4T-4

Permutation-free infinite ICA による 周波数領域ブラインド音源分離

柳楽 浩平　　大塚 琢馬　　奥乃 博
京都大学 大学院情報学研究科 知能情報学専攻

1. はじめに

音に含まれる様々な情報を抽出し，音環境を理解しようという音環境理解 (CASA) システム [1] では，様々な要素技術が不可欠である (図 1)．まず，マイクアレイに入力された混合音声信号から各音源のアクティビティの推定と各音声の分離を行い，その後，分離された音声の発話認識や音源の種別判定などを行う．この中で特に音源アクティビティ推定 (SAD) と音源分離は音源からの情報抽出の前処理として重要である．本研究の目的は SAD と音源分離の同時推定である．

実環境では，音源数や残響の影響を事前に与えることは難しいので，そのような事前情報を極力抑えても処理可能なブラインド音源分離 (BSS) が不可欠である．また，残響への対応のために周波数領域での処理を行う場合，従来の BSS の多くは周波数帯域ごとに独立して処理している結果，得られた分離音から同一音源の音を集めて全周波数領域の音を復元するためには Permuation 問題を解決する必要がある．

本稿では BSS・SAD・Permutation 問題を同時解決する Permutation-free infinite ICA (PF-IICA) を提案する．そのキーアイデアは，1) 全周波数帯域で統一的な音源アクティビティ，および 2) 周波数ごとの音源アクティベーション確率の導入である．

2. BSS 問題の設定と従来法の課題

本稿で扱う問題の設定を次にまとめる．

入力: D 本のマイクで観測される混合音
出力: K 個の音源のアクティビティと各音源信号
仮定: マイク数は音源の数以上存在する

SAD と BSS を同時に達成する手法の一つ周波数領域 infinte sparse factor analysis (FD-ISFA) [2] はノンパラメトリックベイズ理論に基づく推論により音源数の仮定せずに分離処理が可能である．ただし，FD-ISFA は各周波数帯域ごとに独立に処理を行うため，音源信号の復元のためには Permutation 問題の解決が不可欠となる．

3. ベイズ理論に基づく分離処理

本手法は Permutation 問題を回避するために，生成モデルに全帯域で共通する時間アクティビティ行列と全時間フレームで共通する各帯域ごとのアクティベーション確率行列を導入し，各帯域ごとの時間周波数アクティビティ行列の推定結果の出力順序を制御する．

3.1 分離処理の流れ

入力された混合音声信号に対して短時間フーリエ変換 (STFT) を適用し，周波数領域で処理を行う．各帯域毎に白色化処理を行った後，全帯域をまとめて分離処理を施す．分離された結果に対して Projection back を用いて分離信号の振幅を修正し，逆 STFT を適用して分離信号を得る．以下，本手法のベイズモデルについて説明する．

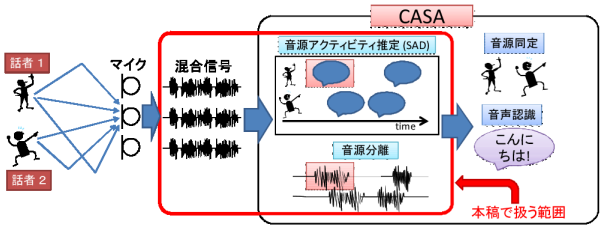

図 1: CASA システムの概要と本手法の位置づけ

3.2 生成モデル

K, D, F, T をそれぞれ音源数，マイク数，周波数帯域の数，信号の時間フレーム数とする．ある周波数帯域 f での瞬時混合モデルは以下のように表現できる．

$$\mathbf{X}_f = \mathbf{A}_f(\mathbf{Z}_f \odot \mathbf{S}_f) + \mathbf{E}_f \quad (f = 1, \cdots, F), \tag{1}$$

$\mathbf{X}_f$, $\mathbf{A}_f$, $\mathbf{Z}_f$, $\mathbf{S}_f$, $\mathbf{E}_f$ はそれぞれ観測信号 x_{fdt}，混合行列 a_{fdk}，各点の音源アクティビティ z_{fkt}，音源信号 s_{fkt}，雑音信号 ε_{fkt} をまとめた行列を表す．音源アクティビティ z_{kft} は二値変数であり，時刻 t フレーム目，f 番目の帯域で音源 k が鳴っている場合には $z_{kft} = 1$，そうでない場合には $z_{kft} = 1$ となる．演算子 $\odot$ は要素ごとの積を表す．

