4U-1

Complex Infinite Sparse Factor Analysisによる周波数領域での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 知能情報学専攻

1. はじめに

音源分離技術は，ハンズフリー音声認識やロボット聴覚システムなどのために重要である．実環境では，マイクへの入力信号は複数話者の音声，反射音や残響などの混合音となる．この混合音からの各話者の音声認識には前処理として音源分離技術が必要となる．

音声信号の音源分離に対する主たる要求条件は3つある: (1) 事前情報を用いない分離 (利用環境を選ばない), (2) アクティビティの推定 (発話区間推定の詳細版), (3) 時間遅れ信号に対する頑健性 (反射音や残響への対応).

音源位置やマイク配置などの事前情報を用いない音源分離はブラインド音源分離 (BSS) と呼ばれる．Infinite sparse factor analysis (ISFA) [1] はノンパラメトリックベイズに基づいたBSS手法である．ISFAは音源分離とアクティビティの同時推定を行うため，(1)，(2) を満たす．しかし従来のISFAは反射音や残響などを含んだ混合音声をモデル化しておらず，その分離が困難であるため (3) を満たさない．本稿ではISFAを複素拡張したComplex ISFAを用いた，上記3点の要求条件を満たすBSS手法について報告する．

2. Complex ISFAによるBSS

2.1 BSSの問題設定

本稿で扱うブラインド音源分離問題は以下のように要約できる．

入力:　D 本のマイクに入力される K 音源の混合信号
出力:　元の K 個の音源信号とそれらのアクティビティ
仮定:　音源数 $K \leq$ マイク数 D

分離の際は音源方向やインパルス応答などの事前情報を用いない．

2.2 音声信号のBSS

反射音や残響など時間遅れ信号を考慮した音声信号の混合過程は時間領域での畳み込み混合モデルとして定式化できる. 畳み込み混合信号のBSSを解く際には短時間フーリエ変換 (Short time Fourier transform: STFT) がよく利用される．STFTにより，時間領域の畳み込み混合は周波数領域の複素信号の瞬時混合に変換される．STFTの後，各周波数ごとに独立に分離処理を行い，逆STFTによって元の音声信号を復元する．

2.3 Complex ISFAのモデル

Infinite sparse factor analysis [1] はノンパラメトリックベイズに基づくBSS手法である．K，D，N をそれぞれ音源数，マイクの数，音源信号の長さとする．ISFAの混合モデルは

$$\mathbf{X} = \mathbf{A}(\mathbf{Z} \odot \mathbf{S}) + \mathbf{E} \tag{1}$$

の式で表される．ここで，$\mathbf{Z} = [\mathbf{z}_1, \cdots \mathbf{z}_N]$, $\mathbf{X} = [\mathbf{x}_1, \cdots \mathbf{x}_N]$, $\mathbf{S} = [\mathbf{s}_1, \cdots \mathbf{s}_N]$, $\mathbf{E} = [\varepsilon_1, \cdots \varepsilon_N]$, $\mathbf{x}_t = [x_{1t}, x_{2t}, \cdots, x_{Dt}]^{\mathrm{T}}$ は時刻 t での混合信号ベクトル，$\mathbf{s}_t = [s_{1t}, s_{2t}, \cdots, s_{Kt}]^{\mathrm{T}}$ は音源信号ベクトル，$\varepsilon_t = [\varepsilon_{1t}, \varepsilon_{2t}, \cdots, \varepsilon_{Dt}]^{\mathrm{T}}$ はガウス性雑音のベクトルとする．また，$\mathbf{A}$ は $D \times K$ の混合行列，$\mathbf{z}_t = [z_{1t}, z_{2t}, \cdots, z_{Kt}]^{\mathrm{T}}$ は時刻 t での各音源のアクティビティを表す．z_{kt} は二値の変数であり，音源 k が時刻 t で音が鳴っている場合は $z_{kt} = 1$ となり，そうでない場合は $z_{kt} = 0$ となる．演算子 $\odot$ は要素ごとの積を表している．ISFAは観測信号 $\mathbf{X}$ のみを用いて音源信号 $\mathbf{S}$ とアクティビティ $\mathbf{Z}$，混合行列 $\mathbf{A}$ などの各パラメータを推定する．

