3—7—4

実環境での混合音声に対する周波数領域ブラインド音源分離手法の性能限界 *

◎荒木 章子†, 牧野 昭二† (NTT CS 基礎研), 西川 剛樹‡, 猿渡 洋‡ (奈良先端大)

1. はじめに

ブラインド音源分離 (Blind Source Separation: BSS) とは, 観測できる混合信号の情報と各音源信号同士の独立性の仮定のみから音源信号を推定する手法であり, 近年多くの手法が提案されている [1, 2, 3]. この手法は, ノイズに頑健な音声認識の前処理として, また聴覚情景解析への一手法として注目されている. しかし, BSS の実環境音声信号への適用例はまだ少なく, 分離性能は必ずしも十分とはいえない.

本稿では, 周波数領域 BSS をとりあげ, 残響を持つ環境における混合音声信号に対する分離性能を考察する. 従来, 逆フィルタを推定する手法では部屋のインパルス応答長より長い分析フレームが必要だと考えられている [4, 5]. これに対し, 周波数領域 BSS では長いフレーム長を用いては却って分離性能が低下することを実験で示し, 周波数領域 BSS では逆フィルタの推定はできていないことを指摘する [6].

2. 周波数領域 BSS

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たものが, マイクロホン j で信号 x_j として観測されるとする.

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

h_{ji} は音源 i からマイクロホン j への P タップのインパルス応答である. 本稿では, 簡単のため, $N = M = 2$ とする (Fig.1).

周波数領域 BSS 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.

$$\boldsymbol{X}(\omega, m) = \boldsymbol{H}(\omega)\boldsymbol{S}(\omega, m). \tag{2}$$

これにより, 式 (1) の重畳信号の混合を, 各周波数での瞬時和として表現でき, 問題を簡単化できる.

次に, 各周波数において $Y_1(\omega, m)$ と $Y_2(\omega, m)$ が互いに独立となるよう, 逆混合行列 $\boldsymbol{W}(\omega)$ を推定する.

$$\boldsymbol{Y}(\omega, m) = \boldsymbol{W}(\omega)\boldsymbol{X}(\omega, m). \tag{3}$$

推定は Kullback-Leibler divergence を最小にする手法を用いて式 (4) により行う [7, 8].

$$\boldsymbol{W}_{i+1} = \boldsymbol{W}_i + \eta\left[\mathrm{diag}\left(\langle\Phi(\boldsymbol{Y})\boldsymbol{Y}^{\mathrm{H}}\rangle\right) - \langle\Phi(\boldsymbol{Y})\boldsymbol{Y}^{\mathrm{H}}\rangle\right]\boldsymbol{W}_i \tag{4}$$

ここで $\langle\cdot\rangle$ は期待値演算, i は更新回数, η はステップサイズ. また $\Phi(\cdot)$ は以下の非線型関数である.

$$\Phi(\boldsymbol{Y}) = \frac{1}{1+\exp(-\boldsymbol{Y}^{(\mathrm{R})})} + j\,\frac{1}{1+\exp(-\boldsymbol{Y}^{(\mathrm{I})})} \tag{5}$$

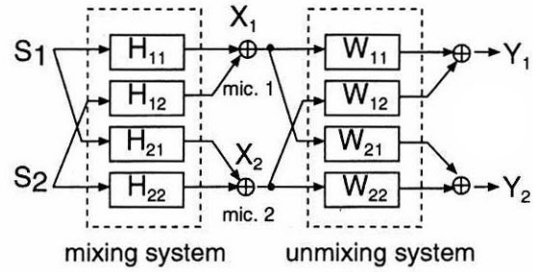

Fig. 1 BSS system configuration.

3. 実験

逆フィルタを推定するには, 部屋のインパルス応答長より長いフレームが必要だと考えられている. ここでは BSS においてどのようなフレーム長が適しているのかを実験で確かめる.

3.1 実験方法

使用した信号は, 実測したインパルス応答を計算機上で音声信号に畳み込んだものである. インパルス応答は, Fig.2 に示す環境で測定した. 部屋のインパルス応答長 T_R は, 0 ms, 150 ms ($P = 1200$), 300 ms ($P = 2400$) の 3 種類である. 音声信号は, 男女それぞれ 2 名ずつが発話した ASJ 研究用連続音声コーパスの中の 2 文であり, 長さは約 8 秒である. この冒頭 3 秒を用いて式 (4) における学習を行い, 8 秒全体の信号を分離した.

実験において, DFT のフレーム長 T を 32 から 2048 に変化させた. サンプリング周波数は 8 kHz, フレームシフトは $T/2$, 分析窓は Hamming 窓である. 周波数領域 BSS において問題となる permutation の問題については, 栗田らの手法で解決した [8].

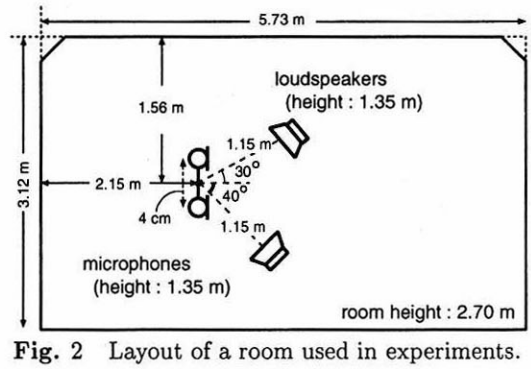

Fig. 2 Layout of a room used in experiments.