PF-IICA は F 組の周波数帯域を同時に処理する．全周波数帯域のアクティビティを束ねるため，以下のようなモデルを導入する．

$$z_{kft} = b_{kt}\phi, \qquad \phi \sim \mathcal{B}ern(\psi_{kf}), \tag{2}$$

ここで，$\mathcal{B}ern$ は Bernoulli 分布を，b_{kt} は音源 k の t フレーム目での全帯域で共通の音源アクティビティを，ψ_{kf} は音源 k の f 番目の帯域のアクティベーション確率を表す．$\mathbf{B}$ は b_{kt} を，Ψ は ψ_{kf} をまとめた行列である．

3.3 事前分布の設計

構築したモデルをベイズ推論するために，各変数ごとに与える適切な事前分布の設計について考える．まず，音源信号の事前分布は以下の 2 種類を考える．

$$\textbf{PF-IICA:} \quad s_{kft} \sim \mathcal{T}_C(\nu, \xi),$$
$$\textbf{PF-ISFA:} \quad s_{kft} \sim \mathcal{N}_C(0, 1)$$

$\mathcal{T}_C$, $\mathcal{N}_C$ は複素 Student-t 分布，一変量複素正規分布を表す．ν, ξ は自由度とスケールを表すパラメータである．複素 Student-t 分布に従う変数は Scale mixtures of Gaussian[3] を利用して生成する．Student-t 分布は優ガウス分布であり，音声信号の複素スペクトルの分布は優ガウス性を持つと分かっているので，事前分布としては PF-ISFA よりも PF-IICA の方が音声信号の生成モデルに適合することが期待される．その他の各変数の事前分布は以下のよう

Blind Source Separation in Frequency domain by Permutation-free Infinite ICA: *Kohei Nagira, Takuma Otsuka, and Hiroshi G. Okuno (Kyoto Univ.)*

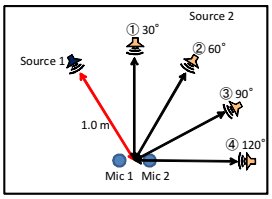

図 2: 音源とマイクの配置

表 1: 実験条件

音源数 K	2
マイク数 D	2
サンプリング周波数	16 kHz
STFT 窓幅	64 ms
STFT シフト幅	32 ms
データベース	JNAS
データ数	20

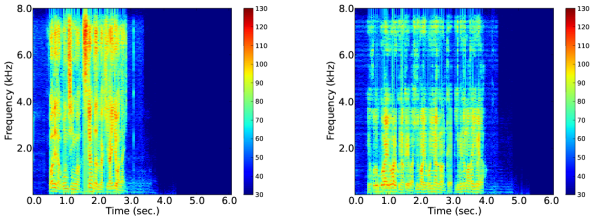

図 3: 音源信号のスペクトログラム

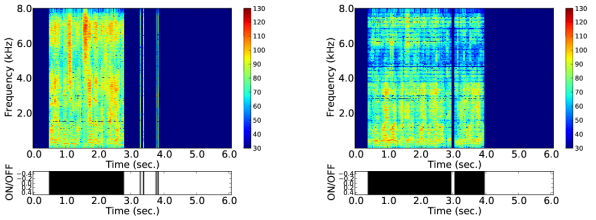

図 4: 分離信号のスペクトログラムと SAD 結果

に定めた．

$$\varepsilon_{ft} \sim \mathcal{N}_C(0, \sigma_\varepsilon^2 \mathbf{I}),\ \sigma_\varepsilon^2 \sim IG(p_\varepsilon, q_\varepsilon), \tag{3}$$

$$\mathbf{a}_{kf} \sim \mathcal{N}_C(0, \sigma_\mathbf{A}^2 \mathbf{I}),\ \sigma_\mathbf{A}^2 \sim IG(p_A, q_A), \tag{4}$$

$$\mathbf{B} \sim IBP(\alpha),\ \alpha \sim \mathcal{G}(p_\alpha, q_\alpha), \tag{5}$$

$$\Psi \sim Beta(\beta/K, \beta(K-1)/K). \tag{6}$$

$\mathbf{a}_{fk}$ は $\mathbf{A}_f$ の k 番目の行を表す．p_ε，q_ε，p_A，q_A，p_α，q_α，β はハイパーパラメータである．$\mathcal{G}$，IG，$Beta$，IBP はガンマ分布，逆ガンマ分布，ベータ分布，Indian buffet process[4] を表す．

