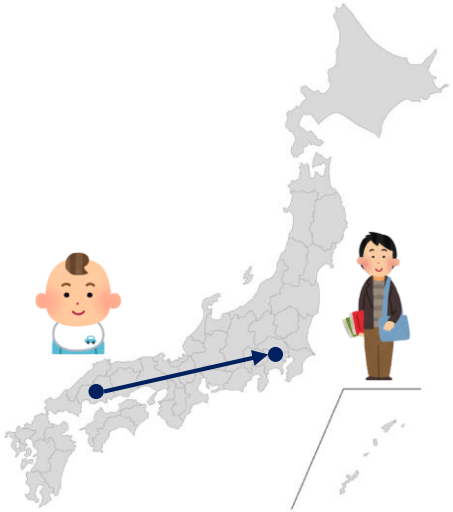
音楽音響信号に対するウェーブレット変換を用いた音源分離

中村　友彦
東京大学　大学院情報理工学系研究科
システム情報学専攻　猿渡・小山研究室　特任助教

自己紹介

- 名前
  - 中村 友彦

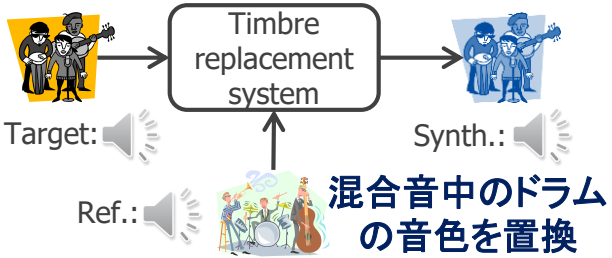
- 経歴
  - 2016/03：東京大学　大学院情報理工学系研究科
    システム情報学専攻　博士課程　修了
  - 2016/04 ~ 2019/08：セコム株式会社　IS研究所
  - 2019/09 ~ 現在：東京大学　大学院情報理工学系研究科
    システム情報学専攻　猿渡・小山研究室　特任助教



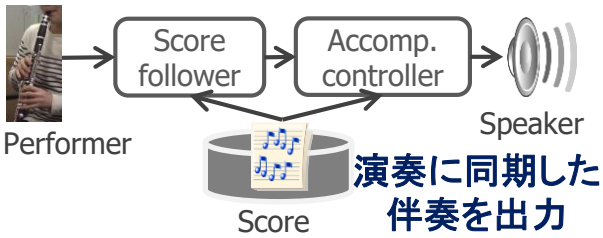
- 専門
  - 音楽・音響信号処理，音楽情報処理

音楽・音響信号処理に従事

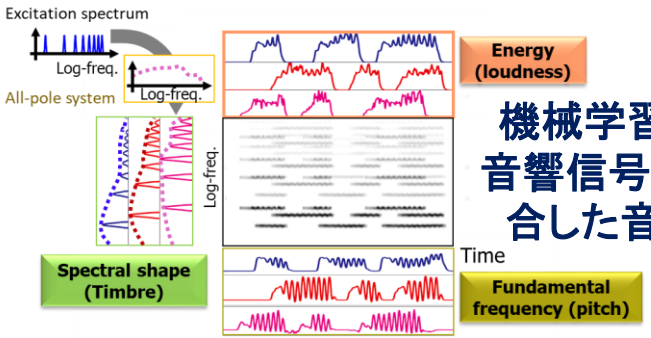
- 主にモノラル音響信号が対象
- 最近はマルチチャンネル音響信号処理も触り始めた.



ドラム音色置換システム [Nakamura+2014]



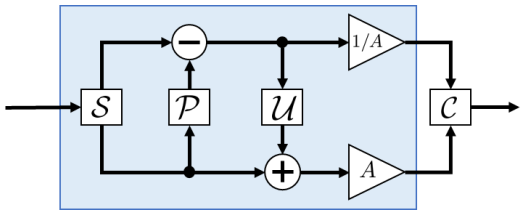
自動伴奏システム [Nakamura+2016]



機械学習・聴覚・音響信号処理を融合した音源分離

調波時間因子分解 [Nakamura+2020]

多重解像度解析にヒントを得たDNN音源分離



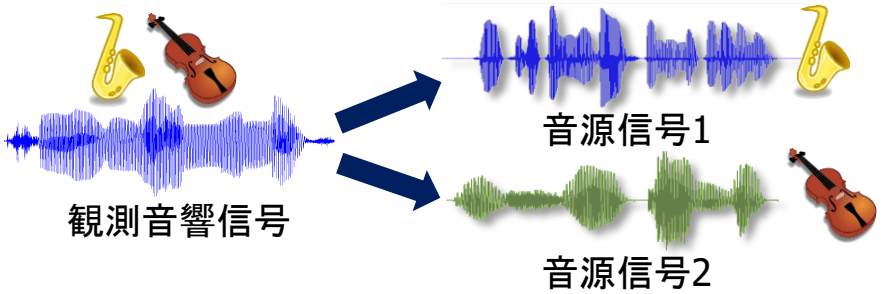
多重解像度深層分析 [Nakamura+2021]

Outline

- 自己紹介
- 音源分離とは？
- 調波時間因子分解：聴覚・機械学習・音響信号処理の知見を融合した音源分離手法
- 多重解像度深層分析：多重解像度解析にヒントを得た深層学習ベース音源分離手法
- まとめ

音源分離とは？

- 観測音響信号を各音源信号に分解する技術
  - 人間のもつ選択的に音を聴き分ける能力を計算機で模擬



- 様々な音メディア処理で有用
  - 実環境では妨害音が不可避に混入 ⇒ 音源分離により目的音を分離・強調
  - 音源毎に分離することで，各々の音源信号に対して別々の処理を適用可能
  - 音声認識，自動採譜，ユーザによる既存楽曲のリミックス，etc.

音声認識

自動採譜


音響信号加工

音源分離の難しさ

- 音源分離は不良設定問題
  - 観測音響信号を説明可能な音源信号の組み合わせは無数に存在

解を限定するために何らかの手がかりが必要

- マルチチャネルの場合は，空間的な情報（e.g., 音源の到来時間差）が利用可能
- モノラルや実質的に空間的な情報が利用できない場合は？

いかに手がかりを得るか？

- 方針1：分離対象の性質を数理モデルとして表現
  - 対象の性質を反映したモデルを構築し，データをよりよく説明するようなパラメータを推定 ⇒ 教師なしで分離も可能

聴覚・機械学習・音響信号処理の知見を融合した音源分離手法（調波時間因子分解） [Nakamura+2020]

- 方針2：大量に学習データが用意可能な場合
  - 深層ニューラルネットワーク（DNN）などを用いて，データ駆動で手がかりを抽出し利用

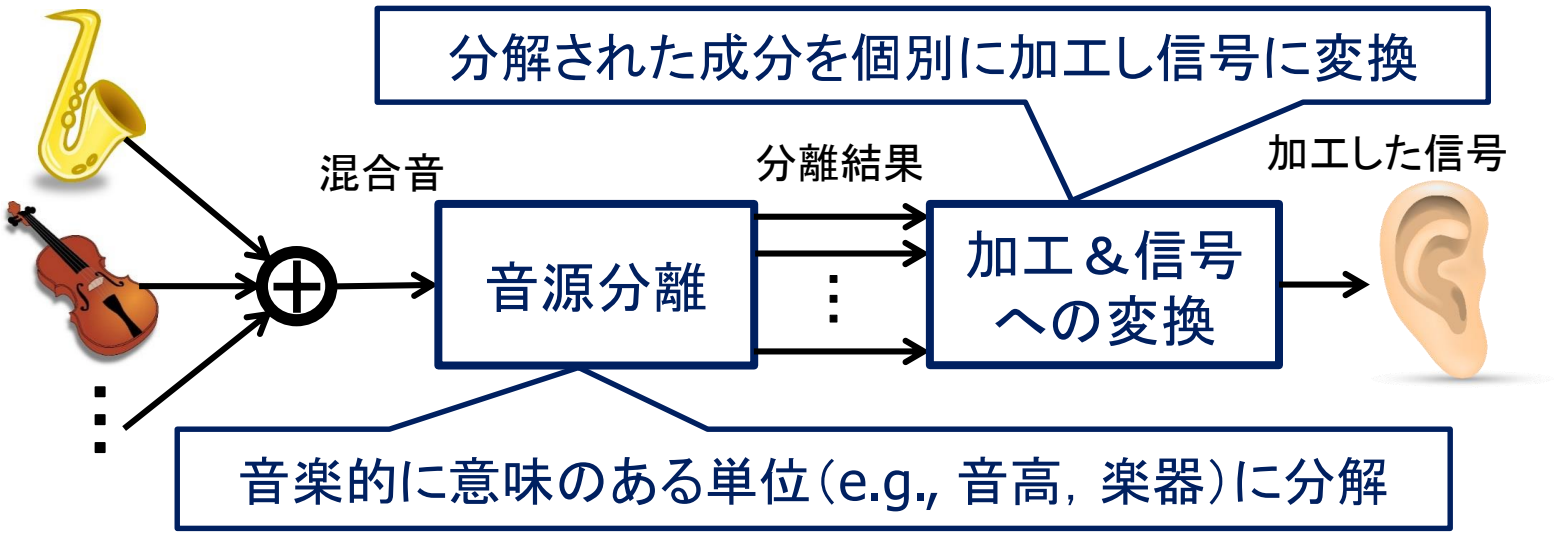
多重解像度解析にヒントを得た深層学習ベース音源分離手法（多重解像度深層分析） [Nakamura+2021]

