調波打撃音分離の時間周波数マスクを用いた線形ブラインド音源分離*

☆大藪宗一郎，北村大地（香川高専），矢田部浩平（早稲田大）

1 はじめに

ブラインド音源分離（blind source separation: BSS）とは，マイクや音源の位置等の事前情報が無いという条件下で，複数の信号源が混合した混合音から，混合前の分離音を推定する技術である．観測マイク数が元の音源数以上となる優決定条件下での分離音推定には，音源信号の統計的独立性の仮定に基づく手法が広く用いられている．例えば，独立成分分析（independent component analysis: ICA）[1] を周波数領域で適用した周波数領域 ICA（frequency-domain ICA: FDICA）[2] や，FDICA での周波数毎の分離後に発生するパーミュテーション問題を分離と同時に解決するために提案された独立ベクトル分析（independent vector analysis: IVA）[3, 4]，IVA と非負値行列因子分解 [5] を融合させた手法である独立低ランク行列分析（independent low-rank matrix analysis: ILRMA）[6, 7] 等が提案されている．

上記の BSS は音源信号に関する事前知識（音源モデル）に基づいてパーミュテーション問題を解決しており，音源モデルの妥当性によって手法の良し悪しが特徴づけられていると解釈できる．例えば，IVA は同一音源の全周波数成分が同時に強いパワーを持つことを仮定しており，ILRMA は同一音源の時間周波数構造が低ランクになることを仮定している．即ち，より良い音源モデルを BSS に導入できれば，より高品質な分離音源が得られる可能性がある．これを探求するには異なる音源モデルの比較が重要であり，幅広い音源モデルに対応できる最適化アルゴリズムが存在するならば音源モデルの比較のコスト緩和に繋がる．

この最適化アルゴリズムの必要性に応じて，近接分離最適化法 [8]–[11] を用いて幅広い音源モデルを統一的に扱える BSS アルゴリズムが提案された [12]．この手法では，近接作用素が計算できる音源モデルであればどのようなモデルでも扱うことができる．そして，この近接作用素は多くの有用な音源モデルにおいて閾値処理として与えることができ，時間周波数マスキングとして再解釈可能である．この解釈に基づく BSS が時間周波数マスキングに基づく優決定 BSS（time-frequency-masking-based determined BSS: TFMBSS）[13] であり，これまで簡便な応用例として IVA の音源モデルにスパース性を追加したスパース IVA の効果が検証されている．なお，TFMBSS と類似する手法として，補助関数に基づく IVA の分散に時間周波数パワーの推定値を用いるモデルベース IVA [14] が提案されているが，TFMBSS は (a) 最適化に近接分離法を用いる点，及び (b) 独立性最大化という統計的枠組みを超える点の 2 点で大きく異なる．

TFMBSS は，時間周波数マスクに基づいて線形の（歪みの少ない）多チャネル音源分離が可能である．この利点を活かして，本稿では時間周波数マスクの一例として調波打撃音分離（harmonic/percussive source separation: HPSS）[15] を用いた TFMBSS を提案する．これは，HPSS に基づいていることから，調波音と打撃音の多チャネル音源分離に利用可能であり，音楽信号の解析（コード・テンポ・音階等の推定）等に応用できる．また，提案手法では，TFMBSS の反復最適化に時間周波数マスクのスムージングを新たに導入することで，より安定した音源分離が可能となることを示す．

2 従来手法

2.1 定式化

音源数と観測チャネル数をそれぞれ N 及び M とし，各時間周波数における音源信号，観測信号，分離信号をそれぞれ

$$\boldsymbol{s}_{ij} = (s_{ij1},\ \cdots,\ s_{ijN})^{\mathrm{T}} \in \mathbb{C}^N \tag{1}$$
$$\boldsymbol{x}_{ij} = (x_{ij1},\ \cdots,\ x_{ijM})^{\mathrm{T}} \in \mathbb{C}^M \tag{2}$$
$$\boldsymbol{y}_{ij} = (y_{ij1},\ \cdots,\ y_{ijN})^{\mathrm{T}} \in \mathbb{C}^N \tag{3}$$

