観測信号ベクトル正規化とクラスタリングによる音源分離手法とその評価*

○荒木章子，澤田宏，向井良，牧野昭二 (NTT 研究所)

1 はじめに

ブラインド音源分離 (BSS) で用いられる主な方法としては独立成分分析による方法 (e.g.,[1]) と、信号のスパース性を利用する方法がある (e.g.,[2])。前者は普通、センサ数 $M \geq$ 信号数 N の場合のみを扱い、$M < N$ の場合へ直接適用することはできない。一方後者は、$M < N$ の場合にも適用できるが、従来手法は 2 センサの観測信号の振幅比や位相差の片方または双方を用いるため、3 個以上のセンサ情報を効果的に利用できない。3 個以上のセンサへの拡張も試みられてはいるが [3]、センサの直線配置を仮定しており、センサの位置補正の正確さに性能が左右される。更に例えば、Fig. 1 に示す [Ex.] のようにいろいろな方向に信号が存在する場合に対して、2 個あるいは数個の直線配置センサを用いると、Fig. 1 では S_1 と S_4 は鏡像の関係から分離が困難である。

そこで本稿では、センサと信号源の数や位置によらずに適用でき、3 個以上のセンサ情報を全て利用しながらブラインド音源分離を可能とする方法を提案する。また、本手法を残響下 ($T_{60} = 130$ ms) における音声信号の分離に適用したところ、非直線配置の 3 個以上のセンサを用いても高い分離性能を得られることを確認したので報告する。

2 問題設定

本稿では、信号の畳み込み混合問題を扱う。センサ j による観測信号 x_j は、$\mathrm{x}_j(n) = \sum_{k=1}^{N} \sum_{t=1} \mathrm{h}_{jk}(t)\mathrm{s}_k(n-t+1)$ $(j = 1, \cdots, M)$ とモデル化される。ここで N は信号数、 M はセンサ数、s_k は信号源 k からの信号、h_{jk} は信号源 k とセンサ j 間のインパルス応答である。BSS の目的は、原信号 s_k やインパルス応答 h_{jk} を知らずに原信号の推定 y_k を求めることである。

本稿では、時間領域で観測した信号 $\mathrm{x}_j(n)$ $(j = 1, \ldots, M)$ に短時間フーリエ変換 (STFT) を適用し、時間周波数領域にて信号を取り扱う。時間周波数領域における観測信号 $x_j(f, \tau)$ は近似的に

$$x_j(f, \tau) \approx \sum_{k=1}^{N} h_{jk}(f) s_k(f, \tau) \qquad (1)$$

と書ける。ここで $s_k(f, \tau)$ $(k = 1, \ldots, N)$ は原信号の STFT 結果、$h_{jk}(f)$ は信号源 k からセンサ j への周波数応答である。また全てのセンサにおける観測信号をまとめた $\mathbf{x} = [x_1, \ldots, x_M]^T$ を本稿では観測信号ベクトルと呼ぶ。

またここでは、信号がスパースであることを仮定する。これにより各時間周波数 (f, τ) において原信号のうちの 1 つ s_k のみが支配的である、すなわち式 (1) を $x_j(f, \tau) \approx h_{jk}(f) s_k(f, \tau)$ と近似できる。これは音声信号などで確認される。

3 従来法

スパース信号の分離は、各時間周波数 (f, τ) でどの s_k が支配的かを判定し、ある s_k が支配的な時間周波数 (f, τ) の観測のみを集めることで実現できる。

これを従来は、各時間周波数 (f, τ) で観測した信号の位置情報に基づき行っていた。具体的には、2 つのセンサのみで観測された信号の振幅比 $\frac{|x_1(f,\tau)|}{|x_2(f,\tau)|}$ や位相差 $\angle \frac{x_1(f,\tau)}{x_2(f,\tau)}$ (またはこれから推定される信号の到来方向) を特徴量としてクラスタリングを行い、それぞれのクラスタに属する時間周波数 (f, τ) の観測信号を再構成してそれぞれの分離信号を推定していた。

しかし従来法では、1 章で述べたように、3 個以上のセンサを利用する場合 (特に非直線配置の場合) への拡張が難しかった、あるいは、有効的ではなかった。

4 提案手法

そこで本稿では 3 つ以上のセンサを用いる時も全ての観測信号を有効に利用し、さらにシステムに自由なセンサ配置を許す分離手法を提案する。従来、2 センサで観測した信号を、振幅比や位相差などの 1 次元または 2 次元の情報に落としてクラスタリングしていたのに対し、この方法では多次元 (M 次元) の複素空間にてクラスタリングを行う所が異なる。

1. 観測信号の正規化： 観測信号ベクトル $\mathbf{x} = [x_1, \ldots, x_M]^T$ そのものを多次元複素空間でクラスタリングしようとしても、STFT のフレームの位置により位相が回ってしまっていたり周波数依存性があったりして、うまくクラスタを形成しない。そこで本手法では、まず観測信号ベクトル $\mathbf{x} = [x_1, \ldots, x_M]^T$ の各要素について、位相の正規化 (センサ J における

*Blind source separation method with observation vector normalization and clustering and its evaluation. by ARAKI, Shoko, SAWADA, Hiroshi, MUKAI, Ryo, MAKINO, Shoji (NTT Communication Science Laboratories, NTT Corporation.)

