卒業研究論文

論 文 題 目

時間微分スペクトログラムに基づく ブラインド音源分離

提出年月日	令和５年　２月　２２日
学　　科	電 気 情 報 工 学 科
氏　　名	綾野 翔馬　　印
指導教員(主査)	北村 大地 講師　　印
副　　査	雛元 洋一 助教　　印
学　科　長	辻 正敏 教授　　印

香川高等専門学校

Blind Audio Source Separation Based on Time Differential of Spectrograms

Shoma Ayano

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

Abstract

Blind source separation (BSS) is a technique for estimating each original audio source from an observed signal, and various methods have been proposed so far. Many BSS methods are estimating demixing system that achieves source separation in time-frequency domain. This time-frequency expression is called complex-valued spectrogram and has complex-valued elements consisting of amplitude and phase. Nevertheless, in almost BSS methods including independent vector analysis and harmonic/percussive source separation estimate sources on the basis of a source model of amplitude spectrograms. In this thesis, to improve performance of BSS, we consider a source model of not only amplitude but also phase spectrograms. As a method that enables us to model both the amplitude and phase, phase-corrected spectrogram was proposed. In order to apply BSS methods using the phase-corrected spectrogram, complex-valued spectrogram of each separated source is required. In this thesis, as a fundamental research, we analyze characteristics of time differential of complex-valued spectrogram. Also, we consider about the method that converts time differential of complex-valued spectrogram to time-domain signal. In addition, we apply time differential of complex-valued spectrogram to conventional BSS methods and experimentally examine separation performance.

Keywords: blind source separation, phase spectrogram, time differential of complex-valued spectrogram

（和訳）

ブラインド音源分離（BSS）は，複数の音源が混合した観測信号から混合前の個々の音源を推定する技術であり，さまざまな手法が提案されている． 多くの BSS 手法では，時間周波数領域で音源分離を実現する分離系を推定している． この時間周波数領域の特徴量は複素スペクトログラムと呼ばれ，振幅と位相から成る複素数の成分を持っている． しかしながら，独立ベクトル分析や調波・打撃音分離を含むほとんどの BSS 手法では，複素スペクトログラムの振幅値のみをモデル化し，分離を行っている． 本論文では，BSS の性能を押し上げるために，振幅だけでなく位相も音源のモデル化に用いる手法について検討する． 振幅と位相の両方の構造をモデル化する一つの方法として，修正位相スペクトログラムが提案されている． 修正位相スペクトログラムに BSS を適用する場合，分離された各音源の時間微分複素スペクトログラムが必要となる． 本論文では，基礎的な検討として，まず時間微分複素スペクトログラムの特徴の解析と，時間領域の信号に変換する方法について検討する．
また，時間微分複素スペクトログラムを既存の BSS 手法に適用し，その分離性能について実験的に調査する．

目次

第 1 章

緒言

1.1　本論文の背景

音源分離とは，複数の音源信号が混合した状態で観測された信号から，混合前の音源を推定する技術である．具体的な利用例を Figs. 1.1 および 1.2 に示す．Fig. 1.1 では，観測された音声信号を目的の音声と背景雑音に分離し，背景雑音のみを除去することで目的の音声の強調を行っている．強調された音声は例えば音声認識の入力として用いられる．Fig. 1.2 では，観測された音楽信号を調波楽器音（ピアノやヴァイオリンなど弦楽器の音）と打楽器音（ドラムなど打楽器の音）に分離している．調波楽器音の高さを調べることで音楽のメロディーやコードなどを推定することができ，打楽器音の時間周期を調べることで音楽のテンポを推定することができる．

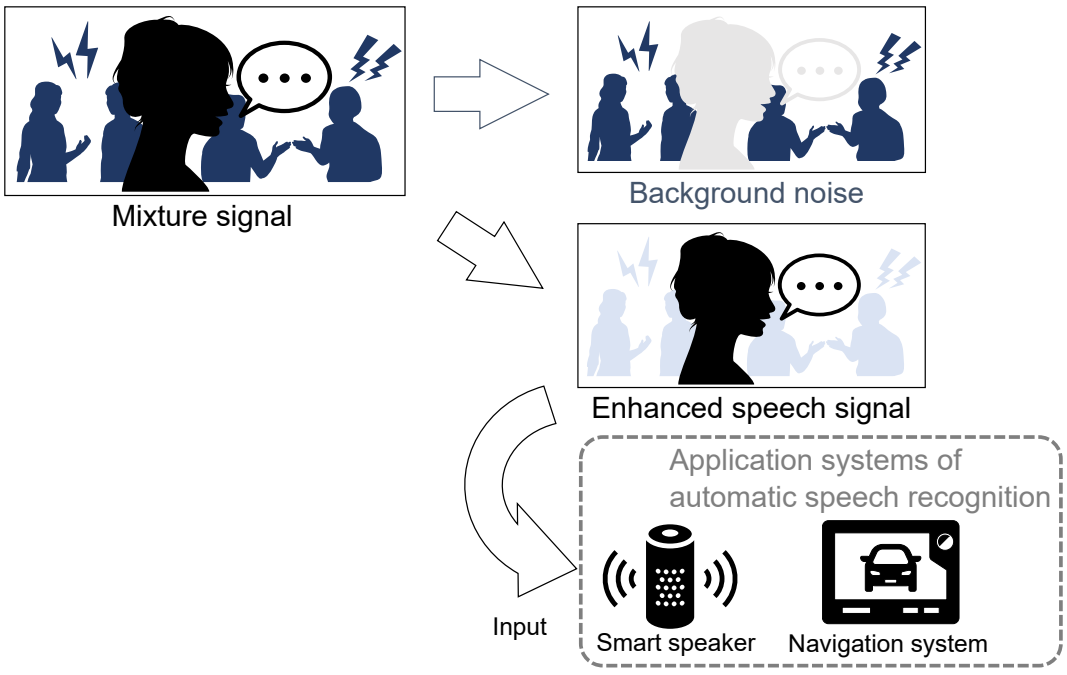


Fig. 1.1: Example of applications using speech source separation.

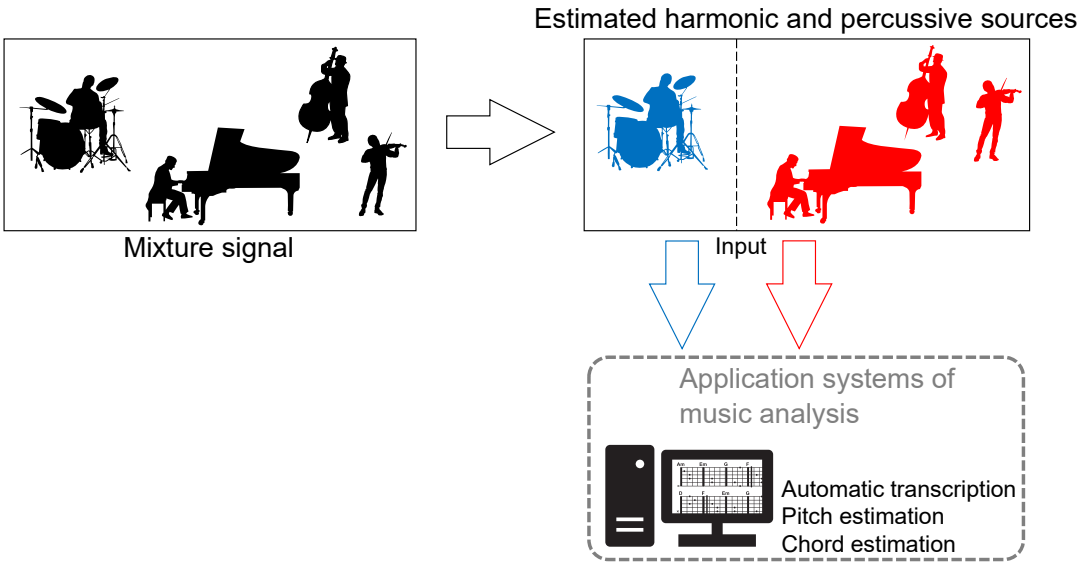


Fig. 1.2: Example of applications using music source separation.

音源分離技術の一例として，ブラインド音源分離（blind source separation: BSS）[1] が挙げられる．BSS の概要を Fig. 1.3 に示す．BSS は混合後の観測信号のみが与えられ，マイクや音源の位置，音源の種類やその学習データなどが与えられない条件で混合前の音源信号を推定する不良設定逆問題である．事前情報を一切必要としないため，実用化と応用が期待されている．

特に，観測信号のチャネル数（即ち，録音時のマイクロホン数）が混合している音源数以上となる観測信号を扱う BSS を優決定条件 BSS という．優決定条件 BSS では，音源間の統計的な独立性に基づく数理アルゴリズムである独立成分分析（independent component analysis: ICA）[2] の登場以降，盛んに研究されている．残響が生じているような観測信号においても BSS を適用できるように ICA を周波数毎に適用した周波数領域 ICA（frequency-domain ICA: FDICA）[3] が提案され，FDICA で推定される周波数毎の分離信号の順番を適切に並び替えるパーミュテーション問題の解決法が近年に至るまで検討されている [4, 5, 6, 7, 8, 9]．また，FDICA に対して音源の時間周波数構造を導入し，パーミュテーション問題を回避しながら分離信号を推定する独立ベクトル分析（independent vector analysis: IVA）[10, 11] が提案された．さらに，IVA で用いられた音源モデルを導入する手法を拡張し，非負値行列因子分解 [12] に基づく低ランク時間周波数構造を ICA の音源モデルに取り入れた独立低ランク行列分析（independent low-rank matrix analysis: ILRMA）[13, 14] が提案されている．このように，優決定条件 BSS では近年までにさまざまな手法が登場している．本論文では，これらの手法を独立性に基づく多チャネル BSS と呼ぶ．

独立性に基づく多チャネル BSS のうち，FDICA，IVA および ILRMA では短時間フーリエ変換（short-time Fourier transform: STFT）によって得られる時間周波数特徴量を用いている．この時間周波数特徴量は複素数の行列であり，複素スペクトログラムと呼ばれる．複素

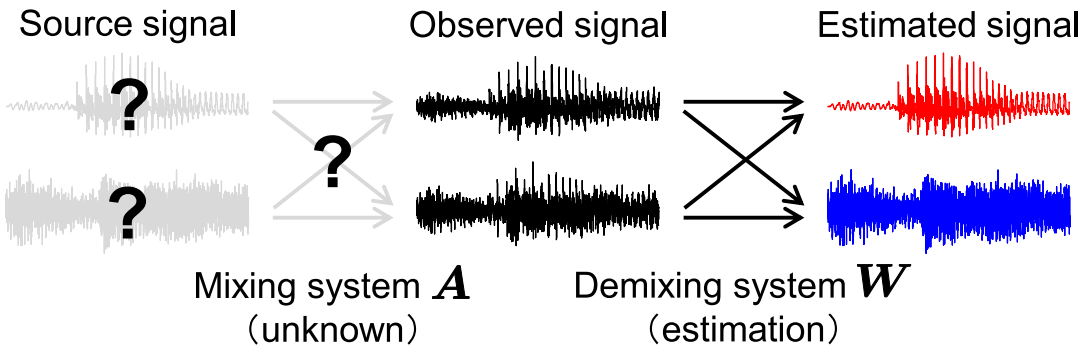


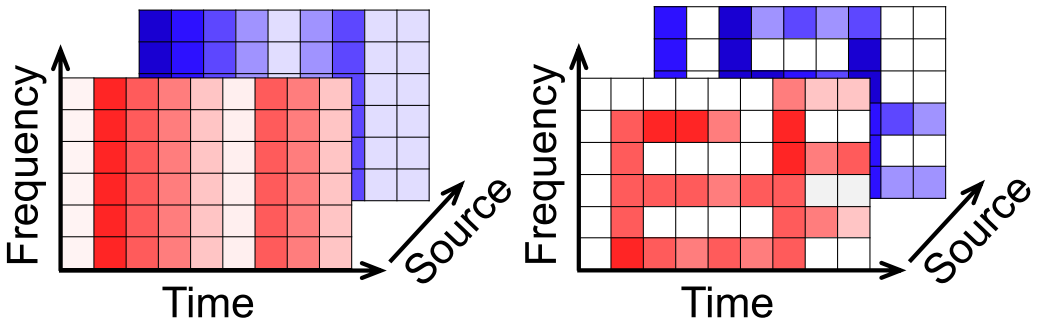
Fig. 1.3: Overview of BSS.

数の性質より複素スペクトログラムは振幅と位相に分けることができる．本論文では，複素スペクトログラムの各要素の振幅および位相をとった時間周波数行列をそれぞれ振幅スペクトログラムおよび位相スペクトログラムと表記する．さらに，振幅の 2 乗値を要素毎にとった時間周波数行列をパワースペクトログラムと表記する．

BSS では「音源モデル」と呼ばれる仮定をヒントに音源分離を達成する．この音源モデルとは，混合する前の個々の音源の時間周波数領域の構造や特徴を数式で表したものであり，不良設定逆問題である BSS を解くには必要不可欠である．一例として，IVA と ILRMA の音源モデルを Fig. 1.4 に，分離メカニズムの概要をそれぞれ Figs. 1.5 および 1.6 に示す．IVA では，「振幅スペクトログラムにおいて，同一音源の全周波数成分は連動して生起する傾向にある」という仮定をもとに分離を行っている．また ILRMA では，「パワースペクトログラムにおいて，同一音源の全周波数成分はよりサイズの小さい 2 つの行列の積で近似できる」という仮定をもとに分離を行っている．このように，IVA および ILRMA では，振幅又はパワースペクトログラムに対して混合前の音源の時間周波数構造仮定を置くことで分離を行っている．しかしながら，位相スペクトログラムに関しては，IVA や ILRMA やそのほかの BSS においても，これまで何も仮定されてこなかった．その理由として，次の 2 点が挙げられる．

1. 人間の聴覚は振幅スペクトログラムの変化には敏感であるが，位相スペクトログラムの変化はあまり認知できないこと
2. 位相は $-\pi$ から π rad の範囲で周期を持つように包み込まれており，数学的なモデルとして取り扱いづらいこと

そのため，独立性に基づく多チャネル BSS において位相スペクトログラムの音源モデルを陽に取り扱った例は極めて少なく，ほとんど未開拓の領域となっている．その一方で，振幅のみを用いた音源モデルに基づく BSS の性能の限界は問題となりつつある．例えば，録音時の残響時間が極端に長い場合，混合している音源の数が極端に多い場合，音源モデルが混合している音源などでは独立性に基づく多チャネル BSS の性能が極端に劣化することが知られている [15].



(a) IVA　　(b) ILRMA

Fig. 1.4: Source models assumed in IVA and ILRMA.

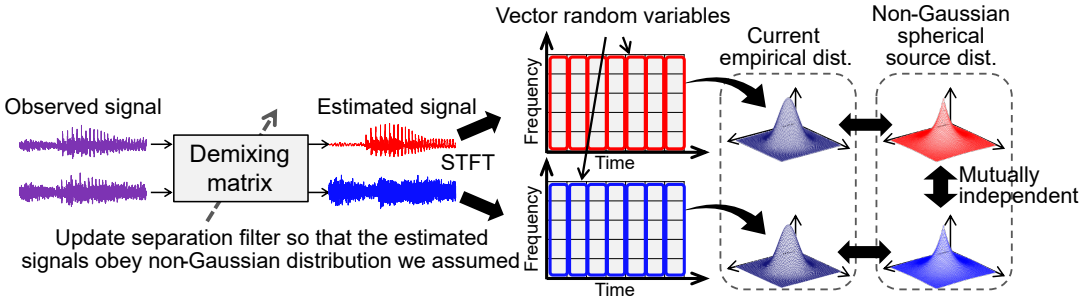


Fig. 1.5: Mechanism of IVA and its source model assumption.

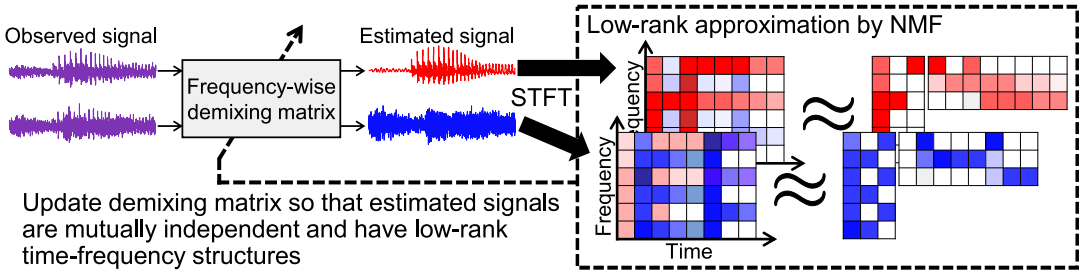


Fig. 1.6: Mechanism of ILRMA and its source model assumption.

1.2　本論文の目的

修正位相スペクトログラム [16] と呼ばれる時間周波数表現では，振幅が強く現れる部分では位相が連続となる性質が見られる．これを用いて，位相の時間周波数構造に関する音源モデルを考えることができる．つまり，これまで未開拓であった位相の時間周波数構造に関するモデルを取り扱うことができるようになる．しかし，修正位相スペクトログラムそのものを BSS する場合，分離された修正位相スペクトログラムを時間信号へと逆変換する際に分離された各音源の時間微分複素スペクトログラムが必要となる．したがって，修正位相スペクトログラム

を BSS に適用するためには時間微分複素スペクトログラムそのものを分離することが必要となる．

本論文では，位相に関する時間周波数特徴量を考慮した BSS の基礎的な検討を目的として，複素スペクトログラムの時間微分である時間微分複素スペクトログラムという特徴量を BSS に適用することを提案する．一般に，時間微分複素スペクトログラムは，微分窓関数（通常の窓関数の導関数）を用いた STFT を音響信号に適用することで得ることができる．しかし，音響信号処理の分野で一般的に用いられる STFT の条件では，時間微分複素スペクトログラムから音源信号への逆変換が不可能である．この問題を STFT の理論から説明し，新たに逆変換可能な時間微分複素スペクトログラムの計算方法を説明する．実験では，逆変換可能な時間微分複素スペクトログラムを，前述の IVA と，音楽の BSS で良く用いられる調波打撃音分離（harmonic/percussive source separation: HPSS）の 2 手法に適用し，性能を実験的に調査する．得られた実験結果から，複素スペクトログラムと時間微分複素スペクトログラムの音源分離の性能の差を比較し，時間微分複素スペクトログラムの特徴を調査する．

1.3 本論文の構成

2 章では，提案手法を説明する上で重要な音響信号の時間周波数領域への変換，および音源分離技術である独立性に基づく多チャネル BSS および HPSS の理論について述べる．3 章では，微分窓関数を用いた STFT を適用する提案手法の動機および詳細について述べる．具体的には，複素スペクトログラムと時間微分複素スペクトログラムの比較および音響信号への逆変換の注意点について述べる．4 章では，複素スペクトログラムおよび時間微分複素スペクトログラムに対して音源分離を行い，各性能指標を比較し，微分窓関数が音源分離に与える影響について述べる．最後に 5 章で本論文の結果についてまとめ，今後の課題を述べる．

第 2 章

基礎理論

2.1 まえがき

本章では，音響信号処理の時間周波数特徴量および BSS に関する基礎理論について説明する．2.2 節では，音響信号を時間周波数領域で表現するためのさまざまな手法について述べる．具体的には，STFT やその他のさまざまな時間周波数表現の特徴を説明し，どのような分野で用いられるかを紹介する．2.3 節では，独立性に基づく多チャネル BSS の理論について説明する．独立性に基づく多チャネル BSS である ICA，FDICA，および IVA について定式化する．2.4 節では，HPSS の理論について説明する．HPSS の概要を説明し，最適化に基づく HPSS およびメディアンフィルタに基づく HPSS を定式化する．最後に，2.5 節で本章をまとめる．

2.2 音響信号の時間周波数領域での表現

2.2.1 さまざまな変換

音響信号を時間周波数領域で表現することは，時間領域の音響信号を捉える上で非常に有効な手段である．一例として，振幅スペクトログラムを見ることで特定の音高の成分がどの程度含まれているかを知ることができる．音響信号を時間周波数領域で表現する方法として，STFT，修正離散コサイン変換（modified discrete cosine transform: MDCT），定 Q 変換（constant-Q transform: CQT），メル周波数スペクトログラム，およびメル周波数ケプストラム係数（Mel-frequency cepstrum coefficient: MFCC）の時系列などが挙げられる．CQT を除くこれらの手法の共通点として，Fig. 2.1 に示すように一定時間のフレーム毎に信号を切り出し，各時間フレームで時間領域から周波数領域の変換を行った後に変換後のフレームを時間方向に並べている．

STFT は，音響信号処理において最も利用されている時間周波数領域での表現であり，これについては 2.2.2 項で詳しく説明する．信号に対して STFT を適用すると複素数の行列が現れる．Fig. 2.2 に示す時間波形に STFT を適用した振幅スペクトログラムを Fig. 2.3 に示す．Fig. 2.2 は 0〜1 s で 440Hz の正弦波，1.5〜2.5 s で 220 Hz，440 Hz および 880 Hz の正弦

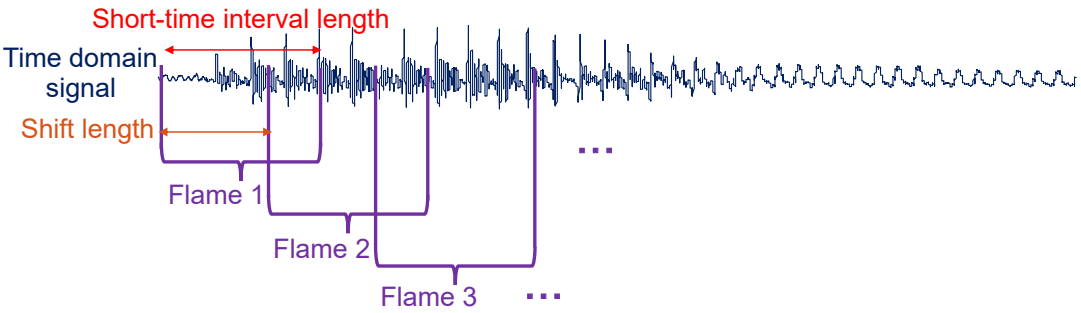


Fig. 2.1: Windowing (splitting) time-domain signal into short-time frame signals.

波，3 s にドラムのスネア音，4.5 s〜7 s にピアノの A4 音を含む時間波形である．Fig. 2.3 は線形周波数軸と線形時間軸を持ち，正弦波やピアノ音などの調波構造を持つ音に対しては横筋状の構造，ドラムのスネア音のような打撃音に対しては縦筋状の構造が見られる．ただし，この図は振幅スペクトログラムだけを表示しており，実際は時間周波数毎の波形の時間シフトを表す位相スペクトログラムも同時に存在する点に注意されたい．位相スペクトログラムについては 2.2.3 項で詳しく確認する．

MDCT は，信号の末尾にそのフレームの信号を時間反転した信号を付け加えてフーリエ変換を行う変換である．このような信号は偶関数となるため，フーリエ変換の基底関数を複素正弦波ではなく余弦波のみとして直交基底分解できるようになる．したがって，時間信号に MDCT を適用すると実数の行列が現れる．Fig. 2.2 に示す時間波形に MDCT を適用した行列を Fig. 2.4 に示す．Fig. 2.4 は線形周波数軸と線形時間軸を持ち，振幅スペクトログラムと比較すると定常な音響信号でも時間周波数成分の強度が時間的に細かく変動していることが分かる．これは，MDCT が STFT における位相（時間シフト）を含む実数特徴量に変換していることに由来する．これらの MDCT の性質を利用して，深層ニューラルネットワークと組み合わせた信号処理 [17] や情報符号化のアルゴリズム [18] が提案されている．

CQT は，Q 値と呼ばれる周波数毎の局在波の広がりを表す定数を全周波数で一定としたウェーブレット変換である．信号に対して CQT を適用すると複素数の行列が現れる．Fig. 2.2 に示す時間波形に CQT を適用した行列を Fig. 2.5 に示す．Fig. 2.5 は対数周波数軸と線形時間軸を持ち，正弦波やピアノ音に対しては横線の構造が見られる．音の周波数が 4 倍になると人間が感じる音高は約 3 倍（2 オクターブ上の音）になることから，人間が感じる音高は周波数の対数にある程度比例する．このような性質から，CQT は音楽音響処理の分野でよく利用される [19]．

極端に高いまたは低い周波数では，人間が感じる音高は周波数の対数に比例しないことが知られており，その感覚的なずれを補正した尺度をメル周波数と呼ぶ．メル周波数スペクトログラムは，メル周波数に基づいた音響特徴量であり，複素スペクトログラムの絶対値 2 乗であるパワースペクトログラムに対してメルフィルタバンクと呼ばれる複数の周波数フィルタを適用して得られる複素数の行列である．Fig. 2.3 に示す振幅スペクトログラムにメルフィルタバンクを適用したメル周波数スペクトログラムを Fig. 2.6 に示す．MFCC は上記メル周波数スペ

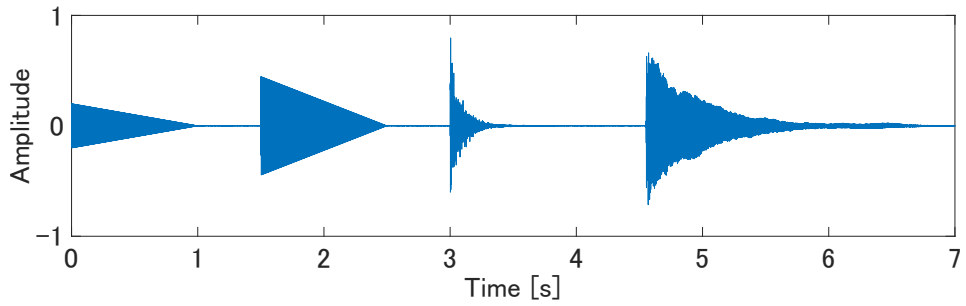


Fig. 2.2: Time-domain signal.