Outline

- 自己紹介
- 音源分離とは？
- 調波時間因子分解：聴覚・機械学習・音響信号処理の知見を融合した音源分離手法
- 多重解像度深層分析：多重解像度解析にヒントを得た深層学習ベース音源分離手法
- まとめ

音源分離と音楽音響信号加工

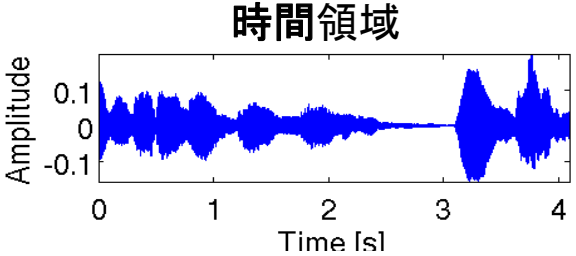
- 音源分離を用いた音楽音響信号加工システム
  - 分離性能が加工後の音質に直結 ⇒ 高精度な分離技術が必要

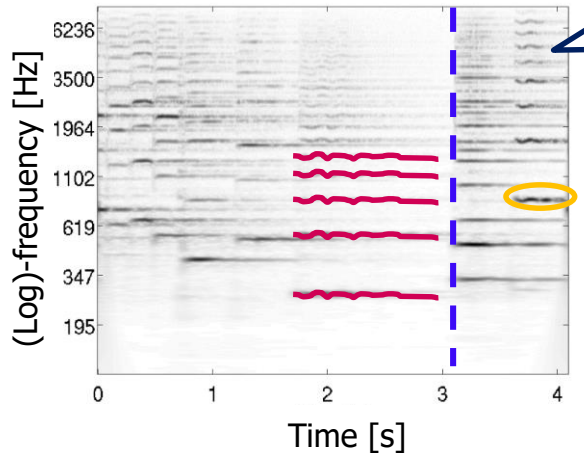


- 課題：音楽的要素（音高，リズム，楽器）が時々刻々と変化する音響信号をいかに分離するか？

音楽音響信号の分離に有用な手がかり

- 時間周波数領域では様々な手がかりが利用可能

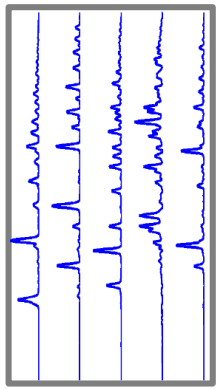




局所的な構造
- 調波性
- 基本周波数（F_0）やパワーの連続性
- 調波成分の共通の立ち上がり
- 調波成分の振幅や周波数の共通変化

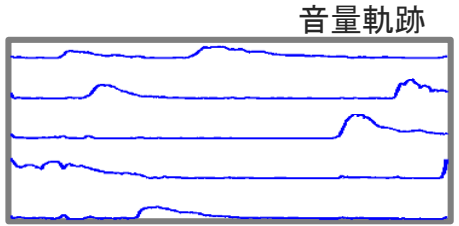
大局的な構造
- 限られた種類のスペクトルが繰り返し出現

≈



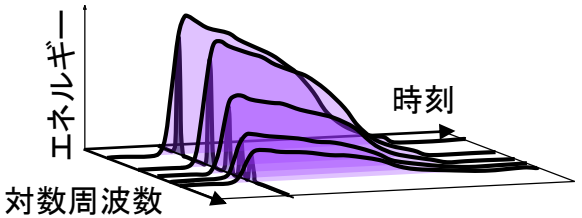
繰り返し出現する
スペクトルパターン

×

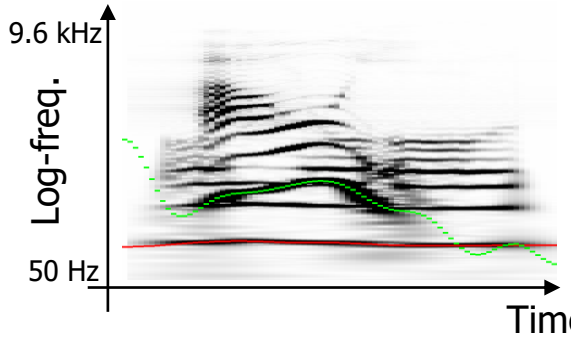


スペクトログラムの局所的構造に基づくアプローチ

- 調波時間構造化クラスタリング（HTC） [Kameoka+2005]
  - 人間は音を聴き分ける際に，特定の条件を満たす時間周波数成分をひとまとまりの音（音脈）として捉える
  - スペクトログラムの局所的制約として表現し，音脈の数理的なモデルを提案
    - 調波性
    - 調波成分の同時性
    - 調波成分の振幅包絡の同期性

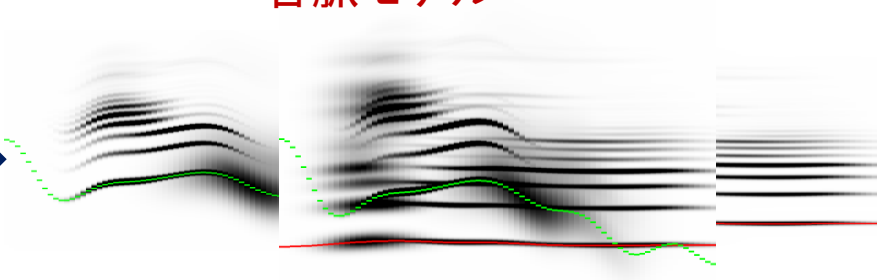


音脈モデル



観測スペクトログラム

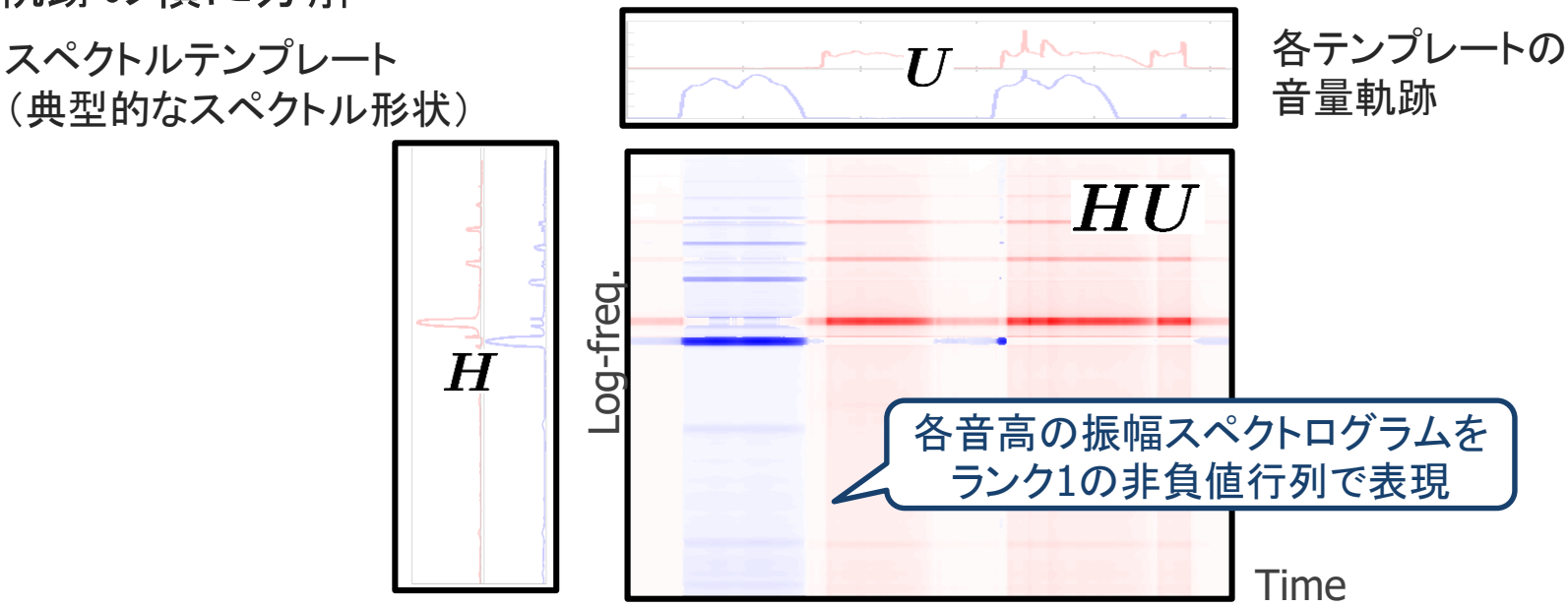
音源数分の音脈モデルをfitting



スペクトログラムモデル

スペクトログラムの大局的構造に基づくアプローチ

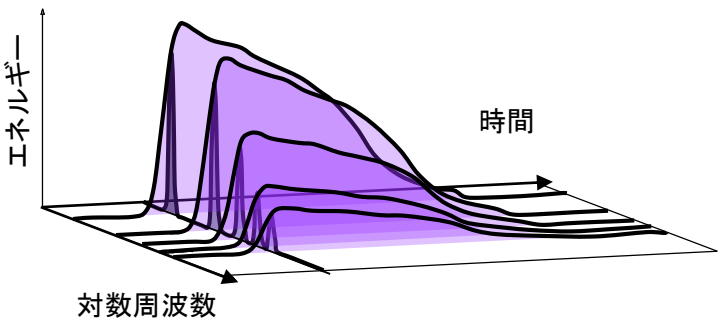
- 非負値行列因子分解（NMF）[Lee+1999, Smaragdis+2003]
  - 少数の限られた音高が繰り返し出現
  - 観測振幅スペクトログラムを非負値行列とみなし，スペクトルテンプレートと音量軌跡の積に分解