と表す．ここで，$i=1,\cdots,I$ は周波数インデクス，$j=1,\cdots,J$ は時間インデクス，$n=1,\cdots,N$ は音源インデクス，$m=1,\cdots,M$ はチャネルインデクスを示し，$\cdot^{\mathrm{T}}$ は転置を表す．また，各信号の複素スペクトログラムを $\boldsymbol{S}_n \in \mathbb{C}^{I\times J}$，$\boldsymbol{X}_m \in \mathbb{C}^{I\times J}$，及び $\boldsymbol{Y}_n \in \mathbb{C}^{I\times J}$ で表す．これらの行列の要素はそれぞれ s_{ijn}，x_{ijm}，及び y_{ijn} である．混合系が線形時不変であり，時間周波数領域での複素瞬時混合で表現できると仮定すると，周波数毎の時不変な複素混合行列 $\boldsymbol{A}_i = (\boldsymbol{a}_{i1},\ \cdots,\ \boldsymbol{a}_{iN}) \in \mathbb{C}^{M\times N}$（ここで $\boldsymbol{a}_{in} = (a_{in1},\ \cdots,\ a_{inM})^{\mathrm{T}}$ は各音源のステアリングベクトル）が定義でき，観測信号と音源信号の関係を次式で表現できる．

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

この混合モデルは，時不変混合系の残響時間が短時間フーリエ変換（short-time Fourier transform: STFT）の窓長よりも十分短い場合に近似的に成立する．このとき，$M=N$ かつ $\boldsymbol{A}_i$ が正則であれば，分離ベクトル $\boldsymbol{w}_{in} = (w_{in1}\ \cdots\ w_{inM})^{\mathrm{T}}$ で構成される分離行列 $\boldsymbol{A}_i^{-1} = \boldsymbol{W}_i = (\boldsymbol{w}_{i1}\ \cdots\ \boldsymbol{w}_{iN})^{\mathrm{H}} \in \mathbb{C}^{N\times N}$ が存在し，分離信号は次式で与えられる．

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

ここで，$\cdot^{\mathrm{H}}$ はエルミート転置を示す．優決定条件 BSS では，式 (5) 中の分離行列 $\boldsymbol{W}_i$ を全周波数において推定することが最終的な目標となる．本稿では，以後常に決定的な系（$M=N$）を考える．

*Linear blind source separation using time-frequency mask obtained by harmonic/percussive source separation. By Soichiro OYABU, Daichi KITAMURA (NIT Kagawa), and Kohei YATABE (Waseda Univ.).

Algorithm 1 TFMBSS

Input: X, $\mathbf{w}^{[1]}$, $\mathbf{y}^{[1]}$, μ_1, μ_2, α
Output: $\mathbf{w}^{[k+1]}$

1: **for** $k = 1, \cdots, K$ **do**
2: $\quad \widetilde{\mathbf{w}} = \mathrm{prox}_{\mu_1 \mathcal{I}} \left[\mathbf{w}^{[k]} - \mu_1 \mu_2 X^{\mathrm{H}} \mathbf{y}^{[k]} \right]$
3: $\quad \mathbf{z} = \mathbf{y}^{[k]} + X(2\widetilde{\mathbf{w}} - \mathbf{w}^{[k]})$
4: $\quad \widetilde{\mathbf{y}} = \mathbf{z} - \mathcal{M}(\mathbf{z}) \odot \mathbf{z}$
5: $\quad \mathbf{y}^{[k+1]} = \alpha \widetilde{\mathbf{y}} + (1-\alpha)\mathbf{y}^{[k]}$
6: $\quad \mathbf{w}^{[k+1]} = \alpha \widetilde{\mathbf{w}} + (1-\alpha)\mathbf{w}^{[k]}$
7: **end for**

