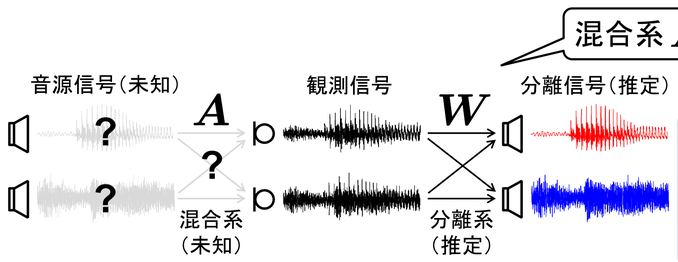
1-1P-2

# 解像度の異なる複数の時間周波数表現を用いた独立低ランク行列分析

☆細谷泰稚，北村大地 (香川高専)，矢田部浩平 (早稲田大)

## 1. 研究背景

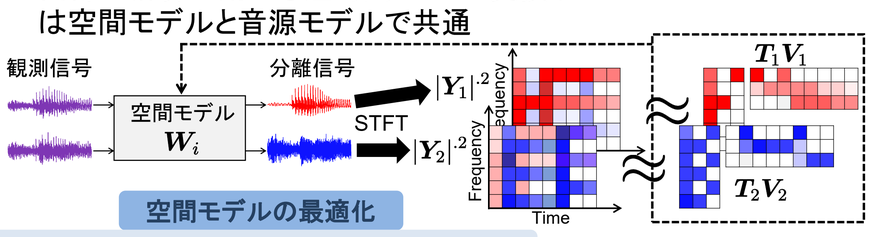
- ブラインド音源分離（BSS）：元の音源信号や混合系が未知である音源分離




**代表的な従来手法**

- 独立成分分析（ICA）[Comon, 1994]
- 独立ベクトル分析（IVA）[Kim+, 2007]
- 独立低ランク行列分析（ILRMA）[Kitamura+, 2016]

- 独立低ランク行列分析（independent low-rank matrix: ILRMA）
  - 空間モデルの最適化と音源モデルの最適化を交互に行う
  - 内部の計算で用いる短時間Fourier変換（short-time Fourier transform: STFT）の窓関数は空間モデルと音源モデルで共通




**音源モデルの最適化**

非負値行列因子分解（NMF）[Lee, 1999]により，分離信号の時間周波数構造（パワースペクトログラム）$|\boldsymbol{Y}_n|^{.2}$ を2つの行列 $\boldsymbol{T}_n$ 及び $\boldsymbol{V}_n$ の積で低ランク近似