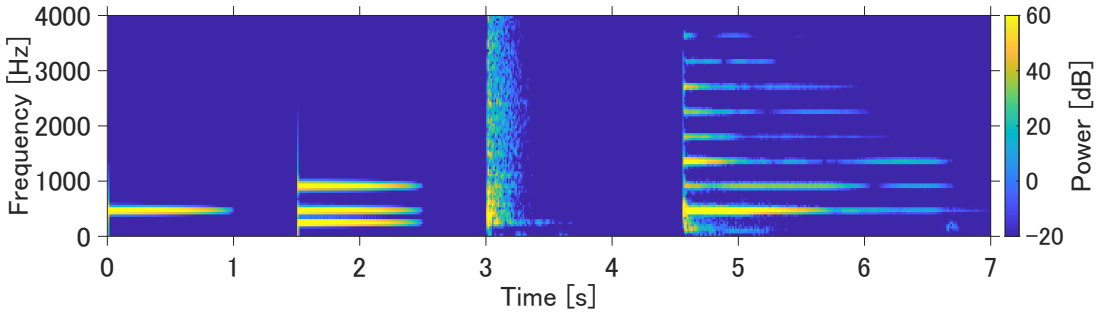


Fig. 2.3: Amplitude spectrogram obtained by STFT.

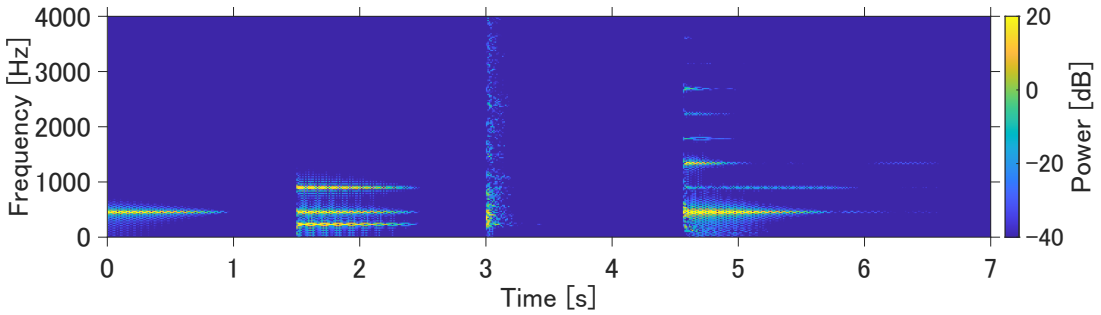


Fig. 2.4: Time-frequency matrix obtained by MDCT.

クトログラムの各時間フレームに対して DCT を行ったもので，実数の行列となる．Fig. 2.6 に示すメル周波数スペクトログラムの各時間フレームに DCT を適用した行列を Fig. 2.7 に示す．MFCC は音響信号の音高や音量に依らない音色成分（スペクトル包絡の概形）をよく表している．メル周波数スペクトログラムおよび MFCC は音声認識でよく用いられる [20] が，音楽信号に対する適用例も見られる [21, 22].

2.2.2　STFT

STFT は，2.2.1 項で述べたように，時間領域の信号から時間的に変化する複素スペクトルを得る手法である．STFT の概要を Fig. 2.8 に示す．ただし，図中の行列に対する絶対値記号

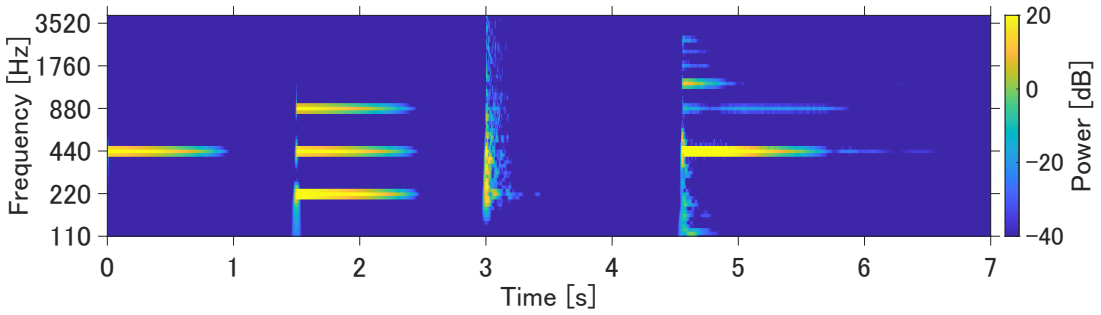


Fig. 2.5: Time-scale (frequency) matrix obtained by CQT.

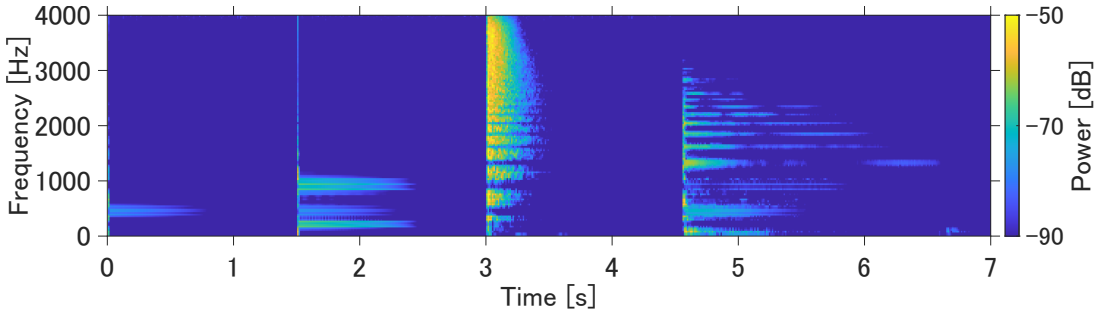


Fig. 2.6: Mel spectrogram.

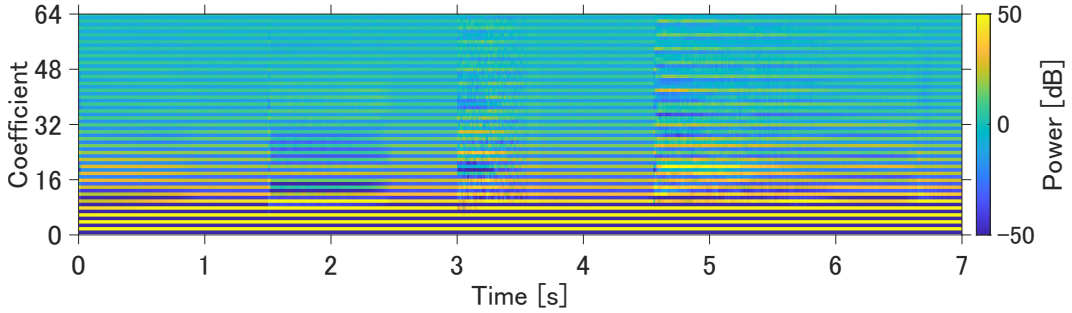


Fig. 2.7: Time-series MFCC obtained by applying DCT to mel spectrogram.

$|\cdot|$ は要素ごとの絶対値をとった行列を表し，また行列に対する偏角記号 $\mathrm{Arg}(\cdot)$ は要素ごとの偏角（ただし，$-\pi$ から π rad）をとった行列を表す．一定時間ごとに時間波形を切り出す窓関数を乗算し，得られた時間波形を離散フーリエ変換によって周波数表現へと変換する．得られた周波数表現を時間方向に並べることで複素スペクトログラムを得ることができる．

信号長 L の信号 $\boldsymbol{z} = [z[1], z[2], \cdots, z[l], \cdots, z[L]]^{\mathrm{T}} \in \mathbb{R}^L$ の STFT を考える．ここで，$l = 1, 2, \cdots, L$ は離散時間インデクスである．STFT において，時間領域から周波数領域への変換時の窓関数の長さおよびシフト長をそれぞれ Q および τ とする（Fig. 2.8 参照）．時間フレーム数 J は次式で求められる．

$$J = \frac{L}{\tau} \tag{2.1}$$

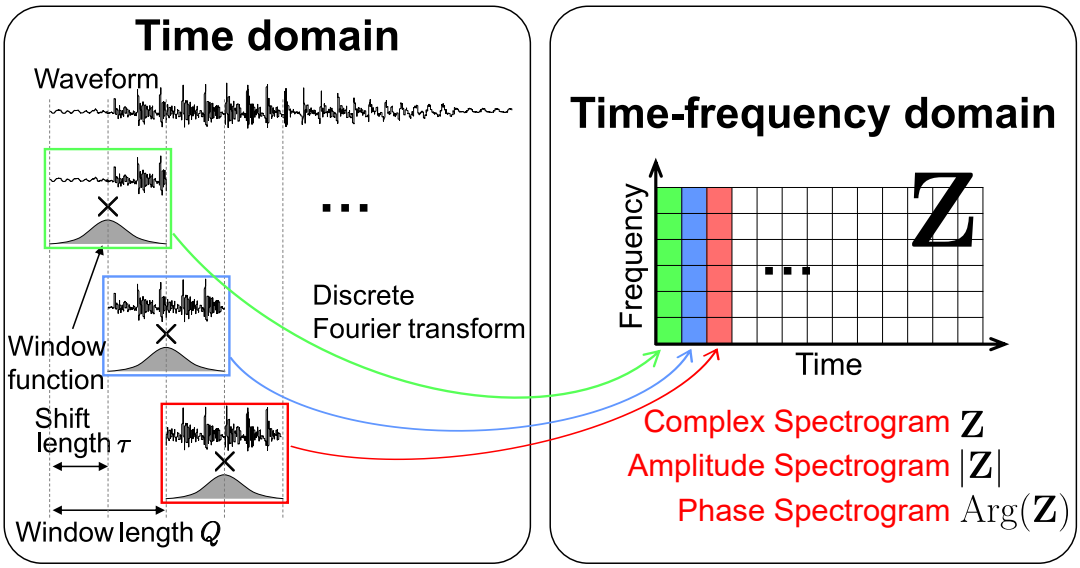


Fig. 2.8: Mechanism of STFT.

時間フレーム数 J が整数となるように信号 $\boldsymbol{z}$ の終端にゼロを挿入する処理（ゼロパディング）が施されている（ゼロパディング後の信号長を改めて L と解釈する）．周波数ビンの数 I は $I = \lfloor \frac{Q}{2} \rfloor + 1$ を満たす整数（$\lfloor \cdot \rfloor$ は床関数）となる．STFT によって得られる行列 $\mathbf{Z} \in \mathbb{C}^{I \times J}$ の (i, j) 番目の要素は次式で表される．

$$\mathrm{z}_{ij} = \sum_{q} z[q + \tau(j-1)]\omega_a[q]\exp\left\{-2\pi\iota\frac{(q-1)(i-1)}{Q}\right\} \tag{2.2}$$

ここで，ι は虚数単位を示している．また，$\boldsymbol{\omega}_a = [\omega_a[1], \omega_a[2], \cdots, \omega_a[Q]]^{\mathrm{T}} \in \mathbb{R}^Q$ は STFT で用いる窓関数であり，$\omega_a[q]$ は $1 \leq q \leq Q$ 以外では 0 をとるように定義する．このように，時間領域の信号は一定幅の短時間毎に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 $\mathbf{Z}$ で表すことができる．また，$\mathbf{Z}$ は J 個の各時間フレームに対して Q 個のサンプルをもとに計算されているため，$J \times Q$ 個のサンプルの情報を含んでいる．もとの信号が含む情報と変換によって得られる信号が含む情報の比は冗長度と呼ばれ，STFT は冗長度 $J \times Q/L = Q/\tau$ の変換である．

2.2.3　修正位相スペクトログラム

本節では，STFT で得られる複素スペクトログラムの位相を，より時間周波数構造が見えるように変換する方法として提案された，修正位相スペクトログラム [16] を説明する．修正位相スペクトログラムは，STFT によって得られた複素スペクトログラムの各時間周波数の位相を適切に回転することによって得られる複素スペクトログラムである．したがって，振幅に関しては通常の（STFT で得られる）スペクトログラムと修正位相スペクトログラムで共通であるが，位相に関しては異なる．修正位相スペクトログラムには複素数の観点で低ランクな構造

が現れることが示されており [23]，振幅と位相の両方を用いた信号処理が提案されている．そのため，位相を用いた音源モデルを考える上で有用な時間周波数表現である．

式 (2.2) とは異なる STFT の定義を次式に示す．

$$\mathrm{z}_{ij}^{(\mathrm{org})} = \sum_{q} z[q]\omega_a[q-\tau(j-1)]\exp\left\{-2\pi\iota\frac{(q-1)(i-1)}{Q}\right\} \tag{2.3}$$

ここで，$\mathbf{Z}^{(\mathrm{org})} \in \mathbb{C}^{I\times J}$ は式 (2.3) によって計算される複素スペクトログラムである．式 (2.2) は窓関数と基底のインデクスが同じになっているが，式 (2.3) は時間信号と基底のインデクスが同じになっている．まず，複素スペクトログラム $\mathbf{Z}^{(\mathrm{org})}$ の時間微分である時間微分複素スペクトログラム $\mathbf{Z}^{(\mathrm{diff})} \in \mathbb{C}^{I\times J}$ を定義する．この $\mathbf{Z}^{(\mathrm{diff})}$ は，窓関数 $\omega_a[q]$ を時間 q に関して微分した導関数（以後，微分窓関数と呼ぶ）$\omega_a^{(\mathrm{diff})}[q]$ を用いて，次式で計算される．

$$\mathrm{z}_{ij}^{(\mathrm{diff})} = -\sum_{q} z[q]\omega_a^{(\mathrm{diff})}[q-\tau(j-1)]\exp\left\{-2\pi\iota\frac{(q-1)(i-1)}{Q}\right\} \tag{2.4}$$

各時間周波数における位相の回転量は瞬時周波数 $\boldsymbol{\phi} \in \mathbb{R}^{I\times J}$ を用いて計算される．瞬時周波数 $\boldsymbol{\phi}$ は次式で表される．

$$\varphi_{ij} = \mathrm{Im}\left[\frac{\mathrm{z}_{ij}^{(\mathrm{diff})}}{\mathrm{z}_{ij}^{(\mathrm{org})}}\right] \tag{2.5}$$

ここで，$\mathrm{Im}\,[\cdot]$ は複素数の虚部のみを返す関数である．複素スペクトログラム $\mathbf{Z}$ の (i,j) 番目の要素での位相の回転量 θ_{ij} は次式で表される．

$$\theta_{ij} = -\frac{2\pi\tau}{Q}\sum_{q=1}^{j}\varphi_{iq} \tag{2.6}$$

これを用いて，修正位相スペクトログラム $\mathbf{Z}^{(\mathrm{IPC})} \in \mathbb{R}^{I\times J}$ は次式で求められる．

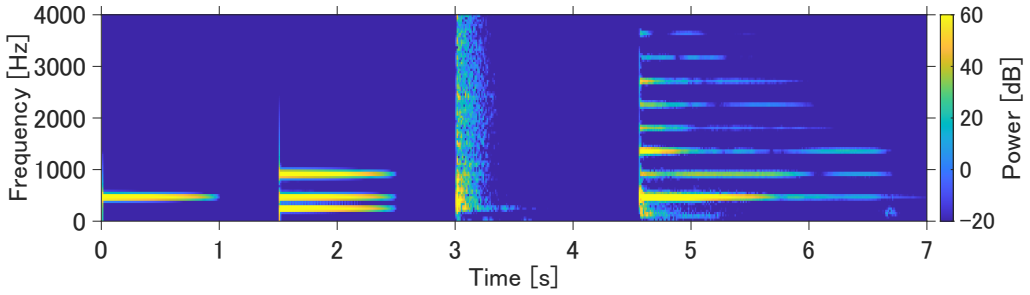
$$\mathrm{z}_{ij}^{(\mathrm{IPC})} = \mathrm{z}_{ij}^{(\mathrm{org})}\exp\left(\iota\theta_{ij}\right) \tag{2.7}$$

Fig. 2.2 の時間信号に対して振幅スペクトログラム，式 (2.2) で計算される位相スペクトログラムおよび修正位相スペクトログラムの位相を Fig. 2.9 に示す．位相スペクトログラムに対し，修正位相スペクトログラムの位相は正弦波やピアノ音が含まれる部分で横線の構造が見られる．

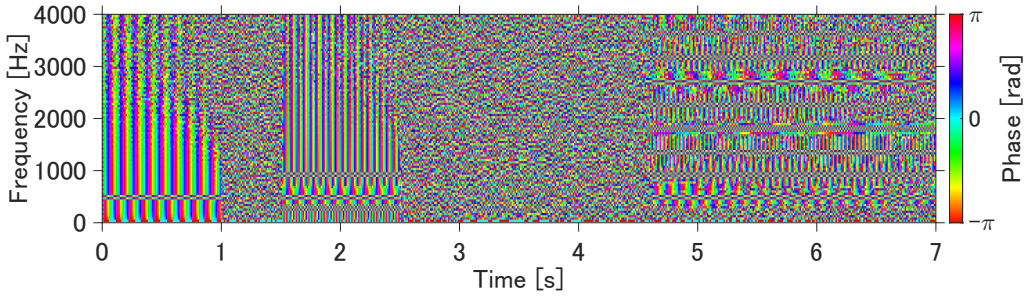
2.3 独立性に基づく多チャネル BSS

2.3.1 ICA

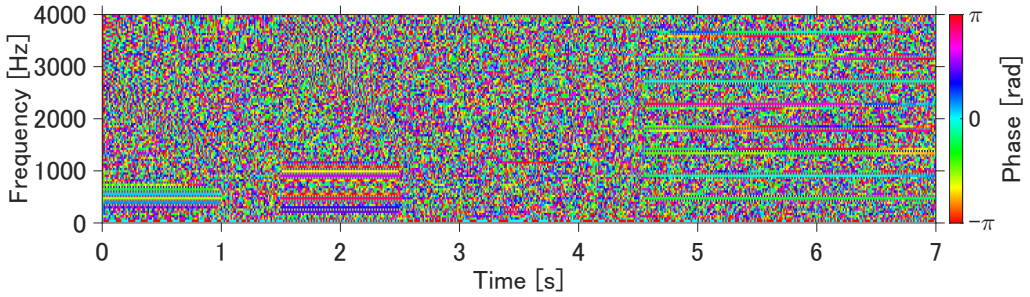
独立性に基づく多チャネル BSS の基礎的な数理理論である ICA [2] は，時間領域での瞬時混合および音源信号間の統計的性質を仮定した手法である．ここでは N を音源信号数，M を



(a) Amplitude spectrogram.



(b) Phase spectrogram.



(c) Phase of instantaneous phase corrected spectrogram.

Fig. 2.9: Amplitude spectrogram and phase spectrograms.

観測チャネル数（ただし $M = N$）として説明する．音源信号 $\boldsymbol{s}[l] = \left[s_1[l], s_2[l], \cdots, s_N[l]\right]^{\mathrm{T}}$ の瞬時混合は次式で表される．

$$\begin{cases} x_1[l] = a_{11}s_1[l] + a_{12}s_2[l] + \cdots + a_{1N}s_N[l] \\ x_2[l] = a_{21}s_1[l] + a_{22}s_2[l] + \cdots + a_{2N}s_N[l] \\ \vdots \\ x_M[l] = a_{M1}s_1[l] + a_{M2}s_2[l] + \cdots + a_{MN}s_N[l] \end{cases} \tag{2.8}$$

ここで，$x_m[l]$ は m 番目の観測信号である．多チャネルの観測信号を $\boldsymbol{x}[l] = [x_1[l], x_2[l], \cdots, x_M[l]]^{\mathrm{T}} \in \mathbb{R}^M$，$a_{mn}$ を要素に持つ行列を $\boldsymbol{A} \in \mathbb{R}^{M\times N}$ とすると，式

(2.8) は次式で表される．

$$\boldsymbol{x}[l] = \boldsymbol{A}\boldsymbol{s}[l] \tag{2.9}$$

もし $\boldsymbol{A}$ が正則ならば，$\boldsymbol{W} = \boldsymbol{A}^{-1} \in \mathbb{R}^{N\times N}$ を推定することで観測信号 $\boldsymbol{x}[l]$ から音源信号 $\boldsymbol{s}[l]$ を求めることができる．この $\boldsymbol{W}$ を求めることが ICA の目的である．

$\boldsymbol{W}$ を求めるために，ICA では以下の統計的性質を仮定する．

1. 各分離信号は統計的に独立
2. 各分離信号は非ガウス分布に従う

これらの仮定より，分離信号ができるだけ独立となる分離行列 $\boldsymbol{W}$ を求めるための ICA の最適化問題は次式で表される．

$$\min_{\boldsymbol{W}} \mathcal{J}(\boldsymbol{W}) \tag{2.10}$$

ここで，最適化のコスト関数 $\mathcal{J}(\boldsymbol{W})$ は次式で定義される．

$$\mathcal{J}(\boldsymbol{W}) = \int_{-\infty}^{\infty} p(\boldsymbol{y}) \log \frac{p(\boldsymbol{y})}{\prod_{n=1}^{N} p(y_n)} d\boldsymbol{y} \tag{2.11}$$

また，$p(y_n)$ は n 番目の分離信号 $y_n[l]$ の生成モデルであり，$p(\boldsymbol{y})$ は同時分布である．

前述の最適化問題を用いて分離行列を推定した場合，信号の大きさや順序を決定できないという問題があり，ICA によって推定される分離行列 $\hat{\boldsymbol{W}} \in \mathbb{R}^{N\times N}$ には，以下の任意性が存在する．

1. 分離信号のチャネルの順序の任意性
2. 分離信号のスケールの任意性

これらの任意性は分離信号に対して Fig. 2.10 のように現れる．求めたい $\boldsymbol{W}$ と推定した $\hat{\boldsymbol{W}}$ の関係は次式となる．

$$\hat{\boldsymbol{W}} = \boldsymbol{D}\boldsymbol{\Pi}\boldsymbol{W} \tag{2.12}$$

ここで，$\boldsymbol{D} \in \mathbb{R}^{N\times N}$ は信号のスケールを変える対角行列，$\boldsymbol{\Pi} \in \mathbb{R}^{N\times N}$ は信号の順序を変える置換行列（パーミュテーション行列）である．上記の任意性 1 は $\boldsymbol{\Pi}$ に現れ，任意性 2 は $\boldsymbol{D}$ に現れる．任意性 1 に関しては，$\boldsymbol{D}^{-1}$ を解析的に求めるプロジェクションバック（projection back: PB）法 [24] と呼ばれる補正手法が提案されており，次式で計算される．

$$\hat{y}_n[l] = \hat{\boldsymbol{W}}^{-1}(\boldsymbol{e}_n \odot \boldsymbol{y}[l]) \tag{2.13}$$

ここで，$\boldsymbol{e}_n \in \mathbb{R}^N$ は n 番目の要素が 1，その他の要素が 0 のベクトルであり，$\odot$ はアダマール積を示す．式 (2.13) で $\hat{\boldsymbol{W}}$ に含まれる $\boldsymbol{D}$ の影響のみを打ち消すような計算を行っている．

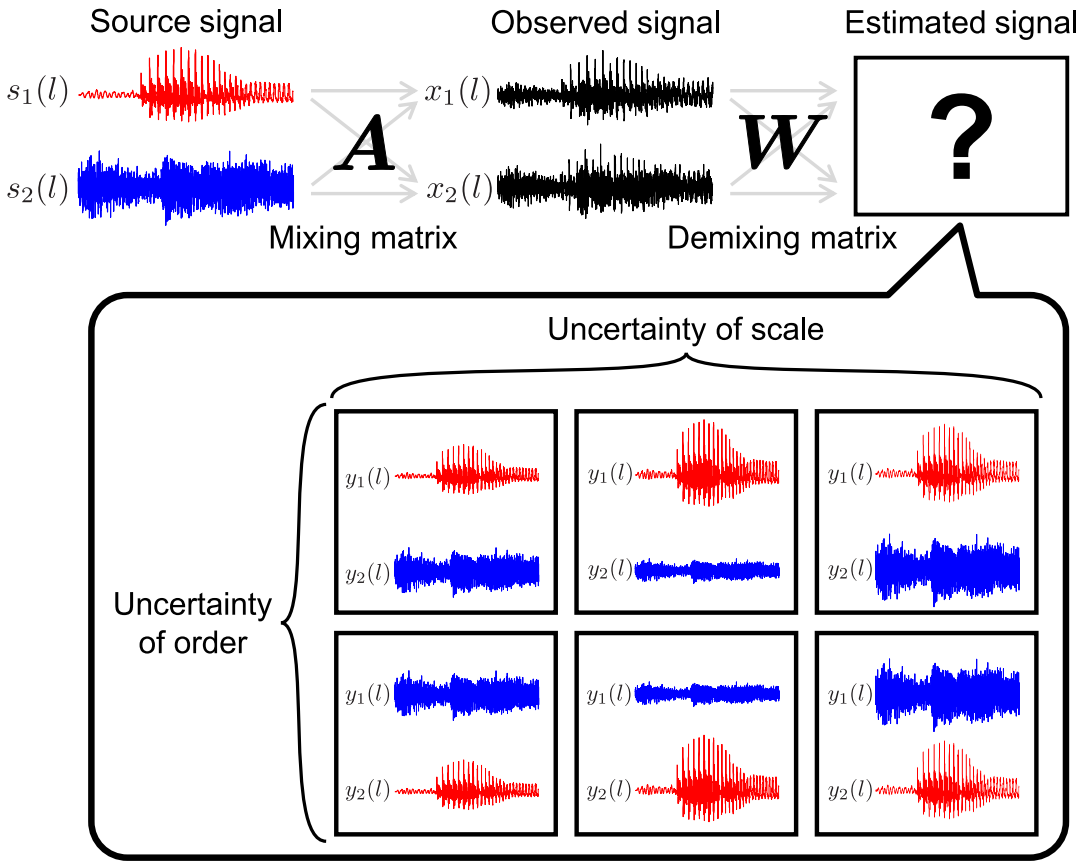


Fig. 2.10: Uncertainty in ICA.

