Linear Multichannel Blind Source Separation Based on Time-Frequency Mask Obtained by Harmonic/Percussive Sound Separation

2381

Soichiro Oyabu, Daichi Kitamura (National Institute of Technology, Kagawa College, Japan), and Kohei Yatabe (Waseda University, Japan)

1. Background

- **Blind source separation (BSS)**

Audio source separation problem without any prior information or training

We only aim at the separation of harmonic and percussive audio sources

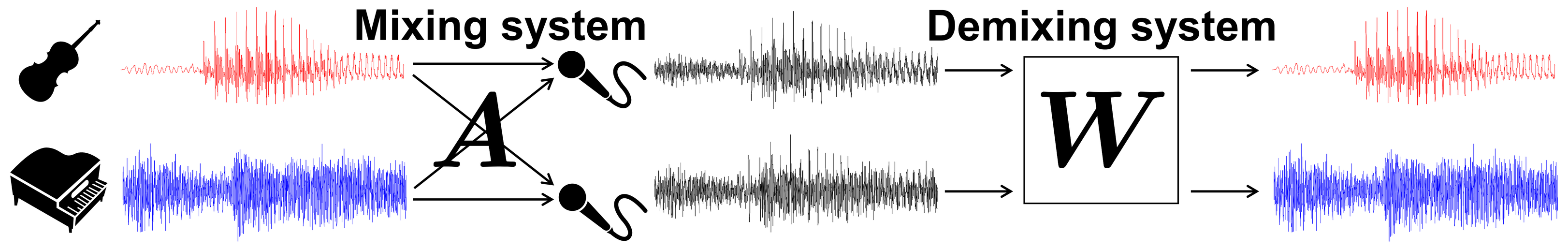
- **Single-channel BSS**

BSS problem for monaural-recorded audio signals (difficult)

Harmonic/percussive sound separation (HPSS) [Ono+, 2008], etc.

- **Multichannel BSS**

Blind estimation of demixing system $\boldsymbol{W}$ (inverse of mixing system $\boldsymbol{A}$)



High separation quality because of utilization of spatial features

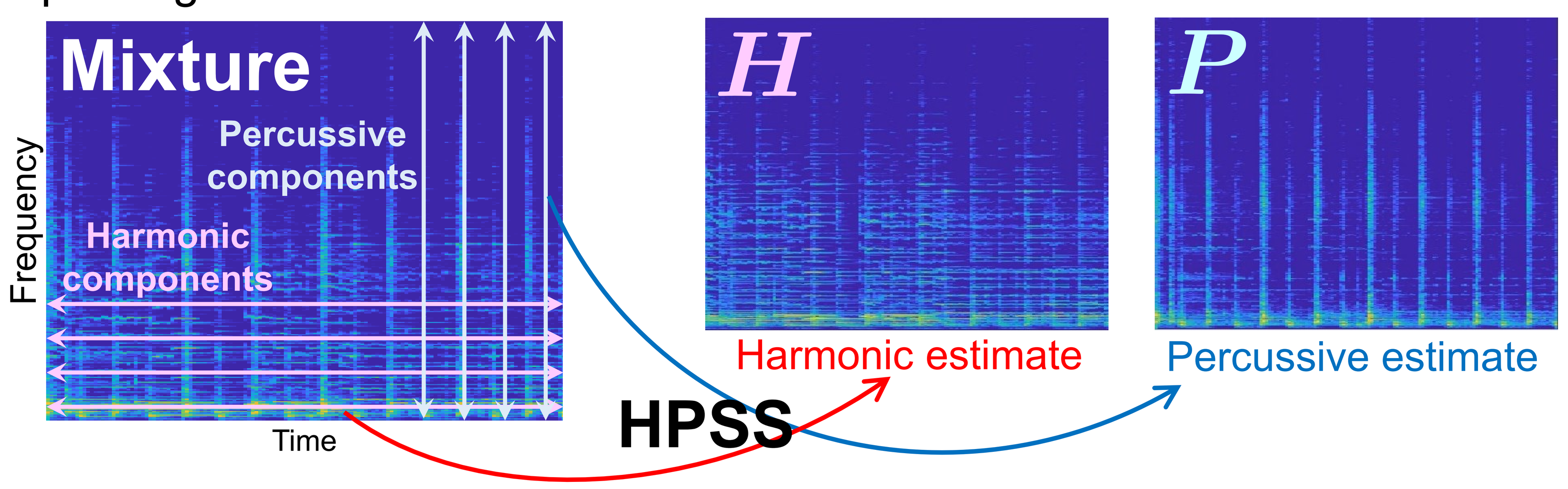
Independent vector analysis (IVA) [Hiroe+, 2006], [Kim+, 2006], [Ono, 2011]