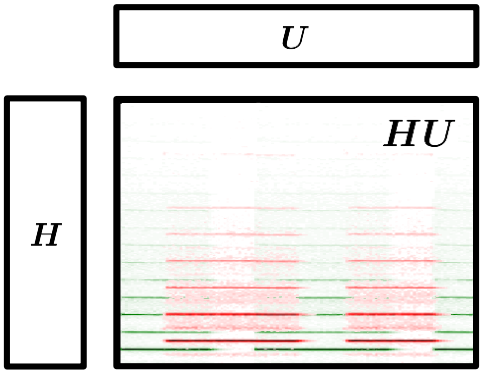
調波時間因子分解(HTFD)[Nakamura+2020]

- HTCとNMFは，それぞれ異なる手がかりを用いて分離
  ⇒ 高精度な音源分離の実現には両者ともに重要

HTC：スペクトログラムの局所的構造に着目



NMF：スペクトログラムの大局的構造に着目



- 両者の手がかりは互いに矛盾しないことに着眼し，HTCとNMFを融合したスペクトログラムの確率的生成モデルを構築

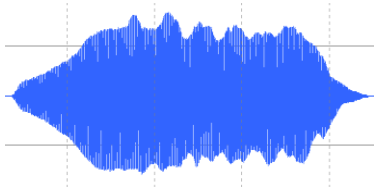
調波楽器音の時間信号モデルとCWT(1/2)

- 時間信号モデル：疑似周期信号
  - 局所的には周期的，滑らかに周期・振幅が変化
  - 調波性，調波成分の同期的変化を保証

$$f_k(u) = \sum_{n=1}^{N} a_{k,n}(u) e^{j(n\theta_k(u)+\varphi_{k,n})}$$

瞬時複素振幅

調波成分の瞬時位相
（基本周波数（F_0）成分に対応）



k：音源index

n：調波index

- 疑似周期信号の連続ウェーブレット変換（CWT）
  - CWTは対数周波数解像度をもち，音高の分布に適合

時刻

対数周波数

$$W_k(x,t) \triangleq \left\langle f_k(u), \psi_{x,t}(u) \right\rangle_{u\in(-\infty,\infty)}$$

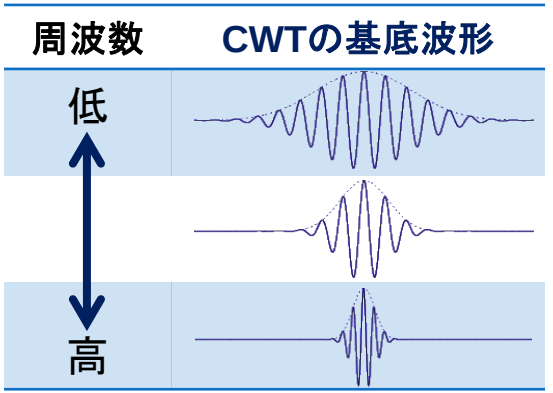
CWTの基底波形

$$= \left\langle F_k(\omega), \Psi_{x,t}(\omega) \right\rangle_{\omega\in(-\infty,\infty)}$$

角周波数

Parseval's theorem

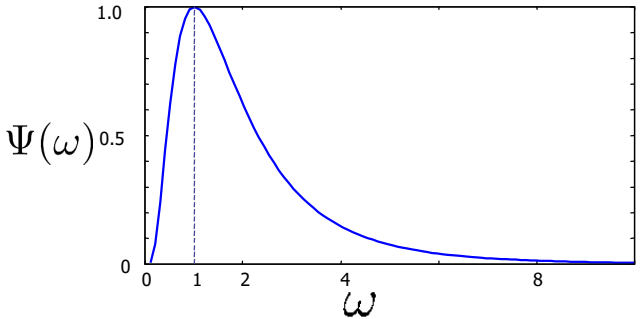
$f_k(u)$のFourier変換　CWTの基底波形のFourier変換



調波楽器音の時間信号モデルとCWT(2/2)

- 対数正規分布型のウェーブレット$\Psi(\omega)$を導入

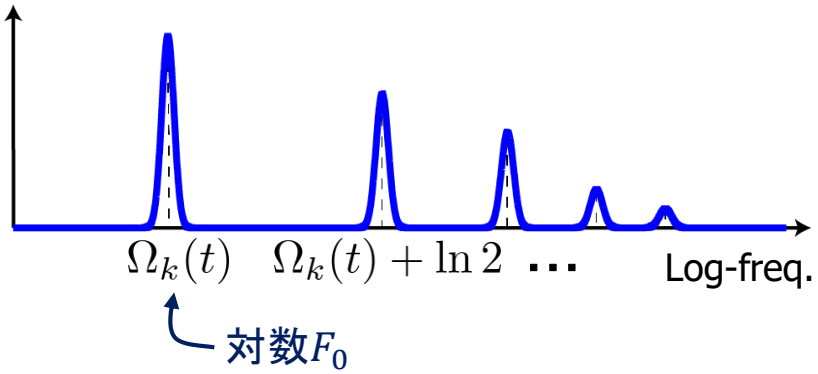
$$\Psi(\omega) = \begin{cases} \exp\left[-\dfrac{(\log\omega)^2}{2\sigma^2}\right] & (\omega > 0) \\ 0 & (\omega \le 0) \end{cases}$$



- 疑似周期信号の振幅CWTは，近似的に対数周波数軸上で混合Gaussモデル(GMM)型の関数により表現可能

対数周波数　時刻　対数F_0

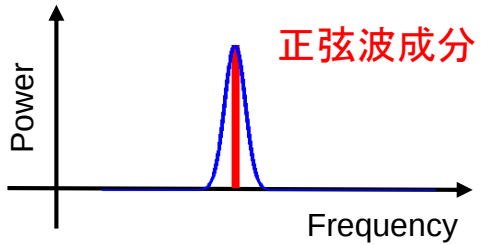
$$|W_k(x,t)| \simeq \sum_n |a_{k,n}(t)| e^{-\frac{(x-\Omega_k(t)-\ln n)^2}{2\sigma^2}}$$