2.4 事前分布の設計

各パラメータの事前分布は以下の通りである．

$$\varepsilon_t \sim \mathcal{N}_C(0, \sigma_\varepsilon^2 \mathbf{I}) \qquad \sigma_\varepsilon^2 \sim \mathcal{IG}(p_1, p_2), \tag{2}$$

$$s_{kt} \sim \mathcal{N}_C(0, 1), \tag{3}$$

$$\mathbf{a}_k \sim \mathcal{N}_C(0, \sigma_{\mathbf{A}}^2 \mathbf{I}) \qquad \sigma_{\mathbf{A}}^2 \sim \mathcal{IG}(p_3, p_4), \tag{4}$$

$$\mathbf{Z} \sim \mathrm{IBP}(\alpha) \qquad \alpha \sim \mathcal{G}(p_5, p_6). \tag{5}$$

ここで，$\mathbf{a}_k$ は $\mathbf{A}$ の k 番目の列，p_1，p_2，p_3，p_4，p_5，p_6 はハイパーパラメータである．$\mathcal{N}_C(\mu, \sigma^2)$ は平均 μ，分散 σ^2 の一変量複素正規分布

$$\mathcal{N}_C(x; \mu, \sigma^2) = \frac{1}{\pi\sigma^2} \exp\left(-\frac{|x-\mu|^2}{\sigma^2} \right) \tag{6}$$

を表す．$\mathcal{G}(b, \theta)$ と $\mathcal{IG}(b, \theta)$ は形状母数 b，尺度母数 θ のガンマ分布と逆ガンマ分布を表す．

$\mathrm{IBP}(\alpha)$ はパラメータ α の Indian buffet process (IBP) [2] を表す．IBPに基づくアクティビティの事前分布は $P(z_{kt} | \mathbf{z}_{-\mathbf{kt}}) = \frac{m_{k,-t}}{N}$ になる．ただし，$m_{k,-t} = \sum_{s \neq t} z_{ks}$ を表し，$\mathbf{z}_{-kt}$ は $\mathbf{z}_t$ の要素のうち z_{kt} を取り除いたものを表す．

2.5 尤度関数

Complex ISFAの尤度関数は以下のように表される．

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

ここで，$\mathbf{E} = \mathbf{X} - \mathbf{A}(\mathbf{Z} \odot \mathbf{S})$ であり，各時刻でのデータは独立同分布であると仮定している．

2.6 事後分布の推論

各パラメータの事後分布は，ベイズの定理に基づいて，上記の各事前分布と尤度関数の積から推論する．ここで得られた事後分布からMetropolis-Hastingsアルゴリズムに基づくサンプリングにより各変数を順に更新する．このパラメータ更新を繰り返して分離処理を進める．

2.7 後処理

周波数領域での分離では，スケーリング問題とパーミュテーション問題について考えなければならない．これらの問題は，本手法では各周波数帯域で独立に分離を行うために，各帯域での出力信号の振幅および出力順序を揃える必要があるというものである．

Blind Source Separation for Speech Signals in Frequency domain by Using Complex Infinite Sparse Factor Analysis: *Kohei Nagira, Toru Takahashi, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)*

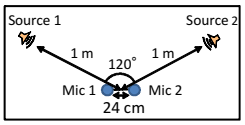

図 1: 音源とマイクの位置

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 [msec]
シフト幅	32 [msec]

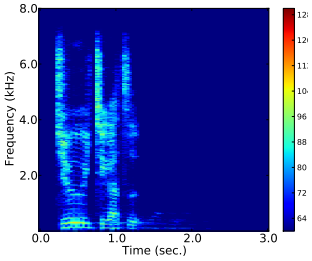

図 2: 音源信号

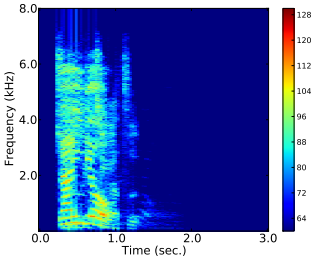

図 3: 混合信号

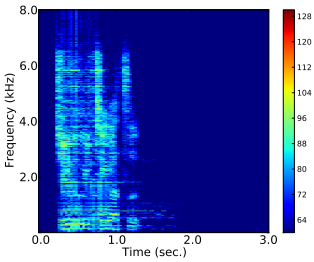

図 4: 本手法の分離結果

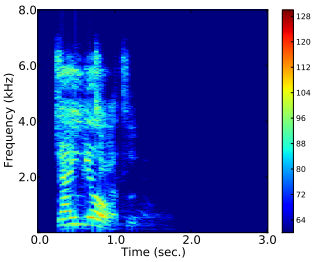

図 5: ベース手法の分離結果

スケーリング問題は projection back [3] により解決する．パーミュテーション問題は混合前の原信号を用いて，分離信号との相関をとることで解いている．これは ISFA の複素拡張自身の分離性能を評価するためである．

