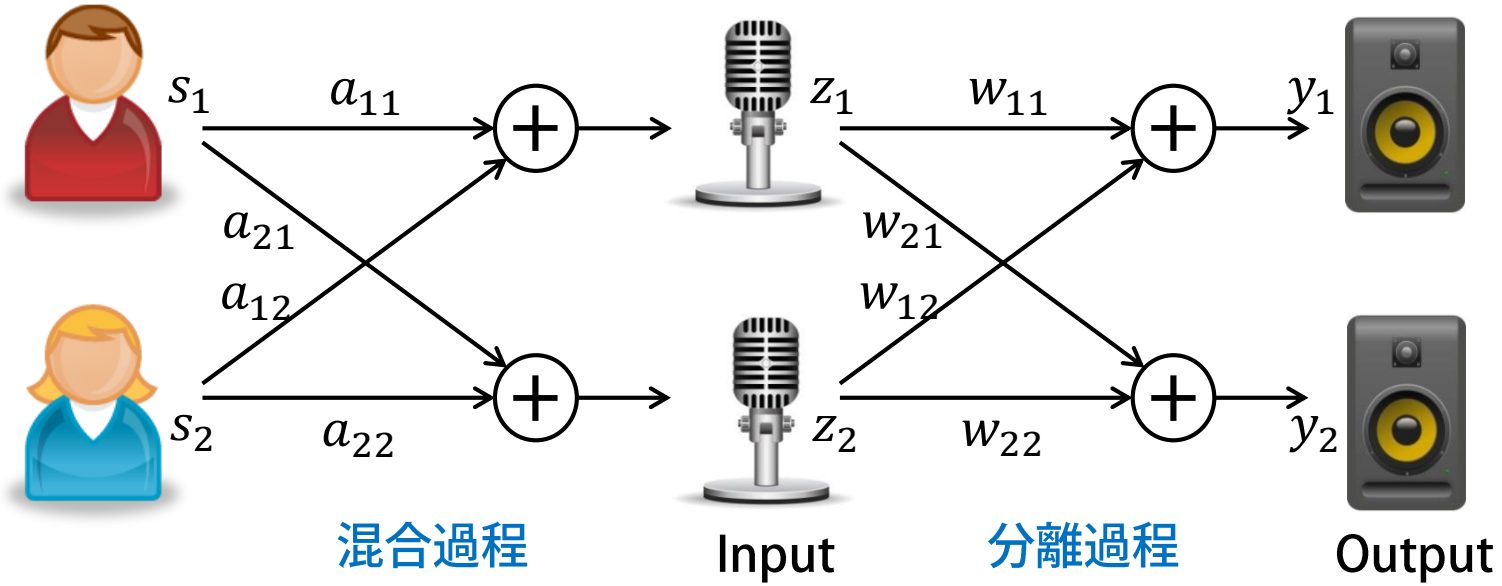
ロボットのための
マイクアレイ音響信号処理

京都大学 大学院情報学研究科 知能情報学専攻
理研 AIP 音響情景理解チーム (兼任)
吉井 和佳
yoshii@kuis.kyoto-u.ca.jp

マイクアレイ処理の目標

- マルチチャネルの混合音に対する音源分離と音源定
  - 入力：$z_1, z_2, \cdots, z_N$ 出力：$y_1, y_2, \cdots, y_M$ $(\approx s_1, s_2, \cdots, s_M)$
    - 混合過程：$s_1, s_2, \cdots, s_M \rightarrow$ 観測 $z_1, z_2, \cdots, z_N$
    - 条件：A が既知 ↔ A が未知 (ブラインド)



マイクアレイ処理の例

- ブラインド条件かどうかで使用できる技術が異なる
  - 非ブラインド条件
    - ビームフォーミング
    - MUSIC (multiple signal classification) 法
  - ブラインド条件
    - 独立成分分析 (ICA)・独立ベクトル分析 (IVA)
    - マルチチャネル非負値行列分解 (NMF)
    - 時間周波数クラスタリング・マスキング
  - より先進的なトピック
    - 統一的な確率モデルによる分離・定位・残響除去
    - ノンパラベイズモデルによる音源数の自動推定

