香川高専

卒業研究論文

論文題目

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

提出年月日	令和 2 年 2 月 28 日
学　　科	電気情報工学科
氏　　名	大藪 宗一郎　　印
指導教員（主査）	北村 大地 助教　　印
副　　査	村上 幸一 准教授　　印
学　科　長	辻 正敏 教授　　印

香川高等専門学校

Linear blind source separation using time-frequency mask obtained by harmonic/percussive source separation.

Soichiro Oyabu

Department of Electrical and Computer Engineering
National Institute of Technology, Kagawa College

Abstract

Time-frequency-masking-based determined BSS (TFMBSS) is capable of liner (distortion-less) multichannel source separation based on a time-frequency mask, which is given as a source model. In this thesis, I aim to achieve linear audio source separation that employs conventional monaural source separation techniques. In particular, I propose a combination of TFMBSS and harmonic/percussive sound separation (HPSS) to achieve the linear separation of drums and other music instruments. The proposed method alternately updates the separation filter in TFMBSS and the time-frequency mask in HPSS as an iterative optimization. However, since the time-frequency mask drastically changes in each parameter update, the source separation performance becomes unstable. To solve this problem, I propose a smoothing process of the time-frequency mask.

The source-to-distortion ratio (SDR) improvement was 4.68 dB with conventional nonlinear HPSS. In the proposed method using TFMBSS and HPSS, it is experimentally shown that SDR improvement increased to 11.0 dB with the benefit of the linear separation mechanism.

Keywords: time-frequency mask, smoothing, audio source separation

（和訳）

時間周波数マスキングに基づく優決定 BSS（TFMBSS）は，音源モデルとして与えられる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．本論文では，モノラルの音源分離手法で生成する時間周波数マスクを TFMBSS の音源モデルとして活用することにより，モノラルの音源分離手法の時間周波数マスクを構成するアイディアを活かしつつ高音質な音源分離を達成することを目的とする．特に，TFMBSS と調波打撃音分離（HPSS）を組み合わせ，ドラムと他の楽器の線形分離を実現することを提案する。提案手法では，反復計算を行い HPSS による時間周波数マスクを逐次的に更新している．この時，各反復間のマスクが大きく変動するため音源分離性能が安定しないという問題が発生する．この問題を解決するために，時間周波数マスクのスムージング処理を提案する．

従来の非線形な HPSS では，信号対歪み比（SDR）改善量が 4.68 dB であったが，TFMBSS を用いた提案法では，線形分離化の恩恵によって 11.0 dB まで音質が向上した．

目次

第 1 章

緒言

1.1　音源分離の背景

音源分離とは，観測したある混合音源から，混合前の信号を推定する技術である．具体的な利用例を Figs. 1.1 及び 1.2 に示す．音源分離の例としてまず音声信号に対する分離が挙げられる．音声信号に対する分離では，混合信号から雑音を除去して音声だけを抽出する雑音と音声の分離や，複数人が会話を行っている状況下で各個人ごとに分離し抽出するような音声と音声の分離などがある．これらの研究のモチベーションとしては，スマートスピーカーやナビゲーションシステムなど音声認識技術を用いたものが近年増えている中，雑音などが含まれる混合信号では入力信号の認識精度が著しく低下するという問題が存在し，雑音の混合がないクリアな音声音源が入力として求められていることなどが挙げられる．

もう一つの例として音楽信号に対する分離がある．音声信号に対する分離は，ある観測した音楽信号からピアノ，ギター，ドラム ··· というように各楽器ごとに分離するという操作が主である．これらの研究のモチベーションとしては，近年では，既存楽曲のオーディオの再編集を行うようなリミックス文化が形成されており，オーディオ編集を行うようなユーザは各楽器ごとの高品質な分離音源を必要としていること等が挙げられる．

Fig. 1.1. Application examples of source separation for speech signals.

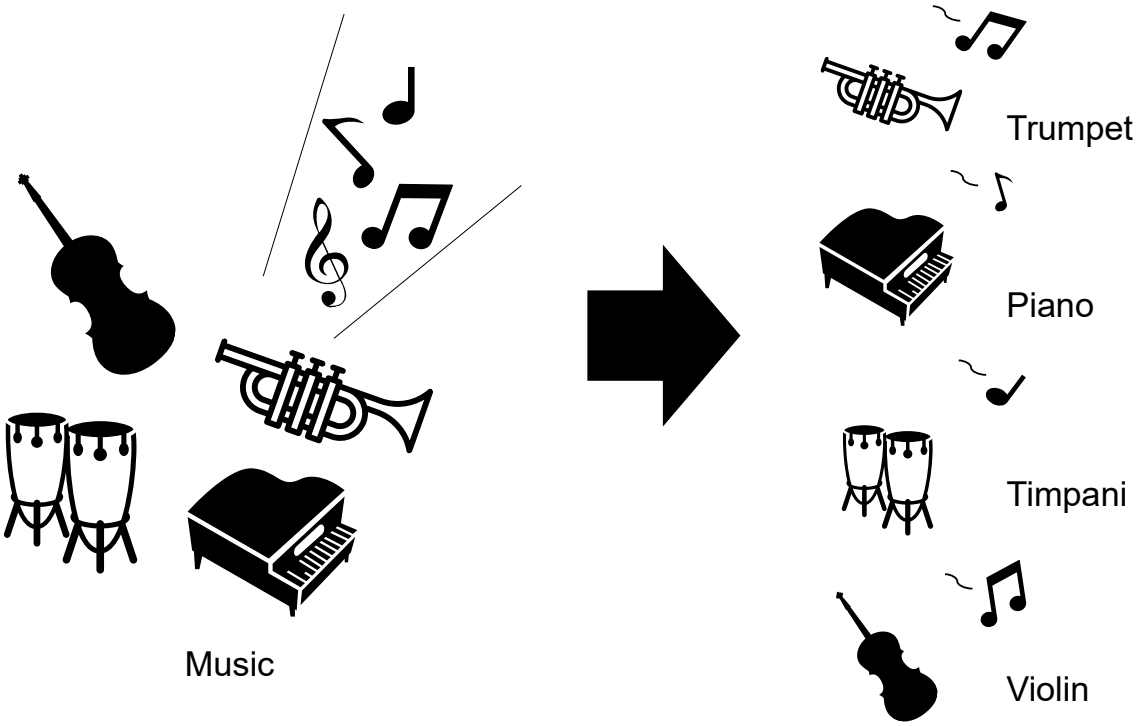


Fig. 1.2. Application example of source separation for music signals.

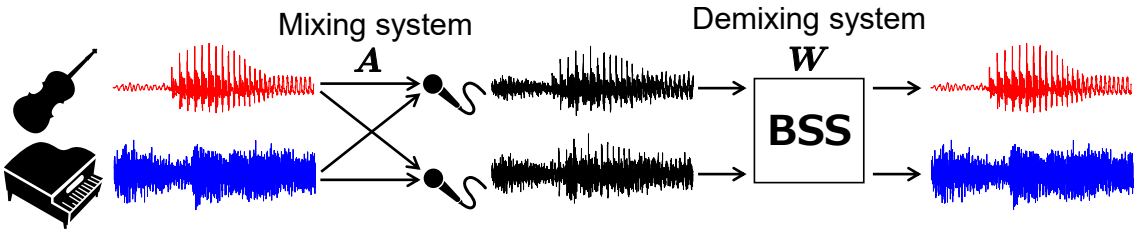


Fig. 1.3. Overview of blind source separation.

上記のように，音源分離技術は近年ニーズが高まっており，これらのタスクを満足するには高精度な音源分離手法が求められる．この経緯から 1990 年代から今日まであらゆる音源分離手法が提案されてきた．その音源分離手法の中でも，マイクロホン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ような分離手法をブラインド音源分離（blind source separation: BSS）という．ブラインド音源分離の概要を Fig. 1.3 に示す．未知の混合系 $\boldsymbol{A}$（マイクロホンや音源位置や部屋の形状などに依存して変化）から混合信号が生成される．これに対して混合系 $\boldsymbol{A}$ の逆系である分離系を推定し混合系 $\boldsymbol{A}$ に適用することで元の音源を推定するという仕組みである．

観測マイクロホン数が元の音源数以上となる優決定条件下での音源分離には，音源信号間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] を周波数毎に適用した周波数領域 ICA（frequency-domain

ICA: FDICA）[2] や，FDICA におけるパーミュテーション問題 [3] を分離と同時に解決する独立ベクトル分析（independent vector analysis: IVA）[4, 5] 及び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 は非負値行列因子分解（non negative matrix factorization: NMF）[8, 9] を用いることで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 [10]–[13] 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 [14]．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）[15] 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 [5] 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 [16] 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．

1.2 本論文における主題

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と前節で述べたが，この利点を活かして，本論文では時間周波数マスクの一例として混合音を調波音と打撃音に分離するモノラルの音源分離手法である調波打撃音分離（harmonic/percussive source separation: HPSS）[17] を用いた TFMBSS を提案する．この提案手法の既存手法に対する立ち位置を Fig. 1.4 に示す．HPSS 単体での分離では性能が悪く音の歪みが激しいため，音楽としての芸術性が大きく損なわれてしまう．そこで TFMBSS を用いて，線形な分離フィルタである多チャネル音源分離として分離することで芸術性を失わないような自然な分離音を得ることを目的とする．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．

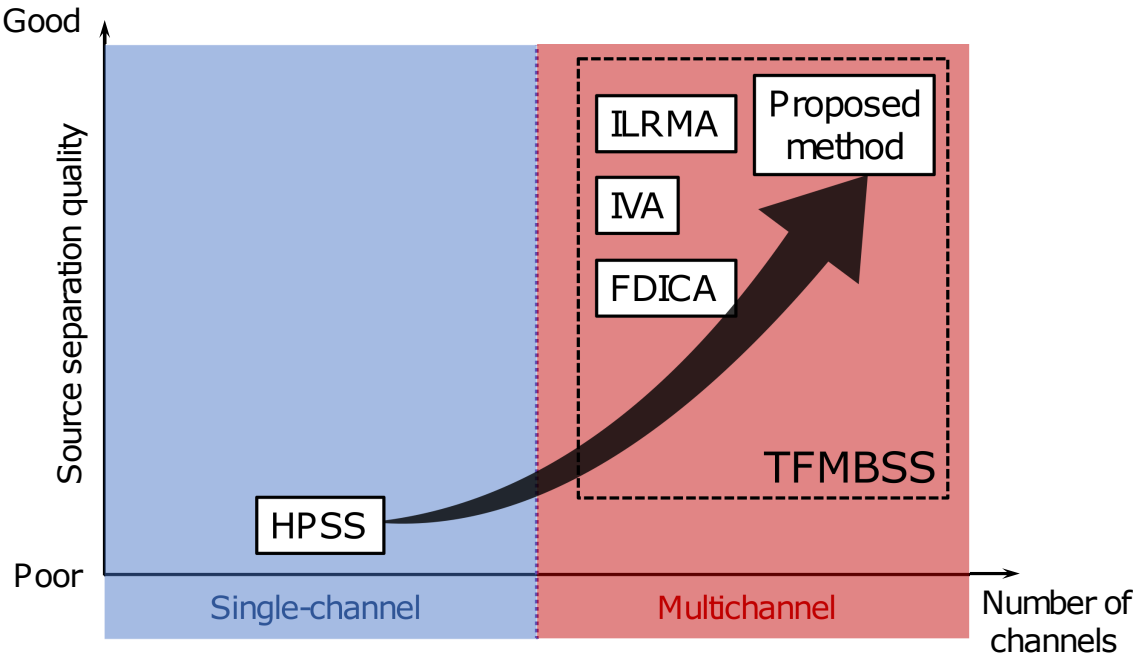


Fig. 1.4. Scope of this thesis.

1.3　本論文の構成

まず，2 章では，音響信号処理において基本的な変換である短時間フーリエ変換（short-time Fourier transform: STFT）について説明し，さらに関連の既存手法として HPSS 及び TFMBSS の概要について述べる．3 章では，本論文の提案手法における 2 種類のアルゴリズムの概要について述べ，比較検討を行う．4 章では，3 章の結果を元に新たな操作としてスムージングを実装しこの操作による差分の検証を行う．5 章では，3 章及び 4 章を踏まえた最終的な提案手法と他の従来手法と性能比較実験を行い評価する．最後に 6 章では, すべての章を総括した結言を述べる．

第 2 章

従来の音源分離手法

2.1 STFT

STFT とは，1 次元の時間信号を 2 次元の時間周波数信号に変換することである．STFT の処理の概要を Fig. 2.1 に示す．時間領域の信号を任意の長さ（フーリエ変換長）に分割し，各分割した信号をオーバーラップを持たせながら任意の長さ（シフト長）ごとにシフトし離散フーリエ変換を適応する．この時窓関数を乗ずることで両端の不連続性をスムーズにしている．これらの操作より，時間軸ごとに並んだ複素数の周波数ベクトルが得られる．このベクトルを時間軸で連結し行列として扱うことで時間と周波数両方の情報を持った 2 次元信号が得られる．これをスペクトログラムと呼ぶ．音源分離技術においては，このスペクトログラムを信号処理の対象とするのが一般的である．

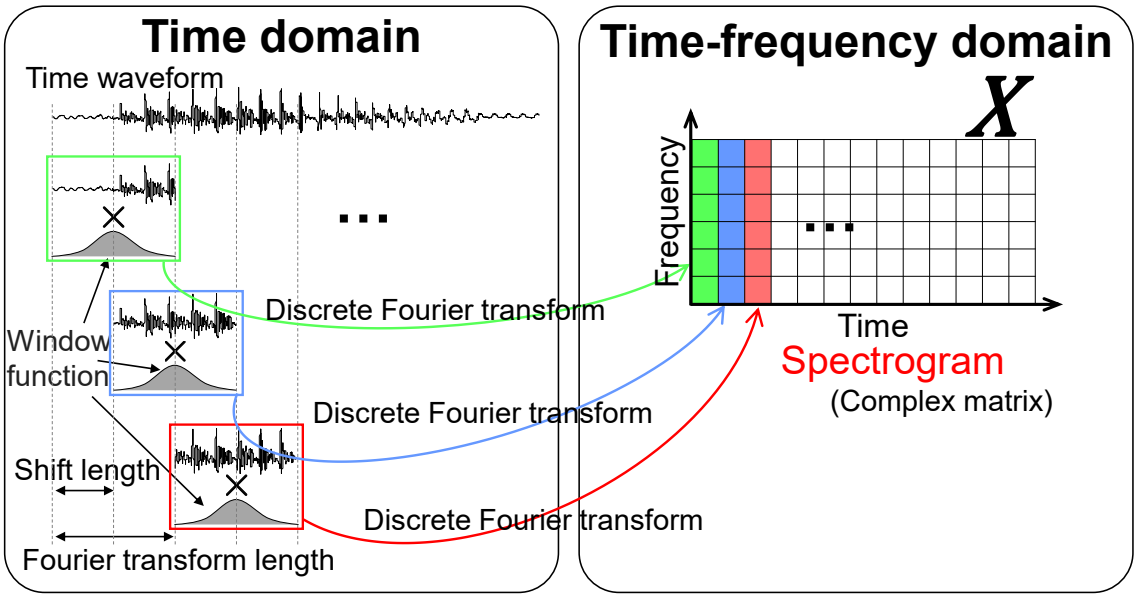


Fig. 2.1. Overview of STFT.

2.2　定式化

音源数と観測チャネル数（マイクロホン数）をそれぞれ N 及び M とし，多チャネルの時間信号を STFT して得られる時間周波数毎の音源信号，観測信号及び分離信号をそれぞれ

$$\boldsymbol{s}_{ij} = (s_{ij1},\ \cdots,\ s_{ijN})^{\mathrm{T}} \in \mathbb{C}^N \tag{2.1}$$

$$\boldsymbol{x}_{ij} = (x_{ij1},\ \cdots,\ x_{ijM})^{\mathrm{T}} \in \mathbb{C}^M \tag{2.2}$$

$$\boldsymbol{y}_{ij} = (y_{ij1},\ \cdots,\ y_{ijN})^{\mathrm{T}} \in \mathbb{C}^N \tag{2.3}$$

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．ここで，$\mathbb{C}$ は複素数全体の集合である．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．ここで，$\boldsymbol{A}_i$ は部屋の形状，マイクロホンの位置及び部屋の残響度合いなどの情報から成る線形システムである．

$$\begin{aligned}\boldsymbol{x}_{ij} &= \boldsymbol{A}_i \boldsymbol{s}_{ij} \\ &= \begin{pmatrix} a_{i11} & \cdots & a_{iN1} \\ \vdots & & \vdots \\ a_{i1M} & \cdots & a_{iNM} \end{pmatrix} \begin{pmatrix} s_{ij1} \\ \vdots \\ s_{ijM} \end{pmatrix}\end{aligned} \tag{2.4}$$

この混合モデルは，時不変混合系の残響時間が 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．

$$\begin{aligned}\boldsymbol{y}_{ij} &= \boldsymbol{W}_i \boldsymbol{x}_{ij} \\ &= \begin{pmatrix} \overline{w}_{i11} & \cdots & \overline{w}_{i1M} \\ \vdots & & \vdots \\ \overline{w}_{iN1} & \cdots & \overline{w}_{iNM} \end{pmatrix} \begin{pmatrix} x_{ij1} \\ \vdots \\ x_{ijM} \end{pmatrix}\end{aligned} \tag{2.5}$$

ここで，$\cdot^{\mathrm{H}}$ はエルミート転置，$\overline{\cdot}$ は複素共役を示す．優決定条件 BSS では，式 (2.5) 中の分離行列 $\boldsymbol{W}_i$ を全周波数 $(i = 1,\cdots,I)$ において推定することが最終的な目標となる．分離信号 $\boldsymbol{y}_{ij}$ は，混合信号 $\boldsymbol{x}_{ij}$ に対して周波数毎の分離行列を乗じることで推定されるため，式 (2.5) による音源分離は線形フィルタによる処理と等価であり，自然性の高い分離が可能である．但し，$\boldsymbol{W}_i$ は $N = M$ の条件を満たさなければ存在しないため，式 (2.5) の音源分離は優決定条件（$N \leq M$）でのみ適用可能である（$N < M$ では主成分分析等の次元圧縮を適用して $N = M$ とする）．

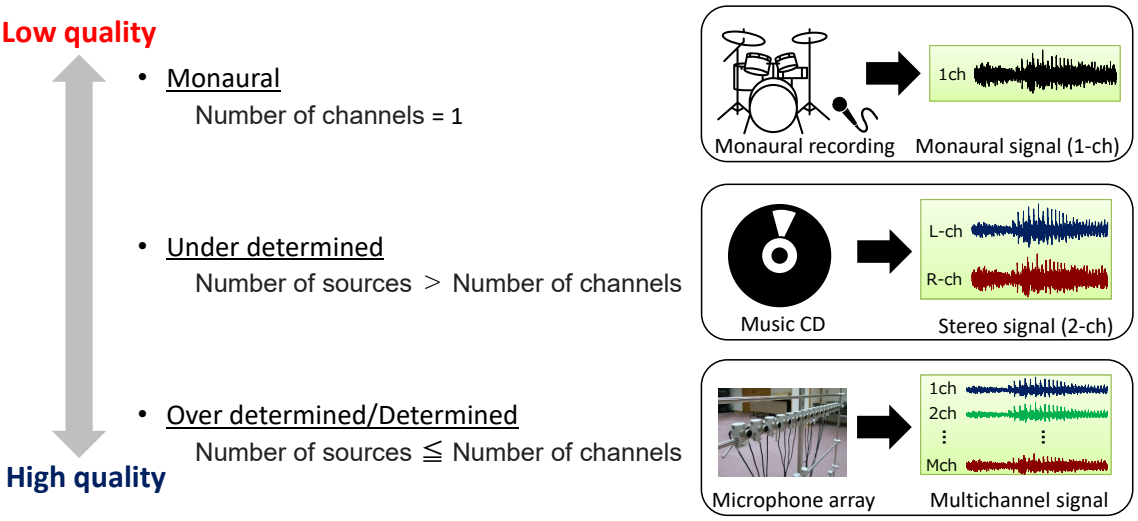


Fig. 2.2. Classification of audio source separation techniques based on number of channels of observed signals.

2.3　観測信号のチャネル数における音源分離手法の区分

音源分離において分離対象の音源のチャネル数の違いによる区分が存在しその概要を Fig. 2.2 に示す．チャネル数が一つの場合，モノラルの音源分離という．例として HPSS や罰則条件付き半教師あり NMF [18] などが挙げられる．一方，複数チャネルで録音された信号を対象とする場合は，チャネル数が音源数より少ない場合（$N > M$）と，逆にチャネル数が音源以上の場合（$N \leq M$）が考えられる．前者は劣決定条件，後者は優決定条件と呼ばれ，特に優決定条件の音源分離は式 (2.5) のように線形分離フィルタ $\boldsymbol{W}_i$ が構成できるため比的高品質である．優決定条件の音源分離の例としては，ICA，IVA，ILRMA などが挙げられる．チャネル数が少ないということは，解に対する情報量が少ないということであるため，チャネル数が少ないほど音源分離は困難であり音質が劣化する．そのため，芸術性が重要となる音楽信号等では，低品質になりがちなモノラルの音源分離手法を適用しても，音源分離信号を利用できない場合がある．

2.4　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．HPSS の概要を Fig. 2.3 に示す．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（Fig. 2.3 中の赤矢印）であり，打撃音は非定常的でかつ周波数方向に滑らか（Fig. 2.3 中の黄色矢印）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 [17] の HPSS では，

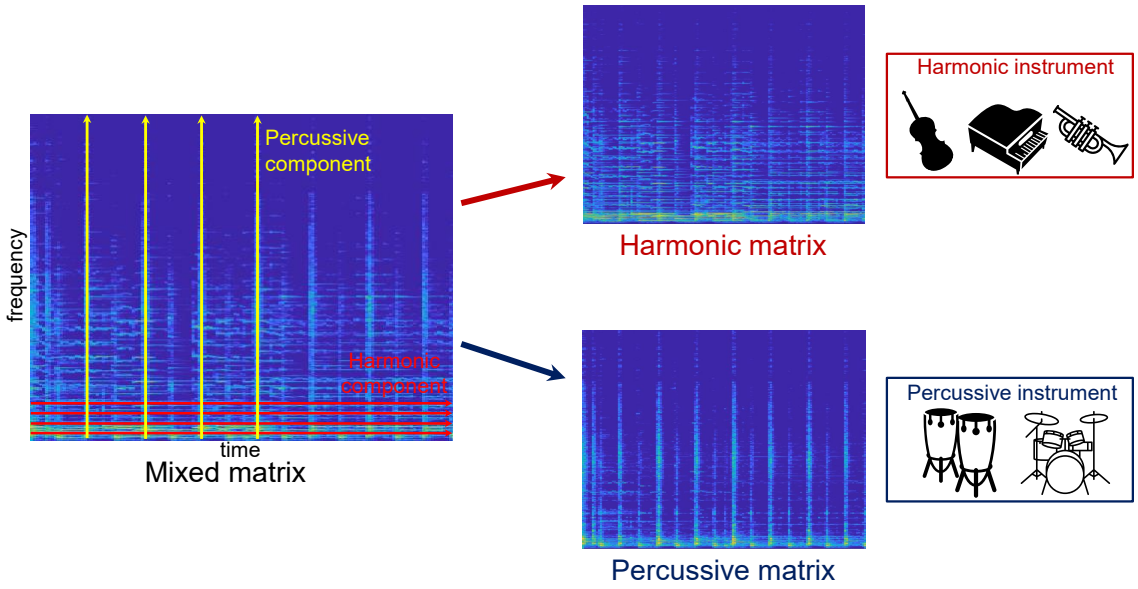


Fig. 2.3. Separation principle of HPSS.

