振幅の滑らかさを考慮した凸最適化による調波打撃音分離 *

◎赤石夏輝, 山田宏樹, 矢田部浩平 (農工大)

1 まえがき

調波打撃音分離（HPSS）は，信号の正弦波成分と打撃音成分を分離する処理である。HPSS の手法の一つに，振幅スペクトログラムの滑らかさを考慮した最適化に基づく手法がある [1]。この手法では，一般的に非凸となる HPSS の問題を複素スペクトログラムの振幅のみを考慮することで凸最適化問題として定式化している。しかし，位相の情報を扱えていないため，分離の性能に上限が存在する。本稿では，位相の情報を失わないようにしながら，凸最適化問題で定式化された HPSS を提案し，性能の向上を確認した。

2 振幅の滑らかさを考慮した HPSS

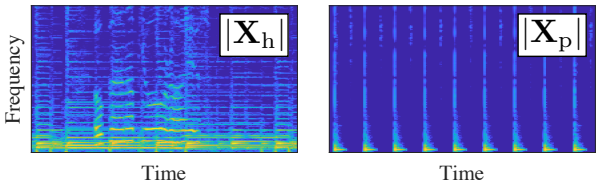
音響信号には正弦波成分と打撃音成分が含まれることが多い。正弦波成分（図–1 左）と打撃音成分（図–1 右）の振幅スペクトログラムはそれぞれ時間方向と周波数方向に滑らかであるという性質がある。この振幅スペクトログラムの滑らかさの異方性に基づいて，正弦波成分と打撃音成分の振幅がそれぞれ時間・周波数方向に滑らかさを強調することで HPSS を行う手法が提案されている [1]。

この手法は，複素スペクトログラムの振幅の時間・周波数方向の差分のエネルギーを最小化する以下の最適化問題を解いている。

$$\begin{aligned} \underset{\mathbf{H},\mathbf{P}}{\text{Minimize}} \quad & \lambda_{\mathrm{h}} \left\| \mathscr{D}_\tau(\mathbf{H}) \right\|_{\mathrm{F}}^2 + \lambda_{\mathrm{p}} \left\| \mathscr{D}_\omega(\mathbf{P}) \right\|_{\mathrm{F}}^2 \\ \text{subject to} \quad & \mathbf{H} \geq \mathbf{O}, \quad \mathbf{P} \geq \mathbf{O}, \quad \mathbf{H} + \mathbf{P} = |\mathbf{X}|^2 \end{aligned} \tag{1}$$

ただし $\|\cdot\|_{\mathrm{F}}$ はフロベニウスノルム，$|\cdot|$ は要素ごとの絶対値，$\mathbf{H}, \mathbf{P}$ はそれぞれ正弦波成分と打撃音成分の振幅スペクトログラムに対応する最適化変数，$\mathscr{D}_\tau, \mathscr{D}_\omega$ はそれぞれ時間，周波数方向の差分をとる作用素，$\lambda_{\mathrm{h}}, \lambda_{\mathrm{p}} > 0$ はそれぞれ正弦波成分と打撃音成分のバランス調整に関するパラメータ，$\mathbf{O}$ は要素が全て 0 の行列，$\geq$ は要素ごとの不等号，$\mathbf{X}$ は元信号の複素スペクトログラムである。

この手法は振幅スペクトログラム $\mathbf{H}, \mathbf{P}$ を最適化変数とすることで，HPSS を凸最適化問題として定式化している。これによって性能が安定する一方で，振幅のみを扱うため位相の分離ができていない。このとき，分離信号を得る際には元信号の位相を用いるが，この操作によって分離の性能が制限されてしまう。