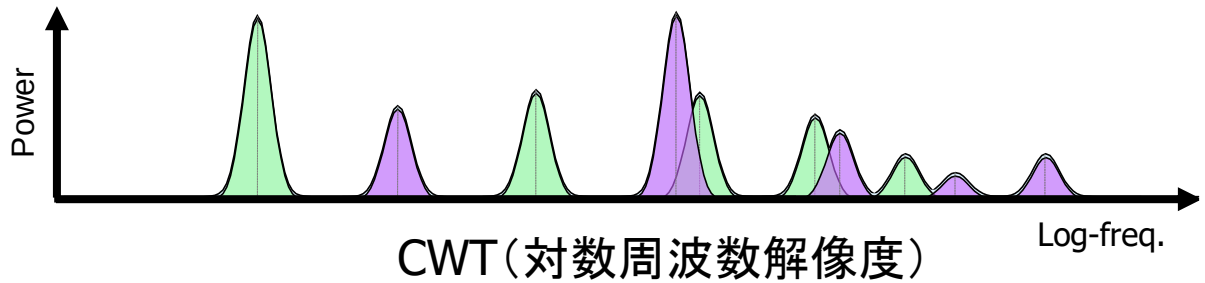
- HTCの音脈モデルと同型
- Gauss関数がスペクトル漏れに相当

音源分離の手がかりとしてのスペクトル漏れ

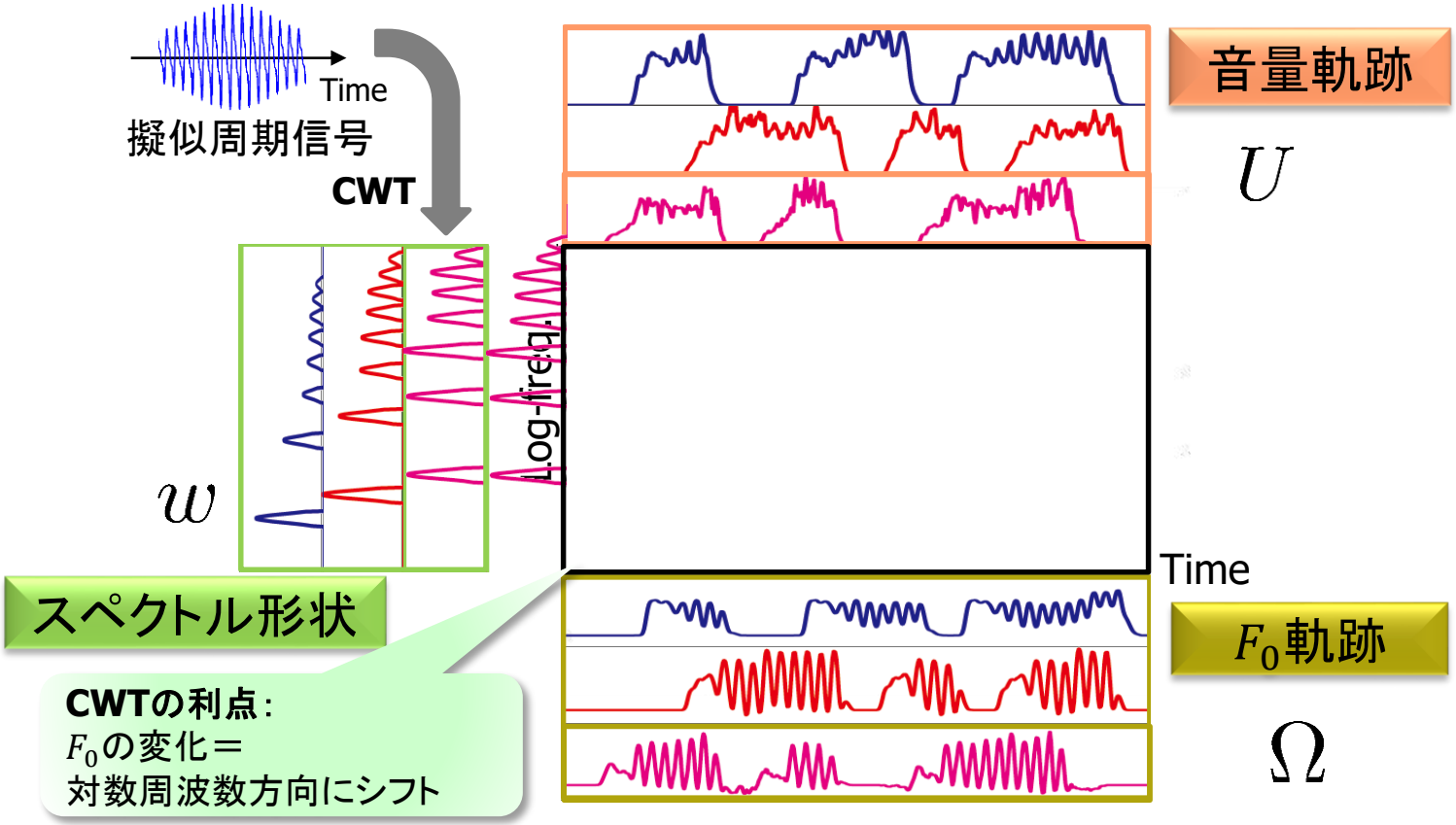
- スペクトル漏れ
  - 基底波形の影響で周波数方向に
    エネルギーが拡散



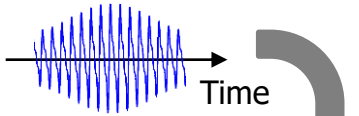
- スペクトル漏れが楽音のスペクトル形状を制限
  ⇒ 異なる音源の周波数が近接した場合，スペクトル漏れが分離の手がかりとなりうる．



HTFDのスペクトログラムモデル

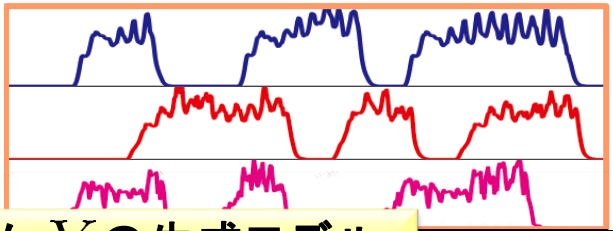


HTFDのスペクトログラムモデル



擬似周期信号

CWT



音量軌跡

U

観測スペクトログラム Y の生成モデル

$$Y_{l,m} \sim \mathrm{Poisson}\left(Y_{l,m}; \sum_{k}\left(\sum_{n} w_{k,n} e^{-\frac{(x_l - \Omega_{k,m} - \ln n)^2}{2\sigma^2}}\right) \times U_{k,m}\right)$$

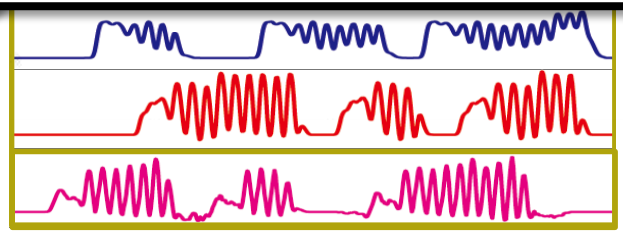
$Y_{l,m}$ の l：対数周波数index

$U_{k,m}$ の m：時刻index

- スペクトル形状がGMM型に拘束
- F_0によるスペクトル変動も表現可能

スペクトル形状

CWTの利点：
F_0の変化＝
対数周波数方向にシフト



F_0軌跡

Ω

事後確率最大化問題としての定式化

- 与えられた観測スペクトログラム Y に対し，以下を推定
  - F_0軌跡 Ω
  - 音量軌跡 U
  - スペクトル形状 W

  } Θ：パラメータ
- 最大事後確率推定

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}} \ln p(\Theta|Y) = \underset{\Theta}{\operatorname{argmax}} \left(\ln p(Y|\Theta) + \ln p(\Theta) \right)$$

観測スペクトログラムの生成モデル

事前分布
- F_0軌跡は連続的に変化，対応する音高の周りに分布
- 音量軌跡はスパース

  - 補助関数法 [Ortega+1970, Hunter+2000] を適用し，閉形式の更新式からなる反復アルゴリズムを導出可能

楽音分離実験(1/2)

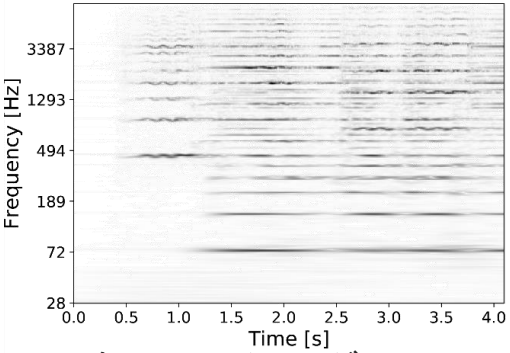
- 目的:HTFDによる音高毎への分離性能の確認
- 実験条件
  - データ:RWCクラシック音楽データベース [Goto2004]
    - テストデータ:No. 1から7の冒頭30 s
    - ハイパーパラメータ決定用検証データ:No. 13, 14, 15の冒頭30 s
    - MIDIシンセサイザーを用いて，標本化周波数16 kHzで合成
  - CWT:高速近似CWT(対数正規分布型ウェーブレット($\sigma = \ln(2)/36$)を使用) [亀岡+2008]
    - 時間シフト間隔:約2 ms(計算時間削減のため，パラメータ推定時には10 msとなるよう間引き)
    - 中心周波数:27.5 Hzから7040 Hz(100/3 cent間隔)
  - 比較手法:Hennequin [Hennequin+2010], HTC [Kameoka+2007], Harmonic NMF (HNMF) [Raczynski+2007], HTFD
- 評価指標:Signal-to-distortion ratio (SDR) [Vincent+2006]
  - 音源分離の標準的な評価指標．値が高いほど分離性能が高い．

楽音分離実験(2/2)

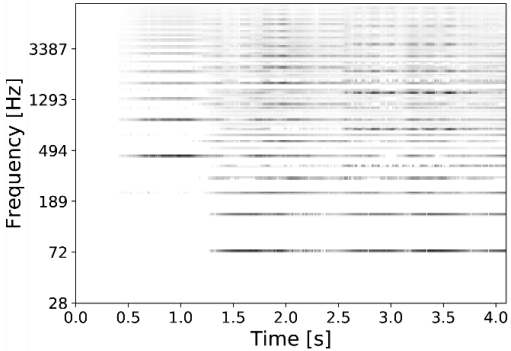
- HTFDが様々な楽曲で最高性能を達成

Method	Window Length	Excerpt No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 7
[Hennequin+2010]	64 ms	9.08/9.76	9.75/11.00	7.31/7.06	7.98/8.00	9.22/11.45	9.97/10.80	10.89/12.20
	128 ms	9.50/10.02	11.22/11.94	9.20/7.97	9.52/9.95	10.65/12.80	10.94/11.62	11.68/12.63
	256 ms	−6.59/ − 3.41	11.46/12.37	9.85/9.66	10.31/11.03	9.99/11.26	10.80/11.65	10.20/12.15
HTC [Kameoka+2007]	-	13.38/11.36	10.44/10.95	10.68/10.98	8.82/10.80	9.07/12.21	11.70/13.30	4.68/8.75
HNMF [Raczynski+2007]	-	18.47/15.76	15.10/15.60	13.26/14.77	17.10/16.95	16.08/17.17	20.91/20.96	16.82/17.14
HTFD	-	**19.21/16.28**	**16.99/18.35**	**13.75/17.19**	**18.23/17.78**	**18.01/17.92**	**22.27/22.24**	**17.89/17.62**

