香川高専

卒業研究論文

論文題目

スペクトログラム無矛盾性に基づく
独立低ランク行列分析

提出年月日	令和 2 年 2 月 26 日
学　　科	電気情報工学科
氏　　名	豊島 直　　印
指導教員（主査）	北村 大地 助教　　印
副　　査	柿元 健 准教授　　印
学　科　長	辻 正敏 教授　　印

香川高等専門学校

Independent Low-Rank Matrix Analysis Based On Spectrogram Consistency

Nao Toshima

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

Abstract

Blind source separation (BSS) is a technique for estimating individual audio sources from an observed mixture signal. BSS is utilized for a preprocess of automatic speech recognition, hearing-aid systems, and so on. For the (over-)determined case (the number of microphones is equal to or greater than the number of source signals in the mixture), independent low-rank matrix analysis (ILRMA) can achieve relatively high BSS performance. ILRMA estimates separated source signals that satisfy the following two conditions based on a statistical BSS algorithm called independent component analysis (ICA) and low lank approximation of matrix called nonnegative matrix factorization (NMF): (a) maximizing statistical independence among separated source signals by ICA and (b) low-rank approximation of the spectrogram (time-frequency matrix of an acoustic signal) of each separated source signal by NMF. However, conventional ILRMA does not consider a principal property called "spectrogram consistency": the consistent spectrogram directly corresponds to one acoustic signal in the time domain, whereas the inconsistent spectrogram does not. In the previous study, it was revealed that the spectrogram consistency can assist BSS algorithm and improves the separation accuracy of the source signals. In this thesis, I propose a new ILRMA algorithm that introduces the spectrogram consistency in addition to the above-mentioned two BSS principles (a) and (b). Also, I investigate how much the spectrogram consistency can improve the source separation accuracy in ILRMA-based BSS. The experimental results show that the source-to-distortion ratio is improved about 4 dB between conventional and proposed ILRMA in terms of a median value, which reveals that the spectrogram consistency greatly contributes to improve the performance of ILRMA.

Keywords: blind audio source separation, independent low-rank matrix analysis, spectrogram consistency

（和訳）

ブラインド音源分離（blind source separation: BSS）とは，複数の音源が混合した観測信号から，混合前の音源信号を推定する技術である．BSS は，音声認識の前段処理や補聴器などに活用される．混合している音源数以上のマイクロフォン数で観測される場合においては，独立低ランク行列分析（independent low-rank matrix analysis: ILRMA）と呼ばれる手法が比較的高性能である．ILRMA は，独立成分分析（independent component analysis: ICA）と呼ばれる統計的 BSS アルゴリズムと，非負値行列因子分解（nonnegative matrix factorization: NMF）と呼ばれる行列の低ランク近似に基づき，次の 2 項目を満たす分離信号を推定する：(a) ICA による分離信号間の統計的独立性の最大化，及び (b) NMF による各分離信号のスペクトログラム（音響信号の時間周波数行列）の低ランク近似．しかしながら，従来の ILRMA ではスペクトログラム無矛盾性と呼ばれる性質が考慮されていない．スペクトログラム無矛盾性とは，時間周波数表現された音響信号が時間領域の音響信号と直接対応している性質であり，これを満たさない時間周波数領域の信号をスペクトログラム矛盾と呼ぶ．先行研究より，BSS においてスペクトログラム無矛盾性を考慮することで分離信号の推定精度を向上させることができることが明らかになっている．そこで本論文では，スペクトログラム無矛盾性を考慮した新しい ILRMA を提案する．ILRMA が仮定する前述の 2 項目に加えて，スペクトログラム無矛盾性を新たに担保した場合に，どの程度音源分離精度が向上するかを調査する．実験結果より，従来手法と提案手法では信号対歪み比の中央値においておよそ 4 dB もの改善がみられ，スペクトログラム無矛盾性が ILRMA においても大きく性能向上に寄与する事実を明らかにした．

目次

第 1 章

緒言

1.1 本論文の背景

音源分離とは，Fig. 1.1 に示す例のように，複数の音源が混合した観測信号から，混合前の音源を推定する技術である．この技術は，音声認識の前段処理，雑音抑制などに応用されており，カーナビゲーションシステム，スマートフォン，補聴器などの多くのデバイスに組み込まれている．特に，音源位置やマイクロフォン位置等が未知の条件で音源分離を達成する技術はブラインド音源分離（blind source separation: BSS）と呼ばれる．観測信号のチャネル数（マイクロフォン数）と混合されている音源数が等しい条件下では，独立成分分析（independent component analysis: ICA）[1] に基づく BSS として，周波数領域 ICA（frequency-domain ICA: FDICA）[2]，独立ベクトル分析（independent vector analysis: IVA）[3, 4]，及び独立低ランク行列分析（independent low-rank matrix analysis: ILRMA）[5, 6, 7] 等が提案されてきた．IVA や ILRMA はいずれも，FDICA におけるパーミュテーション問題 [8]（周波数毎に得られる分離信号の順序を整列する問題）を解決するための FDICA の拡張である．特に ILRMA は，その分離性能の高さや各パラメータの初期値に対する頑健性等の利点から，数多くの改良手法や一般化手法へと発展している（例えば [9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22] 等）．

上記の BSS 含む多くの音響信号処理では，短時間フーリエ変換（short-time Fourier transform: STFT）を用いて時間領域の信号（時間波形）を時間周波数領域の信号（スペクトログラム）に変換する．このとき，STFT の過程で短時間信号には窓関数が乗じられる．これは，短時間信号のスペクトルに窓関数のスペクトルを畳み込むことに相当するため，スペクトログラム上では信号のパワーが周波数方向に滲む．また，STFT の過程では短時間信号をオーバーラップさせながらシフトして抽出するため，スペクトログラム上では信号のパワーが時間方向にも滲む．これらの滲みに起因して，時間周波数領域では，近傍時間周波数に一貫した共起性が生じる．結果的に，時間周波数領域では，近傍時間周波数間で一貫した共起性が生じていることが自然である．スペクトログラムに対して，BSS 等の信号処理を施した場合，通常はこの一貫した共起性が崩される．このようにスペクトログラムにおける近傍時間周波数グリッドの共起性が人工的に崩された状態の時間周波数領域の信号を「矛盾したスペクトログラム」と呼

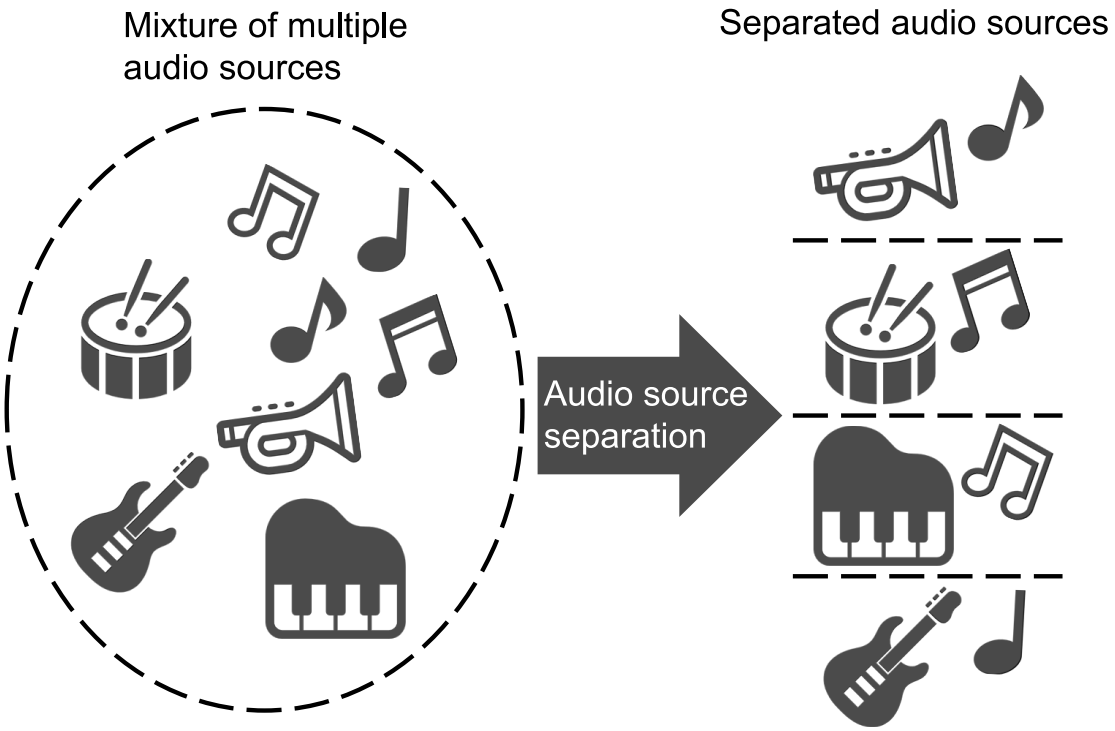


Fig. 1.1. Example of audio source separation, separation of musical instruments.

ぶ．逆に，スペクトログラム全体で一貫した共起性が保たれているスペクトログラムを「無矛盾なスペクトログラム」と呼ぶ [23, 24]．矛盾したスペクトログラムは，直接対応する時間領域の波形が存在しない．しかしながら，矛盾したスペクトログラムに逆 STFT を適用すると，時間周波数領域で無矛盾なスペクトログラムに射影されたうえで時間領域に変換される．すなわち，いかなるスペクトログラムも，逆 STFT を適用して一度時間領域に戻し，再び STFT を適用してスペクトログラムを得ることで，無矛盾なスペクトログラムに変換することができ，スペクトログラム無矛盾性を担保できる．近年，このスペクトログラム無矛盾性を BSS の音源分離の最適化途中で常に担保する FDICA が提案された [25]．FDICA に基づく BSS では，パーミュテーション問題の解決のために，音源由来の何らかの基準（音源モデル）を導入する必要があるが，文献 [25] ではスペクトログラム無矛盾性がパーミュテーション問題解決の新しい基準となることが示されている．その原理は，分離信号のパーミュテーションが近傍周波数で異なる場合に，スペクトログラム無矛盾性，すなわち近傍時間周波数の一貫した共起性が大きく損なわれるという性質に基づいている．

1.2　本論文の目的

本論文では，文献 [25] で示された結果に基づき，より高精度な BSS 達成を目的として，スペクトログラム無矛盾性を ILRMA に導入した音源分離手法及びその最適化アルゴリズムを新たに提案する．また，従来の ILRMA と本論文で提案する ILRMA との比較実験において，

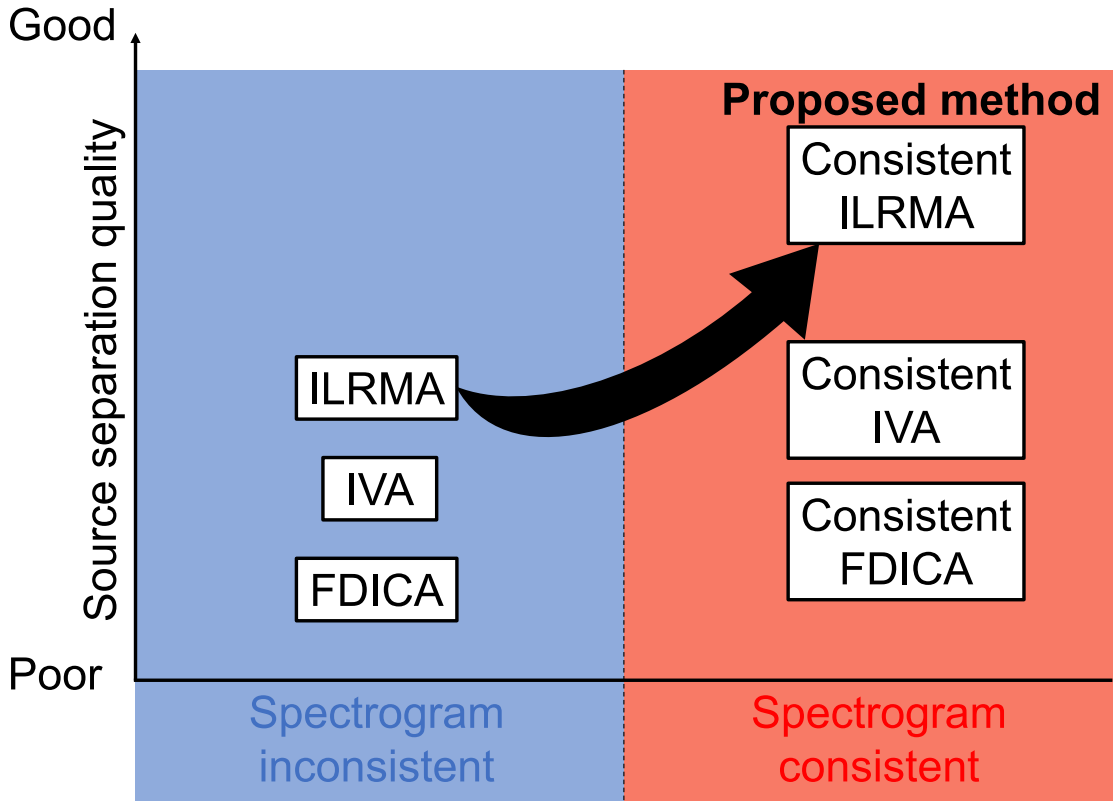


Fig. 1.2. Scope of this thesis.

音楽信号と音声信号の音源分離を行い，スペクトログラム無矛盾性の考慮の有無が分離性能に与える影響について調査する．以後，提案手法であるスペクトログラム無矛盾性を ILRMA に導入した音源分離手法を Consistent ILRMA と呼ぶ．Fig. 1.2 は本論文の立ち位置を表した図である．Consistent FDICA 及び Consistent IVA [25] はそれぞれ，FDICA 及び IVA においてスペクトログラム無矛盾性を考慮した際の音源分離手法を表す．また，横軸はスペクトログラム無矛盾性の有無を，縦軸は音源分離精度を表す．Consist FDICA や Consistent IVA において，従来の BSS より音源分離性能が向上したことと同様に，本論文では Consistent ILRMA においても同様の性能向上が得られることを期待している．

1.3 本論文の構成

2 章では，ILRMA とその関連手法及びスペクトログラム無矛盾性について述べる．3 章では，ILRMA へのスペクトログラム無矛盾性の適用について述べる．4 章では，Consistent ILRMA と従来手法との比較実験と結果について述べる．5 章では，本論文の結果について述べる．

第 2 章

従来手法

2.1 はじめに

本章では，2.2 節で STFT の原理について述べ，2.3 節で定式化を行う．また，2.4 節で FDICA の原理及びパーミュテーション問題について述べる．そして，2.5 節及び 2.6 節でそれぞれ IVA 及び ILRMA の原理について述べる．その後，2.7 節及び 2.8 節でそれぞれスペクトログラム無矛盾性について及び BSS へのスペクトログラム無矛盾性の適用について説明し，2.9 節で本章をまとめる．

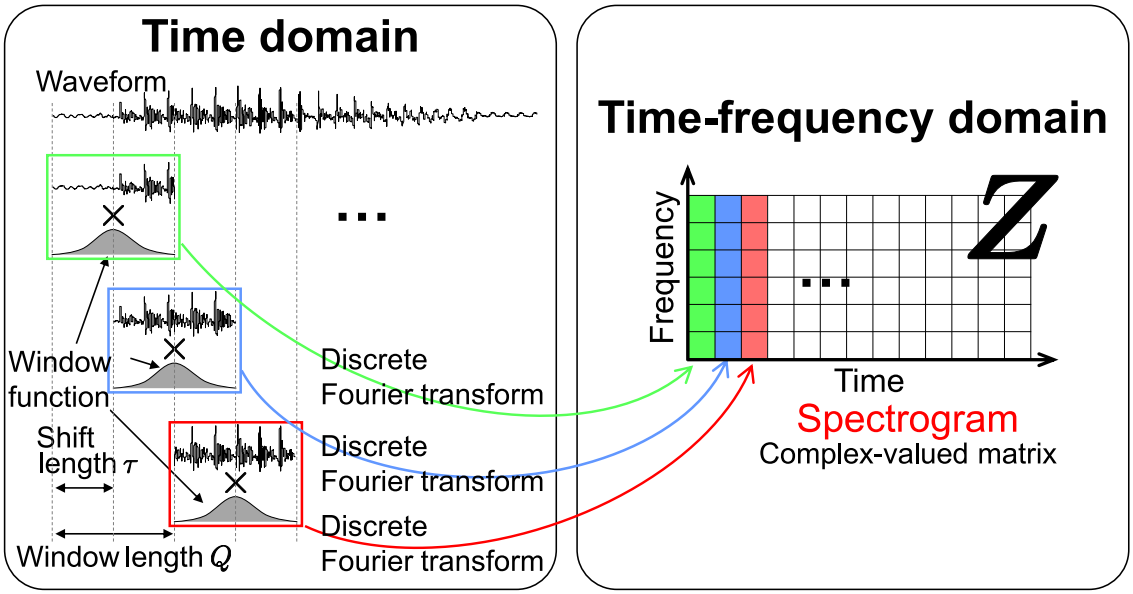


Fig. 2.1. Mechanism of STFT.

2.2 STFT

STFT は，Fig. 2.1 に示すように，時間領域の信号を時間周波数領域の信号に変換する手法である．これによって，時間的に変化するスペクトル成分を表現することができる．STFT の分析窓関数の長さ及びシフト長をそれぞれ Q 及び τ としたとき，時間領域の信号 $z[l]$ の j 番目の短時間区間（時間フレーム）の信号は次式で表される．

$$\boldsymbol{z}^{(j)} = \left[z\left[(j-1)\tau+1\right], z\left[(j-1)\tau+2\right], \cdots, z\left[(j-1)\tau+Q\right]\right]^{\mathrm{T}} \tag{2.1}$$

$$= \left[z^{(j)}[1], z^{(j)}[2], \cdots, z^{(j)}[q], \cdots, z^{(j)}[Q]\right]^{\mathrm{T}} \in \mathbb{R}^{Q} \tag{2.2}$$

ここで，$j = 1, 2, \cdots, J$ 及び $q = 1, 2, \cdots, Q$ は，それぞれ時間フレーム及び時間フレーム内のサンプルのインデックスを示す．また，$\cdot^{\mathrm{T}}$ はベクトルや行列の転置を表す．また，時間フレーム数 J は次式によって与えられる．

$$J = \frac{L}{\tau} \tag{2.3}$$

式 (2.1) で定義される全時間フレームの信号を全ての j についてまとめた全時間フレームの信号を $\boldsymbol{z} = \left[z^{(1)}, z^{(2)}, \cdots, z^{(j)}, \cdots, z^{(J)}\right] \in \mathbb{R}^{Q\times J}$ と表記すると，STFT の処理は次式のように表される．

$$\boldsymbol{Z} = \mathrm{STFT}_{\boldsymbol{\omega}}(\boldsymbol{z}) \in \mathbb{C}^{I\times J} \tag{2.4}$$

ここで，$\boldsymbol{\omega} = [\omega[1], \omega[2], \cdots, \omega[q], \cdots, \omega[Q]]^{\mathrm{T}} \in \mathbb{R}^{Q}$ は短時間区間の滑らかな周期性を担保するために乗じられる窓関数であり，式 (2.5) で示す Hann window，式 (2.6) で示す hamming window，及び式 (2.7) で示す Blackman window がよく用いられる．

$$\omega_{\mathrm{Hann}}[q] = 0.5 - 0.5\cos 2\pi\frac{q}{Q} \tag{2.5}$$

$$\omega_{\mathrm{Hamming}}[q] = 0.54 - 0.46\cos 2\pi\frac{q}{Q} \tag{2.6}$$

$$\omega_{\mathrm{Blackman}}[q] = 0.42 - 0.5\cos 2\pi\frac{q}{Q} + 0.08\cos 4\pi\frac{q}{Q} \tag{2.7}$$

また，スペクトログラム $\boldsymbol{Z}$ の (i, j) 番目の要素は次式で表される．

$$z_{ij} = \sum_{q=1}^{Q} \omega[q] z^{(j)}[q] \exp\left\{\frac{-\iota 2\pi(q-1)(i-1)}{F}\right\} \tag{2.8}$$

ここで，F は $\lfloor\frac{F}{2}\rfloor + 1 = I$ を満たす整数（$\lfloor\cdot\rfloor$ は床関数）を，$i = 1, 2, \cdots, I$ は周波数ビンのインデクスを，ι は虚数単位を示している．このように，時間領域の信号は一定幅の短時間ごとに分析窓関数を乗じて離散フーリエ変換を行うことで，横軸が時間，縦軸が周波数のスペクトログラムと呼ばれる複素行列 $\boldsymbol{Z}$ で表すことができる．Figs. 2.2 及び 2.3 にそれぞれ音声信号と音楽信号のパワースペクトログラムを例として示している．ここで，色の違いは青色ほどパワーが小さく，黄色ほどパワーが大きいことを示している．パワースペクトログラムとは，式 (2.4) で求めた複素スペクトログラム $\boldsymbol{Z}$ の要素ごとの絶対値 2 乗を取ったものである．

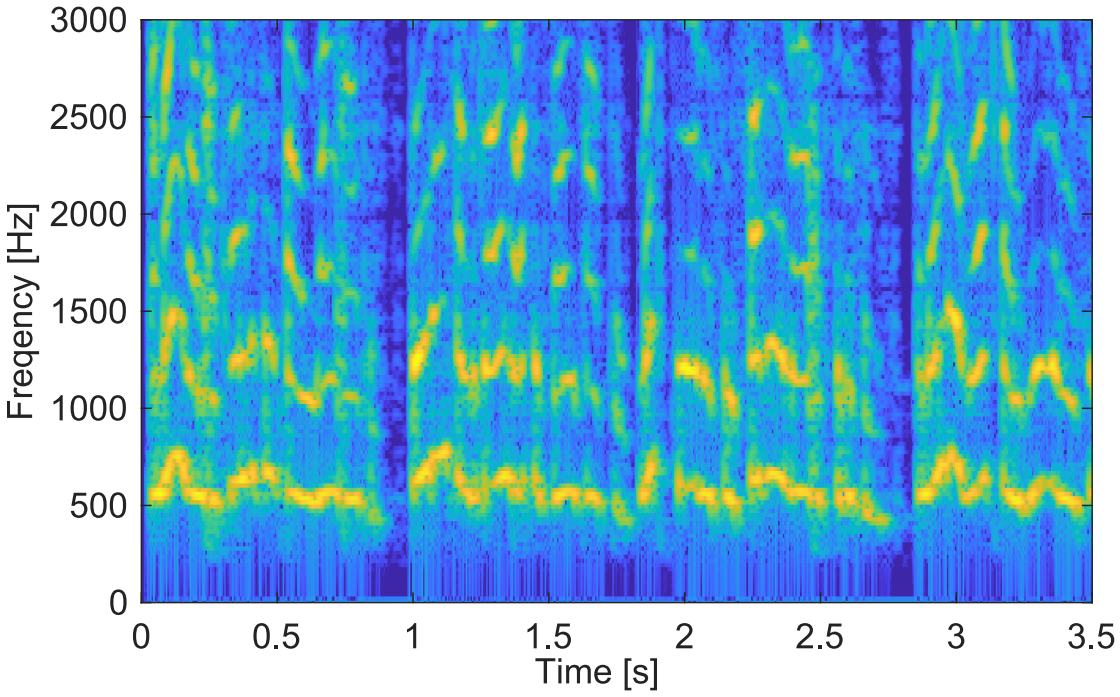


Fig. 2.2. Power spectrogram of speech signal.

