^b(0))^{-1} R_y^b(k) \} W_i(z) \quad (3)$$

*Time-domain blind source separation with utilization of null beamforming, by S. Araki, R. Aichner, S. Makino (NTT Communication Science Laboratories, NTT Corporation), T. Nishikawa, and H. Saruwatari (Nara Institute of Science and Technology).

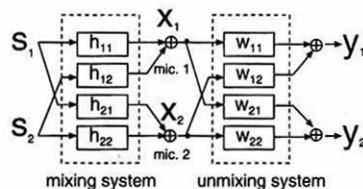


図 1 BSS の構成

ここで $y(n) = [y_1(n), y_2(n)]^T$ は出力信号、 $R_y^b(\tau)$ は出力信号の共分散行列 $(= \langle y(n) y^T(n-\tau) \rangle_b)$ 、 $\langle \cdot \rangle_b$ は時間ブロック b ($b=1, \dots, B$) についての時間平均、 α はステップサイズ、 $W_{ij}(z) = \sum_{k=0}^{Q-1} w_{ij}(k) z^{-k}$ 、 z^{-1} はディレイオペレータである。

4. 時間領域 BSS の初期値設定: 死角型ビームフォーマの利用

時間領域 BSS は、音源の分離動作だけでなくデコンボリューションまで行って、信号自身の時間相関まで下げながら信号同士を独立にするという複雑な動作を行う [4]。そのため、時間領域アルゴリズムを収束に導くことは非常に難しく、収束先での分離性能は初期値に大きく依存する。

そこで本稿では、時間領域 BSS の学習を分離動作方向へ導くための初期値の設定を考える。我々は周波数領域 BSS がデコンボリューションは行わず分離動作のみを行い、さらにこの分離が妨害音方向からの音を抑圧する動作により実現されている、という知見を得ている [2]。これを基に、時間領域 BSS の分離フィルタの学習の初期値 $w_0(k)$ として、ある角度方向へ鋭い空間的死角を作る死角型ビームフォーマ (Null Beam Former: NBF) を用いることを提案する。

NBF は、到来音波が平面波であることを仮定し、ある角度方向からの妨害音波が各マイクへ到達する時間差を補償するフィルタを用い、各マイクで観測した妨害音を同相化し減算することで妨害音出力を抑圧する。

初期値 1

図 2 に 2 マイクの場合について各マイクに到達した音波の位相を揃えて減算し、NBF を実現する方法を示す。図 2 より、 $w_0(k)_{ij}$ ($i \neq j$) にはマイク i とマイク j (それぞれ座標 d_i, d_j にあるとする) への音波の到達時間差 τ_{ji} を実現する sinc 関数を、 $w_0(k)_{ij}$ ($i=j$) にはデルタ関数を利用すると良いことが分かる。この時、非因果成分を実現するために適当なディレイ D を入れる。即ち、

$$w_0(k)_{ij} = -\text{sinc}(\omega_{\max}(k - \tau_{ji}) - D) \quad (\text{for } i \neq j) \quad (4)$$

$$w_0(k)_{ij} = \delta(k - D) \quad (\text{for } i=j) \quad (5)$$

ここで $\omega_{\max} = 2\pi f_s/2$ 、 f_s はサンプリング周波数であり、 $D = \frac{Q}{2}$ とした。

初期値 2

ここでは、音源 i からマイク j へ到達する音波と音源

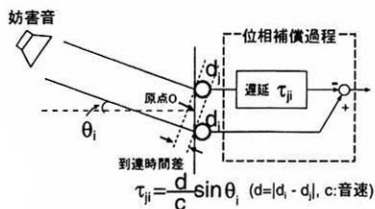


図2 NBFの構成と、初期値1の説明。

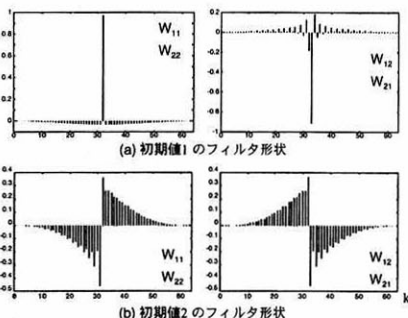


図3 初期値 $w_0(k)$ の形状。フィルタ長 $Q=64$ の場合。62.5Hz 以下をカットしている。

i から原点 O へ到達する音波の時間差 $\tau_{ji} = \frac{d_j}{c} \sin(\theta_i)$ を用いる。この時混合系 $H(\omega)$ が、マイク間の到達時間差のみを表すと仮定すると、 $H(\omega)$ は周波数領域で

$$H(\omega) = \begin{bmatrix} \exp(j\omega\tau_{11}) & \exp(j\omega\tau_{12}) \\ \exp(j\omega\tau_{21}) & \exp(j\omega\tau_{22}) \end{bmatrix} \quad (6)$$