混合信号 $\boldsymbol{B}$ から $\boldsymbol{H}$ と $\boldsymbol{P}$ を推定するために，式 (2.6) の目的関数を $\boldsymbol{H}$ 及び $\boldsymbol{P}$ に関して最小化する

$$J(\boldsymbol{H},\boldsymbol{P}) = \sum_{i,j} \left\{ \gamma_H \left(|h_{i(j+1)}|^{0.5} - |h_{ij}|^{0.5} \right)^2 + \gamma_P \left(|p_{(i+1)j}|^{0.5} - |p_{ij}|^{0.5} \right)^2 \right\} \tag{2.6}$$

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，$\gamma_H > 0$ 及び $\gamma_P > 0$ は各項への重み係数である．なお，式 (2.6) の最小化においては，式 (2.7) 及び (2.8) に示される拘束条件が課せられている．

$$|b_{ij}| = |h_{ij}| + |p_{ij}| \tag{2.7}$$

$$\arg\, b_{ij} = \arg\, h_{ij} = \arg\, p_{ij} \tag{2.8}$$

式 (2.6) の最小化する h_{ij} 及び p_{ij} は，次式の反復更新式を全ての i 及び j について繰り返し計算することで推定できる [17]．

$$|h_{ij}|^{0.5} = \frac{\gamma_H \left(|h_{(i+1)j}|^{0.5} + |h_{(i-1)j}|^{0.5} \right) |b_{ij}|^{0.5}}{\sqrt{{\gamma_H}^2 \left(|h_{(i+1)j}|^{0.5} + |h_{(i-1)j}|^{0.5} \right)^2 + {\gamma_P}^2 \left(|p_{i(j+1)}|^{0.5} + |p_{i(j-1)}|^{0.5} \right)^2}} \tag{2.9}$$

$$|p_{ij}|^{0.5} = \frac{\gamma_P \left(|p_{i(j+1)}|^{0.5} + |p_{i(j-1)}|^{0.5} \right) |b_{ij}|^{0.5}}{\sqrt{{\gamma_H}^2 \left(|h_{(i+1)j}|^{0.5} + |h_{(i-1)j}|^{0.5} \right)^2 + {\gamma_P}^2 \left(|p_{i(j+1)}|^{0.5} + |p_{i(j-1)}|^{0.5} \right)^2}} \tag{2.10}$$

Algorithm 1 TFMBSS

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

Output: $\mathbf{w}^{[k+1]}$

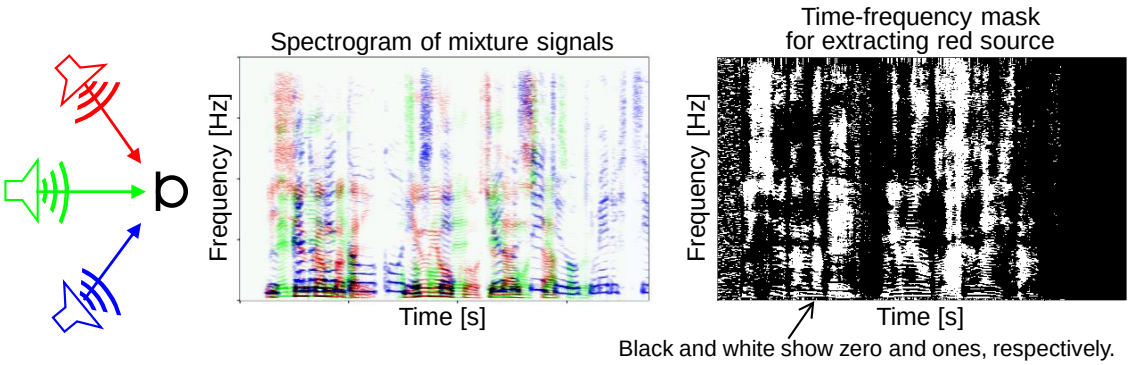
1: **for** $k = 1, \cdots, K$ **do**
2: $\quad \widetilde{\mathbf{w}} = \mathrm{prox}_{\mu_1 \mathcal{I}} \left[\mathbf{w}^{[k]} - \mu_1 \mu_2 X^{\mathrm{H}} \mathbf{y}^{[k]} \right]$
3: $\quad \mathbf{z} = \mathbf{y}^{[k]} + X(2\widetilde{\mathbf{w}} - \mathbf{w}^{[k]})$
4: $\quad \widetilde{\mathbf{y}}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

2.5 TFMBSS

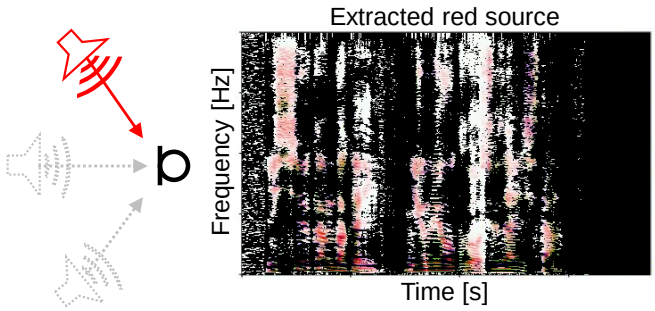
2.5.1 TFMBSS の概要

文献 [14] では，優決定条件での新しい音源分離フレームワークが提案されている．このフレームワークでは，FDICA に何らかの音源モデルを導入してパーミュテーション問題 [3] 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 で活用できるアルゴリズムが導出されている．本手法では，近接分離最適化法 [10]–[13] 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 [15] 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ク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．本論文ではこれを TFMBSS と呼ぶ．TFMBSS 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 [14, 15] 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 [15]．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．



(a) Mixture signal and time-frequency mask for red source



(b) Extracted red source by time-frequency mask

Fig. 2.4. Overview of time-frequency mask.

2.5.2　時間周波数マスク

時間周波数マスクとは，観測信号のある要素に対して目的の分離信号が存在しているかどうかを表すマスク行列である．ソフトマスクであれば 0 から 1 までの値で構成され，バイナリマスクであれば 0 または 1 の値で構成される．この概要を Fig. 2.4 に示す．Fig. 2.4（a）において，赤，緑，青の音源から成る混合信号から赤の音源のみを取り出したいとき，赤の音源の時間周波数成分を 1，それ以外を 0 とするようなマスクを作成し，要素毎に積を取ることで赤の音源のみを分離することが可能である．

モノラル信号や劣決定条件における音源分離は，式 (2.5) のような分離フィルタ $\boldsymbol{W}_i$ を構成できない（式 (2.4) における混合行列 $\boldsymbol{A}_i$ が可逆ではない）ため，何らかの方法により Fig. 2.4 のような時間周波数マスクを作成して音源分離手法を実現するアルゴリズムが大多数である．前節の HPSS も同様であり，式 (2.9) 及び (2.10) で推定した $\boldsymbol{H}$ と $\boldsymbol{P}$ からソフトマスクを構成し，混合信号のスペクトログラムに適用する．しかしながら，時間周波数マスクによる音源分離は非線形処理であることから，極端な音質の劣化や人工的な歪みを招く問題がある．TFMBSS は，この問題を解決するために提案された手法である．すなわち，HPSS のような「時間周波数マスクを構成するアイディア」 を活かしつつ，式 (2.5) の線形分離フィルタ $\boldsymbol{W}_i$ を推定することで，既存の音源分離手法を踏襲した線形な優決定条件音源分離を実現している．

2.5.3 近接作用素

Algorithm 1 の 2 行目では prox と呼ばれる関数が使用されているが，これは近接作用素（proximal operator: prox）と呼ばれる最適化関数である．prox 関数の詳細式を式 (2.11) に示す．prox 関数は最小化と射影を織り交ぜた関数であり，式 (2.11) は，$\delta \to \infty$ のとき式 (2.12) 及び $\delta \to 0$ のとき式 (2.13) となる．ここで，$\boldsymbol{u}$ はある D 次元集合 $\mathbb{R}^D$ の部分集合 C に属する要素であり，$\boldsymbol{v}$ は集合 $\mathbb{R}^D$ に属する任意の値である．そして，δ は後述の度合いを設定する任意パラメータであり，$\mathrm{dom}(f)$ は関数 f の実効定義域である．実効定義域は式 (2.14) のように定義する．前者の場合，式 (2.11) の第 2 項が消えて式 (2.12) となる．$\boldsymbol{u}^*$ は $f(\boldsymbol{u})$ の最小値を与える点であり，これは最小化単体の意味合いを持つ．後者の場合，式 (2.11) の第 1 項が消えて式 (2.13) となりこれは射影単体の意味合いを持つ．これより，δ の数値によって最小化と射影の役割の度合いを重みづけしていると分かる．

$$\mathrm{prox}_{\delta f}(\boldsymbol{v}) \triangleq \underset{\boldsymbol{u}\in\mathrm{dom}(f)}{\arg\min} \left\{ f(\boldsymbol{u}) + \frac{1}{2\delta} ||\boldsymbol{u}-\boldsymbol{v}||_2{}^2 \right\} \tag{2.11}$$

$$\mathrm{prox}_{\delta f}(\boldsymbol{v}) = \underset{\boldsymbol{u}\in\mathrm{dom}(f)}{\arg\min}\, f(\boldsymbol{u}) = \boldsymbol{u}^* \tag{2.12}$$

$$\mathrm{prox}_{\delta f}(\boldsymbol{v}) = \underset{\boldsymbol{u}\in\mathrm{dom}(f)}{\arg\min} \frac{1}{2\delta} ||\boldsymbol{u}-\boldsymbol{v}||_2{}^2 = \prod_C (\boldsymbol{v}) \tag{2.13}$$

$$\mathrm{dom}(f) \triangleq \{\boldsymbol{u} \in \mathbb{R}^D : f(\boldsymbol{u}) < \infty\} \tag{2.14}$$

prox 関数の概要図を Fig. 2.5 に示す．ここで図の楕円は $f(\boldsymbol{u})$ の高等線である．式 (2.13) における $\prod_C$ は L_2 ノルムであり，射影作用素とも呼ばれ，$\mathrm{dom}(f)$ 内に $\boldsymbol{v}$ が存在しない場合，$\mathrm{dom}(f)$ 内に $\boldsymbol{v}$ を写像する．そして，δ の数値に従って $f(\boldsymbol{v})$ を最小化する $\boldsymbol{u}^*$ に近づいていくという動作をする．以上より，Algorithm 1 の 2 行目は微分不可能性や非有限性を持った関数あっても最小化するという役割を担っている．

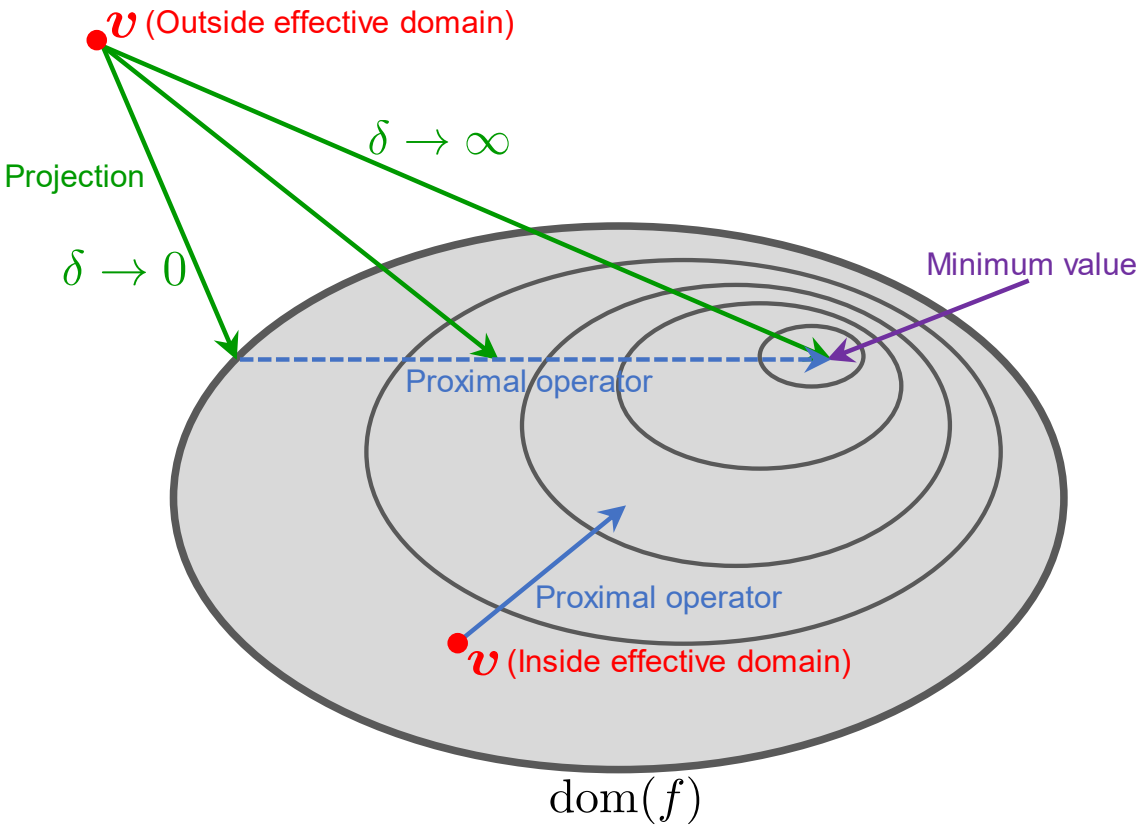


Fig. 2.5. Overview of proximal operator.

2.6　本章のまとめ

本章では，本論文において関連した音源分離手法として HPSS 及び TFMBSS を解説した．次章以降では，本章で解説した音源分離手法を元にした新しい音源分離手法を提案し実験考察を行う．

第 3 章

HPSS による時間周波数マスクに基づくブラインド音源分離

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．この問題は，音楽信号のように芸術的価値が重要な信号においては深刻である．一方，観測信号が優決定条件（$M \geq N$）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．

そこで本論文では，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．そして HPSS を導入した音源分離手法として 2 種類のアルゴリズムを提案し実験検討を行う．提案手法 1 では忠実にモノラルの HPSS の分離方法を踏襲したアルゴリズムであるのに対し，提案手法 2 では HPSS をフィルタとして捉え，調波音成分でないもの，打撃音成分でないものを排他的に取り除いていくアルゴリズムである．以降では，これら 2 種類のアルゴリズムを比較しながら評価を行う．

3.2 提案手法 1 の概要

提案手法 1 の処理のブロック図を Fig. 3.1 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 そのものとなっている．

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 (2.9) 及び (2.10) 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，次章で詳しく述べるように，1

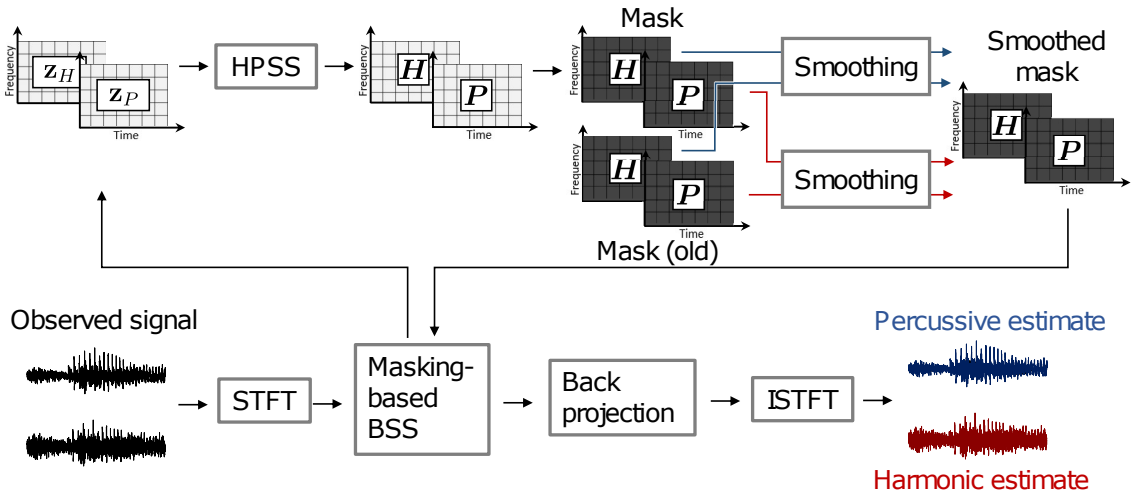


Fig. 3.1. Block diagram of proposed method 1, where $\mathbf{z}_H$ and $\mathbf{z}_P$ are parameters that corresponds to harmonic and percussive components, respectively.

反復前で用いた時間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，分離行列 $\boldsymbol{W}_i$ の推定後に式 (2.5) で得られる分離信号 $\boldsymbol{y}_{ij}$ に対して，プロジェクションバック法 [19] を適用し，周波数毎のスケールを復元する．元々，式 (2.5) における分離信号 $\boldsymbol{y}_{ij}$ は，分離信号のスケール（音量）までは推定できず，適当なスケールで推定されてしまう問題がある．全周波数で一様なスケールの任意性であれば問題ないが，周波数毎の分離行列 $\boldsymbol{W}_i$ がスケール不定であるために，分離信号 $\boldsymbol{y}_{ij}$ は周波数 i 毎にバラバラなスケールになってしまい，このまま逆 STFT（inverse STFT: ISTFT）を適用して時間信号に変換しても正しい分離信号は得られない．この問題に対する解決策としてプロジェクションバック法が導入されている．ここで，$\boldsymbol{y}'_{ijn}$ を次式のように定義する．

$$\boldsymbol{y}'_{ijn} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ y_{ijn} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{3.1}$$

すなわち，$\boldsymbol{y}'_{ijn}$ は $\boldsymbol{y}_{ij}$ の n 番目の要素のみを残し，他を 0 とした分離信号である．プロジェクションバック法では，式（3.1）の分離信号に対して，推定済の分離行列 $\boldsymbol{W}_i$ の逆行列 $\boldsymbol{W}_i^{-1}$

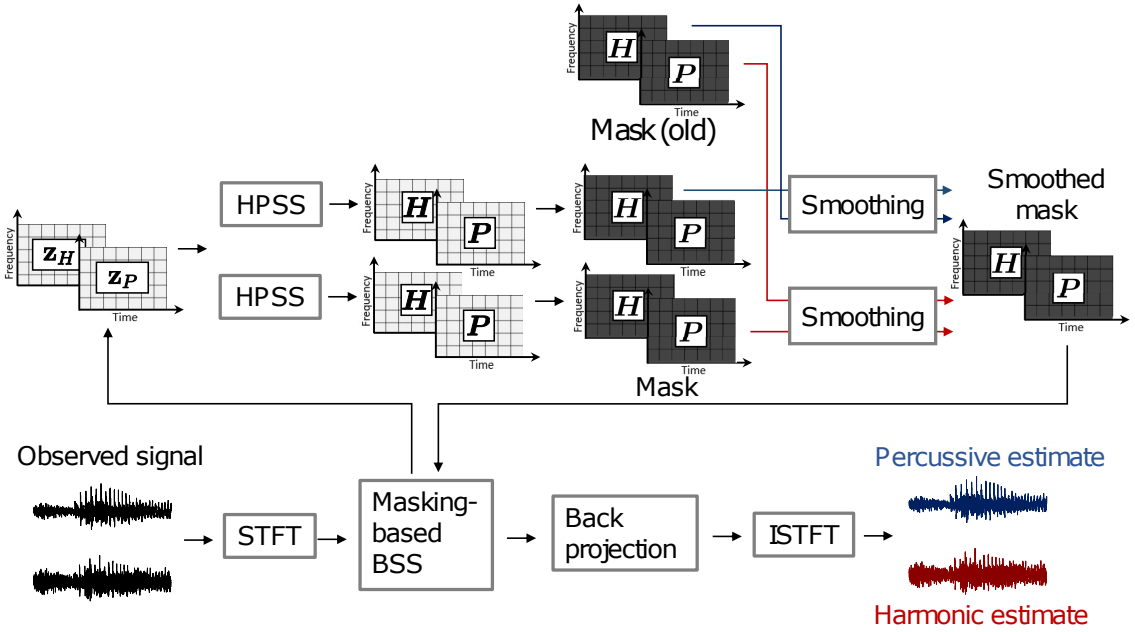


Fig. 3.2. Block diagram of proposed method 2, where $\mathbf{z}_H$ and $\mathbf{z}_P$ are parameters that corresponds to harmonic and percussive components, respectively.