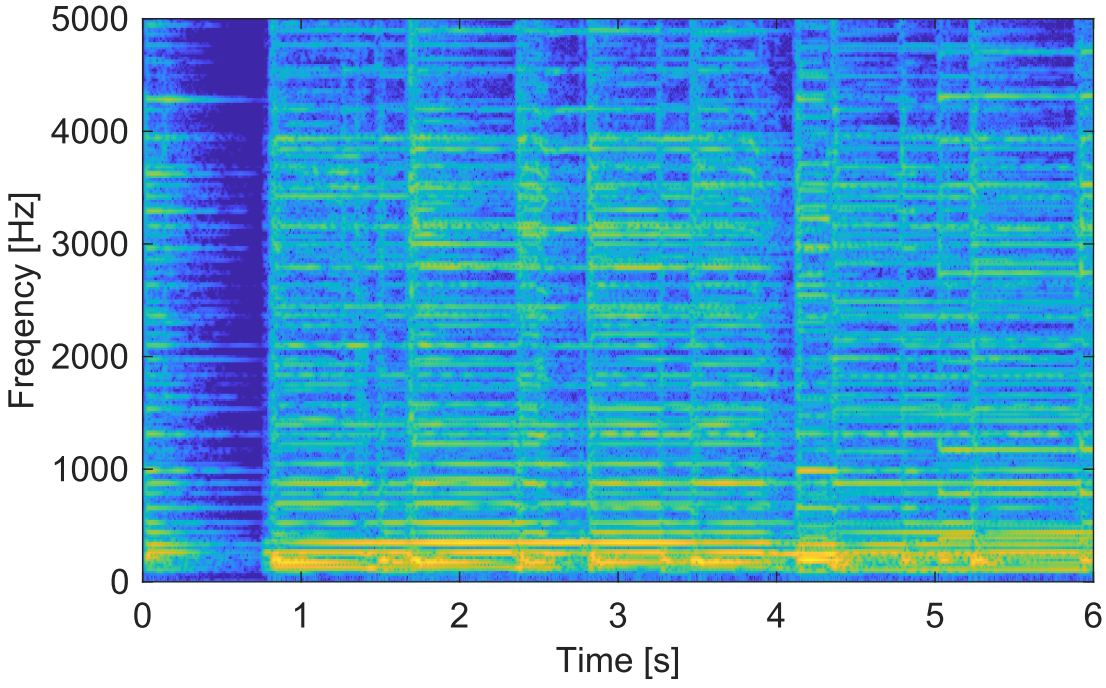


Fig. 2.3. Power spectrogram of music signal.

2.3 周波数領域の BSS における定式化

離散時間信号の l 番目のサンプルを $x[l]$ のように表記し，N 個の音源信号が M 個のマイクロフォンで観測される状況を考える．多チャネルの音源信号，観測信号，及び分離信号をそれぞれ次式で表す．

$$\boldsymbol{s}[l] = \left[\, s_1[l], s_2[l], \cdots, s_n[l], \cdots s_N[l] \,\right]^{\mathrm{T}} \qquad \in \mathbb{R}^N \tag{2.9}$$

$$\boldsymbol{x}[l] = \left[\, x_1[l], x_2[l], \cdots, x_m[l], \cdots x_M[l] \,\right]^{\mathrm{T}} \qquad \in \mathbb{R}^M \tag{2.10}$$

$$\boldsymbol{y}[l] = \left[\, y_1[l], y_2[l], \cdots, y_n[l], \cdots y_N[l] \,\right]^{\mathrm{T}} \qquad \in \mathbb{R}^N \tag{2.11}$$

ここで，$n = 1, 2, \cdots, N$，$m = 1, 2, \cdots, M$，及び $l = 1, 2, \cdots, L$ はそれぞれ音源，マイクロフォン（チャネル），及び離散時間のインデクスである．BSS では，音源信号 $\boldsymbol{s}$ に近い分離信号 $\boldsymbol{y}$ を，観測信号 $\boldsymbol{x}$ から推定することが目的となる．

FDICA，IVA，及び ILRMA 等の周波数領域 BSS では，信号を時間周波数領域で取り扱う．合成時の窓関数を $\tilde{\boldsymbol{\omega}}$ とおくとき，逆 STFT を $\mathrm{ISTFT}_{\tilde{\boldsymbol{\omega}}}(\cdot)$ と表記する．本論文では，$\boldsymbol{\omega}$ と $\tilde{\boldsymbol{\omega}}$ のペアが次式の完全再構成条件を満たすことを仮定する．

$$\boldsymbol{z} = \mathrm{ISTFT}_{\tilde{\boldsymbol{\omega}}}(\mathrm{STFT}_{\boldsymbol{\omega}}(\boldsymbol{z})) \qquad \forall \boldsymbol{z} \in \mathbb{R}^{Q \times J} \tag{2.12}$$

各チャネルに STFT を適用して得られる音源信号，観測信号，及び分離信号のスペクトログラムの (i, j) 番目の要素をそれぞれ次式で表す．

$$\boldsymbol{s}_{ij} = \left[\, s_{ij1}, s_{ij2}, \cdots, s_{ijn}, \cdots s_{ijN} \,\right]^{\mathrm{T}} \qquad \in \mathbb{C}^N \tag{2.13}$$

$$\boldsymbol{x}_{ij} = \left[\, x_{ij1}, x_{ij2}, \cdots, x_{ijm}, \cdots x_{ijM} \,\right]^{\mathrm{T}} \qquad \in \mathbb{C}^M \tag{2.14}$$

$$\boldsymbol{y}_{ij} = \left[\, y_{ij1}, y_{ij2}, \cdots, y_{ijn}, \cdots y_{ijN} \,\right]^{\mathrm{T}} \qquad \in \mathbb{C}^N \tag{2.15}$$

また，式 (2.13)–(2.15) の時間周波数行列をそれぞれ $\boldsymbol{S}_n \in \mathbb{C}^{I \times J}$，$\boldsymbol{X}_m \in \mathbb{C}^{I \times J}$，及び $\boldsymbol{Y}_n \in \mathbb{C}^{I \times J}$ と定義する．周波数領域 BSS では，次式の瞬時混合を仮定する．

$$\boldsymbol{x}_{ij} = \boldsymbol{A}_i \boldsymbol{s}_{ij} \tag{2.16}$$

ここで，$\boldsymbol{A}_i \in \mathbb{C}^{M \times N}$ は周波数毎の混合行列である．式 (2.16) は，混合系の残響時間が窓長よりも十分短い場合に近似的に成立する．

以後，本論文では決定的な系（$M = N$）のみを考える．この場合，BSS は $\boldsymbol{A}_i$ の逆行列を推定する問題となる．この逆行列を $\boldsymbol{W}_i \approx \boldsymbol{A}_i^{-1}$ とすると，分離信号は次式となる．

$$\boldsymbol{y}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij} \tag{2.17}$$

ここで，$\boldsymbol{W}_i = [\boldsymbol{w}_{i1}, \boldsymbol{w}_{i2}, \cdots, \boldsymbol{w}_{in}, \cdots, \boldsymbol{w}_{iN}]^{\mathrm{H}} \in \mathbb{C}^{N \times M}$ は分離行列と呼ばれ，$\cdot^{\mathrm{H}}$ はベクトルや行列のエルミート転置を表す．

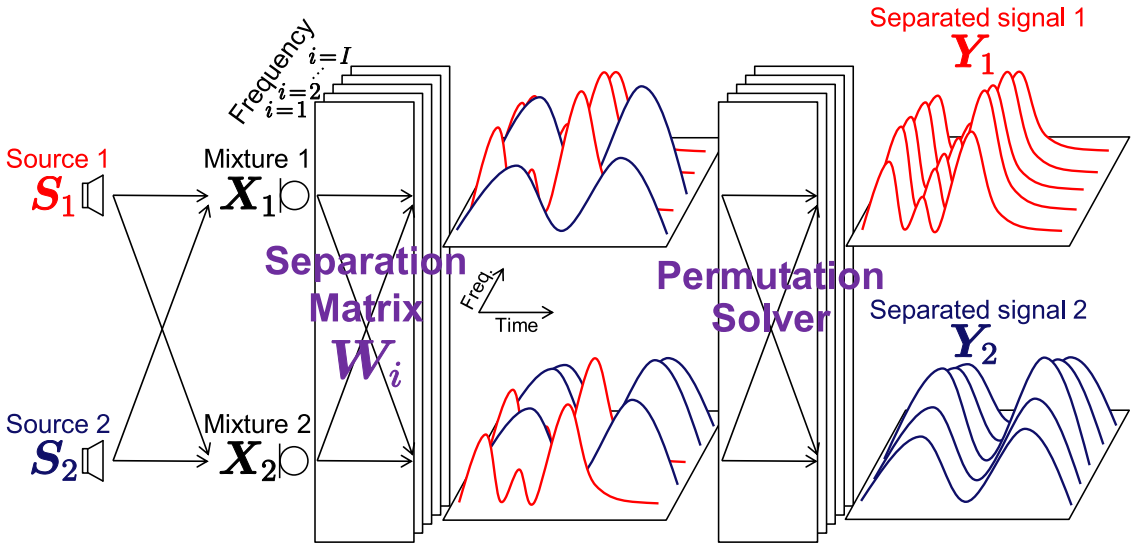


Fig. 2.4. Permutation problem in FDICA.

ICA に基づく BSS では，分離音源のスケールや順序が不定であるため，次式の任意性が存在する．

$$\hat{\boldsymbol{y}}_{ij} = \hat{\boldsymbol{W}}_i \boldsymbol{x}_{ij} \quad (\hat{\boldsymbol{W}}_i = \boldsymbol{D}_i \boldsymbol{P}_i \boldsymbol{W}_i) \tag{2.18}$$

ここで，$\boldsymbol{D}_i \in \mathbb{C}^{N\times N}$ 及び $\boldsymbol{P}_i \in \{0,1\}^{N\times N}$ はそれぞれ任意の対角行列及びパーミュテーション行列（置換行列）である．

2.4　FDICA とパーミュテーション問題

FDICA は残響を含むことによって生じた音源の畳み込み混合を分離するため，観測信号を STFT し，得られたスペクトログラムの周波数ビン毎の複素時系列観測信号 $\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \cdots, \boldsymbol{x}_{iJ}$ に対し，個別に ICA を適用することにより分離信号のスペクトログラムを推定する．しかし，式 (2.18) で示したように，ICA の分離信号にはスケールや順序の不定性があり，これらは他の周波数とは無関係に起こるため，観測信号に FDICA を適用しただけでは分離信号を正しく推定することができない．周波数毎のスケールの任意性についてはプロジェクションバック法 [26] によって解析的に復元可能であるが，Fig. 2.4 のような周波数毎の分離信号の順序の整列はパーミュテーション問題と呼ばれる．これはすなわち，$\boldsymbol{P}_1 = \boldsymbol{P}_2 = \cdots = \boldsymbol{P}_I$ となるように分離信号を整列することであり，大きな課題である．

2.5　IVA

IVA は 2.4 節で述べたパーミュテーション問題を回避しながら音源分離を行う手法である [3, 4]．IVA では，全周波数成分をまとめたベクトル $\bar{\boldsymbol{y}}_{jn} = [y_{1jn}, y_{2jn}, \cdots, y_{ijn}, \cdots, y_{Ijn}]$ を

考え，このベクトルを確率変数としたときの確率分布が球対称分布になるという仮定を置き，周波数毎に分離行列を推定する．この仮定を置くことは，分離信号のある周波数が大きな値をとるとそれに連動して他の周波数も大きな値をとる，という音源モデルに対応し，これによって，パーミュテーション問題を可能な限り回避しながら，分離行列 $\boldsymbol{W}_i$ を推定することができる．音のスペクトルは，基本周波数とその整数倍の周波数で連動して大きな値をとるため，前述の仮定により，連動する周波数成分を同一の音源のスペクトログラムとみなすことができ，結果的に IVA ではパーミュテーション問題をある程度回避することができる．

2.6 ILRMA

IVA では，分離行列 $\boldsymbol{W}_i$ を推定する際に，「同一音源の異なる周波数の成分は連動して大きな値をとる」という音源モデルを仮定することで，パーミュテーション問題をある程度回避することができる．この音源モデルを一般化させ，より高精度にパーミュテーション問題を回避しながら BSS を行う手法として，ILRMA が提案されている [5, 6, 7]．ILRMA では，「同一音源の時間周波数構造は低ランク行列で近似できる」というモデルを仮定しており，この低ランク近似には非負値行列因子分解（nonnegative matrix factorization: NMF）[27, 28] が用いられている．

ILRMA は，次式の複素ガウス分布を音源信号の生成モデルとして仮定する．

$$\begin{aligned} p(\bar{\boldsymbol{y}}_{j1}, \bar{\boldsymbol{y}}_{j2}, \cdots, \bar{\boldsymbol{y}}_{jn}, \cdots, \bar{\boldsymbol{y}}_{jN}) &= \prod_n p(\bar{\boldsymbol{y}}_{jn}) \\ &= \prod_{n,i} \frac{1}{\pi r_{ijn}} \exp\left(-\frac{|y_{ijn}|^2}{r_{ijn}}\right) \end{aligned} \tag{2.19}$$

ここで，r_{ijn} は，各音源の時間周波数毎の分散であり，$r_{ijn} = \mathrm{E}\left[|y_{ijn}|^2\right]$ である．式 (2.19) に基づく観測信号の負対数尤度関数は

$$\mathcal{L} = -2J \sum_i \log|\det \boldsymbol{W}_i| + \sum_{i,j,n} \left(\frac{1}{\pi r_{ijn}} + \log r_{ijn}\right) \tag{2.20}$$

で与えられる．さらに，分散 r_{ijn} は次式のように NMF で低ランク分解される．

$$r_{ijn} = \sum_k t_{ikn} v_{kjn} \tag{2.21}$$

$$\boldsymbol{R}_n = \boldsymbol{T}_n \boldsymbol{V}_n \tag{2.22}$$

ここで，$k = 1, 2, \cdots, K$ は NMF における基底のインデックスである．したがって，各音源の時間周波数構造の基底数は共通の K 本となる（ランク K の非負行列で近似される）．また，$\boldsymbol{T}_n \in \mathbb{R}_{\geq 0}^{I \times K}$ 及び $\boldsymbol{V}_n \in \mathbb{R}_{\geq 0}^{K \times J}$ は NMF における基底行列及びアクティベーション行列であり，t_{ikn} 及び v_{kjn} はそれぞれ $\boldsymbol{T}_n$ 及び $\boldsymbol{V}_n$ の要素である．さらに，$\boldsymbol{R}_n \in \mathbb{R}_{\geq 0}^{I \times J}$ は時間周波数分散行列であり，r_{ijn} は $\boldsymbol{R}_n$ の要素である．

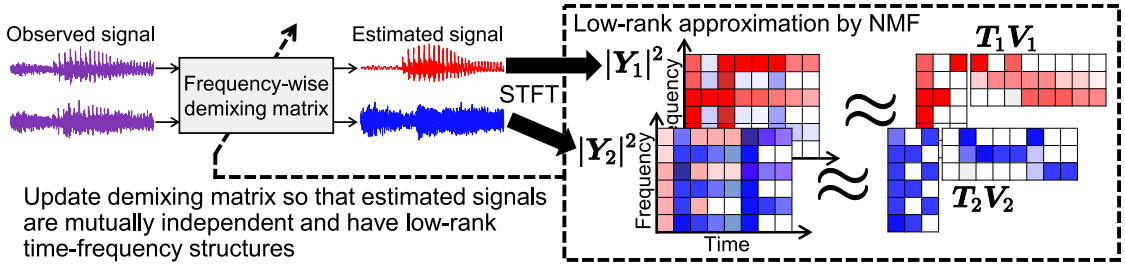


Fig. 2.5. Principle of BSS based on IRLMA.

