ノンパラメトリックベイズによる時間周波数領域における音声信号のブラインド音源分離

○柳楽浩平　高橋徹　尾形哲也　奥乃博 (京都大学大学院 情報学研究科)

1. はじめに

近年，よりよいヒューマンロボットインタラクションのために，実環境下でのロボット聴覚システムの研究が行われている [1, 2, 3]．ロボット聴覚システムにより，ロボットは同時に話す複数話者の定位をして話者方向を向いたり，混合音声信号を分離したり，音声認識したりできるようになる．ロボット聴覚システムはロボットに搭載された複数マイクから入力される信号を用いて，音源定位，音源分離，音声認識を行う．

実環境下でのロボット聴覚システムが満たすべき要求条件は以下の通りである．

1. 事前情報を用いないブラインド音源分離
2. Active な音源数が未知な状況での音源分離
3. 残響と反射への複素信号による対応 (図 1)

マイクロホンアレイシステムを用いた音源分離，特に，ビームフォーマは，音源方向をパラメータとして必要とする．音源方向求める音源定位には MUSIC (Multiple Signal Classification) 法がよく利用される．MUSIC 法は，音源数は事前に与えられた状態で音源定位を行う．さらに，各方向のインパルス応答などの事前情報の利用により定位性能が向上する．以上から，音源定位を用いた音源分離システムは条件 1，2 を満たさない．

音源定位情報などの事前情報を利用しない音源分離はブラインド音源分離 (BSS) と呼ばれる．よく知られている BSS の手法である独立成分分析 (ICA) は音源数がマイクロホン数と同じであると仮定して実環境での BSS を達成する [4]．よって ICA を用いたシステムでは条件 2 を満たさない．また，条件 3 は遠隔音声認識の分野で活発に議論されているトピックである．

本研究の目的は，音源数をパラメータとして与えずに，音源 ON/OFF 情報を同時に推定する，実環境下での BSS 手法開発である．ノンパラメトリックベイズに基づく BSS 手法として infinite sparse Factor Analysis (isFA) [5] が提案されており，音源 ON/OFF 情報の推定と音源分離を同時に達成するものの，分離が可能な信号は実数領域の信号のみであった．これは従来の isFA の能力が実信号の瞬時混合問題に限定されることを意味する．つまり，残響やマイク間での到来時間差などを扱うための畳込み混合モデルに対して従来の isFA をそのままでは適用できない．

本稿では実数 isFA を複素信号に適用することにより畳込み混合問題を解決する BSS 手法を提案する．鍵となるアイデアは 2 つあり，(1) 複素信号の実部・虚部への分解と，(2) 音源 ON/OFF 情報に基づく実数 isFA 出力信号の分類である．

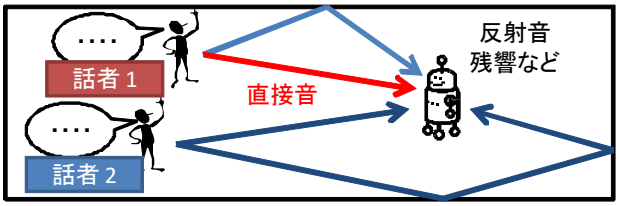

図 1 実環境でのロボットのマイクへの入力信号

2. 本研究で扱う問題

2·1 問題設定

まずはじめに，本研究で扱う問題についてまとめる．

入力	マイク D 本の信号 (K 音源の混合信号)
出力	K 個の音源信号
	各音源の ON/OFF 情報
仮定	$K \leq D$,
	残響は STFT の窓長より短い
	音源の位置は固定

K 個の音源からの混合信号を D 本のマイクを用いてとらえ，音源の混合過程などの事前情報を使わずに，もとの K 個の音声を分離する．

本稿では $K \leq D$ を仮定している．この仮定は，たとえ D 個以上の音源がその場に存在していたとしてもそのすべてが常に ON になっているわけではなく，ON になるのは高々D 個であることを意味する．言い換えると，各時間フレームで OFF になっている音源が数多く存在する．つまり，ON になっている音源数が分かっていないという状況とは矛盾しない．