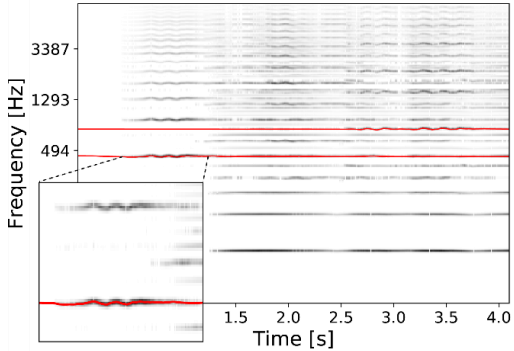
- HTFDはF_0を適切に追跡可能



観測スペクトログラム



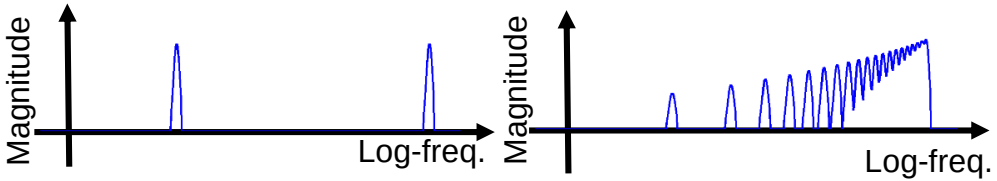
HNMFの推定値



HTFDの推定値

HTFDの拡張：ソースフィルタモデルの導入

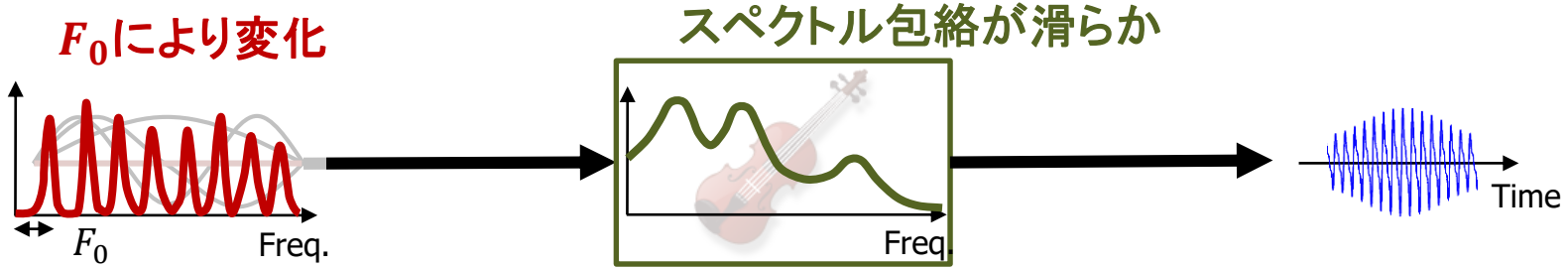
- 疑似周期信号の仮定
  - 局所的には周期的
  - 滑らかに周期・振幅が変化



⇒ 実際の楽音スペクトルと大きく異なる形状も許容

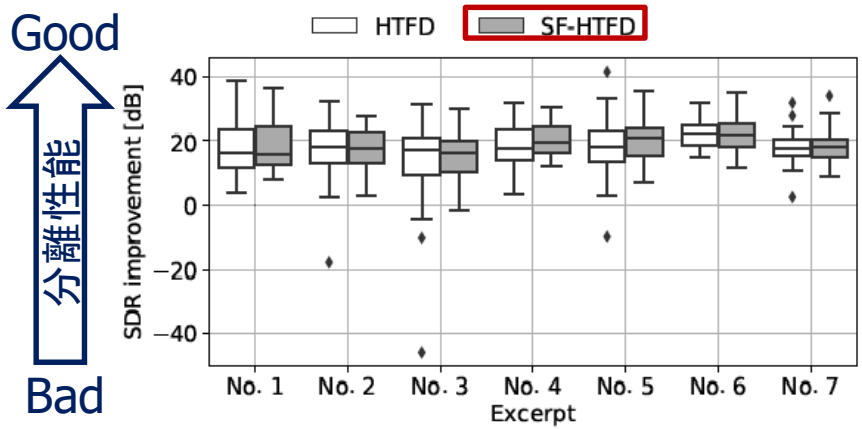
- ソースフィルタモデルにより，スペクトル包絡に制約を導入
  - 楽音の生成過程を振動体と共鳴体に分離し表現

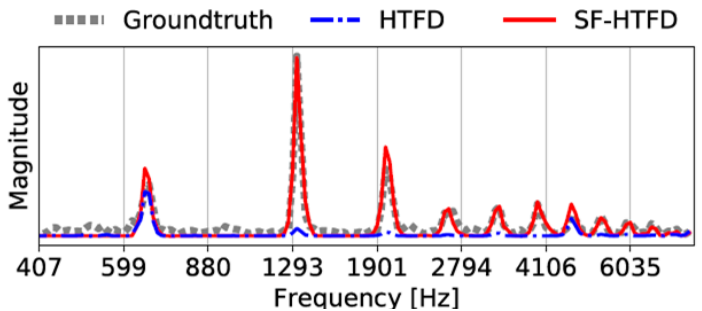
⇒ F_0とスペクトル包絡に関する制約を別々に導入可能



実験：ソースフィルタモデル導入の効果

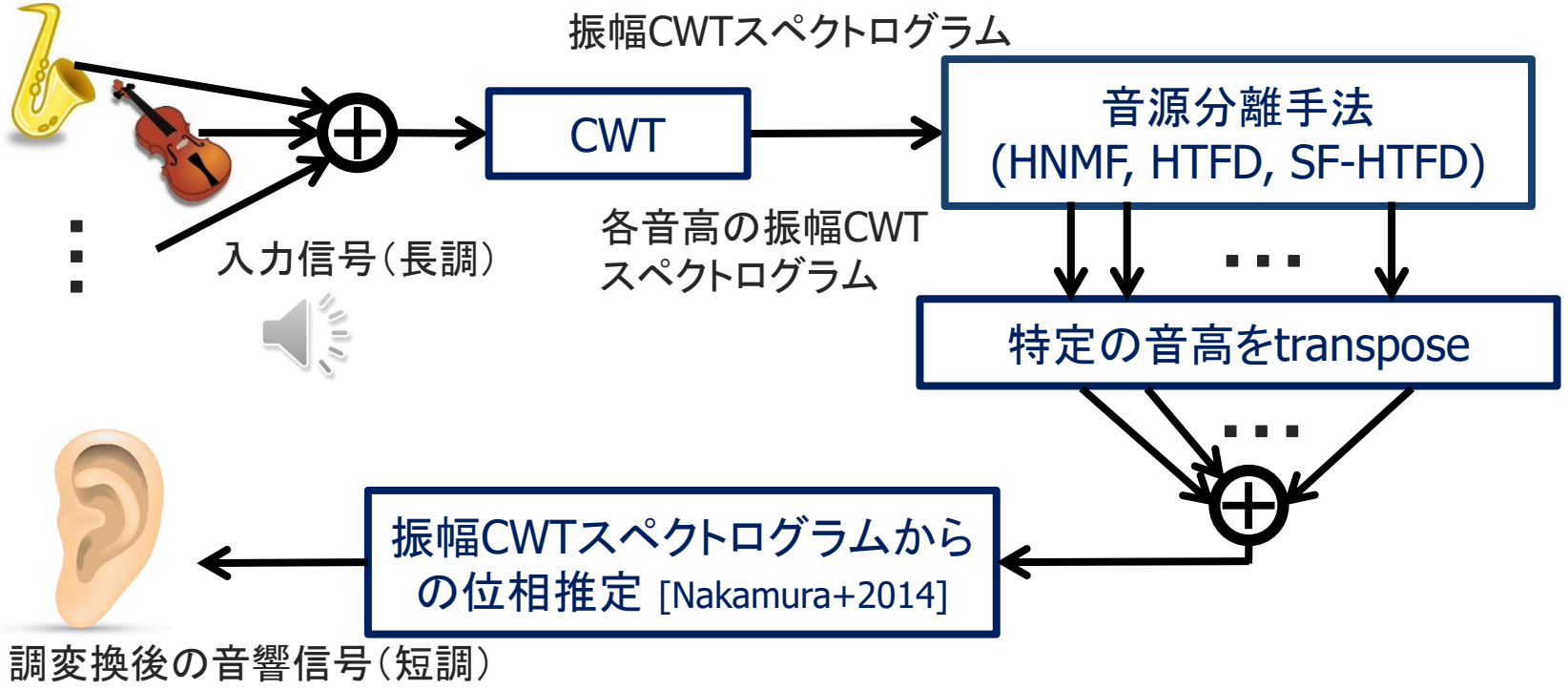
- 実験条件：HTFDの実験と同一
  - HTFDとソースフィルタモデルを導入したHTFD（SF-HTFD）を比較
- ソースフィルタモデルの導入により**ロバストに分離可能**
  - 低SNRな入力でも，HTFDに比べ大きくSDRが下がることが少ない．
  - 比較的不自然なスペクトル形状が推定されにくい．





推定スペクトルの比較

デモ：調変換への応用



HNMF：　HTFD：　SF-HTFD：

https://tomohikonakamura.github.io/Tomohiko-Nakamura/demo/HTFD/

Outline

- 自己紹介
- 音源分離とは？
- 調波時間因子分解：聴覚・機械学習・音響信号処理の知見を融合した音源分離手法
- 多重解像度深層分析：多重解像度解析にヒントを得た深層学習ベース音源分離手法
- まとめ

背景：DNNに基づくend-to-end音源分離

- DNNを用いたモノラル音源分離
  - 振幅・パワースペクトログラムから時間周波数マスクを推定する手法が代表的 [Hershey+2016, Jansson+2017, Takahashi+2017, Hennequin+2020, Takahashi+2021]

各時間周波数ビンに含まれる各音源成分の割合

- マスク推定における位相処理の問題点
  - 多くの場合，混合音の位相を分離音の位相として使用
  - 再構成した時間信号の振幅スペクトログラムが，振幅スペクトログラムドメインでの分離結果と必ずしも一致しない，i.e., 無矛盾な位相とならない [LeRoux+2008]