と書ける。長さ Q の分離フィルタの初期値を得るには、はじめに周波数分解能を $2\pi f_s/Q$ (f_s はサンプリング周波数) として式 (6) を計算し、 $W(\omega) = H^{-1}(\omega)$ により NBF の伝達関数を求める。次にこれを逆 FFT を用いて時間領域表現を得る。

$$w'_0(k)_{ij} = \text{IFFT}[W_{ij}(\omega)] \quad (7)$$

このようにして得られる $w'_0(k)_{ij}$ は非因果成分が波形の後半部分にまわりこんだ形で得られるため、 $w'_0(k)_{ij}$ の波形の前半と後半を入れ替える必要がある。よって分離フィルタの初期値 $w_0(k)_{ij}$ は

$$w_0(k)_{ij} = w'_0(k + \frac{Q}{2})_{ij} \quad (\text{for } k = 0, \dots, \frac{Q}{2} - 1) \quad (8)$$

$$w_0(k)_{ij} = w'_0(k - \frac{Q}{2})_{ij} \quad (\text{for } \frac{Q}{2}, \dots, Q - 1) \quad (9)$$

となる。初期値 2 では、目的音方向の利得が 1 に拘束される。図 3 に初期値 1、2 によるフィルタ形状をそれぞれ示す。

5. 実験結果

実験に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 $T_R = 300\text{ms}$ である。サンプリング周波数は 8 kHz であり、300ms の残響時間は 2400 タップに相当する。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込んで作成した。音声の組合せは男女 2 名ずつの音声 6 組である。学習データ、分離データの長さはそれぞれ 3 秒である。分離性能の評価尺度は、Signal to Interference Ratio (SIR = 出力 SNR - 入力 SNR) である。

分離フィルタの初期値には $\pm 60^\circ$ 方向へ死角を作る NBF を用いた。初期値 1 および時間領域 BSS により

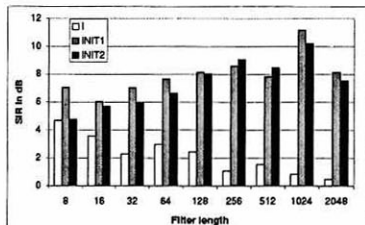


図4 式 (10) と NBF のそれぞれの初期値を用いた場合の各フィルタ長における分離性能。数字は SIR [dB]。

分離信号のスペクトルが平坦になる白色化が生じるため、分離後に補正を行っている [5]。時間領域 BSS アルゴリズムのステップサイズ α は初期値 I では 0.1、初期値 2 では 0.002、ブロック数 B は両者とも 20 を用いた。また学習の反復回数は、白色化補正をする前に最も良い SIR を与える反復回数を選んだ。

分離結果を図 4 に示す。図中の I は、初期値を通常用いられる

$$w_0(k) = \begin{bmatrix} \delta(k - \frac{Q}{2}) & 0(k) \\ 0(k) & \delta(k - \frac{Q}{2}) \end{bmatrix} \quad (10)$$

$$(k = 0, \dots, Q - 1)$$

とした場合の分離性能を示している。式 (10) の初期値では、残響下で分離を行う為に必要な長い分離フィルタの推定は難しいことが分かる。

一方、図 4 の INIT1 は、初期値として NBF を実現する初期値 1 を用いた結果であり、INIT2 は、初期値 2 を用いた結果である。分離フィルタ長が長い場合でも推定が可能となり、高い分離性能を得ることができている。なお初期値の段階における SIR の値は、 $Q=1024$ の場合で約 5dB であった。

初期値 2 を用いる場合は、初期値による信号の白色化は起こらないが、時間領域 BSS による白色化があるため、依然として分離後の補正が必要である。

6. まとめ

本稿では、初期値に敏感で収束させることが難しい時間領域ブラインド音源分離 (BSS) に対し、BSS の分離が妨害音方向への死角形成で実現されていることを利用して、分離フィルタの学習の初期値に死角型ビームフォーマ (NBF) を利用することを提案した。この初期値により時間領域 BSS で実環境での分離に十分な長さの分離フィルタを推定することに成功した。

参考文献

- [1] S. Haykin, *Unsupervised adaptive filtering*. John Wiley & Sons, 2000.
- [2] 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.
- [3] T. Nishikawa, H. Saruwatari, K. Shikano, "Comparison of blind source separation methods based on time-domain ICA using nonstationarity and multi-stage ICA," *IEICE Tech. Rep.*, Jan. 2002.
- [4] X. Sun and S. Douglas, "A natural gradient convolutional blind source separation algorithm for speech mixtures," in *Proc. Conference Indep. Compon. Anal. Signal. Sep.*, Dec. 2001, pp. 59-64.
- [5] R. Aichner, S. Araki, S. Makino, T. Nishikawa, and H. Saruwatari, "Time domain blind source separation of non-stationary convolved signals by utilizing geometric beamforming," in *NNSP2002*, 2002, (accepted).