音源信号と観測信号の関係

- 各チャネルの観測信号は音源信号の時間遅れ
  - 音源信号 $s(t)$ を M 個のマイクで観測するとする
  - 各マイク m では遅れ τ_m が発生する

$$\boldsymbol{z}(t) = \begin{bmatrix} z_1(t) \\ z_2(t) \\ \vdots \\ z_M(t) \end{bmatrix} = \begin{bmatrix} s(t-\tau_1) \\ s(t-\tau_2) \\ \vdots \\ s(t-\tau_M) \end{bmatrix} \xrightarrow{\text{フーリエ変換}} \boldsymbol{z}(\omega) = \begin{bmatrix} Z_1(\omega) \\ Z_2(\omega) \\ \vdots \\ Z_M(\omega) \end{bmatrix}$$

観測信号

$$Z_m(\omega) \equiv \int_{-\infty}^{\infty} z_m(t) e^{-j\omega t} dt = \int_{-\infty}^{\infty} s(t-\tau_m) e^{-j\omega t} dt = e^{-j\omega\tau_m} S(\omega)$$

$$S(\omega) \equiv \int_{-\infty}^{\infty} s(t) e^{-j\omega t} dt \qquad \boldsymbol{a}(\omega) = \begin{bmatrix} a_1 \\ \vdots \\ a_M \end{bmatrix} \equiv \begin{bmatrix} e^{-j\omega\tau_1} \\ \vdots \\ e^{-j\omega\tau_M} \end{bmatrix}$$

ステアリングベクトル

$$\boldsymbol{z}(\omega) = \boldsymbol{a}(\omega) S(\omega)$$

観測　関係　音源

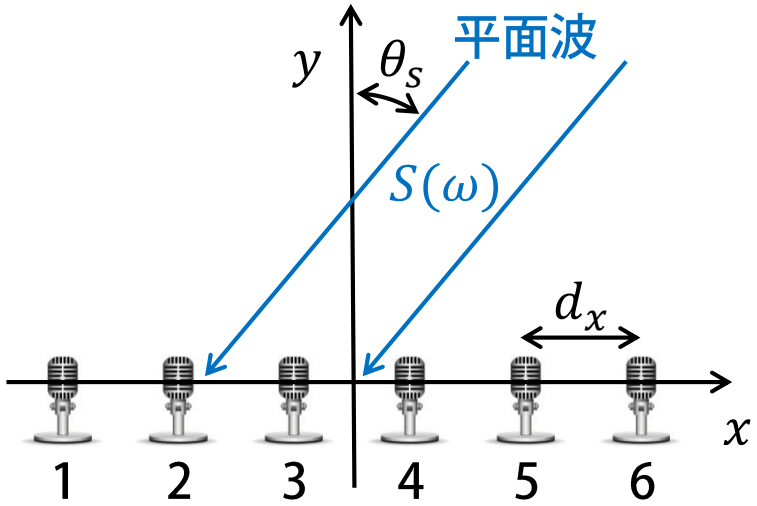
ストレート型のマイクアレイ

- マイク位置

$$\boldsymbol{p_m} = \left[\left((m-1) - \frac{M-1}{2}\right) d_x, 0, 0\right]^T$$