ILRMA における音源分離の原理を Fig. 2.5 に示す．分離行列 $\boldsymbol{W}_i$ 及び NMF 音源モデル $\boldsymbol{T}_n$ 及び $\boldsymbol{V}_n$ の最適化の過程では，分離信号のパワースペクトログラム $|\boldsymbol{Y}_n|^2$ を低ランク行列 $\boldsymbol{T}_n\boldsymbol{V}_n$ としてモデル化しながら，その時間周波数構造を共変関係として加味した分離行列 $\boldsymbol{W}_i$ を推定する．混合前の各音源のパワースペクトログラム $|\boldsymbol{S}_n|^2$ が低ランクであれば，混合信号のパワースペクトログラム $|\boldsymbol{X}_m|^2$ のランクは基本的に増加することから，ILRMA は分離信号の時間周波数構造を低ランク行列に誘導することで，パーミュテーション問題を IVA よりも高精度に回避しつつ，互いに独立となる分離信号を推定することができる．ここで，行列に対する $|\cdot|^2$ は要素毎の絶対値二乗を表す．

ILRMA における分離行列 $\boldsymbol{W}_i$ の最適化では，補助関数法に基づく反復最適化手法を用いることができる．この最適化手法は反復射影法（iterative projection: IP）[4, 29] と呼ばれ，従来の自然勾配法 [3, 30] よりも高速かつ安定に収束することが実験的に示されている．ILRMA における分離ベクトル $\boldsymbol{w}_{in}$ の更新式は IP を用いて次式で与えられる．

$$\boldsymbol{U}_{in} \leftarrow \frac{1}{J}\sum_j \frac{1}{\sum_k t_{ikn}v_{kjn}}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^{\mathrm{H}} \tag{2.23}$$

$$\boldsymbol{w}_{in} \leftarrow (\boldsymbol{W}_i\boldsymbol{U}_{in})^{-1}\boldsymbol{e}_n \tag{2.24}$$

$$\boldsymbol{w}_{in} \leftarrow \boldsymbol{w}_{in}\left(\boldsymbol{w}_{in}^{\mathrm{H}}\boldsymbol{U}_{in}\boldsymbol{w}_{in}\right)^{-\frac{1}{2}} \tag{2.25}$$

ここで，$\boldsymbol{e}_n \in \mathbb{R}^N_{\{0,1\}}$ は n 番目の要素が 1，他要素が 0 のベクトルである．分離ベクトル更新後は，分離信号を次式で更新する．

$$y_{ijn} \leftarrow \boldsymbol{w}_{in}^{\mathrm{H}}\boldsymbol{x}_{ij} \tag{2.26}$$

音源モデル $\boldsymbol{T}_n$ 及び $\boldsymbol{V}_n$ は次の乗算型反復更新式で最適化できる．

$$t_{ikn} \leftarrow t_{ikn}\sqrt{\frac{\sum_j |y_{ijn}|^2\left(\sum_{k'} t_{ik'n}v_{k'jn}\right)^{-2}v_{kjn}}{\sum_j\left(\sum_{k'} t_{ik'n}v_{k'jn}\right)^{-1}v_{kjn}}} \tag{2.27}$$

$$v_{kjn} \leftarrow v_{kjn}\sqrt{\frac{\sum_i |y_{ijn}|^2\left(\sum_{k'} t_{ik'n}v_{k'jn}\right)^{-2}t_{ikn}}{\sum_i\left(\sum_{k'} t_{ik'n}v_{k'jn}\right)^{-1}t_{ikn}}} \tag{2.28}$$

$\boldsymbol{T}_n$ 及び $\boldsymbol{V}_n$ の更新後は推定分散 $\boldsymbol{R}_n$ を次式で更新する．

$$\boldsymbol{R}_n \leftarrow \boldsymbol{T}_n\boldsymbol{V}_n \tag{2.29}$$

以上より，これらの分離行列と音源モデルの更新式を交互に反復することで，式 (2.20) を最小化できる．全変数の推定後は，次式のプロジェクションバック法 [26] を適用する．

$$\tilde{\boldsymbol{y}}_{ijn} = \boldsymbol{W}_i^{-1} (\boldsymbol{e}_n \circ \boldsymbol{y}_{ij}) = y_{ijn} \boldsymbol{\lambda}_{in}, \tag{2.30}$$

ここで，$\tilde{\boldsymbol{y}}_{ijn} = [\, \tilde{y}_{ijn1}, \tilde{y}_{ijn2}, \cdots, \tilde{y}_{ijnm}, \cdots, \tilde{y}_{ijnM} \,]^{\mathrm{T}} \in \mathbb{C}^M$ はスケール補正後の分離信号の (i,j) 番目の成分，$\boldsymbol{\lambda}_{in} = [\, \lambda_{in1}, \lambda_{in2}, \cdots, \lambda_{inm}, \cdots, \lambda_{inM} \,]^{\mathrm{T}} \in \mathbb{C}^M$ はスケール補正係数，及び $\circ$ は要素毎の積を表す．

2.7 スペクトログラム無矛盾性

Fig. 2.6 の左上図のように，スペクトログラムのある時間周波数グリッドが大きなパワー値を持つ状況を考える．STFT を適用した際に乗じた窓関数は，時間周波数領域では窓関数のスペクトルを周波数方向に畳み込むことに相当する．したがって，ある時間周波数グリッドの大きなパワー値は周波数に滲み，周波数方向に共起性が生まれる．また，STFT を適用した際の時間フレーム間のオーバーラップは，時間周波数領域では時間方向の冗長性（隣接する時間フレームが互いに共通の情報を一部含んでいる性質）を生じさせる．これによって，ある時間周波数グリッドの大きなパワー値は時間方向にも滲み，時間方向にも共起性が生まれる．結果的に，時間周波数領域では，近傍時間周波数間で一貫した共起性が生じていることが自然であり，この共起性をスペクトログラム無矛盾性と呼ぶ．この共起性が，時間周波数領域での何らかの信号処理によって崩された状態のことをスペクトログラム矛盾と呼ぶ．従って，Fig. 2.6 の左上図は共起性に一貫性のない矛盾したスペクトログラムであり，この信号に対応する無矛盾なスペクトログラムは Fig. 2.6 の右上図である．時間と周波数の両方向に信号のパワーの滲みが生じていることが確認できる．

スペクトログラム無矛盾性は，任意のスペクトログラムに対して，Fig. 2.7 のように，逆 STFT 及び STFT を続けて適用することで，担保することができる．今，Fig. 2.7 中の $\boldsymbol{s}$ が時間領域のある信号である．これを STFT すると，時間周波数領域の $\boldsymbol{S}$ の点に写像される．このスペクトログラム $\boldsymbol{S}$ に対し，何らかの信号処理を加えた場合，$\boldsymbol{S}'$ というスペクトログラムが得られるが，この $\boldsymbol{S}'$ は一貫した共起性が崩され，矛盾したスペクトログラムとなっている．この $\boldsymbol{S}'$ に直接対応する時間波形は時間領域に存在しない．この矛盾したスペクトログラム $\boldsymbol{S}'$ に対して，逆 STFT を適用すると，時間周波数領域において無矛盾なスペクトログラム $\boldsymbol{S}''$ に射影されたうえで時間領域に変換され，$\boldsymbol{s}''$ が得られる．したがって，矛盾・無矛盾問わず時間周波数領域のいかなるスペクトログラムも，一度逆 STFT を適用して時間領域に戻し，再び STFT を適用して時間周波数領域の信号に変換することで，無矛盾なスペクトログラムに復元することができる．

前述の通り，逆 STFT は矛盾したスペクトログラムを無矛盾なスペクトログラムに復元する．即ち，スペクトログラム $\boldsymbol{Z}$ の無矛盾性は

$$\mathcal{E}(\boldsymbol{Z}) = \boldsymbol{Z} - \mathrm{STFT}_{\boldsymbol{\omega}}(\mathrm{ISTFT}_{\tilde{\boldsymbol{\omega}}}(\boldsymbol{Z})) \tag{2.31}$$

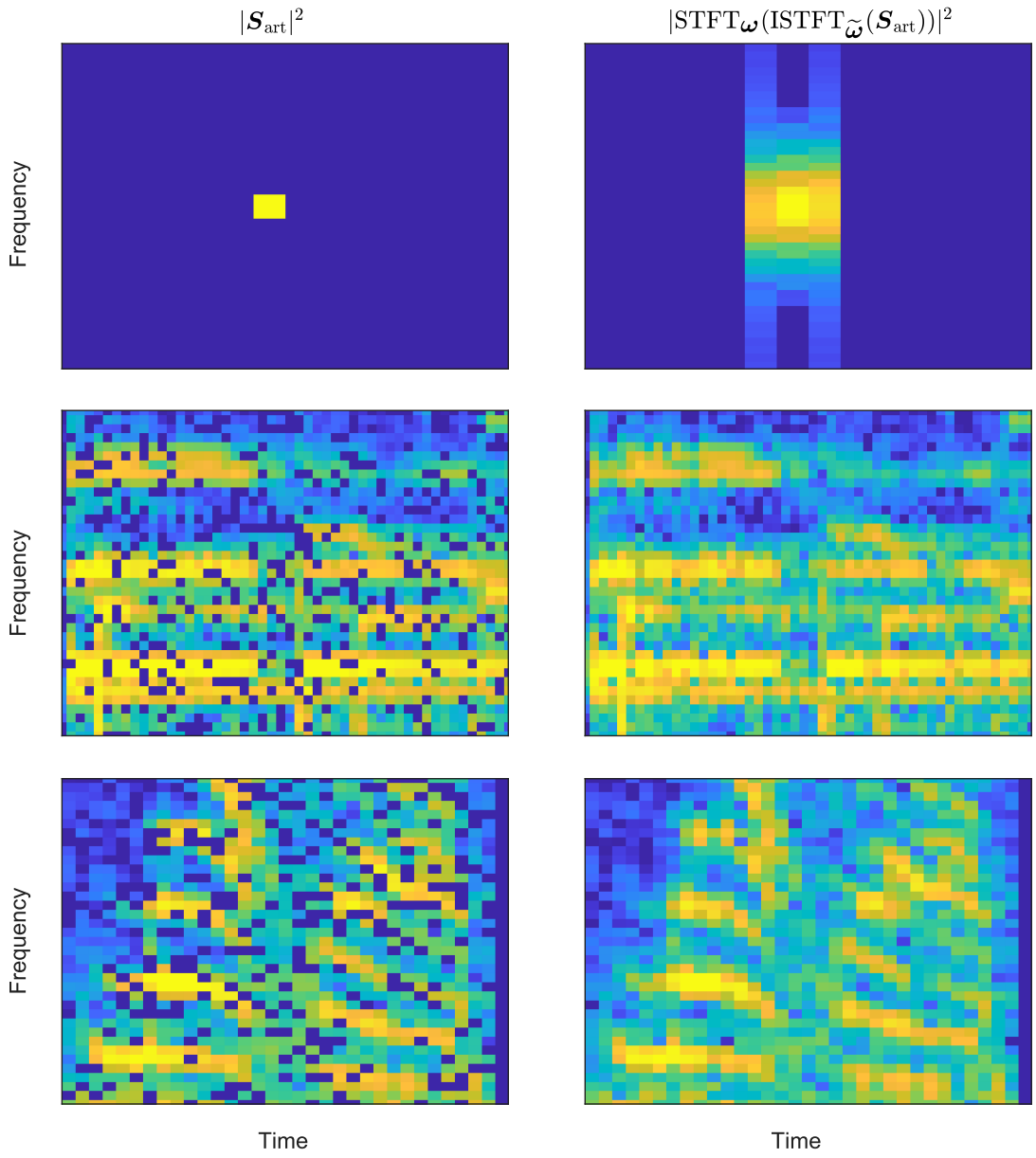


Fig. 2.6. Inconsistent power spectrograms $|\boldsymbol{S}_{\mathrm{art}}|^2$ (left column) and their consistent version (right column) obtained by applying inverse STFT and STFT. The top-left spectrogram is artificially produced with random phase. The middle-left and the bottom-left spectrograms are music and speech signals with random dropout. Enforcing spectrogram consistency can be viewed as a smoothing process of the inconsistent spectrogram along both time and frequency axes.

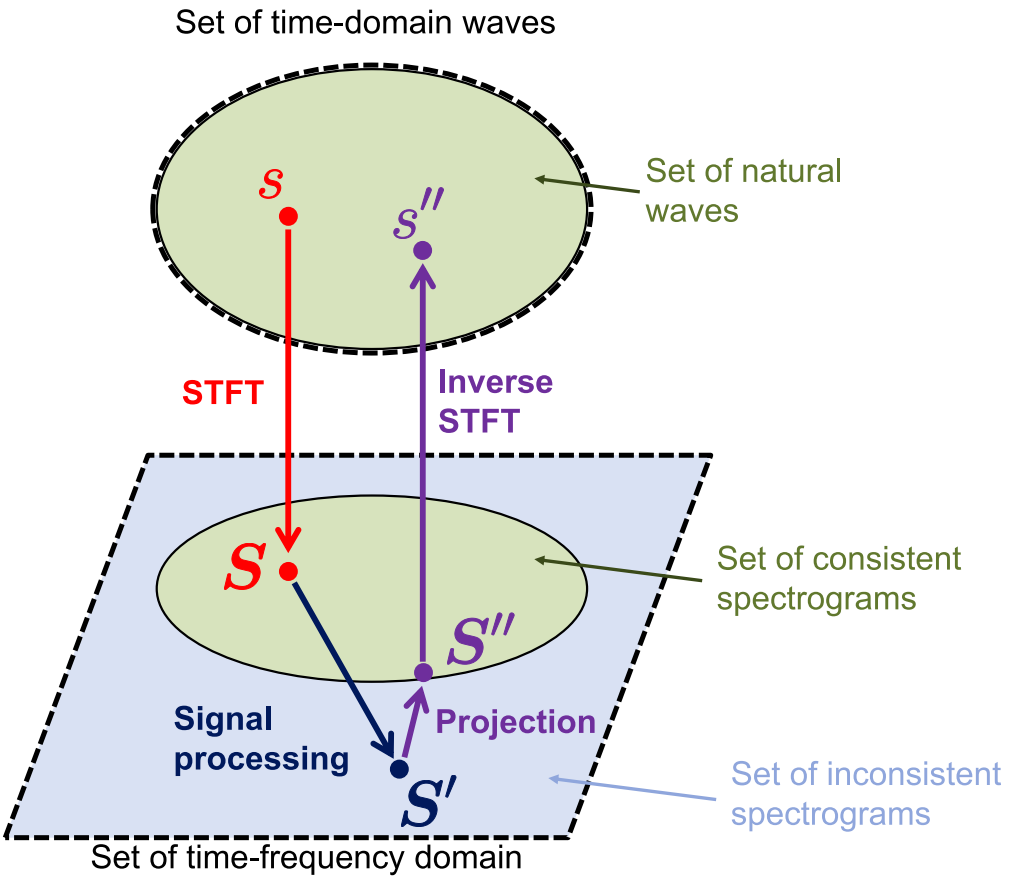


Fig. 2.7. Spectrogram (in)consistency with STFT and inverse STFT.

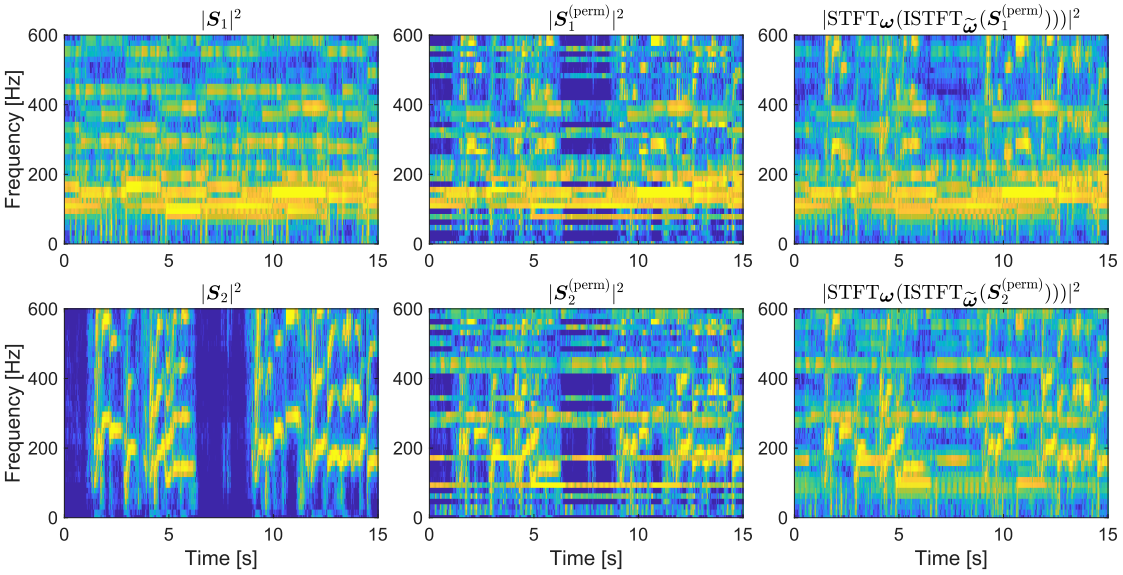
のノルム $\|\mathcal{E}(\boldsymbol{Z})\|$ によって特徴付けられ，それが 0 となるスペクトログラム $\boldsymbol{Z}$ を無矛盾と呼ぶ．大雑把に言えば Fig. 2.6 に示すように，逆 STFT はスペクトログラムの時間及び周波数方向への滲みがない成分に対して，スムージングをかけるような処理となる．$\boldsymbol{S}_{\mathrm{art}}$ は，人工的に作られたスペクトログラム（すなわち，一貫した共起性が考慮されていない矛盾したスペクトログラム）を表す．Fig. 2.6 の左上図の $\boldsymbol{S}_{\mathrm{art}}$ は，中心部分の時間周波数グリッドのみが大きなパワー値を持つスペクトログラムである．Fig. 2.6 の左中図及び左下図の $\boldsymbol{S}_{\mathrm{art}}$ はそれぞれ，音楽信号と音声信号の無矛盾なスペクトログラムに対してランダムに選んだ時間周波数グリッドのパワーを 0 にしたスペクトログラムである．これらも，人工的な処理を加えているため矛盾したスペクトログラムとなっている．これらに対して，逆 STFT 及び STFT を適用したものが Fig. 2.6 の右列である．この時に生じる周波数方向のスムージングは，時間領域での窓関数の乗算が周波数領域では畳み込みになることに起因しており，また時間方向のスムージングは STFT の際のオーバーラップシフトが原因で起こる．この結果より，スペクトログラム無矛盾性とは，スペクトログラムの近傍時間及び周波数成分の連動性と解釈することもできる．

2.8　スペクトログラム無矛盾性に基づく BSS

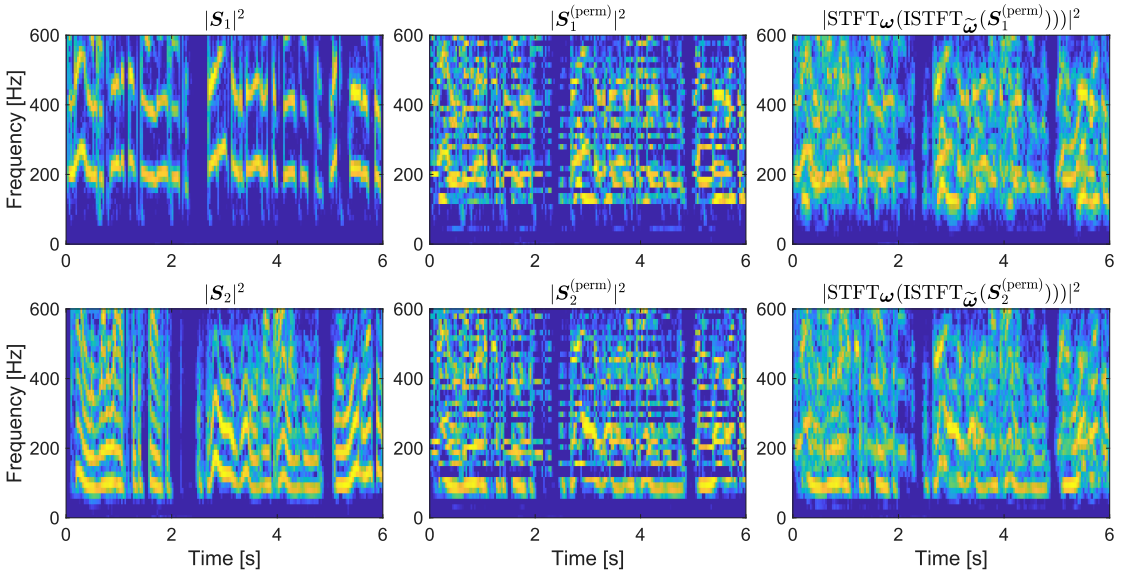
前節で述べたスペクトログラム無矛盾性を BSS に適用すると，パーミュテーション問題が少し改善し，音源分離が進むことが明らかにされている [25]．Fig. 2.8(a) と Fig. 2.8(b) の $|\boldsymbol{S}_n|^2$ は，それぞれ音楽信号と音声信号のパワースペクトログラムを表している．また，Fig. 2.8 の $|\boldsymbol{S}_n^{(\mathrm{perm})}|^2$ は，$|\boldsymbol{S}_n|^2$ に対して周波数ビンを $|\boldsymbol{S}_1|^2$ と $|\boldsymbol{S}_2|^2$ でランダムに入れ替えることで，パーミュテーション問題を人工的に起こしたものである．さらに，Fig. 2.8 の $|\mathrm{STFT}_{\boldsymbol{\omega}}(\mathrm{ISTFT}_{\tilde{\boldsymbol{\omega}}}(\boldsymbol{S}_n^{(\mathrm{perm})}))|^2$ は，$|\boldsymbol{S}_n^{(\mathrm{perm})}|^2$ を無矛盾なスペクトログラムに変換したときのスペクトログラムである．周波数方向の連続性が強調され，パーミュテーション問題が少し緩和され，左列の $|\boldsymbol{S}_n|^2$ にわずかに近づいていることが分かる．このようなパーミュテーション問題の緩和が FDICA や IVA などの BSS において，音源分離性能の向上をサポートすることが，実験的に確認されている．

2.9　本章のまとめ

本章では，Consistent ILRMA においてベースとなる音源分離アルゴリズムである ILRMA の理論及びスペクトログラム無矛盾性について述べた．次章では，ILRMA へのスペクトログラム無矛盾性の適用について述べる．



(a) Music signals: guitar (top row) and vocals (bottom row)



(b) Speech signals: female (top row) and male (bottom row)

Fig. 2.8. Smoothing effect of spectrogram consistency applied to permutation misaligned signals: (a) music and (b) speech signals. The left column shows the original source signals $|\boldsymbol{S}_n|^2$, and the center column shows their randomly permuted versions, which simulates the permutation problem and is denoted as $\boldsymbol{S}_n^{(\mathrm{perm})}$. The right column shows the consistent versions of $\boldsymbol{S}_n^{(\mathrm{perm})}$.