2.2 TFMBSS

文献 [12] では，FDICA に音源モデルを導入してパーミュテーション問題を回避する BSS（IVA や ILRMA 等）を統一的に解釈し，音源モデルを plug-and-play で活用できる新しい音源分離フレームワークが提案されている．本手法では，近接分離最適化法 [8]–[11] と呼ばれる最適化アルゴリズムを適用しており，例えば IVA で仮定される音源モデルを用いた BSS では，従来の IVA と同程度の音源分離を高速に達成している．

さらに文献 [13] では，上記の音源分離フレームワーク中の音源モデルに依存する箇所が時間周波数マスキングとして解釈できることに着目し，時間周波数マスクで表現される音源モデルを plug-and-play で活用可能な BSS を新たに提案している．本手法のアルゴリズムを Algorithm 1 に示す．ここで，Algorithm 1 中の X は多チャネル観測信号の複素スペクトログラム（$\boldsymbol{X}_1, \cdots, \boldsymbol{X}_M$）から構成される複素行列であり，$\mathbf{w}$ は全周波数の分離行列（$\boldsymbol{W}_1, \cdots, \boldsymbol{W}_I$）をベクトル化した複素ベクトルである．また，$\odot$ は要素毎の積を表す．これらを含む Algorithm 1 中の各変数・演算の詳細な定義は文献 [12, 13] に詳しい．また，Algorithm 1 の 4 行目の $\mathcal{M}(\mathbf{z})$ が，TFMBSS で用いられる時間周波数マスクである．このアルゴリズムでは，中間変数 $\mathbf{z}$ を引数とし分離をさらに促進するような時間周波数マスクを返す関数 $\mathcal{M}$ を音源モデルとして活用することで，そのモデルに即した音源分離が達成される．これは，マスクの情報 $\mathcal{M}(\mathbf{z})$ を事前分布においた事後確率最大化推定法としても解釈できる [13]．従って，TFMBSS では，音源分離を促進するような時間周波数マスクを返す関数 $\mathcal{M}(\mathbf{z})$ を自由に入れ替えることで，様々な音源モデルを導入した BSS が実現される．

2.3 HPSS

HPSS とは，調波楽器及び打楽器の音の振幅スペクトログラムの特徴に着目して，混合音を調波音と打撃音に分離する手法である．具体的には，振幅スペクトログラムが調波音は時間方向に滑らかであり，打撃音は非定常的でかつ周波数方向に滑らかである，という点に着目して分離を行う．ここで，モノラルの混合信号，分離された調波信号，分離された打撃信号の複素スペクトログラムをそれぞれ $\boldsymbol{B} \in \mathbb{C}^{I \times J}$，$\boldsymbol{H} \in \mathbb{C}^{I \times J}$，及び $\boldsymbol{P} \in \mathbb{C}^{I \times J}$ と表す．文献 [15] の HPSS では，混合信号 $\boldsymbol{B}$ から $\boldsymbol{H}$ と $\boldsymbol{P}$ を推定するために，次式の目的関数を $\boldsymbol{H}$ 及び $\boldsymbol{P}$ に関して最小化する．

$$J(\boldsymbol{H}, \boldsymbol{P}) = \sum_{i,j} \Big\{ \gamma_H \left(|h_{i(j+1)}|^{0.5} - |h_{ij}|^{0.5} \right)^2 + \gamma_P \left(|p_{(i+1)j}|^{0.5} - |p_{ij}|^{0.5} \right)^2 \Big\} \quad (6)$$