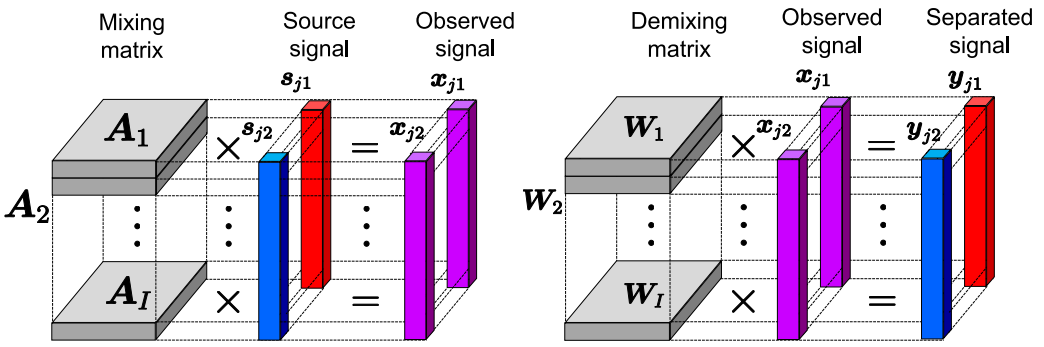
2.3.2　FDICA

残響を含む信号に対する音源分離は，畳み込み混合系の逆系を推定することで達成できる．しかし，時間領域における畳み込み混合系の逆系の推定は困難である．2.2.2 項で述べた STFT では，畳み込みは積和に変換される．つまり，時間領域での畳み込み混合を時間周波数領域での瞬時混合として扱うことが可能となる．そこで，観測信号を STFT して得られた複素スペクトログラムの周波数ビン i での複素時系列信号に対して，独立な ICA を適用し，分離信号の複素スペクトログラムを推定する手法である FDICA [3] が提案された．この手法では，次式のような時間周波数領域における混合および分離モデルを仮定する．

$$\mathrm{x}_{ijn} = \mathbf{A}_i \mathrm{s}_{ijn} \tag{2.14}$$

$$\mathrm{y}_{ijn} = \mathbf{W}_i \mathrm{x}_{ijn} \tag{2.15}$$

ここで，$\mathbf{S} \in \mathbb{C}^{I \times J \times N}$ は音源信号の複素スペクトログラム，$\mathbf{X} \in \mathbb{C}^{I \times J \times N}$ は混合信号の複素スペクトログラム，$\mathbf{Y} \in \mathbb{C}^{I \times J \times N}$ は分離信号の複素スペクトログラムを表し，$\mathbf{A}_i \in \mathbb{C}^{N \times N}$ および $\mathbf{W}_i \in \mathbb{C}^{N \times N}$ は i 番目の周波数ビンに対する混合行列および分離行列である．式 (2.14) および式 (2.15) を図示したものが Fig. 2.11 である．



(a) Mixing model (b) Demixing model

Fig. 2.11: Mixing and demixing models assumed in FDICA.

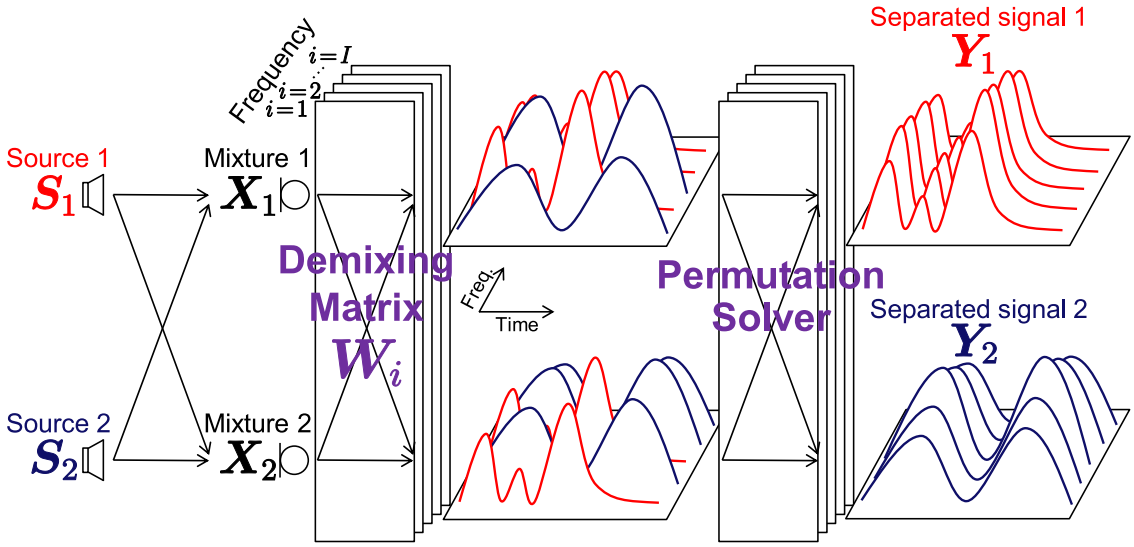


Fig. 2.12: Permutation problem in FDICA.

2.3.1 項で述べたように，ICA の分離信号には順序およびスケールの任意性がある．FDICA では周波数毎に独立な ICA が適用されるため，各周波数ビンで推定した分離信号の複素スペクトログラムの並びとスケールがバラバラになるという問題が生じる．周波数毎のスケールの任意性については，式 (2.13) の PB 法により解析的に復元可能である．一方で，Fig. 2.12 に示すような周波数毎のパーミュテーション行列の問題を解決することは困難である．この問題はパーミュテーション問題と呼ばれており，これを解決することが FDICA における大きな課題である．

2.3.3 IVA

2.3.2 項では，FDICA にはパーミュテーション問題という大きな問題が存在することを述べた．それに対して，IVA [10] では，Fig. 1.4 (a) に示すように FDICA の音源モデルに「同

一音源の全周波数成分は連動して生起する傾向にある」という仮定を導入し，全周波数成分をまとめてベクトル変数とすることでベクトル間の独立性を最大化するようなモデルとなっている．したがって，実際に複数の周波数ビンで同時に共起する成分が同一音源としてまとめられるような分離行列が推定される．

IVA の最適化問題は次式のように表される．

$$\min_{\boldsymbol{W}_1,\cdots,\boldsymbol{W}_I} -2J\sum_i \log|\det \boldsymbol{W}_i| + \sum_{j,n}\sqrt{\sum_i |y_{ijn}|^2} \tag{2.16}$$

ここで，$\boldsymbol{W}_i = [\boldsymbol{w}_{i1}^{\mathrm{H}},\cdots,\boldsymbol{w}_{in}^{\mathrm{H}},\cdots,\boldsymbol{w}_{iN}^{\mathrm{H}}]^{\mathrm{H}} \in \mathbb{C}^{N\times N}$ は i 番目の周波数ビンに対する分離行列であり，$y_{ijn} = \boldsymbol{w}_{in}^{\mathrm{H}}\boldsymbol{x}_{ij}$ は n 番目の推定信号の複素スペクトログラム $\boldsymbol{Y}_n \in \mathbb{C}^{I\times J}$ の (i,j) 成分である（y_{ijn} が最適化変数 $\boldsymbol{W}_i$ の行ベクトルを含む点に注意）．式 (2.16) の解を求める方法として，補助関数 IVA（auxiliary-function-based IVA: AuxIVA）[11] を用いた反復音源ステアリング法（iterative source steering: ISS）[31] と呼ばれる数値的に安定な分離行列の更新式が提案されている．次式に更新式を示す．

$$r_{kj} \leftarrow \sqrt{\sum_i |y_{kij}|^2},\ \forall k,j \tag{2.17}$$

$$u_m \leftarrow \frac{1}{2Jr_{mj}}\sum_j y_{mij}y_{kij}^*,\ \forall m\neq k \tag{2.18}$$

$$d_m \leftarrow \frac{1}{2Jr_{mj}}\sum_j |y_{kij}|^2,\ \forall m \tag{2.19}$$

$$v_{mk} \leftarrow \frac{u_m}{d_m},\ \forall m\neq k \tag{2.20}$$

$$v_{kk} \leftarrow 1 - d_k^{-\frac{1}{2}} \tag{2.21}$$

$$\boldsymbol{y}_{ij} \leftarrow \boldsymbol{y}_{ij} - \boldsymbol{v}_k y_{kij} \tag{2.22}$$

上記の更新式に対してすべての $k = 1,2,\cdots,M$ および $i = 1,2,\cdots,I$ について計算することを 1 回の更新とし，複数回反復計算を行うことで IVA の局所解を求めることができる．

2.4　HPSS

HPSS は調波楽器および打楽器の音源が持つ，対極的な振幅スペクトログラムの構造を仮定とした音源モデルに基づいて，次式のように，混合信号を調波音および打撃音に分離する手法である．

$$\boldsymbol{B} = \boldsymbol{H} + \boldsymbol{P} \tag{2.23}$$

ここで，$\boldsymbol{B} \in \mathbb{C}^{I\times J}$，$\boldsymbol{H} \in \mathbb{C}^{I\times J}$，および $\boldsymbol{P} \in \mathbb{C}^{I\times J}$ は，それぞれモノラルの混合信号の複素スペクトログラム，分離された調波信号の複素スペクトログラム，および分離された打撃信号の複素スペクトログラムである．

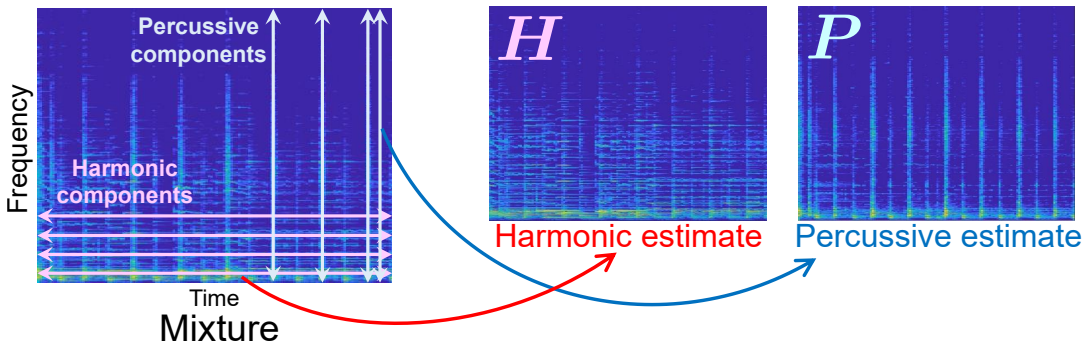


Fig. 2.13: Separation principle of HPSS.

HPSS の概要を Fig. 2.13 に示す．HPSS では，調波音は時間方向に連続であり，打撃音は非定常的でかつ周波数方向に連続である，という振幅スペクトログラム構造の特徴に着目している．本節では，代表的な 2 種類の HPSS である最適化に基づく HPSS（optimization-based HPSS: OHPSS）[25] とメディアンフィルタに基づく HPSS（median-filter-based HPSS: MHPSS）[26] を説明する．

2.4.1 OHPSS

OHPSS では，前節で説明した調波音 $\boldsymbol{H}$ および打撃音 $\boldsymbol{P}$ の構造の違いを目的関数の最適化問題とし，これを解くことで $\boldsymbol{H}$ および $\boldsymbol{P}$ を推定する．文献 [25] に示す HPSS では，混合された複素スペクトログラム $\boldsymbol{B}$ から $\boldsymbol{H}$ および $\boldsymbol{P}$ を推定するために，次式の最小化問題の解として $\boldsymbol{H}$ および $\boldsymbol{P}$ を推定する．

$$\min_{\boldsymbol{H},\boldsymbol{P}} \mathcal{J}(\boldsymbol{H},\boldsymbol{P}) \ \text{ s.t. } |b_{ij}| = |h_{ij}| + |p_{ij}| \ \forall\, i,j,\ \mathrm{Arg}(b_{ij}) = \mathrm{Arg}(h_{ij}) = \mathrm{Arg}(p_{ij}) \ \forall\, i,j \tag{2.24}$$

ここで，最適化のコスト関数 $\mathcal{J}(\boldsymbol{H},\boldsymbol{P})$ は次式で定義される．

$$\mathcal{J}(\boldsymbol{H},\boldsymbol{P}) = \sum_{i,j} \left\{ \left(|h_{i(j+1)}|^C - |h_{ij}|^C\right)^2 + \left(|p_{(i+1)j}|^C - |p_{ij}|^C\right)^2 \right\} \tag{2.25}$$

また，h_{ij} および p_{ij} はそれぞれ $\boldsymbol{H}$ および $\boldsymbol{P}$ の (i,j) 番目の要素であり，C はドメインを指定するパラメータである．計算の簡単化のため，本論文では以下 $C = 0.5$ とする．最小化問題 (2.24) の（局所的な）解である h_{ij} および p_{ij} は，次式の反復更新式を全ての i および j について繰り返し計算することで推定できる．

$$|h_{ij}|^{0.5} = \frac{|h_{(i+1)j}|^{0.5} + |h_{(i-1)j}|^{0.5}}{\sqrt{\left(|h_{(i+1)j}|^{0.5} + |h_{(i-1)j}|^{0.5}\right)^2 + \left(|p_{i(j+1)}|^{0.5} + |p_{i(j-1)}|^{0.5}\right)^2}} |b_{ij}|^{0.5} \tag{2.26}$$

$$|p_{ij}|^{0.5} = \frac{|p_{i(j+1)}|^{0.5} + |p_{i(j-1)}|^{0.5}}{\sqrt{\left(|h_{(i+1)j}|^{0.5} + |h_{(i-1)j}|^{0.5}\right)^2 + \left(|p_{i(j+1)}|^{0.5} + |p_{i(j-1)}|^{0.5}\right)^2}} |b_{ij}|^{0.5} \tag{2.27}$$

2.4.2　MHPSS

MHPSS では，振幅スペクトログラムの時間方向および周波数方向にそれぞれメディアンフィルタを適用する．メディアンフィルタは，フィルタを適用する方向のスパイク状の成分を除去するため，非線形かつ強力な平滑化が各方向に施される．したがって，時間方向および周波数方向の滑らかさを強調した信号を推定することができ，調波音 $\boldsymbol{H}$ および打撃音 $\boldsymbol{P}$ が得られる．

MHPSS では，フィルタサイズ $2D+1$ の移動メディアンフィルタをシフト長 1 点でずらしながら適用する．メディアンフィルタを適用するベクトルは，次式のように混合信号の振幅スペクトログラム $|\boldsymbol{B}|$ の行ベクトル $\boldsymbol{b}_{ij}^{(\mathrm{r})}$ および列ベクトル $\boldsymbol{b}_{ij}^{(\mathrm{c})}$ となる．

$$\boldsymbol{b}_{ij}^{(\mathrm{r})} = \left[|b_{i(j-D)}|, |b_{i(j-D+1)}|, \cdots, |b_{ij}|, \cdots, |b_{i(j+D-1)}|, |b_{i(j+D)}|\right] \in \mathbb{R}_{\geq 0}^{2D+1} \tag{2.28}$$

$$\boldsymbol{b}_{ij}^{(\mathrm{c})} = \left[|b_{(i-D)j}|, |b_{(i-D+1)j}|, \cdots, |b_{ij}|, \cdots, |b_{(i+D-1)j}|, |b_{(i+D)j}|\right] \in \mathbb{R}_{\geq 0}^{2D+1} \tag{2.29}$$

これらのベクトルにメディアンフィルタを適用することで，h_{ij} および p_{ij} が推定できる [25]．

$$|h_{ij}| = \mathrm{median}\left(\boldsymbol{b}_{ij}^{(\mathrm{r})}\right) \tag{2.30}$$

$$|p_{ij}| = \mathrm{median}\left(\boldsymbol{b}_{ij}^{(\mathrm{c})}\right) \tag{2.31}$$

ここで，$\mathrm{median}(\cdot)$ は入力されたベクトルの中央値のみをスカラーとして返す関数である．

2.5　本章のまとめ

本章では，音響信号の時間周波数領域への変換と基本的な音響信号処理について説明した．次章では，従来とは異なる時間周波数領域での表現として，時間微分複素スペクトログラムを提案する．さらに，提案する表現がもつ性質，役割，および変換における注意点などについて詳しく述べる．

第 3 章

提案手法

3.1 まえがき

本章では，本論文の提案手法である微分窓関数を用いた時間微分複素スペクトログラムとその逆変換の計算方法について述べ，さらに複素スペクトログラムと時間微分複素スペクトログラムを比較して BSS への適用可能性について議論する．3.2 節では，時間微分複素スペクトログラムの必要性について述べる．微分窓関数を用いて得られる時間微分複素スペクトログラムをどのような目的で用いるかについて述べる．3.3 節では，時間微分複素スペクトログラムから時間信号へ復元する方法について述べる．具体的には，時間微分複素スペクトログラムが時間信号へと変換できる条件を数式で示す．3.4 節では，振幅スペクトログラムと時間微分振幅スペクトログラムを図示し，比較する．窓関数と微分窓関数の周波数特性を比較し，BSS に導入する際に考えられる影響について述べる．最後に，3.5 節で本章をまとめる．

3.2 時間微分複素スペクトログラムの必要性

本論文では，位相を考慮した音源分離を考えるために，2.2.3 項で説明した修正位相スペクトログラムを BSS へ導入することを目的としている．修正位相スペクトログラムの計算に用いる時間微分複素スペクトログラムは複素スペクトログラムを時間方向に微分したもので，文献 [16] では，微分窓関数を用いて STFT を行うと時間微分複素スペクトログラムを得ることができることが示されている．具体的には，式 (2.3) の右辺の時間微分が式 (2.4) で表せるこ

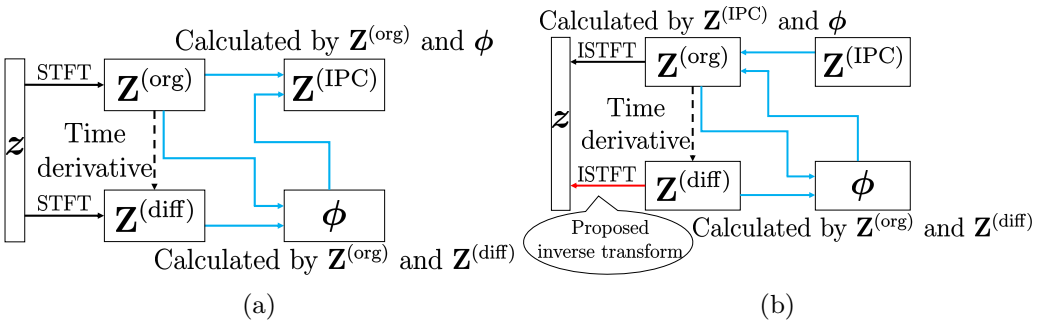


Fig. 3.1: Calculation of instantaneous phase corrected spectrogram

とに由来しており，次式で計算される．

$$\begin{aligned}
\frac{d}{d(\tau j)}\mathbf{z}_{ij}^{(\mathrm{org})} &= \frac{d}{d(\tau j)}\sum_q z[q]\omega_a[q-\tau(j-1)]\exp\left\{-2\pi\iota\frac{(q-1)(i-1)}{Q}\right\} \\
&= \sum_q z[q]\frac{d}{d(\tau j)}\omega_a[q-\tau(j-1)]\exp\left\{-2\pi\iota\frac{(q-1)(i-1)}{Q}\right\} \\
&= -\sum_q z[q]\omega_a^{(\mathrm{diff})}[q-\tau(j-1)]\exp\left\{-2\pi\iota\frac{(q-1)(i-1)}{Q}\right\} \\
&= \mathbf{z}_{ij}^{(\mathrm{diff})}
\end{aligned} \tag{3.1}$$

時間信号から修正位相スペクトログラムを計算するためには Fig. 3.1 (a) に示す 2.2.3 項の計算が必要となる．一方で，修正位相スペクトログラムから時間信号を得るためには Fig. 3.1 (b) に示す計算が必要となる．修正位相スペクトログラムを音源分離に用いた場合，分離された修正位相スペクトログラムから分離信号への変換には理想的に分離された分離信号の複素スペクトログラムおよび時間微分複素スペクトログラムが必要となる．しかし，理想的に分離された分離信号を得ることはできないため，分離された修正位相スペクトログラムを分離信号へ変換することはできない．

この問題を解決するために，音源分離手法を用いて複素スペクトログラムを分離し，分離された複素スペクトログラムおよび時間微分複素スペクトログラムを用いて修正位相スペクトログラムを分離信号に変換することを考えている．分離された複素スペクトログラムを得ることはできるが，同時に分離された時間微分複素スペクトログラムを得る必要がある．時間微分複素スペクトログラムに対して BSS を適用した例はなく，分離された時間微分複素スペクトログラムを得ることができるかは不明である．そこで，本論文では分離された時間微分複素スペクトログラムを得るために，時間微分複素スペクトログラムに対して BSS を適用することを考える．

時間微分複素スペクトログラムに対して BSS を適用するにあたって，音源分離の性能を測る指標が必要となる．音源分離の性能を測るためには，複素スペクトログラムに対応する時間信号を得る必要がある．しかし，STFT でよく利用されるシフト長では時間微分複素スペクト

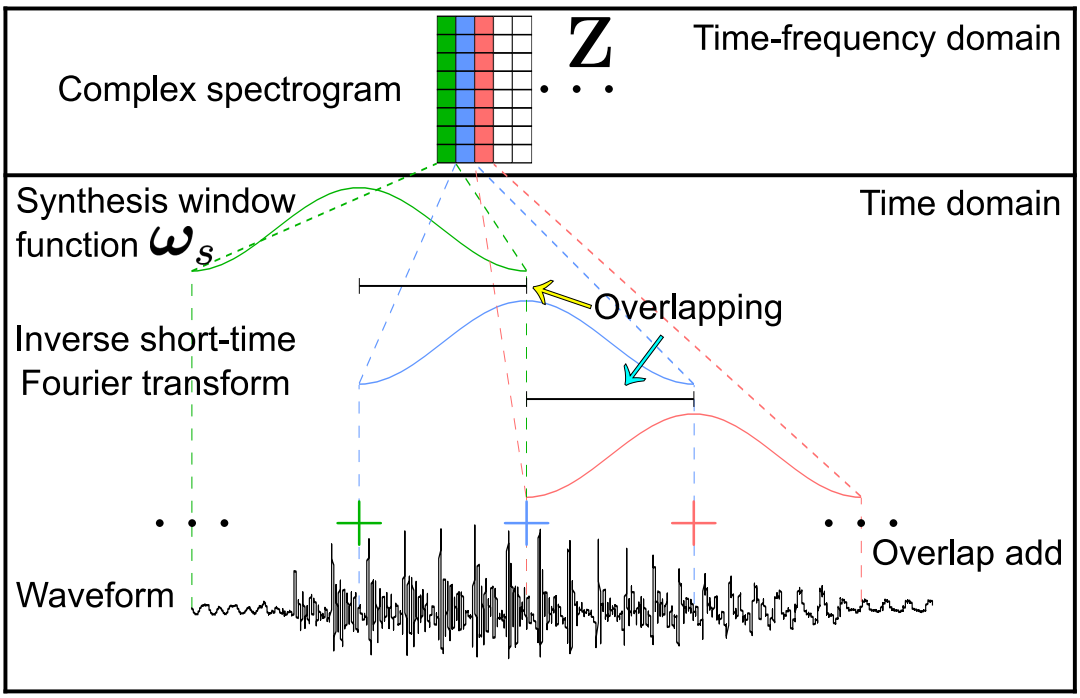


Fig. 3.2: Mechanism of inverse STFT.