第 3 章

提案手法

3.1 はじめに

本章では，3.2 節でスペクトログラム無矛盾性に基づく ILRMA を提案した動機を述べる．また，3.3 節でスペクトログラム無矛盾性に基づく ILRMA について説明する．そして，3.4 節でスペクトログラム無矛盾性の適用によるプロジェクションバックの適用について述べる．その後，3.5 節で提案手法である Consistent ILRMA のアルゴリズムを示し，3.6 節で本章をまとめる．

3.2 動機

文献 [25] では，スペクトログラム無矛盾性を活用した FDICA 及び IVA が提案されている．この手法では，FDICA や IVA における分離行列 $\boldsymbol{W}_i$ の反復最適化において，分離信号 $\boldsymbol{Y}_n$ のスペクトログラム無矛盾性を，反復更新する度に毎回担保している．このスペクトログラム無矛盾性の担保によって，2.8 節で述べたように，パーミュテーション問題をより高精度に回避でき，結果的に従来の FDICA や IVA よりも高精度な BSS が可能となる．この結果は，スペクトログラムを加工するあらゆる音源分離手法に対して，スペクトログラム無矛盾性の担保が音源分離精度を向上させる可能性を示唆している．そこで本論文では，FDICA や IVA 等の既存の BSS 手法と比べて，より高精度な音源分離を実現している ILRMA に対し，文献 [25] と同様にスペクトログラム無矛盾性を考慮することが，更なる音源分離精度の向上をもたらすか否かについて比較実験を通して確認する．

3.3 スペクトログラム無矛盾性に基づく ILRMA

Consistent ILRMA では，スペクトログラム無矛盾性を担保する処理を，ILRMA の最適化アルゴリズムの毎反復に導入する．従来の ILRMA の分離行列の反復最適化更新式

(2.23)–(2.26) 及び NMF 音源モデルの反復最適化計算の式 (2.27)，(2.28) において，

$$\boldsymbol{Y}_n \leftarrow \mathrm{STFT}_{\boldsymbol{\omega}}(\mathrm{ISTFT}_{\tilde{\boldsymbol{\omega}}}(\boldsymbol{Y}_n)) \tag{3.1}$$

なる演算を挿入することで，毎回の反復においてスペクトログラム無矛盾性を担保する．式 (3.1) は，Fig. 3.1 に示すように，分離信号のスペクトログラム $\boldsymbol{Y}_n$ を無矛盾スペクトログラムの集合へと射影していることに対応する．ここで，赤色の矢印は STFT，青色の矢印は ILRMA における分離行列 $\boldsymbol{W}_i$ の反復更新，橙色の矢印は Consistent ILRMA におけるスペクトログラム無矛盾性の担保，紫色の矢印は ISTFT での射影をそれぞれ表している．また，$\boldsymbol{x}$ は観測信号のスペクトログラム，$\boldsymbol{X}$ は $\boldsymbol{x}$ のスペクトログラム，$\boldsymbol{Y}$ は分離信号，$\boldsymbol{S}$ は音源信号をそれぞれ表している．従って，もし $\boldsymbol{Y}_n$ が無矛盾であれば式 (3.1) は何もしておらず，$\boldsymbol{Y}_n$ に矛盾があれば，Figs. 2.6 及び 2.8 に示すように，時間及び周波数の両方向にスムージングがかかる．

上記の新しい処理の導入により，変数の毎回の反復更新において，矛盾したスペクトログラムが無矛盾なスペクトログラムに射影される．これによって，Consistent ILRMA では Fig. 3.1 に示すように，従来手法より真の音源信号に近づきながら音源分離を進められることが期待できる．

3.4 反復毎のプロジェクションバックの適用

ILRMA の推定は FDICA や IVA と同様に，式 (2.18) に示す任意性がある．とくに，任意の対角行列 $\boldsymbol{D}_i$ に起因する周波数毎のスケール不定性は，それ自身がスペクトログラム無矛盾性を崩してしまう．Consistent ILRMA では，$\boldsymbol{D}_i$ に起因する矛盾成分の影響を最小限に抑えるため，毎回の反復計算において式 (3.1) を行う直前に，式 (2.30) のプロジェクションバック法 [26] を適用する．また，式 (2.30) によって目的関数 (2.20) の値が変動することを防ぐために，他の変数も次のように補正する．

$$\boldsymbol{w}_{in} \leftarrow \boldsymbol{w}_{in}\lambda_{inm_{\mathrm{ref}}} \tag{3.2}$$

$$y_{ijn} \leftarrow \boldsymbol{w}_{in}^{\mathrm{H}}\boldsymbol{x}_{ij} \tag{3.3}$$

$$t_{ikn} \leftarrow t_{ikn}|\lambda_{inm_{\mathrm{ref}}}|^2 \tag{3.4}$$

ここで，m_{ref} はプロジェクションバック法で用いるリファレンスチャネルのインデクスである．

3.5 アルゴリズム

Algorithm 1 は提案手法である Consistent ILRMA のアルゴリズムである．3 行目では，式 (3.1) に示す，分離信号 $\boldsymbol{Y}_n$ の無矛盾性の担保を行っている．4，5 行目ではそれぞれ，式 (2.27)，(2.28) に示す，NMF 音源モデルの非負行列 $\boldsymbol{T}_n$ と $\boldsymbol{V}_n$ の更新を行っている．6–9 行目では，式 (2.23)–(2.26) に示す，分離行列 $\boldsymbol{W}_i$ の更新を行っている．10 行目では，式 (2.30)

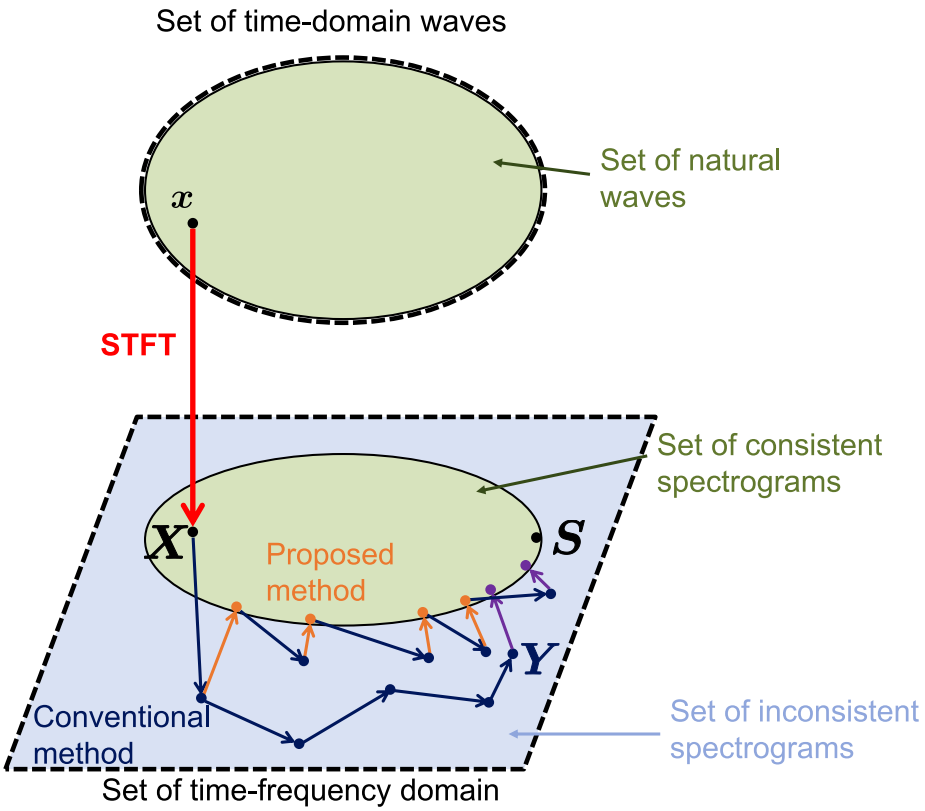


Fig. 3.1. Comparison between conventional and proposed ILRMAs, where arrows in red show STFT, arrows in dark blue show iterative update of parameters in ILRMA, arrows in orange show ensuring process of spectrogram consistency, and arrows in purple show projection onto set of consistent spectrograms applied in inverse STFT. Separated signal estimated by proposed ILRMA tends to approach to oracle source signal $\boldsymbol{S}$.

に示す，周波数毎のスケール不定性を解消するためのプロジェクションバックを行っている．11–13 行目では，式 (3.2)–(3.4) に示す，プロジェクションバックによる目的関数 (2.20) の値の変動を防ぐための，変数 $\boldsymbol{W}_i$，$\boldsymbol{Y}_n$，及び $\boldsymbol{T}_n$ の補正を行っている．したがって，従来の ILRMA と Consistent ILRMA の違いは，3 行目及び 10–13 行目の有無である．

3.6　本章のまとめ

本章では，スペクトログラム無矛盾性を担保した ILRMA の原理と実装について述べた．次章では，本章で述べた提案手法であるスペクトログラム無矛盾性を考慮した ILRMA と，従来のスペクトログラム無矛盾性を考慮していない ILRMA との比較実験の結果を示す．

Algorithm 1 Consistent ILRMA

Input: $\{\boldsymbol{x}_{ij}\}_{i=1,j=1}^{I,J}$, maxIter

Output: $\{\boldsymbol{y}_{ij}\}_{i=1,j=1}^{I,J}$

Initialize $\{\boldsymbol{T}_n\}_{n=1}^{N}, \{\boldsymbol{V}_n\}_{n=1}^{N}, \{\boldsymbol{W}_i\}_{i=1}^{I}$

for iter $= 1, 2, \cdots,$ maxIter **do**

$\boldsymbol{Y}_n \leftarrow \mathrm{STFT}_{\boldsymbol{\omega}}(\mathrm{ISTFT}_{\widetilde{\boldsymbol{\omega}}}(\boldsymbol{Y}_n))\ \forall n$

$t_{ikn} \leftarrow t_{ikn}\sqrt{\frac{\sum_j |y_{ijn}|^2 \left(\sum_{k'} t_{ik'n} v_{k'jn}\right)^{-2} v_{kjn}}{\sum_j \left(\sum_{k'} t_{ik'n} v_{k'jn}\right)^{-1} v_{kjn}}}\ \forall i, j, k, n$

$v_{kjn} \leftarrow v_{kjn}\sqrt{\frac{\sum_i |y_{ijn}|^2 \left(\sum_{k'} t_{ik'n} v_{k'jn}\right)^{-2} t_{ikn}}{\sum_i \left(\sum_{k'} t_{ik'n} v_{k'jn}\right)^{-1} t_{ikn}}}\ \forall i, j, k, n$

$\boldsymbol{U}_{in} \leftarrow \frac{1}{J}\sum_j \frac{1}{\sum_k t_{ikn} v_{kjn}} \boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^{\mathrm{H}}\ \forall i, j, n$

$\boldsymbol{w}_{in} \leftarrow (\boldsymbol{W}_i \boldsymbol{U}_{in})^{-1} \boldsymbol{e}_n\ \forall i, n$

$\boldsymbol{w}_{in} \leftarrow \boldsymbol{w}_{in}\left(\boldsymbol{w}_{in}^{\mathrm{H}} \boldsymbol{U}_{in} \boldsymbol{w}_{in}\right)^{-\frac{1}{2}}\ \forall i, n$

$y_{ijn} \leftarrow \boldsymbol{w}_{in}^{\mathrm{H}} \boldsymbol{x}_{ij}\ \forall i, j, n$

$\tilde{\boldsymbol{y}}_{ijn} = y_{ijn}\boldsymbol{\lambda}_{in}\ \forall i, j, n$

$\boldsymbol{w}_{in} \leftarrow \boldsymbol{w}_{in}\lambda_{inm_{\mathrm{ref}}}\ \forall i, n$

$y_{ijn} \leftarrow \boldsymbol{w}_{in}^{\mathrm{H}} \boldsymbol{x}_{ij}\ \forall i, j, n$

$t_{ikn} \leftarrow t_{ikn}|\lambda_{inm_{\mathrm{ref}}}|^2\ \forall i, k, n$

end for

第 4 章

実験

4.1 はじめに

本章では 4.2 節で，従来のスペクトログラム無矛盾性を担保しない ILRMA と提案手法であるスペクトログラム無矛盾性を担保する ILRMA との比較実験の条件を示す．また 4.3 節で，実験の結果と考察を述べる．そして，4.4 節で，本章をまとめる．

4.2 実験条件

提案手法であるスペクトログラム無矛盾性を担保する ILRMA の有効性を確認するために，BSS の性能を従来のスペクトログラム無矛盾性を担保しない ILRMA と比較した．ここでは，純粋にスペクトログラム無矛盾性の担保の有無に起因する性能の違いのみを評価するため，従来手法は Consistent ILRMA から式 (3.1) の演算のみを省いたものとし，その他の条件は全て Consistent ILRMA と統一した．実験には，文献 [31] と同じく，Table. 4.1 に示す音楽信号 4 種と音声信号 4 種及び混合系（インパルス応答）2 種を用いた．Fig. 4.1 にインパルス応答の収録環境を示す．評価値は音源対歪み比（source-to-distortion ratio: SDR）[32] の改善量を用いた．その他の実験条件は Table 4.2 に示す通りである．

4.3 実験結果

Fig. 4.2 は，Consistent ILRMA の各反復における矛盾成分のエネルギー値の反復毎の変化の一例である．ここで，縦軸 $\|\mathcal{E}(\mathsf{Y})\|_2^2/\|\mathsf{X}\|_2^2$ 中の X 及び Y はそれぞれ $\mathsf{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2]$ 及び $\mathsf{Y} = [\boldsymbol{Y}_1, \boldsymbol{Y}_2]$ を表し，$\mathcal{E}$ の定義は式 (2.31) に示されている．矛盾成分のエネルギー値は反復初期において大きく増加するものの，反復後半では一定の値に収束している．このことから，Consistent ILRMA の最適化ができるだけ無矛盾な解へと誘導する様子を確認できる．

Fig. 4.3 は，提案手法の各反復におけるコスト関数値の例である．コスト関数は式 (2.20) に示されている．コスト関数は，わずかに上昇することはあっても，大きく上昇することはなく，反復の半ばで一定の値に収束している．このことから，Consistent ILRMA において，音

Table 4.1. Music and speech sources obtained from SiSEC2011 dataset

Signal	Data name	Source (1/2)	Lnegth [s]
Music	bearlin-roads	acoustic guit main/vocals	14.6
Music	another_dreamer-the_ones_we_love	guitar/vocals	25.6
Music	fort minor-remember_the_name	violins synth/vocals	24.6
Music	ultimate_nz_tour	guitar/synth	18.6
speech	dev1_female4	src_1/src_2	10.0
speech	dev1_female4	src_3/src_4	10.0
speech	dev1_male4	src_1/src_2	10.0
speech	dev1_male4	src_3/src_4	10.0

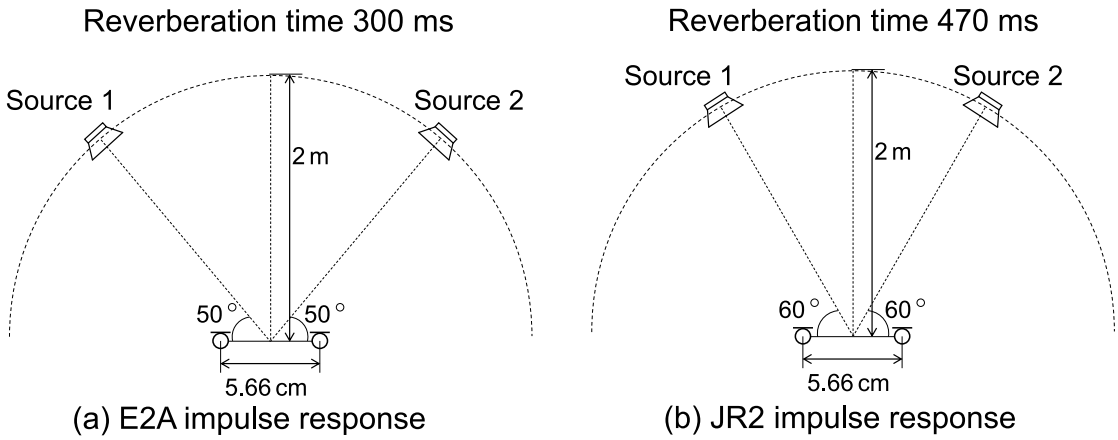


Fig. 4.1. Impulse responses used in experiment.

Table 4.2. Experimental conditions

Window function	Hann, Hamming, and Blackman window
Window length	64, 128, 256, 512, 768, 1024 ms
Window shift length	1/16, 1/8, 1/4, or 1/2 of window length
Number of bases K for each source in ILRMA	10 for music signals and 2 for speech signals
Initialization of parameters	$\boldsymbol{W}_i$: identity matrix $\boldsymbol{T}_n$ and $\boldsymbol{V}_n$: random values in the rang $(0, 1)$
Number of iterations	100
Number of trials	5 with different random seeds
Reference channel m_{ref}	1

源分離が順調に進んでいることが確認できる．

Figs. 4.4–4.7 は窓関数に Hann window を使用した場合の各シフト長及び窓長での平均 SDR 改善量の比較を示している．Hamming window 及び Blackman window を使用した場合の平均 SDR 改善量の比較については付録 A の Figs. A.1–A.32 にまとめている．図中のラベル Conv. 及び Prop. はそれぞれスペクトログラム無矛盾性を担保しない ILRMA（従来手法）と提案手法である Consistent ILRMA に対応しており，異なる初期値に対する試行や 4 種類の信号に対する結果を全てまとめて箱ひげ図で示している．Figs. 4.4–4.7 より，従来の ILRMA の SDR 改善量が低い場合に，提案法である Consistent ILRMA の SDR 改善量が従来の ILRMA の SDR 改善量を下回ることがある（例えば Fig. 4.4 の全てのシフト長における窓長が 64 ms の場合）．このことより従来の ILRMA が動かなかった場合は，スペクトログラム無矛盾性を担保することが，音源分離精度の悪化を招いているものと考えられる．Figs. 4.4–4.7 より，Consistent ILRMA の従来の ILRMA からの SDR 改善量の向上は，音声信号より音楽信号の方が大きいことが確認できる．また，Fig. 4.2 より，音声信号より音楽信号の方が無矛盾成分量が小さいことが確認できる．前述の 2 項より，無矛盾成分量が小さい信号は無矛盾成分量の大きな信号に比べて，スペクトログラム無矛盾性の担保によって分離性能は向上すると推測できる．Figs. 4.4–4.7 より，シフト長の変化による SDR 改善量の向上については大きな変化が確認できない．これは，スペクトログラム無矛盾性によるパーミュテーション問題の解決には，窓関数の畳み込みによる周波数方向の共起性が大きく影響しており，シフト長のオーバーラップによる時間方向の共起性による影響は小さいからだと考えられる．Figs. 4.4 及び 4.5 より，音楽信号では，式 (2.16) が成立しやすい長い窓長において ILRMA は高い性能を発揮することが確認できる．また Figs. 4.6 及び 4.7 より，音声信号では窓長が 512 ms のときが最も分離性能が良く，それよりも長い窓長になると分離性能が低下していることが確認できる．また Consistent ILRMA は，従来手法の分離性能が比較的高い条件において性能が向上する傾向が確認でき，場合によっては 2–3 dB 程度の改善が見られた．これは，音源分離が成功するほど，分離信号 $\boldsymbol{Y}_n$ は本来の音源信号 $\boldsymbol{S}_n$ に近づき，Consistent ILRMA においてスペクトログラム無矛盾性を担保する効果が高くなるためと推測される．

4.4　本章のまとめ

本章では，従来の ILRMA と Consistent ILRMA の音源分離実験を行い，式 (3.1) の有無による様々な窓長，シフト長及び窓長による SDR の改善量を比較した．Consistent ILRMA では，必ずしも分離精度が向上するとは言えるわけではないが，全体としては大きく向上している．本比較実験により，ILRMA において，スペクトログラム無矛盾性を考慮することによって，音源分離精度が向上する事が示された．

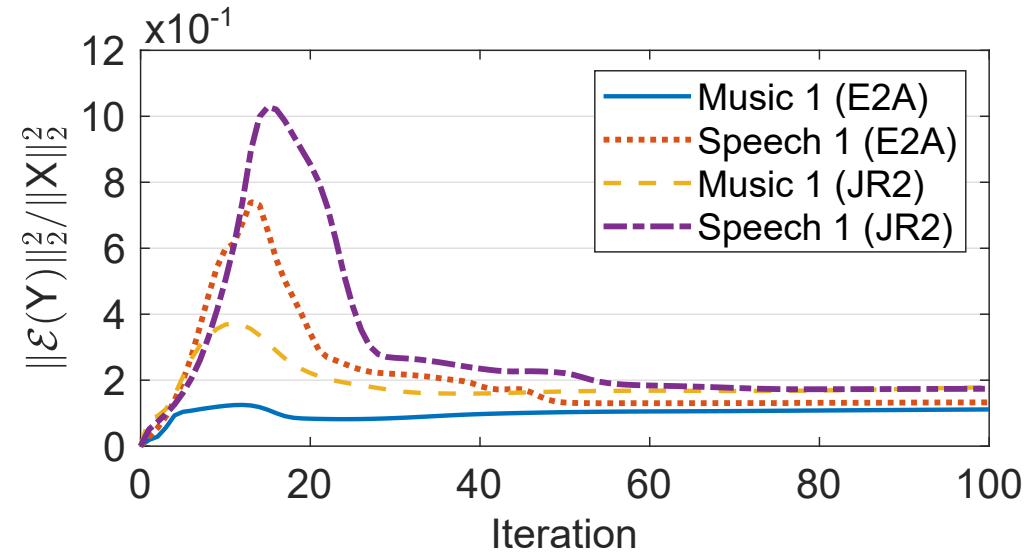


Fig. 4.2. Values of negative log-likelihood function (2.20) of proposed consistent IL-RMA+BP (window length, 256 ms; shift length, 32 ms).

