周波数・時間・到来方向のスパースネスを統合した近接分離最適化による IVA *

☆牛島隆裕 (大分大), 太刀岡勇気 (デンソー IT ラボ), 上ノ原進吾, 古家賢一 (大分大)

1 はじめに

音源分離とは、複数の音源が混合されて観測された信号から、単一の音源を含む信号を復元する技術である。音源分離が適用される場面として、雑音に頑健な音声認識を実現するために使用者の音声のみを取り出して雑音を抑圧する音声強調が例に挙げられる。

音源分離に用いられる手法は様々存在しており、例として独立成分分析 [1]、独立ベクトル分析 (IVA)[2]、同期同時対角化 [3]、独立低ランク行列分析 [4]、マルチチャネル非負値行列因子分解 [5] が存在している。これらの手法に用いる最適化手法として補助関数法による最適化が提案されており、高速でよい解に収束することが挙げられている。

近年、補助関数法以外の最適化手法を用いた音源分離手法として、近接分離最適化によるブラインド音源分離 (BSS) が提案されている [6]。近接分離最適化による BSS は音源の生成モデルを表す音源項と分離行列のスケールを調整する項を独立に設定することができる利点がある。音源項の近接写像を求めることが可能ならば、容易に音源項を変更した結果を検証可能である。[6] では例として音源項の生成モデルにラプラス球対称確率分布を用いたラプラス IVA が提案されている。また異なる応用として、音源分離に必須の前処理である白色化を行った場合、スパース構造が崩れて発生する雑音の影響を軽減するため、スパース構造を復元する重みを導入したスパース IVA[7] が提案されている。従来法の検討 [7] では音声データに対する有効性が示されているが、音源によっては分離性能が低下することを本稿の実験により示す。

本稿では、種々の音源での分離性能の向上のため、従来のスパース IVA で考えられている周波数方向のスパース性の他に、時間方向と到来方向のスパース性も考慮したスパース IVA を提案する。そしてそれらの複数のスパースネスを組み合わせることによって、従来のスパース IVA に比べて様々な音源に対応したスパースネス統合 IVA の提案を行う。実験では従来法で検討されてきた音声データのほかに音楽データに対し検討を行い提案法の有効性を示す。

2 近接分離最適化による BSS

2.1 音源分離の概要

音源信号 $\boldsymbol{s}_{ij}$、観測信号 $\boldsymbol{x}_{ij}$ 及び分離信号 $\boldsymbol{y}_{ij}$ を短時間フーリエ変換したスペクトログラムにより表す。

$$\boldsymbol{s}_{ij} = (s_{ij,1}, s_{ij,2}, ..., s_{ij,N})^T \tag{1}$$

$$\boldsymbol{x}_{ij} = (x_{ij,1}, x_{ij,2}, ..., x_{ij,M})^T \tag{2}$$

$$\boldsymbol{y}_{ij} = (y_{ij,1}, y_{ij,2}, ..., y_{ij,N})^T \tag{3}$$

ただし、各インデックスを周波数ビン i $(1 \leq i \leq I)$、時間フレーム j $(1 \leq j \leq J)$、分離対象となる音源の ID を n $(1 \leq n \leq N)$、及び観測チャネル番号を m $(1 \leq m \leq M)$ として表す。また、^{T} は転置を表す。混合系の音源分離では、音源から観測マイクへの混合の過程を表す混合行列 $A_i \in C^{M \times N}$ を用いて観測信号、音源との関係を次式で表せる。

$$\boldsymbol{x}_{ij} = A_i \boldsymbol{s}_{ij} \tag{4}$$

$M = N$ かつ A_i が正則である場合、$W_i = A_i^{-1}$ となる分離行列 $W_i \in \mathbb{C}^{N \times M}$ が存在し、観測信号から音源信号への分離を次式で表せる。

$$\boldsymbol{y}_{ij} = W_i \boldsymbol{x}_{ij} \tag{5}$$

IVA では各音源が互いに独立でありかつ何らかの確率分布に従って生成される場合、観測信号 $\boldsymbol{x}_{ij}$ から分離行列 W_i を推定できる。分離行列 W_i の推定は以下の最小化問題を解くことで達成される。

$$\underset{\{W_i\}_{i=1}^I}{\text{Minimize}} \quad \sum_{i=1}^{I} \mathcal{P}\left(W_i \boldsymbol{x}_{ij}\right) - \sum_{i=1}^{I} \log \left|\det\left(W_i\right)\right| \tag{6}$$