**空間モデルの最適化**

分離信号が「互いに独立」かつ「低ランクな時間周波数構造」を持つように分離行列 $\boldsymbol{W}_i$ を更新

### 本研究の動機

- 過去の実験的な調査によって確認されている二つの事実

空間モデル $\boldsymbol{W}_i$ の最適化に用いるSTFTの窓長は観測信号の残響よりも十分長くあるべき

音源モデル $\boldsymbol{T}_n, \boldsymbol{V}_n$ の最適化に用いるSTFTの窓長はパワースペクトログラムの近似精度を左右する

ILRMAが周波数領域での瞬時混合を仮定しているため

STFTの窓長によって，スペクトログラムの時間周波数解像度が変化するため

➡ 「音源モデルでの最適な窓長」と「空間モデルでの最適な窓長」が存在し，それらが互いに異なっている可能性がある

しかし，従来のILRMAでは二つのモデルに同一の窓長を用いている

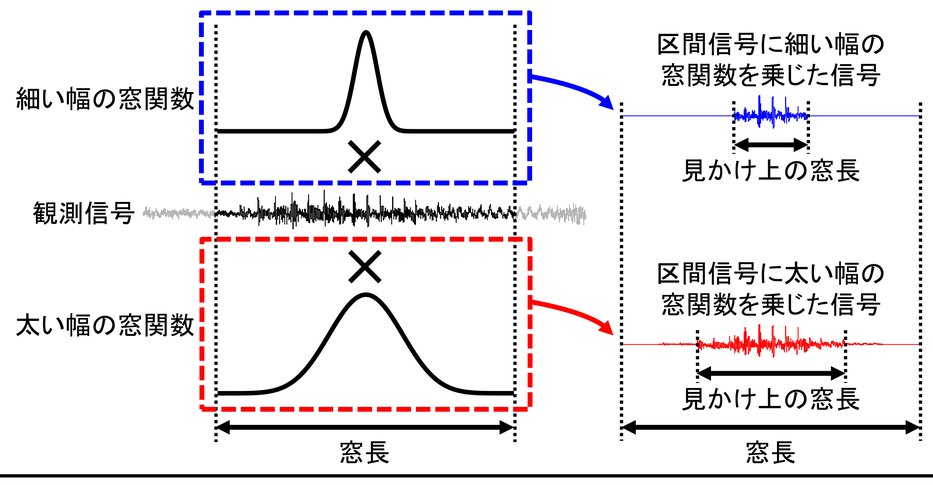
### 本研究の目的

- ILRMAの空間モデル及び音源モデルにそれぞれ独立した解像度の時間周波数表現を導入する手法を提案
- 各モデルで用いる時間周波数解像度を変化させ，それぞれの場合における音源分離性能を調査

## 2. 提案手法

- **多重解像度時間周波数表現に基づくILRMA**
  - ILRMAの空間モデル及び音源モデルにおける（見かけ上の）窓長をそれぞれ独立して設定可能な手法
  - 各モデルにそれぞれ独立した解像度の時間周波数表現を導入

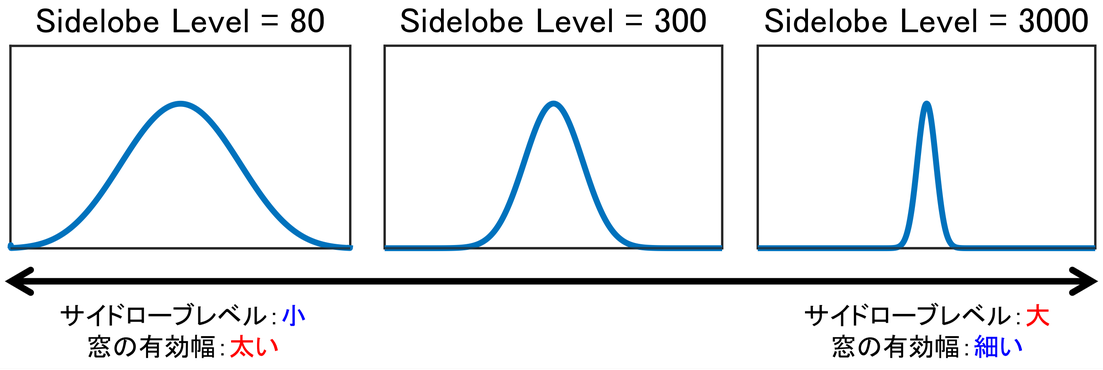
- 窓長の変更方法
  - STFTを適用する際の短時間信号に掛ける窓関数の有効幅を設定
    - 細い幅の窓関数：短い窓長での時間周波数表現
    - 太い幅の窓関数：長い窓長での時間周波数表現
  - スペクトログラムのサイズを変化させずに，解像度のみを変更可能