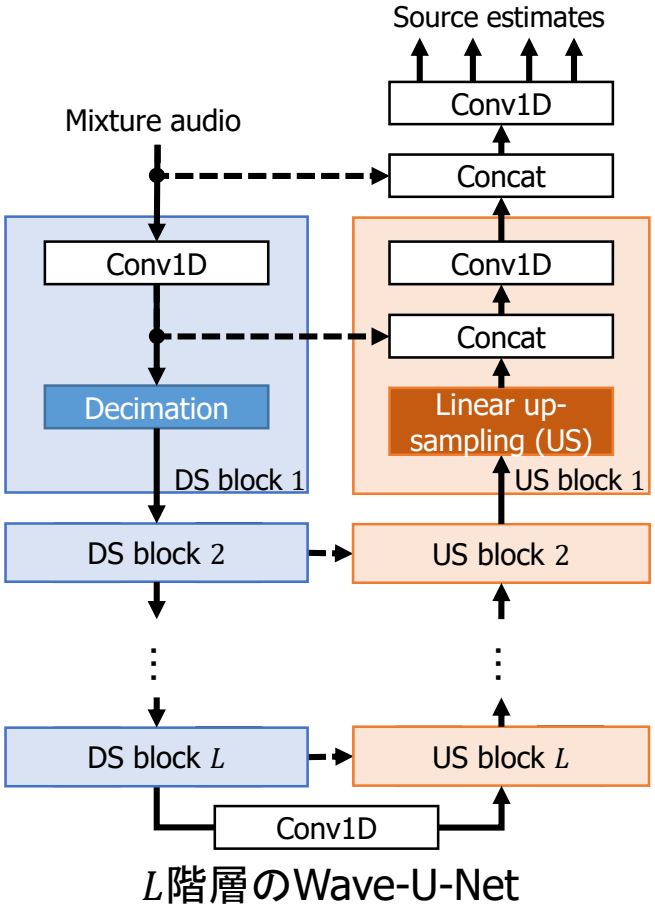
- End-to-endアプローチによる音源分離 [Venkataramani+2018, Stoller+2018, Luo+2019, Luis+2019, Kavalerox+2019, Samuel+2020, Nakamura+2021]
  - 時間波形を直接DNNに入力し，分離波形を直接出力
    ⇒ 位相を分離過程で直接考慮

Wave-U-Net [Stoller+2018]

- 音源分離のためのend-to-end DNN

- U-Net構造 [Ronneberger+2015]
  - エンコーダ：間引き（Decimation）により，特徴量を繰り返しダウンサンプリング（DS）
  - デコーダ：線形補間により，特徴量を繰り返しアップサンプリング（US）

- U-Net構造の利点
  - 指数的に畳み込み層の受容野を拡大
    ⇒ 音源信号の長期依存性を捉えるように促進

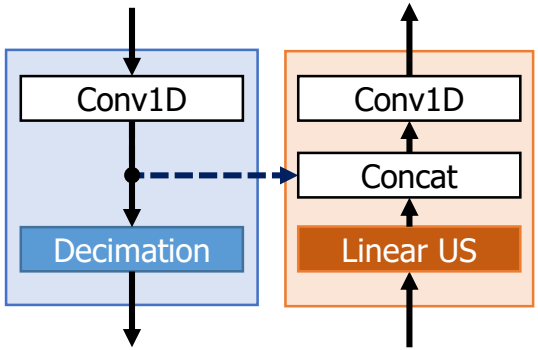


L階層のWave-U-Net

Decimation層の2つの問題点

(i) アンチエイリアシングフィルタの欠如

- 特徴量を時間方向に間引く処理のみ
  ⇒ エイリアシングが生じる
- 特徴量ドメインのエイリアシングは性能低下要因
  - e.g., 画像処理 [Mairal+2014, Zeiler+2014, Zhang+2019], 音素認識 [Gong+2018]



Wave-U-NetのDS&USブロック

(ii) 特徴量に含まれる情報の破棄

- 破棄された特徴量にも分離に有用な情報が含まれる可能性あり
- スキップ接続からDS前の特徴量が得られるものの，後続の畳み込み層は位置によって不変な処理
  ⇒ どの特徴量が破棄または保持されているかを同定不能

信号処理として"リーズナブル"なDS層

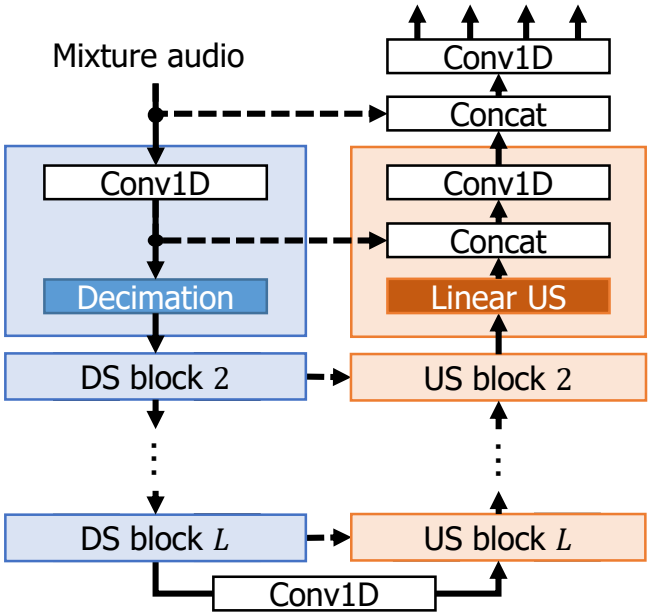
- デシメーション層の問題点の解決方針：学習 vs. 構造の変更
  - 2つの問題点による分離性能への影響が，学習により軽減できるかは無保証
  - 問題の原因はデシメーション自体
    ⇒ 信号処理の観点からDS層の構造を見直すことで解決できないか？

目的：以下の性質を満たすDS層を設計

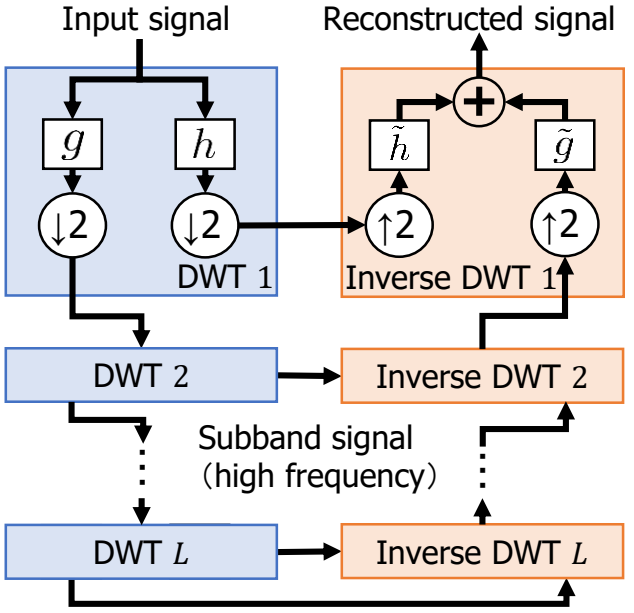
- アンチエイリアシングフィルタを持つ
- DSの前後で特徴量の全情報を保持
  i.e., DS層が完全再構成性を持つ

着眼：U-Netと多重解像度解析の構造の類似性

- 多重解像度解析（MRA）[Mallat+1989]
  - 離散ウェーブレット変換（DWT）により繰り返し信号をDS
  - 逆DWTにより，サブバンド信号から入力信号を再構成可能



L階層のWave-U-Net

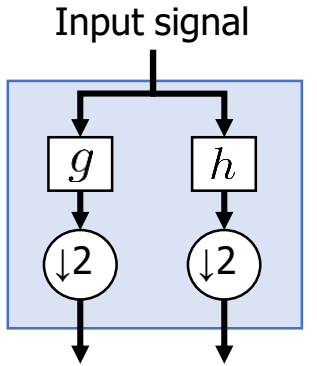


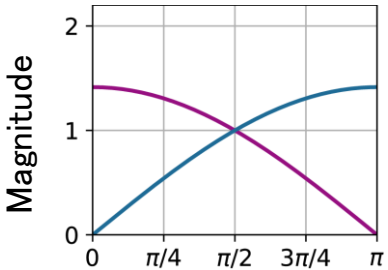
L 階層の多重解像度解析・合成

DWT層：DWTを用いたDS層 [Nakamura+2021]

- DWTの性質
  - ローパス・ハイパスフィルタ(g, h)からなる2チャネルフィルタバンク
    ⇒ アンチエイリアシングフィルタをもつ
  - 完全再構成性をもつ
    ⇒ DSによる情報の欠落がない