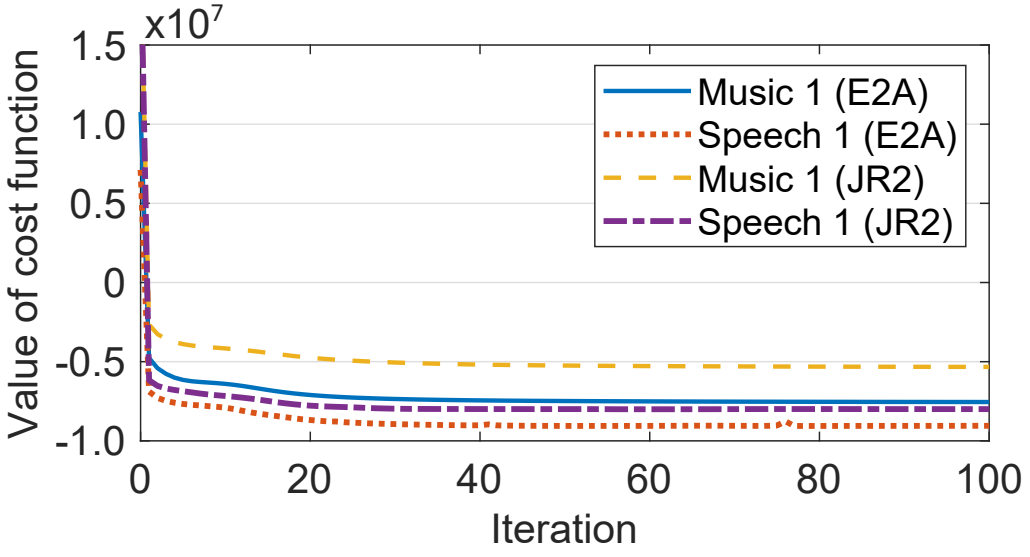


Fig. 4.3. Examples of normalized energy of inconsistent components ($||\boldsymbol{\varepsilon}(Y)||_2^2/||X||_2^2$) of proposed consistent ILRMA for music 1: **a** 256-ms-long window and 32-ms shifting and **b** 1024-ms-long window and 512-ms shifting, where X = [$\boldsymbol{X}_1$, $\boldsymbol{X}_2$], Y = [$\boldsymbol{Y}_1$, $\boldsymbol{Y}_2$], and $\boldsymbol{\varepsilon}(\cdot)$ is in (2.31).

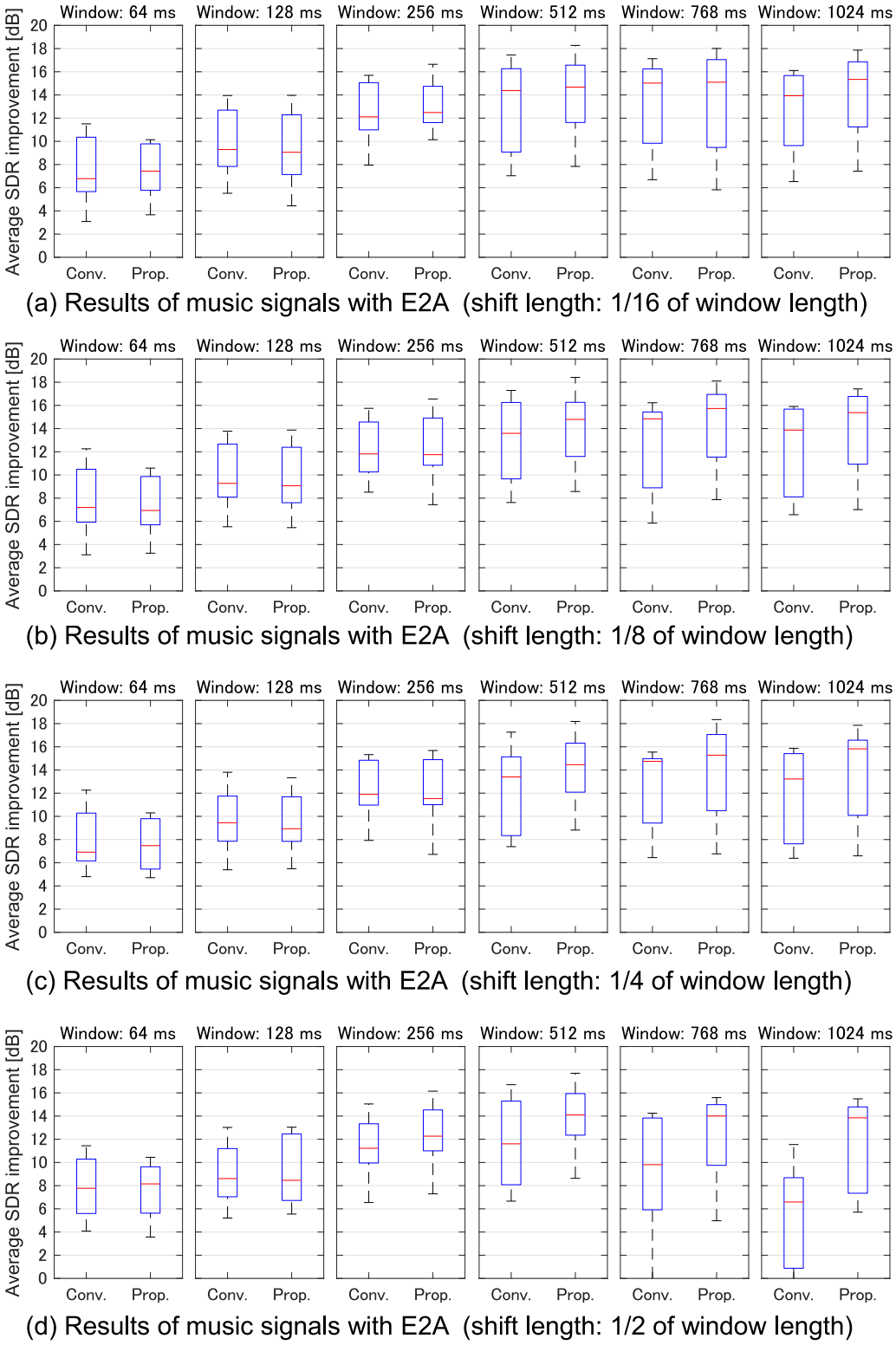


Fig. 4.4. SDR improvements for music with E2A (a) with 1/16 window shifting, (b) with 1/8 window shifting, (c) with 1/4 window shifting, and (d) with 1/2 window shifting, where "Conv." and "Prop." respectively indicate conventional and proposed methods.

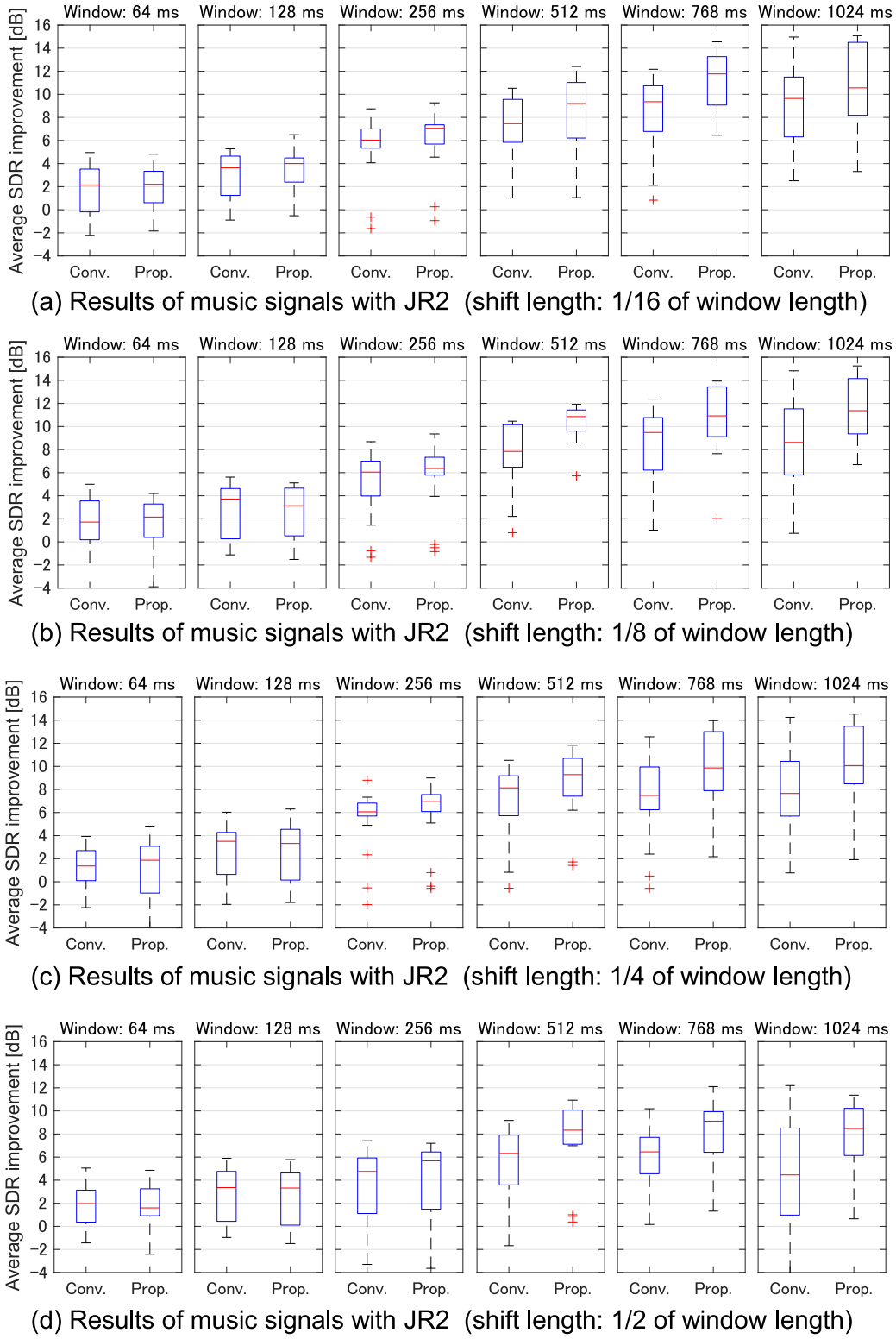


Fig. 4.5. SDR improvements for music with JR2 (a) with 1/16 window shifting, (b) with 1/8 window shifting, (c) with 1/4 window shifting, and (d) with 1/2 window shifting, where "Conv." and "Prop." respectively indicate conventional and proposed methods.

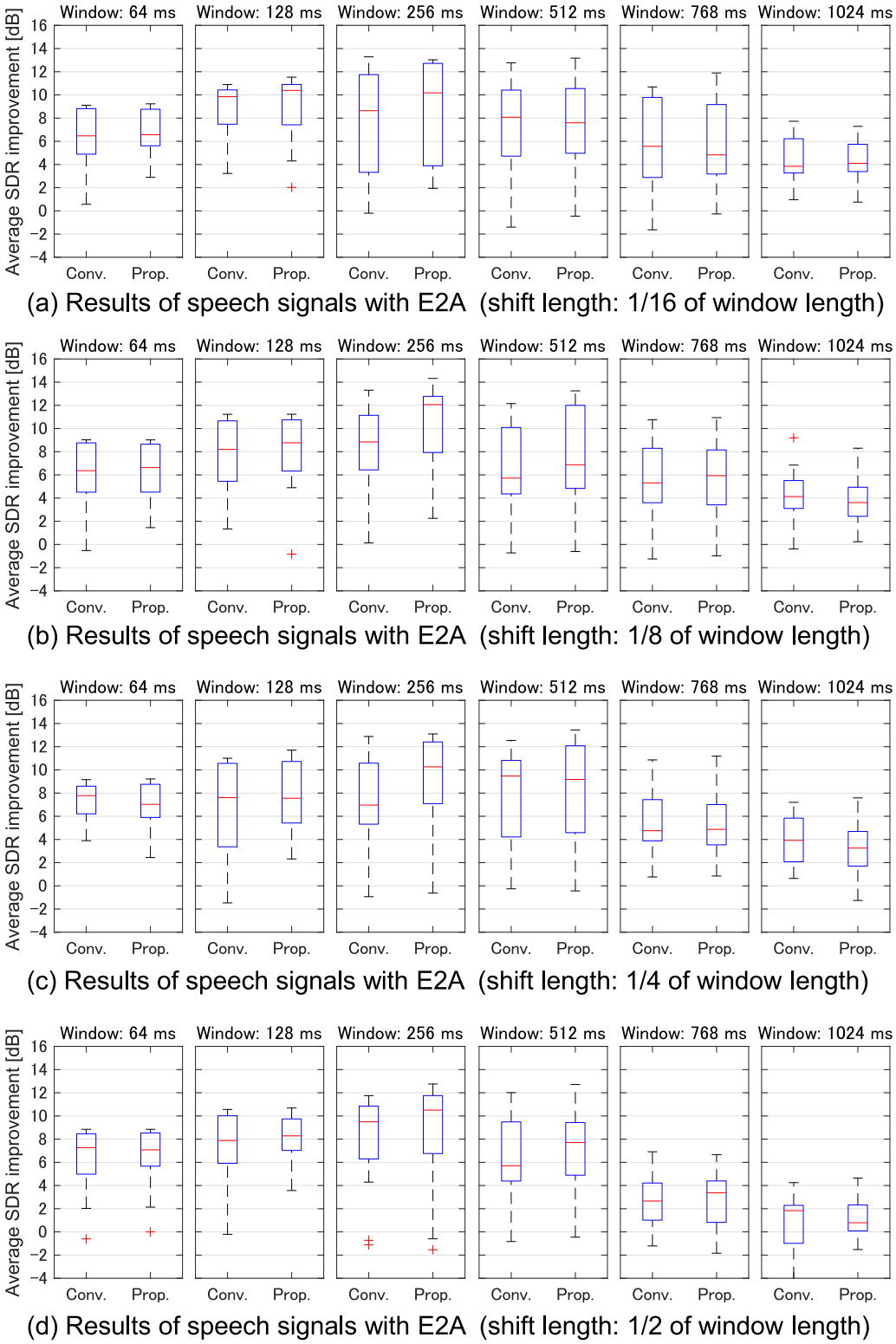


Fig. 4.6. SDR improvements for speech with E2A (a) with 1/16 window shifting, (b) with 1/8 window shifting, (c) with 1/4 window shifting, and (d) with 1/2 window shifting, where "Conv." and "Prop." respectively indicate conventional and proposed methods.

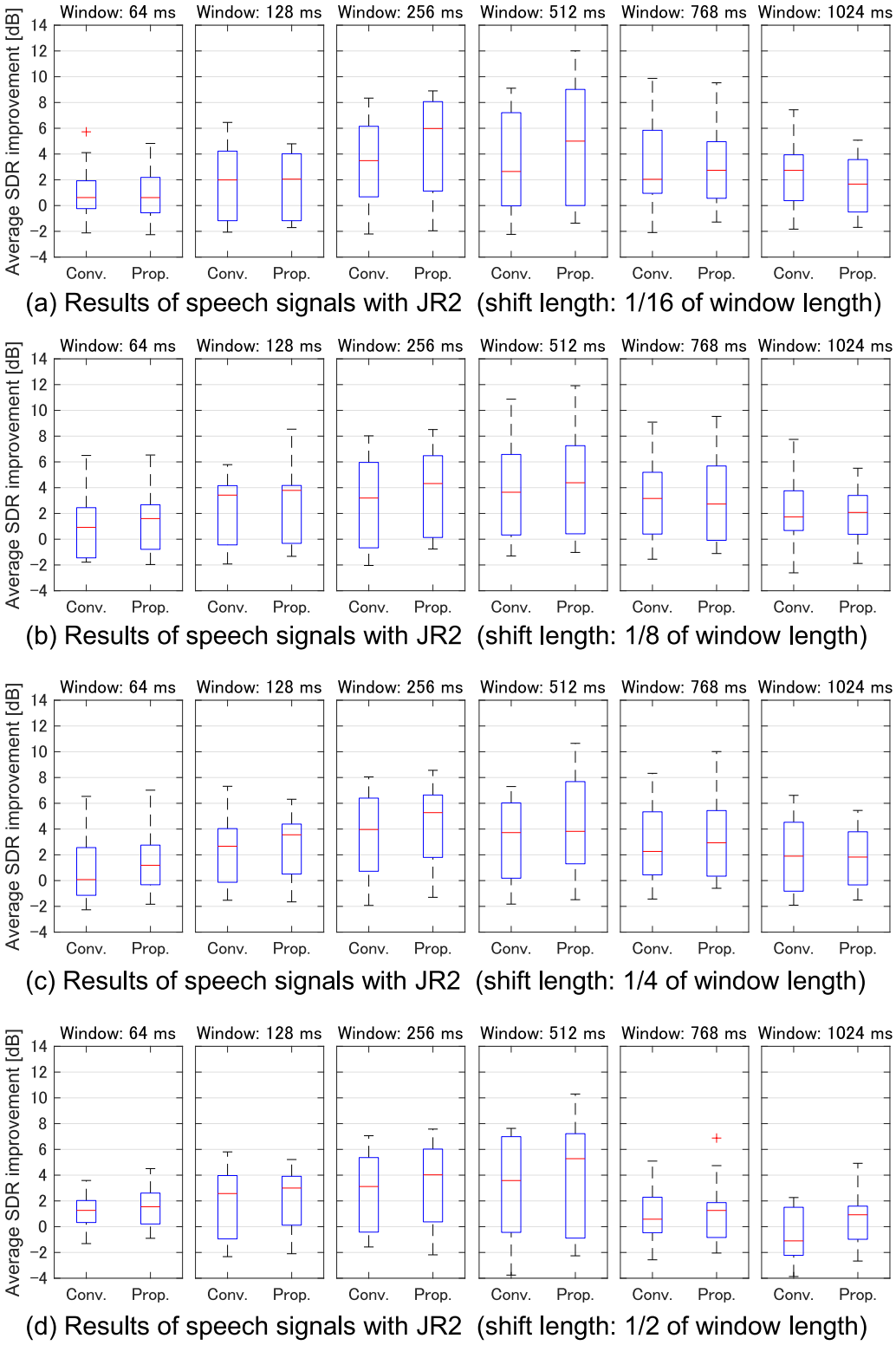


Fig. 4.7. SDR improvements for speech with JR2 (a) with 1/16 window shifting, (b) with 1/8 window shifting, (c) with 1/4 window shifting, and (d) with 1/2 window shifting, where "Conv." and "Prop." respectively indicate conventional and proposed methods.