Independent low-rank matrix (ILRMA) [Kitamura+, 2016]

Time-frequency-masking-based BSS (TFMBSS) [Yatabe+, 2019]

- **HPSS** [Ono+, 2008], [FitzGerald, 2010], [Duong+, 2011], [Tachibana+, 2012], etc.

BSS focusing on "**smoothness**" along with time or frequency directions in spectrogram



Estimate $\boldsymbol{H}$ and $\boldsymbol{P}$ by iteratively minimizing **smoothness cost function**

- **TFMBSS** [Yatabe+, 2019]

Linear multichannel BSS with **plug-and-play source models**

Source model is input as a time-frequency mask

Algorithm 1 TFMBSS

Input: $X, \mathbf{w}^{[1]}, \mathbf{y}^{[1]}, \mu_1, \mu_2, \alpha$

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

1: **for** $k = 1, \cdots, K$ **do**

2: $\quad \widetilde{\mathbf{w}} = \mathrm{prox}_{\mu_1 \mathcal{I}}[\mathbf{w}^{[k]} - \mu_1\mu_2 X^{\mathrm{H}}\mathbf{y}^{[k]}]$

3: $\quad \mathbf{z} = \mathbf{y}^{[k]} + X(2\widetilde{\mathbf{w}} - \mathbf{w}^{[k]})$

4: $\quad \mathcal{M} = \mathtt{generateMask}(\mathbf{z})$ — Generate time-frequency mask based on temporal estimated sources $\mathbf{z}$

5: $\quad \widetilde{\mathbf{y}} = \mathbf{z} - \mathcal{M} \odot \mathbf{z}$ — Masking process

6: $\quad \mathbf{y}^{[k+1]} = \alpha\widetilde{\mathbf{y}} + (1-\alpha)\mathbf{y}^{[k]}$

7: $\quad \mathbf{w}^{[k+1]} = \alpha\widetilde{\mathbf{w}} + (1-\alpha)\mathbf{w}^{[k]}$