ログラムを時間信号に変換することができないという問題があり，音源分離の性能を測るためにはこの問題を解決する必要がある．

3.3　時間微分複素スペクトログラムの信号への復元

STFT によって得られた複素スペクトログラムは逆 STFT によって時間信号へと変換される．逆 STFT の概要を Fig. 3.2 に示す．複素スペクトログラム $\mathbf{Z}$ は Fig. 2.8 のように計算される．したがって，逆変換でははじめに複素スペクトログラムの各時間フレームを逆離散フーリエ変換し．各フレームの時間信号を得る．さらに，フレームごとの時間信号から全体の時間信号を得るために適切な位置に各フレームを配置し，足し合わせるという処理になる．この処理はオーバーラップ加算（overlap add: OLA）と呼ばれる [27]．OLA によりもとの時間信号が複数回足し合わされるため，窓関数 $\boldsymbol{\omega}_a$ による影響とオーバーラップ幅による影響を打ち消すための適切な窓関数 $\boldsymbol{\omega}_s \in \mathbb{R}^Q$（以降，合成窓関数と呼ぶ）を各フレームに乗算した後に足し合わせる必要がある．

時間信号を STFT して得られた複素スペクトログラムが逆 STFT によってもとの時間信号に戻る条件，つまり完全再構成条件は次式で表される．

$$\sum_m \omega_a[t - m\tau]\omega_s[t - m\tau] = 1 \tag{3.2}$$

左辺がシフト長 τ の周期を持つ周期関数となっていることから，窓関数と合成窓関数に対する τ 個の条件が導かれることとなる．$\boldsymbol{\omega}_s$ は値を Q 個持つため，$\boldsymbol{\omega}_a$ に対する完全再構成条件を

満たす $\boldsymbol{\omega}_s$ は一意には定まらない．

時間微分複素スペクトログラムに対して完全再構成条件を満たす合成窓関数が存在するかを確認するために，微分窓関数の一例として，1 項の cos 関数で表されるハン窓 $\boldsymbol{\omega}_a$ の微分窓関数 $\boldsymbol{\omega}_a^{(\mathrm{diff})}$ を考える．Fig. 3.3 (a) に示すハン窓は次式で表される窓関数である．

$$\omega_a[q] = \begin{cases} 0.5 - 0.5\cos\left(\frac{2\pi(q-1)}{Q}\right) & (1 \leq q \leq Q) \\ 0 & (\text{otherwise}) \end{cases} \tag{3.3}$$

したがって，Fig. 3.3 (b) に示すハン窓の微分窓関数は次式で表される．

$$\omega_a^{(\mathrm{diff})}[q] = \begin{cases} \sin\left(\frac{2\pi(q-1)}{Q}\right) & (1 \leq q \leq Q) \\ 0 & (\text{otherwise}) \end{cases} \tag{3.4}$$

STFT で利用されるシフト長は窓長の 1/2 であることが多く，ハーフオーバーラップと呼ばれる．ハーフオーバーラップの条件下でハン窓の微分窓関数を考えると，完全再構成条件である式 (3.2) は次のように表される．

$$\sum_m \omega_a^{(\mathrm{diff})}\left[t - m\frac{Q}{2}\right]\omega_s^{(\mathrm{diff})}\left[t - m\frac{Q}{2}\right] = 1 \tag{3.5}$$

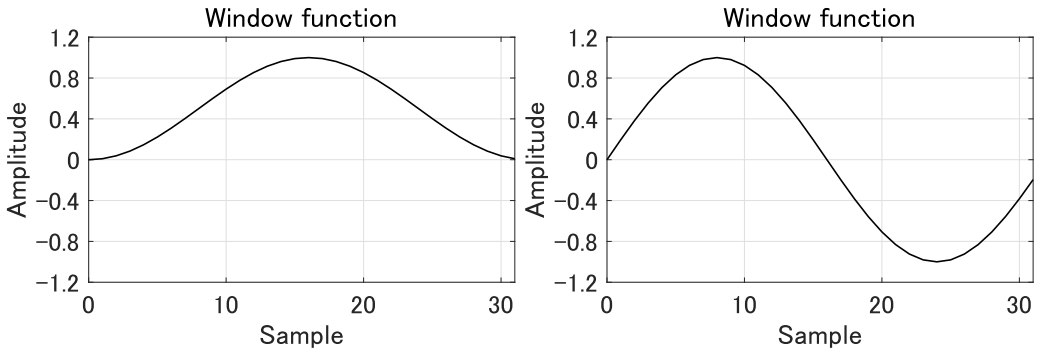
ここで，$\boldsymbol{\omega}_s^{(\mathrm{diff})} \in \mathbb{R}^Q$ は，$\boldsymbol{\omega}_a^{(\mathrm{diff})}$ に対する合成窓関数である．$t = 1$ の場合に注目すると，次式の条件が導かれる．

$$\omega_a^{(\mathrm{diff})}[1]\,\omega_s^{(\mathrm{diff})}[1] + \omega_a^{(\mathrm{diff})}\left[1 + \frac{Q}{2}\right]\omega_s^{(\mathrm{diff})}\left[1 + \frac{Q}{2}\right] = 1 \tag{3.6}$$

完全再構成条件を満たす合成窓関数が存在するためには，すべての t に対して式 (3.2) が成り立つ必要がある．しかし，式 (3.4) より，$\omega_a^{(\mathrm{diff})}[1] = 0$ かつ $\omega_a^{(\mathrm{diff})}[1 + Q/2] = 0$ であるため，上式を満たすような合成窓関数 $\boldsymbol{\omega}_s^{(\mathrm{diff})}$ は存在しない．この事実はハン窓の場合のみではなく，音響信号処理で一般的に用いられる窓関数の微分窓関数のほとんどはハーフオーバーラップの条件下での完全再構成条件を満たす合成窓関数が存在しない．より具体的には，ある t について式 (3.2) の窓関数の部分がすべて 0 であるときのみ完全再構成条件を満たす合成窓関数が存在しない．逆に，完全再構成条件を満たす合成窓関数が存在するための窓関数の条件を考えると，式 (3.2) より次式となる．

$$\prod_{t=1}^{\tau}\left\{\sum_m |\omega_a[t - m\tau]|\right\} \neq 0 \tag{3.7}$$

式 (3.7) を満たす変換は，例えば STFT におけるシフト長が $\tau = Q/2$ のハーフオーバーラップではなく，$\tau = Q/4$ のクォーターオーバーラップシフト等が考えられる．実際に本論文の次章の実験では，$\tau = Q/4$ の条件で時間微分複素スペクトログラムを計算し，逆 STFT によって復元できることを実験的に示す．



(a) Hann window.　(b) Differential Hann window.

Fig. 3.3: Window functions.

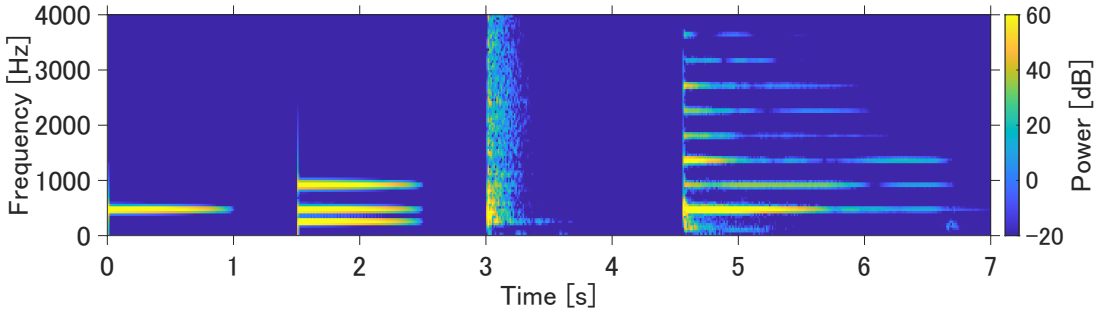


Fig. 3.4: Spectrogram using Hann window.

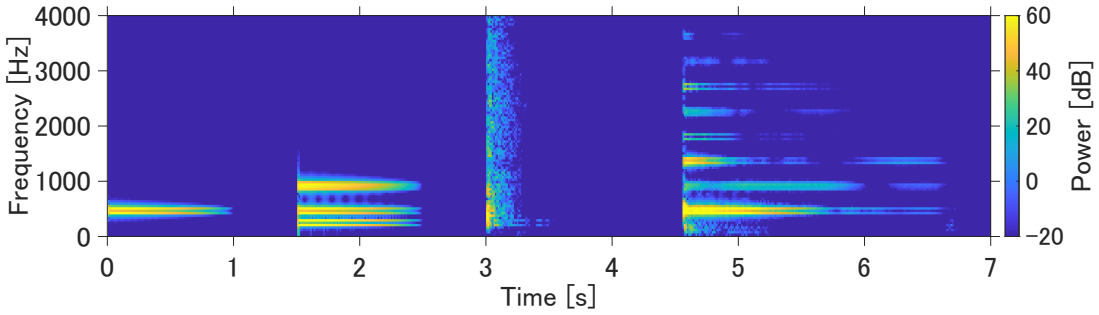
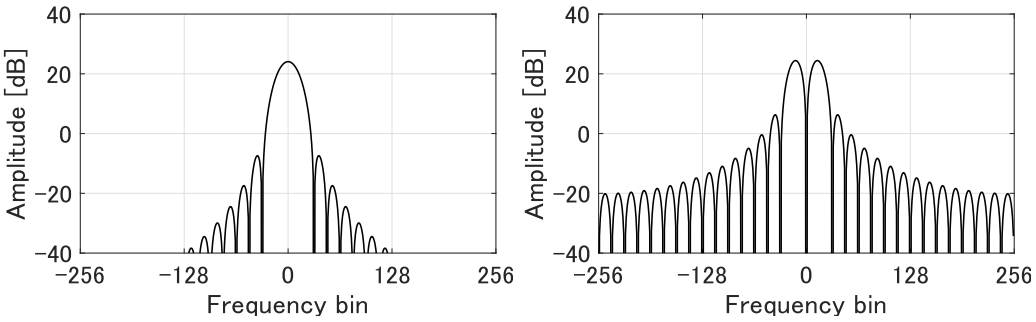


Fig. 3.5: Spectrogram using differential Hann window.

3.4　窓関数による振幅スペクトログラムの違い

2.2 項で示したように，時間周波数領域への変換手法によって時間周波数特徴量の見た目は大きく異なることがわかる．Fig. 2.2 の時間信号に対して STFT を適用して得られる振幅スペクトログラムおよび時間微分振幅スペクトログラムをそれぞれ Fig. 3.4 および Fig. 3.5 に示す．



(a)　　(b)

Fig. 3.6: Frequency response.

Fig. 3.4 と比較すると，Fig. 3.5 は周波数方向に 2 本の横線を並べたような構造が見られる．これは，窓関数の周波数特性の違いによるものである．ハン窓および微分ハン窓の周波数応答を Figs. 3.6 (a) および Fig. 3.6 (b) に示す．時間領域で窓関数を乗算する操作は周波数領域では畳み込み演算を行う操作となるので，複素スペクトログラムの各時間フレームでは窓関数の周波数応答が畳み込まれている．ハン窓および微分ハン窓の周波数特性を比較した場合，微分ハン窓は左右にメインローブが分かれており，サイドローブレベルも高くなっている．さらに，中心周波数付近で非常に小さな値をとっていることに起因して，正弦波 Figs. 3.4 及び 3.5 の 0–1 s の純粋な正弦波成分は，時間微分複素スペクトログラムでは 2 重の横線となっている．

周波数特性において，サイドローブレベルが高くなるこは，振幅スペクトログラムの周波数方向へのにじみが大きく現れることを表す．このようにサイドローブレベルが高いことは，一般的な時間周波数解析においてはデメリットとされるが，2.3.3 項で説明したように IVA の音源モデルは周波数方向の共起性を分離の手がかりに用いるため，総合的にはよい分離性能が得られると考えた．また，前述のように正弦波成分が 2 重の横筋状の構造として現れることは，2.4 項で説明した HPSS の横筋状・縦筋状の構造を陽に用いた音源モデルへの適合度合いを高め，調波成分の分離性能の向上が期待される．以上の仮説について詳しく検証するため，4 章では時間微分複素スペクトログラムに IVA および HPSS を適用する．

3.5　本章のまとめ

本章では，時間微分複素スペクトログラムを比較して BSS への適用可能性について議論した．微分窓関数を用いて計算される時間微分複素スペクトログラムを時間信号へと復元する場合，式 (3.7) を満たすような窓関数を選ぶことが必要であることを示した．また，IVA や HPSS において高い分離性能を得ることができると考えた．次章では，実際に時間微分複素スペクトログラムを用いた音源分離の性能についての実験をまとめる．

第 4 章

実験

4.1 まえがき

本実験では，複素スペクトログラムと時間微分複素スペクトログラムに対して 2 章で述べた IVA，OHPSS，および MHPSS を適用し，分離性能を比較する．4.2 節では本実験で用いる評価指標について述べる．具体的には，信号の分離度合いを示す客観的指標を数式を用いて表す．4.3 節では本実験で用いる実験条件を示す．HPSS の実験では，実験条件を決定した後，トレーニングデータを用いてハイパーパラメータを探索し，テストデータで用いる実験条件について決定する．4.4 節では 4.3 節で決定した実験条件を用いて音源分離を行い，各性能指標を比較する．HPSS の実験では，テストデータを用いて実験を行い，各性能指標を比較する．最後に，4.5 節で本章をまとめる．

4.2 評価指標

本実験では，評価指標として信号対歪み比（source-to-distortion ratio: SDR）[33]，SDR 改善量（SDR improvement: SDRi），信号対干渉比（source-to-interference ratio: SIR）[33]，および信号群対人工歪み比（sources-to-artificial ratio: SAR）[33] を用いた．今，n 番目の目的音源 $\boldsymbol{s}_n \in \mathbb{R}^L$ における推定信号 $\hat{\boldsymbol{y}}_n \in \mathbb{R}^L$ は以下の式に示すように分解することができる．

$$\hat{\boldsymbol{y}}_n = \boldsymbol{s}_{\text{target}} + \boldsymbol{e}_{\text{interf}} + \boldsymbol{e}_{\text{artif}} \tag{4.1}$$

ここで，$\boldsymbol{s}_{\text{target}} \in \mathbb{R}^L$，$\boldsymbol{e}_{\text{interf}} \in \mathbb{R}^L$，および $\boldsymbol{e}_{\text{artif}} \in \mathbb{R}^L$ はそれぞれ $\boldsymbol{s}_n$ の成分，$\hat{\boldsymbol{y}}_n$ に残留した非目的音源成分，及び音源分離によって生じた人工的な歪み成分を表す．

SIR は，音源分離度合いを示した指標であり，$\boldsymbol{s}_{\text{target}}$，$\boldsymbol{e}_{\text{interf}}$，および $\boldsymbol{e}_{\text{artif}}$ を用いて次式で表される．

$$\text{SIR} = 10 \log_{10} \frac{\|\boldsymbol{s}_{\text{target}}\|_2}{\|\boldsymbol{e}_{\text{interf}}\|_2} \ [\text{dB}] \tag{4.2}$$

ここで，$\|\cdot\|_2$ は L_2 ノルムを表している．SIR が高い場合，よく分離された信号であることを表している．

SAR は，音源分離度合いを示した指標であり，$\boldsymbol{s}_{\mathrm{target}}$，$\boldsymbol{e}_{\mathrm{interf}}$，および $\boldsymbol{e}_{\mathrm{artif}}$ を用いて次式で表される．

$$\mathrm{SAR} = 10\log_{10}\frac{\|\boldsymbol{s}_{\mathrm{target}} + \boldsymbol{e}_{\mathrm{interf}}\|_2}{\|\boldsymbol{e}_{\mathrm{artif}}\|_2}\ [\mathrm{dB}] \tag{4.3}$$

SAR が高い場合，信号処理によって生じた人工的な歪みを含んでいない信号であることを表している．

SDR は，SIR と SAR の両方を加味した総合的な指標であり，SDR の値が高いほど総合的によい分離を行っているということができる．

$$\mathrm{SDR} = 10\log_{10}\frac{\|\boldsymbol{s}_{\mathrm{target}}\|_2}{\|\boldsymbol{e}_{\mathrm{interf}} + \boldsymbol{e}_{\mathrm{artif}}\|_2}\ [\mathrm{dB}] \tag{4.4}$$

したがって，高い SDR 値を達成するには，$\boldsymbol{e}_{\mathrm{interf}}$，および $\boldsymbol{e}_{\mathrm{artif}}$ が少なく，$\boldsymbol{s}_{\mathrm{target}}$ が高精度に推定されている必要がある．

4.3　各手法の実験条件

すべての実験で DGTtool [29, 30] を用いて複素スペクトログラムおよび時間微分複素スペクトログラムを計算した．DGTtool は，STFT を適用できる扱いやすい MATLAB 用のライブラリであり，入力信号に対する適切なゼロパディング，振幅スペクトログラムの表示，微分窓関数の生成，および最適合成窓関数の生成などを自動で行うことができる．また，STFT のシフト長は全ての実験において共通の $\tau = Q/4$ と設定した．この設定は，3.3 節で述べた通り，時間微分複素スペクトログラムを時間領域の信号に復元する際の完全再構成条件を満たすことを目的としている．

4.3.1　IVA

SiSEC2011 [34] のデータセットを音源信号として使用し，サンプリング周波数を 16 kHz とした．Table 4.1 に実験で用いた音源信号の詳細を示す．新情報処理開発機構（Real World Computing Partnership: RWCP）データベース [32] 収録のインパルス応答 E2A（$T_{60} = 300$ ms）による 2 音源の畳み込み混合を行い，10 音楽と 10 音声の 2 チャネル観測信号を生成した．ここで用いたインパルス応答 E2A の収録条件は Fig. 4.1 に示す通りである．

IVA の周波数毎の分離行列の初期値は単位行列とした．分離には式 (2.22) で示す AuxIVA および ISS に基づく更新式を用いた．100 回の更新によって得られた分離行列 $\mathbf{W}_i$ を用いて分離後の複素スペクトログラムを計算し，さらに式 (2.13) に示す PB 法を適用して周波数毎の信号のスケールを補正した．評価指標には 4.2 節で説明した SDR，SIR，および SAR を用いた．

Table 4.1: Sources obtained from SiSEC 2011 dataset used as dry sources

Data name	Source (1/2)	Length [s]
another_dreamer-the_ones_we_love	drums/guitar	25.6
another_dreamer-the_ones_we_love	guitar/vocals	25.6
bearlin-roads	acoustic_guit_main/vocals	14.6
bearlin-roads	drums/bass	14.6
bearlin-roads	piano/acoustic_guit_main	14.6
fort_minor-remember_the_name	violins_synth/vocals	24.6
fort_minor-remember_the_name	vocals/drums	24.6
tamy-que_pena_tanto_faz	guitar/vocals	13.6
ultimate_nz_tour	drums/vocals	18.6
ultimate_nz_tour	guitar/synth	18.6
female4	no. 1/no. 2	10.0
female4	no. 1/no. 4	10.0
female4	no. 2/no. 3	10.0
female4	no. 2/no. 4	10.0
female4	no. 3/no. 4	10.0
male4	no. 1/no. 2	10.0
male4	no. 1/no. 4	10.0
male4	no. 2/no. 3	10.0
male4	no. 2/no. 4	10.0
male4	no. 3/no. 4	10.0

4.3.2 OHPSS

HPSS に関する実験では，SiSEC2016 [35] の DSD100 データセットを音源信号として使用した．DSD100 はトレーニングデータ（Dev）とテストデータ（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．各曲はさまざま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），およびその他の音源（other）が音源毎に収録されている．本実験では，drums を打撃音成分，その他すべての瞬時混合を調波成分として分離性能の評価を行った．また，音源信号のサンプリング周波数は 44.1 kHz とした．

t 回目の反復における $|\boldsymbol{H}|$ を $\left|\boldsymbol{H}^{(t)}\right|$，$|\boldsymbol{P}|$ を $\left|\boldsymbol{P}^{(t)}\right|$ と定義する．OHPSS の初期値 $\left|\boldsymbol{H}^{(0)}\right|$

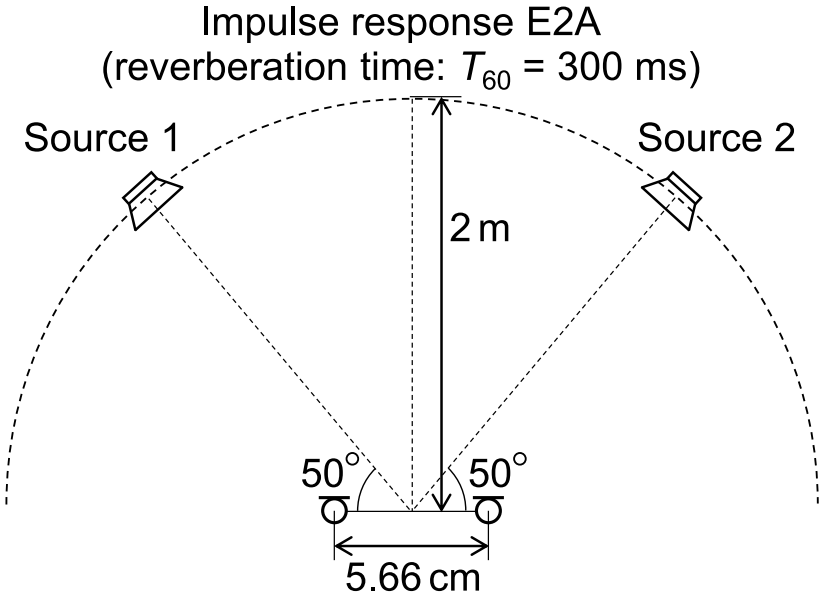


Fig. 4.1: Recording conditions of impulse response E2A.

および $\left|\boldsymbol{P}^{(0)}\right|$ は次式のように決定した．

$$|\boldsymbol{H}^{(0)}| = |\boldsymbol{P}^{(0)}| = \left(\frac{|\boldsymbol{B}|^C}{2}\right)^{\frac{1}{C}} \tag{4.5}$$

式 (2.26) および式 (2.27) の反復回数は 15 回とした．推定した $\left|\boldsymbol{H}^{(15)}\right|$ および $\left|\boldsymbol{P}^{(15)}\right|$ から位相を含めた複素スペクトログラム $\boldsymbol{H}$ および $\boldsymbol{P}$ を得るために次式の Wiener フィルタを適用した．

$$h_{ij} = \frac{\left|h_{ij}^{(15)}\right|^2}{\left|h_{ij}^{(15)}\right|^2 + \left|p_{ij}^{(15)}\right|^2} b_{ij} \tag{4.6}$$

$$p_{ij} = \frac{\left|p_{ij}^{(15)}\right|^2}{\left|h_{ij}^{(15)}\right|^2 + \left|p_{ij}^{(15)}\right|^2} b_{ij} \tag{4.7}$$

トレーニングデータを用いて各複素スペクトログラムに対して分離を行い，分離信号を得た後，4.2 節で説明した性能指標を計算した．計算された SDR が最も高くなるような窓関数の長さ Q およびドメインパラメータ C を決定し，テストデータに対して実験を行った．テストデータに対する評価指標には 4.2 節で説明した SDRi，SIR，および SAR を用いた．

4.3.3　MHPSS

MHPSS の実験についても，音源信号は 4.3.2 項の実験条件と同様である．OHPSS と同様に推定した $\left|\boldsymbol{H}^{(\mathrm{med})}\right|$ および $\left|\boldsymbol{P}^{(\mathrm{med})}\right|$ から複素スペクトログラム $\boldsymbol{H}$ および $\boldsymbol{P}$ を得るために

式 (4.8) および式 (4.9) の Wiener フィルタを適用した.

$$h_{ij} = \frac{\left|h_{ij}^{(\mathrm{med})}\right|^2}{\left|h_{ij}^{(\mathrm{med})}\right|^2 + \left|p_{ij}^{(\mathrm{med})}\right|^2} b_{ij} \tag{4.8}$$

$$p_{ij} = \frac{\left|p_{ij}^{(\mathrm{med})}\right|^2}{\left|h_{ij}^{(\mathrm{med})}\right|^2 + \left|p_{ij}^{(\mathrm{med})}\right|^2} b_{ij} \tag{4.9}$$

トレーニングデータを用いて各複素スペクトログラムに対して分離を行い，分離信号を得た後，4.2 節で説明した性能指標を計算した．計算された SDR が最も高くなるような窓関数の長さ Q およびフィルタ長 M を決定し，テストデータに対して実験を行った．テストデータに対する評価指標には 4.2 節で説明した SDRi，SIR，および SAR を用いた.

4.4 実験結果

4.4.1 IVA

IVA での実験結果の平均 SDR，平均 SIR，平均 SAR に関するヴァイオリンプロットを Figs. 4.2–4.4 に示す．ここで，図中の青色（左側に表示したヴァイオリンプロット）および赤色（右側に表示したヴァイオリンプロット）は，それぞれハン窓および微分ハン窓に対する実験結果である．図中の青色または赤色の点は各データに対する実験結果の値，曲線領域内に同色で示す水平線は全点の平均値，中央の白色の点は中央値，灰色の太い垂直線は全点の四分位範囲を，曲線はカーネル密度推定に基づく分布を表している.

短い窓長では平均 SDR には大きな差は現れなかったが，長い窓長では時間微分複素スペクトログラムは複素スペクトログラムよりも低い SDR を示した．SDR は総合的な分離性能を表すため，総合的な分離性能では複素スペクトログラムのほうが高いと言える．一方で，512 点の窓長を除いて，時間微分複素スペクトログラムは複素スペクトログラムより低い SIR を示し，高い SAR を示した．複素スペクトログラムには人工的な歪みを抑制するような効果があることが読み取れる.

4.4.2 OHPSS

トレーニングデータに対する全実験結果は付録 A.1 に示す．付録 A.1 より，テストデータでは窓関数の長さ Q を 4096 点（92.80 ms），ドメインパラメータ C を 1.6 と固定して実験を行った．テストデータに対する実験結果を Table 4.2 に示す.