- 窓関数の有効幅の変更方法
  - 窓関数にChebyshev窓を使用

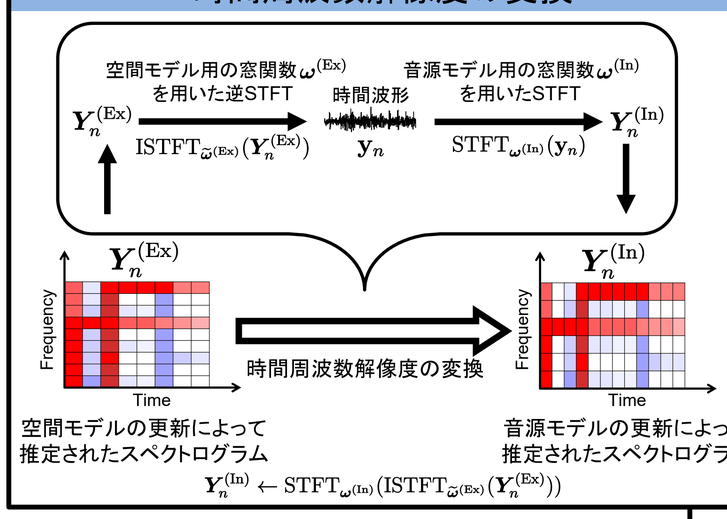
周波数領域における窓関数のサイドローブピークの最大値

  - 「サイドローブレベル」というパラメタにより，時間領域での窓関数の有効幅を変更可能
  - 上記のパラメタの増加に伴い，窓関数の有効幅は細くなる




- アルゴリズム

### 時間周波数解像度の変換

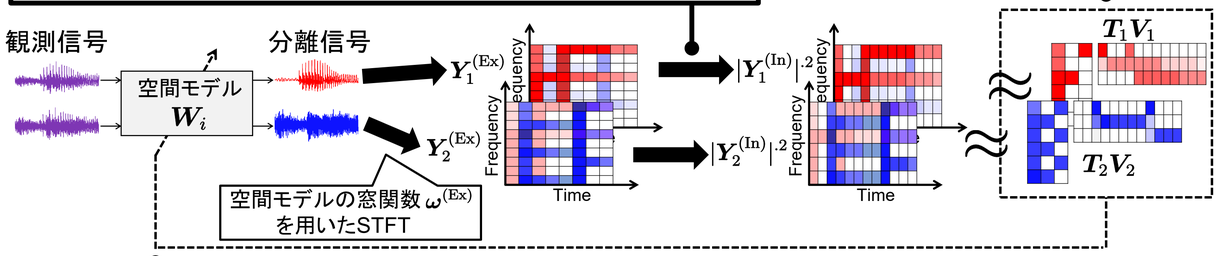



### 音源モデルの更新

NMFによる低ランク近似

$$\boldsymbol{T}_n \leftarrow \boldsymbol{T}_n \odot \left\{ \frac{\left[ |\boldsymbol{Y}_n^{(\mathrm{In})}|^{.2} \odot (\boldsymbol{T}_n\boldsymbol{V}_n)^{\cdot -2} \right] \boldsymbol{V}_n^{\mathrm{T}}}{(\boldsymbol{T}_n\boldsymbol{V}_n)^{\cdot -1} \boldsymbol{V}_n^{\mathrm{T}}} \right\}^{\cdot \frac{1}{2}}$$

$$\boldsymbol{V}_n \leftarrow \boldsymbol{V}_n \odot \left\{ \frac{\boldsymbol{T}_n^{\mathrm{T}} \left[ |\boldsymbol{Y}_n^{(\mathrm{In})}|^{.2} \odot (\boldsymbol{T}_n\boldsymbol{V}_n)^{\cdot -2} \right]}{\boldsymbol{T}_n^{\mathrm{T}} (\boldsymbol{T}_n\boldsymbol{V}_n)^{\cdot -1}} \right\}^{\cdot \frac{1}{2}}$$




### 空間モデルと分離信号の更新

$$\boldsymbol{U}_{in} \leftarrow \frac{1}{J} \sum_j \frac{1}{[\boldsymbol{T}_n\boldsymbol{V}_n]_{i,j}} \boldsymbol{x}_{ij}^{(\mathrm{Ex})} \boldsymbol{x}_{ij}^{(\mathrm{Ex})\mathrm{H}}$$

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

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

$$\lambda_{in} \leftarrow \left[ \boldsymbol{W}_i^{-1} \right]_{m_{\mathrm{ref}},n}$$

$$\boldsymbol{w}_{in} \leftarrow \lambda_{in} \boldsymbol{w}_{in}$$

$$y_{ijn}^{(\mathrm{Ex})} \leftarrow \boldsymbol{w}_{in}^{\mathrm{H}} \boldsymbol{x}_{ij}^{(\mathrm{Ex})}$$

$$[\boldsymbol{T}_n]_{i,k} \leftarrow |\lambda_{in}|^2 [\boldsymbol{T}_n]_{i,k}$$