8: **end for**

$\odot$: entrywise product

Our research aim

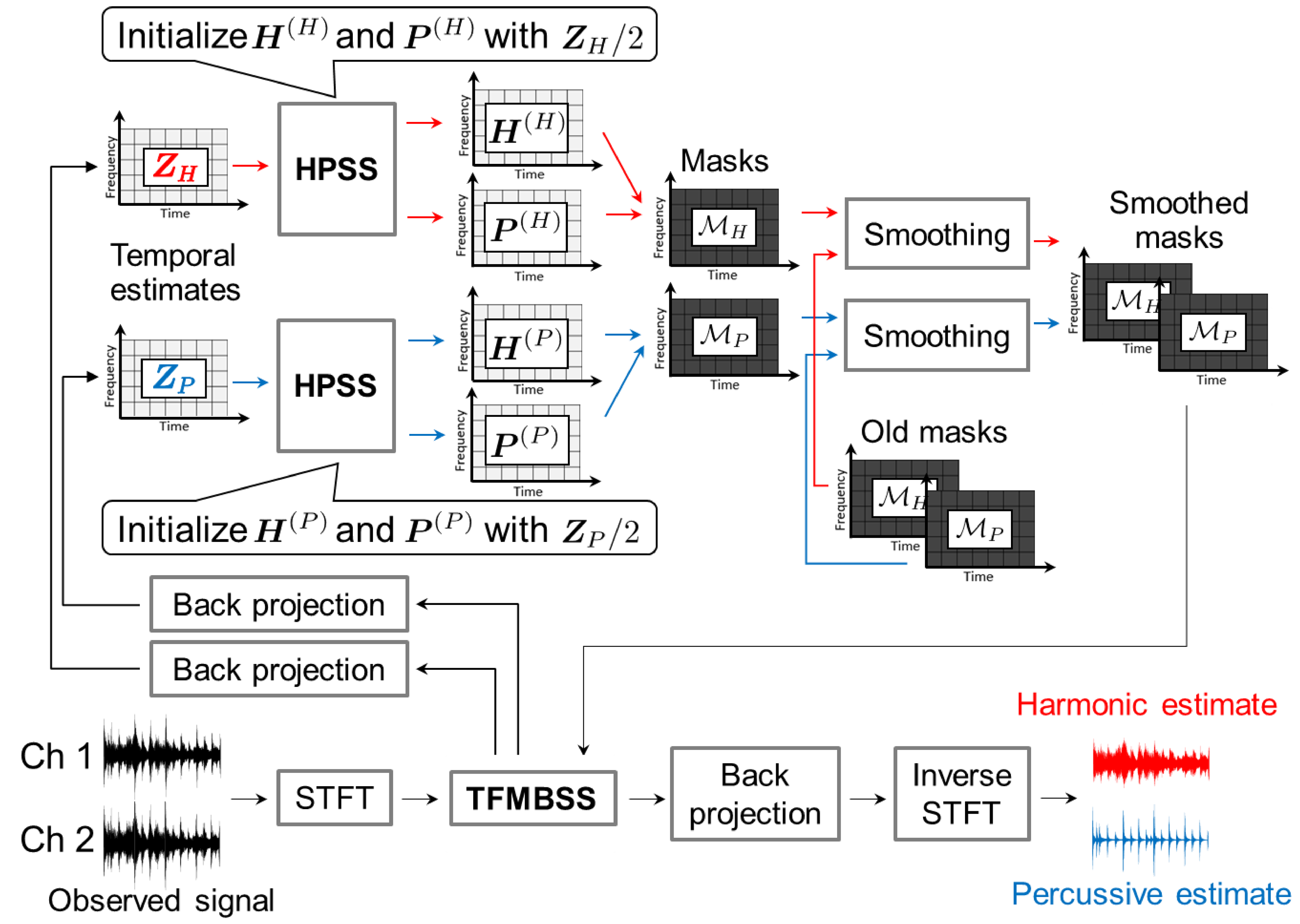
HPSS-based source model is effective, but its separation mechanism is **non-linear**, resulting in the **generation of artificial distortions**

For multichannel signals, **linear distortion-less separation** can be achieved by estimating the spatial demixing system $\boldsymbol{W}$

We propose high-quality multichannel blind HPSS

2. Proposed Method

- **Process flow (overview)**



TFMBSS iteratively optimizes the linear spatial demixing system $\boldsymbol{W}$

In each iteration of TFMBSS, **HPSS-based new mask calculation** and **mask smoothing** are performed

After TFMBSS is converged, the estimated signals are obtained via inverse STFT

- **HPSS-based new mask calculation**

Two HPSS are independently applied to each of temporarily estimated signals $\boldsymbol{Z}_{\mathrm{H}}$ and $\boldsymbol{Z}_{\mathrm{P}}$

Two Wiener-like masks $\mathcal{M}_H$ and $\mathcal{M}_P$ are constructed using the results of HPSS

$$\mathcal{M}_H = \frac{|\boldsymbol{H}^{(H)}|^2}{|\boldsymbol{H}^{(H)}|^2 + |\boldsymbol{P}^{(H)}|^2} \qquad \mathcal{M}_P = \frac{|\boldsymbol{P}^{(P)}|^2}{|\boldsymbol{H}^{(P)}|^2 + |\boldsymbol{P}^{(P)}|^2}$$

These masks enhance the harmonic or percussive components / by eliminating the other components