ここで，h_{ij} 及び p_{ij} はそれぞれ $\boldsymbol{H}$ 及び $\boldsymbol{P}$ の要素であり，γ_H 及び γ_P は各項への重み係数である．この時，$\gamma_H > 0$ 及び $\gamma_P > 0$ である．なお，式 (6) の最小化においては，次に示される拘束条件が課せられている．

$$|b_{ij}| = |h_{ij}| + |p_{ij}| \quad (7)$$

$$\arg b_{ij} = \arg h_{ij} = \arg p_{ij} \quad (8)$$

次式の反復更新式を計算することで式 (6) の最小化問題を解く．

$$|h_{ij}|^{0.5} = \frac{\gamma_H \left(|h_{(i+1)j}|^{0.5} + |h_{(i-1)j}|^{0.5} \right) |b_{ij}|^{0.5}}{\sqrt{c_{ij}^{(H)} + c_{ij}^{(P)}}} \quad (9)$$

$$|p_{ij}|^{0.5} = \frac{\gamma_P \left(|p_{i(j+1)}|^{0.5} + |p_{i(j-1)}|^{0.5} \right) |b_{ij}|^{0.5}}{\sqrt{c_{ij}^{(H)} + c_{ij}^{(P)}}} \quad (10)$$

$$c_{ij}^{(H)} = {\gamma_H}^2 \left(|h_{(i+1)j}|^{0.5} + |h_{(i-1)j}|^{0.5} \right)^2 \quad (11)$$

$$c_{ij}^{(P)} = {\gamma_P}^2 \left(|p_{i(j+1)}|^{0.5} + |p_{i(j-1)}|^{0.5} \right)^2 \quad (12)$$

3 提案手法

3.1 動機

モノラル信号の音源分離手法である HPSS では，調波音と打撃音を良く分離することができる反面，非線形な音源分離であることに起因する音質の劣化が問題となる．例えば，音源分離の誤差成分が局所的に残留することによりミュージカルノイズ等の人工的な歪みが発生する場合がある．一方，観測信号が多チャネルである場合は，IVA や ILRMA のように線形な空間分離フィルタ（分離行列 $\boldsymbol{W}_i$）を推定することで，歪みの少ない自然な音源分離が可能となる．そこで本稿では，HPSS による調波打撃音分離を利用しつつ，線形な音源分離を達成する手法として，TFMBSS の時間周波数マスク関数 $\mathcal{M}$ に HPSS を導入した音源分離手法を新たに提案する．

3.2 提案手法の概要

提案手法の処理のブロック図を Fig. 1 に示す．本手法では，TFMBSS の最適化反復中に，中間変数 $\mathbf{z}$ に対して HPSS を適用し，その結果から新たな時間周波数マスクを生成して再び TFMBSS で利用することを繰り返す．即ち，時間周波数マスクを決める関数 $\mathcal{M}(\mathbf{z})$ が HPSS となっている．

より具体的には，まず中間変数 $\mathbf{z}$ 中の調波音と打撃音に対応する要素をそれぞれ HPSS の変数 $\boldsymbol{H}$ 及び $\boldsymbol{P}$ の初期値とし，式 (9) 及び (10) を反復的に計算する．次に，得られた $\boldsymbol{H}$ と $\boldsymbol{P}$ の推定結果から時間周波数マスクを作成する．さらに，1 反復前で用いた時間