2·2 本手法の概要

本手法の流れを図 2 に示す．最初に (1) D チャンネルの時間領域の入力信号を短時間フーリエ変換 (STFT) によって時間周波数領域での複素スペクトルに変換する．(2) 得られた D 個の複素スペクトルを実部と虚部に分け，$2D$ 個の実数信号として扱う．(2) から (6) は各周波数ビンごとに独立に処理を行う．次に (3) 白色化により無相関化を行い，(4) 実数領域 isFA を用いて $2K$ 個の出力信号を得る．$2K$ 個の信号はそれぞれ K 個の音源の実部と虚部に対応しているので，それらの組合せを探し出す必要がある．(5) isFA のもう一つの出力である ON/OFF 情報の類似度を利用して，$2K$ 個の出力信号を K 個のクラスに分類する．一つのクラスに含まれる二つの信号はそれぞれ同一音源の実部と虚部に対応している．最後に，(6) これらの信号から音源の実部と虚部を復元する．さらに，この時点で，出力信号の振幅は元信号の振幅とは異なっている．これは ICA においてスケーリング問題としてよく知られて

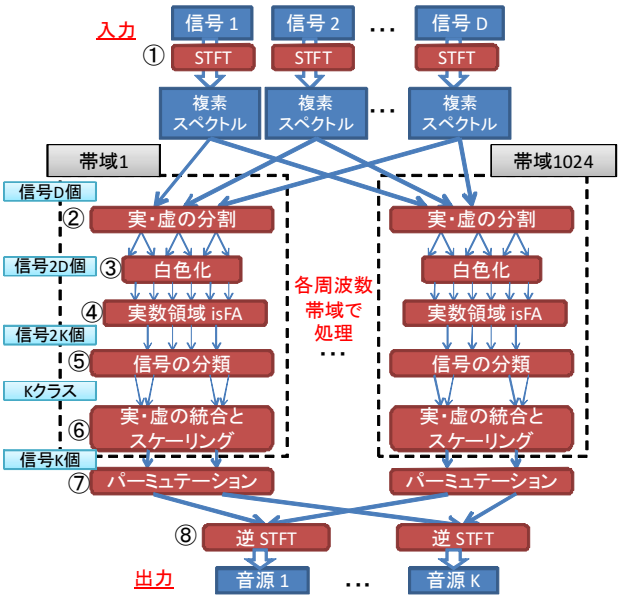

図 2 本手法の流れ

いる．これらの課題を projection back [6] を用いて解決する．

全周波数帯域での分離結果が得られた後，(7) 各周波数帯域でのパーミュテーション問題を解いて複素スペクトルを復元し，(8) 逆 STFT をして元音源の音声を復元する．

3. 本手法の詳細

3·1 前処理

STFT で得られる複素スペクトルに対して実数領域 isFA を適用するために，分離対象である音声の複素スペクトルの性質を考える．図 3 は JNAS200 文を結合した十分に長い音声を窓長 1024 点，シフト長 512 点で STFT した時の各周波数帯域の実部と虚部の相関係数を表したヒストグラムである．この図をみると多くの周波数帯域において 0.0 の近傍値となっているので，音声信号の実部と虚部は無相関といえる．つまり，これらの実部と虚部を独立した確率変数と考え，音声信号を実部と虚部に分け，それらを独立した信号とみなせる．

これにより，時間周波数領域での複素信号の瞬時混合問題

$$x_d^f(t) = \sum_{k=1}^{K} a_{d,k}^f s_k^f(t)$$

は，式 (1) のような実数領域のベクトル形式で表せる．

$$\begin{pmatrix} \mathrm{Re}\, x_d^f(t) \\ \mathrm{Im}\, x_d^f(t) \end{pmatrix} = \sum_{k=1}^{K} \begin{pmatrix} \mathrm{Re}\, a_{d,k}^f & -\mathrm{Im}\, a_{d,k}^f \\ \mathrm{Im}\, a_{d,k}^f & \mathrm{Re}\, a_{d,k}^f \end{pmatrix} \begin{pmatrix} \mathrm{Re}\, s_k^f(t) \\ \mathrm{Im}\, s_k^f(t) \end{pmatrix} \quad (1)$$