を次式のように適用する．

$$\begin{pmatrix} \hat{y}_{ijn1} \\ \vdots \\ \hat{y}_{ijnm} \\ \vdots \\ \hat{y}_{ijnM} \end{pmatrix} = \boldsymbol{W}_i^{-1} \boldsymbol{y}'_{ijn} \tag{3.2}$$

ここで，得られた信号 $\hat{y}_{ijnm}$ は m 番目のマイクロホンで観測されたスケールに合わせた n 番目の分離信号である．式（3.2）で表されるプロジェクションバック法により，スケールを全周波数で合わせた分離信号 $\hat{y}_{ijnm}$ が得られる．本論文でのプロジェクションバック法は，常に $\hat{y}_{ijn1}$ を出力する．その後，ISTFT を用いて，分離信号を時間信号に変換する．

3.3　提案手法 2 の概要

提案手法 2 の処理のブロック図を Fig. 3.2 に示す．本手法では提案手法 1 と同様に，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別の HPSS に入力として与え，その要素をそれぞれ二分の一した値を HPSS の変数 $\boldsymbol{H}$ 及び $\boldsymbol{P}$ の初期値とし，式 (2.9) 及び (2.10) を反復的に計算する．ここで，HPSS の入力変数が二つ必要であること及び

制約条件式 (2.8) の遵守という二つの理由から，二分の一した値を HPSS の初期値として採用している．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，中間変数 $\mathbf{z}$ 中の調波音信号 $\mathbf{z}_H$ で作成したマスクでは 1 反復前で用いた時間周波数マスク Mask（old）のうち $\boldsymbol{H}$ に対応するマスクとのスムージングを施す．一方，中間変数 $\mathbf{z}$ 中の打撃音信号 $\mathbf{z}_P$ で作成したマスクでは 1 反復前で用いた時間周波数マスク Mask（old）のうち $\boldsymbol{P}$ に対応するマスクとのスムージングを施す．これを新たな時間周波数マスクとして TFMBSS に返す．これらの操作において，スムージングを施さない方のマスクは前者であれば調波成分ではないもの，後者であれば打撃成分ではないものに該当する．これに基づいて中間変数 $\mathbf{z}$ 中の調波音で作成したマスクの $\boldsymbol{P}$ に対応するマスク及び中間変数 $\mathbf{z}$ 中の打撃音で作成したマスクの $\boldsymbol{H}$ に対応するマスクは使用せず破棄する．該当しないものを取り除くというように排他的に動作させることで片方のマスクはより $\boldsymbol{P}$ の成分が取り除かれたマスクとなり，もう片方のマスクはより $\boldsymbol{H}$ の成分が取り除かれたマスクとなる．提案手法 1 と同様に，分離信号にはプロジェクションバック法を適用し，その後 ISTFT を用いて，分離信号を時間信号に変換する．

3.4　本章のまとめ

本章では，本論文における二つの提案手法について説明した．次章では，提案手法の不安定性という問題に対する解決としてのスムージング処理を詳しく解説し，有用性を検証する．

第 4 章

時間周波数マスクのスムージング

4.1　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．

$$[\mathcal{M}_H]_{ij} = \frac{|h_{ij}|}{|h_{ij}| + |p_{ij}|} \tag{4.1}$$

$$[\mathcal{M}_P]_{ij} = \frac{|p_{ij}|}{|h_{ij}| + |p_{ij}|} \tag{4.2}$$

ここで，$\mathcal{M}_H \in \boldsymbol{R}_{[0,1]}^{I\times J}$ 及び $\mathcal{M}_P \in \boldsymbol{R}_{[0,1]}^{I\times J}$ 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．

4.2　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，最適化としての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，本論文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．

$$\mathcal{M} = \mathcal{M}^{\beta} \odot {\mathcal{M}_{\mathrm{old}}}^{\beta_{\mathrm{old}}} \tag{4.3}$$

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 (4.3) の処理を $\mathcal{M}_H$ 及び $\mathcal{M}_P$ のそれぞれに施す．1 反復前の時間周波数マスクと現在のマスクを β 及び β_{old} に応じて要素ごとの積を取ることで，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．

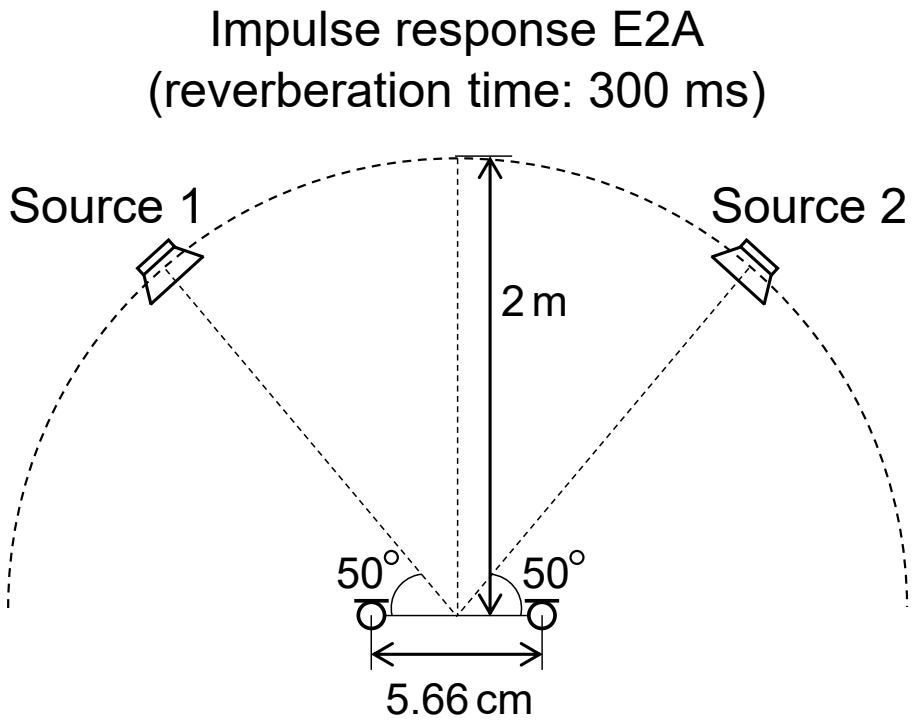


Fig. 4.1. Convolution system in E2A.

4.3　マスクのスムージングにおける影響の検証

4.3.1　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（後述のその他の音源（other）に該当）の音源分離実験を行った．本実験では，SiSEC2016 [20] の DSD100 データセットを使用した．DSD100 はトレーニングセット（Dev）とテストセット（Test）の２つのデータセットが存在し各 50 曲が収録されている．各曲は様々なスタイルの楽曲で構成されており，ボーカル音源（vocals），ベース音源（bass），ドラム音源（drums）及びその他の音源（other）が音源ごとに収録されている．ここで，その他の音源（other）は，鍵盤楽器，管楽器及び弦楽器などの音階を持った楽器の混合音であり，その編成は楽曲に依存する．DSD100 のテストセットの中でアルファベット順で並べた音源から 3〜6，8，11，13〜19，36 番目の楽曲 14 曲，DSD100 のトレーニングセットの中でアルファベット順で並べた音源から 1，3，9，15，23，27 番目の楽曲 6 曲のドラム音源（drums）とその他の音源（other）を選び，それぞれ Song ナンバー 1〜14 及び 15〜20 に割り当てた．選定の基準は，楽曲自体に打撃音と調波音がバランス良く存在しているか又はシンセドラムなどではなく一般的なドラム音かどうかである．これらのドライソースを，Fig. 4.1 に記載のマイクロホン間隔 5.66 cm 及びマイクロホンを中点とした半径 2m の円上で音源方位 50°&130° の E2A インパルス応答 [21]（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 4.1 に示す．評価指標に信号対歪み比（source-to-distortion ratio: SDR）[22] を用いた．

Table. 4.1. Experimental conditions

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

4.3.2 実験結果

提案手法 1 においての，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 改善量を Figs. 4.2–4.21 に示す．同様に，提案手法 2 においての，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 改善量を Figs. 4.22–4.41 に示す．提案手法 1，2 共に 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．提案手法 1，2 を比較すると提案手法 1 では β の値がある程度高く設定されていても安定した推移が見られるが収束点が低く，提案手法 2 では β の値が極めて高くなければ推移が安定しないが収束点は高いことが観測された．しかしながら，両手法で全体的に推移が安定しており，スムージングの効果がみられた．β を極端に低くした場合推移は非常に安定であるが，収束点が若干低くなる傾向も観測された．本論文におけるスムージングの目的は SDR の推移の安定によって収束値の安定，増加であるため SDR の安定を取ることで収束点が低くなるのは本末転倒である．そこで以降の実験では，最も安定した β_{old}=0.45 及び β=0.05 の条件ではなく，安定性は下がるが収束点が最も高い条件の β_{old}=0.375 及び β=0.125 を採用する．

4.4 本章のまとめ

本章では，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い提案手法の β_{old} 及び β のみを変えた場合の各反復ごとの SDR 改善量を比較した．提案手法 1 はスコアの推移が比較的安定だが収束点が低く，逆に提案手法 2 はスコアの推移は安定しないが収束点は高いことが確認された．両手法ともに β_{old} 及び β の変化に応じて確実に推移の安定が見込まれた．この実験より，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．次章ではこのスムージングの概念を取り入れ，他の従来手法と性能比較実験を行い本手法の有用性を確立する．

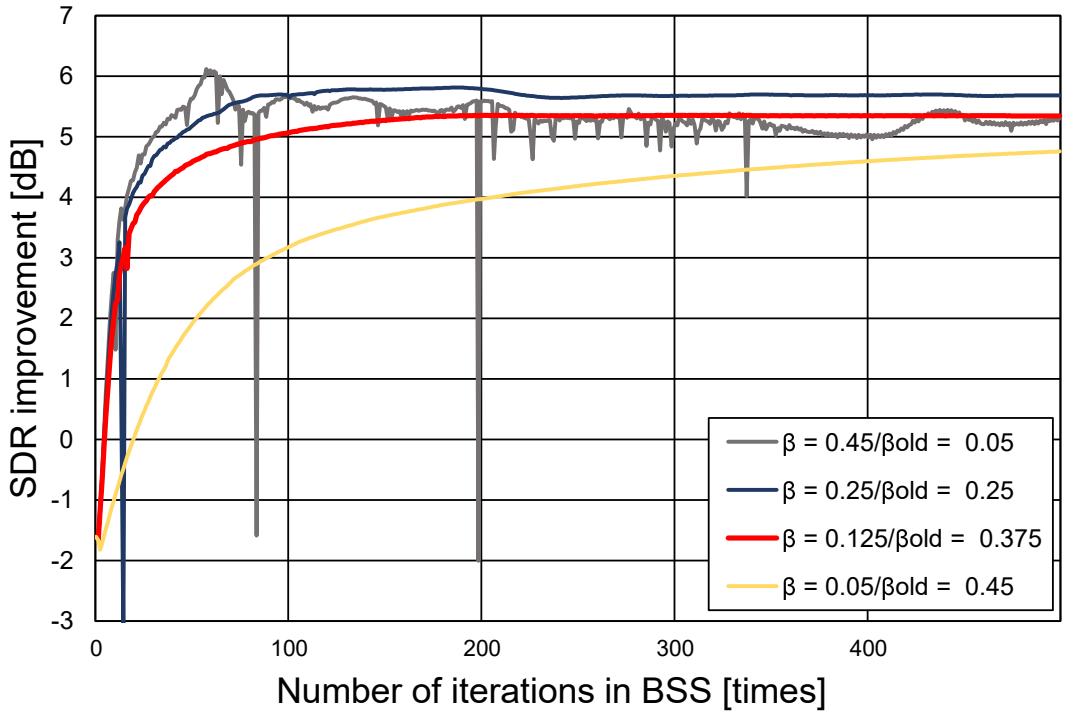


Fig. 4.2. Example of convergence behaviors of proposed method 1 with various β_{old} and β (song no. 1).

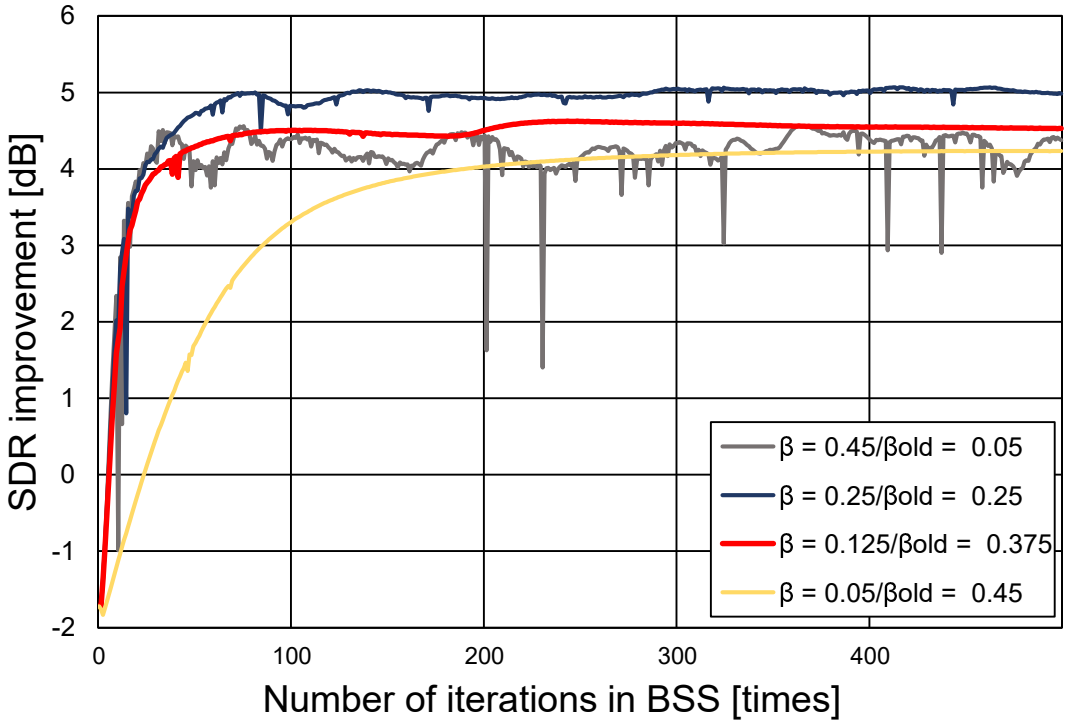


Fig. 4.3. Example of convergence behaviors of proposed method 1 with various β_{old} and β (song no. 2).

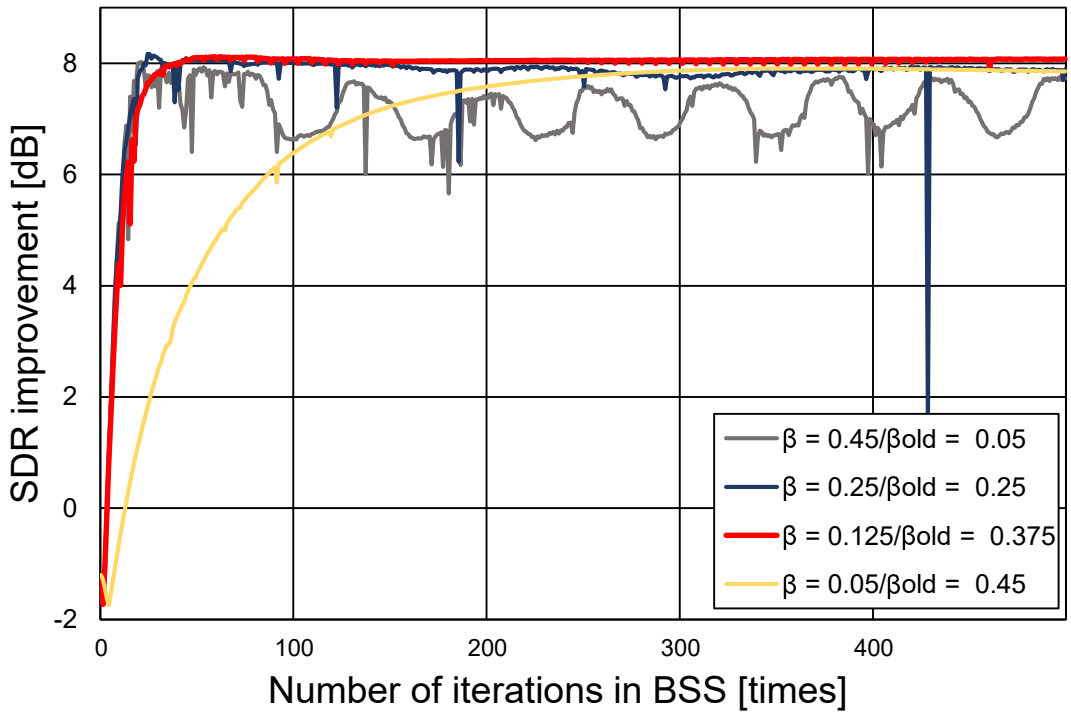


Fig. 4.4. Example of convergence behaviors of proposed method 1 with various β_{old} and β (song no. 3).

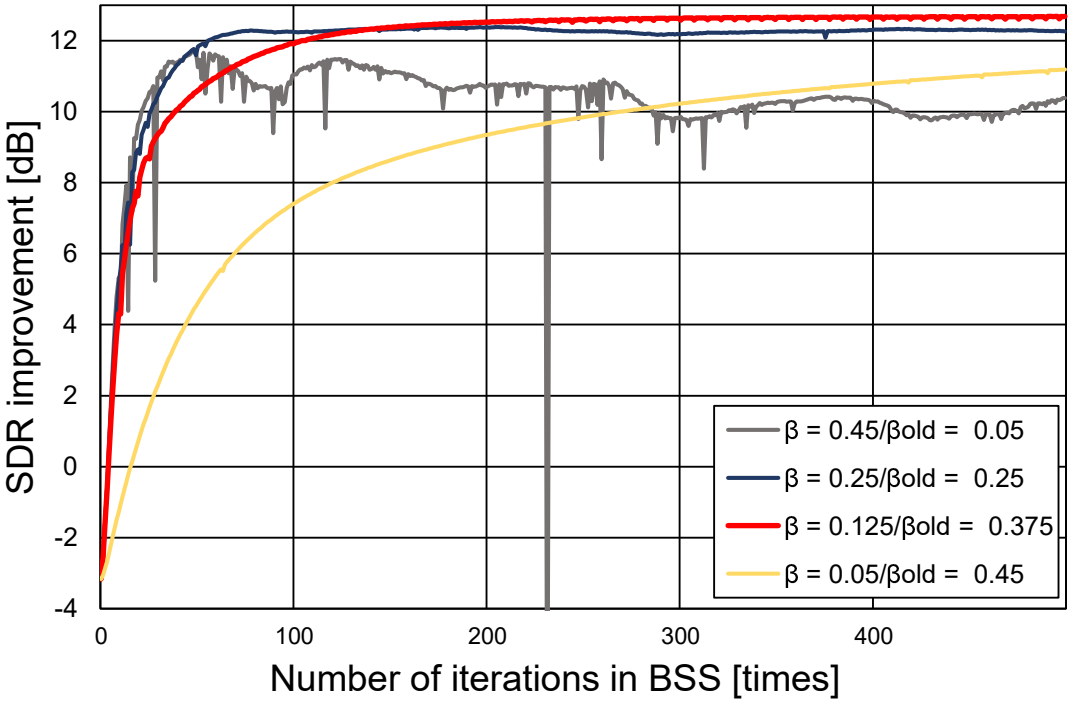


Fig. 4.5. Example of convergence behaviors of proposed method 1 with various β_{old} and β (song no. 4).

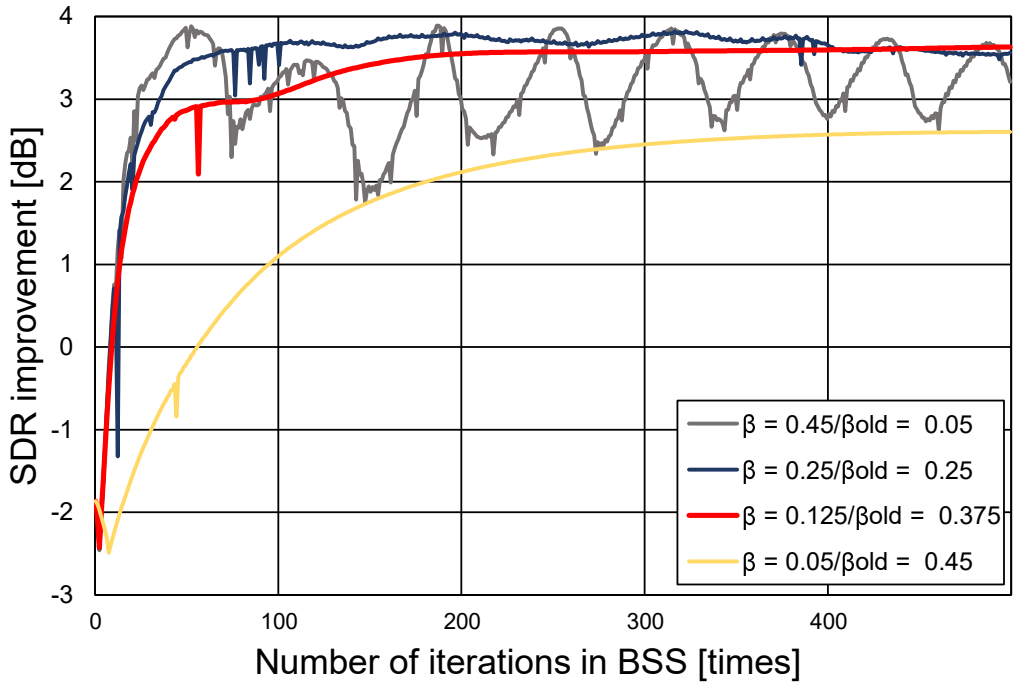


Fig. 4.6. Example of convergence behaviors of proposed method 1 with various β_{old} and β (song no. 5).