テストデータの実験では，複素スペクトログラムおよび時間微分複素スペクトログラムの両方で SDR が最も高く現れる窓関数の長さ Q とドメインパラメータ C が同じとなった．また，テストデータに対する実験においても平均 SDRi には同様の値が現れている．したがって，OHPSS においては複素スペクトログラムおよび時間微分複素スペクトログラムは同様の

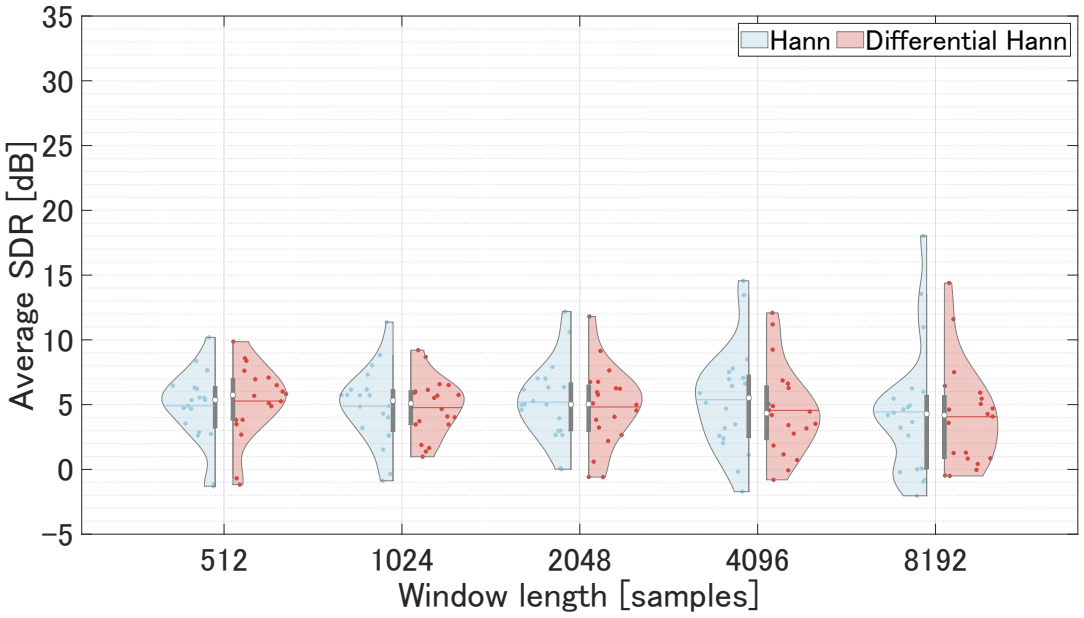


Fig. 4.2: Average SDR values calculated by IVA outputs using Hann and differential Hann window with various window length.

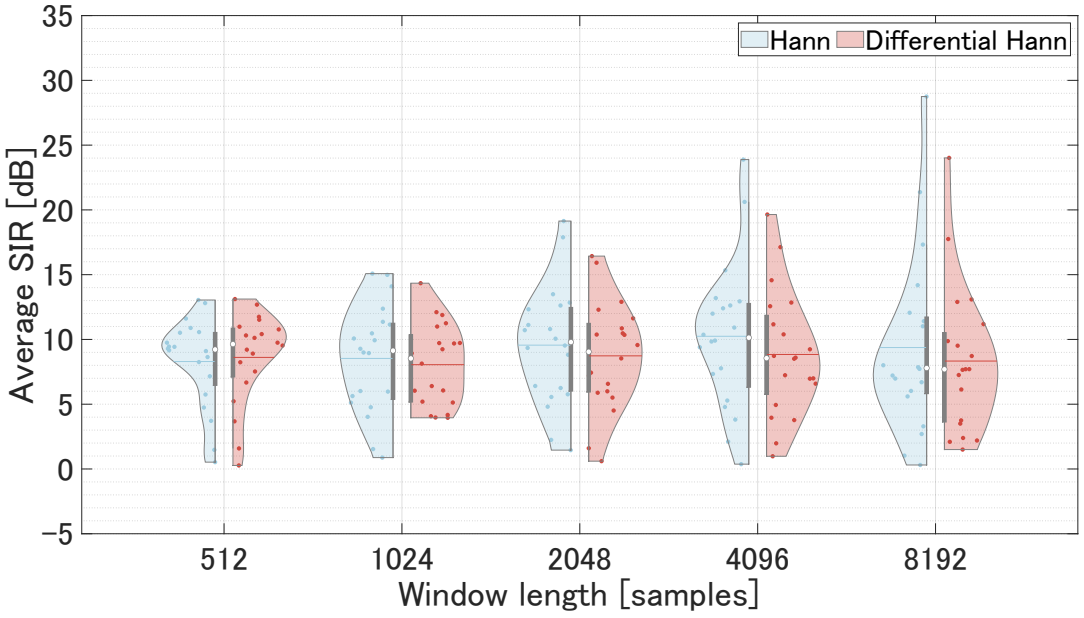


Fig. 4.3: Average SIR values calculated by IVA outputs using Hann and differential Hann window with various window length.

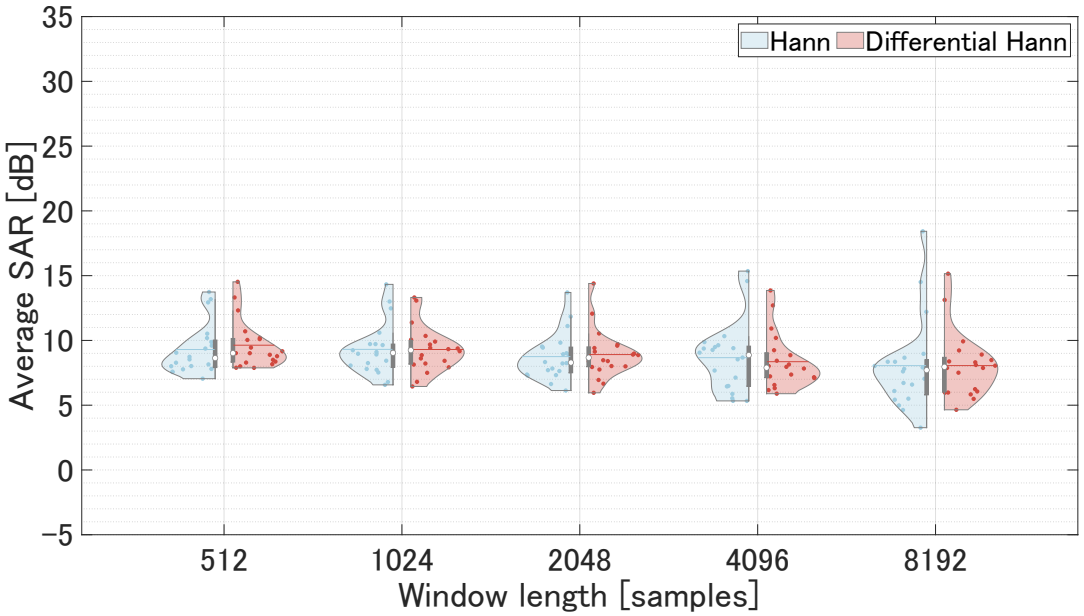


Fig. 4.4: Average SAR values calculated by IVA outputs using Hann and differential Hann window with various window length.

Table 4.2: Evaluation scores of each window in OHPSS

Window	Average SDRi [dB]	Average SIR [dB]	Average SAR [dB]
Hann	-3.280	5.800	6.368
Differential Hann	-3.307	5.137	7.012

Table 4.3: Evaluation scores of each window in MHPSS

Window	Average SDRi [dB]	Average SIR [dB]	Average SAR [dB]
Hann	-2.659	6.114	7.223
Differential Hann	-2.785	5.208	8.157

性質を持っていることがわかる．しかし，SIR および SAR は異なる結果が現れており，複素スペクトログラムでは分離度合いを下げる代わりに人工的な歪みを抑制する特徴があることがわかる．

4.4.3 MHPSS

トレーニングデータに対する全実験結果は付録 A.2 に示す．付録 A.2 より，テストデータでは窓関数の長さ Q を 4096 点（92.80 ms），フィルタ長 M を 17 と固定して実験を行った．テストデータに対する実験結果を Table 4.3 に示す．

MHPSS ではフィルタ長が 3 のとき，実験結果（特に SAR）に大きな差が見られる．これ

は，3.4 節で説明した時間微分振幅スペクトログラムの構造によるものである．具体的には，2 重線の構造に対して長さ 3 のメディアンフィルタを適用すると必ず 1 重線へと変換されることに由来している．長さ 3 のメディアンフィルタを時間微分振幅スペクトログラムに適用することで時間微分振幅スペクトログラムの構造を大きく変化させることになる．一方で，振幅スペクトログラムには 2 重線の構造はないので，長さ 3 のメディアンフィルタはもとのスペクトログラムの構造を大きく変化させることはない．したがって，フィルタ長が 3 のとき，振幅スペクトログラムは高い SAR および低い SIR を示し，時間微分振幅スペクトログラムは極端に低い SAR を示している．その他のフィルタ長では同等の SDR が現れており，総合的な分離性能には大きな変化がないことが読み取れる．

4.5　本章のまとめ

本章では，提案手法の各パラメータの検討および得られた複素スペクトログラムの性能比較を行った．4.4 節では，時間微分複素スペクトログラムは総合的な分離性能では複素スペクトログラムよりも低いまたは同等の性能を示した．時間微分複素スペクトログラムは BSS においては分離度合いを下げる代わりに音質や人工的な歪みを小さくするような性質があることが示された．また，時間微分複素スペクトログラムに対しても BSS を適用することができ，十分分離可能であることがわかった．次章では，本論文の結論をまとめる．

第 5 章

結言

本論文では，位相に関する時間周波数特徴量を考慮した音源分離を目的とし，時間微分複素スペクトログラムを既存の BSS に適用した．特定の条件下での時間微分複素スペクトログラムの逆変換が計算可能なことを示した．また，時間微分複素スペクトログラムの時間周波数表現としての性質を調査し，既存の BSS へ適用した．時間微分複素スペクトログラムの BSS においては，多くの実験において分離度合いが低下し，人工歪みが軽減される傾向が見られた．総合的な分離性能の向上には至らなかったが，時間微分複素スペクトログラムを既存の BSS 手法を用いて分離することは十分可能であることを示した．

最後に今後の展望を述べる．本論文では，修正位相スペクトログラムを用いることで位相を考慮した音源分離を行ことを目的としている．修正位相スペクトログラムをモデル化し，BSS によって分離された時間微分複素スペクトログラムを用いて修正位相スペクトログラムを時間信号へと変換することで，位相を考慮した音源分離の実現が期待できる．

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講師に心より感謝申し上げます．北村大地講師には，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．音響信号処理や数学などの知識や研究に用いる計算機設備の設置および利用など，貴重な経験となりました．

本論文の副査である雛元洋一助教には，論文の構成や記述に関して有益な助言を頂き，大変お世話になりました．ここに厚く御礼申し上げます．

東京農工大学の矢田部浩平准教授には，ミーティングなどを通じ，多数の助言を頂きました．ここに感謝申し上げます．

北村研究室の先輩である専攻科 1 年の川口翔也氏，蓮池郁也氏，溝渕悠朔氏，村田佳斗氏には研究や論文執筆の基礎や MATLAB に関するアドバイスなど数々のご支援をいただきました．また，北村研究室同期の唐渡昂希氏，岸本麗央氏，島田優斗氏には研究のみならず一年間の学校生活を様々な面で支えていただきました．研究室で大変充実した時間を過ごすことができました．心より感謝申し上げます．

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] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, "A review of blind source separation methods: Two converging routes to ILRMA originating from ICA and NMF," *Asia-Pacific Signal and Information Processing Association Transactions on Signal and Information Processing*, vol. 8, no. e12, pp. 1–14, 2019.

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

[3] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1–3, pp. 21–34, 1998.

[4] 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.

[5] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 666-678, 2006.

[6] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS," *IEEE International Symposium on Circuits and Systems*, pp. 3247–3250, 2007.

[7] F. Hasuike, D. Kitamura, and R. Watanabe, "DNN-based frequency-domain permutation solver for multichannel audio source separation," *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 871–876, 2022.

[8] S. Yamaji and D. Kitamura, "DNN-based permutation solver for frequency-domain independent component analysis in two-source mixture case," *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 781–787, 2020.

[9] L. Li, H. Kameoka, S. Seki "HBP: An efficient block permutation solver using Hungarian algorithm and spectrogram inpainting for multichannel audio source separation," *in Proceedings of IEEE International Conference on Acoustics, Speech and Signal*

Processing, pp. 516–520, 2022.

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

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

[12] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, pp. 788–791, 1999.

[13] 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 Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1626–1641, 2016.

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

[15] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, “Relaxation of rank-1 spatial constraint in overdetermined blind source separation,” *European Signal Processing Conference*, pp. 1261–1265, 2015.

[16] K. Yatabe and Y. Oikawa, “Phase corrected total variation for audio signals,” *in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 656–660, 2018.

[17] 小泉悠馬, “深層学習に基づく音源強調と位相制御,” **日本音響学会誌**, vol. 75, no. 3, pp. 156–163, 2019

[18] V. Britanak, K. R. Rao, “An efficient implementation of the forward and inverse MDCT in MPEG audio coding,” *IEEE Signal Processing Letters*, vol. 8, no. 2, pp. 48–51, 2001.

[19] B. Fuentes, A. Liutkus, R. Badeau, and G. Richard, “Probabilistic model for main melody extraction using Constant-Q transform,” *in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5357–5360, 2012.

[20] V. Tiwari, “MFCC and its applications in speaker recognition,” *International journal on emerging technologies*, vol. 1, pp. 19–22, 2010.

[21] M. S. Nagawade and V. R. Ratnaparkhe, “Musical instrument identification using MFCC,” *IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology*, pp. 2198–2202, 2017

[22] S. Kawaguchi and D. Kitamura, “Amplitude spectrogram prediction from mel-frequency cepstrum coefficients and loudness using deep neural networks,” *in Pro-*

ceedings of RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing, 2023 (in press).

[23] Y. Masuyama, K. Yatabe and Y. Oikawa, "Low-rankness of Complex-valued Spectrogram and Its Application to Phase-aware Audio Processing," *in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019.

[24] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," *in Proceedings of International Conference on Latent Variable Analysis and Signal Separation*, pp. 722–727, 2001.

[25] 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," *in Proceedings of 16th European Signal Processing Conference*, 2008, pp. 1–4.

[26] D. FitzGerald, "Harmonic/percussive separation using median filtering," *in Proceedings of the International Conference on Digital Audio Effects*, vol. 13, 2010.

[27] 小野順貴, "短時間フーリエ変換の基礎と応用," **日本音響学会誌**, vol. 72, no. 12, pp. 764–769, 2016

[28] D. W. Griffin and J. S. Lim, "Signal estimation from modifiedshort-time Fourier transform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, pp. 236–243, 1984.

[29] K. Yatabe, DGTtool. Zenodo, 2021, doi:10.5281/ZENODO.5010751.

[30] 矢田部浩平, "短時間フーリエ変換および離散ガボール変換の MATLAB 実装について," **日本音響学会** *2021* **年秋季研究発表会講演論文集**, pp. 253–256, 2021.

[31] S. Robin and N. Ono, "Fast and stable blind source separation with rank-1 updates," *in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.236–240, 2020.

[32] 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," *in Proceedings of International Conference on Language Resources and Evaluation*, pp. 965–968, 2000.

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

[34] S. Araki, F. Nesta, E. Vincent, Z. Koldovsky, G. Nolte, A. Ziehe, and A. Benichoux, "The 2011 Signal Separation Evaluation Campaign (SiSEC2011): - Audio Source Separation -," *in Proceedings of International Conference on Latent Variable Analysis and Signal Separation*, pp. 414–422, 2012.

[35] A. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and

J. Fontecave, “The 2016 signal separation evaluation campaign,” *in Proceedings of Latent Variable Analysis Signal Separation*, pp. 323–332, 2017.

付録 A

トレーニングデータに対する HPSS の全実験結果

A.1　OHPSS

Figs. A.1–A.20 に 4.3.2 項のトレーニングデータに対する実験により得られた結果を示す．Figs. A.1–A.5 は SDR，Figs. A.6–A.10 は SDRi，Figs. A.11–A.15 は SIR，および Figs. A.16–A.20 は SAR を表している．

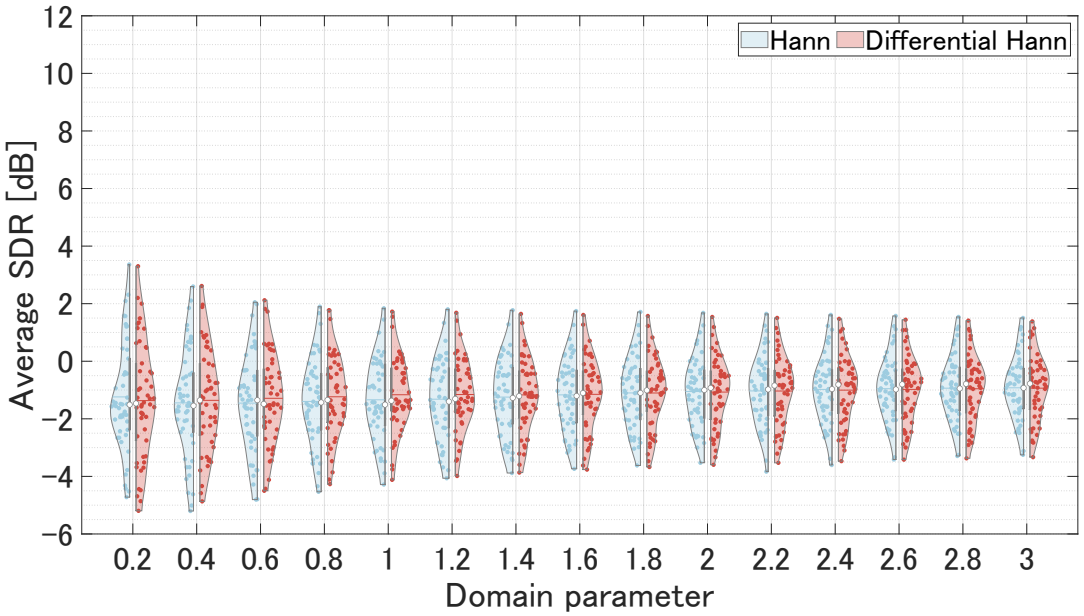


Fig. A.1: OHPSS experiment using dev dataset. (SDR, window length = 512 samples)

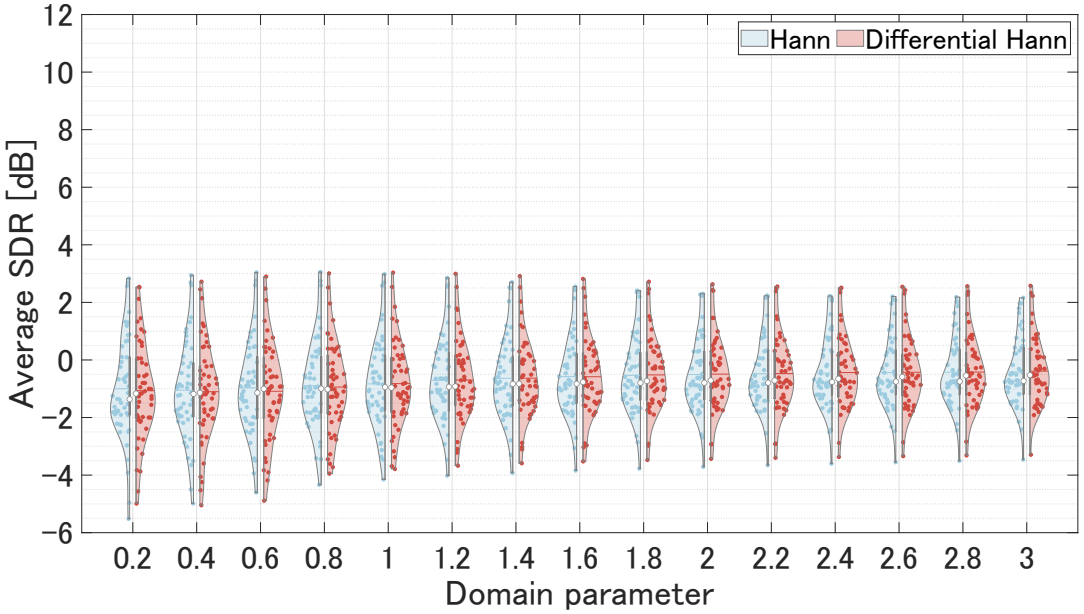


Fig. A.2: OHPSS experiment using dev dataset. (SDR, window length = 1024 samples)

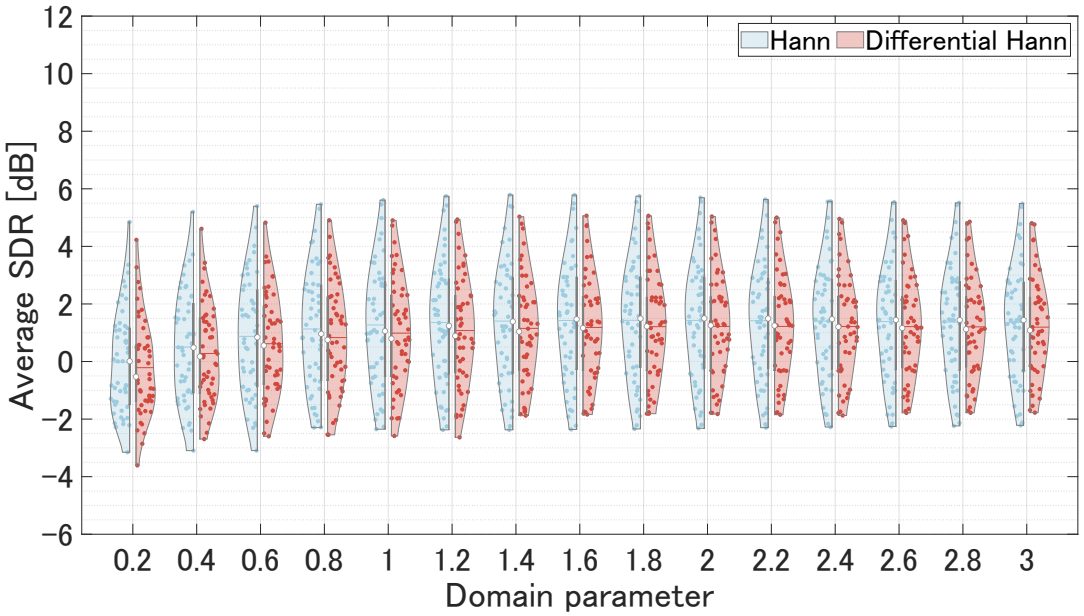


Fig. A.3: OHPSS experiment using dev dataset. (SDR, window length = 2048 samples)

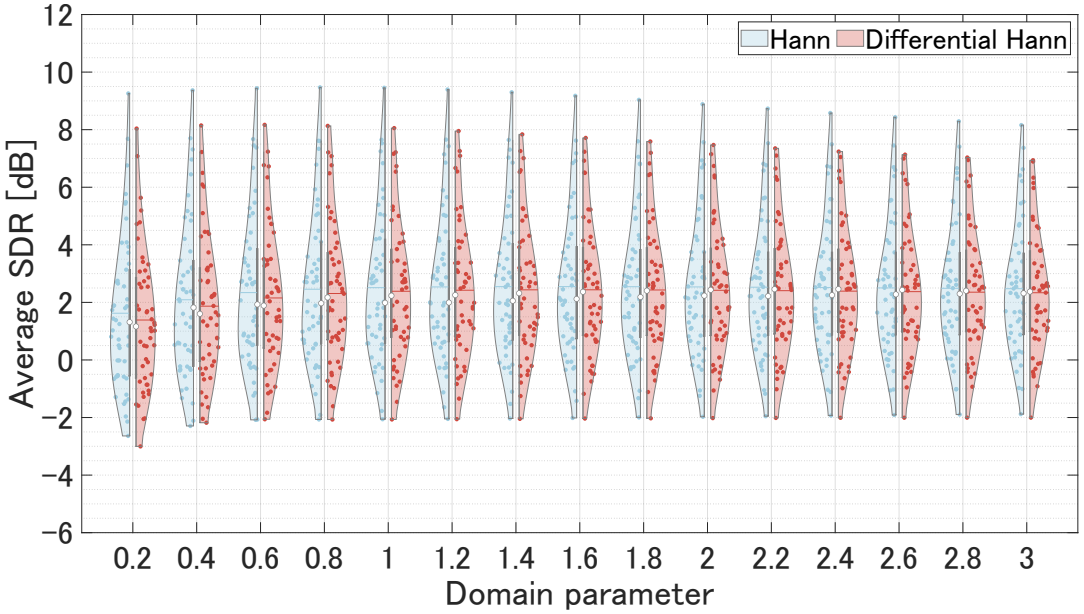


Fig. A.4: OHPSS experiment using dev dataset. (SDR, window length = 4096 samples)

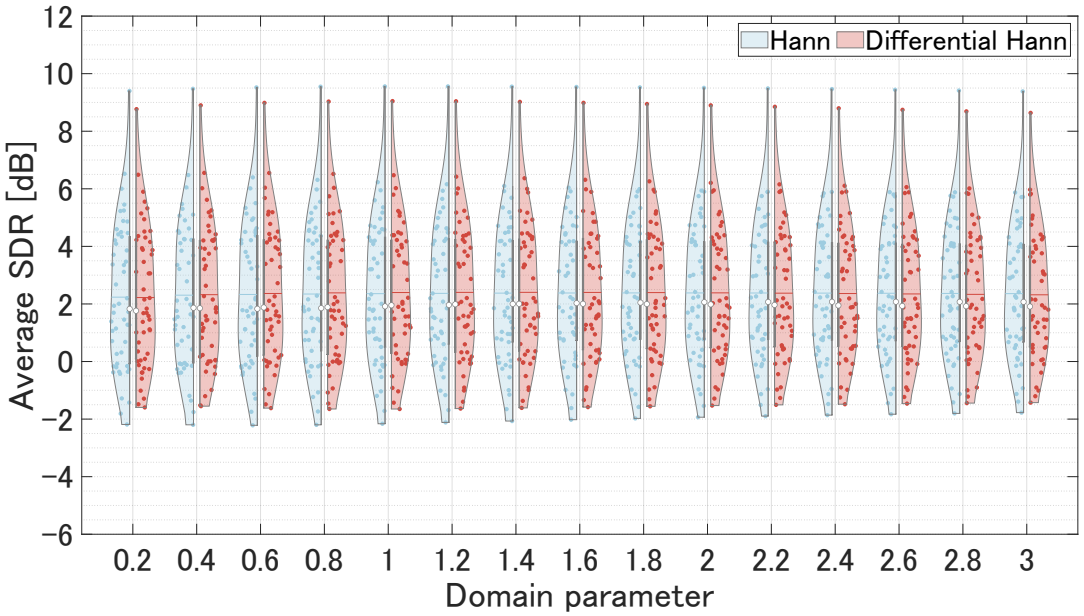


Fig. A.5: OHPSS experiment using dev dataset. (SDR, window length = 8192 samples)