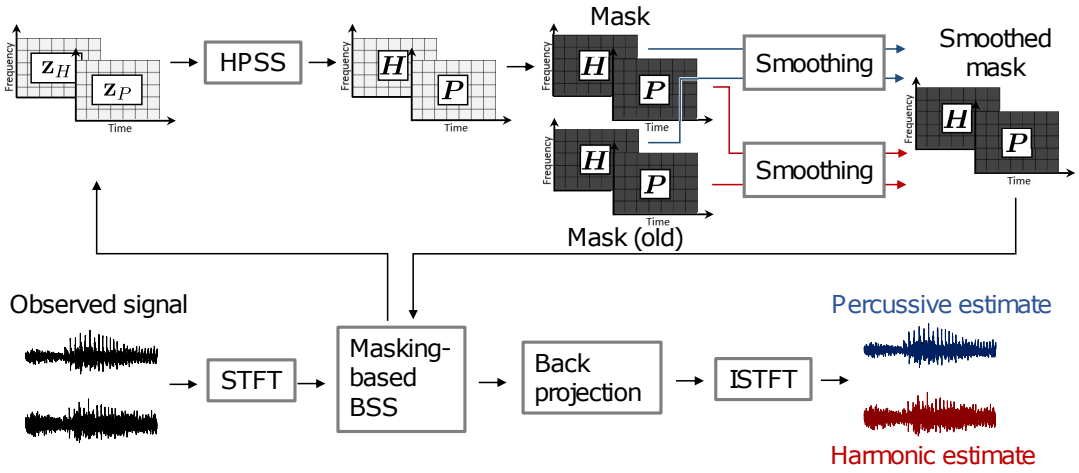

Fig. 1　Block diagram of proposed method, where $\mathbf{z}_H$ and $\mathbf{z}_P$ are parameters that corresponds to harmonic and percussive components, respectively.

周波数マスクとのスムージングを施し，これを新たな時間周波数マスクとして TFMBSS に返す．なお，TFMBSS も IVA や ILRMA と同様に分離信号のスケールの推定はできない為，プロジェクションバック法 [16] を用いて周波数毎のスケールを復元する．その後，逆 STFT（inverse STFT: ISTFT）を用いて，分離信号を時間信号に変換する．

3.3　HPSS による時間周波数マスクの生成

中間変数 $\mathbf{z}$ 中の調波音と打撃音に対応する要素を変数 $\boldsymbol{H}$ 及び $\boldsymbol{P}$ の初期値とした HPSS を行い，推定された $\boldsymbol{H}$ と $\boldsymbol{P}$ から次の時間周波数マスクを生成する．

$$[\mathcal{M}_H]_{ij} = \frac{|h_{ij}|}{|h_{ij}| + |p_{ij}|} \tag{13}$$

$$[\mathcal{M}_P]_{ij} = \frac{|p_{ij}|}{|h_{ij}| + |p_{ij}|} \tag{14}$$

ここで，$\mathcal{M}_H$ 及び $\mathcal{M}_P$ はそれぞれ調波音と打撃音の成分を強調する時間周波数マスクであり，$[\mathcal{M}]_{ij}$ はマスク $\mathcal{M}$ の ij 要素（スカラー）を表す．上記のマスク生成は，TFMBSS での反復毎に行う．

3.4　時間周波数マスクのスムージング

TFMBSS では，時間周波数マスク $\mathcal{M}$ が反復毎に大きく変動する場合，安定した音源分離ができない場合がある．提案手法においても，反復毎に HPSS でマスクの再生成を行うことから，マスクが大幅に変動しており，安定性に欠ける可能性がある．

この問題に対処するために，本稿ではマスクを生成する度に，1 反復前のマスクとのスムージングを施すことで，TFMBSS の最適化を安定させる．このマスクのスムージング処理は次式で表される．

$$\mathcal{M} = \mathcal{M}^{\beta} \odot \mathcal{M}_{\mathrm{old}}{}^{\beta_{\mathrm{old}}} \tag{15}$$

ここで，$\mathcal{M}_{\mathrm{old}}$ は 1 反復前の時間周波数マスクであり，β 及び β_{old} はそれぞれスムージング度合いを決定するパラメータである．式 (15) の処理を $\mathcal{M}_H$ 及び $\mathcal{M}_P$ のそれぞれに施すことで，スムージングを行う．スムージング後のマスクは TFMBSS に返され，中間変数 $\mathbf{z}$ 中の調波音と打撃音に対応する要素にそれぞれ適用される．