---

* Limitation of frequency domain Blind Source Separation for convolutive mixture of speech

† NTT Communication Science Laboratories

‡ Nara Institute of Science and Technology

3.2 実験結果

実験結果を Fig.3 に示す. 分離性能の評価には, *noise reduction rate* (NRR) を用いた. これは, 出力信号の S/N から入力信号の S/N を引いたものとして定義される. Fig.3 では便宜上, y_1, y_2 それぞれに対する NRR の平均値を示した.

Fig.3 (b) より, 残響時間 $T_R = 150$(ms) の時は DFT フレームサイズ $T = 256$ において, $T_R = 300$(ms) の時は $T = 512$ において, 最も良い分離性能が得られていることが分かる.

以上より, 部屋のインパルス応答長が長い場合においても, $P \ll T$ は不適切であり, より短いフレーム長での分析において最も良い分離性能が得られた.

4. 考察

式 (4) の学習が収束した時, W は

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} = \begin{bmatrix} c_1 & 0 \\ 0 & c_2 \end{bmatrix} \tag{6}$$

の解となる. ここで c_1, c_2 は 任意の複素数であり, これより (1)$\boldsymbol{W}$ は必ずしも $\boldsymbol{H}^{-1}$ にはならない. すなわち, 逆システムはできない. (2) 目的音に任意の残響を許す分離が行われる. の 2 点が言える.

また, 2 マイクロホンの BSS システムをマイクロホンアレイとして解釈すると, 妨害音の主に直接音方向に Null を向けて, 妨害音を消す動作を行うことが予想される. Fig.4 は直接音と残響音のパワー比を計算機上で変化させて分離を行った場合の分離性能である. 残響音の多い (直接音寄与率が低い) 場合に分離性能が低下している. これは, 主に直接音に対処する $\boldsymbol{W}$ の動作を示していると考える.

$\boldsymbol{W}$ は必ずしも混合過程の逆システムにならないため, 必ずしも $P \ll T$ を満たす必要はなく, 適切な分析フレーム長が存在するものと考える.

5. まとめ

本稿では, 長いインパルス応答を持つ環境での混合音声の分離において, 従来言われている $P \ll T$ の条件は不適切であることを実験で示した. これは, 周波数領域 BSS の枠組では, 主に妨害音の方向からの信号のみを考慮して分離を行っており, 部屋の逆フィルタを推定することができないためであると考える.

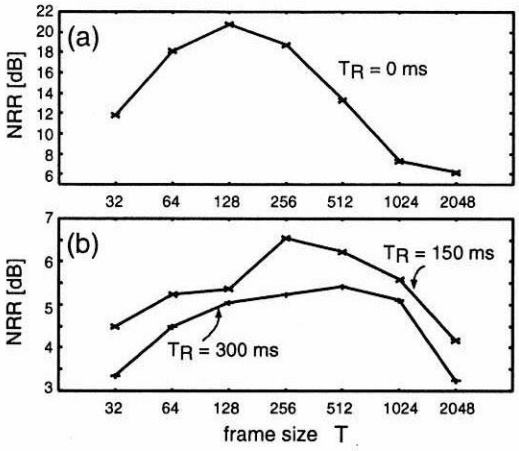

Fig. 3 Result of NRR for different frame sizes.

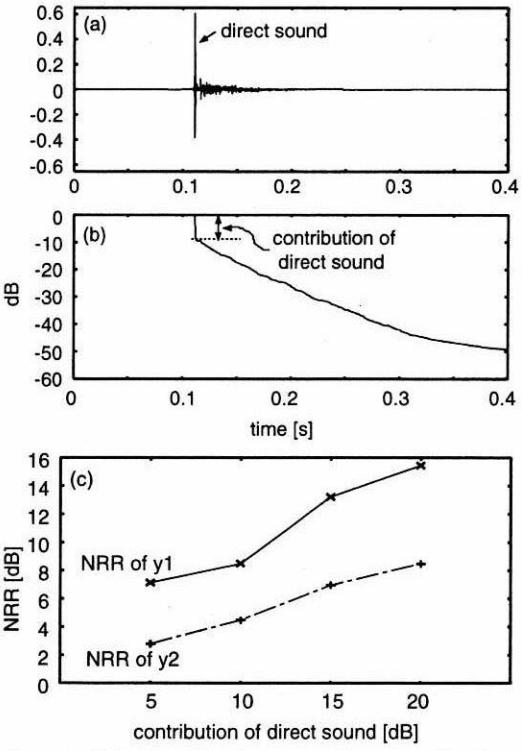

Fig. 4 Relationship between the contribution of the direct sound and separation performance. $T_R = 300$ms, $T = 256$. (a) Example of the impulse response and (b) Schroeder curve. (c) Performance.

参考文献

[1] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no.6, pp. 1129–1159, 1995.

[2] S. Haykin, "Unsupervised adaptive filtering," A wiley-interscience publication, 2000.

[3] T. W. Lee, "Independent component analysis -Theory and applications," Kluwer academic publishers, 1998.

[4] 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.

[5] M. Z. Ikram and D. R. Morgan, "Exploring permutation inconsistency in blind separation of speech signals in a reverberant environment," *Proc. ICASSP2000*, pp. 1041-1044, Jun. 2000.

[6] S. Araki, S. Makino, T. Nishikawa, and H. Saruwatari, "Fundamental limitation of frequency domain blind source separation for convolutive mixture of speech," *Proc. ICASSP2001*, May, 2001. (accepted)

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

[8] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," *Proc. ICASSP2000*, pp. 3140-3143, Jun. 2000.