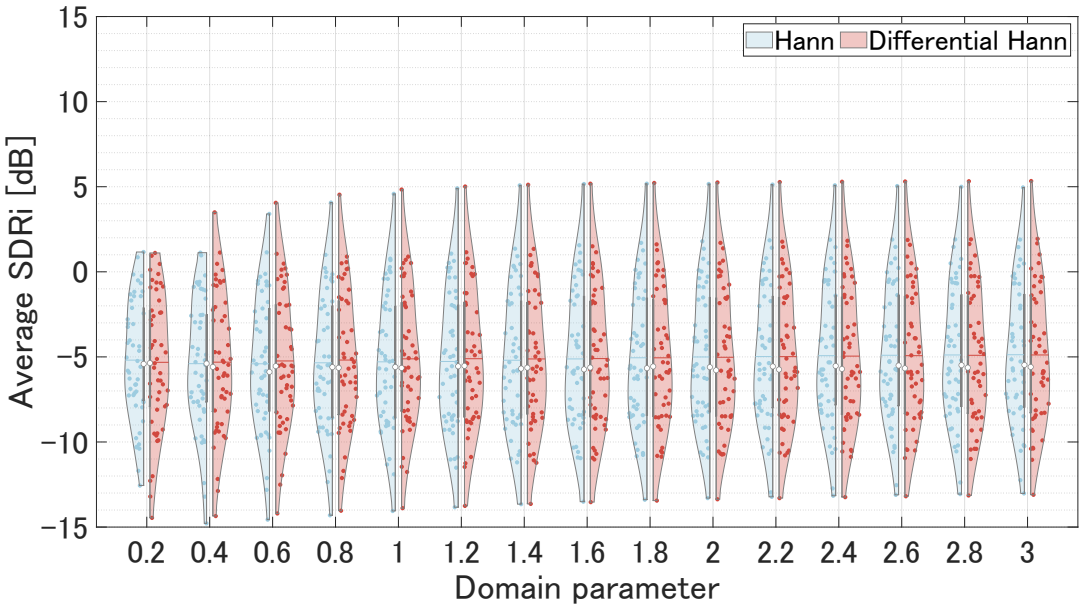


Fig. A.6: OHPSS experiment using dev dataset. (SDRi, window length = 512 samples)

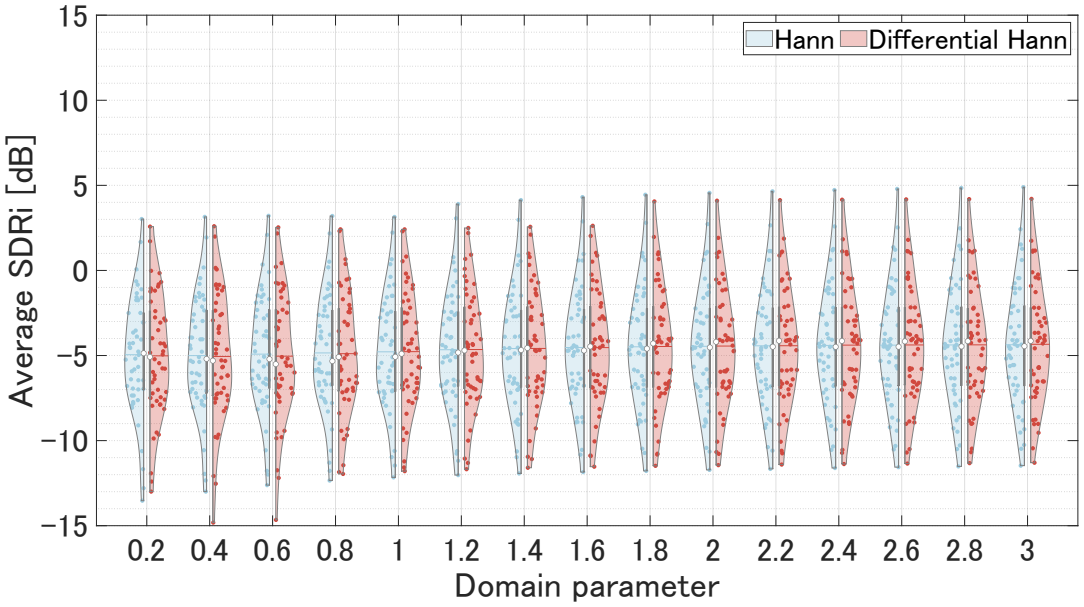


Fig. A.7: OHPSS experiment using dev dataset. (SDRi, window length = 1024 samples)

A.2　MHPSS

Figs. A.21–A.40 に 4.3.3 項のトレーニングデータに対する実験により得られた結果を示す．Figs. A.21–A.25 は SDR，Figs. A.26–A.30 は SDRi，Figs. A.31–A.35 は SIR，および Figs. A.36–A.40 は SAR を表している．

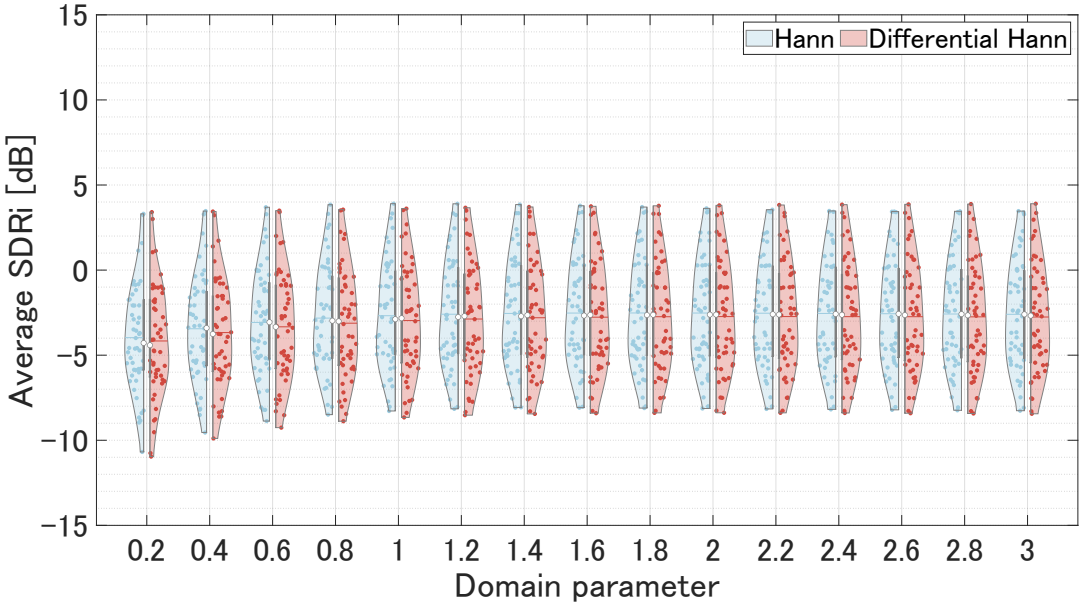


Fig. A.8: OHPSS experiment using dev dataset. (SDRi, window length = 2048 samples)

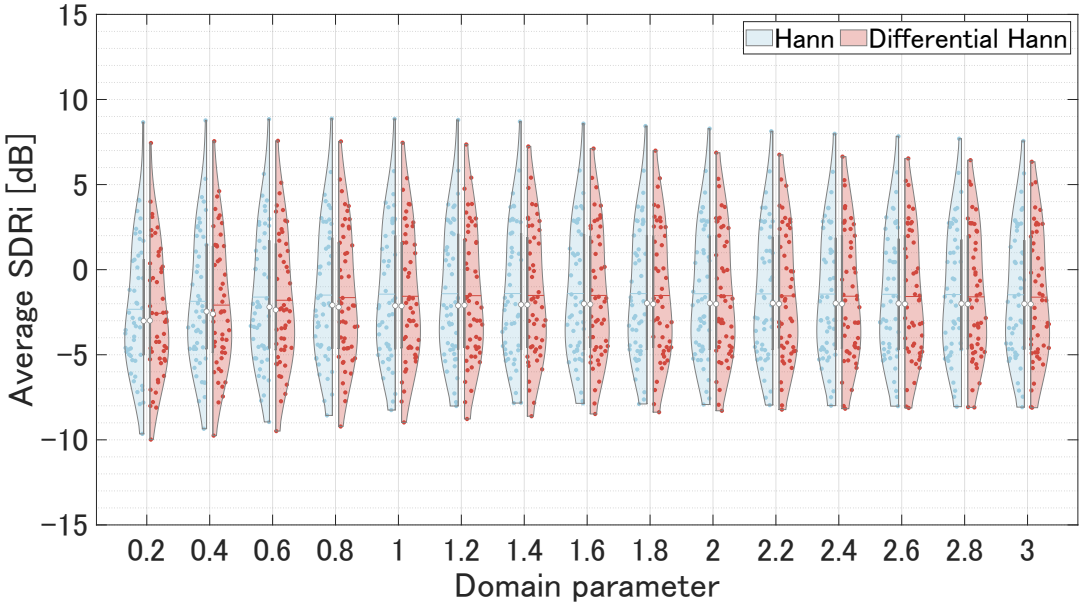


Fig. A.9: OHPSS experiment using dev dataset. (SDRi, window length = 4096 samples)

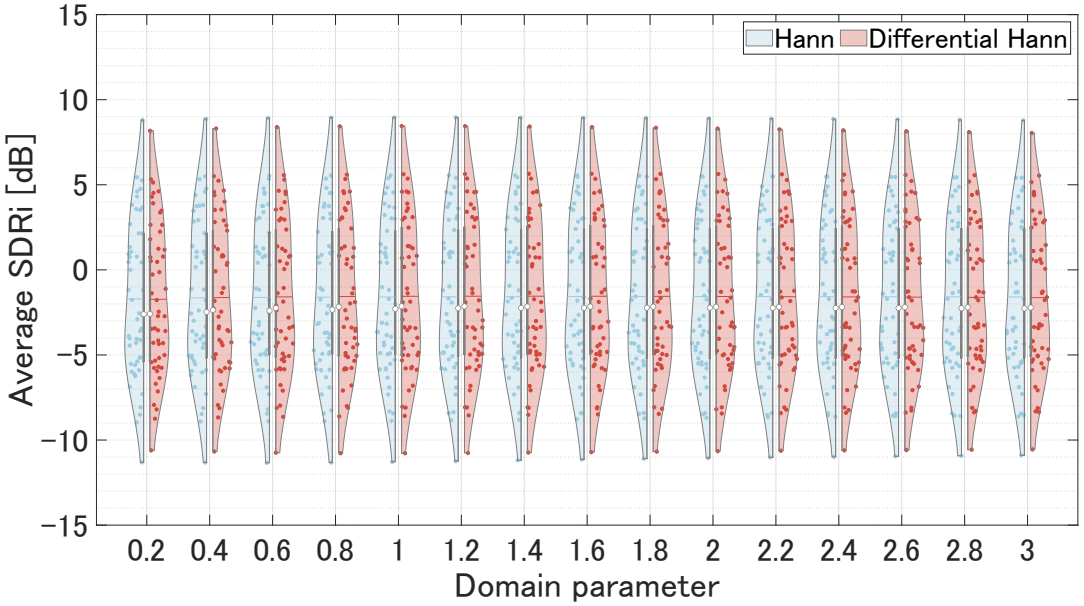


Fig. A.10: OHPSS experiment using dev dataset. (SDRi, window length = 8192 samples)

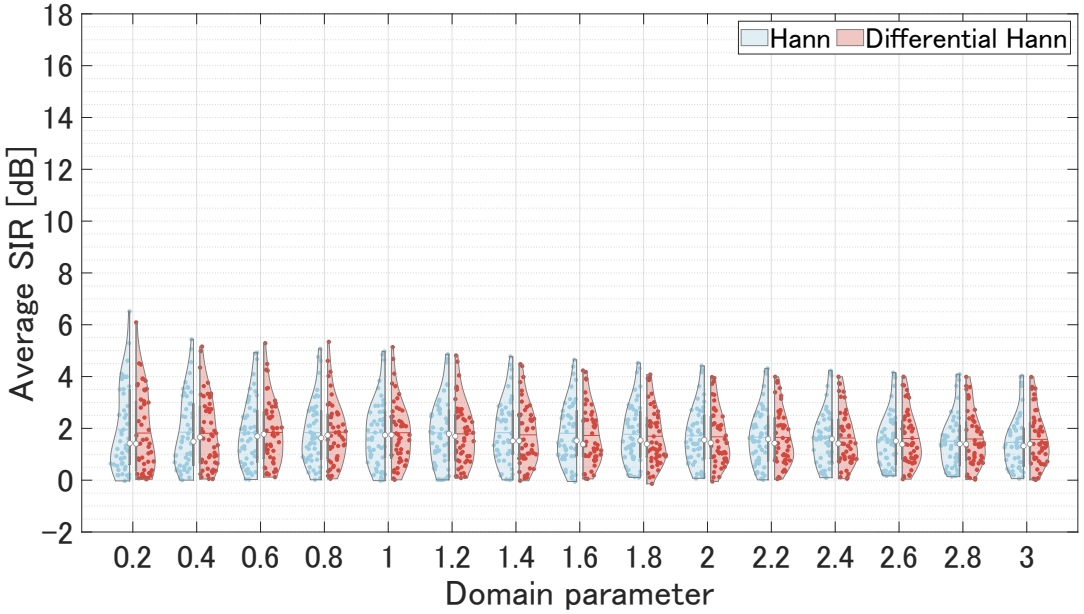


Fig. A.11: OHPSS experiment using dev dataset. (SIR, window length = 512 samples)

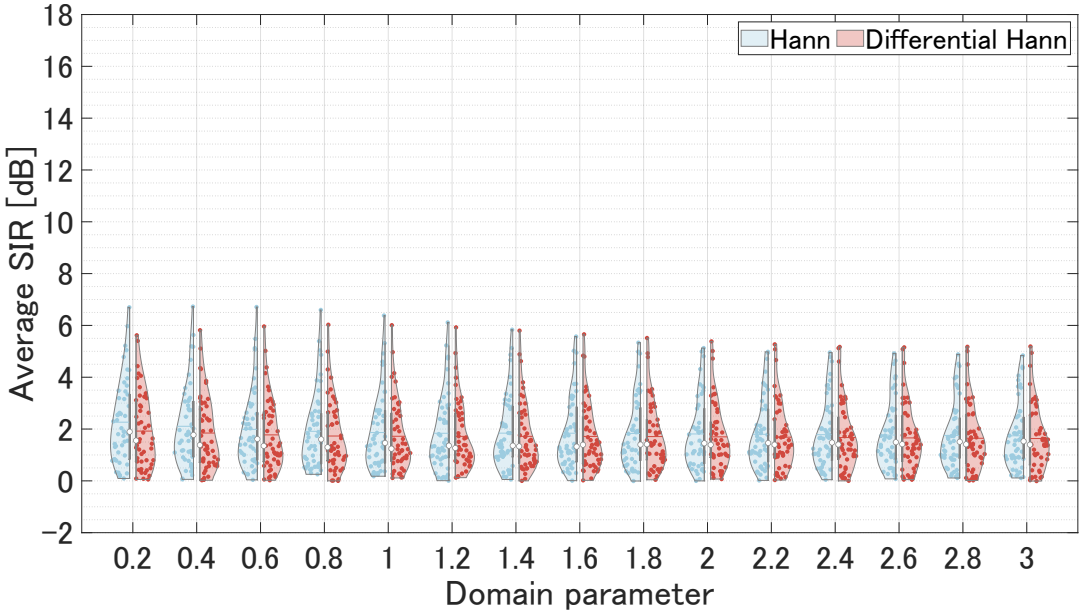


Fig. A.12: OHPSS experiment using dev dataset. (SIR, window length = 1024 samples)

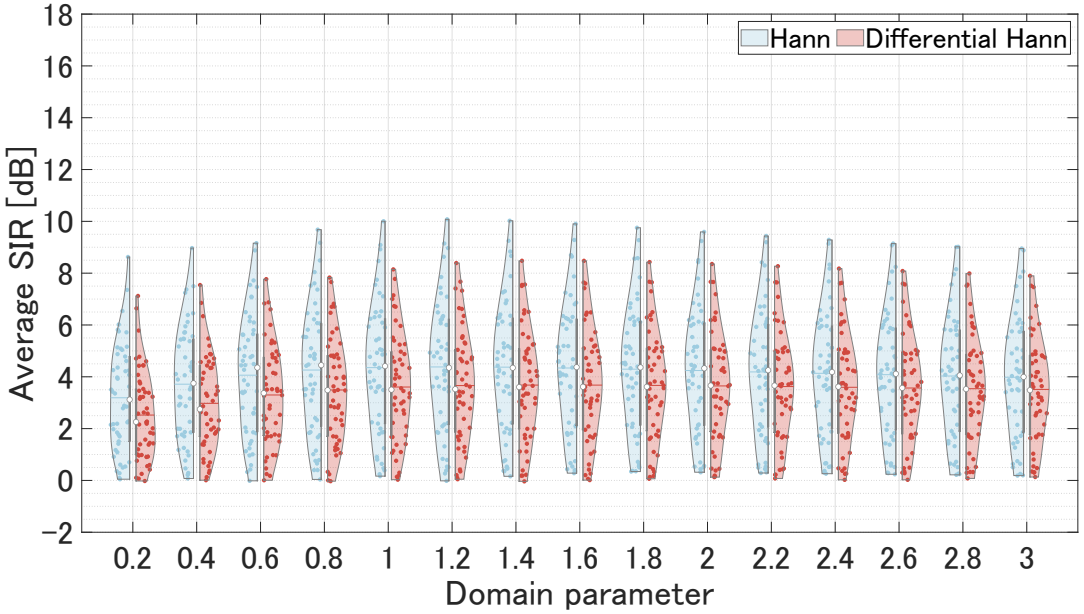


Fig. A.13: OHPSS experiment using dev dataset. (SIR, window length = 2048 samples)

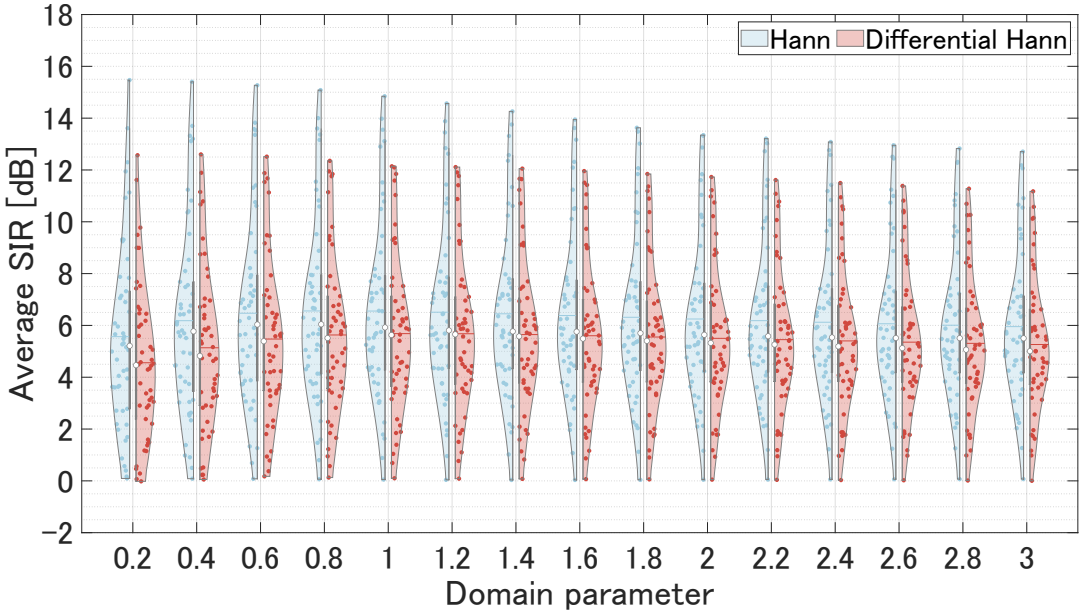


Fig. A.14: OHPSS experiment using dev dataset. (SIR, window length = 4096 samples)

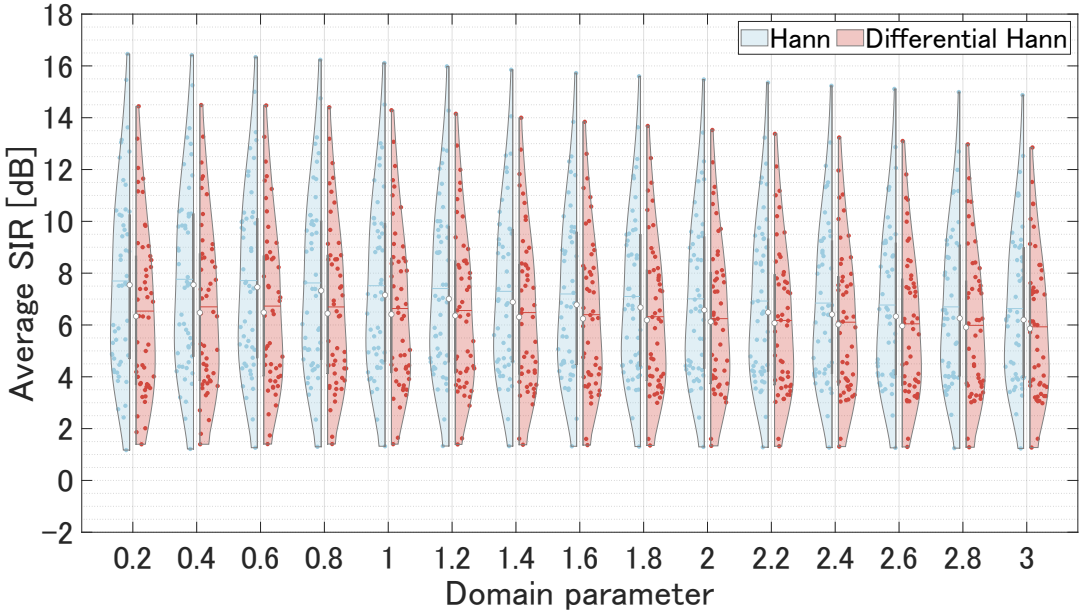


Fig. A.15: OHPSS experiment using dev dataset. (SIR, window length = 8192 samples)

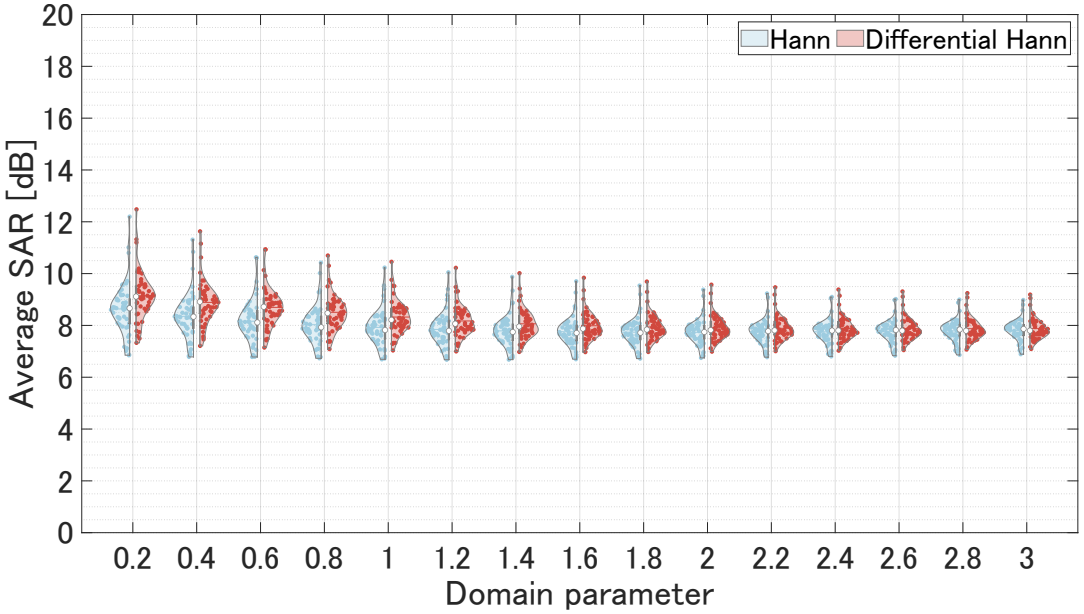


Fig. A.16: OHPSS experiment using dev dataset. (SAR, window length = 512 samples)