第 1 項 $\mathcal{P}\left(W_i x_{ij}\right)$ は音源の生成モデルを表す確率分布から導かれる実数値関数であり以下音源項と呼ぶ。第 2 項は分離行列のスケールを調整するための正規化項である。

2.2 近接分離最適化ブラインド音源分離

近接分離最適化ブラインド音源分離は, 式 (6) の最小化問題を主-双対近接分離法によって解いていく手法である。Fig. 1 に最適化の流れを示している。まず、式 (6) を主-双対近接分離法が適用できるような形に変形すると以下の式のようになる。

$$\underset{\mathbf{w}}{\text{Minimize}} \quad \mathcal{P}(\mathrm{X}\mathbf{w}) + \mathcal{L}(\mathbf{w}) \tag{7}$$

$$\mathcal{L}(\mathbf{w}) = -\sum_{i=1}^{I} \sum_{m=1}^{M} \log \sigma_m(\mathfrak{M}(\mathbf{w})_i) \tag{8}$$

行列 X は観測信号のスペクトログラム $\boldsymbol{x}$ を $IMJ \times IMN$ の形状に拡張したブロック対角行列である。ベクトル $\mathbf{w}$ は $I \times M \times N$ の形状の分離行列 W を $1 \times IMN$

*IVA based on proximal splitting optimization integrating sparseness of frequency, time and direction of arrival by Takahiro Ushijima (Oita University), Yuuki Tachioka (Denso IT Laboratory), Shingo Uenohara, Ken'ichi Furuya (Oita University)

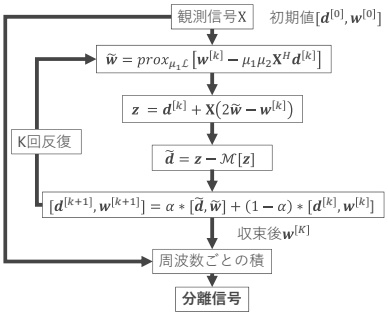

Fig. 1　近接分離最適化音源分離の流れ

の形にベクトル化したものである。また、$\mathfrak{M}(\mathbf{w})$ は $\mathbf{w}$ を分離行列 W の形に戻すための処理であり、W_i と一致する。$\sigma_m(\mathfrak{M}(\mathbf{w})_i)$ は W_i の m 番目の特異値である。$\mathcal{L}(\mathbf{w})$ は分離行列のスケールの正規化項を近接写像が適用できる形に変形したものであり、これは特異値の近接写像の計算を行うことで求めることが可能である。$\mathcal{P}(\mathrm{X}\mathbf{w})$ は音源の生成モデルを表す音源項である。

2.3　スパース IVA[7]

スパース IVA は近接分離最適化の枠組みにおいて、音源項を工夫することで、白色化により失われた周波数方向のスパース性を復元させる手法である。これは混合 $L_{2,1}$ ノルムと L_1 ノルムの和の近接写像 $\mathrm{prox}_{\lambda_1\|\cdot\|_{2,1}+\lambda_2\|\cdot\|_1}[\mathbf{z}]$ にスパース構造を復元する重みを適用している。以下の式で定式化され Fig. 1 中の $\mathcal{M}[\boldsymbol{z}]$ に使用される。

$$(\mathcal{M}[\boldsymbol{z}])_{ijn}^{Freq} = \zeta_{ijn}^{z,\kappa} \times \Xi_\kappa\left[\left(1 - \frac{\lambda_1}{\left(\Sigma_{i=1}^{I}\left(\Theta_\eta^{Freq}[\boldsymbol{x}]\right)_i \left|\zeta_{ijn}^{z,\kappa} z_{ijn}\right|^2\right)^{1/2}}\right)_+\right] \quad (9)$$