観測信号の位相で正規化) と周波数正規化を行う。

$$\bar{x}_j(f,\tau) \leftarrow |x_j(f,\tau)| \exp\left[\jmath \frac{\arg[x_j(f,\tau)/x_J(f,\tau)]}{4fc^{-1}d_{\max}}\right] \quad (2)$$

ここで、c は伝播速度である。また $d_{\max}$ はあるセンサ J と $^\forall j \in \{1,\ldots,M\}$ の距離の最大値であるが、正確な値が不明な場合は適当な大きめの値を用いることができる。

これにより、正規化された観測ベクトル $\bar{\mathbf{x}} = [\bar{x}_1,\ldots,\bar{x}_M]^T$ を得る。この正規化は、前述した観測ベクトルの位相不定性や周波数依存性を排除し、信号源の位置に関連する情報を顕在化させる (詳細は [4]) ので、次のクラスタリングで $\bar{\mathbf{x}}$ は信号源毎にクラスタを形成する。

2. クラスタリング： 次に、正規化された観測信号ベクトル $\bar{\mathbf{x}}(f,\tau)$ を、全ての (f,τ) について同時に、観測信号ベクトルと同じ次元 (M 次元複素空間) にてクラスタリングする。これにより信号源数と同じ N 個のクラスタ $C_1,\ldots,C_N$ が形成される。クラスタリングの詳細な手順と結果については [4] を参照されたい。

3. 信号分離： そしてそれぞれのクラスタ C_k が個々の信号に相当するので、

$$y_k(f,\tau) = \begin{cases} x_J(f,\tau) & \bar{\mathbf{x}}(f,\tau) \in C_k \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

により分離信号 $y_k(f,\tau)$ を得ることができる。

5 実験と結果

Fig. 1 に示す環境にて実験を行った。図に示すように、センサは非直線配置である。混合信号は、5 秒間の英語音声に、図 1 の環境にて計測したインパルス応答を畳み込んで得た。サンプリング周波数は 8kHz、フレーム長 L は 512、フレームシフトは、$L/8$〜$L/2$ とした。ここでは、正規化された観測信号ベクトル $\bar{\mathbf{x}}$ をさらにノルム正規化 $\bar{\mathbf{x}}(f,\tau) \leftarrow \bar{\mathbf{x}}(f,\tau) / ||\bar{\mathbf{x}}(f,\tau)||$ し、超球面上における k-means 法にてクラスタリングを行った。

分離結果を、信号対妨害音比 (SIR) 改善量と信号対歪比 (SDR) で評価した。また本手法では式 (3) の非線形処理により、可聴な非線型歪み (ミュージカルノイズ) が発生するため、これについても被験者 10 名による主観評価 (MOS 値) にて確認した。

結果を Tables 1, 2 に示す。センサ数 $M <$ 信号数 N の時や、センサ配置が不規則な時でも、高い分離性能を得ることができている。また [5] では、2 センサを用いたスパース性に基づく音源分離において、細かいフレームシフトがミュージカルノイズを抑制すること

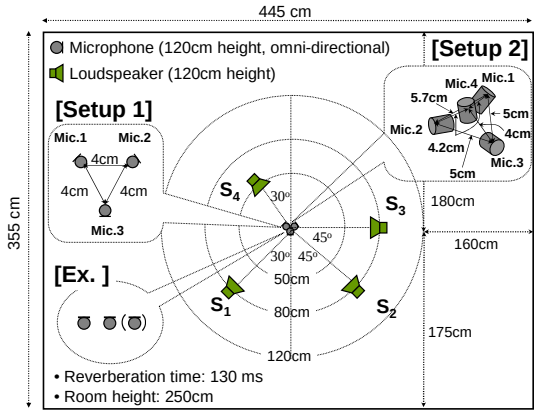

Fig. 1 実験条件

Table 1 [Setup 1] における分離結果,

		y_1	y_2	y_3	y_4	MOS
InputSIR$_i$		-7.4	-6.2	-6.0	-1.0	-
Shift $L/2$	SIR$_i$	16.4	10.5	14.2	11.1	1.9
	SDR$_i$	4.7	3.9	4.8	7.9	
Shift $L/4$	SIR$_i$	17.4	11.6	15.5	12.0	2.6
	SDR$_i$	5.5	4.7	5.6	8.7	
Shift $L/8$	SIR$_i$	17.9	11.8	15.9	12.3	2.9
	SDR$_i$	5.6	4.8	5.8	8.8	

Table 2 [Setup 2] における分離結果,

		y_1	y_2	y_3	y_4	MOS
InputSIR$_i$		-8.1	-5.3	-6.6	-0.6	-
Shift $L/2$	SIR$_i$	17.8	15.6	9.6	15.7	1.9
	SDR$_i$	4.2	5.7	3.6	11.5	
Shift $L/4$	SIR$_i$	19.0	16.5	9.9	17.0	2.7
	SDR$_i$	4.7	6.1	4.1	12.1	
Shift $L/8$	SIR$_i$	19.3	17.3	10.1	17.4	2.7
	SDR$_i$	4.9	6.6	4.2	12.2	

を確認しているが、本手法においても、フレームシフトを細かくすることで SIR や SDR を劣化させることなく MOS 値を向上できる (ミュージカルノイズを抑制できる) ことが分かった。

このような条件における分離を実現したのは本提案が初めてである。[6] にて分離音声の例を試聴できる。

参考文献

[1] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[2] Ö. Yılmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on SP*, vol. 52, no. 7, pp. 1830–1847, 2004.

[3] J. Rosca, C. Borss and R. Balan, "Generalized sparse signal mixing model and application to noisy blind source separation," *Proc. ICASSP2004*, vol. III, pp. 877–880, 2004.

[4] S. Araki, H. Sawada, R. Mukai and S. Makino, "A novel blind source separation method with observation vector clustering," in *Proc. IWAENC2005*, Sept. 2005.

[5] S. Araki, S. Makino, H. Sawada, and R. Mukai, "Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask," in *Proc. ICASSP2005*, vol. III, pp. 81–84, Mar. 2005.

[6] http://www.kecl.ntt.co.jp/icl/signal/araki/xcluster_fine.html