必要となるのはパワースペクトログラムの近似行列（分散行列）
音源モデルの情報をそのまま計算に用いることができる

➡ 音源モデルから空間モデルへと向かう方向の時間周波数解像度の変換を考える必要がなくなる

➡ 音源モデルを時間領域の信号へと変換する際に付与する位相について議論する必要がなくなる

解像度の変換は常に空間モデル→音源モデルの一方向となる

## 3. 比較実験

- **実験条件**
  - RWCPデータベース収録のインパルス応答E2Aによって2音源の畳み込み混合を行い，生成した10曲の混合信号に対して提案手法を適用
  - 提案手法の空間モデル及び音源モデルのサイドローブレベルには，以下の表に示す値を設定し，source-to-distortion ratio（SDR）の改善量（SDRi）で評価

| | |
|---|---|
| 窓長 | 256 ms（4096点） |
| シフト長 | 32 ms（512点，窓長の1/8） |
| Chebyshev窓のサイドローブレベル | {20, 30, 40, 50, 60, 70, 80, 90, 100, 120, 150, 200, 300, 500, 700, 1000, 1500, 2000, 3000} |
| NMFの基底数 | 10 |
| 初期値 | $\boldsymbol{W}_i$：単位行列<br>$\boldsymbol{T}_n$ 及び $\boldsymbol{V}_n$：乱数行列 |
| 反復回数 | 200 回 |
| 試行回数 | 異なる乱数シードで10回 |
| 参照マイクチャネル $m_{\mathrm{ref}}$ | 1 |

空間モデルと音源モデルのサイドローブレベルが一致する場合は従来のILRMAの結果に対応
実験結果の図では，**太枠で囲んだ対角部分が従来のILRMAの結果**となる

- **実験結果**
- ドラムとボーカルの混合音源に対する平均SDRi

| Average SDRi | 空間モデルのサイドローブレベル [dB] | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 音源モデルのサイドローブレベル [dB] | 40 | 50 | 60 | 70 | 80 | 90 | 100 | 120 | 150 | 200 | 500 | 1000 | 1500 |
| 40 | 1.32 | 8.62 | 5.02 | 4.58 | 4.60 | 4.61 | 4.68 | 4.78 | 4.94 | 4.80 | 5.79 | 5.56 | 5.03 |
| 50 | 1.93 | 10.03 | 10.07 | 10.10 | 10.37 | 10.25 | 10.21 | 10.25 | 10.00 | 9.87 | 9.81 | 9.90 | 9.42 |
| 60 | 2.06 | 1.77 | 12.19 | 10.13 | 7.60 | 7.88 | 8.63 | 9.02 | 9.32 | 10.00 | 9.32 | 10.63 | 8.02 |
| 70 | 1.78 | 1.05 | 13.64 | 6.33 | 5.81 | 5.54 | 5.09 | 4.91 | 3.86 | 3.40 | 6.51 | 9.54 | 7.03 |
| 80 | 1.58 | -0.17 | 14.25 | 8.07 | 5.86 | 6.80 | 5.84 | 4.50 | 4.61 | 3.51 | 4.87 | 6.48 | 6.99 |
| 90 | 1.35 | 0.74 | 9.63 | 13.17 | 3.78 | 5.45 | 5.57 | 4.30 | 3.98 | 3.30 | 4.45 | 6.62 | 6.20 |
| 100 | -0.12 | 1.38 | 9.78 | 8.21 | 10.59 | 4.74 | 4.92 | 4.55 | 3.52 | 3.17 | 4.53 | 5.95 | 6.21 |
| 120 | 0.31 | -0.08 | 13.65 | 15.32 | 12.84 | 11.30 | 4.49 | 3.95 | 2.59 | 2.56 | 4.25 | 5.28 | 6.26 |
| 150 | -1.96 | -0.36 | 12.88 | 11.96 | 6.78 | 2.90 | 6.37 | 3.74 | 2.44 | 2.24 | 3.29 | 4.20 | 4.51 |
| 200 | -2.20 | 0.22 | 12.75 | 10.26 | 10.68 | 12.53 | 11.68 | 3.70 | 2.07 | 2.04 | 3.29 | 3.96 | 4.46 |
| 500 | -0.09 | 4.27 | 4.15 | 11.16 | 3.84 | 9.17 | 4.61 | 5.59 | 11.12 | 11.92 | 3.40 | 3.21 | 2.67 |
| 1000 | 1.08 | 6.96 | 13.10 | 12.43 | 12.41 | 11.74 | 11.64 | 10.31 | 10.49 | 10.39 | 2.83 | 3.39 | 3.26 |
| 1500 | 1.65 | 9.04 | 12.67 | 11.61 | 12.15 | 12.30 | 12.05 | 10.92 | 8.81 | 3.20 | 5.24 | 3.46 | 3.37 |