Table 1　Experimental conditions

Window function in STFT	Hann window
Window length in STFT	128 ms
Shift length in STFT	64 ms
Parameters in HPSS	$\gamma_H = 1.02$ $\gamma_P = 1.01$
# of iterations in HPSS	15 times
Parameters in masking-based BSS	$\alpha = 0.25$ $\mu_1 = \mu_2 = 1.0$
# of iterations in BSS	500 times

4　実験

4.1　実験条件

提案手法の有効性を確認するために，音楽信号中のドラムとそれ以外の楽器音の音源分離実験を行った．本実験では，SiSEC2016 [17] の DSD100 データセット中のドラム音源（drums）とその他の音源（other）を 20 曲選んだ．これらのドライソースを，文献 [18] に記載のマイク間隔 5.66 cm 及び音源方位 50°&130° の E2A インパルス応答 [19]（残響長 300 ms）で畳み込み，多チャネル混合信号を作成した．その他の実験条件は Table 1 に示す．評価指標には，信号対歪み比（source-to-distortion ratio: SDR）[20] を用いた．

4.2　実験結果

提案手法の β_{old} 及び β のみを変えた場合の各反復ごとの SDR 改善量の一例を Fig. 2 に示す．β を高く設定した場合，SDR の推移が安定せず収束点も低くなることが観測された．一方，β_{old} を高く設定した場合，推移は安定するが収束が遅れることが観測された．SDR 推移の安定と収束速度はトレードオフであるためこの点を考慮したパラメータ設定が必要となる．

次に，データセット中の 3 曲を例にとって，各従来手法と比較した結果を Fig. 3 に示す．ここで，

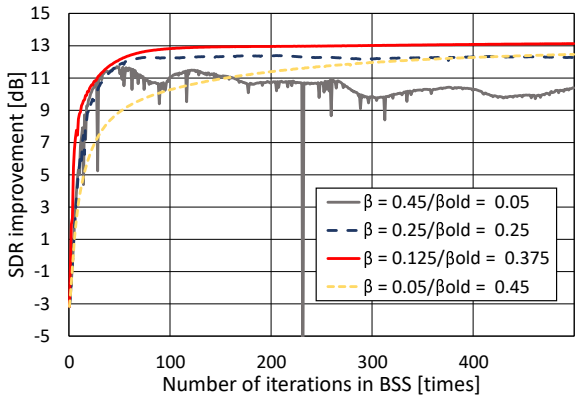

Fig. 2　Example of convergence behaviors of proposed method with various β_{old} and β.

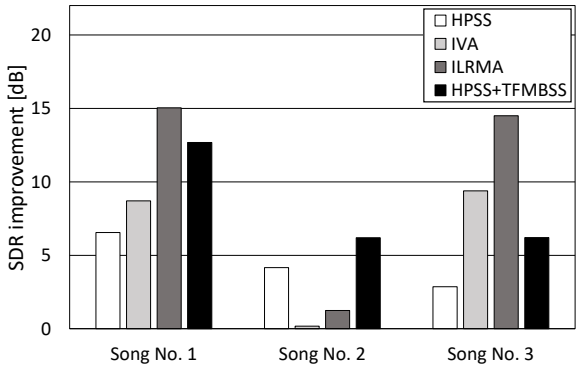

Fig. 3　Example of SDR improvements of ILRMA, IVA, conventional HPSS, and proposed method.

HPSS+TFMBSS が提案手法を示す．Fig. 3 での β_{old} 及び β はそれぞれ 0.125 及び 0.375 である．提案手法では HPSS によって作成されたマスクを元に分離するため，従来の HPSS の得手不得手が反映されているものの，線形分離化されたことによる恩恵は十分に見受けられる．ILRMA や IVA の SDR 改善量が振るわない楽曲であっても高い性能を出す例が存在した．

Table 2 は，データセット 20 曲全てにおける各従来手法との SDR 改善量の平均値の比較である．従来手法の HPSS と比較すると音質の向上は明らかであるが，調波音と打撃音の区別がはっきりとした楽曲以外には弱く他の従来法には平均スコアでは下回った．