- DWTを用いればDecimation層の問題を同時に解決可能
  ⇒ DWTを用いたDS層（DWT層）を提案

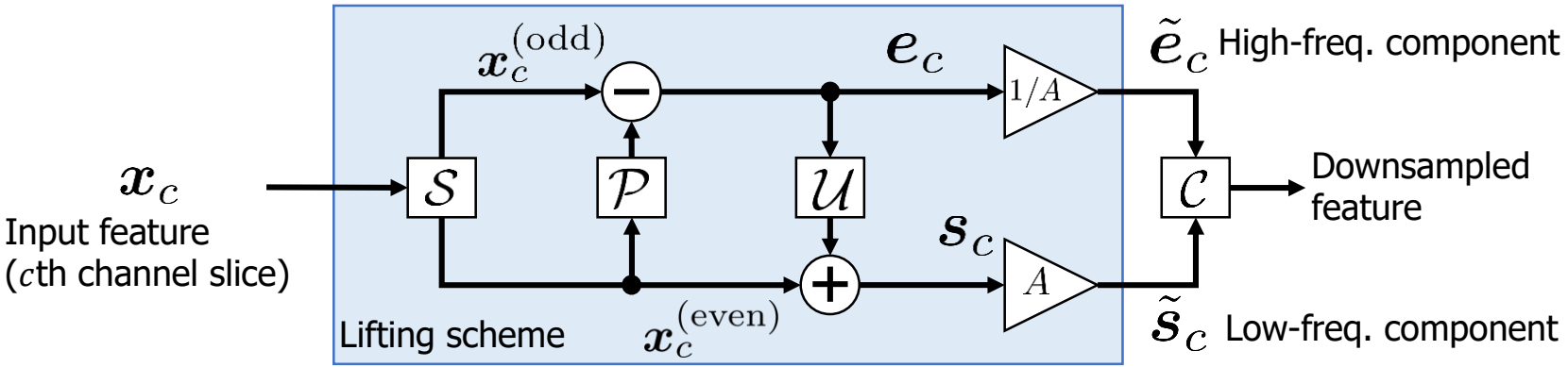




Haarウェーブレットを用いたDWTのローパス・ハイパスの周波数応答

DWT層の実装(1/2)

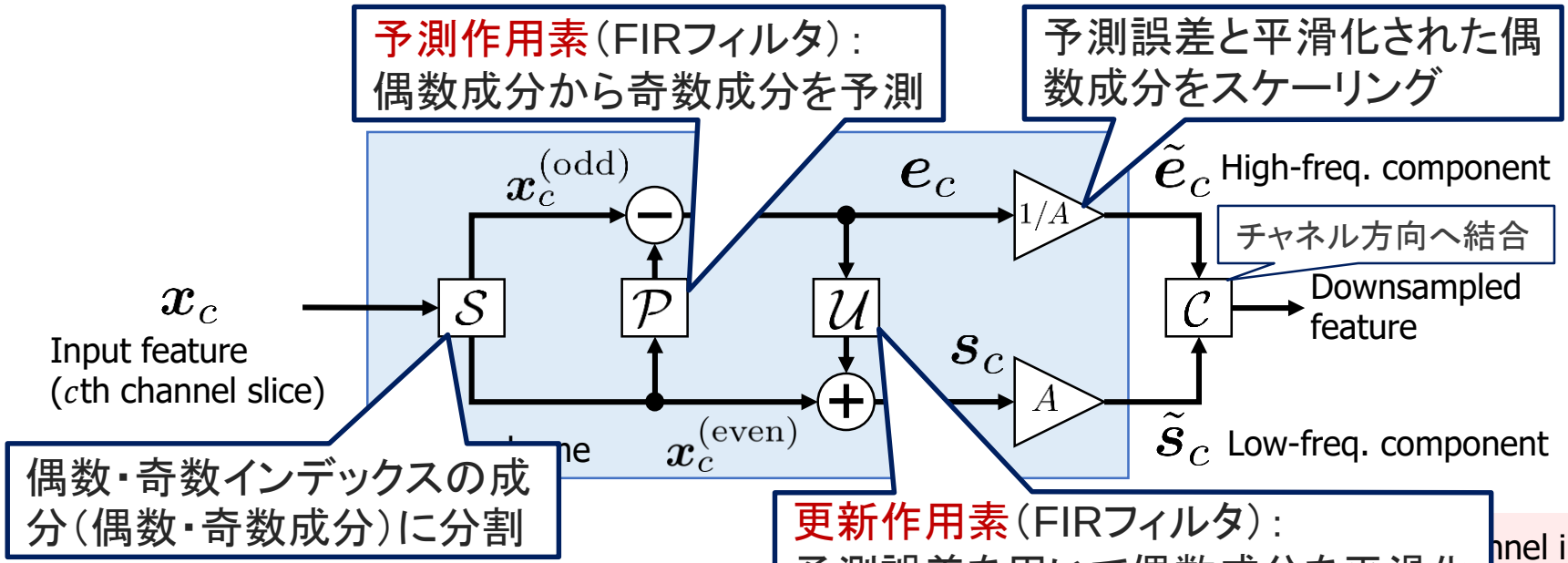
- DWT層の処理
  - 入力された特徴量の各チャネルスライス $\boldsymbol{x}_c$ を信号をみなし，リフティングスキーム [Swelden+1996] によりDWTを適用
  - 得られた低周波・高周波成分 $\{\tilde{\boldsymbol{s}}_c\}_c, \{\tilde{\boldsymbol{e}}_c\}_c$ をチャネル方向へ結合して，DSされた特徴量 $\tilde{\boldsymbol{x}}_c$ を出力



c: Channel index

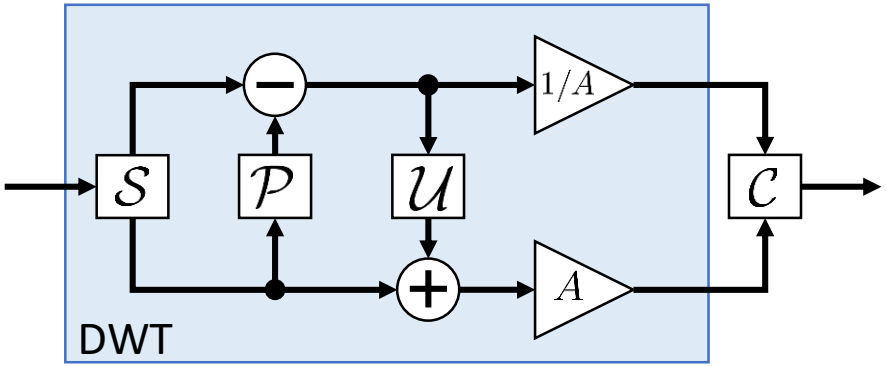
DWT層の実装(2/2)

- リフティングスキーム [Swelden+1996]
  - DWTと逆DWTの効率的な計算方法
  - 予測作用素と更新作用素を変更することで周波数特性を変更可能

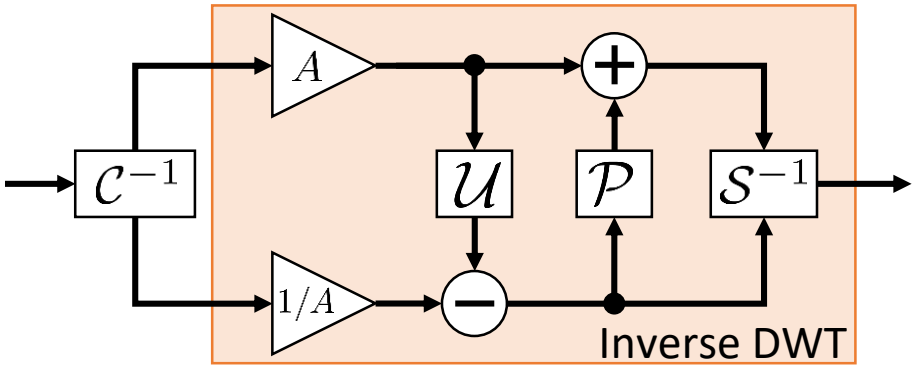


逆DWT層：逆DWTに基づくUS層

- DWT層を構成するステップは全て可逆
  ⇒ 完全再構成性は構造から保証可能
- 逆DWTに基づくUS層（逆DWT層）は，DWT層の処理を逆順に行うことで構成可能



DWT層



逆DWT層

従来のDS層との関連

- 従来のDS層：アンチエイリアシングフィルタまたは完全再構成性を欠く

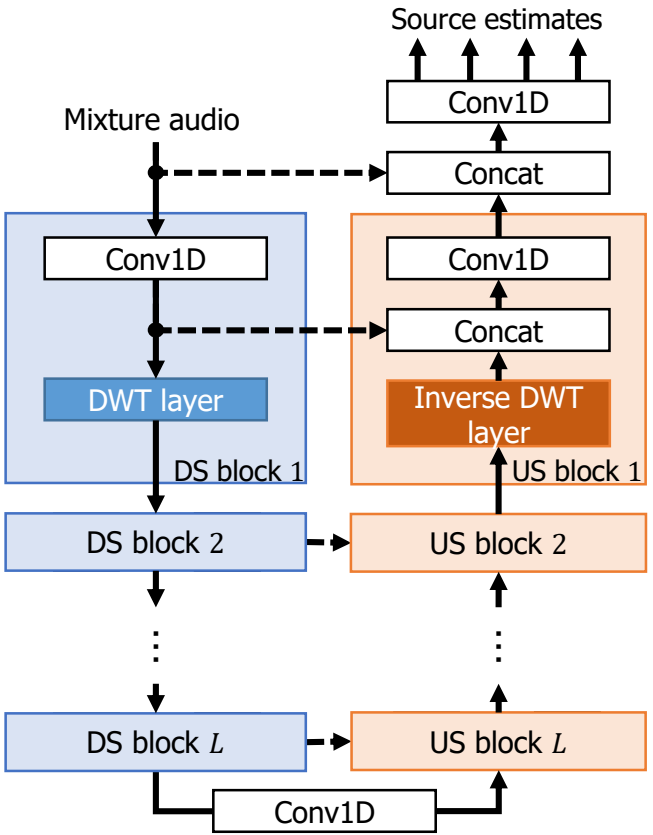
DS layer	Anti-aliasing filters	Perfect reconstruction property
Decimation layer	No	No
Average pooling	**Yes**	No
Squeezing operation [Dinh+2017]	No	**Yes**
DWT layer	**Yes**	**Yes**