第 5 章

結言

本論文では，高い音源分離性能を誇る ILRMA に対し，スペクトログラム無矛盾性を取り入れた ILRMA を提案し，従来の ILRMA と Consistent ILRMA で音源分離実験を行い，音源分離精度の向上を比較した．そして，ILRMA においてもスペクトログラム無矛盾性を考慮することによって，音源分離精度が向上することを実験的に示した．

Consistent ILRMA では，多くの場合において，従来手法を上回る音源分離精度を実現した．しかし，分離精度が向上しないばかりか，低下する場合も見られた．まず，Consistent ILRMA は従来手法が十分に性能を発揮した場合に，さらなる精度向上が期待できる手法である．そのため，従来手法が性能を十分に発揮することができないような条件設定では，もちろん Consistent ILRMA での精度向上は期待できない．

Consistent ILRMA は，分離信号 $\boldsymbol{Y}_n$ に対して，スペクトログラム無矛盾性を担保した後に，NMF 音源モデル $\boldsymbol{T}_n$ 及び $\boldsymbol{V}_n$ を用いて分離行列 $\boldsymbol{W}_i$ を更新している．よって，NMF 音源モデル $\boldsymbol{T}_n$ 及び $\boldsymbol{V}_n$ が，スペクトログラム無矛盾性の担保の有無で具体的にどのような変化が表れているのかを調査すれば，また一歩進んだ研究ができると考えられる．そのため，前述のような変化を調査することが求められる．

謝辞

本論文は，独立行政法人 国立高等専門学校機構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，論文執筆や研究に関する議論など，細部にわたるまで丁寧にご指導いただきました．前期は，コロナ禍による突然のリモートでの研究になりましたが，素早い対応により，快適に研究をすることができました．また，身の回りの設備には出費を惜しまず整えてもらうなど，他にも数えればキリがないですが，本当にさまざまな支援をしていただきました．感謝してもしきれません．

本論の副査である柿元健准教授には，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．別の視点からの指摘には気づかされることが多く，とてもためになりました．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，共同研究を通じ多数のご支援とご助言をいただきました．香川高専とはまた違った雰囲気で研究を進めることができ，とても新鮮でいい経験ができました．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 2 年の大島風雅氏には，研究発表会での助言など，また山地修平氏には，論文執筆の仕方や，その際の TEX の使い方など，数々のご支援をいただきました．深謝いたします．

また，北村研究室同期の香西海斗氏・多田敏貴氏にはゼミや日頃のディスカッションのほか，1 年に亘る研究室生活を様々な面で支えていただきました．最も身近で関わってきたかけがえのない存在であり，研究以外のことでも数えきれないほど助けてもらいました．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 Process.*, 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. Audio, Speech, Lang. Process.*, 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. Workshop on Applications of Signal Process. to Audio and Acoust.*, pp. 189–192, 2011.

[5] 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. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1626–1641, 2016.

[6] 北村大地, 小野順貴, 沢田宏, 亀岡弘和, 猿渡洋, “独立低ランク行列分析に基づくブラインド音源分離,” *IEICE Technical Report*, EA2017-56, vol. 117, no. 255, pp. 73–80, 2017.

[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] 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 Audio Process.*, vol. 12, no. 5, pp. 530–538, Sep. 2004.

[9] Y. Mitsui, D. Kitamura, S. Takamichi, N. Ono, and H. Saruwatari, “Blind source separation based on independent low-rank matrix analysis with sparse regularization for time-series activity,” *Proc. Int. Conf. Acoust., Speech and Signal Process.*, pp. 21–25, 2017.

[10] H. Kagami, H. Kameoka, and M. Yukawa, “Joint separation and dereverberation of reverberant mixtures with determined multichannel non-negative matrix factorization,” *Proc. Int. Conf. Acoust., Speech and Signal Process.*, pp. 31–35, 2018.

[11] R. Ikeshita and Y. Kawaguchi, "Independent low-rank matrix analysis based on multivariate complex exponential power distribution," *Proc. Int. Conf. Acoust., Speech and Signal Process.*, pp. 741–745, 2018.

[12] D. Kitamura, S. Mogami, Y. Mitsui, N. Takamune, H. Saruwatari, N. Ono, Y. Takahashi, and K. Kondo, "Generalized independent low-rank matrix analysis using heavy-tailed distributions for blind source separation," *EURASIP J. Adv. Signal Process.*, vol. 2018, 2018.

[13] K. Yoshii, K. Kitamura, Y. Bando, E. Nakamura, and T. Kawahara, "Independent low-rank tensor analysis for audio source separation," *Proc. European Signal Process. Conf.*, pp. 1657–1661, 2018.

[14] R. Ikeshita, "Independent positive semidefinite tensor analysis in blind source separation," *Proc. European Signal Process. Conf.*, pp. 1652–1656, 2018.

[15] R. Ikeshita, N. Ito, T. Nakatani, and H. Sawada, "Independent low-rank matrix analysis with decorrelation learning," *Proc. Workshop on Applications of Signal Process. to Audio and Acoust.*, pp. 288–292, 2019.

[16] N. Makishima, S. Mogami, N. Takamune, D. Kitamura, H. Sumino, S. Takamichi, H. Saruwatari, and N. Ono, "Independent deeply learned matrix analysis for determined audio source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 10, pp. 1601–1615, 2019.

[17] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara, "Semi-supervised multichannel speech enhancement with a deep speech prior," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2197–2212, 2019.

[18] S. Mogami, N. Takamune, D. Kitamura, H. Saruwatari, Y. Takahashi, K. Kondo, and N. Ono, "Independent low-rank matrix analysis based on time-variant sub-Gaussian source model for determined blind source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 503–518, 2020.

[19] Y. Takahashi, D. Kitahara, K. Matsuura, and A. Hirabayashi, "Determined source separation using the sparsity of impulse responses," *Proc. Int. Conf. Acoust., Speech and Signal Process.*, pp. 686–690, 2020.

[20] M. Togami, "Multi-channel speech source separation and dereverberation with sequential integration of determined and underdetermined models," *Proc. Int. Conf. Acoust., Speech and Signal Process.*, pp. 231–235, 2020.

[21] S. Kanoga, T. Hoshino, and H. Asoh, "Independent low-rank matrix analysis-based automatic artifact reduction technique applied to three BCI paradigms," *Front. Hum. Neurosci.*, vol. 14, 2020.

[22] K. Yatabe and D. Kitamura, "Time-frequency-masking-based determined BSS with application to sparse IVA," *Proc. Int. Conf. Acoust., Speech and Signal Process.*, pp.

715–719, 2019.

[23] J. Le Roux and E. Vincent, “Consistent Wiener filtering for audio source separation,” *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 217–220, 2013.

[24] K. Yatabe, Y. Masuyama, T. Kusano, and Y. Oikawa, “Representation of complex spectrogram via phase conversion,” *Acoust. Sci. & Tech.*, vol. 40, no. 3, pp. 170–177, 2019.

[25] K. Yatabe, “Consistent ICA: Determined BSS meets spectrogram consistency,” *IEEE Signal Process. Lett.*, vol. 27, pp. 870–874, 2020.

[26] K. Matsuoka and S. Nakashima, “Minimal distortion principle for blind source separation,” *Proc. Int. Conf. Independent Comp. Anal. Blind Signal Separation*, pp. 722–727, 2001.

[27] 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.

[28] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization” *Proc. Annual Conf. Neural Info. Process. Syst.*, pp. 556–562, 2000.

[29] N. Ono and S. Miyabe, “Auxiliary-function-based independent component analysis for super-Gaussian sources,” *Proc. Int. Conf. Latent Variable Anal. Signal Separation.*, pp. 165–172, 2010.

[30] S. Douglas and S. Amari, “Natural-gradient adaptation,” in *Unsupervised adaptive filtering*, Ed. S. Haykin, vol. I, pp. 13-61, Wiley, 2000.

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

[32] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

発表文献一覧

国内学会

1. **豊島直**, 北村大地, 矢田部浩平, “スペクトログラム無矛盾性を用いた独立低ランク行列分析,” 日本音響学会講演論文集, 2-R2-13, pp. 291–294, 2020.

受賞

1. 日本音響学会 2020 年秋季研究発表会 学生優秀発表賞

付録 A

Hamming window 及び Blackman window を用いた場合の比較実験結果

Figs. A.1–A.16 に Hamming window での比較実験結果の箱ひげ図を，Figs. A.17–A.32 に Blackman window での比較実験結果の箱ひげ図を示す.

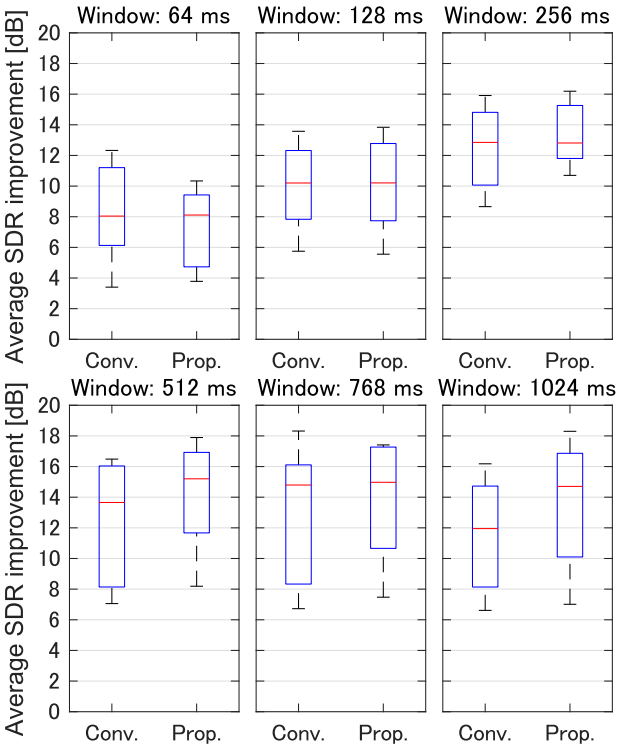


Fig. A.1. Average SDR improvements for synthesized music mixtures (music 1–4) with E2A, where Hamming window is used with 1/16 length in STFT.

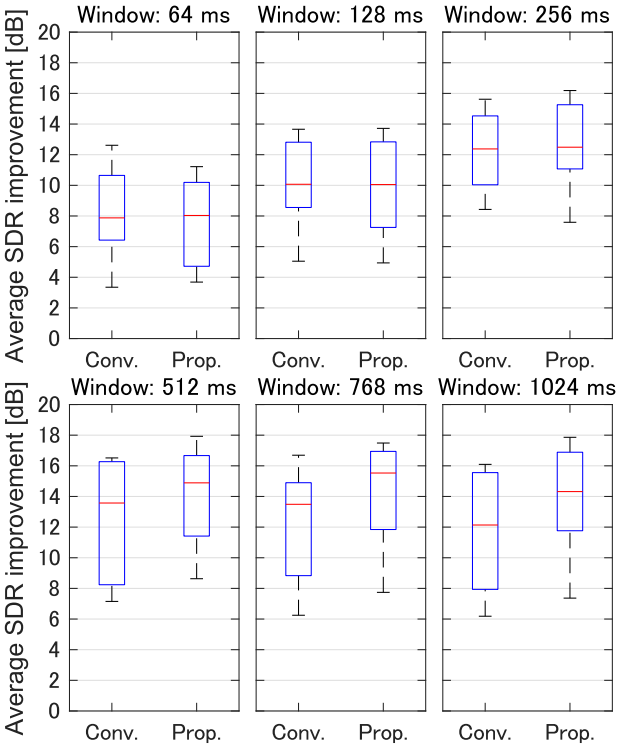


Fig. A.2. Average SDR improvements for synthesized music mixtures (music 1–4) with E2A, where Hamming window is used with 1/8 length in STFT.

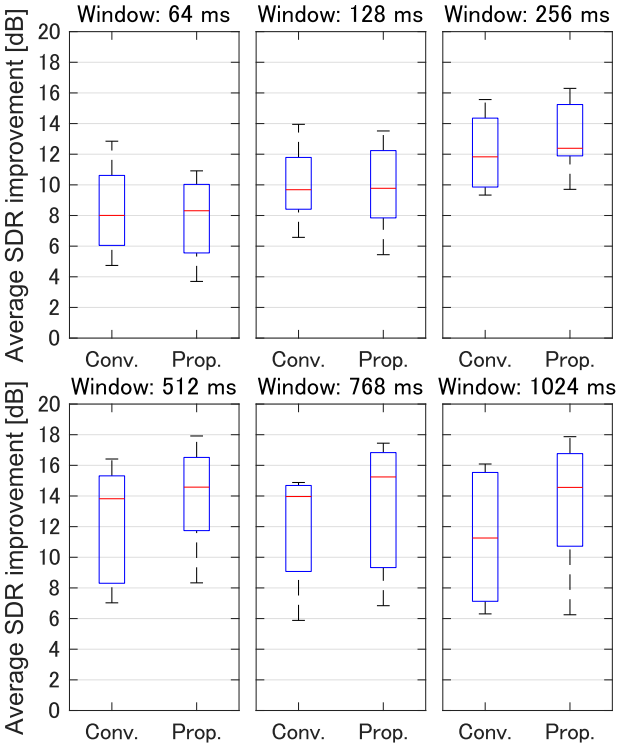


Fig. A.3. Average SDR improvements for synthesized music mixtures (music 1–4) with E2A, where Hamming window is used with 1/4 length in STFT.

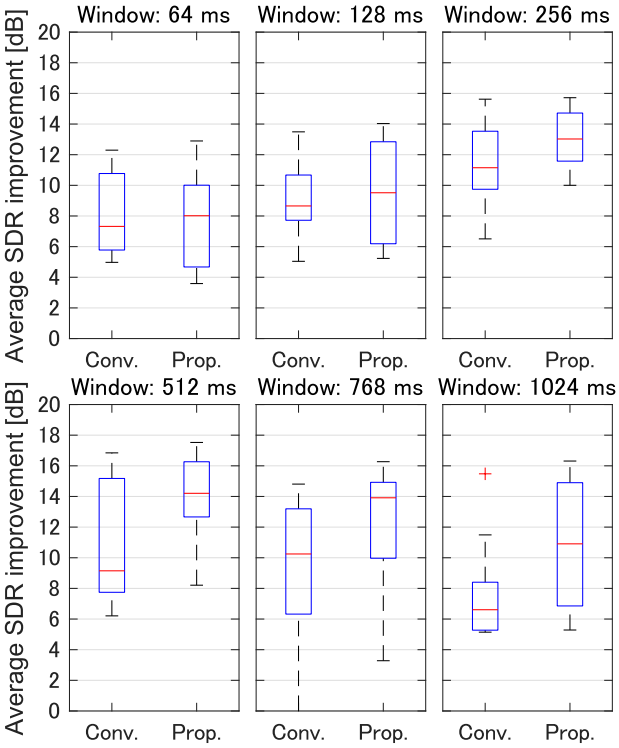


Fig. A.4. Average SDR improvements for synthesized music mixtures (music 1–4) with E2A, where Hamming window is used with 1/2 length in STFT.

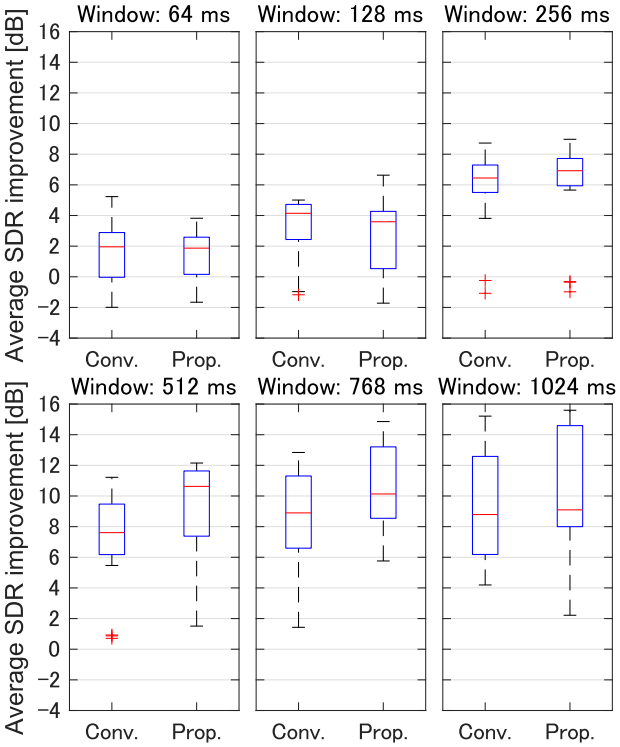


Fig. A.5. Average SDR improvements for synthesized music mixtures (music 1–4) with JR2, where Hamming window is used with 1/16 length in STFT.

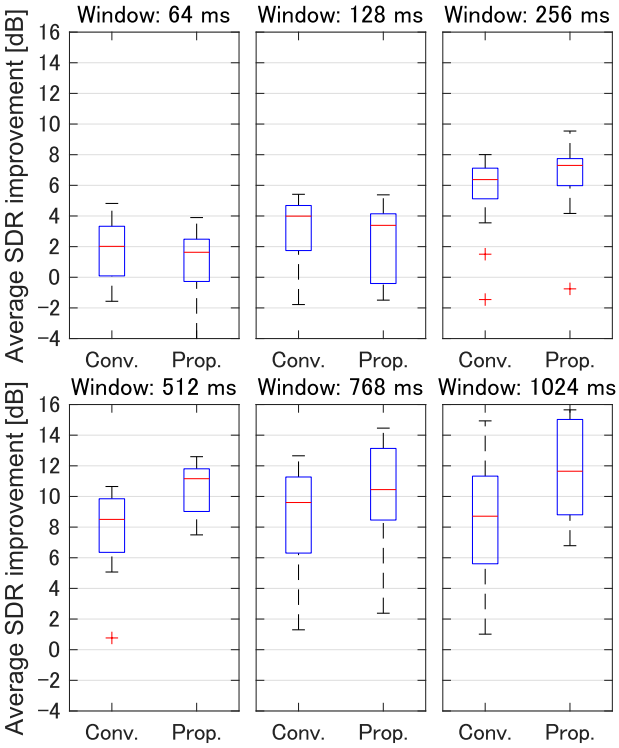


Fig. A.6. Average SDR improvements for synthesized music mixtures (music 1–4) with JR2, where Hamming window is used with 1/8 length in STFT.

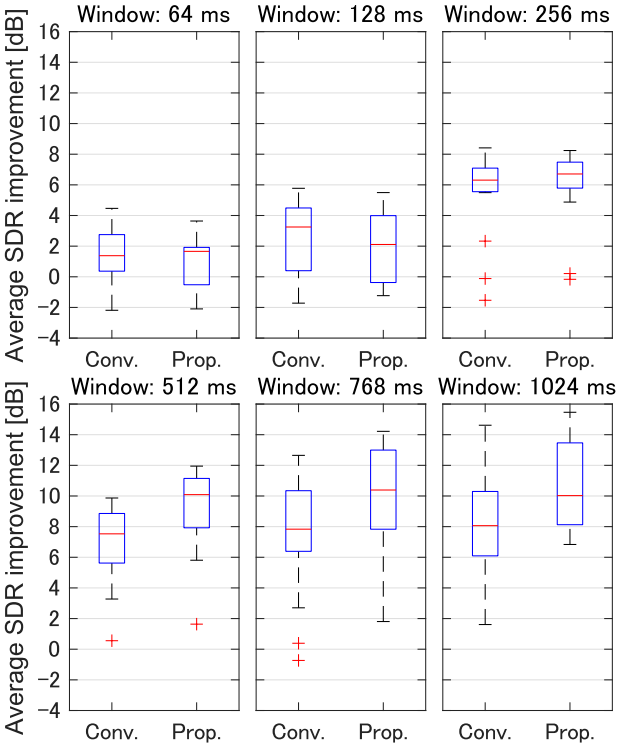


Fig. A.7. Average SDR improvements for synthesized music mixtures (music 1–4) with JR2, where Hamming window is used with 1/4 length in STFT.