- 遅延

$$\tau_m = -\left((m-1) - \frac{M-1}{2}\right)\frac{d_x}{c}\sin\theta_s$$



$$\boldsymbol{z}(\omega) = [Z_1(\omega), Z_2(\omega), \cdots, Z_6(\omega)]$$

- ステアリングベクトル

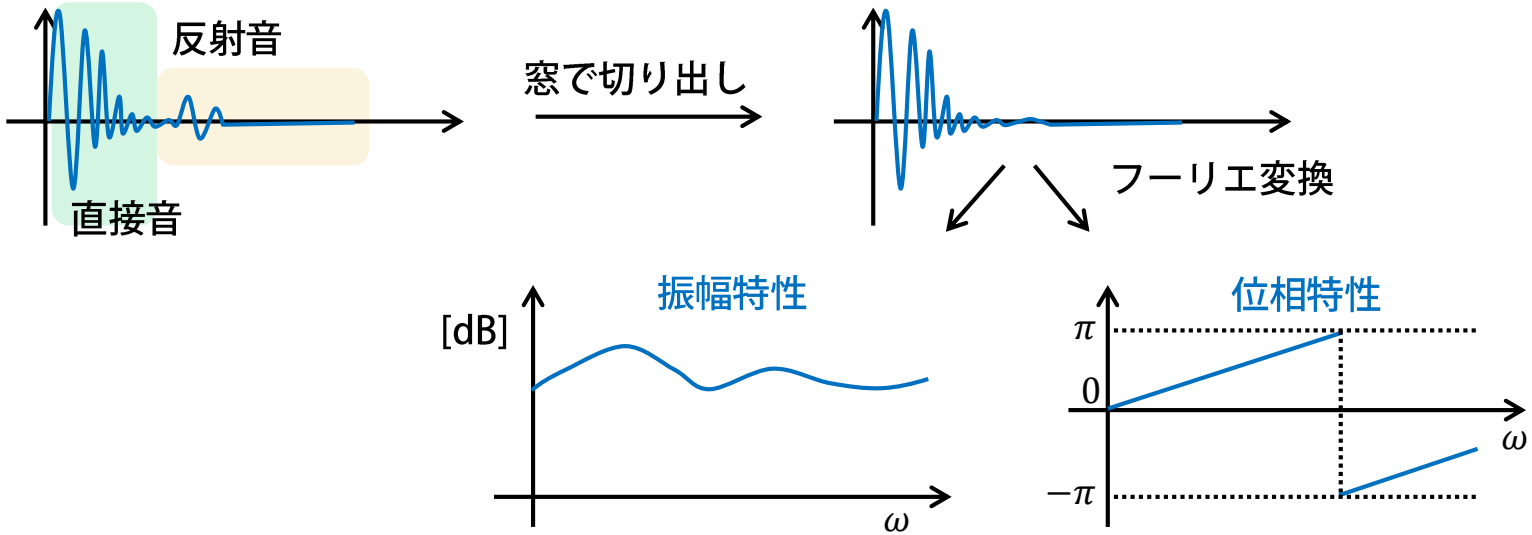
$$a_m(\omega) = \exp\left(j\left((m-1) - \frac{M-1}{2}\right)\frac{2\pi d_x}{\lambda}\sin\theta_s\right)$$

$$\boldsymbol{z}(\omega) = \boldsymbol{a}(\omega)S(\omega)$$

$$\boldsymbol{a}(\omega) = e^{-\frac{j(M-1)\psi}{2}}\left[1, e^{j\psi}, e^{j2\psi}, \cdots, e^{j(M-1)\psi}\right]^T \quad \left(\psi = \frac{2\pi d_x}{\lambda}\sin\theta_s\right)$$