  - Average pooling：Decimationの代わりに移動平均フィルタを利用
  - Squeezing operation: DWT層から予測・更新ステップを除いたDS層

多重解像度深層分析(MRDLA) [Nakamura+2021]

- 多重解像度深層分析(MRDLA)
  - Wave-U-Netの構造がベース
  - Wave-U-NetのDS層とUS層を，それぞれDWT層と逆DWT層へ置換
    ⇒ Decimation層の問題を同時に解決

- DWT層の拡張
  - 様々なウェーブレットでの性能比較 [Kozuka+2020]
  - ウェーブレットとDNNの同時学習 [Nakamura+2021]



L階層のMRDLA

楽音分離実験(1/2):実験条件

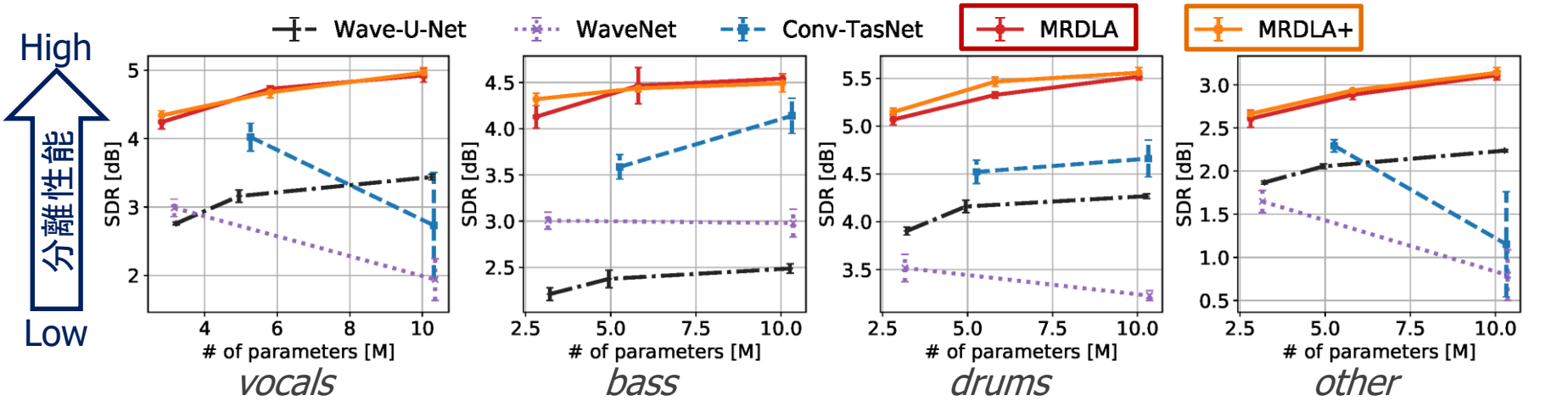
- 目的:MRDLAの音源分離性能の検証

- データセット:MUSDB18 [Rafii+2018]
  - 学習/検証データ:75, 25曲(4分割交差検定を使用)
  - テストデータ:50曲
  - 楽器:*vocal, bass, drums, other*
  - サンプリング周波数:22.05 kHz
- 学習条件(データ拡張以外はWave-U-Netの論文 [Stoller+2018] の条件を使用)
  - ロス関数は平均二乗誤差とし, Adam(学習率:1.0×10^{-4}, バッチサイズ:16)を使用
  - データ拡張:ランダムクロップ・ゲイン, 左右チャネル・楽曲間のデータの入れ替え
- 評価指標:SDR
  - BSSEval v4ツール [Stoter+2018] を用いて計算

楽音分離実験(2/2):結果

- 各楽器・モデルサイズで，MRDLAが最高性能を達成
  - State-of-the-artな手法(Wave-U-Net [Stoller+2018]，WaveNet [Lluıs+2019]，Conv-TasNet [Luo+2019])に比べても，少ないパラメータで高性能に動作
  - 学習可能なDWT層に変更することで，パラメータが少ない場合に若干性能が向上

FixedなDWT層を使用　　学習可能なDWT層を使用



Ablation study（1/2）：実験条件

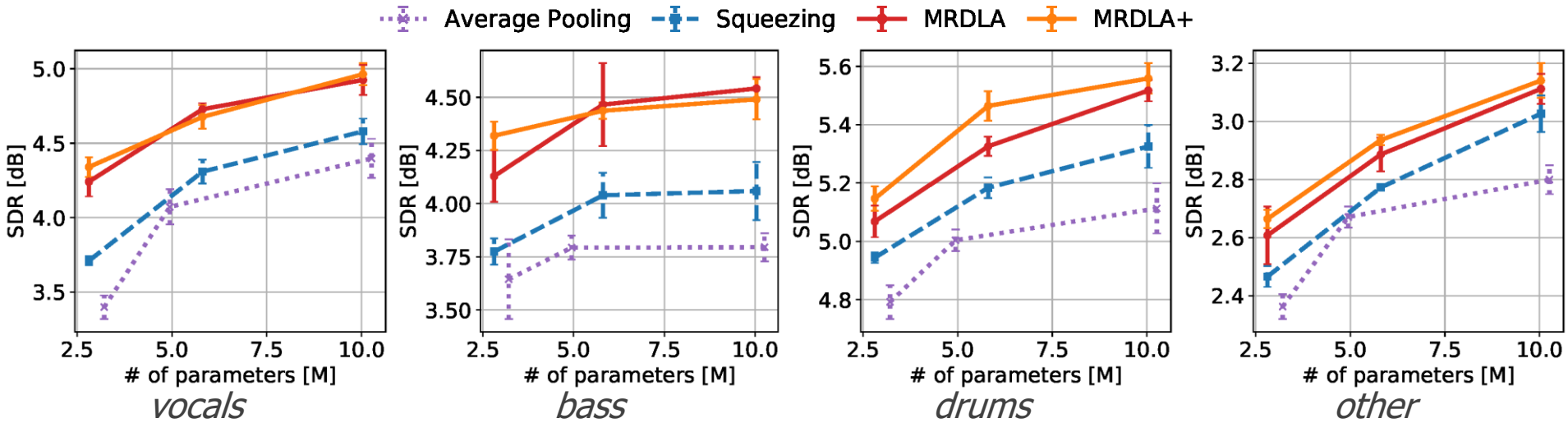
- 目的：アンチエイリアシングフィルタと完全再構成性の効果を分けて評価

- 比較手法

Model	Anti-aliasing filters	Perfect reconstruction property	DS layer
Average Pooling	**Yes**	No	Average pooling
Squeezing	No	**Yes**	Squeezing operation [Dinh+2017]
MRDLA	**Yes**	**Yes**	DWT layer
MRDLA+	**Yes**	**Yes**	WN-TDWT layer

Ablation study（2/2）：結果

- アンチエイリアシングフィルタと完全再構成性のいずれも分離性能に寄与
- アンチエイリアシングフィルタよりも，完全再構成性のほうが比較的重要であることを示唆



主観評価実験：MRDLAの分離品質

- プリファレンステストを実施
  - MUSDB18データセットのテストデータ中10曲（10秒切り出し）をランダムに選択
  - 12人の被験者が，混合音と正解音，各手法の分離音とマイナスワン音（混合音から分離音を除算したもの）を聴取し，分離音の品質が良いものを選択
  - 各曲各楽器で手法に関しランダムな順番で提示

- 結果：MRDLAの分離品質が有意に高いことを確認

Instrument	Preference score [%]		p-value
	Wave-U-Net	MRDLA+	
vocals	29.17	**70.83**	$< 10^{-5}$
bass	6.67	**93.33**	$< 10^{-20}$
drums	19.17	**80.83**	$< 10^{-10}$
other	26.67	**73.33**	$< 10^{-6}$

Instrument	Preference score [%]		p-value
	Conv-TasNet	MRDLA+	
vocals	13.33	**86.67**	$< 10^{-15}$
bass	13.33	**86.67**	$< 10^{-15}$
drums	10.83	**89.17**	$< 10^{-17}$
other	22.50	**77.50**	$< 10^{-8}$

Demo: Secretariat - Over The Top

Observed mixture: 

https://tomohikonakamura.github.io/Tomohiko-Nakamura/demo/MRDLA/

Instrument	Groundtruth	Wave-U-Net [Stoller+2018]	Proposed MRDLA
vocals			
bass			
drums			
other			

まとめ

- 調波時間因子分解（HTFD）[Nakamura+2020]
  - HTC，NMF，ソースフィルタモデルを統合した教師なし音源分離手法
  - 連続ウェーブレット変換領域で，スペクトログラムの局所的，大局的な構造，音源の生成過程を組み込んだ確率的生成モデルを構築
  - 従来の教師なし音源分離手法に比べ，大きく性能が向上

- 多重解像度深層分析（MRDLA）[Nakamura+2021]
  - Wave-U-Net [Stoller+2018] をベースにした時間領域音源分離手法
  - 多重解像度分析 [Mallat+1989] とU-Net構造の類似性に着眼し，離散ウェーブレット変換を用いたダウンサンプリング層（DWT層）を提案
    ⇒ 特徴量領域でのエイリアシングと情報欠落を防ぐことで，分離性能が向上