ただし，f, t はそれぞれ周波数ビンと時間フレームのインデックスを表し，$x_d^f(t)$ は d 番目の観測信号スペクトルを，$s_k^f(t)$ は k 番目の音源信号スペクトルを，$a_{d,k}^f$ は混合行列の要素を表す．これは複素数の積 $r = pq$ が

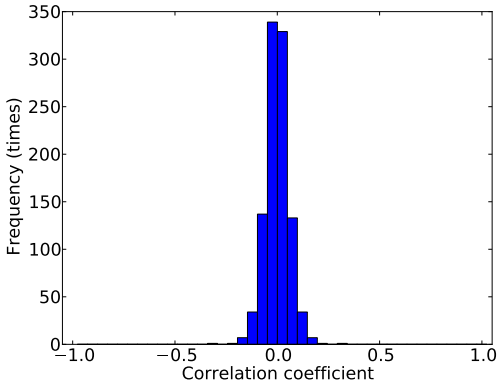

図 3 相関係数のヒストグラム：実部と虚部の独立性

以下の等価な実数の行列積で表せることによる．

$$\begin{pmatrix} \mathrm{Re}\, r \\ \mathrm{Im}\, r \end{pmatrix} = \begin{pmatrix} \mathrm{Re}\, p & -\mathrm{Im}\, p \\ \mathrm{Im}\, p & \mathrm{Re}\, p \end{pmatrix} \begin{pmatrix} \mathrm{Re}\, q \\ \mathrm{Im}\, q \end{pmatrix}$$

このようにして複素信号を実数に変換した後は各信号に白色化を施し無相関化する．これにより後の isFA の収束を速められる．

複素数を実部・虚部に分割して考えることにより，多くのパラメータは実数領域と比較して 2 倍の数のパラメータ推定となるが，混合行列の推定では，1 つの複素数が 4 つの実数となるため，パラメータ数が 4 倍となる事に注意しておく．

3·2 実数 isFA によるパラメータ推定

isFA は潜在的に無限個存在する実数信号の瞬時混合問題を解くノンパラメトリックベイズを利用したブラインド音源分離手法である．isFA では観測された混合信号から元音源信号が得られると同時に各音源の ON/OFF 情報が得られる．この音源 ON/OFF 情報は各時間フレームでその音源が active がどうかを表す二値行列で表される．

この実数領域 isFA は，各パラメータに事前分布を仮定し，潜在的に無限個の音源の存在を仮定した式 (2) のような生成モデルをもとに尤度関数 $P(X|A, Z, S)$ を求め，事前分布と尤度関数を掛けることで得られる事後分布からマルコフ連鎖モンテカルロ法に基づいてサンプリングを行い各パラメータを反復推定する．

$$X = A(Z \odot S) + E \quad (2)$$

X は観測信号，A は混合信号，Z は ON/OFF 情報，S は音源信号，E は雑音をそれぞれ表す．演算子 $\odot$ は要素ごとの積を表す．潜在的に無限個の音源を扱うために音源 ON/OFF 情報の事前分布には Indian Buffet Process (IBP) [7] を用い，音源信号，混合行列，雑音の各パラメータの事前分布は正規分布を仮定する．尤度関数や事後分布等の具体的な導出は Knowles ら [5] によって詳しく説明されている．

3·3 後処理

実数領域 isFA は各 K 個の音源信号の実部・虚部 $2K$ 個を順不同でバラバラに出力するので，各音源ごとにまとめる必要がある．図 4 の上半分に分類方法の概要