$\zeta_{ijn}^{z,\kappa}$ が L_1 ノルムの近接写像、$(\Xi_\kappa[\cdot])$ の引数部分が $L_{2,1}$ ノルムの近接写像に対応している。ただし、$(\cdot)_+ = \max(0,\cdot)$、$\lambda_1, \lambda_2$ は閾値パラメータ、$\kappa \geq 1$ はバイアス低減処理 $(\Xi_\kappa[z])_{ijn}$ に用いるパラメータであり、以下の式のように表される。

$$(\Xi_\kappa[\boldsymbol{z}])_{ijn} = (\kappa z_{ijn} / \max_{ijn}\{z_{ijn}\}) \quad (10)$$

$\zeta_{ijn}^{z,\kappa}$ は要素毎の firm thresholding を表している。

$$\zeta_{ijn}^{z,\kappa} = \Xi_\kappa\left[(1 - \lambda_2 / z_{ijn})_-\right] \quad (11)$$

ただし、$(\cdot)_- = \min(1,\cdot)$ である。Θ_η^{Freq} は周波数帯域毎のスパース性を復元するためのピークを計測するための重みである。この重みは、白色化によって励起した雑音部分の周波数帯域のパワーを軽減する効果がある。

$$\left(\Theta_\eta^{Freq}[\boldsymbol{x}]\right)_i = \Upsilon_\eta\left[\frac{\left(\Sigma_{m=1}^{M}\Sigma_{j=1}^{J}\left|x_{ijm}\right|^2\right)^{\frac{1}{2}}}{\Sigma_{m=1}^{M}\Sigma_{j=1}^{J}\left|x_{ijm}\right|}\right] \quad (12)$$

$\Upsilon_\eta[\cdot]$ は L_1 正規化を表しており以下の形で定式化される。

$$\Upsilon_\eta[\boldsymbol{\xi}] = \boldsymbol{\xi}_\eta / \left(\|\boldsymbol{\xi}_\eta\|_1 / I\right) \qquad \xi_\eta = (\boldsymbol{\xi} - \eta)_+$$

$\boldsymbol{\xi}$ は Υ_η への引数、$\boldsymbol{\xi}_\eta$ は $\boldsymbol{\xi}$ を閾値処理した後の値となる、元々スパース性が低い雑音の支配的な周波数帯域をカットする効果がある。

3　提案法

3.1　概要

スパース IVA では周波数ビン毎のスパース性の計算を行なっている。提案法では周波数方向の他に二つのスパース性の導入を行う。一つ目は信号のパワーの時間変動のスパース性、二つ目は音源の到来方向に対するスパース性である。

3.2　時間方向のスパース性

時間方向のスパース性は以下のように、各時間フレームにおける周波数ビンでの観測信号のパワーの総和を計算している。

$$\left(\Theta_\eta^{Time}[\boldsymbol{x}]\right)_j = \Upsilon_\eta\left[\frac{\left(\Sigma_{m=1}^{M}\Sigma_{i=1}^{I}\left|x_{ijm}\right|^2\right)^{\frac{1}{2}}}{\Sigma_{m=1}^{M}\Sigma_{i=1}^{I}\left|x_{ijm}\right|}\right] \quad (13)$$

提案法では、Θ_η^{Time} を式 (9) の Θ_η^{Freq} の代わりに用いる。

3.3　到来方向のスパース性

到来方向はマイク間の位相差によって求める。到来方向に基づくバイナリマスクは以下のように求めている。

$$\mathcal{D}_{i,j,n} = \begin{cases} 1 & (|\,\theta_{ij} - \theta_n| > \theta_c) \\ \epsilon & else \end{cases} \quad (14)$$