- **Mask smoothing using previous mask**

Optimization of TFMBSS is based on **primal-dual splitting algorithm**

Drastic change of masks in each iteration will cause instability of parameter optimization (see experiment 1)

Introduce **mask smoothing process** based on weighted geometric mean

Mask calculated in the current iteration — Entrywise product — Mask calculated in the previous iteration

$$\mathcal{M} \leftarrow \mathcal{M}^{\beta} \odot \mathcal{M}_{\mathrm{old}}{}^{\beta_{\mathrm{old}}} \quad (\beta + \beta_{\mathrm{old}} = 1,\ \beta, \beta_{\mathrm{old}} \geq 0)$$

Intensity of smoothing can be controlled by β and β_{old}

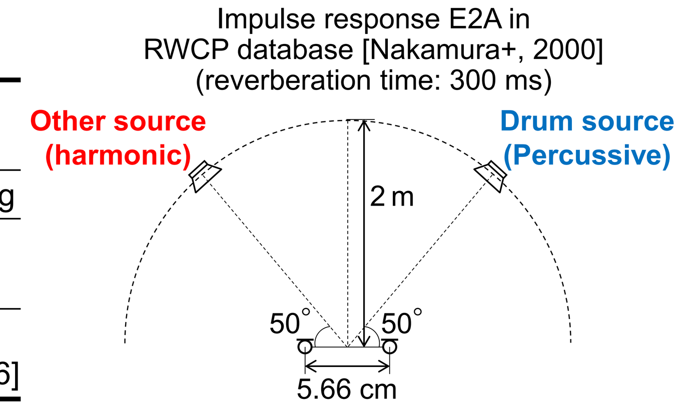
3. Experiments

- **Conducted experiments**

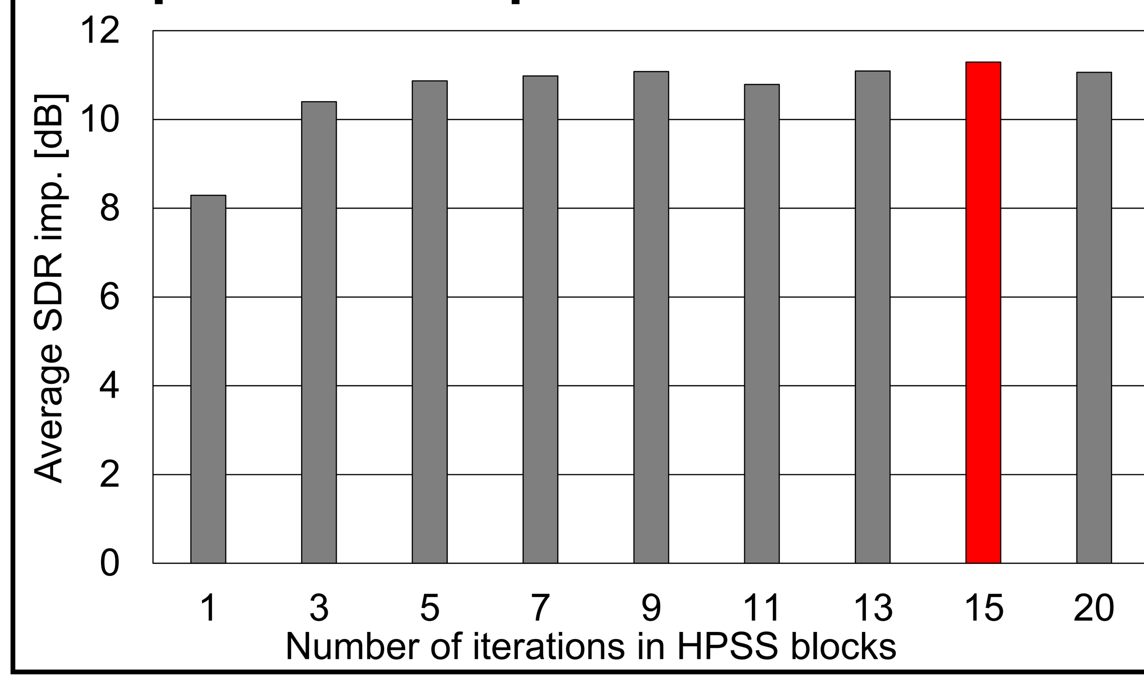
1. Investigation of the optimal number of iterations in HPSS process
2. Investigation of the optimal smoothing parameter in smoothing process
3. Performance comparison with state-of-the-art existing BSS algorithms