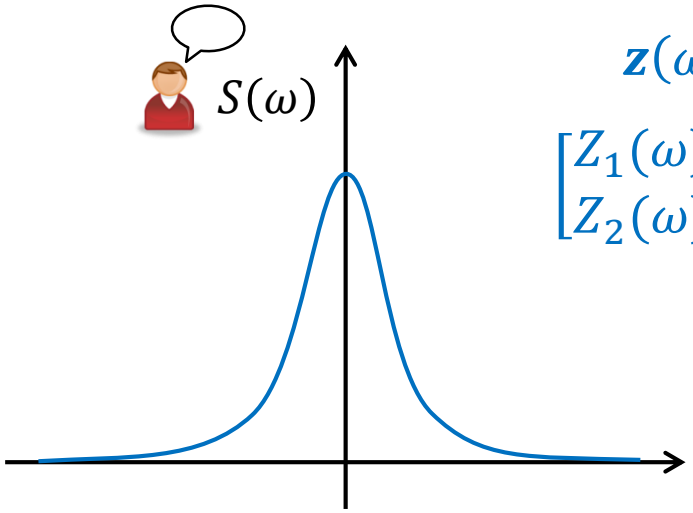
ステアリングベクトルの準備

- 幾何的に計算する (中空を仮定する場合が多い)
  - 公式使うだけ： $a_m(\omega) = \exp(-j\boldsymbol{k}^T\boldsymbol{p}) = \exp\left(j\frac{2\pi}{\lambda}\boldsymbol{u}^T\boldsymbol{p}_m\right)$
    - $\boldsymbol{u}$: 音源位置
    - $\boldsymbol{p}_m$: マイクの位置
- 実際に計測してみる
  - 直接音のみを抽出する
  - フーリエ変換を行って周波数軸に変換しておく



単一音源の観測と確率モデル

- 確率モデル $y = f(x)$ をきめて x の分布を与える
  - 観測可能な y が従う分布が分かる
  - 必要に応じてヤコビアンを計算する



We often assume

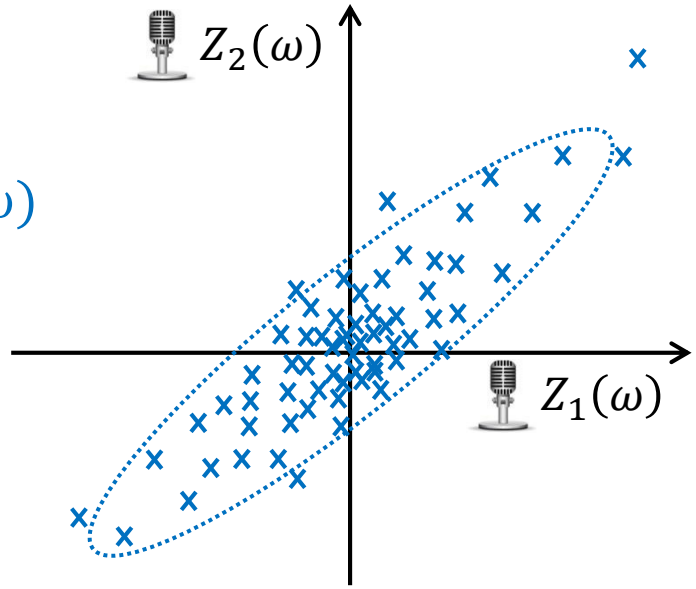
$S(\omega) \sim N_c(0,1)$

$$\boldsymbol{z}(\omega) = \boldsymbol{a}(\omega)S(\omega)$$

$$\begin{bmatrix} Z_1(\omega) \\ Z_2(\omega) \end{bmatrix} = \begin{bmatrix} a_1(\omega) \\ a_2(\omega) \end{bmatrix} S(\omega)$$

$$\boldsymbol{a}(\omega) = \begin{bmatrix} a_1 \\ \vdots \\ a_M \end{bmatrix} \equiv \begin{bmatrix} e^{-j\omega\tau_1} \\ \vdots \\ e^{-j\omega\tau_M} \end{bmatrix}$$