ただし $\mathcal{D}_{i,j,n}$ は音源 n を強調するバイナリマスク、$\theta_{i,j}$ は周波数ビン i と時間フレーム j における推定到来角度、θ_n は音源 n の到来角度、θ_c は到来方向を許容するための定数である。後の実験では $\theta_{i,j}$ のヒストグラムの最頻値から到来方向 θ_n を求めている。

到来方向のスパース性を扱う際、本稿では二通りのマスクを提案する。一つ目は到来方向分離マスクで式 (9) に代わって以下のような重みを用いる。

$$\Theta_{\eta,n}^{DOAsep} = \mathcal{D}_{i,j,n} \quad (15)$$

これは各音源の到来方向に対するバイナリマスクを用いた重みであり、式 (9) に使用することで音源 n ごとに強調するようなマスク $(\mathcal{M}[z])_{ijn}^{DOAsep}$ を生成する。二つ目は到来方向共通マスクで以下のような重みを用いる。

$$\Theta_\eta^{DOAshare} = \max_n \mathcal{D}_{i,j,n} \quad (16)$$

これは直接音を通し、それ以外の音を排除するマスクである。例として 2 方向から音源が到来する場合、

到来方向分離マスクでは 1 方向の音源のみを強調し他の到来方向の音源を排除するマスクを音源数だけ持つが、到来方向共通マスクでは到来する 2 方向をどちらとも通す 1 つのマスクを持っている。これは、直接音の分離は音源分離アルゴリズムに任せ、残響成分や雑音成分をマスクにより除去することを狙っている。

3.4　スパースネスの統合

従来の周波数方向、提案した時間方向と到来方向のスパース性をそれぞれ音源項として式 (7) に加え、複数の音源項がある場合の近接分離最適化による音源分離 [6] として統合することで多様な音源への対応を試みる。スパースネスを統合した際の最適化式を以下に表す。

$$\underset{\mathbf{w}}{\text{Minimize}} \quad \lambda_{Freq}\mathcal{P}^{Freq}(\mathrm{X}\mathbf{w}) + \lambda_{Time}\mathcal{P}^{Time}(\mathrm{X}\mathbf{w}) + \lambda_{DOA}\mathcal{P}^{DOA}(\mathrm{X}\mathbf{w}) + \mathcal{L}(\mathbf{w}) \quad (17)$$

ただし各スパース IVA の音源項を $\mathcal{P}^{[\cdot]}$、音源項を適用する大きさを表す重みを $\lambda_{[\cdot]}$、$[\cdot]$ の添え字は各スパースネスを表し、周波数方向は $Freq$、時間方向は $Time$、到来方向は DOA で表す。

4　評価実験

4.1　スパースネスの違いによる分離性能の比較

4.1.1　実験条件

本実験では 2 音源の混合音を 2 チャネルで音源分離することで精度を確かめる。分離する音源は人間同士の話し声から構成される音声データと、楽器や歌声で構成される音楽データを対象とした。音声データの信号は SiSEC[8] の UND の女声データの内、−50 度と 45 度から到来する 2 つの信号を合成して MixtureA,−10 度と 15 度から到来する信号を合成して MixtureB として使用する。音楽データの信号は RWCP 実環境音声・音響データベースのインパルス応答 (E2A[10]) に、SiSEC で提供されている楽曲 Bearlin-Roads と楽曲 Ultimate-Nz-Tour の各楽器の音源データを Table 2 のように畳み込んで生成した。初期値は分離行列 $\mathbf{W}$ に単位行列、双対変数 $\mathbf{d}$ にゼロベクトルを与えた。スパース IVA において閾値 η は ξ の最大値を超えるような値に設定しまった場合、重み Θ_η の値がすべて零になってしまう恐れがあるため、ここでは ξ の値を利用することで, 閾値を決定する。具体的には、$\xi_i(i = 1, 2, \cdots, I)$ を昇順に並べ替えた $\hat{\xi}_i(i = 1, 2, \cdots, I)$ を導入し以下のように閾値 η を求めている。$\eta = \hat{\xi}_k$、$k = I * \delta$。音源分離性能の評価には、SDR[9] を用いた。値が高いほど分離性能が高いことを表す。