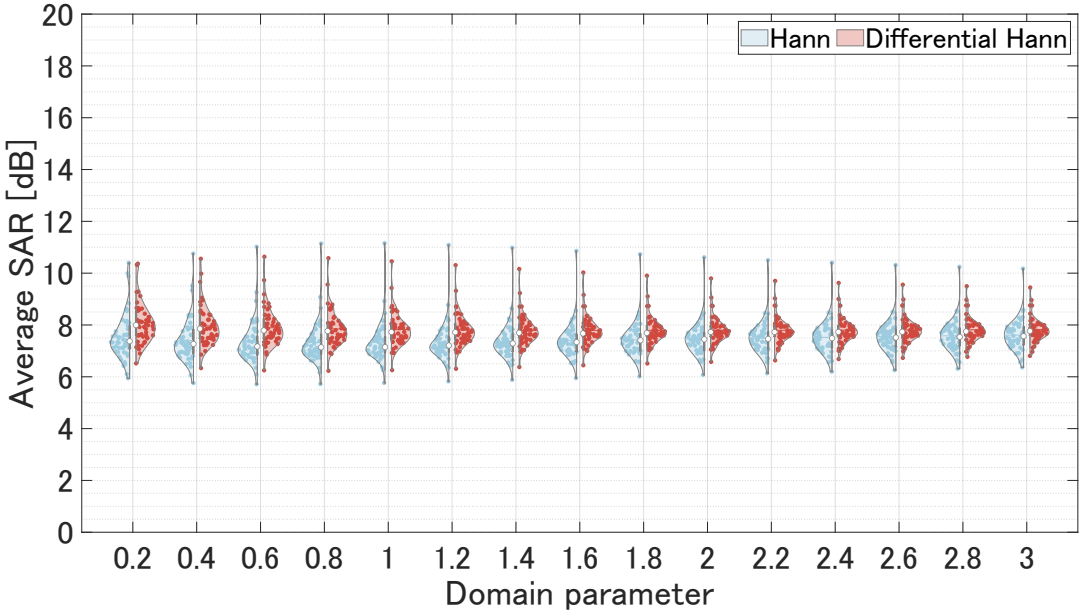


Fig. A.17: OHPSS experiment using dev dataset. (SAR, window length = 1024 samples)

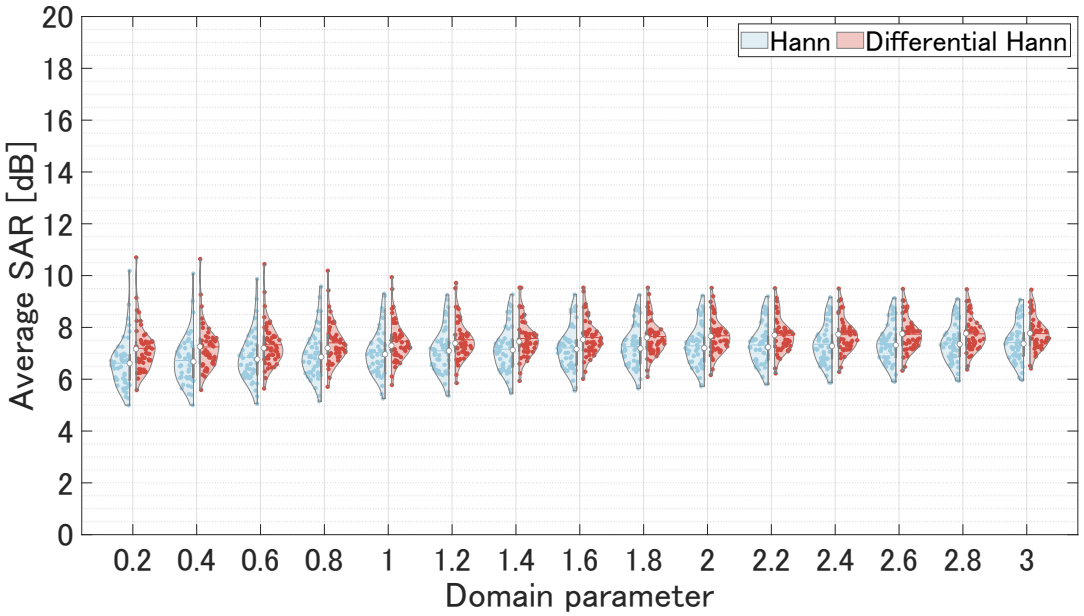


Fig. A.18: OHPSS experiment using dev dataset. (SAR, window length = 2048 samples)

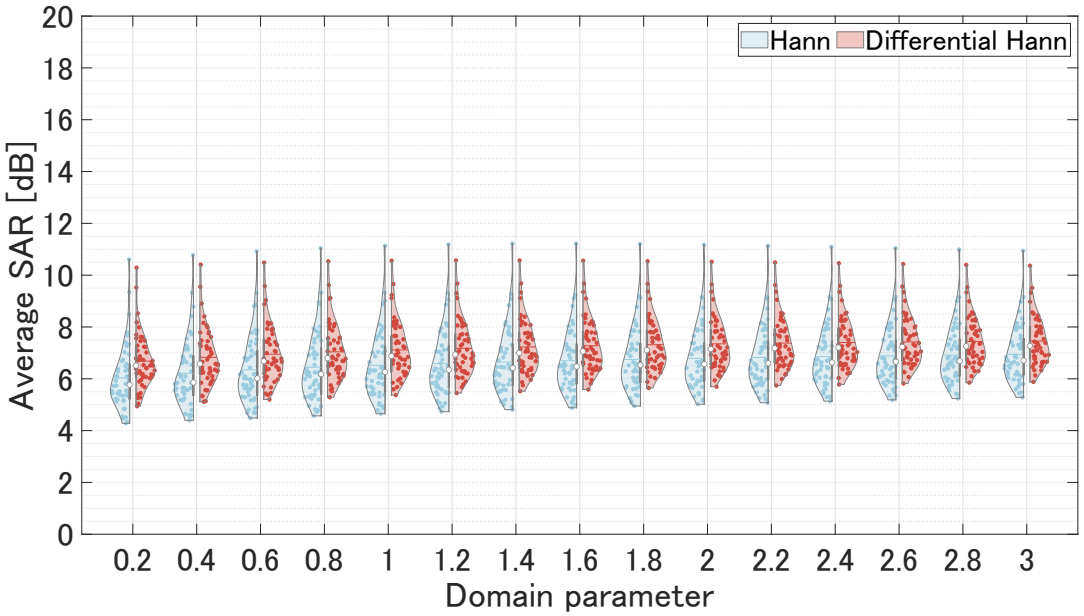


Fig. A.19: OHPSS experiment using dev dataset. (SAR, window length = 4096 samples)

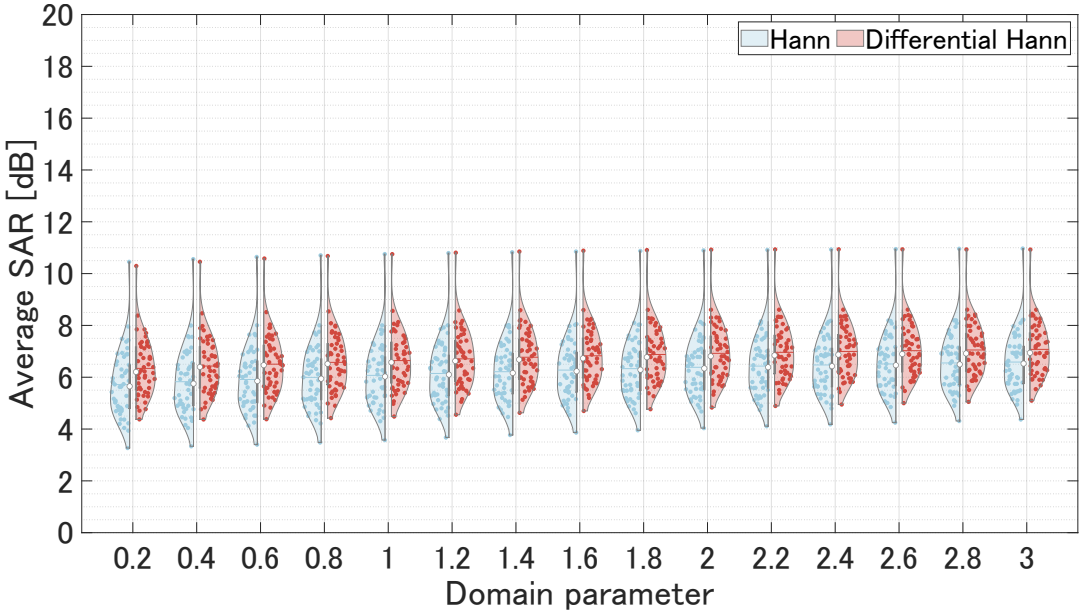


Fig. A.20: OHPSS experiment using dev dataset. (SAR, window length = 8192 samples)

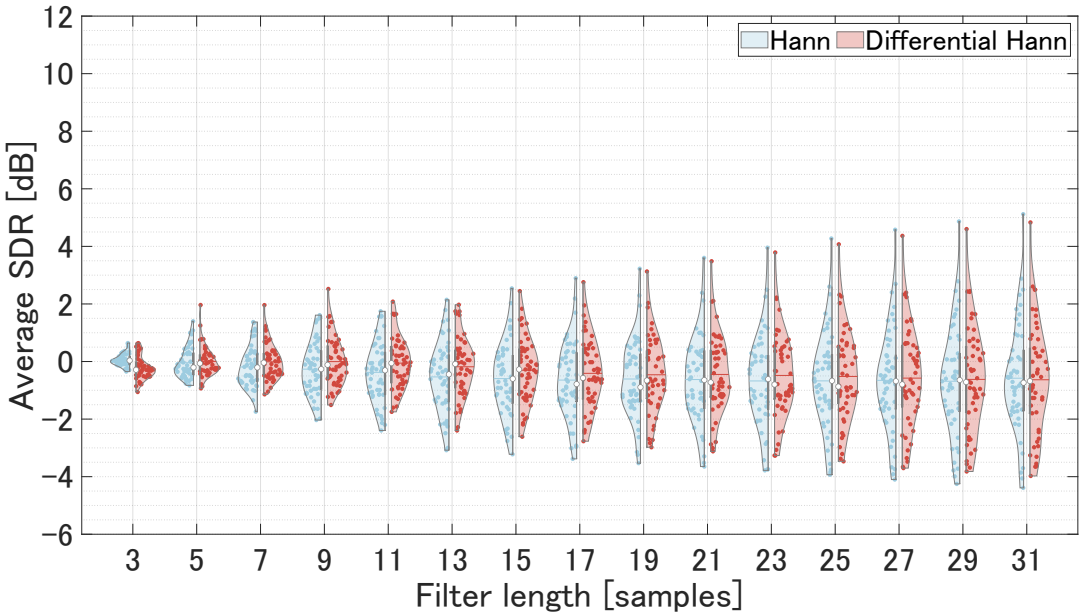


Fig. A.21: MHPSS experiment using dev dataset. (SDR, window length = 512 samples)

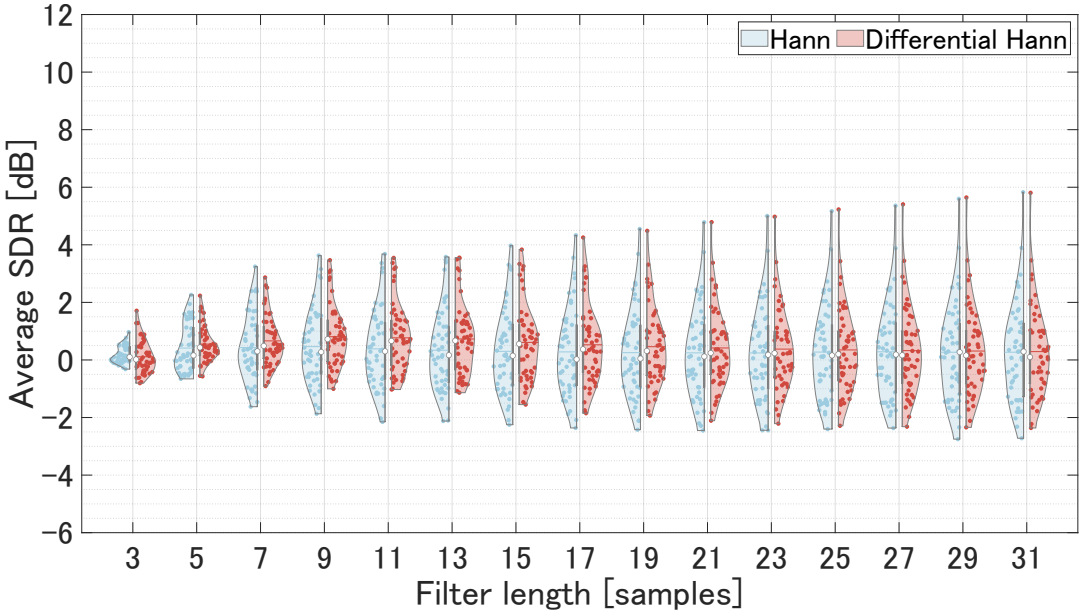


Fig. A.22: MHPSS experiment using dev dataset. (SDR, window length = 1024 samples)

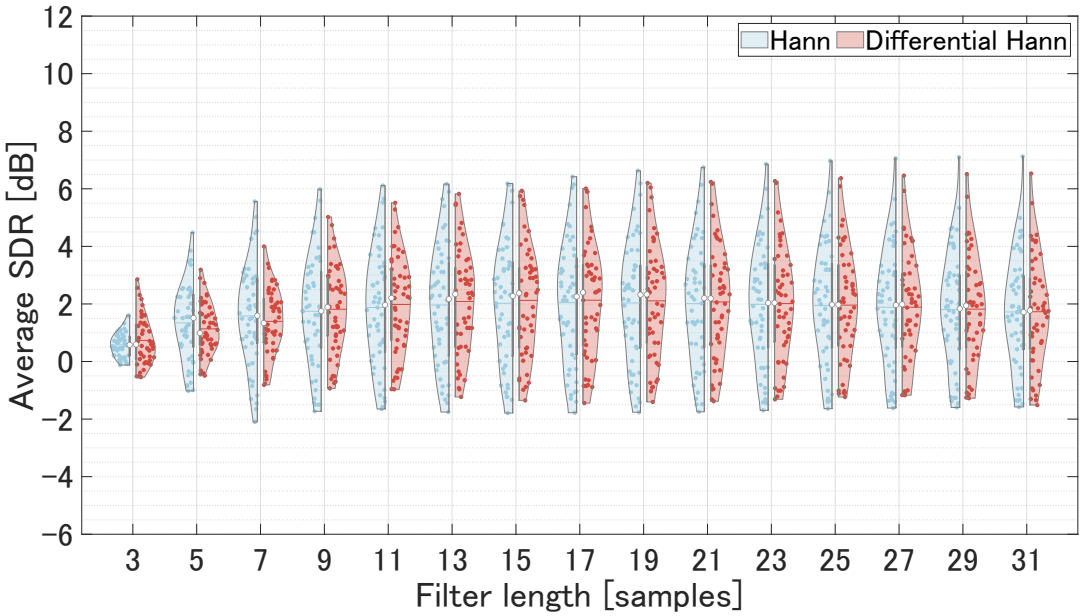


Fig. A.23: MHPSS experiment using dev dataset. (SDR, window length = 2048 samples)

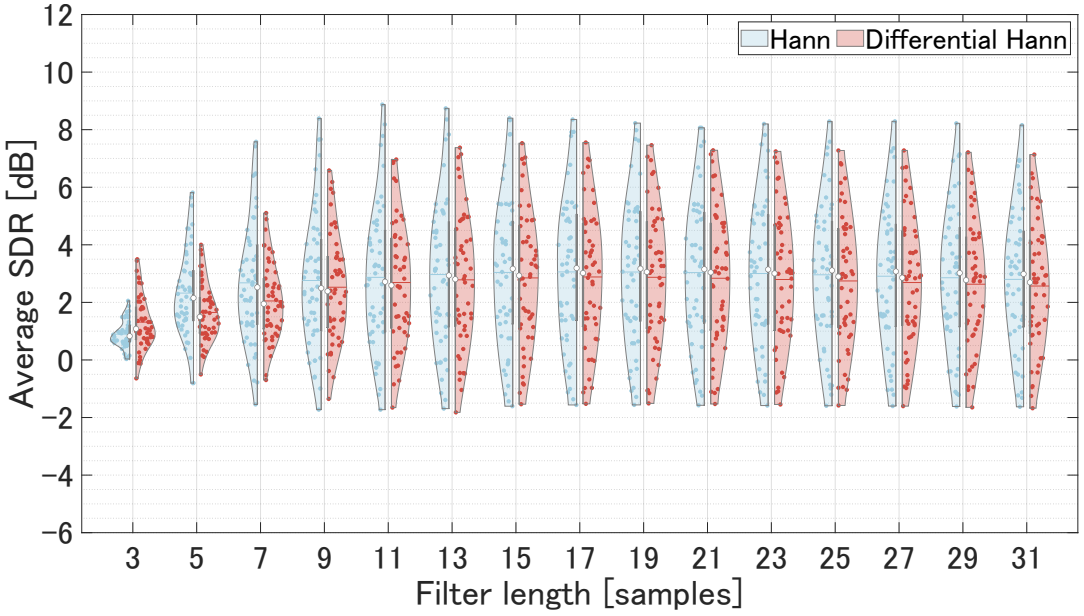


Fig. A.24: MHPSS experiment using dev dataset. (SDR, window length = 4096 samples)

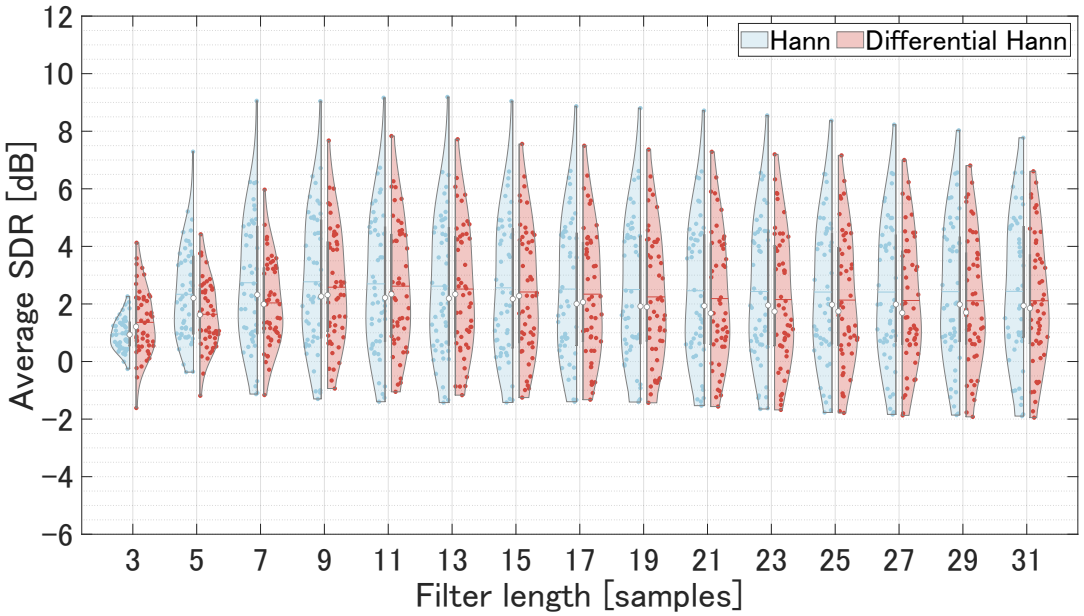


Fig. A.25: MHPSS experiment using dev dataset. (SDR, window length = 8192 samples)

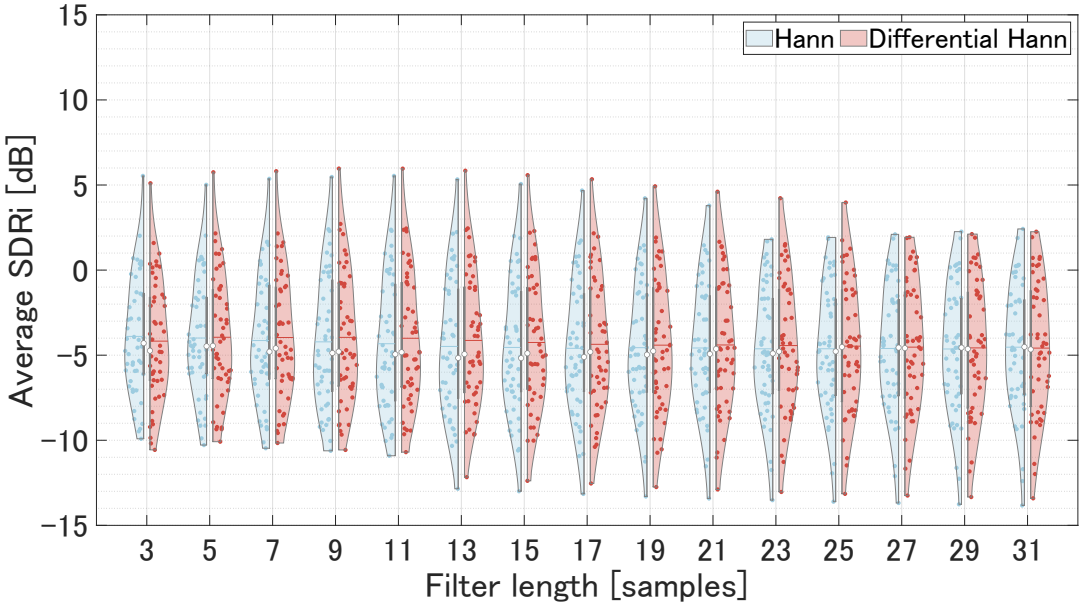


Fig. A.26: MHPSS experiment using dev dataset. (SDRi, window length = 512 samples)

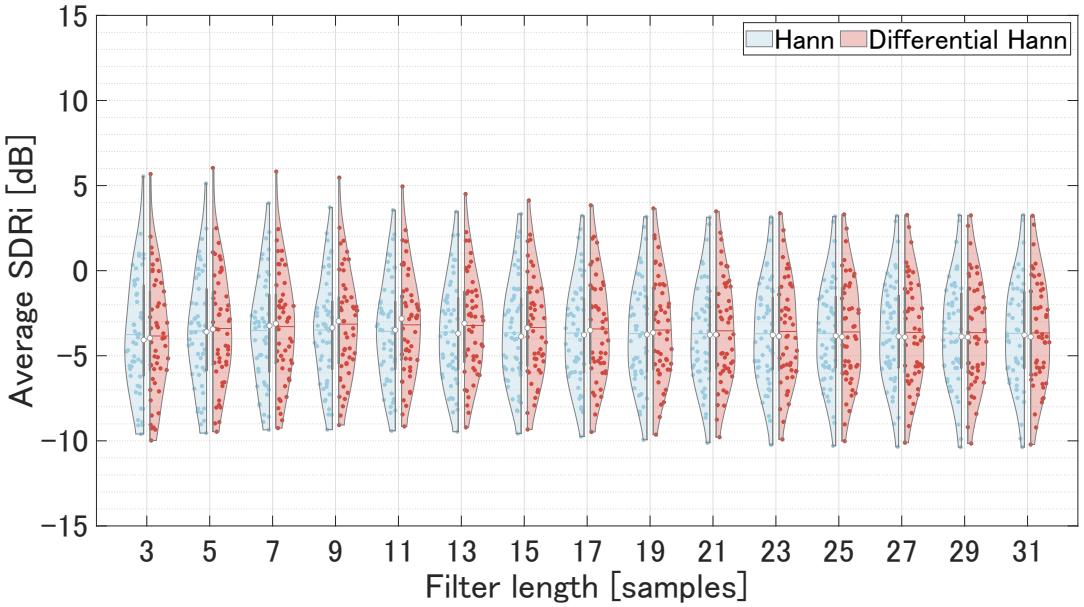


Fig. A.27: MHPSS experiment using dev dataset. (SDRi, window length = 1024 samples)

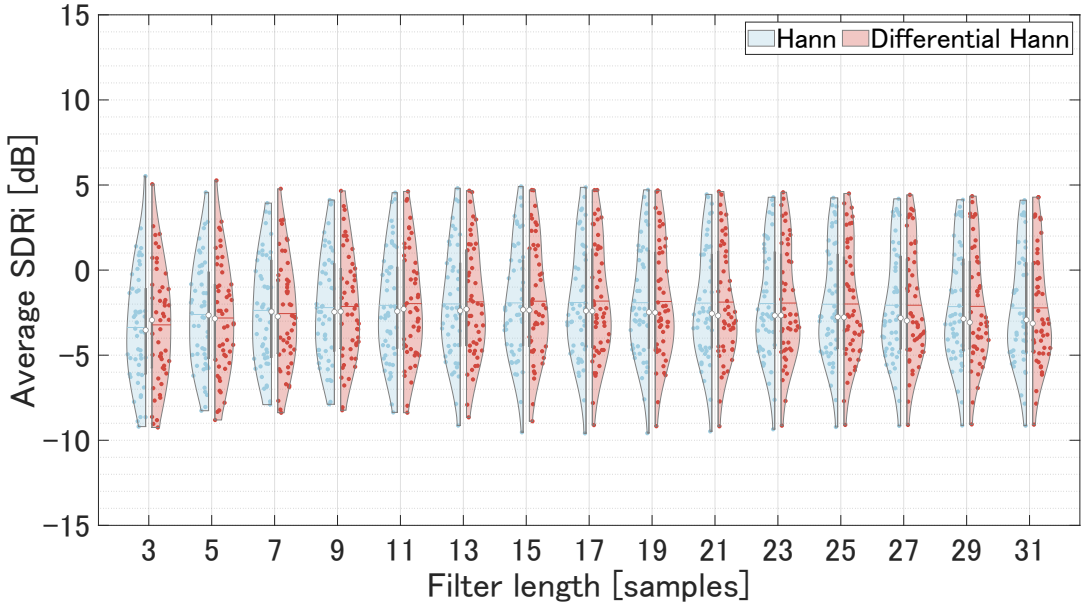


Fig. A.28: MHPSS experiment using dev dataset. (SDRi, window length = 2048 samples)