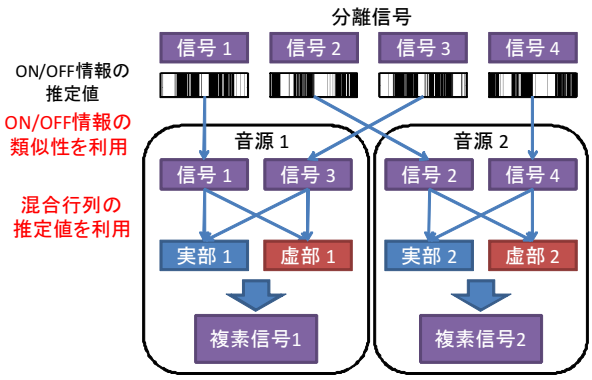

図 4 分離信号のクラス分けと実虚の統合

を示す．各クラスはそれぞれ一音源に対応しており，各クラスに含まれる 2 つの信号が音源の実部と虚部に対応している．

ここでは実数領域 isFA の出力の一つである各音源の ON/OFF 情報の類似性を利用してこれらの $2K$ 個の信号を分類する．同一信号の実部と虚部の推定値の ON/OFF 情報が類似している事に基づいて，$2K$ 個の信号中で ON/OFF 情報が最も近い 2 信号を同一音源の実部と虚部とみなし一つのクラスに分類する．

ここで，音源信号の実部と虚部の無相関性が 3.1 節で述べた ON/OFF 情報の類似性と矛盾しないことに注意しておく．音声信号の実部と虚部の信号波形の相関係数はほぼ 0 であるが，実部と虚部の ON/OFF 情報には関連がある．これは，もしある音源信号が active である場合には，当然その信号の実部も虚部も active となるため，両者は類似したものとなるからである．

二信号間の ON/OFF 情報の類似度は以下のようにして計算される．$\mathbf{z}_{s_1} = [z_{s_1,1}, \cdots, z_{s_1,N}]$，$\mathbf{z}_{s_2} = [z_{s_2,1}, \cdots, z_{s_2,N}]$ は 2 つの信号の ON/OFF 情報を表すとし，式 (3) の $\mathbf{z}_{xor}$ を計算する．

$$\mathbf{z}_{xor} = \mathbf{z}_{s_1} \otimes \mathbf{z}_{s_2} \tag{3}$$

ここで，演算子 $\otimes$ は排他的論理和（XOR）を表す．この $\mathbf{z}_{xor}$ の総和が小さいほど $\mathbf{z}_{s_1}$ と $\mathbf{z}_{s_2}$ は類似していると言える．この類似度を $2K$ 個の信号の全ペアに対して計算し，最小値となったもの同士を一クラスにまとめる．ただし，この段階ではどちらが実部でどちらが虚部かは不明である．この曖昧性はスケーリング問題を解決する際に同時に解決する．

信号の実部と虚部の曖昧性とスケーリング問題を同時に解くために projection back [6] を応用する．k 番目の推定信号波形 $\widehat{\mathbf{s}}_k = (\mathrm{Re}\ \widehat{s}_k, \mathrm{Im}\ \widehat{s}_k)^{\mathrm{T}}$ は，k 番目の真の音源波形 $\mathbf{s}_k = (\mathrm{Re}\ s_k, \mathrm{Im}\ s_k)^{\mathrm{T}}$ を用いて式 (4) のように書ける．

$$\widehat{\mathbf{s}}_k = \mathbf{P}_k \Lambda_k \mathbf{s}_k \tag{4}$$

ただし，

$$\mathbf{P}_k = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} or \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \Lambda_k = \begin{pmatrix} \lambda_{k_1} & 0 \\ 0 & \lambda_{k_2} \end{pmatrix}$$

である．$\mathbf{P}_k$ は実部虚部交換を表す行列，Λ_k はスケーリング行列を表す．ここで，混合モデルを考えること

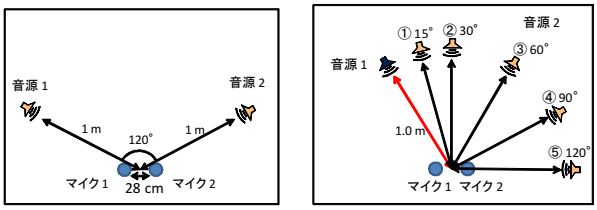

図 5 音源とマイクの配置 左:4.1 節, 右:4.2 節

表 1 実験の条件

音源数 K	2
マイク数 D	2
使用データベース	ASJ-JNAS
サンプリング周波数	16[kHz]
インパルス応答	無響室
STFT 窓長	64[ms]（1024 点）
STFT シフト長	32[ms]（512 点）
反復回数	300

で式 (5) が得られる．

$$\widehat{\mathbf{A}}_{dk}\widehat{\mathbf{s}}_k = \mathbf{A}_{dk}\mathbf{s}_k \tag{5}$$