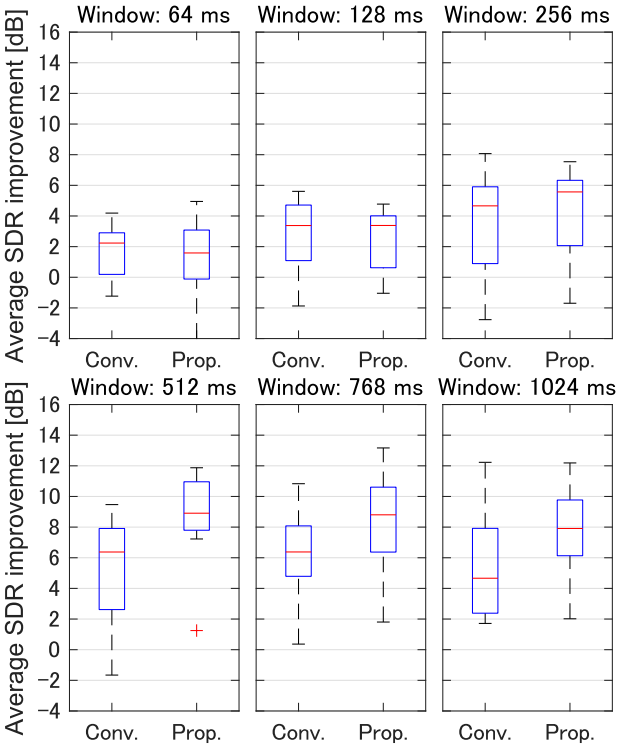


Fig. A.8. Average SDR improvements for synthesized music mixtures (music 1–4) with JR2, where Hamming window is used with 1/2 length in STFT.

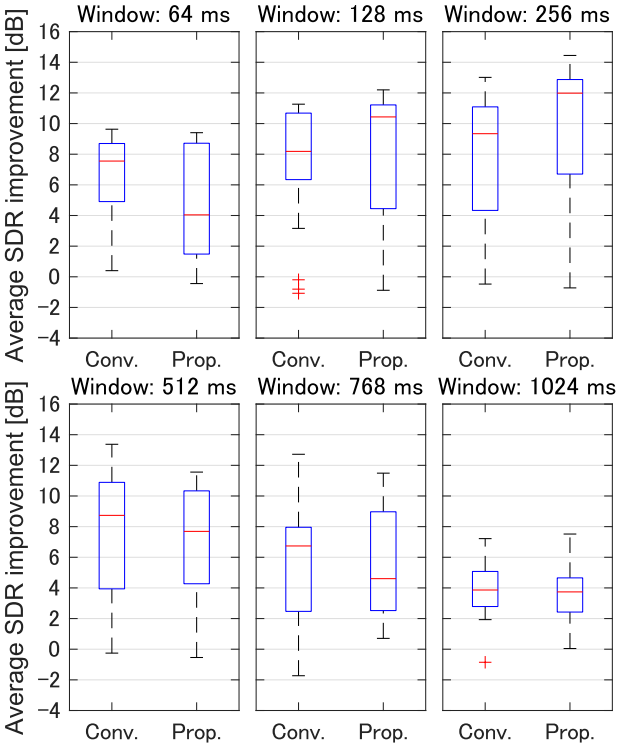


Fig. A.9. Average SDR improvements for synthesized music mixtures (speech 1–4) with E2A, where Hamming window is used with 1/16 length in STFT.

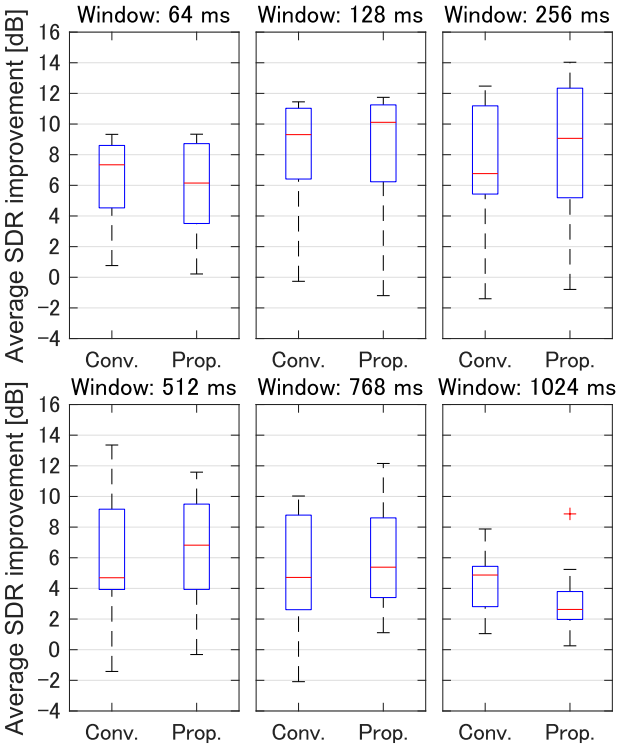


Fig. A.10. Average SDR improvements for synthesized music mixtures (speech 1–4) with E2A, where Hamming window is used with 1/8 length in STFT.

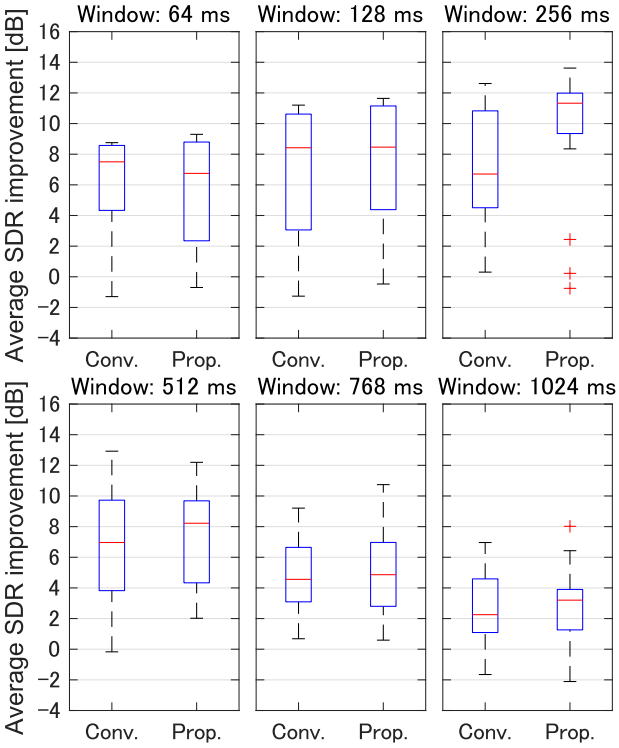


Fig. A.11. Average SDR improvements for synthesized music mixtures (speech 1–4) with E2A, where Hamming window is used with 1/4 length in STFT.

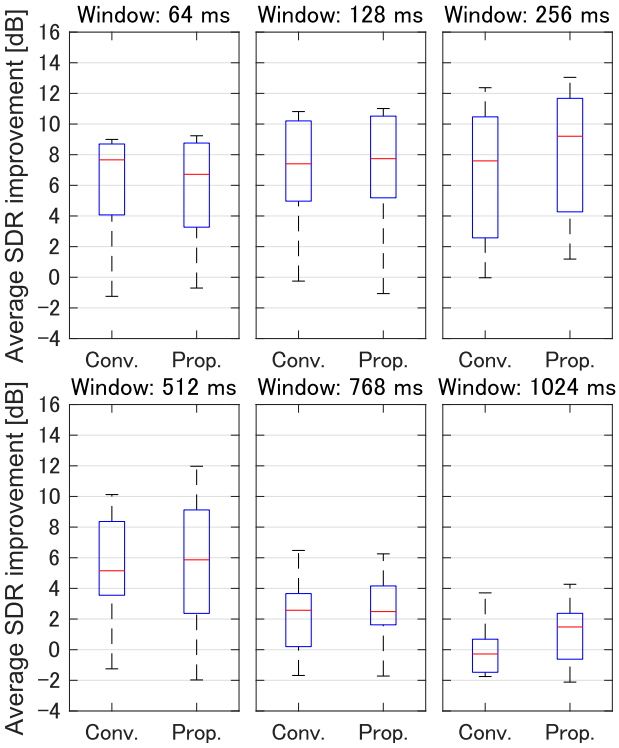


Fig. A.12. Average SDR improvements for synthesized music mixtures (speech 1–4) with E2A, where Hamming window is used with 1/2 length in STFT.

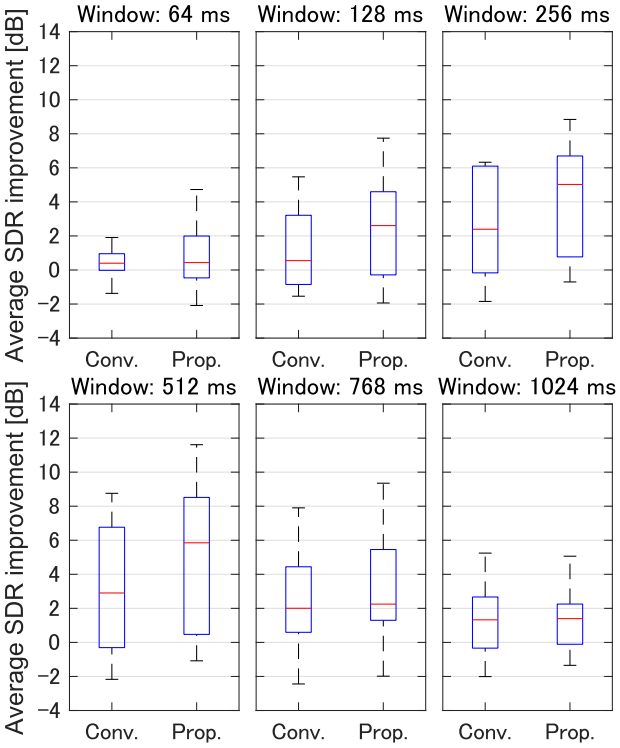


Fig. A.13. Average SDR improvements for synthesized music mixtures (speech 1–4) with JR2, where Hamming window is used with 1/16 length in STFT.

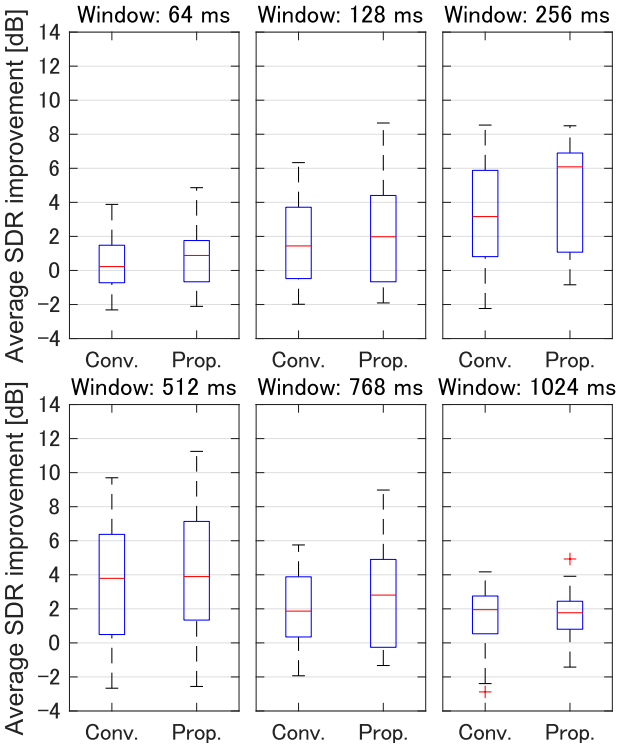


Fig. A.14. Average SDR improvements for synthesized music mixtures (speech 1–4) with JR2, where Hamming window is used with 1/8 length in STFT.

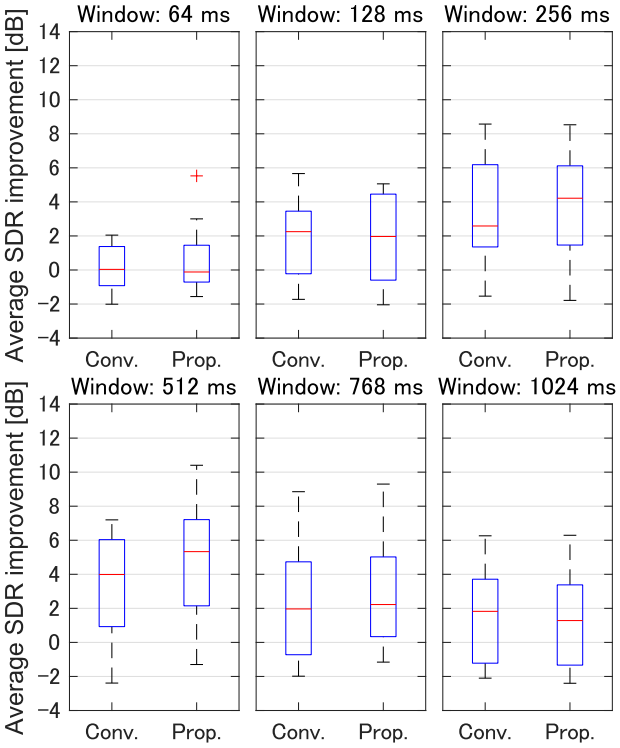


Fig. A.15. Average SDR improvements for synthesized music mixtures (speech 1–4) with JR2, where Hamming window is used with 1/4 length in STFT.

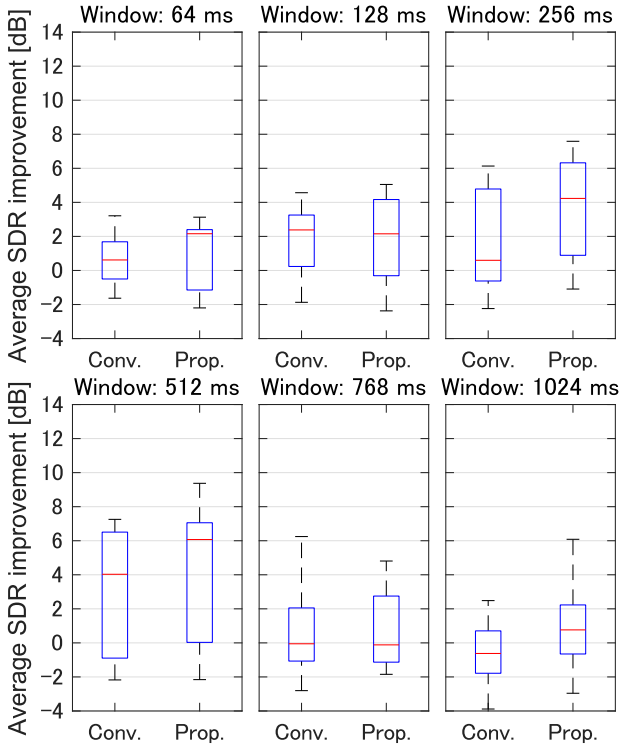


Fig. A.16. Average SDR improvements for synthesized music mixtures (speech 1–4) with JR2, where Hamming window is used with 1/2 length in STFT.

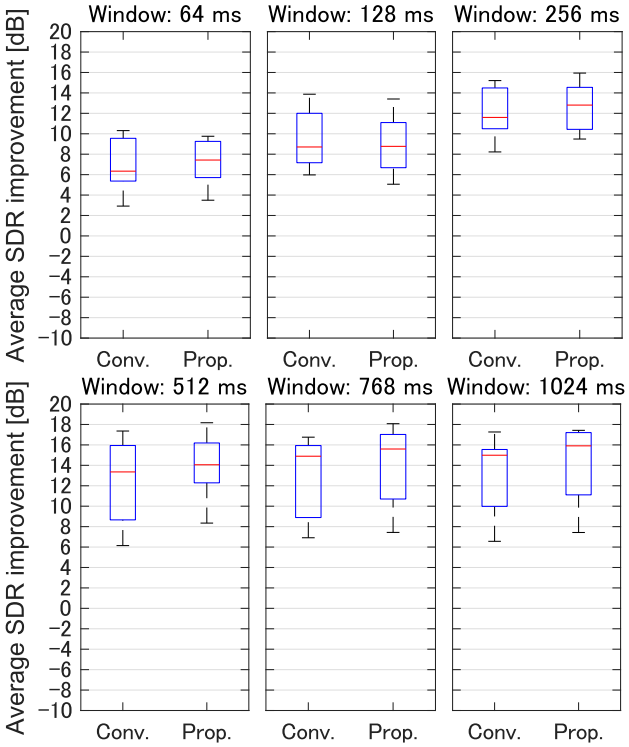


Fig. A.17. Average SDR improvements for synthesized music mixtures (music 1–4) with E2A, where Blackman window is used with 1/16 length in STFT.

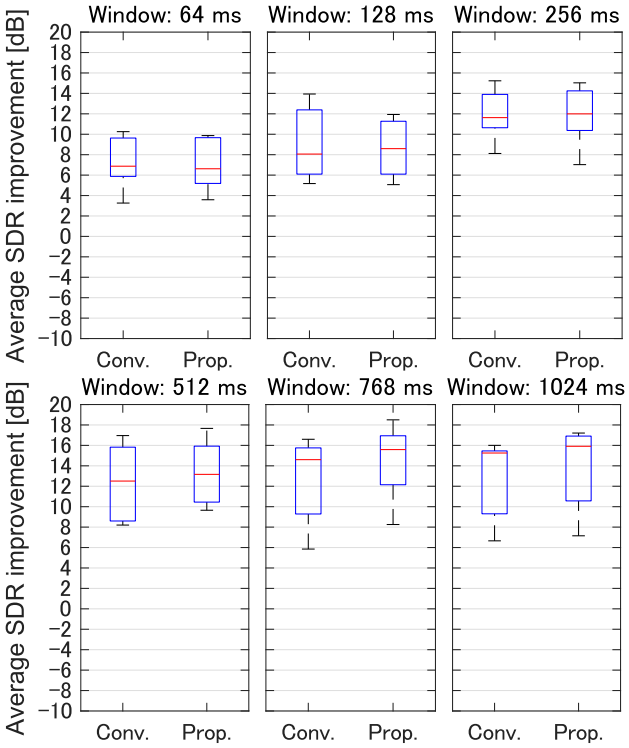


Fig. A.18. Average SDR improvements for synthesized music mixtures (music 1–4) with E2A, where Blackman window is used with 1/8 length in STFT.

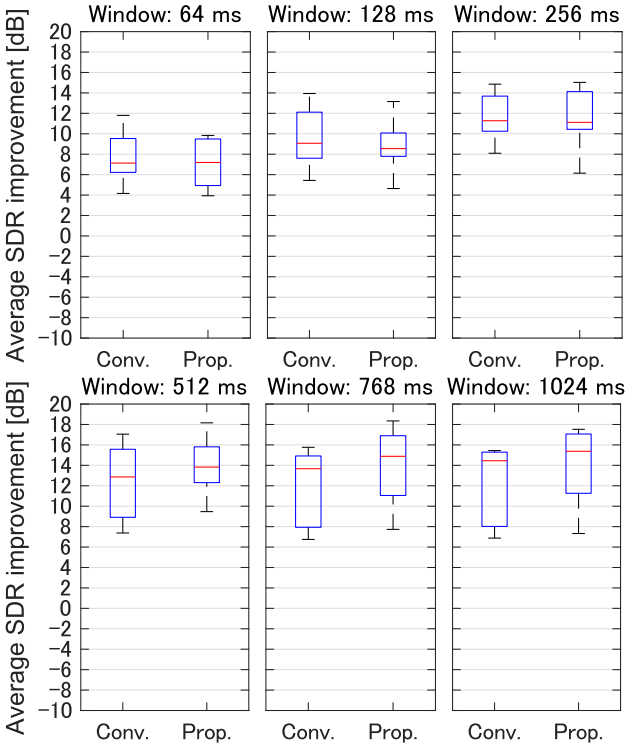


Fig. A.19. Average SDR improvements for synthesized music mixtures (music 1–4) with E2A, where Blackman window is used with 1/4 length in STFT.

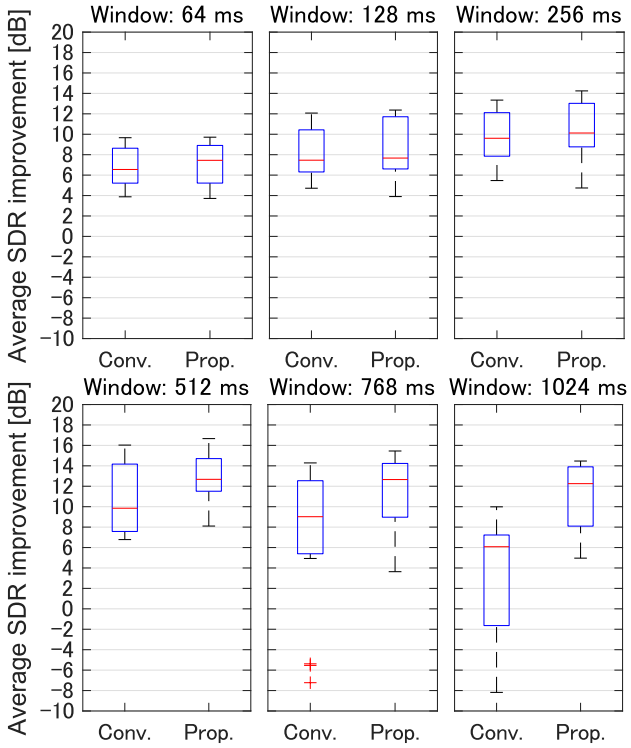


Fig. A.20. Average SDR improvements for synthesized music mixtures (music 1–4) with E2A, where Blackman window is used with 1/2 length in STFT.