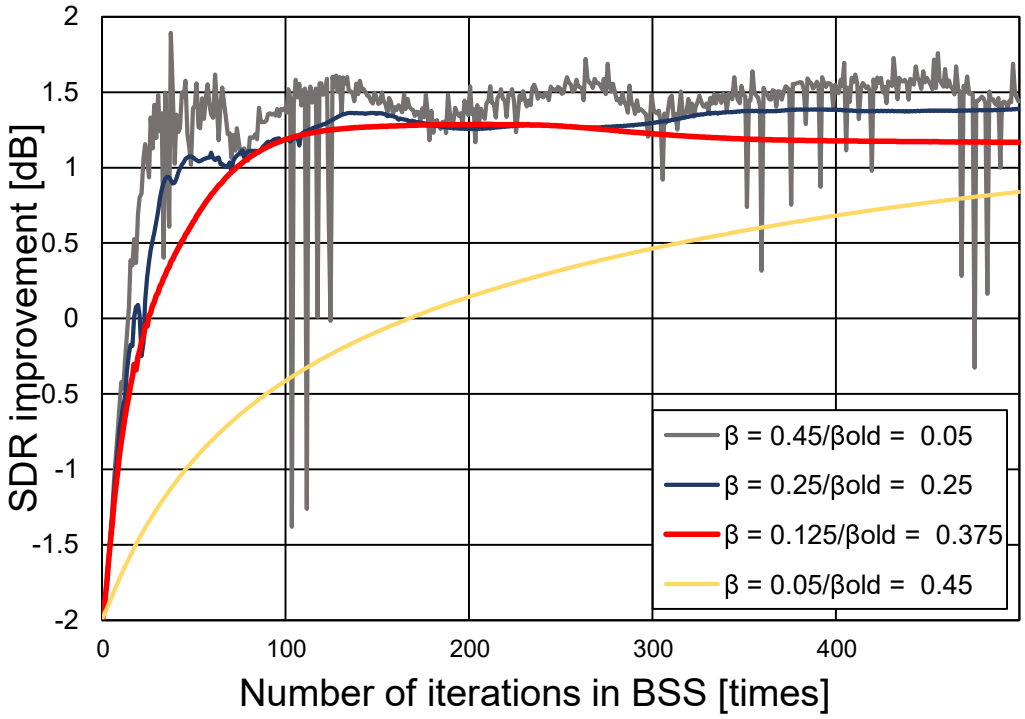


Fig. 4.7. Example of convergence behaviors of proposed method 1 with various β_{old} and β (song no. 6).

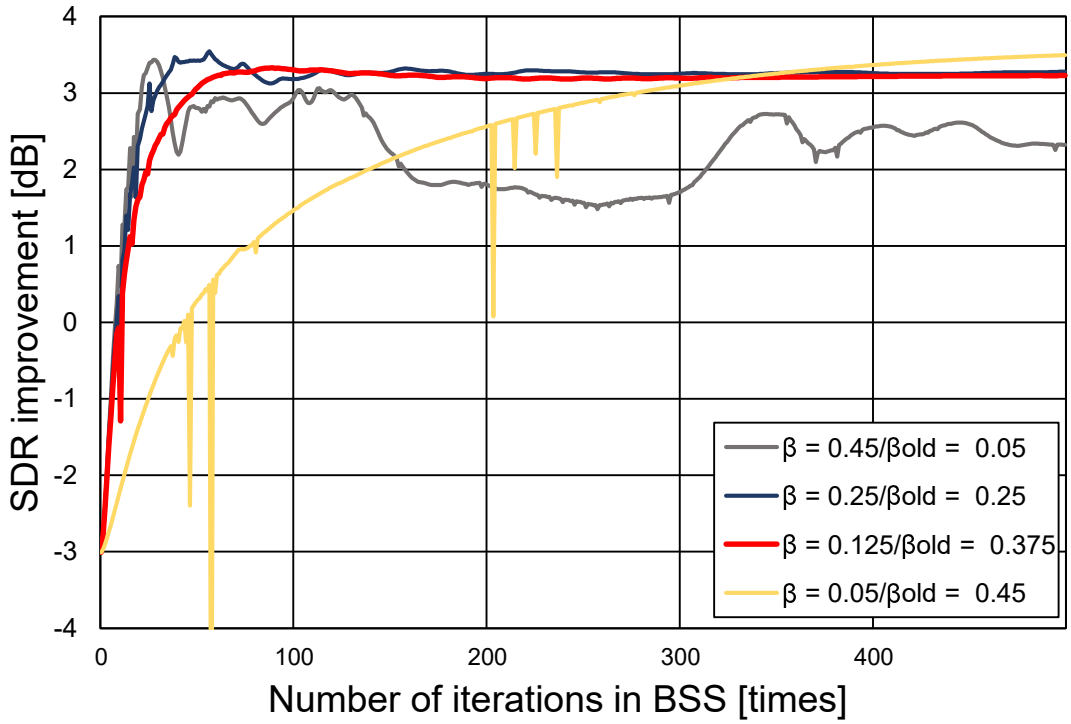


Fig. 4.8. Example of convergence behaviors of proposed method 1 with various β_{old} and β (song no. 7).

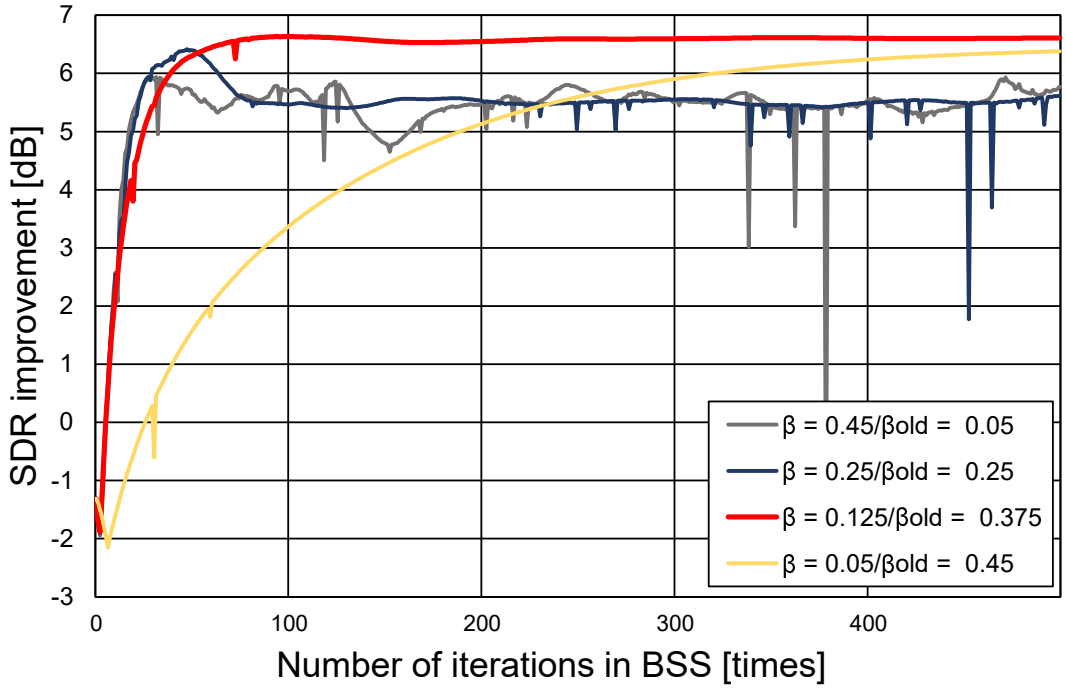


Fig. 4.9. Example of convergence behaviors of proposed method 1 with various β_{old} and β (song no. 8).

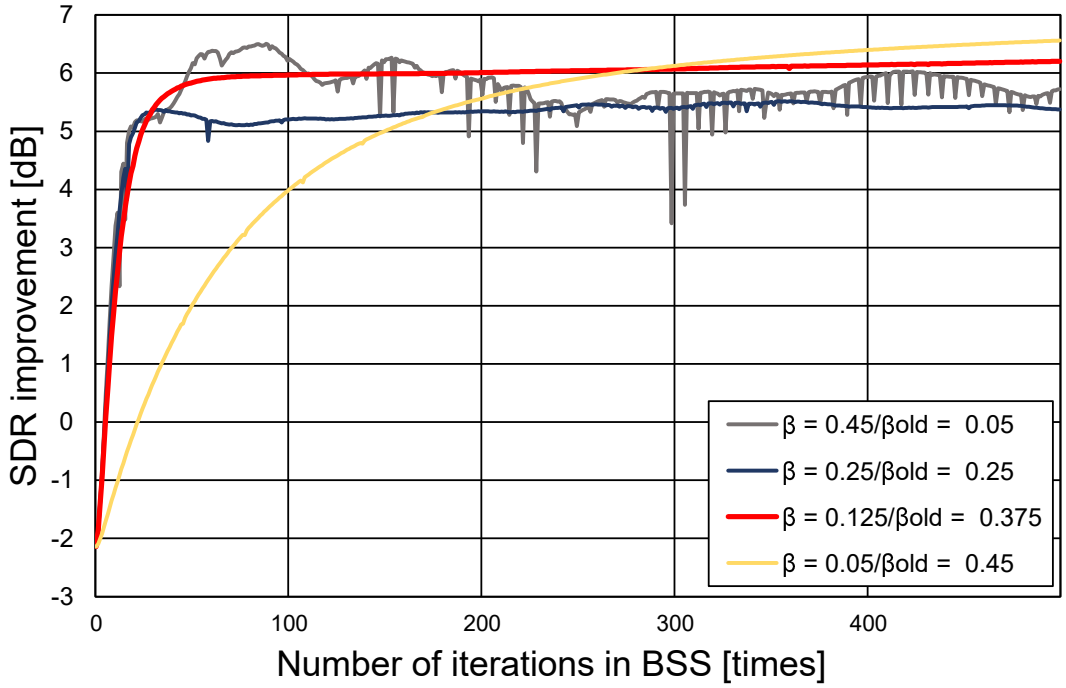


Fig. 4.10. Example of convergence behaviors of proposed method 1 with various β_{old} and β (song no. 9).

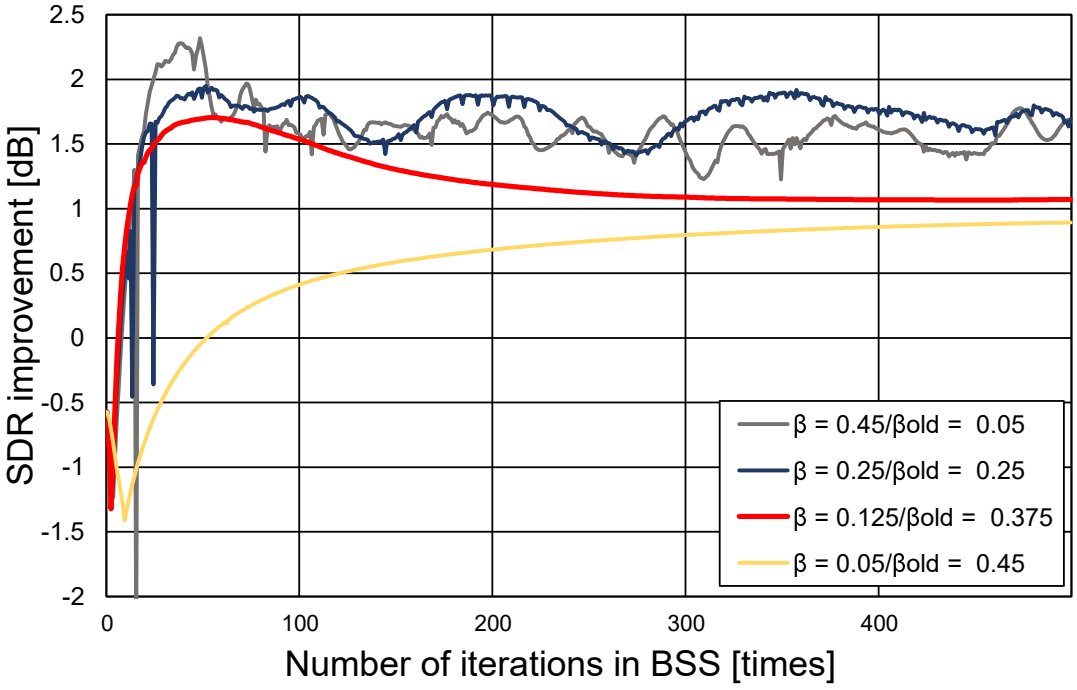


Fig. 4.11. Example of convergence behaviors of proposed method 1 with various β_{old} and β (song no. 10).

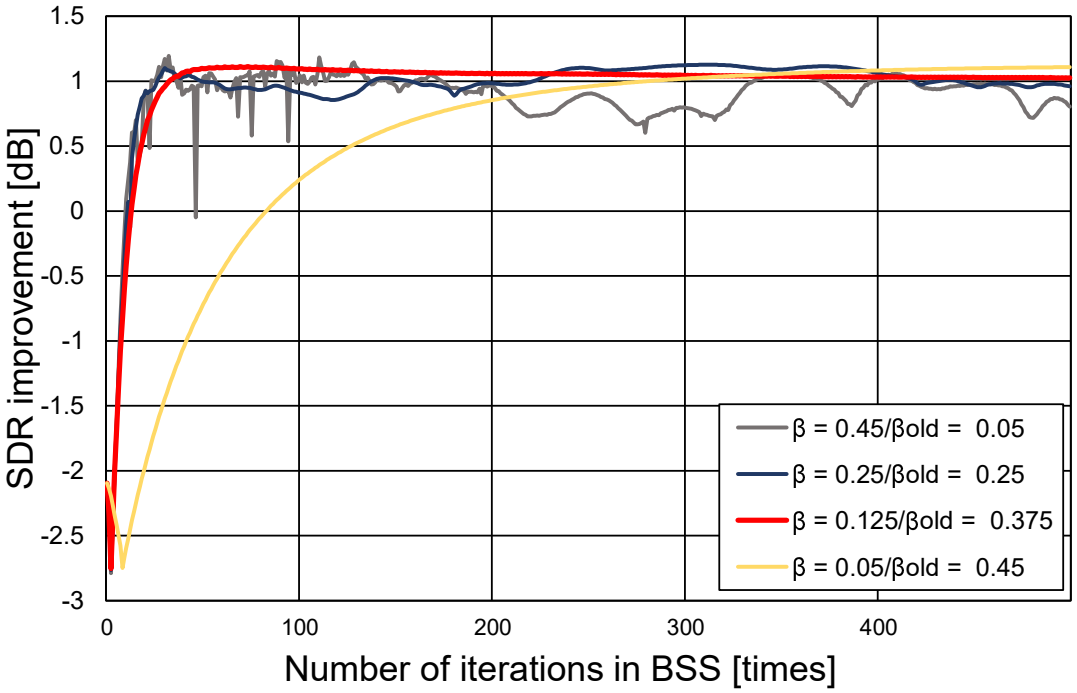


Fig. 4.12. Example of convergence behaviors of proposed method 1 with various β_{old} and β (song no. 11).

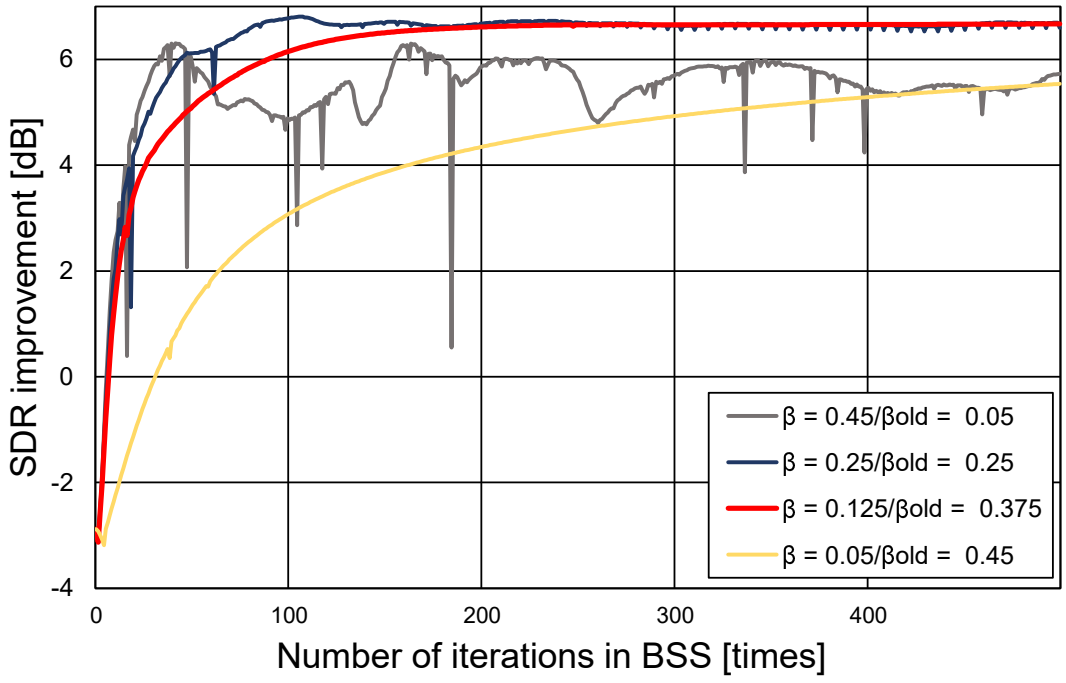


Fig. 4.13. Example of convergence behaviors of proposed method 1 with various β_{old} and β (song no. 12).

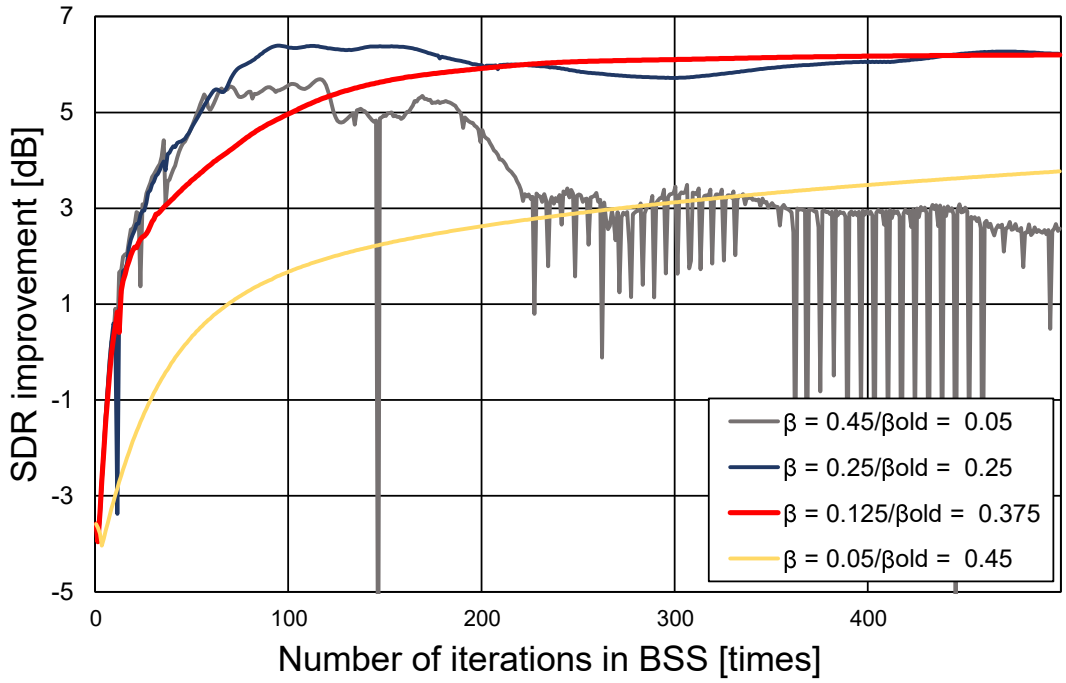


Fig. 4.14. Example of convergence behaviors of proposed method 1 with various β_{old} and β (song no. 13).

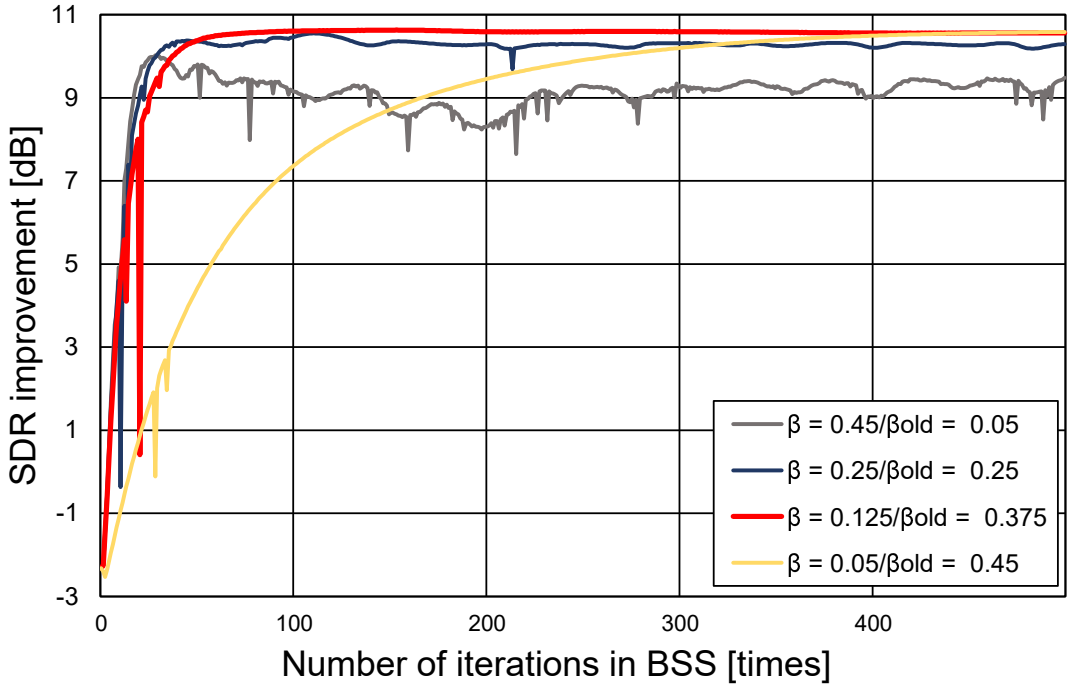


Fig. 4.15. Example of convergence behaviors of proposed method 1 with various β_{old} and β (song no. 14).