$\mathbf{A}_{dk}$ と $\widehat{\mathbf{A}}_{dk}$ はそれぞれ混合行列の真値と推定値の d 番目の観測信号，音源 k に対応する部分を表す．式 (5) の右辺は d 番目の観測信号中の k 番目の音源信号を意味する．つまり，推定された信号に観測信号 d に対応する混合行列を掛けることで，振幅を d 番目の観測信号に合わせられ，実部と虚部の曖昧性も解消される．図 4 の下半分で実部と虚部の統合方法の図を示す．それに加えて，一つの注目する観測信号を決めておくことで全周波数帯域で観測信号の振幅をその観測信号の振幅に合わせられる．

パーミュテーション問題の解法には出力信号と元音声信号との相関を用いる．これは本手法の理想状態での分離性能を測定するためである．パーミュテーション問題の解法は澤田ら [8] などによって提案されているが，使用環境に依存しており，より汎用的な解法を求めて活発に研究が行われている．パーミュテーションを解いた後，逆 STFT で元音源信号を復元する．

4. 実験結果

4·1 ベース手法との比較

本節では無響室のインパルス応答が畳込まれた音声信号を用いた分離実験により本手法を評価する．実験条件は表 1 に，マイクや音源の配置は図 5 の左側に示した通りである．この実験には 17 発話を用いた．

図 6—9 はそれぞれ元の音源，混合音声，本手法による分離音声，ベース手法である実数領域 isFA による分離音声のスペクトログラムである．また、信号対雑音比（SNR）や雑音除去比（NRR）を用いた分離性能の評価結果を表 2 に示す．本手法はベースラインと比較して SNR で 4.45dB の改善がみられた．元音声からのケプストラム距離を用いた分離性能の評価結果を表 3 に示す．ケプストラム距離よりも SNR の方がより大きな改善が見られた．これらの結果から，本手法が畳込み混合の音声分離に効果的であることがわかった．

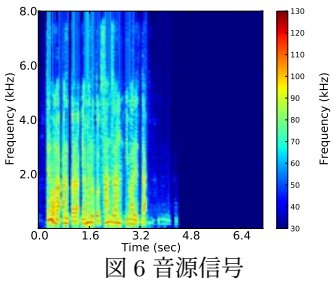

図 6 音源信号

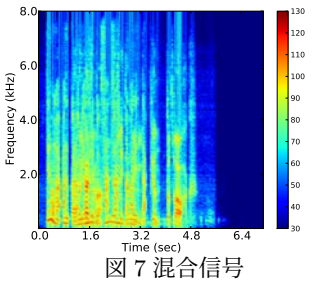

図 7 混合信号

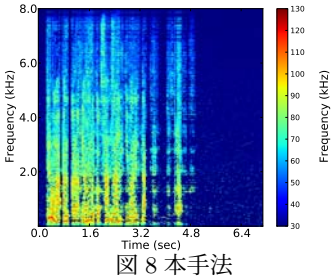

図 8 本手法

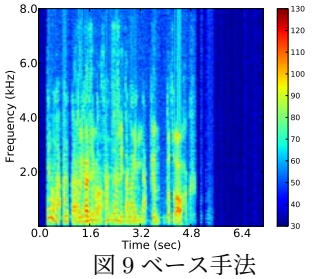

図 9 ベース手法

表 2　4.1 節の結果 (SNR)

	SNR	NRR
分離前	3.31 dB	—
ベース手法	3.10 dB	-0.21 dB
本手法	**7.55 dB**	**4.24 dB**

表 3　4.1 節の結果 (ケプストラム距離)

	分離前	分離後	改善
ベース手法	27.33 dB	29.01 dB	-1.68 dB
本手法	27.33 dB	26.46 dB	**0.87 dB**

表 4　4.2 節の結果 (SNR)

角度 (degree)	15	30	60	90	120
分離前 SNR (dB)	2.93	4.08	2.42	2.62	2.29
分離後 SNR (dB)	7.10	9.49	7.33	6.97	6.31
NRR (dB)	**4.17**	**5.41**	**4.91**	**4.55**	**4.02**