図–1 正弦波成分（左）と打撃音成分（右）の振幅スペクトログラムの例。対数スケールで表示している。

3 提案手法

我々は，位相の情報を失わないように複素スペクトログラムを最適化変数としつつ，凸最適化問題として定式化された手法を提案する。ここで，式 (1) を複素の最適化変数を持つように自然に拡張すると

$$\begin{aligned} \underset{\mathbf{X}_{\mathrm{h}},\mathbf{X}_{\mathrm{p}}}{\text{Minimize}} \quad & \lambda_{\mathrm{h}} \left\| \mathscr{D}_\tau(\mathbf{H}) \right\|_{\mathrm{F}}^2 + \lambda_{\mathrm{p}} \left\| \mathscr{D}_\omega(\mathbf{P}) \right\|_{\mathrm{F}}^2 \\ \text{subject to} \quad & \mathbf{H} = |\mathbf{X}_{\mathrm{h}}|^2, \quad \mathbf{P} = |\mathbf{X}_{\mathrm{p}}|^2 \\ & \mathbf{X}_{\mathrm{h}} + \mathbf{X}_{\mathrm{p}} = \mathbf{X} \end{aligned} \tag{2}$$

のようになる。ただし，$\mathbf{X}_{\mathrm{h}}, \mathbf{X}_{\mathrm{p}}$ はそれぞれ正弦波成分と打撃音成分の複素スペクトログラムに対応する最適化変数である。このとき，$\mathbf{H} = |\mathbf{X}_{\mathrm{h}}|^2, \mathbf{P} = |\mathbf{X}_{\mathrm{p}}|^2$ が非凸制約であるため，目的関数は凸であっても全体として非凸問題となってしまう。そこで，提案法はこの非凸制約を回避するための凸関数を用いることで，凸最適化問題で定式化された HPSS を実現する。

提案手法では，正弦波成分と打撃音成分の複素スペクトログラムのそれぞれの振幅に関する重みを補助的な変数として用いる。その重みが時間・周波数方向にそれぞれ滑らかになるようにすることで，複素スペクトログラムの振幅の構造を間接的に変化させて所望の分離信号を得る。このとき，複素スペクトログラムとその振幅に関する重みを結びつける関数を以下のように置く。

$$\varphi(\mathbf{X}, \mathbf{W}) = \sum_{n=1}^{N} \sum_{m=1}^{M} \phi(X[m,n], W[m,n]) \tag{3}$$

ただし N, M はそれぞれ時間フレーム数と周波数ビン数，ϕ は $|\cdot|^2/2 + 1/2$ のパースペクティブ関数 [2] で，

$$\phi(x, w) = \begin{cases} \frac{|x|^2}{2w} + \frac{w}{2} & (w > 0) \\ 0 & (x = 0 \ \text{ and } \ w = 0) \\ \infty & (\text{otherwise}) \end{cases} \tag{4}$$

で定義される。φ は，重み $\mathbf{W}$ が持つ構造を複素スペクトログラムの振幅 $|\mathbf{X}|$ に反映できる [2]。また，ϕ は凸関数であり，その和である φ は凸関数となる [2]。

* Anisotropic-smoothness-based harmonic/percussive source separation via convex optimization. By Natsuki AKAISHI, Koki YAMADA, and Kohei YATABE (Tokyo University of Agriculture and Technology).

Algorithm 1 提案手法のアルゴリズム

Input: $\mathbf{X}, \mathbf{X}_{\mathrm{h}}^{[0]}, \mathbf{X}_{\mathrm{p}}^{[0]}, \mathbf{W}_{\mathrm{h}}^{[0]}, \mathbf{W}_{\mathrm{p}}^{[0]}, \mathbf{U}_{\mathrm{h}}^{[0]}, \mathbf{U}_{\mathrm{p}}^{[0]}, \mathbf{V}_{\mathrm{h}}^{[0]}, \mathbf{V}_{\mathrm{p}}^{[0]}$, $\lambda_{\mathrm{h}}, \lambda_{\mathrm{p}}, \mu, \nu, \rho$

for $i = 0, 1, 2, \ldots$ **do**

$(\mathbf{X}_{\mathrm{h}}^{[i+\frac{1}{2}]}, \mathbf{W}_{\mathrm{h}}^{[i+\frac{1}{2}]}) = \mathrm{prox}_{\nu\varphi}(\mathbf{X}_{\mathrm{h}}^{[i]} - \nu\mathbf{U}_{\mathrm{h}}^{[i]}, \mathbf{W}_{\mathrm{h}}^{[i]} - \nu\mathscr{D}_{\tau}^{*}(\mathbf{V}_{\mathrm{h}}^{[i]}))$