5　まとめ

本稿では，調波音と打撃音の BSS を目的とし，HPSS に基づく時間周波数マスクを TFMBSS に利用した音源分離手法を新たに提案した．また，TFMBSS の最適化を安定化させるために，時間周波数マスクのスムージングを導入した．実験結果より，線形分離化された提案手法によって，従来の HPSS より音質が向上したことを実験的に示した．そして，各反復間のマスクが大きく変動するため SDR の推移が安定しない問題を適当なパラメータ設定によってスムージングすることで解決出来ることも実験的に示した．

謝辞　本研究の一部は JSPS 科研費 19K20306 の助成を受けたものである。

Table 2　Average SDR for each method

Method	Average SDR [dB]
HPSS	4.68
IVA	7.09
ILRMA	8.56
HPSS+TFMBSS	6.97

参考文献

[1] P. Comon, “Independent component analysis, a new concept?,” *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[2] P. Smaragdis, “Blind separation of convolved mixtures in the frequency domain,” *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[3] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, “Blind source separation exploiting higher-order frequency dependencies,” *IEEE Trans. ASLP*, vol. 15, no. 1, pp. 70–79, 2007.

[4] N. Ono, “Stable and fast update rules for independent vector analysis based on auxiliary function technique,” *Proc. WASPAA*, pp. 189–192, 2011.

[5] D. D. Lee, and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization ,” *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[6] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, “Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization,” *IEEE/ACM Trans. ASLP*, vol. 24, no. 9, pp. 1626–1641, 2016.

[7] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, “Determined blind source separation with independent low-rank matrix analysis,” *In Audio Source Separation,* S. Makino, Ed., pp. 125–155, Springer, Cham, 2018.

[8] P. L. Combettes and J. C. Pesquet, *Proximal Splitting Methods in Signal Processing*, pp. 185–212, Springer, 2011.

[9] N. Parikh and S. Boyd, “Proximal algorithms,” *Found. Trends Optim.*, vol. 1, no. 3, pp. 127–239, 2014.

[10] N. Komodakis and J. C. Pesquet, “Playing with duality: An overview of recent primal-dual approaches for solving large scale optimization problems,” *IEEE Signal Process. Mag.*, vol. 32, no. 6, pp. 31–54, 2015.

[11] M. Burger, A. Sawatzky, and G. Steidl, *First Order Algorithms in Variational Image Processing*, pp. 345–407, Springer, 2016.

[12] K. Yatabe and D. Kitamura, “Determined blind source separation via proximal splitting algorithm,” *Proc. ICASSP*, pp. 776–780, 2018.

[13] K. Yatabe and D. Kitamura, “Time-frequency-masking-based determined BSS with application to sparse IVA,” *Proc. ICASSP*, pp. 715–719, 2019.

[14] A. R. López, N. Ono, U. Remes, K. Palomäki, and M. Kurimo, “Designing multichannel source separation based on single-channel source separation,” *Proc. ICASSP*, pp. 469–473, 2015.

[15] 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,” *Proc. EUSIPCO*, 2008.

[16] N. Murata, S. Ikeda, and A. Ziehe, “An approach to blind source separation based on temporal structure of speech signals,” *Neurocomputing*, vol. 41, no. 1–4, pp. 1–24, 2001.

[17] A. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, “The 2016 signal separation evaluation campaign,” *Proc. LVA/ICA*, pp. 323–332, 2017.

[18] D. Kitamura, N. Ono, and H. Saruwatari, “Experimental analysis of optimal window length for independent low-rank matrix analysis,” *Proc. EUSIPCO*, pp. 1210–1214, 2017.

[19] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, “Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition,” *Proc. LREC*, pp. 965–968, 2000.

[20] E. Vincent, R. Gribonval and C. Fevotte, “Performance measurement in blind audio source separation,” *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.