3.4 尤度関数の設計と事後分布の推論

このモデルの観測信号 $\mathbf{X}$ だけから，音源信号 $\mathbf{S}$，雑音信号 $\mathbf{E}$，時間周波数アクティビティ $\mathbf{Z}$，混合行列 $\mathbf{A}$，全帯域の統一アクティビティ $\mathbf{B}$，各帯域のアクティベーション確率 Ψ の全てをベイズの定理に基づいて推定する．各変数は Metropolis-Hastings アルゴリズムにより更新される．事後分布は事前分布と尤度関数の積で求められる．モデルの尤度関数は以下の通りである．

$$P(\mathbf{X}|\mathbf{A},\mathbf{S},\mathbf{Z}) = \prod_{f=1}^{F} \frac{1}{(\pi\sigma_\varepsilon^2)^{TD}} \exp\left(-\frac{\mathrm{tr}(\mathbf{E}_f^{\mathrm{H}}\mathbf{E}_f)}{\sigma_\varepsilon^2}\right). \tag{7}$$

各データ点は独立同分布に従うと仮定している．PF-ISFA モデルの事後分布の導出は [5] でも示している．

4. 実験

実験では $\mathrm{RT}_{60} = 20,\ 600$ [ms] の 2 種類の残響環境での信号を用いた．マイクと音源の配置は図 2 の通りであり，Source2 は音源間角度が $30, 60, 90, 120$[deg] となる 4 パターンを試した．その他の条件は表 1 の通りである．比較する手法は PF-IICA，PF-ISFA と，FD-ISFA[2] の後に Permutation solver を実行したものの計 3 種類である．

実験結果の一例を図 3，4 に示す．図 3 は音源信号を表し，これら音源信号の混合信号に PF-IICA を適用した結果が図 4 である．スペクトログラムの下にあるグラフが SAD 推定の結果で，黒が ON，白が OFF を表す．これらの図から SAD と BSS を同時に達成し，Permutation 問題を回避していることが分かる．

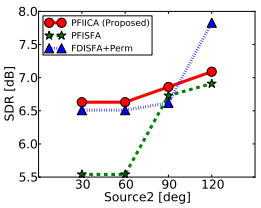

図 5: SDR (RT=20[ms])

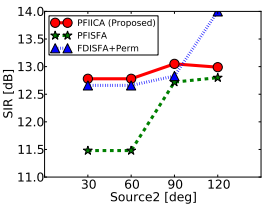

図 6: SIR (RT=20[ms])

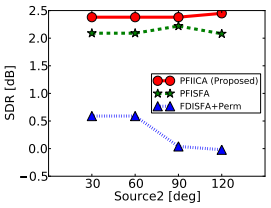

図 7: SDR (RT=600[ms])

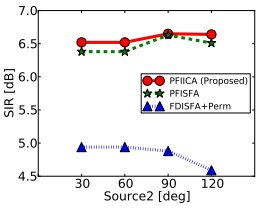

図 8: SIR (RT=600[ms])

図 5–8 は SDR と SIR[6] による評価結果を表す．SDR は音源分離の総合評価指標，SIR は妨害音の抑圧度の評価指標である．図より本手法が多くの場合で FD-ISFA を上回る性能をあげている．これは音源分離と Permutation 解決の逐次実行よりも統一的な枠組みによる同時推定の方が効果的であることを示唆している．さらに，PF-IICA の方が PF-ISFA よりもわずかに性能が向上している．これより音源の事前分布には優ガウス性の分布を用いた方がよいといえる．今回利用した優ガウス性の分布は Student-t 分布であった．一般に音声のパワーの分布はラプラス分布に従うことが知られているので，ラプラス分布の導入を今後検討していく．本手法は膨大な計算量となるため，モデルを簡略化し変数を減らすなど計算量削減を考えなければならない．

5. 結論

本稿では BSS と SAD と Permutation 解決を同時に達成する手法について述べた．ISFA のモデルに各帯域ごとの分離処理を統一的に扱うための変数を導入し，Permutation を回避する．実験により，本手法による Permutation の回避と，モデルにおける音源事前分布に Student-t 分布を用いることによる分離性能の向上を確認した．

謝辞：本研究の一部は，科研費基盤 (S)No.24220006, JST-ANR BINAAHR の支援を受けた．

参考文献

[1] D. Rosenthal, H. G. Okuno Computational auditory scene analysis. CRC press, 1998.

[2] K. Nagira, T. Takahashi, T. Ogata, H. G. Okuno. Complex extension of infinite sparse factor analysis for blind speech separation. *Latent Variable Analysis and Signal Separation*, pages 388–396, 2012.

[3] D. F. Andrews, C. L. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 99–102, 1974

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

[5] K. Nagira, T. Otsuka, H. G. Okuno. Infinite Sparse Factor Analysis for Blind Source Separation in Reverberant Environments *Statistical Techniques of Pattern Recognition*, pages 638–647, 2012.

[6] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. Rosca. First stereo audio source separation evaluation campaign: data, algorithms and results. *Independent Component Analysis and Signal Separation*, pages 552–559, 2007.