$(\mathbf{X}_{\mathrm{p}}^{[i+\frac{1}{2}]}, \mathbf{W}_{\mathrm{p}}^{[i+\frac{1}{2}]}) = \mathrm{prox}_{\nu\varphi}(\mathbf{X}_{\mathrm{p}}^{[i]} - \nu\mathbf{U}_{\mathrm{p}}^{[i]}, \mathbf{W}_{\mathrm{p}}^{[i]} - \nu\mathscr{D}_{\omega}^{*}(\mathbf{V}_{\mathrm{p}}^{[i]}))$

$\mathbf{Y}_{\mathrm{h}} = \mathbf{U}_{\mathrm{h}}^{[i]} + \mu(2\mathbf{X}_{\mathrm{h}}^{[i+\frac{1}{2}]} - \mathbf{X}_{\mathrm{h}}^{[i]})$

$\mathbf{Y}_{\mathrm{p}} = \mathbf{U}_{\mathrm{p}}^{[i]} + \mu(2\mathbf{X}_{\mathrm{p}}^{[i+\frac{1}{2}]} - \mathbf{X}_{\mathrm{p}}^{[i]})$

$(\mathbf{U}_{\mathrm{h}}^{[i+\frac{1}{2}]}, \mathbf{U}_{\mathrm{p}}^{[i+\frac{1}{2}]}) = (\mathbf{Y}_{\mathrm{h}}, \mathbf{Y}_{\mathrm{p}}) - \mu P_{\mathbf{X}}(\mathbf{Y}_{\mathrm{h}}/\mu, \mathbf{Y}_{\mathrm{p}}/\mu)$

$\mathbf{S}_{\mathrm{h}} = \mathbf{V}_{\mathrm{h}}^{[i]} + \mu\mathscr{D}_{\tau}(2\mathbf{W}_{\mathrm{h}}^{[i+\frac{1}{2}]} - \mathbf{W}_{\mathrm{h}}^{[i]})$

$\mathbf{S}_{\mathrm{p}} = \mathbf{V}_{\mathrm{p}}^{[i]} + \mu\mathscr{D}_{\omega}(2\mathbf{W}_{\mathrm{p}}^{[i+\frac{1}{2}]} - \mathbf{W}_{\mathrm{p}}^{[i]})$

$\mathbf{V}_{\mathrm{h}}^{[i+\frac{1}{2}]} = \mathbf{S}_{\mathrm{h}} - \mu\,\mathrm{prox}_{(\lambda_{\mathrm{h}}/\mu)\|\cdot\|_{\mathrm{F}}^{2}}(\mathbf{S}_{\mathrm{h}}/\mu)$

$\mathbf{V}_{\mathrm{p}}^{[i+\frac{1}{2}]} = \mathbf{S}_{\mathrm{p}} - \mu\,\mathrm{prox}_{(\lambda_{\mathrm{p}}/\mu)\|\cdot\|_{\mathrm{F}}^{2}}(\mathbf{S}_{\mathrm{p}}/\mu)$

$(\mathbf{X}_{\mathrm{h}}^{[i+1]}, \mathbf{W}_{\mathrm{h}}^{[i+1]}) = \rho(\mathbf{X}_{\mathrm{h}}^{[i+\frac{1}{2}]}, \mathbf{W}_{\mathrm{h}}^{[i+\frac{1}{2}]}) + (1-\rho)(\mathbf{X}_{\mathrm{h}}^{[i]}, \mathbf{W}_{\mathrm{h}}^{[i]})$

$(\mathbf{X}_{\mathrm{p}}^{[i+1]}, \mathbf{W}_{\mathrm{p}}^{[i+1]}) = \rho(\mathbf{X}_{\mathrm{p}}^{[i+\frac{1}{2}]}, \mathbf{W}_{\mathrm{p}}^{[i+\frac{1}{2}]}) + (1-\rho)(\mathbf{X}_{\mathrm{p}}^{[i]}, \mathbf{W}_{\mathrm{p}}^{[i]})$