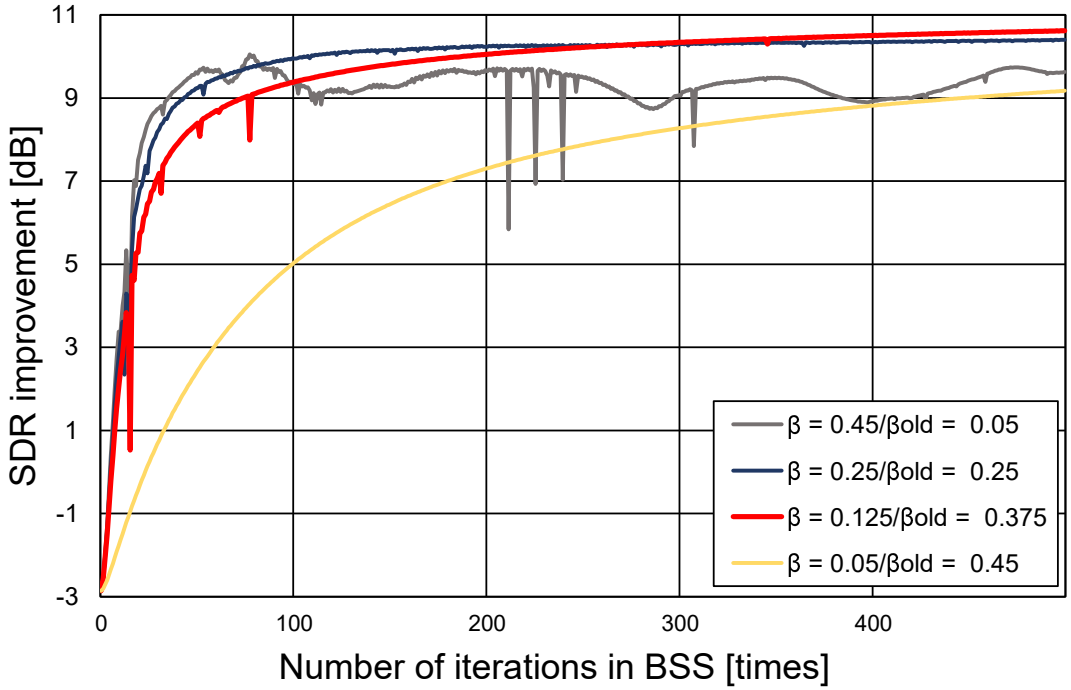


Fig. 4.16. Example of convergence behaviors of proposed method 1 with various β_{old} and β (song no. 15).

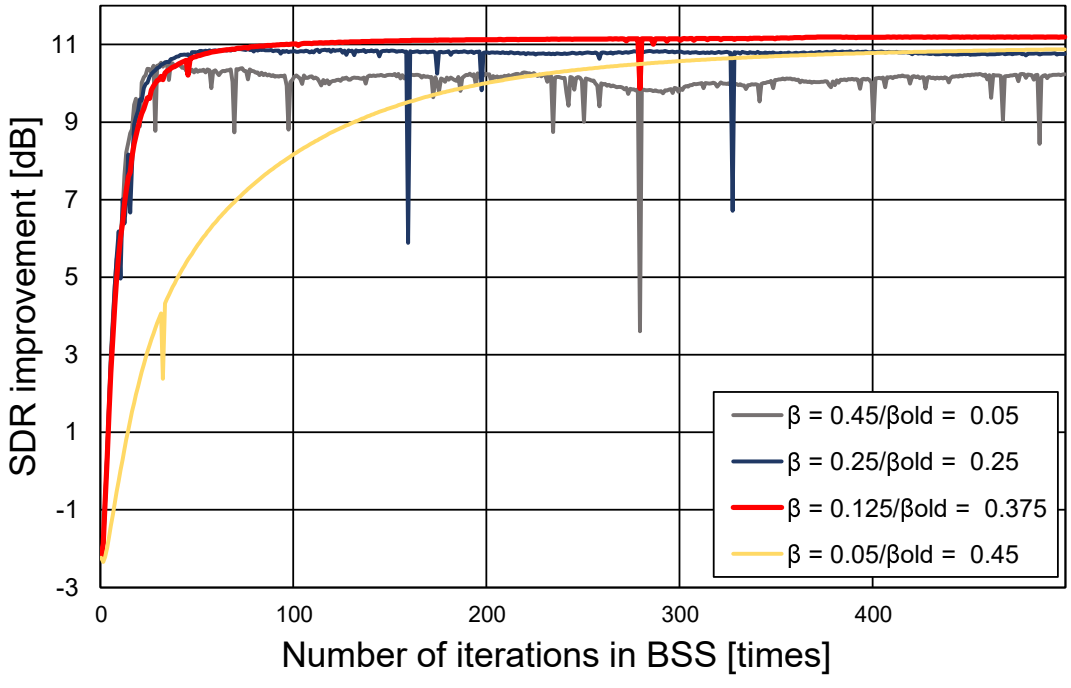


Fig. 4.17. Example of convergence behaviors of proposed method 1 with various β_{old} and β (song no. 16).

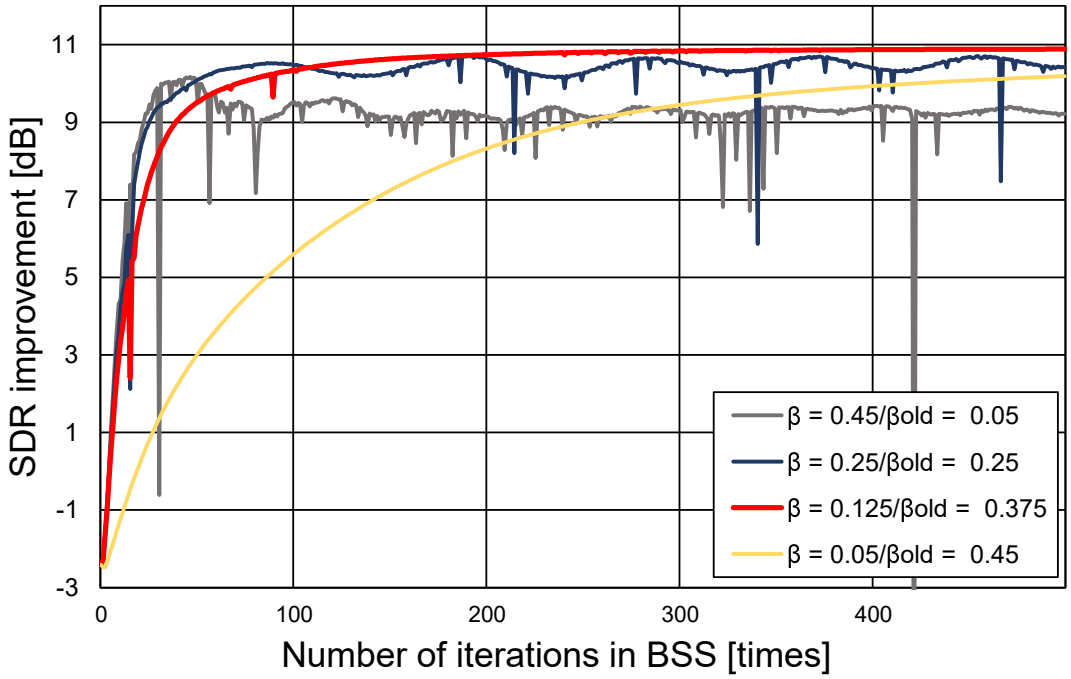


Fig. 4.18. Example of convergence behaviors of proposed method 1 with various β_{old} and β (song no. 17).

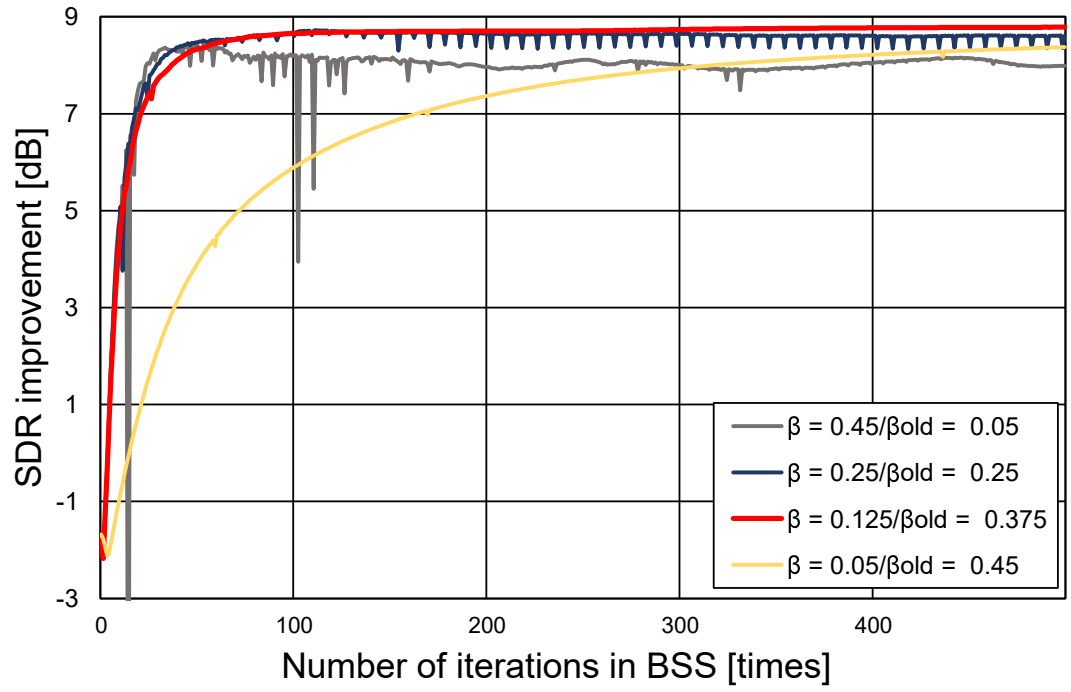


Fig. 4.19. Example of convergence behaviors of proposed method 1 with various β_{old} and β (song no. 18).

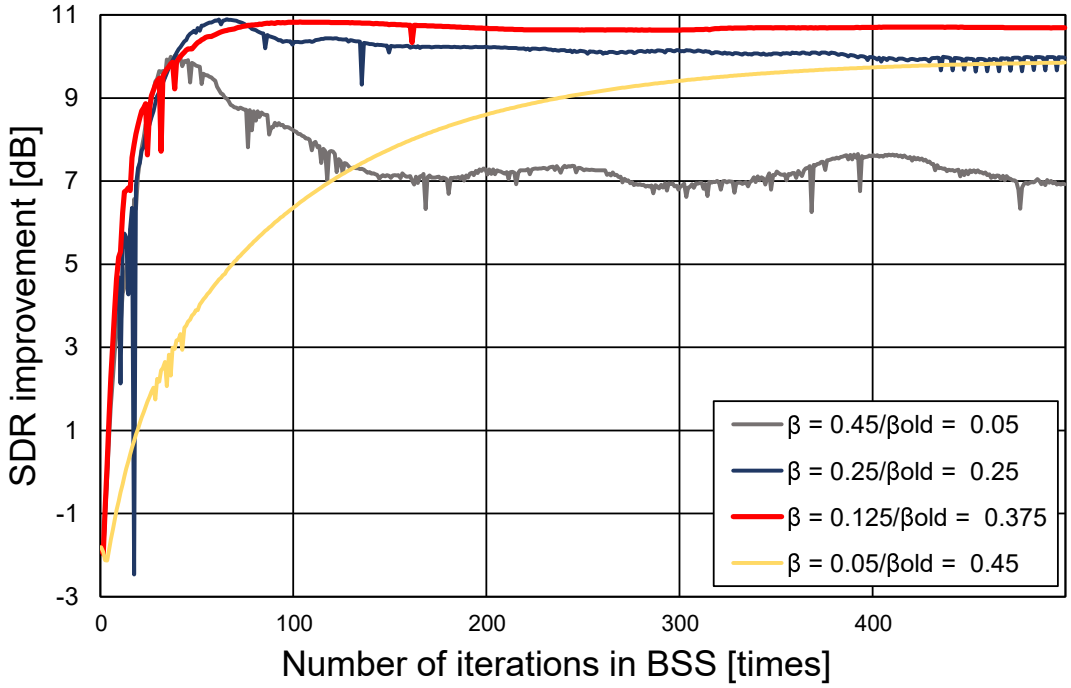


Fig. 4.20. Example of convergence behaviors of proposed method 1 with various β_{old} and β (song no. 19).

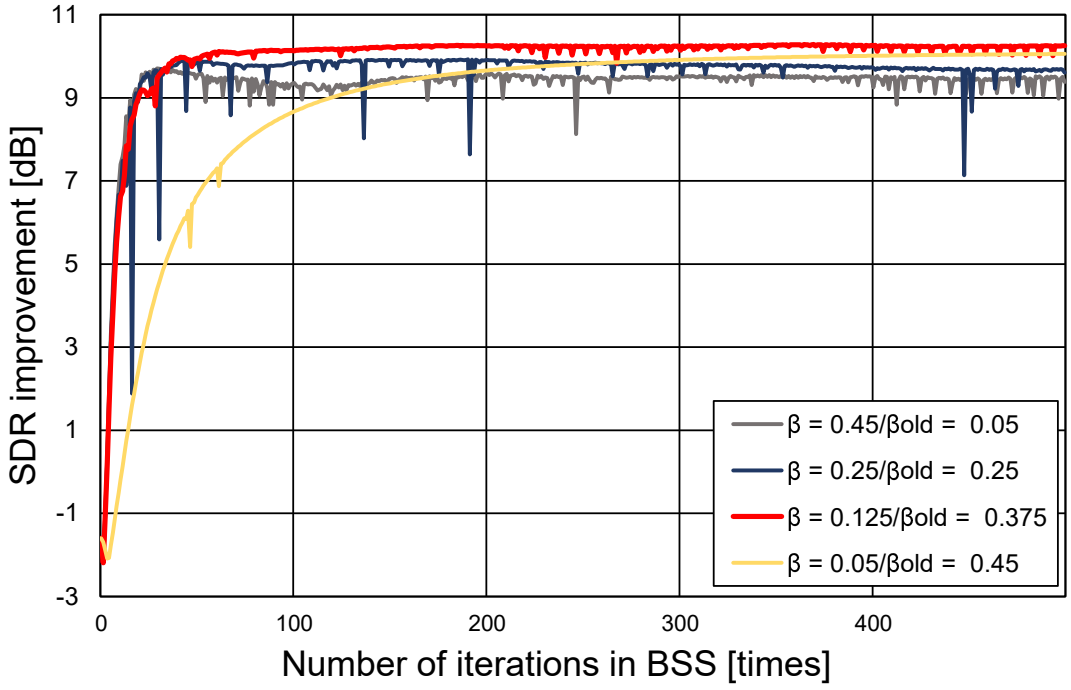


Fig. 4.21. Example of convergence behaviors of proposed method 1 with various β_{old} and β (song no. 20).

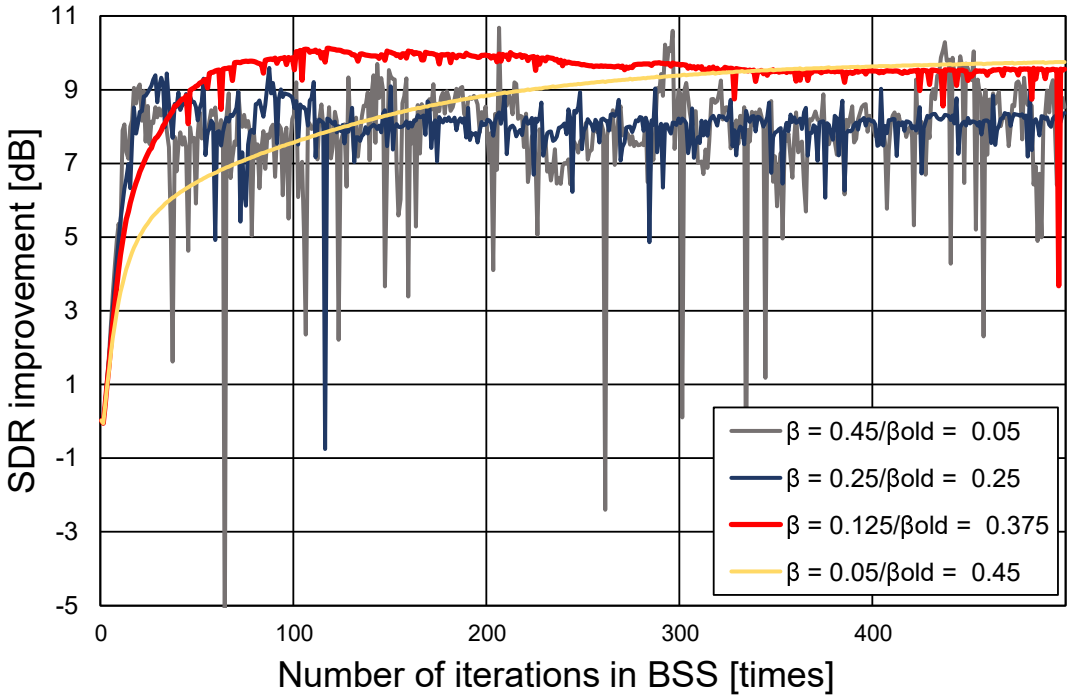


Fig. 4.22. Example of convergence behaviors of proposed method 2 with various β_{old} and β (song no. 1).

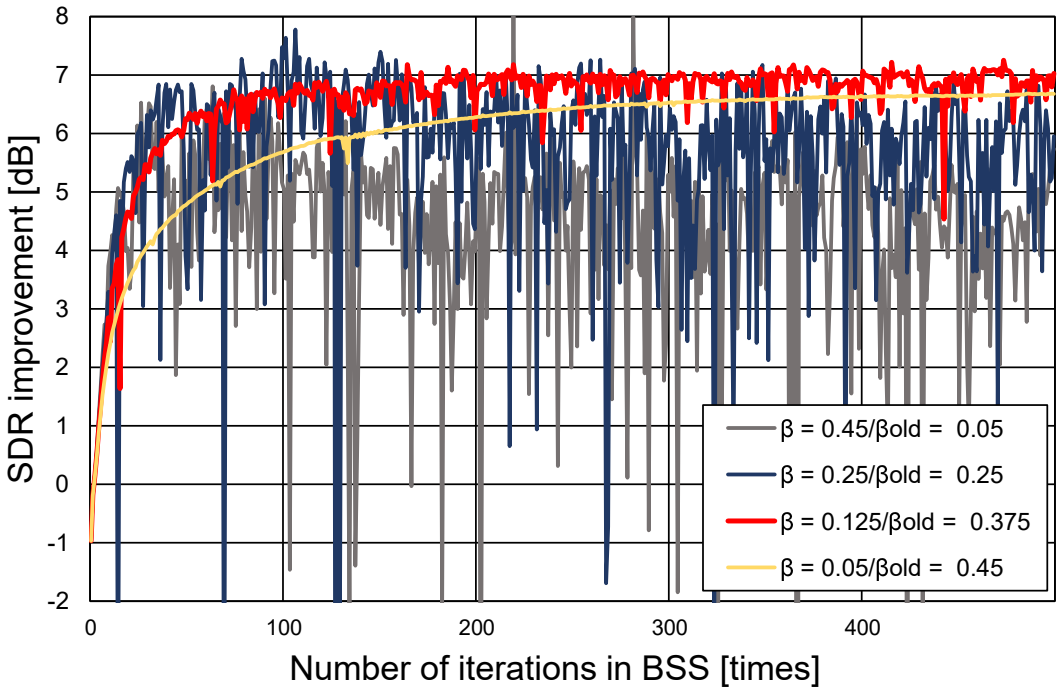


Fig. 4.23. Example of convergence behaviors of proposed method 2 with various β_{old} and β (song no. 2).

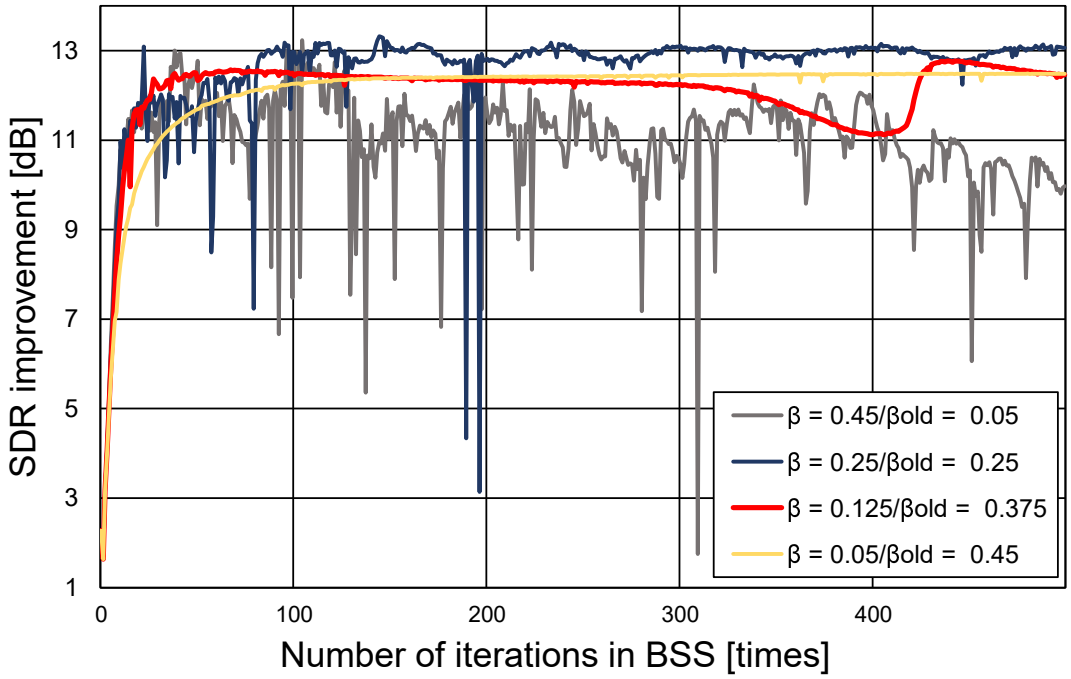


Fig. 4.24. Example of convergence behaviors of proposed method 2 with various β_{old} and β (song no. 3).