- **Conditions**

Music dataset (dry sources)	SiSEC2016 MUS [Liutkus+, 2016] "Drums" and "Other" sources of 20 songs
Windowing in STFT	128-ms-long Hann window with half-overlap shifting
Number of iterations in TFMBSS	500
Subjective evaluation score	Improvement of source-to-distortion ratio (SDR) [Vincent+, 2006]



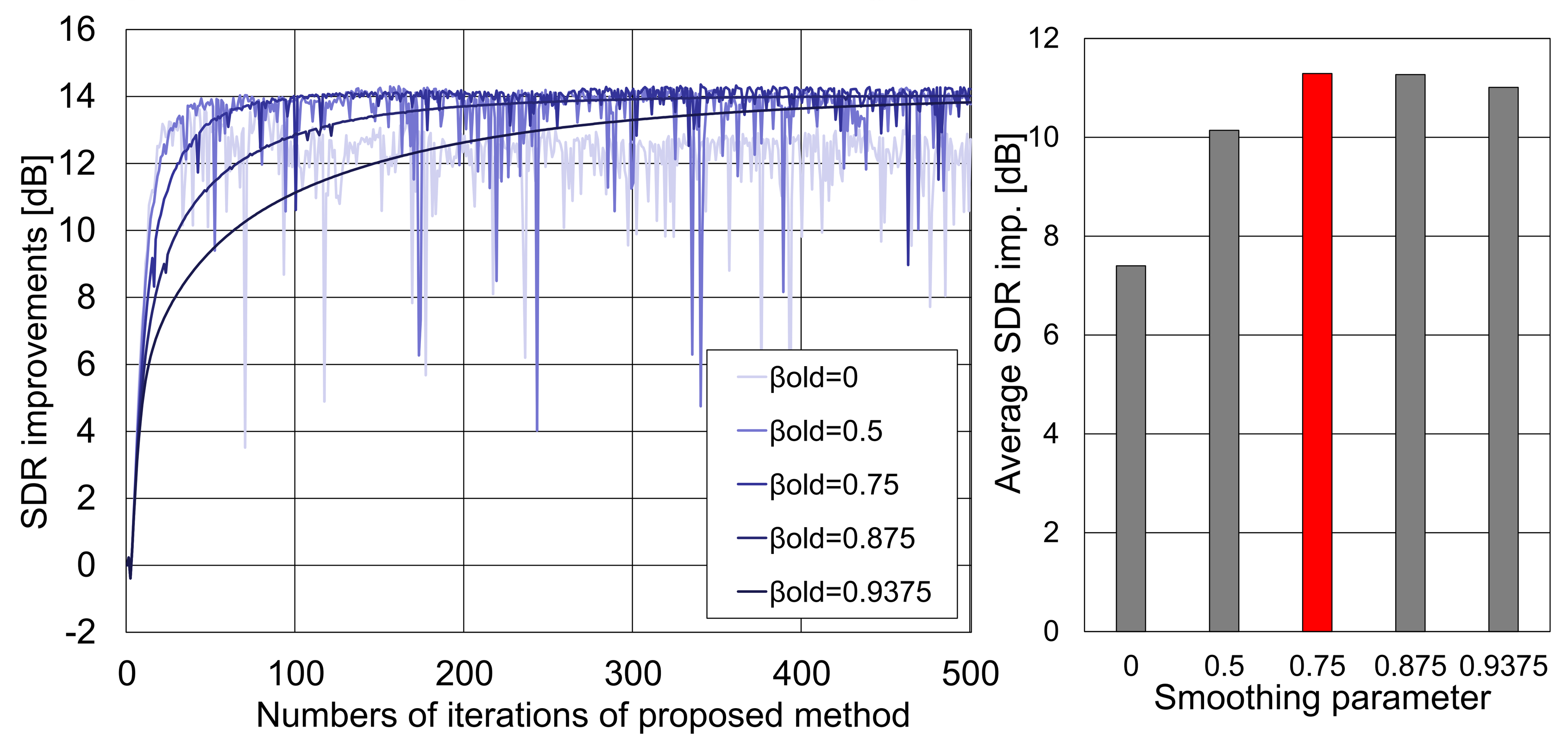
- **Experiment 1: optimal number of iterations in HPSS**