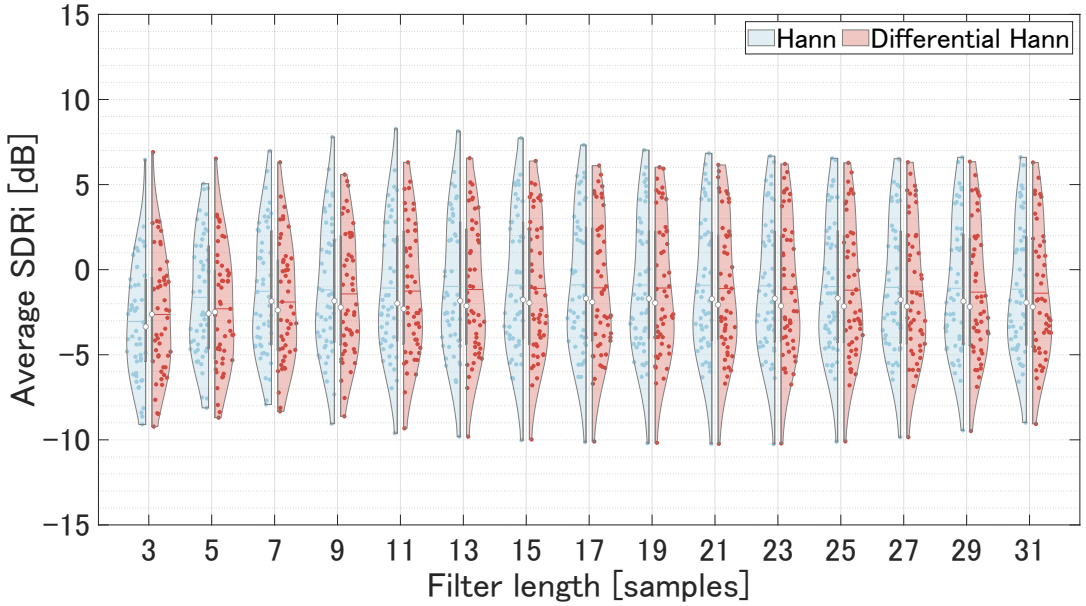


Fig. A.29: MHPSS experiment using dev dataset. (SDRi, window length = 4096 samples)

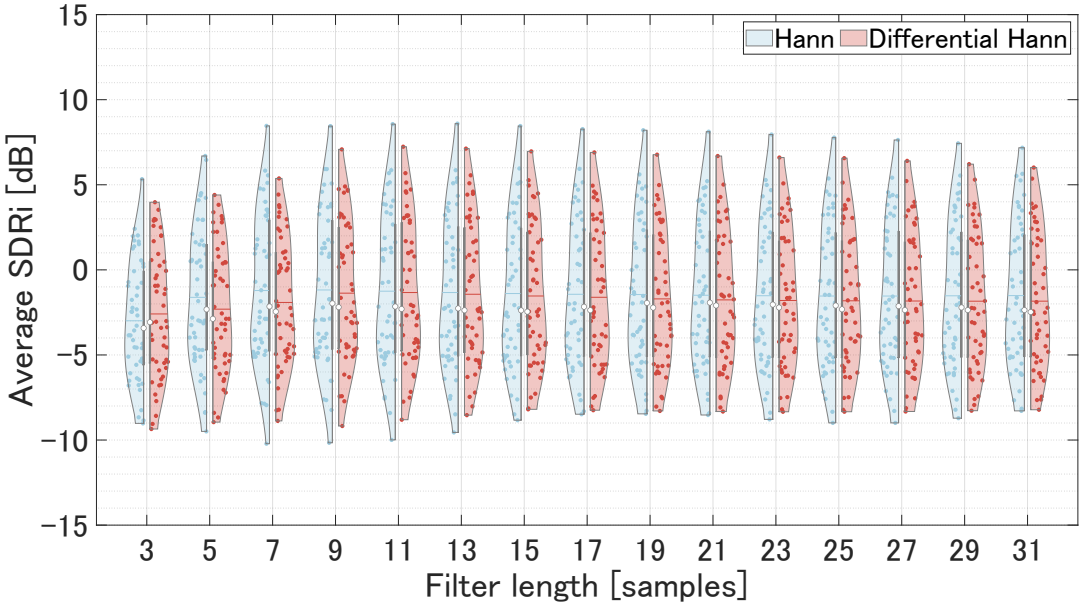


Fig. A.30: MHPSS experiment using dev dataset. (SDRi, window length = 8192 samples)

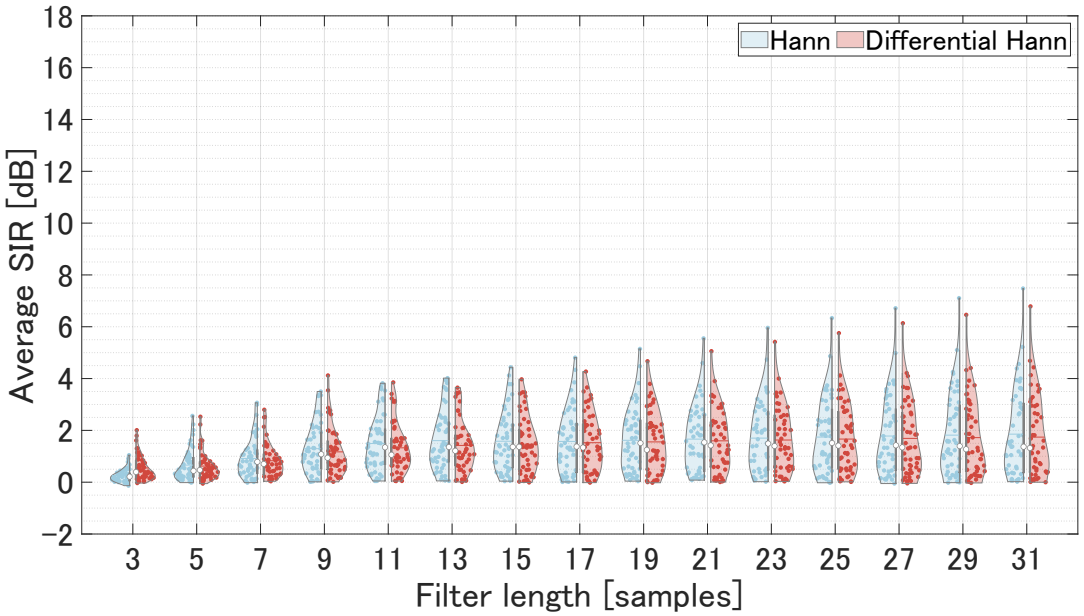


Fig. A.31: MHPSS experiment using dev dataset. (SIR, window length = 512 samples)

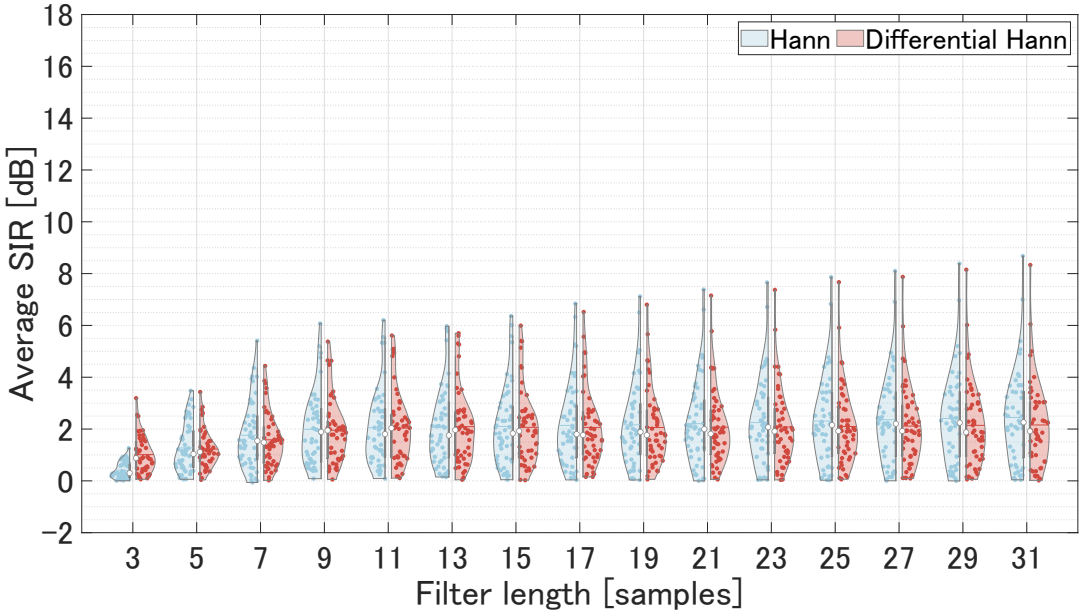


Fig. A.32: MHPSS experiment using dev dataset. (SIR, window length = 1024 samples)

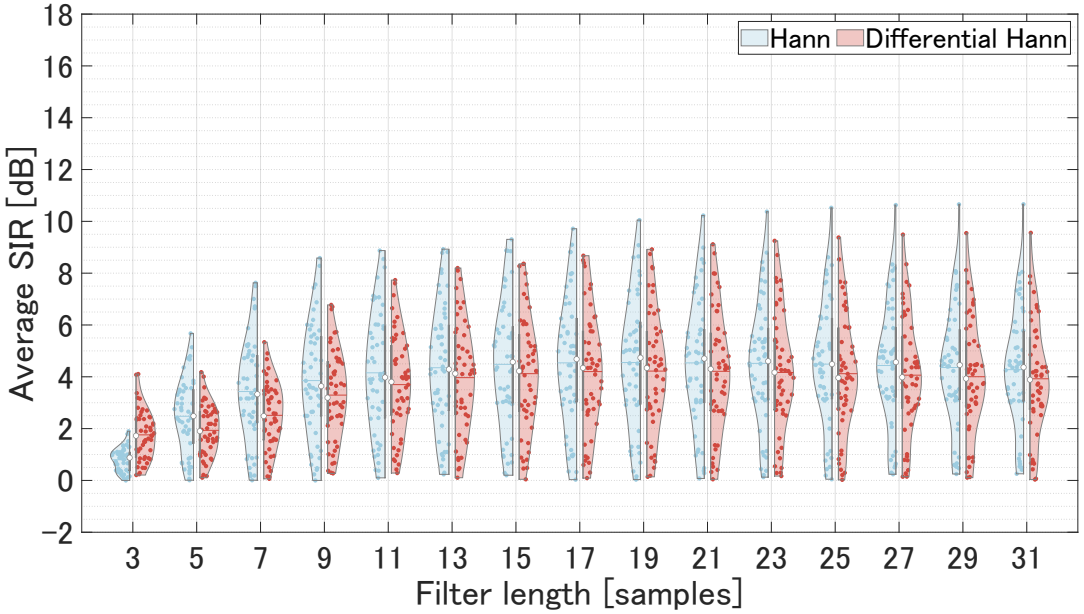


Fig. A.33: MHPSS experiment using dev dataset. (SIR, window length = 2048 samples)

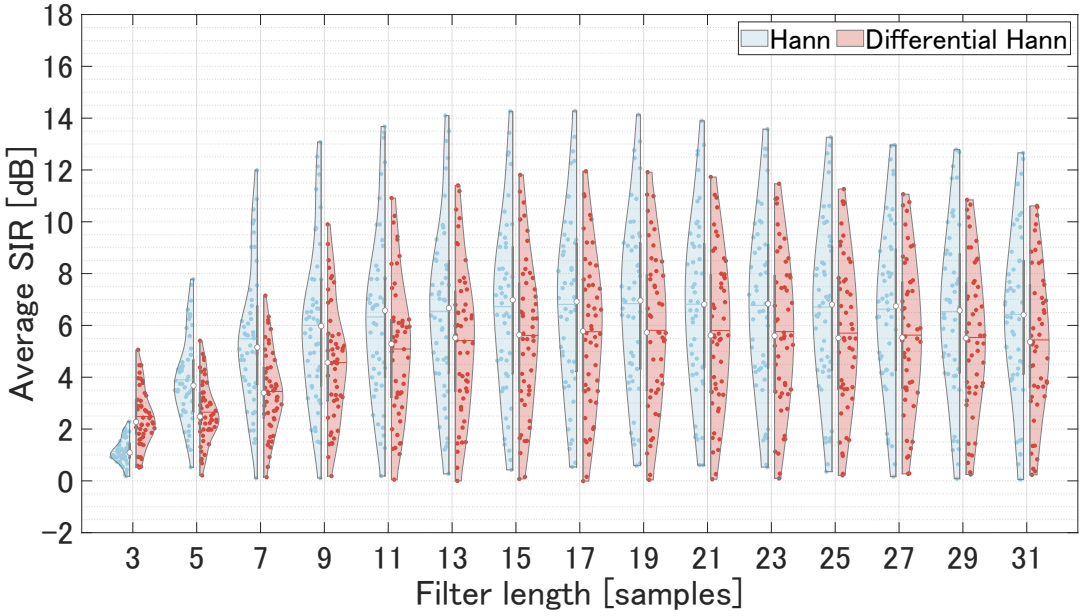


Fig. A.34: MHPSS experiment using dev dataset. (SIR, window length = 4096 samples)

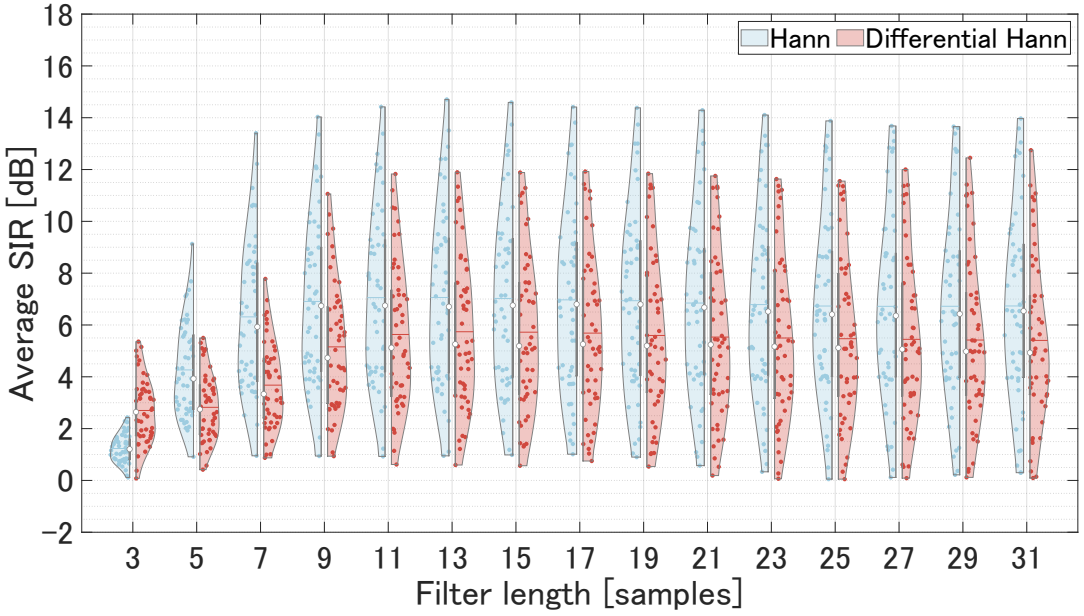


Fig. A.35: MHPSS experiment using dev dataset. (SIR, window length = 8192 samples)

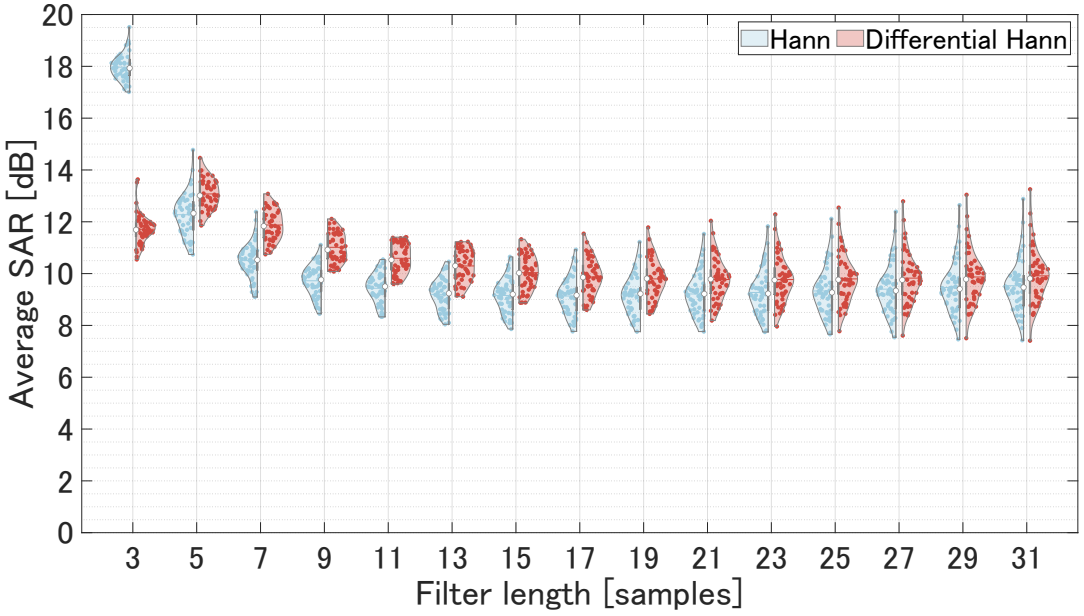


Fig. A.36: MHPSS experiment using dev dataset. (SAR, window length = 512 samples)

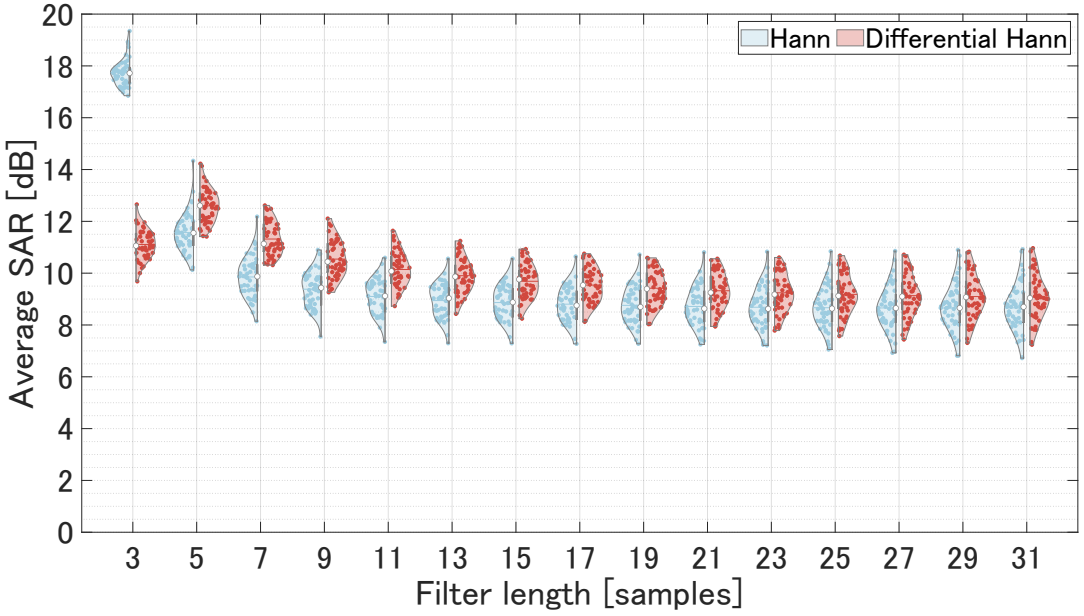


Fig. A.37: MHPSS experiment using dev dataset. (SAR, window length = 1024 samples)

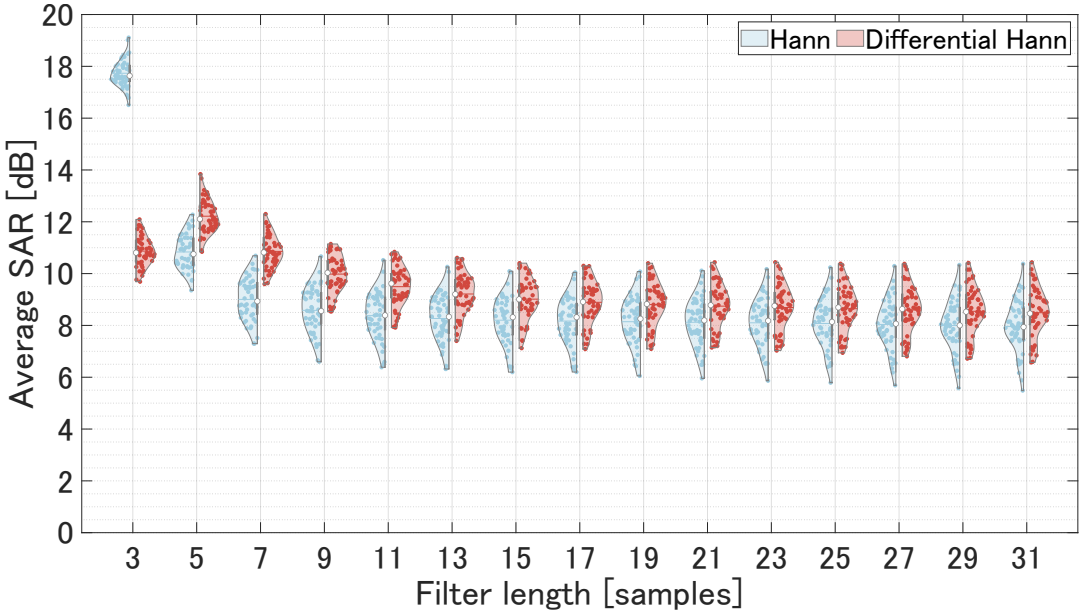


Fig. A.38: MHPSS experiment using dev dataset. (SAR, window length = 2048 samples)

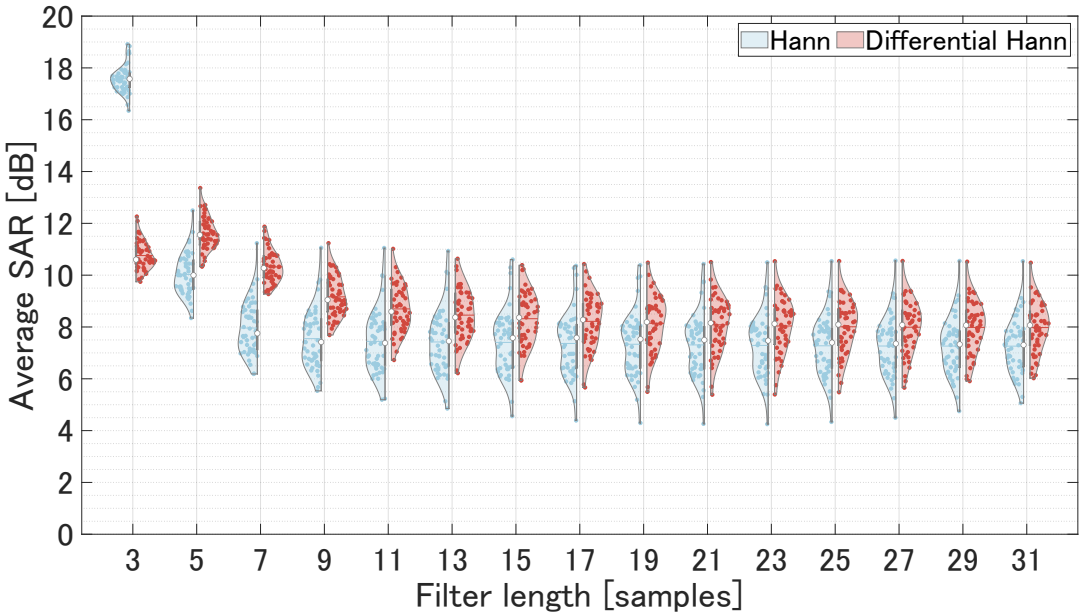


Fig. A.39: MHPSS experiment using dev dataset. (SAR, window length = 4096 samples)

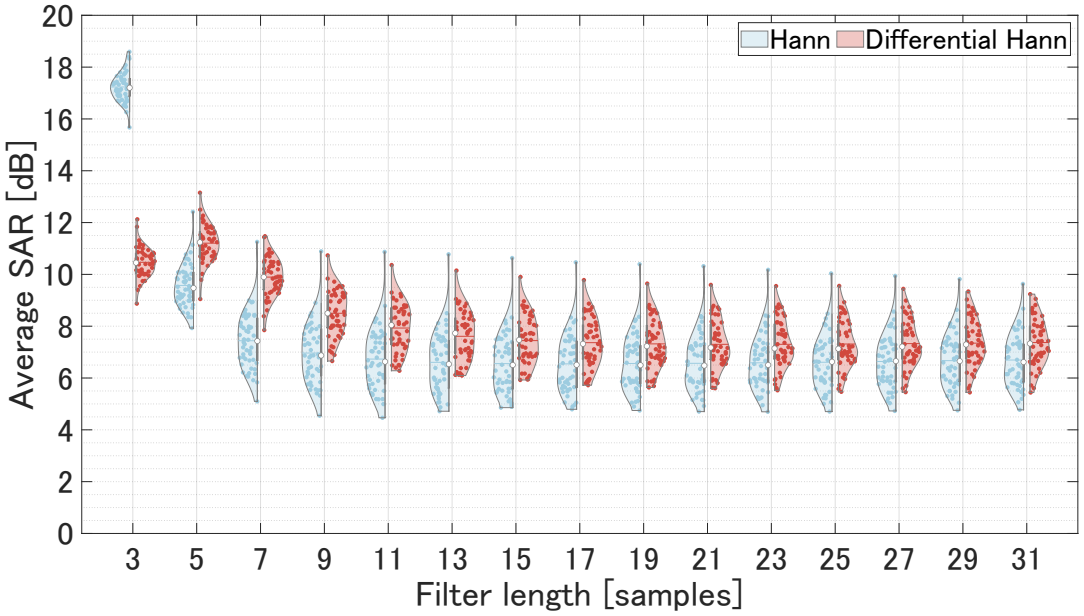


Fig. A.40: MHPSS experiment using dev dataset. (SAR, window length = 8192 samples)