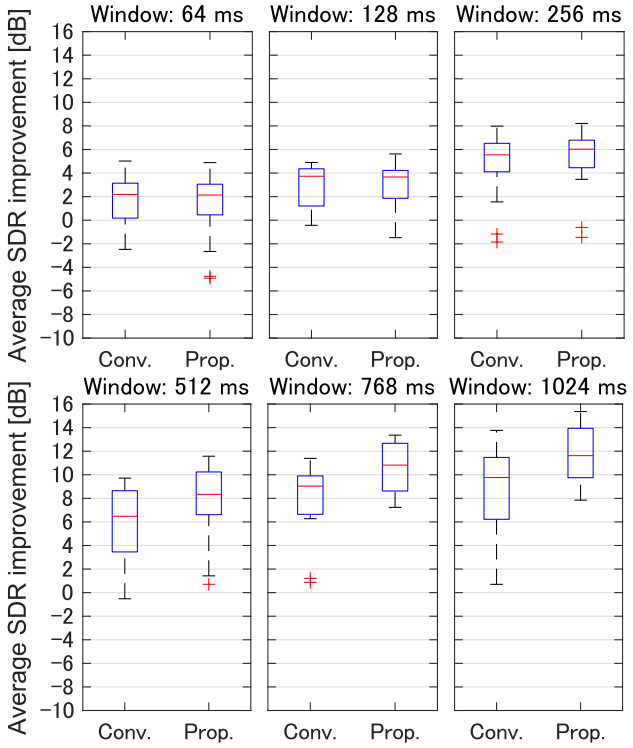


Fig. A.21. Average SDR improvements for synthesized music mixtures (music 1–4) with JR2, where Blackman window is used with 1/16 length in STFT.

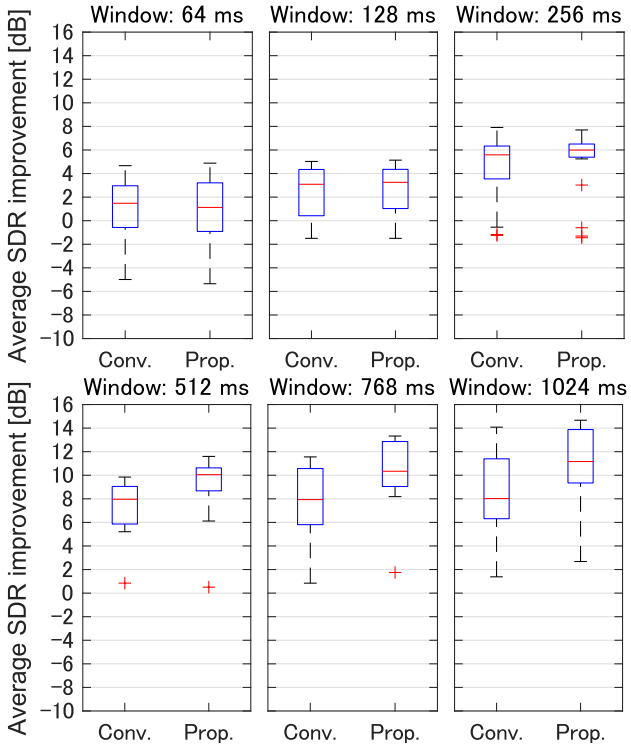


Fig. A.22. Average SDR improvements for synthesized music mixtures (music 1–4) with JR2, where Blackman window is used with 1/8 length in STFT.

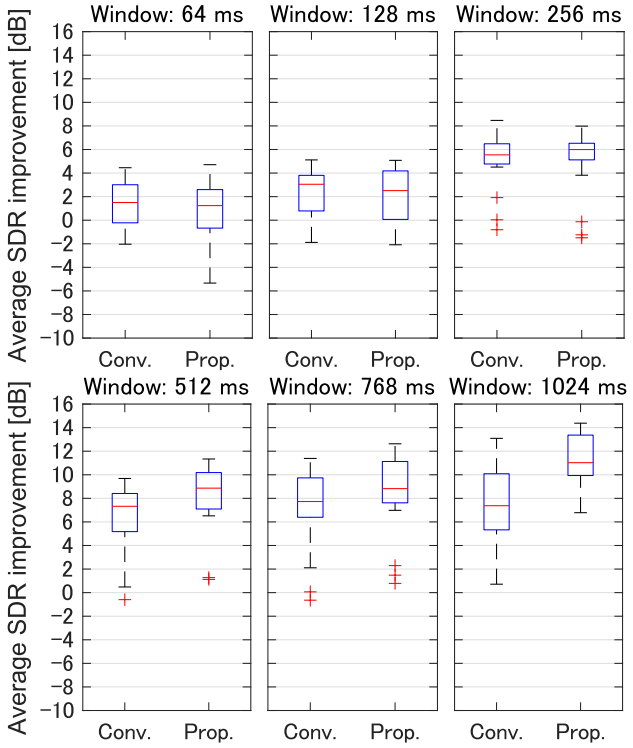


Fig. A.23. Average SDR improvements for synthesized music mixtures (music 1–4) with JR2, where Blackman window is used with 1/4 length in STFT.

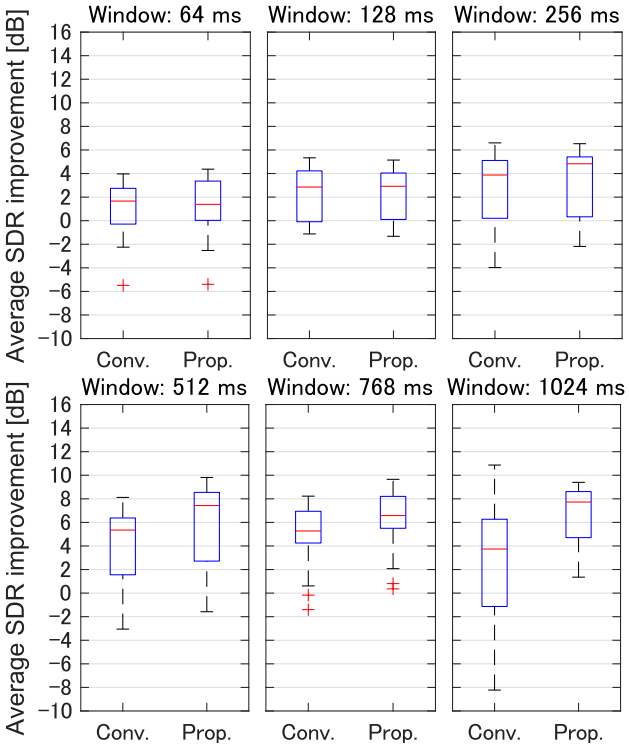


Fig. A.24. Average SDR improvements for synthesized music mixtures (music 1–4) with JR2, where Blackman window is used with 1/2 length in STFT.

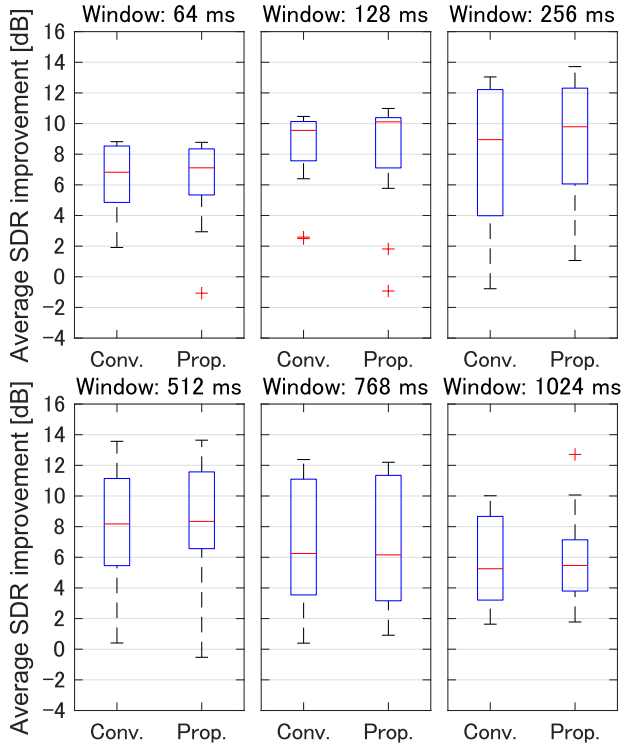


Fig. A.25. Average SDR improvements for synthesized music mixtures (speech 1–4) with E2A, where Blackman window is used with 1/16 length in STFT.

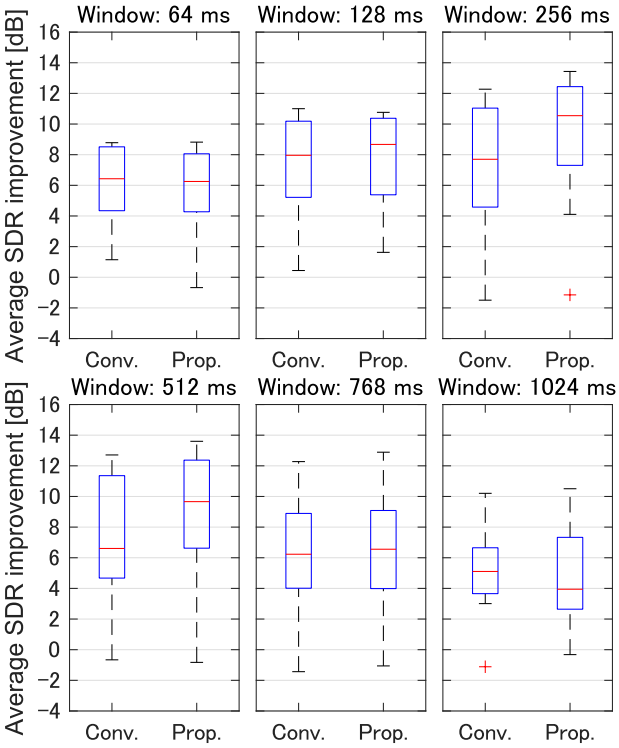


Fig. A.26. Average SDR improvements for synthesized music mixtures (speech 1–4) with E2A, where Blackman window is used with 1/8 length in STFT.

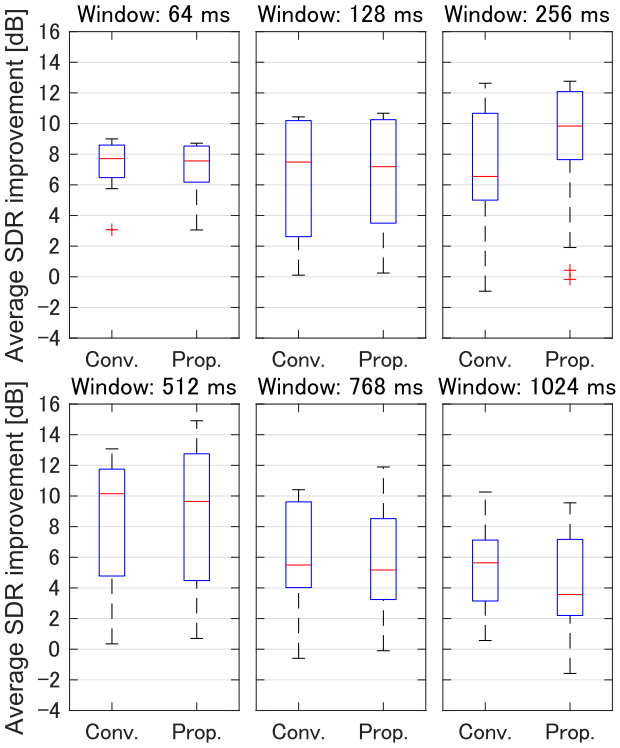


Fig. A.27. Average SDR improvements for synthesized music mixtures (speech 1–4) with E2A, where Blackman window is used with 1/4 length in STFT.

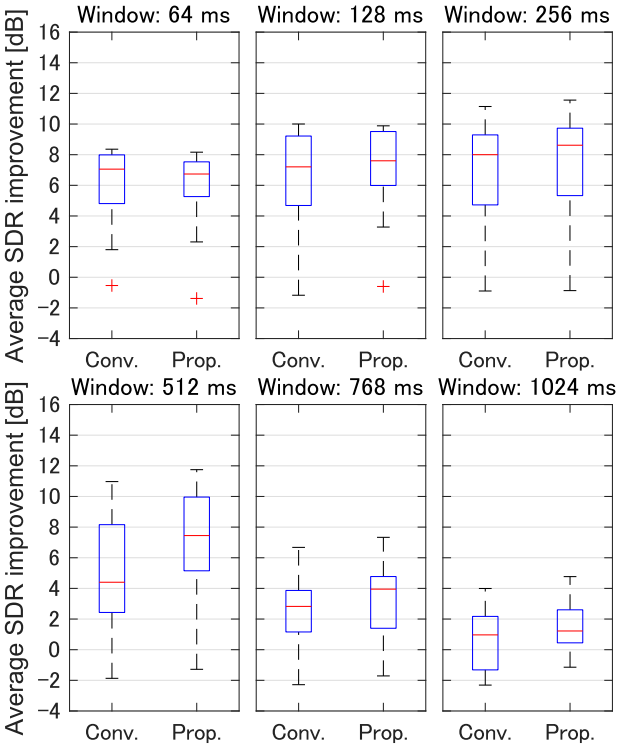


Fig. A.28. Average SDR improvements for synthesized music mixtures (speech 1–4) with E2A, where Blackman window is used with 1/2 length in STFT.

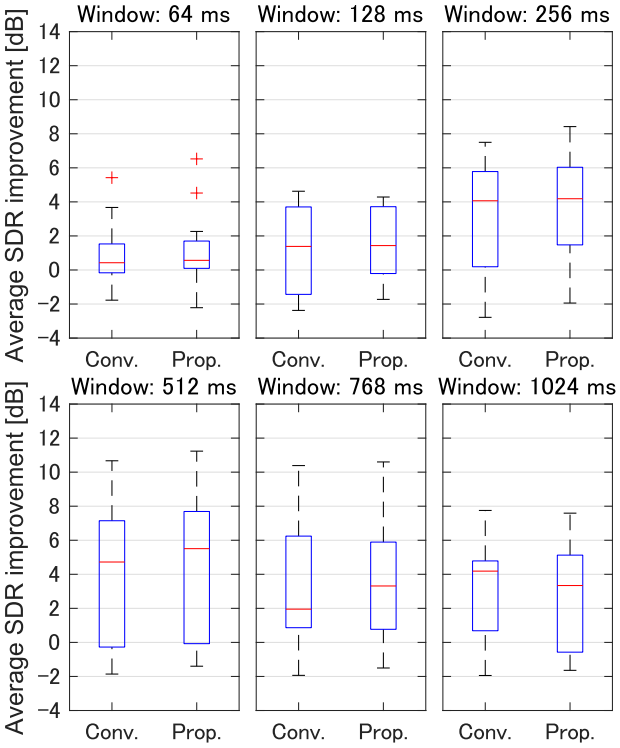


Fig. A.29. Average SDR improvements for synthesized music mixtures (speech 1–4) with JR2, where Blackman window is used with 1/16 length in STFT.

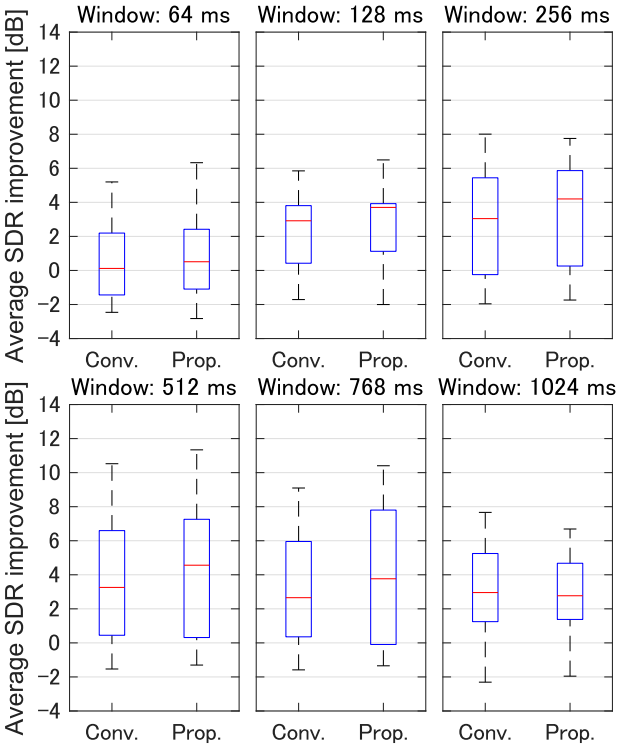


Fig. A.30. Average SDR improvements for synthesized music mixtures (speech 1–4) with JR2, where Blackman window is used with 1/8 length in STFT.

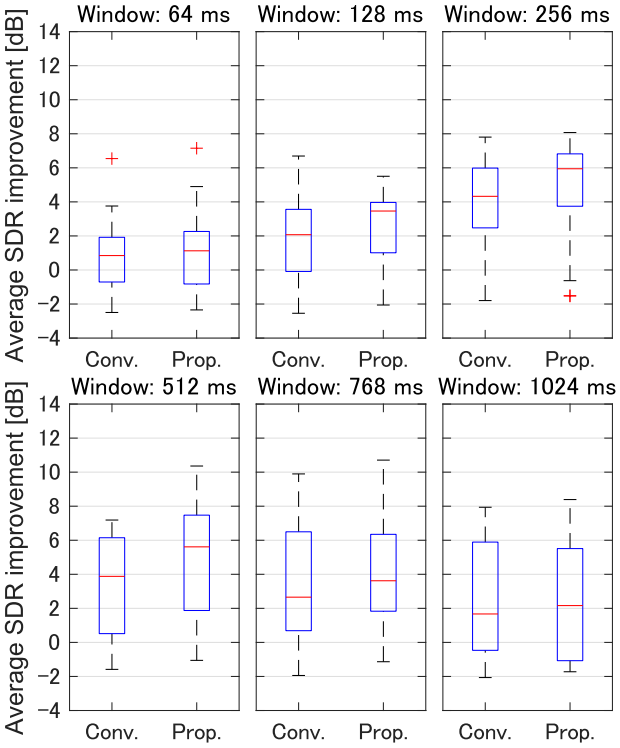


Fig. A.31. Average SDR improvements for synthesized music mixtures (speech 1–4) with JR2, where Blackman window is used with 1/4 length in STFT.

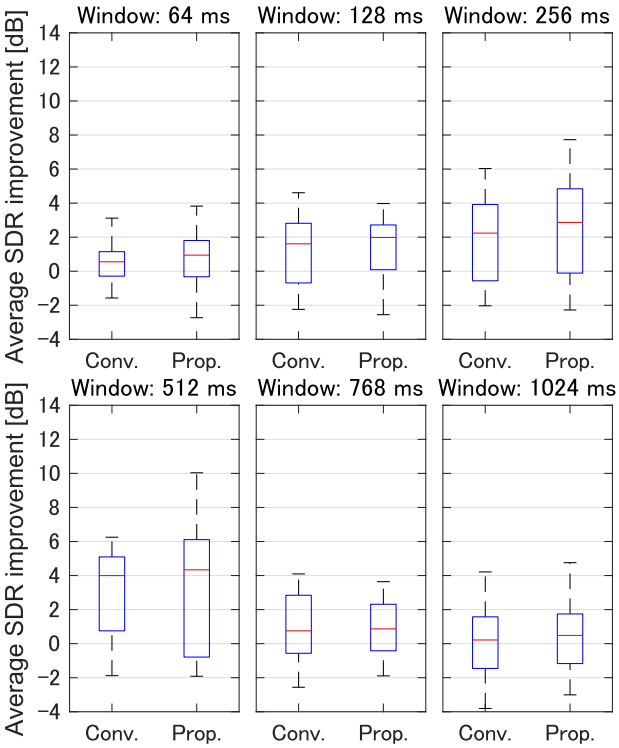


Fig. A.32. Average SDR improvements for synthesized music mixtures (speech 1–4) with JR2, where Blackman window is used with 1/2 length in STFT.