$(\mathbf{U}_{\mathrm{h}}^{[i+1]}, \mathbf{V}_{\mathrm{h}}^{[i+1]}) = \rho(\mathbf{U}_{\mathrm{h}}^{[i+\frac{1}{2}]}, \mathbf{V}_{\mathrm{h}}^{[i+\frac{1}{2}]}) + (1-\rho)(\mathbf{U}_{\mathrm{h}}^{[i]}, \mathbf{V}_{\mathrm{h}}^{[i]})$

$(\mathbf{U}_{\mathrm{p}}^{[i+1]}, \mathbf{V}_{\mathrm{p}}^{[i+1]}) = \rho(\mathbf{U}_{\mathrm{p}}^{[i+\frac{1}{2}]}, \mathbf{V}_{\mathrm{p}}^{[i+\frac{1}{2}]}) + (1-\rho)(\mathbf{U}_{\mathrm{p}}^{[i]}, \mathbf{V}_{\mathrm{p}}^{[i]})$

end for

提案手法では，凸関数 φ を用いて HPSS を

$$\begin{aligned} \underset{\mathbf{X}_{\mathrm{h}},\mathbf{W}_{\mathrm{h}},\mathbf{X}_{\mathrm{p}},\mathbf{W}_{\mathrm{p}}}{\text{Minimize}} \quad & \lambda_{\mathrm{h}}\|\mathscr{D}_{\tau}(\mathbf{H})\|_{\mathrm{F}}^{2} + \lambda_{\mathrm{p}}\|\mathscr{D}_{\omega}(\mathbf{P})\|_{\mathrm{F}}^{2} \\ & + \varphi(\mathbf{X}_{\mathrm{h}}, \mathbf{W}_{\mathrm{h}}) + \varphi(\mathbf{X}_{\mathrm{p}}, \mathbf{W}_{\mathrm{p}}) \quad (5) \\ \text{subject to} \quad & \mathbf{X}_{\mathrm{h}} + \mathbf{X}_{\mathrm{p}} = \mathbf{X} \end{aligned}$$

のように定式化する。ただし，$\mathbf{W}_{\mathrm{h}}, \mathbf{W}_{\mathrm{p}}$ はそれぞれ正弦波成分と打撃音成分の振幅に関する重みである。凸関数 φ は式 (2) における非凸制約 $\mathbf{H} = |\mathbf{X}_{\mathrm{h}}|^2, \mathbf{P} = |\mathbf{X}_{\mathrm{p}}|^2$ を置き換えたものとみなすことができる。これによって，この問題は凸最適化問題となる。この問題では，重み $\mathbf{W}_{\mathrm{h}}, \mathbf{W}_{\mathrm{p}}$ を時間・周波数方向にそれぞれ滑らかにし，関数 φ によってその構造を各成分の複素スペクトログラム $\mathbf{X}_{\mathrm{h}}, \mathbf{X}_{\mathrm{p}}$ の振幅に反映させる。

式 (5) に主双対分離法を適用すると Algorithm 1 が得られる。ただし，$\mathscr{D}_{\tau}^{*}, \mathscr{D}_{\omega}^{*}$ はそれぞれ $\mathscr{D}_{\tau}, \mathscr{D}_{\omega}$ の随伴作用素である。制約条件に関する射影 $P_{\mathbf{X}}$ は

$$P_{\mathbf{X}}(\mathbf{X}_{\mathrm{h}}, \mathbf{X}_{\mathrm{p}}) = (\mathbf{X}_{\mathrm{h}}, \mathbf{X}_{\mathrm{p}}) - (\mathbf{X}_{\mathrm{h}} + \mathbf{X}_{\mathrm{p}} - \mathbf{X})/2 \quad (6)$$