図 8 では周波数帯域によって，分離できている帯域とそうでない帯域がみられる．結果の図を見ると音源が OFF になるはずのところが ON になっている事やその逆も見られるため，ON/OFF 情報の推定ミスが原因であると考えられる．この推定ミスの要因として考えられるは二点あり，一つは反復回数が不足しているため分離が完了していない，もう一つは生成モデルである isFA では音源信号の事前分布を正規分布と仮定しているが，実際の音声信号は優ガウス的な分布となるため，モデルにあわないという可能性が考えられる．

4·2　音源間角度による影響

次に，本手法の空間分解能を評価するために，様々な音源間角度での分離実験を行った．実験条件は一つ目の実験と同様である．(表 1 参照) 図 5 の右側にはマイクと音源の配置が示されている．本実験で用いた音源間角度は 15 °,30 °,60 °,90 °,120 °の 5 通りである．本実験では 50 発話を用いた．

表 4 は SNR と NRR による評価の結果を表している．これをみると，本手法では様々な音源間角度に対して分離性能にあまり変化がみられないことが分かる．言いかえると，本手法は様々な角度に対してロバストな性能を達成していると言える．

5.　結論

本稿では音源 ON/OFF 情報の推定と畳込み混合音声の音源分離を同時に行うノンパラメトリックベイズに基づいたブラインド音源分離システムについて報告した．本手法は複素スペクトルを実部と虚部に分割し，実数領域 isFA を用いて時間周波数領域での複素信号の瞬時混合の分離を行う．

本手法によってベース手法と比較して SNR で 4.45dB，ケプストラム距離で 2.55dB の分離性能の改善を得た．さらに，様々な音源間角度に対して安定した性能となることがわかった．

今後の課題は，まず事前分布に優ガウスな分布を導入したり，反復回数を増やした実験を行うことで，帯域ごとの分離性能差の原因究明が必要である．また，isFA のモデル自身の複素拡張が考えることで，混合行列推定時の不要な推定を排除し，処理の高速化が期待できる．全周波数帯域で統一の ON/OFF 情報の利用により，パーミュテーション問題の解決が不要となる可能性も考えられる．そして，ロボットへの応用を考えると，本手法の実時間処理は喫緊の課題である．

謝辞

本研究の一部は，科研費基盤 (S)，JST-ANR BINAAHR, GCOE の支援を受けた．また，数多くの有益な助言をいただいた武田龍博士に感謝の意を表する．

参考文献

[1] F. Asano, H. Asoh, and T. Matsui. Sound source localization and signal separation for office robot “ jijo-2 ”. In *Multisensor Fusion and Integration for Intelligent Systems, 1999. MFI'99. Proceedings. 1999 IEEE/SICE/RSJ International Conference on*, pages 243–248. IEEE, 1999.

[2] K. Nakadai, T. Takahashi, H.G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino. Design and Implementation of Robot Audition System ”HARK” Open Source Software for Listening to Three Simultaneous Speakers. *Advanced Robotics*, 24(5–6):739–761, 2010.

[3] J.M. Valin, S. Yamamoto, J. Rouat, F. Michaud, K. Nakadai, and H.G. Okuno. Robust recognition of simultaneous speech by a mobile robot. *Robotics, IEEE Transactions on*, 23(4):742–752, 2007.

[4] H. Sawada, R. Mukai, S. Araki, and S. Makino. Polar coordinate based nonlinear function for frequency-domain blind source separation. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP'02)*, pages 1001–1004.

[5] D. Knowles and Z. Ghahramani. Infinite sparse factor analysis and infinite independent components analysis. *Independent Component Analysis and Signal Separation*, pages 381–388, 2007.

[6] N. Murata, S. Ikeda, and A. Ziehe. An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing*, 41(1-4):1–24, 2001.

[7] T. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. *Advances in Neural Information Processing Systems*, 18:475–482, 2006.

[8] H. Sawada, R. Mukai, S. Araki, and S. Makino. A robust and precise method for solving the permutation problem of frequency-domain blind source separation. *IEEE Trans. on Speech and Audio Processing*, 12(5):530–538, 2004.