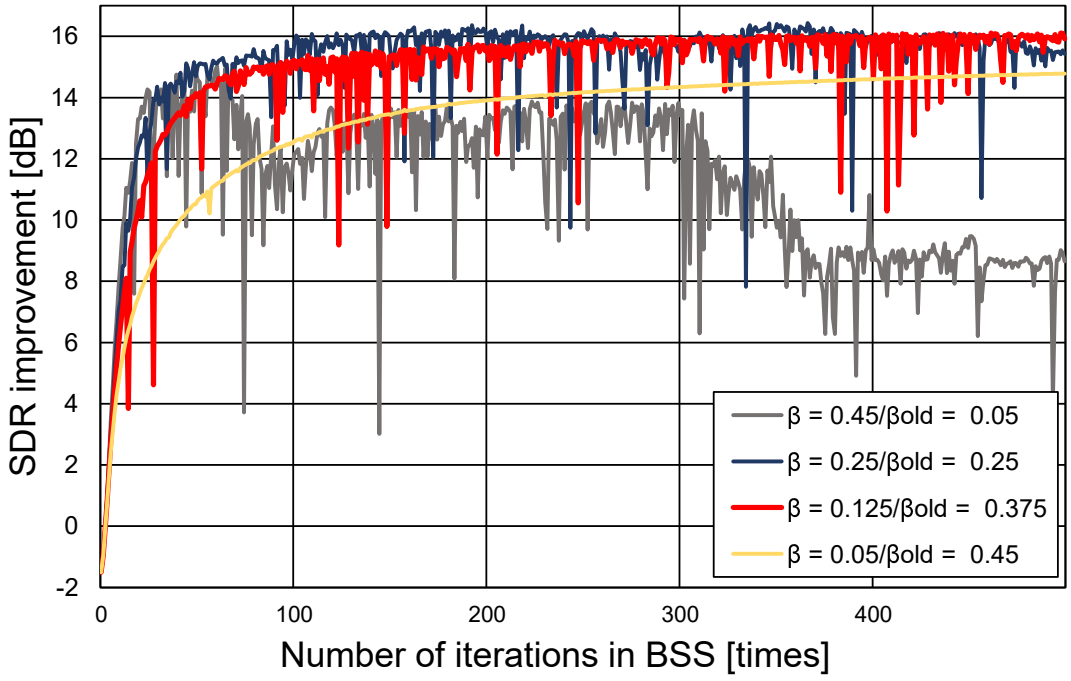


Fig. 4.25. Example of convergence behaviors of proposed method 2 with various β_{old} and β (song no. 4).

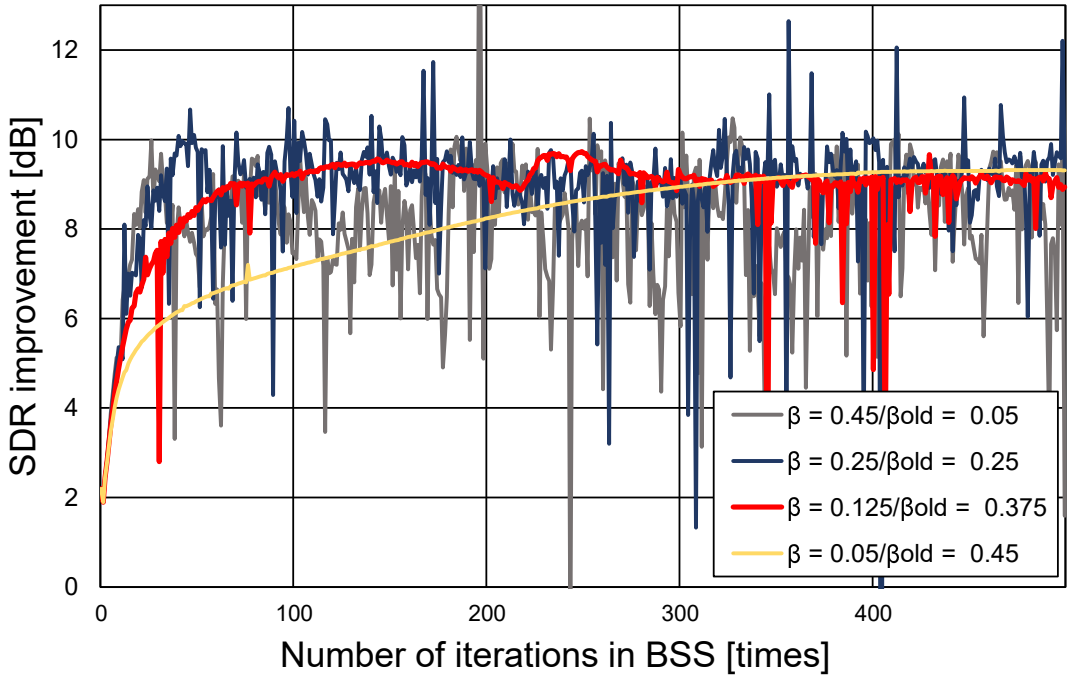


Fig. 4.26. Example of convergence behaviors of proposed method 2 with various β_{old} and β (song no. 5).

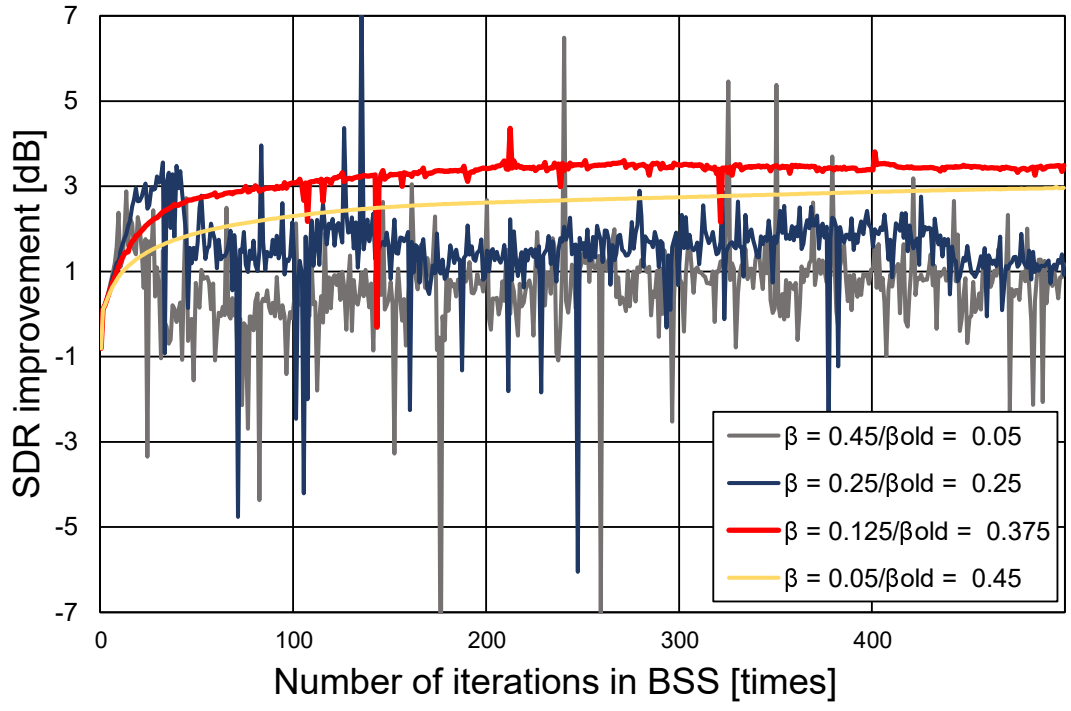


Fig. 4.27. Example of convergence behaviors of proposed method 2 with various β_{old} and β (song no. 6).

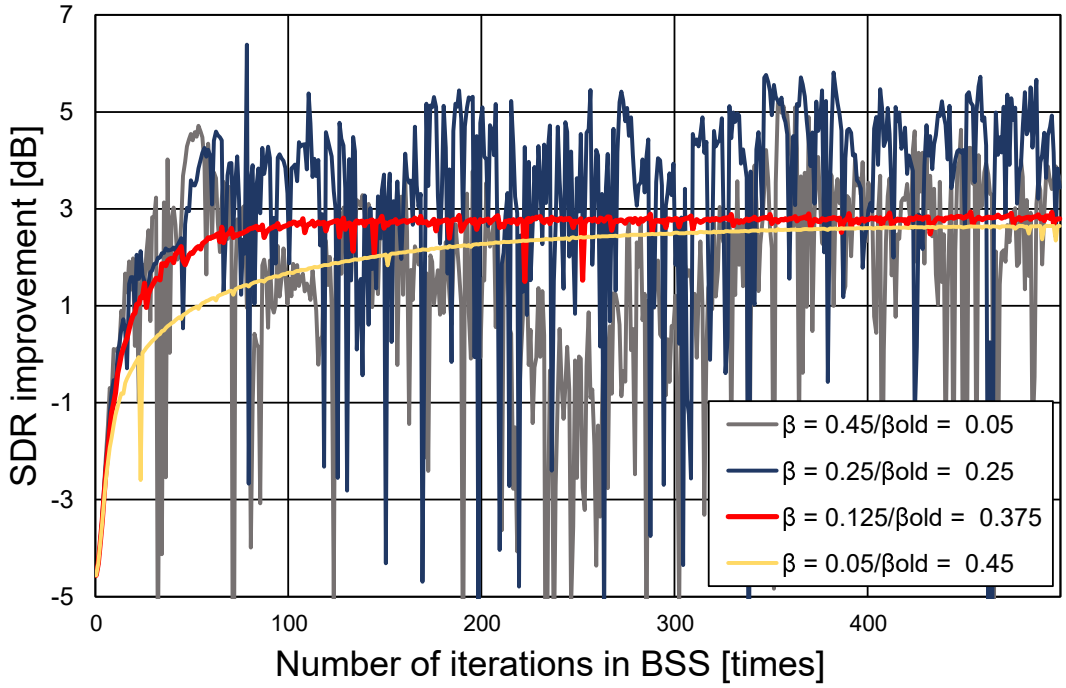


Fig. 4.28. Example of convergence behaviors of proposed method 2 with various β_{old} and β (song no. 7).

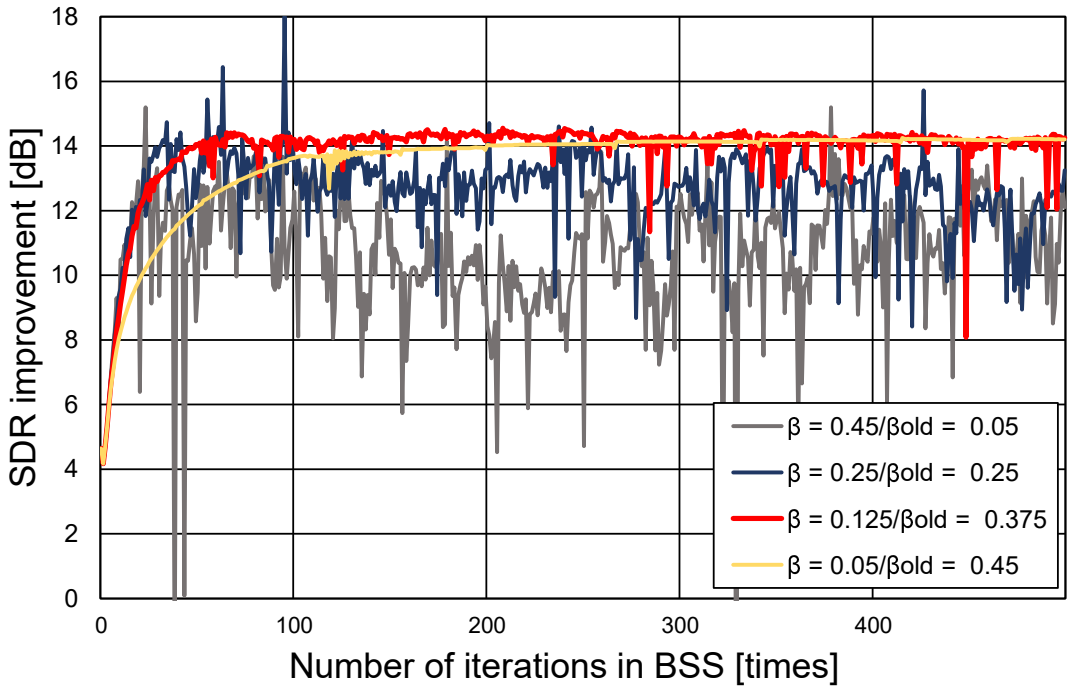


Fig. 4.29. Example of convergence behaviors of proposed method 2 with various β_{old} and β (song no. 8).

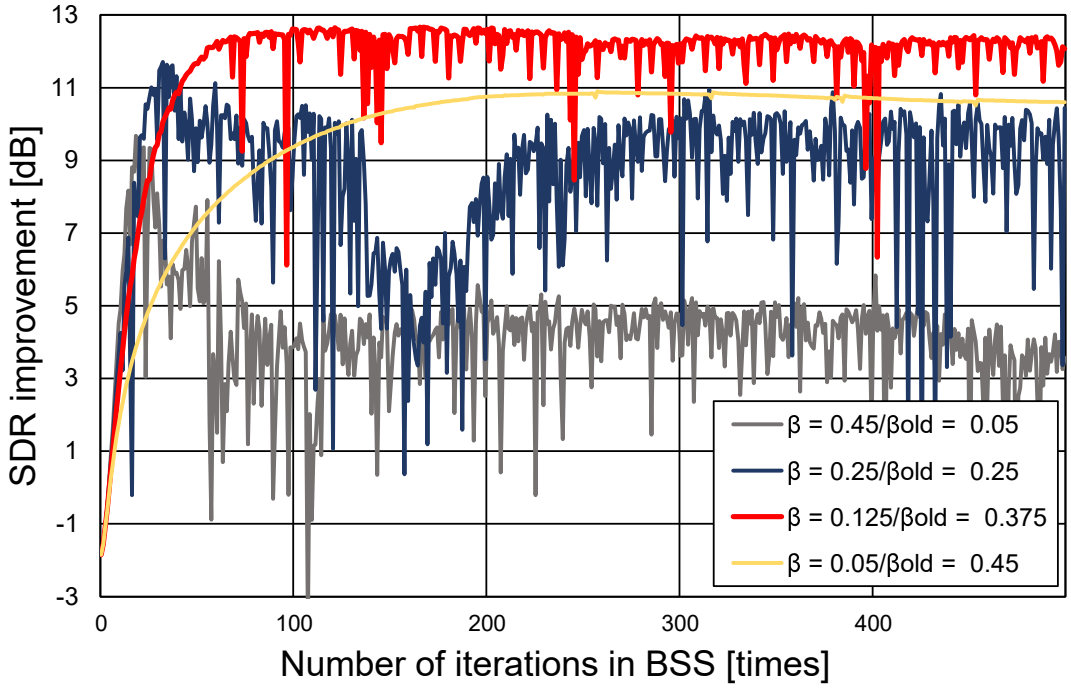


Fig. 4.30. Example of convergence behaviors of proposed method 2 with various β_{old} and β (song no. 9).

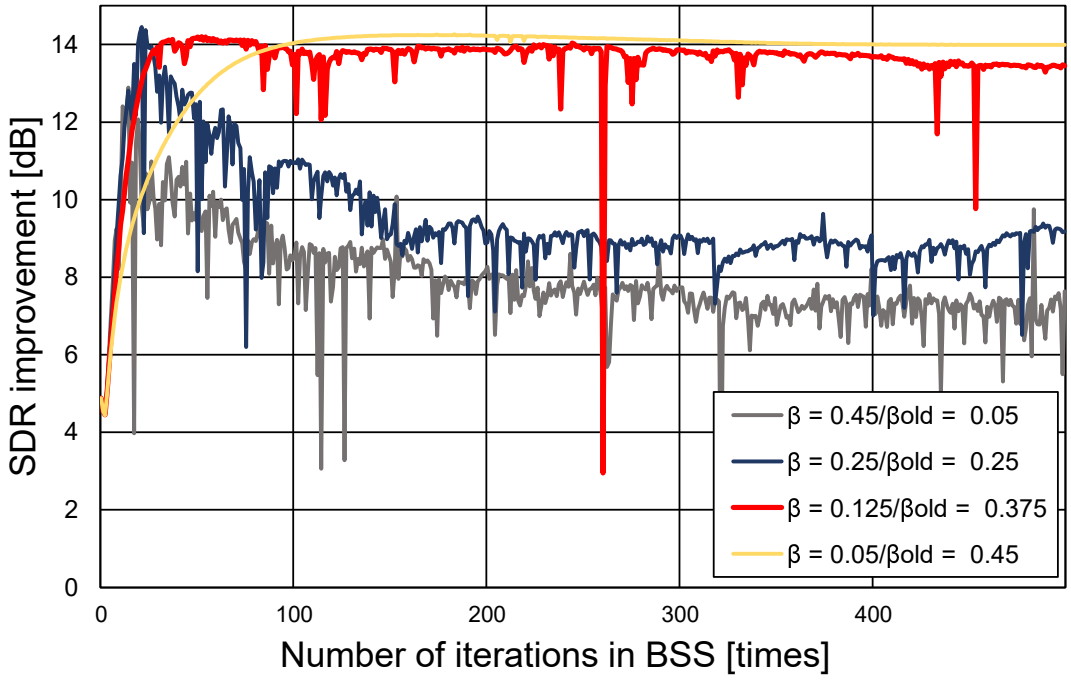


Fig. 4.31. Example of convergence behaviors of proposed method 2 with various β_{old} and β (song no. 10).

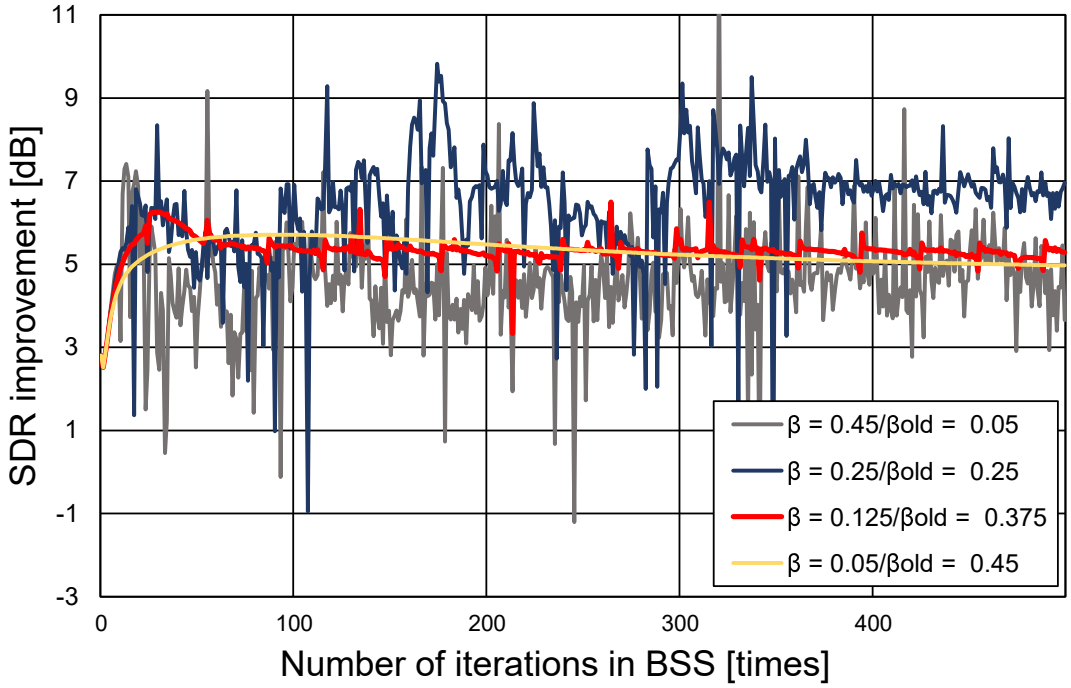


Fig. 4.32. Example of convergence behaviors of proposed method 2 with various β_{old} and β (song no. 11).

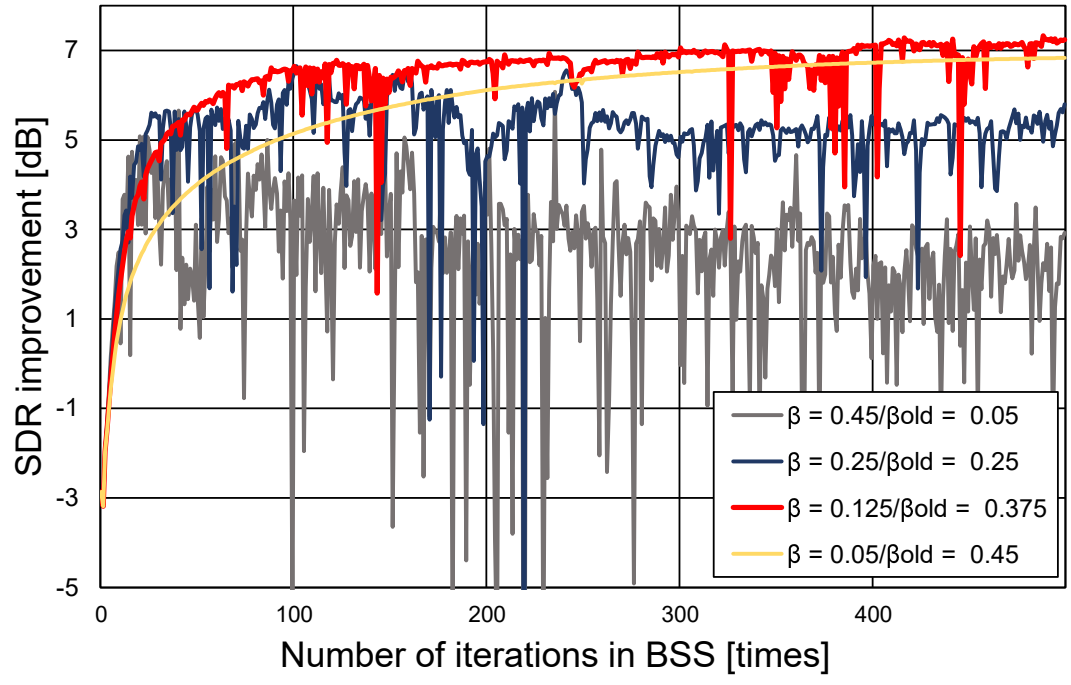


Fig. 4.33. Example of convergence behaviors of proposed method 2 with various β_{old} and β (song no. 12).