のように与えられる。$\mathrm{prox}_{\nu\varphi}$ および $\mathrm{prox}_{(\lambda_{\mathrm{h}}/\mu)\|\cdot\|_{\mathrm{F}}^{2}}$ はそれぞれ $\nu\varphi$ と $(\lambda_{\mathrm{h}}/\mu)\|\cdot\|_{\mathrm{F}}^{2}$ の近接作用素であり，これらは解析的に求められる [2]。

4 実験

提案手法を振幅の滑らかさに基づく HPSS の従来手法（AS）[1] と比較した。また，各手法によって得られる振幅および位相の妥当性を評価するために，真の信号の振幅および位相を各分離信号に用いた場合の性能も評価した。データセットとして，Fraunhofer Institute for IDMT 内の 10 曲の楽曲データ [3] を用いた。各データのサンプリング周波数は 44.1 kHz である。短時間フーリエ変換では 4096 サンプルのハン窓を 1024 サンプルずつシフトした。10 個のデータの SDR，SIR，SAR の中央値によって性能を評価した。従来手法のパラメータは最良のものに設定した。Algorithm 1 の反復回数は 1000 回とし，$\lambda_{\mathrm{h}} = \lambda_{\mathrm{p}} = 1$, $\nu = 0.5$, $\mu = 0.2$, $\rho = 1.99$, $\mathbf{X}_{\mathrm{h}}^{[0]} = \mathbf{X}_{\mathrm{p}}^{[0]} = \mathbf{X}$ とした。

表–1 10 曲に対するスコアの中央値。各条件で良いスコアの方をボールドで示している。

	Harmonic			Percussive		
	SDR	SIR	SAR	SDR	SIR	SAR
(A) AS [1]	7.82	10.34	12.39	−4.78	**1.82**	−1.31
(B) Proposed	**9.49**	**12.08**	**12.49**	**−2.84**	−0.61	**5.57**
(A)+TrueAmp	14.46	**22.33**	15.93	0.97	**16.81**	1.20
(B)+TrueAmp	**16.75**	22.24	**18.80**	**8.65**	15.12	**9.72**
(A)+TruePhase	9.92	23.87	10.05	−2.78	**20.87**	−2.73
(B)+TruePhase	**11.91**	**26.75**	**11.98**	**0.93**	19.34	**1.05**

表–1 に性能評価の結果を示す。(A) が従来手法，(B) が提案手法である。+TrueAmp は各手法によって得られた位相と真の分離信号の振幅を用いた条件，+TruePhase は各手法によって得られた振幅と真の分離信号の位相を用いた条件をそれぞれ表している。なお，従来手法によって位相は得られないため，(A)+TrueAmp は元信号の位相と真の分離信号の振幅を用いた条件となっている。まず表–1 の上段部で (A) と (B) を比べたとき，正弦波成分と打撃音成分の両方に対して，提案手法の SDR と SAR が従来手法よりも大きく向上した。このことから，位相の情報を失わずに処理する提案手法は，従来手法に比べてアーチファクトを軽減できることが示唆される。また，表–1 の中段部および下段部の，真の信号の位相または振幅をそれぞれ用いた条件でも，提案手法の SDR と SAR が従来手法よりも向上した。これらから，提案手法は従来手法よりも品質の良い振幅と位相をそれぞれ得ることができると考えられる。

5 むすび

本稿では，位相の情報を失わずに，凸最適化によって振幅の滑らかさを考慮した HPSS を行う手法を提案した。今後は正弦波成分と打撃音成分の振幅と位相の事前情報をさらに用いた手法を検討する。

参考文献

[1] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka and S. Sagayama, “Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram,” *Eur. Signal Process. Conf. (EUSIPCO)*, pp. 1–4 (2008).

[2] K. Arai, K. Yamada and K. Yatabe, “Versatile time-frequency representations realized by convex penalty on magnitude spectrogram,” *IEEE Signal Process. Lett.*, **30**, 1082–1086 (2023).

[3] E. Cano, M. Plumbley and C. Dittmar, “Phase-based harmonic/percussive separation,” *Interspeech*, pp. 1628–1632 (2014).