4.1.2　実験結果

分離実験の結果を Fig. 2 に示す。比較対象は従来法となるラプラス IVA、周波数方向のスパース IVA、そして提案法となる時間方向のスパース IVA、到来方向の分離マスクと共通マスクとしている。横軸は分離対象の音源、縦軸は分離後の信号の平均 SDR を表している。従来の周波数方向のスパース IVA は、音声データに対しては有効であるものの、音楽データの 2 曲に対しては一転して分離精度が他手法と比べて低くなった。これは楽器音同士のパワーの偏りがあるにもかかわらず閾値処理を行ったため、片方のパワーの小さい音源を削ってしまったためと考えられる。時間方向マスクが最も改善量が大きい。到来方向マスクは分離マスクに比べて共通マスクの方が良好な結果が出た。到来方向分離マスクは各音源をそれぞれ強調するような特徴を持っているが、マスクが強調しようとする音源と音源分離アルゴリズムにより分離された分離音の不整合により、分離が進まない信号があった。そのため到来方向共通マスクの方が優れている結果となった。

Table 1　分離処理に用いるパラメータ

サンプリング周波数	16kHz
フレームサイズ	2048
シフトサイズ	1024
μ_1、μ_2	1，1
λ_1,λ_2	2 ,0.01
κ	1.1
収束速度α	1
閾値パラメータδ	0.5

Table 2　使用した音楽データ

ID	Author/Song	Part 1	Part2
1	bearlin1	Piano(90°)	Vocal(10°)
2	bearlin2	Vocal(10°)	Ambient(150°)
3	bearlin3	Piano(90°)	Ambient(150°)
4	ultimate1	Guitar(90°)	Synth(10°)
5	ultimate2	Synth(10°)	Drum(150°)
6	ultimate3	Guitar(90°)	Drum(150°)