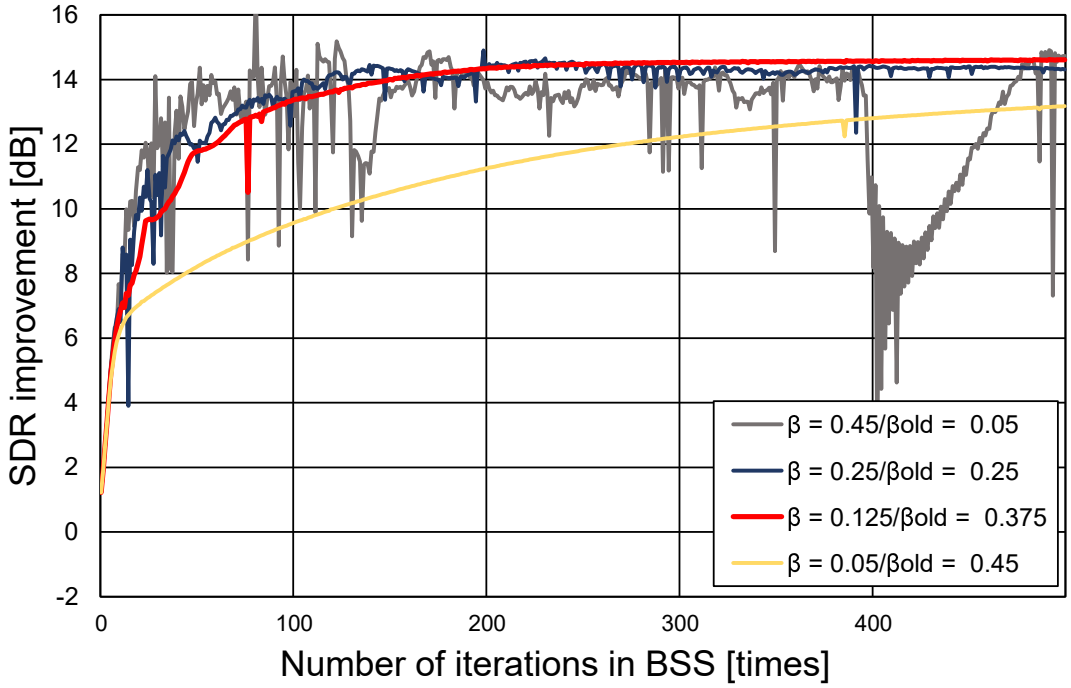


Fig. 4.34. Example of convergence behaviors of proposed method 2 with various β_{old} and β (song no. 13).

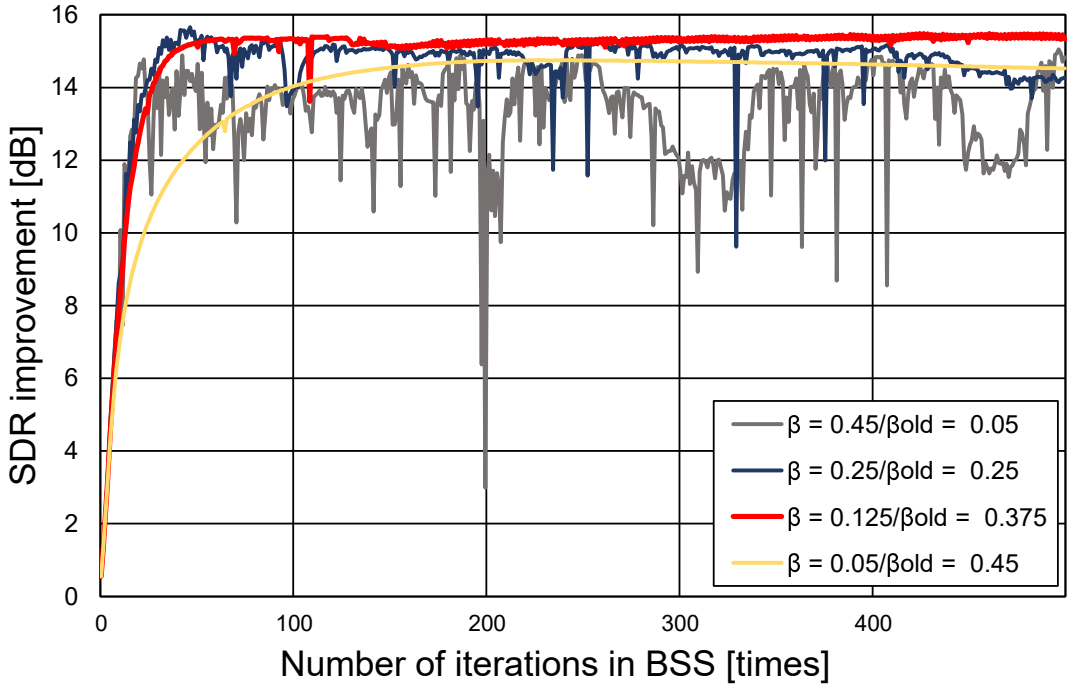


Fig. 4.35. Example of convergence behaviors of proposed method 2 with various β_{old} and β (song no. 14).

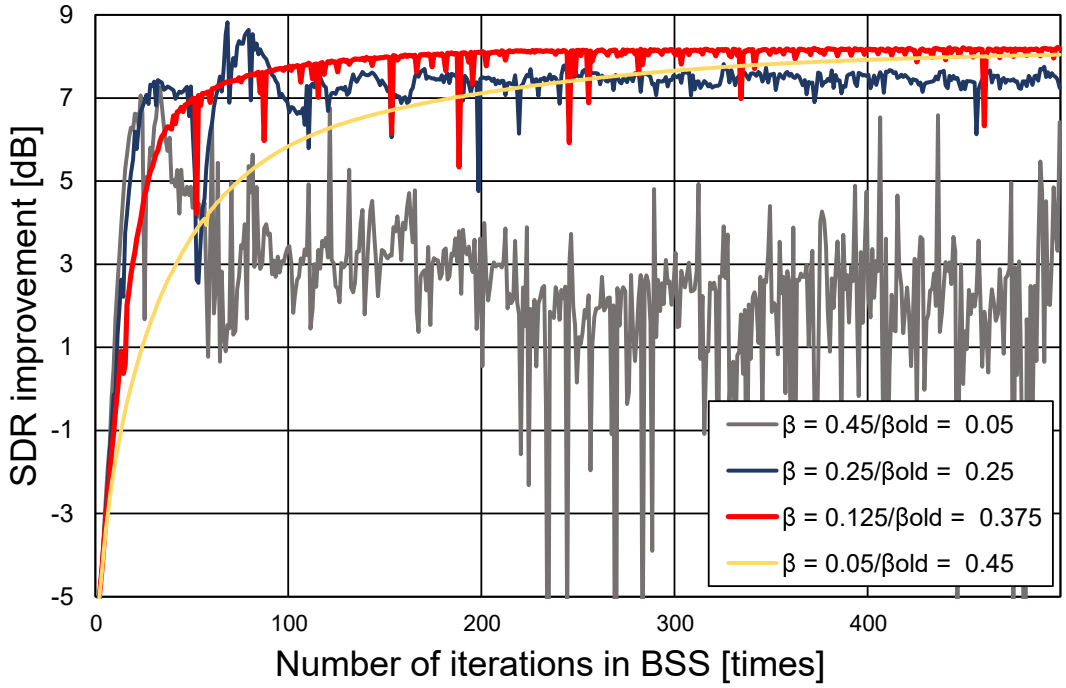


Fig. 4.36. Example of convergence behaviors of proposed method 2 with various β_{old} and β (song no. 15).

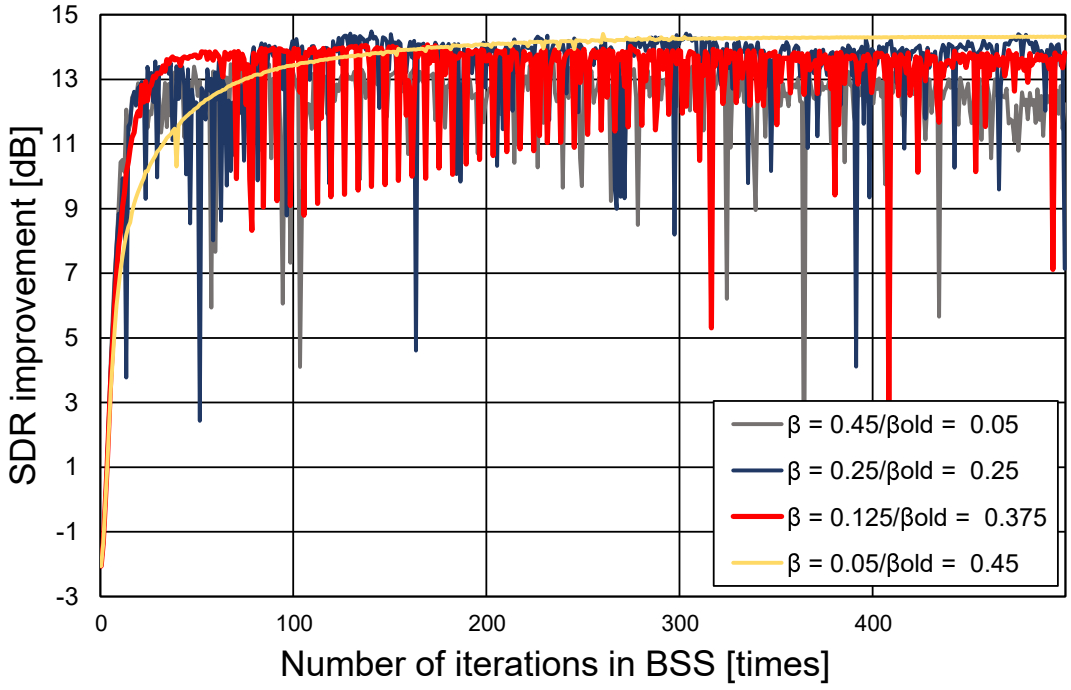


Fig. 4.37. Example of convergence behaviors of proposed method 2 with various β_{old} and β (song no. 16).

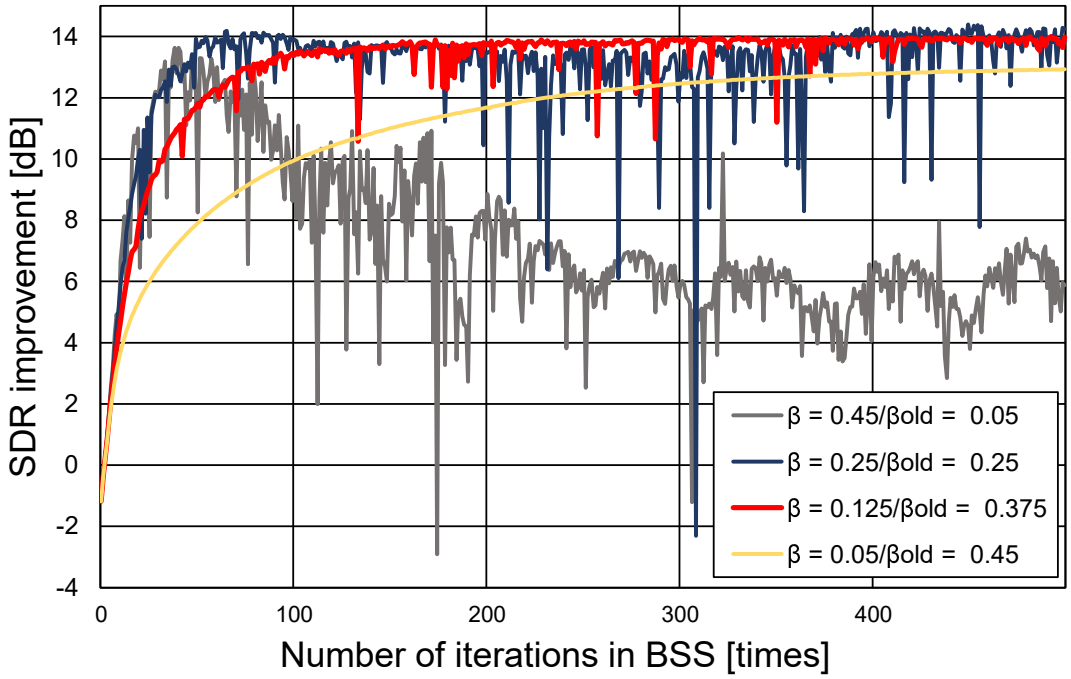


Fig. 4.38. Example of convergence behaviors of proposed method 2 with various β_{old} and β (song no. 17).

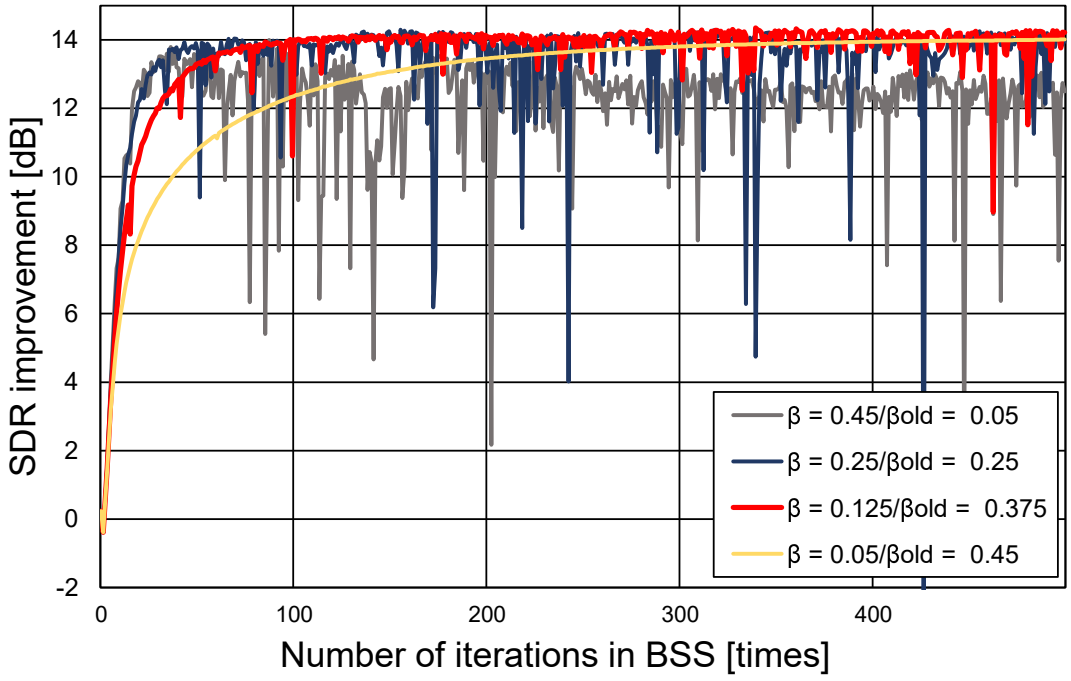


Fig. 4.39. Example of convergence behaviors of proposed method 2 with various β_{old} and β (song no. 18).

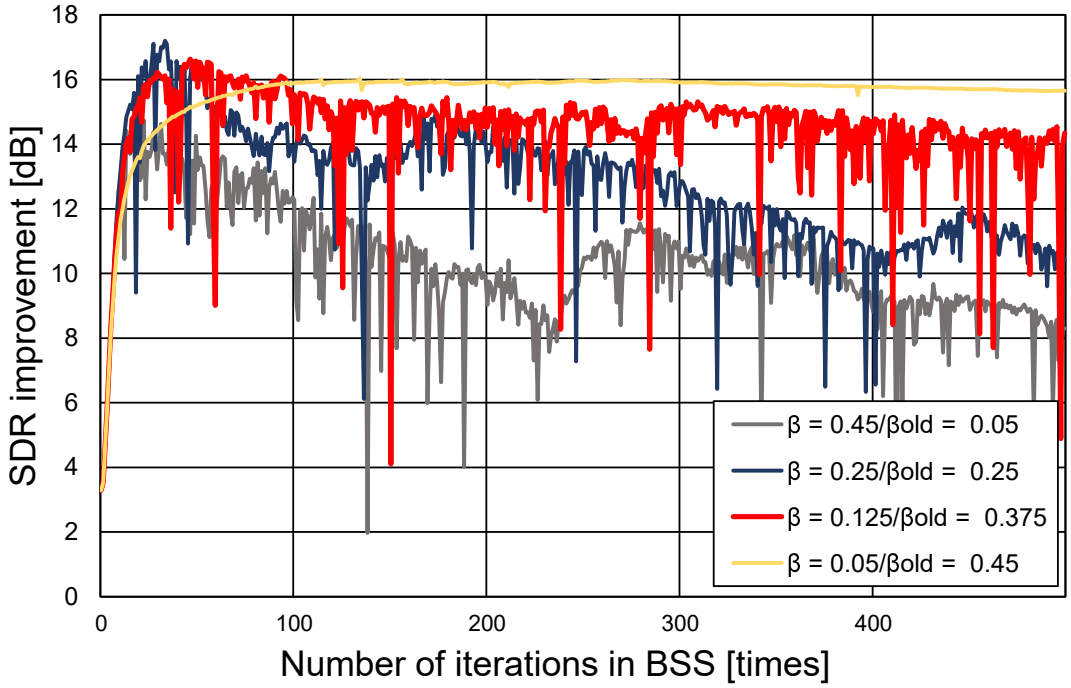


Fig. 4.40. Example of convergence behaviors of proposed method 2 with various β_{old} and β (song no. 19).

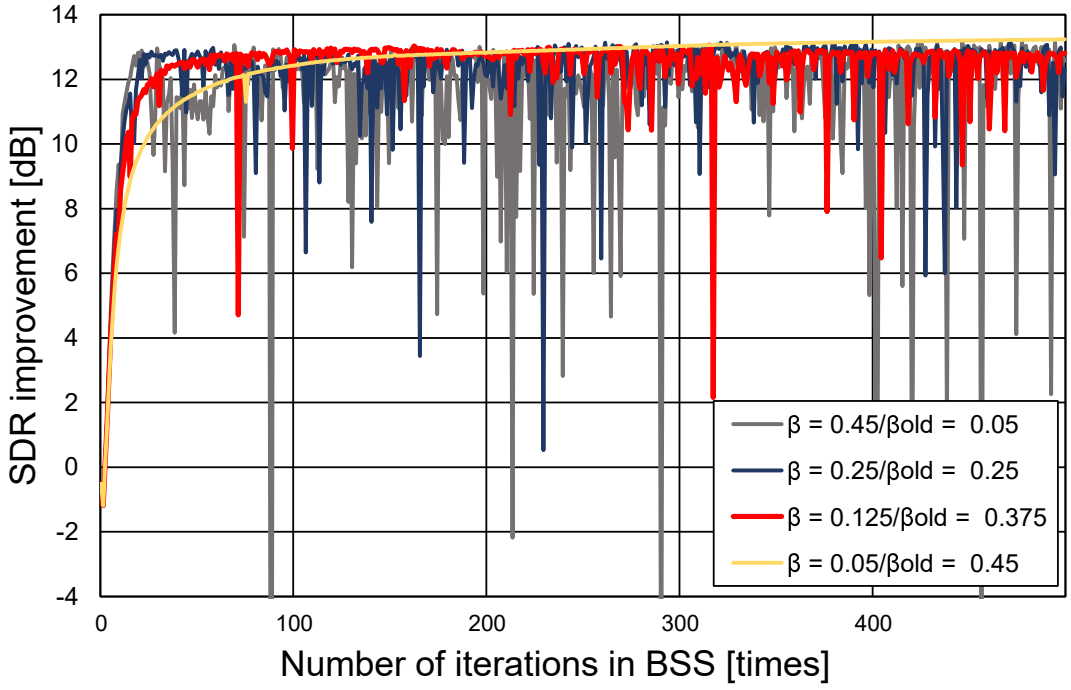


Fig. 4.41. Example of convergence behaviors of proposed method 2 with various β_{old} and β (song no. 20).

第 5 章

他の従来手法との性能比較実験

5.1　性能比較実験

5.1.1　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（前述のその他の音源（other）に該当）の音源分離実験を行い他の既存手法と比較する．本実験では前章の実験と同様に，音源は SiSEC2016 [20] のデータセット中のドラム音源とその他の音源を 20 曲選び，インパルス応答で畳み込みの後，多チャネル混合信号を作成した．その他の実験条件も同様に Table 4.1 に示す通りで．評価指標についても，前章と同様に SDR を用いた．

5.1.2　実験結果

データセット 20 曲を対象とし音源分離実験を行い, 各従来手法と比較した結果を曲ごとに Figs. 5.1–5.7 に示す．ここで，HPSS+TFMBSS が提案手法を示す．Figs. 5.1–5.7 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，全楽曲通して従来の 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．この特性は特に提案手法 1 において色濃く反映されている．提案手法 2 では傾向は見受けられるものの，Song no.9 及び 10 のように HPSS が苦手な楽曲でもある程度分離される，または苦手な楽曲でも高いスコアを出す例も観測された．さらに，Song no.3，5，6 及び 13 のように 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も観測された．提案手法 1 では Song no.7，12 及び 16 において従来法より劣る結果であったが，提案手法 2 ではどの楽曲においても従来法よりも SDR 改善量が上回った．

Table 5.1 は，データセット 20 曲全てにおける提案手法と各従来手法との SDR 改善量の平均値の比較である．従来手法の HPSS と比較すると提案手法 1 及び 2 共に音質の向上を確認した．提案手法 1 では他の従来法に比べ平均スコアは下回った．しかし，提案手法 2 では従来法よりも平均スコアが上回った．

Table. 5.1. Average SDR for each method

Method	Average SDR [dB]
HPSS	4.68
IVA	7.09
ILRMA	8.56
HPSS+TFMBSS(method1)	7.44
HPSS+TFMBSS(method2)	11.0

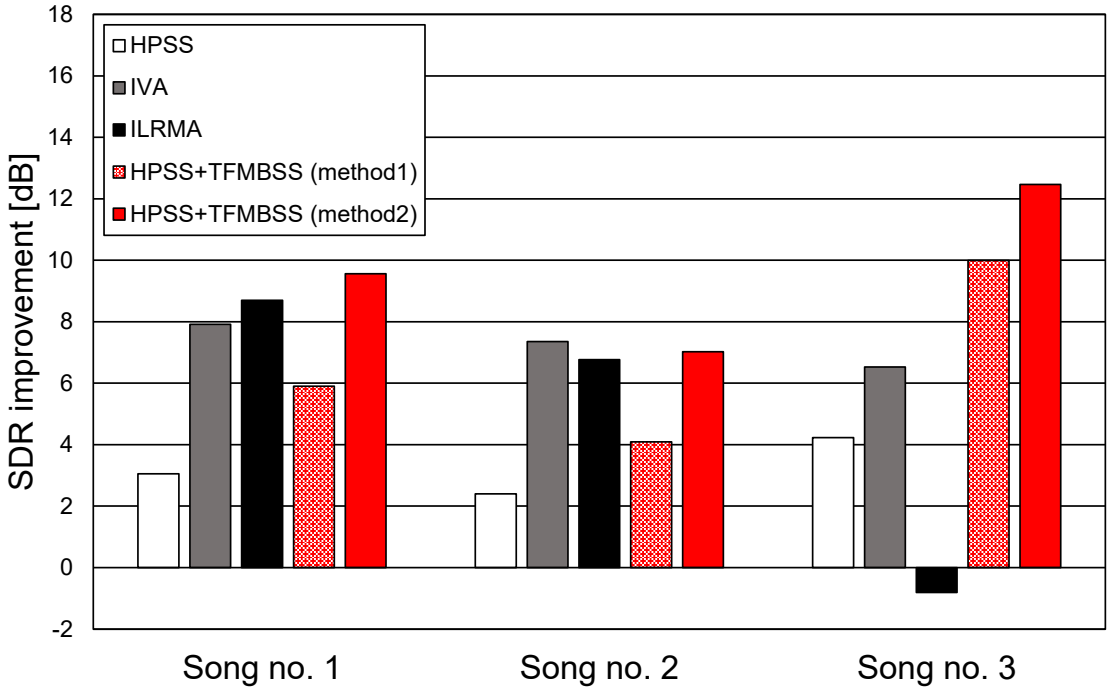


Fig. 5.1. Example of SDR improvements of ILRMA, IVA, conventional HPSS, and proposed methods (song nos. 1–3).