Fewer iterations are not preferable

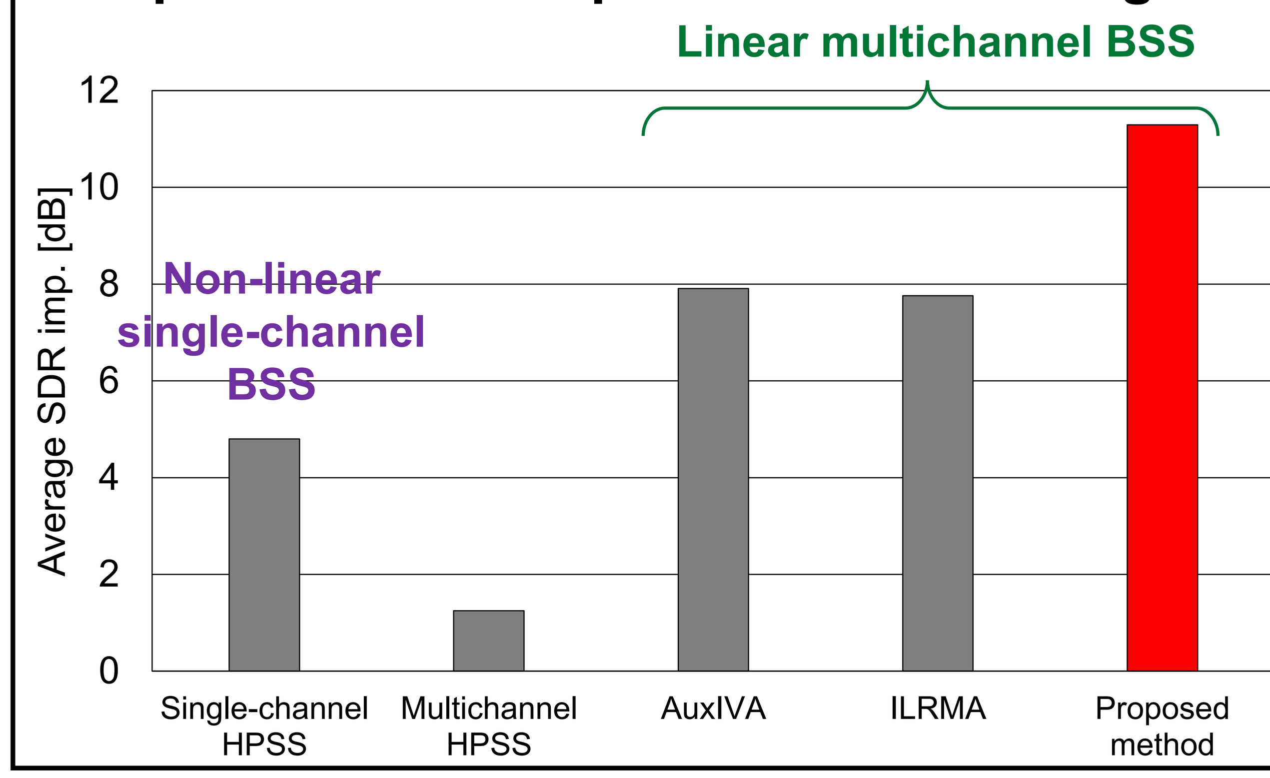
The performance is saturated for more than 9 iterations

- **Experiment 2: optimal smoothing parameter**



Smoothing process can drastically stabilize the SDR behavior

- **Experiment 3: comparison with existing methods**



Greatly outperforms other existing HPSS methods