3. 実験結果

音声信号の分離実験により本手法の有効性を確認する．まず，本手法とベース手法の実数領域 ISFA[1] とを比較する．実験では，無響室，会議室 ($RT_{20} = 430$[ms]) の 2 種類の部屋で録音したインパルス応答の畳み込み混合音声を用いた．表 1 は実験条件を，図 1 はマイクと音源の配置を示している．ATR 音素バランス単語データベース中の 215 単語の発話を用いた．反復回数は 150 回である．各帯域の分離前には前処理として白色化を行っている．

図 2–5 のスペクトログラムはそれぞれ元音源，入力の混合信号，本手法による分離信号，ベース手法による分離信号を表す．SDR (Signal to Distortion Ratio)，ISR (Image to Spatial distortion Ratio)，SIR (Source to Interference Ratio)，SAR (Source to Artifacts Ratio) [4] を用いた定量的な評価も行った．いずれも数値が大きい方がよい分離性能を表す．

結果は表 2 の通りである．時間領域の ISFA が無響室程度の短い残響時間の畳み込み混合でさえ分離不可能であるのに対し，本手法は無響室，会議室での畳み込み混合音声も分離可能である事が分かる．

本手法は無響室環境でベース手法と比較して SDR の平均で 2.93[dB] 改善し，会議室環境でもベース手法に勝る性能が確認された．特に畳み込み混合の音声に対しては，SIR において本手法がベース手法に対し大きな改善が見られた．本手法の SAR の結果がベース手法と比較して悪化しているが，これは周波数領域での分離の際の STFT と逆 STFT によって生じる人工的なノイズによると考えられる．

表 2: ベース手法との比較実験結果:平均分離性能 [dB]

残響		SDR	ISR	SIR	SAR
無響室	分離前	-1.06	1.50	0.93	58.89
	ベース	-0.87	2.64	1.59	**35.89**
	本手法	**2.06**	**3.81**	**8.84**	2.75
会議室	分離前	-2.02	1.01	1.71	58.69
	ベース	-1.97	1.98	2.37	**36.08**
	本手法	**0.55**	**2.94**	**4.95**	3.16

時間領域の ISFA[1] が畳み込み混合音声の分離が困難であるのは，この手法は時間領域での瞬時混合を仮定しており，音声の混合のモデルである時間領域での畳み込み混合には対応できないからだと考えられる．それに対して，本手法は STFT 後の複素スペクトルの瞬時混合の分離が可能であるため，畳み込み混合音声に対しても有効である．

また我々が以前報告した，STFT 後の複素スペクトルの実部と虚部を分け，それぞれを別音源とみなして実数領域 ISFA を適用し，アクティビティの類似度に基づいて複素信号を復元するという手法 [5] との性能比較も行った．実験条件は先ほどと同じであり，データは JNAS 音素バランス文データベース中の 200 文を用いた．反復回数は 100 回である．無響室の場合，本手法が従来手法と比較して SDR で 0.33[dB]，SIR で 1.21[dB] の改善が，会議室の場合も，SDR で 0.44[dB]，SIR で 1.18[dB] の改善がそれぞれ見られた．これは，本手法では ISFA で直接複素信号を扱っているため，従来手法でのアクティビティの類似度に基づく複素信号の復元部分で別音源の実部と虚部が統合される可能性が解消された事に起因していると考えられる．

4. 結論

本稿では実環境での反射音，残響などの時間遅れ信号を考慮した畳み込み混合音声に対する BSS と各音源のアクティビティの同時推定手法について述べた．各周波数帯域ごとに ISFA の複素拡張を用いて複素瞬時混合を分離する．実験により，ベース手法と我々の従来手法に対して，本手法の分離性能の改善が確認された．

今後の課題は，音源のアクティビティの評価と，そのパーミュテーション問題の解法への応用や、ロボットへの応用で重要となるリアルタイム処理を目指した処理速度の向上などがある．

本研究の一部は，科研費 (S)，GCOE の支援を受けた．

参考文献

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

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

[3] 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.

[4] 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.

[5] 柳楽 浩平, 高橋 徹, 尾形 哲也, 奥乃 博. 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. **第 29 回日本ロボット学会学術講演会**, 3A2–5, 2011.