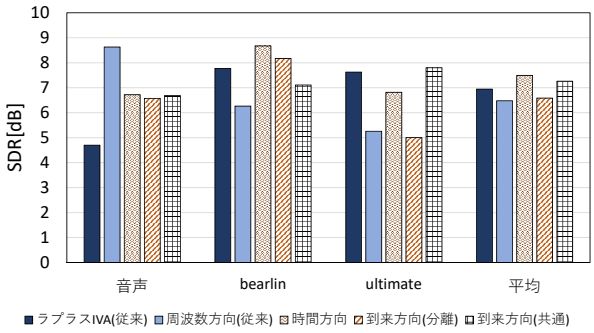

Fig. 2　スパースネスの違いによる分離性能の比較

4.2　スパースネス統合 IVA の実験

4.2.1　実験条件

上記のスパース性の違いによるスパース IVA の分離性能の比較を行った結果から、音声データに対しては従来法の周波数方向のスパース IVA、音楽デー

Table 3　実験に用いた音楽データ

重み＼統合マスク	音声強化	音楽強化
λ_{Freq}	2	0.3
λ_{Time}	0.3	2
λ_{DOA}	0.3	0.3

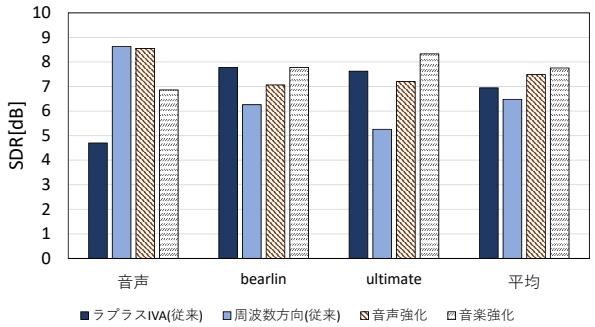

Fig. 3　スパースネスを統合した際の分離結果

タに対しては時間方向のスパース IVA が有効であることが分かった。よって 3.4 節の手法により 3 つのスパースネスを統合した手法により分離実験を行う。ただし到来方向に対するスパース性については、安定性から到来方向共通マスクのみを用いる。統合には各音源項を適用する大きさを調整するための重みを設定する。音声データに対しては周波数方向のスパース IVA の重みを大きくした音声強調マスク、音楽データに対しては時間方向のスパース IVA の重みを大きくした音楽強化マスクを用いて検討する。各音源項を統合する重みは Table 3 に表している。

4.2.2　実験結果

分離実験の結果を Fig. 3 に示す。比較対象は従来法と音声強化マスク、音楽強化マスクである。縦軸は分離後の信号の平均の SDR を表している。音声強化マスクについては、従来の周波数方向のスパース IVA と比べて音声データの SDR がほぼ変わらない性能で、楽器音が平均で約 1.4dB ほど向上した。音楽強化マスクについては、音楽データに強いラプラス IVA に対しても分離性能の若干の向上がみられ、さらに音声データでは約 2.2dB 分離性能が向上した。音声強化マスクや音楽強化マスクの双方で重みが最も大きいマスクの得意な音源に対する性能を維持しつつ、その他の音源に対して分離性能の向上ができた。これは異なるスパースネスを統合することで、単一のスパースネスだけでは考慮されなかった別の特徴を持つスパースネスの利用ができたためと考えられる。

5　まとめ

本稿では、まずラプラス IVA に関して周波数方向のスパース性を導入したスパース IVA は音声データには有効だが、音楽データでは性能が落ちることを実験によって示した。そこで多様な音源に対応するため、時間方向と到来方向の 2 つのスパース性を導入する手法を提案した。分離実験の結果、平均的には時間方向のスパース性の導入が最も有効であることが分かった。また、周波数方向、時間方向、到来方向の 3 つのスパース性を統合し重みを工夫したスパースネス統合 IVA を用いることで音声・音楽の両データの性能が改善した。

参考文献

[1] T.-W. Lee, “Independent Component Analysis-Theory and Applications,” Norwell, MA: Kluwer, 1998.

[2] I. Lee *et al.*,“Fast fixedpoint independent vector analysis algorithms for convolutive blind source separation,” Signal Processing 87(8), 2007

[3] 澤田 宏, “同期同時対角化によるブラインド信号分離,” 第 32 回信号処理シンポジウム講演論文集, pp.332-337, 2017

[4] D. Kitamura *et al.*, “Dtermined blind source separation unifying independent vector analysis and nennegative matrix factorization,” IEEE/ACM Transaction on Audio, Speech, and Language Processing, 24(9):1626-1641, 2016

[5] H. Sawada *et al.*, “Multichannel Extensions of Non-Negative Matrix Factorization with Complex-Valued Data,” IEEE Trans. ASLP, vol.21, no.5, pp.971-982, 2013.

[6] Kohei Yatabe *et al.*, “Determined blind source separation via proximal splitting algorithm,” Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018), pp. 776-780, 2018

[7] Kohei Yatabe *et al.*, ”Time-frequency-masking-based determined BSS with application to sparse IVA,” Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019), pp. 715-719, 2019.

[8] S. Araki, *et al.*, ”The 2011 signal separation evaluation campaign (SiSEC2011): - audio source separation -,” in Proc. Int. Conf. Latent Variable Anal. Signal Separation,pp. 414-422 2012.

[9] E. Vincent *et al.*, “First Stereo Audio Source Separation Evaluation Campaigh: Data Algprithm and Results,” Independent Component Analysis and Signal Separation(Springer, Bearlin, 2007), pp.552-559.

[10] RWCP: “実環境音声・音響データベース (RWCP-SSD)” 音声資源コンソーシアム, http://research.nii.ac.jp/src/ RWCP-SSD.html, 閲覧日:2019/12/19.