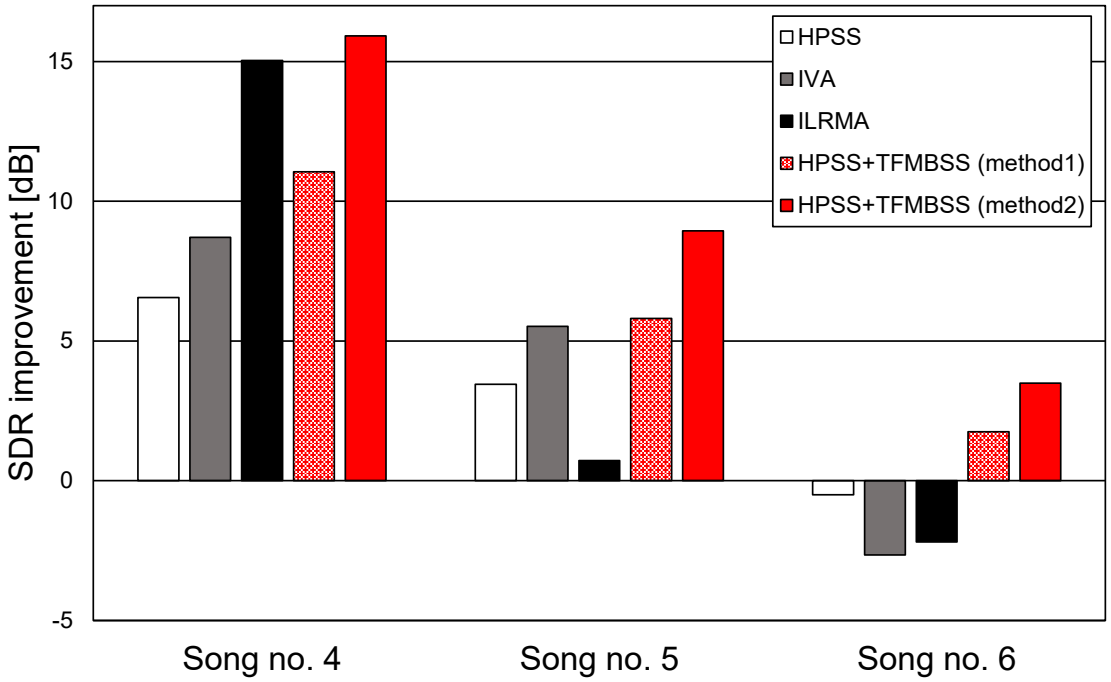


Fig. 5.2. Example of SDR improvements of ILRMA, IVA, conventional HPSS, and proposed methods (song nos. 4–6).

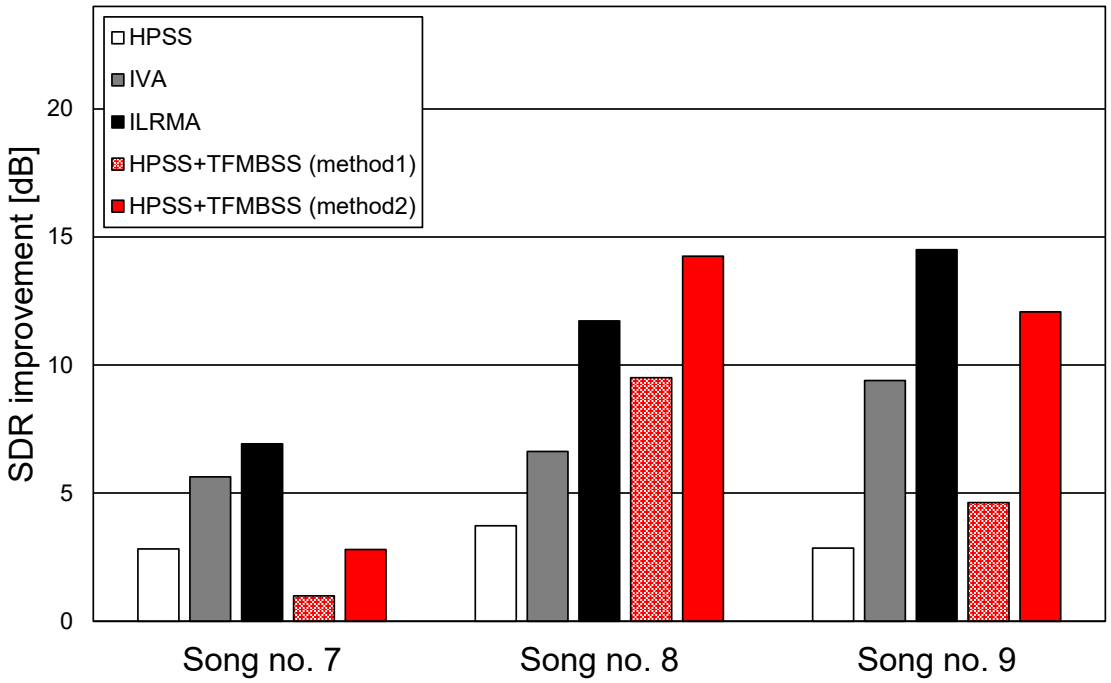


Fig. 5.3. Example of SDR improvements of ILRMA, IVA, conventional HPSS, and proposed methods (song nos. 7–9).

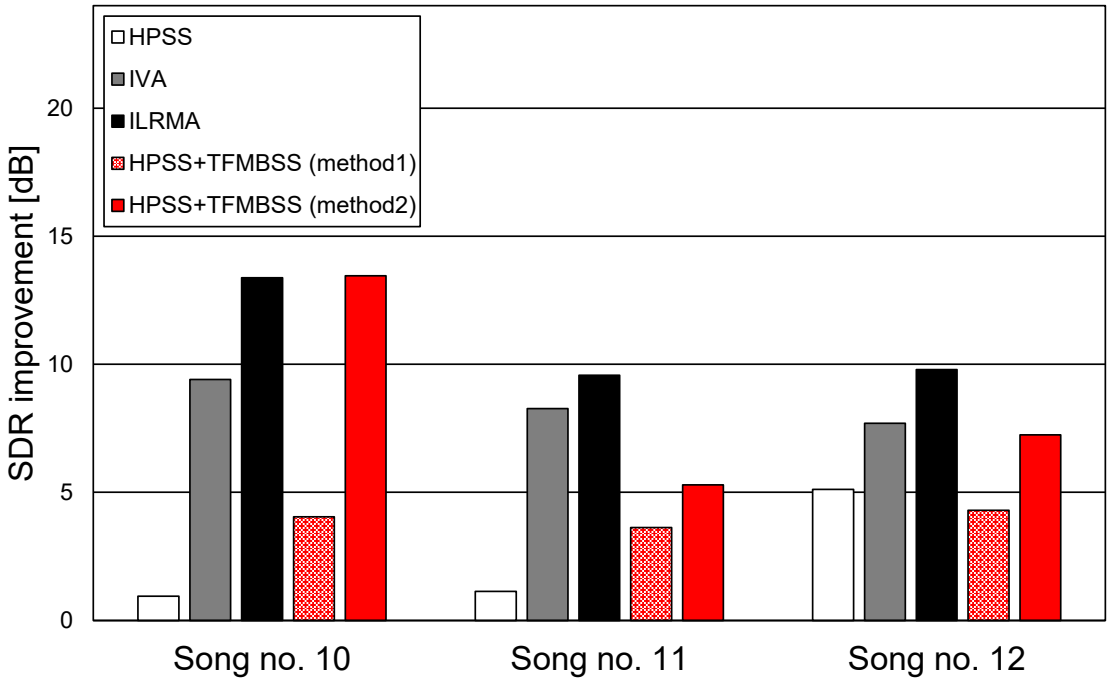


Fig. 5.4. Example of SDR improvements of ILRMA, IVA, conventional HPSS, and proposed methods (song nos. 10–12).

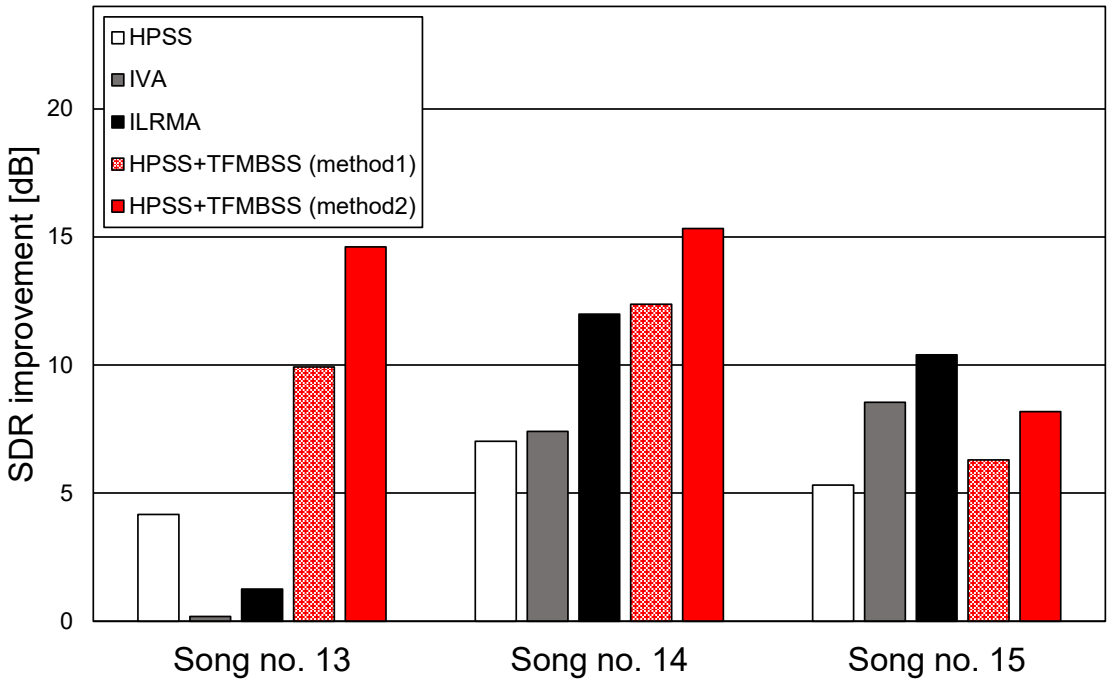


Fig. 5.5. Example of SDR improvements of ILRMA, IVA, conventional HPSS, and proposed methods (song nos. 13–15).

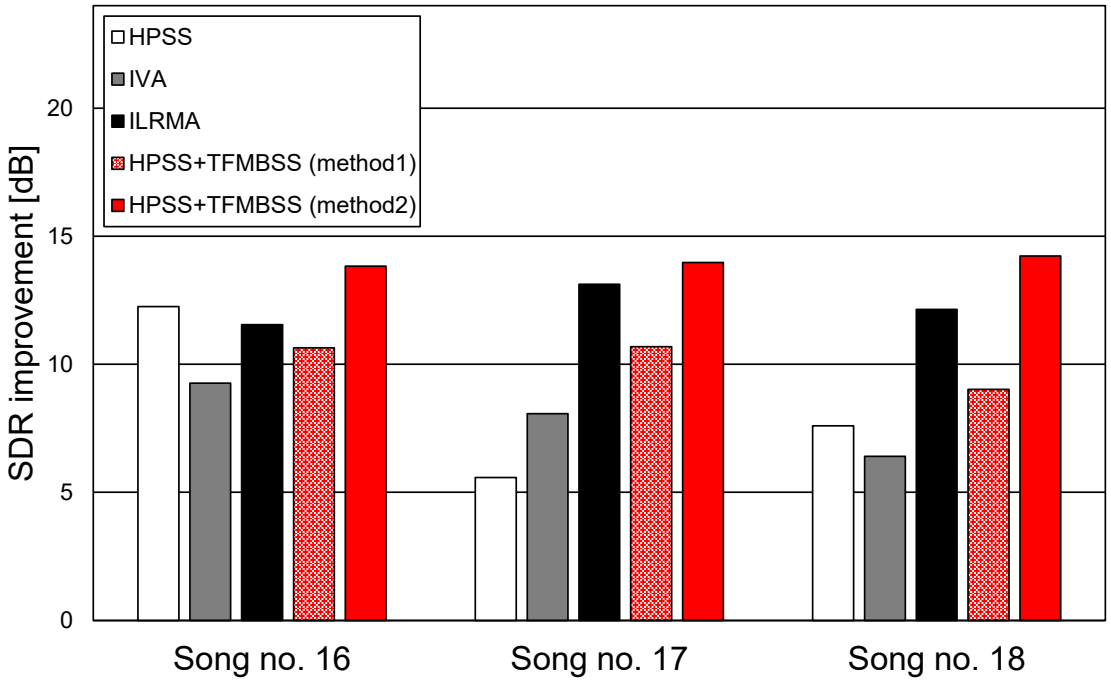


Fig. 5.6. Example of SDR improvements of ILRMA, IVA, conventional HPSS, and proposed methods (song nos. 16–18).

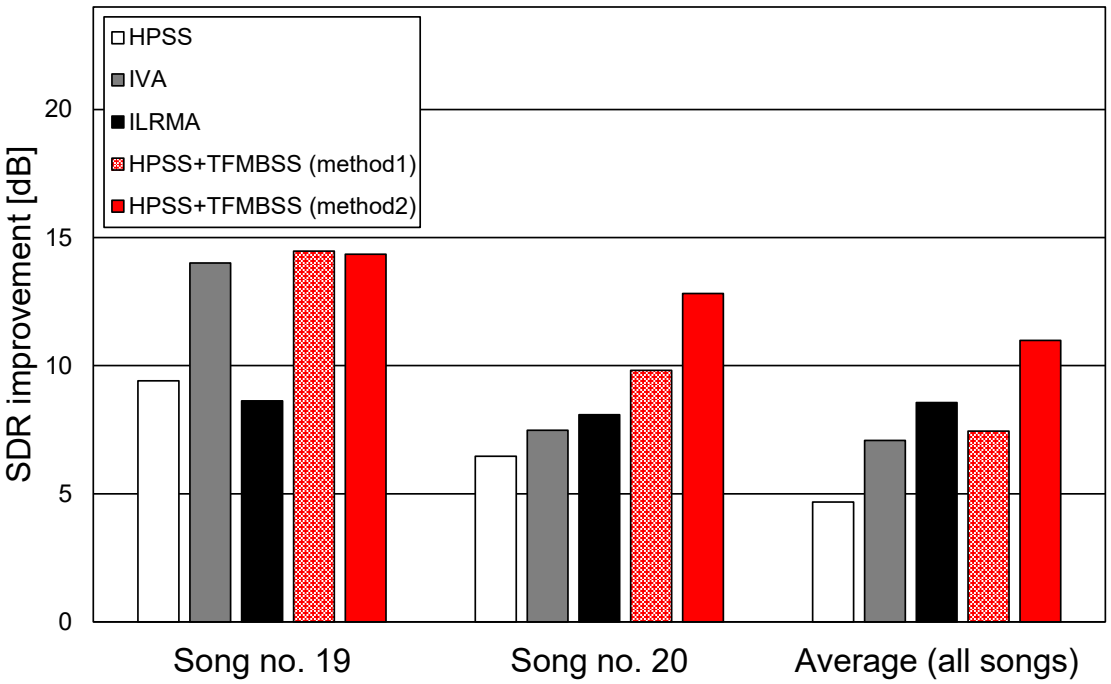


Fig. 5.7. Example of SDR improvements of ILRMA, IVA, conventional HPSS, and proposed methods (song nos. 19 and 20 and average of all songs).

5.2　本章のまとめ

本章では，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い他の手法と比較した．２つの提案手法共に，HPSS の分離特性を反映するため HPSS が分離できない楽曲はスコアが振るわない結果となった．しかし，提案手法１では最終的な平均スコアが他の従来手法と比べて下回ったものの，提案手法２では上回る結果となった．次章では，本論文における総括とした結論を述べる．

第 6 章

結言

本論文では，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 SDR 推移の安定と収束速度のトレードオフを考慮した適当な β_{old} 及び β を設定すること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．

最後に今後の展望を述べる．今回では HPSS を使って調波音と打撃音に分離したが他のモノラル音源分離アルゴリズム用いて調波音と打撃音に分離することも可能である．そのため，本論文の実験結果と他のアルゴリズム用いて分離した結果を比較し，より良い調波音と打撃音における音源モデルの探求が求められる．

謝辞

本論文は，香川高等専門学校電気情報工学科北村研究室にて行われた研究に基づくものです．

まず，本研究を進めるにあたり，ご多忙のところ熱心にご指導くださいました指導教員の北村大地助教に心より感謝申し上げます．北村大地助教には，本研究分野における基礎的な知識から研究に関する詳細な議論など，細部にわたるまで丁寧にご指導いただきました．さらに，細かなミーティングや論文執筆のスケジュール段取り及び，手厚い指導によりとても良い環境下で研究を進めることができました．

本論の副査である村上幸一准教授には，論文の構成や記述に関して大変有益な助言を頂き，大変お世話になりました．ここに厚く御礼申し上げます．

早稲田大学の矢田部浩平講師には，共同研究を通じミーティング及び論文添削での細かな指摘や，多数の知識のご教授を頂きました．心より感謝申し上げます．

北村研究室の先輩である専攻科 1 年の山地修平氏には，研究に対する心持ちや向き合い方など先輩研究生としての経験に基づくアドバイス等をはじめ，日頃の生活面でも数々のご支援をいただきました．また，北村研究室同期の渡辺瑠伊氏には，普段から専門的知識及び情報分野における一般的な知見の確立に多くのアドバイスを頂きました．更に，書面作成におけるまとまったデザイン性や研究におけるソフトウェア環境の参考として多くのヒントを日頃から頂きました．ここに感謝申し上げます．

最後になりますが，現在に至るまで私の学生生活を金銭的に支え，暖かく見守って下さった両親には感謝の念に堪えません．これまで本当にありがとうございました．

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] H. Sawada, R. Mukai, S Araki, and S. Makino, “A robust and precise method for solving the permutation problem of frequency-domain blind source separation,” *IEEE Trans. Speech and Audio Processing*, 12(5), pp. 530–538, 2004.

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

[5] N. Ono, “Stable and fast update rules for independent vector analysis based on auxiliary function technique,” *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 189–192, 2011.

[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. Workshop on Applications of Signal Processing to Audio and Acoustics*, 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] 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.

[9] D. D. Lee, and H. S. Seung, “Algorithms for non-negative matrix factorization ,” *Proc. Neural Information Processing Systems*, pp. 556–562, 2000.

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

[11] N. Parikh and S. Boyd, “Proximal algorithms,” *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[12] N. Komodakis and J. C. Pesquet, “Playing with duality: An overview of recent

primal-dual approaches for solving large scale optimization problems," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 31–54, 2015.

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

[14] K. Yatabe and D. Kitamura, "Determined blind source separation via proximal splitting algorithm," *Proc. International Conference on Acoustics, Speech, and Signal Processing*, pp. 776–780, 2018.

[15] K. Yatabe and D. Kitamura, "Time-frequency-masking-based determined BSS with application to sparse IVA," *Proc. International Conference on Acoustics, Speech, and Signal Processing*, pp. 715–719, 2019.

[16] A. R. López, N. Ono, U. Remes, K. Palomäki, and M. Kurimo, "Designing multichannel source separation based on single-channel source separation," *Proc. International Conference on Acoustics, Speech, and Signal Processing*, pp. 469–473, 2015.

[17] 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. European Signal Processing Conference*, 2008.

[18] D. Kitamura, H. Saruwatari, K. Yagi, K. Shikano, Y. Takahashi, and K. Kondo, "Music signal separation based on supervised non-negative matrix factorization with orthogonal-ity and maximum-divergence penalties," *IEICE Trans. Fundamentals*, vol. E97-A, no. 5, pp.1113–1118, 2014.

[19] 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.

[20] 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. 13th International Conference on Latent Variable Analysis and Signal Separation*, pp. 323–332, 2017.

[21] 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. Language Resources and Evaluation Conference*, pp. 965–968, 2000.

[22] E. Vincent, R. Gribonval and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

発表文献一覧

国内学会

1. 大藪宗一郎, 北村大地, 矢田部浩平, “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,” 日本音響学会 2020 年春季研究発表会講演論文集, 3-1-16, 2020 (to appear).