Poor — Good

対角部分の分離性能を上回るものが非対角部分に存在

空間モデルと音源モデルのサイドローブレベルがそれぞれ70 dBと120 dBの部分に最大の平均SDRiが現れている

対角部分よりも下側に比較的高い平均SDRiがみられる

- 10曲全てに対する平均SDRi

| Average SDRi | 空間モデルのサイドローブレベル [dB] | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 音源モデルのサイドローブレベル [dB] | 40 | 50 | 60 | 70 | 80 | 90 | 100 | 120 | 150 | 200 | 500 | 1000 | 1500 |
| 40 | 3.25 | 4.55 | 1.21 | 1.24 | 1.66 | 1.64 | 1.80 | 1.70 | 1.61 | 1.69 | 2.25 | 1.73 | 0.89 |
| 50 | 3.89 | 6.86 | 4.77 | 4.73 | 4.76 | 4.43 | 4.33 | 4.53 | 5.05 | 5.74 | 5.82 | 4.62 | 3.96 |
| 60 | 2.68 | 6.83 | 9.17 | 7.92 | 7.87 | 6.88 | 6.78 | 6.80 | 6.99 | 6.91 | 6.34 | 5.29 | 4.22 |
| 70 | 2.02 | 5.87 | 6.96 | 8.42 | 6.96 | 6.85 | 6.50 | 6.20 | 5.80 | 5.41 | 5.29 | 5.19 | 4.07 |
| 80 | 1.18 | 3.28 | 7.20 | 8.10 | 7.58 | 7.66 | 7.08 | 6.63 | 6.28 | 5.84 | 4.96 | 4.73 | 3.82 |
| 90 | 0.58 | 1.16 | 6.61 | 8.39 | 7.62 | 7.61 | 7.32 | 6.80 | 6.47 | 6.07 | 4.79 | 4.52 | 3.66 |
| 100 | 0.22 | 0.80 | 5.86 | 7.72 | 8.19 | 7.62 | 7.38 | 6.89 | 6.41 | 5.82 | 4.93 | 4.42 | 3.44 |
| 120 | -0.45 | 0.00 | 4.60 | 6.68 | 8.30 | 8.03 | 7.38 | 6.75 | 6.33 | 5.70 | 5.01 | 4.33 | 3.46 |
| 150 | -1.32 | -0.56 | 4.08 | 4.20 | 6.56 | 6.23 | 7.49 | 6.65 | 6.21 | 5.64 | 4.84 | 4.22 | 3.16 |
| 200 | -2.41 | -1.92 | 5.19 | 4.29 | 4.93 | 6.73 | 7.12 | 6.15 | 6.22 | 5.67 | 4.61 | 4.14 | 3.26 |
| 500 | 0.34 | 2.31 | 4.46 | 4.13 | 4.35 | 4.19 | 3.63 | 2.93 | 3.53 | 4.80 | 4.25 | 3.88 | 3.24 |
| 1000 | -0.49 | 3.70 | 4.74 | 4.33 | 4.32 | 3.97 | 3.70 | 3.45 | 3.55 | 2.95 | 4.44 | 3.65 | 2.96 |
| 1500 | 1.43 | 3.79 | 4.53 | 3.90 | 3.96 | 4.21 | 3.93 | 4.10 | 3.50 | 2.38 | 3.82 | 3.48 | 2.94 |

Poor — Good

空間モデルのサイドローブレベルが80～100 dBの場合において，対角部分の分離性能を上回るものが非対角部分に存在

二つの実験結果（片方は10曲の全体平均）を比べると，「空間モデルと音源モデルのサイドローブレベルがそれぞれ70 dBと120 dBの部分」の平均SDRiに隔たりがある

提案手法による音源分離の性能は，音源毎のばらつきが大きいと推測される

**空間モデルと音源モデルの最適化に同一の解像度の時間周波数表現を用いることが，必ずしも最良の分離性能を与えるとは限らない**