ステアリングベクトル

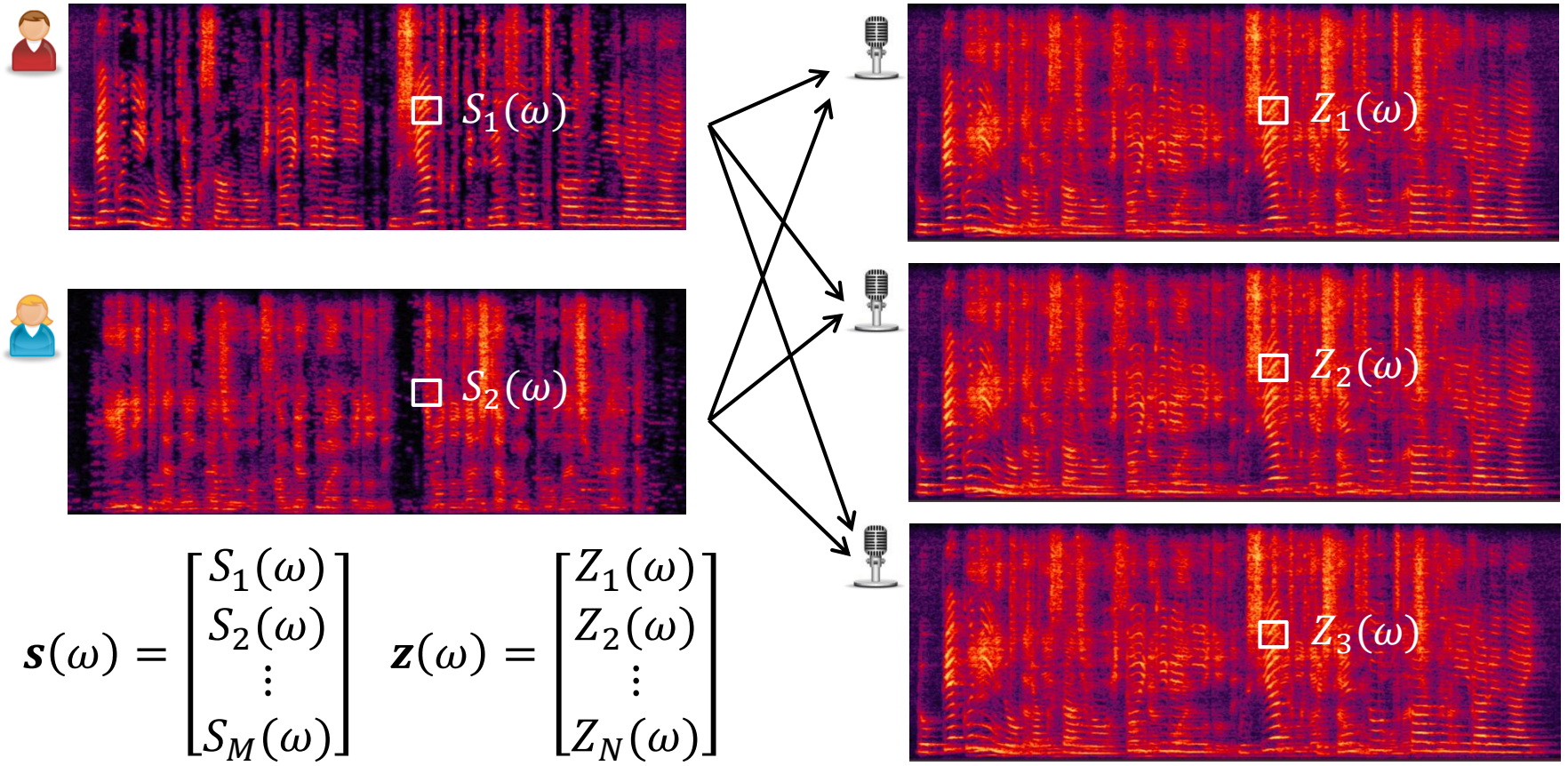


$\boldsymbol{z}(\omega) \sim N_c(0, \boldsymbol{a}(\omega)\boldsymbol{a}^H(\omega))$

空間相関行列

複数マイクによる混合音の観測

- N 個の音源を M 個のマイクで観測する



$$\boldsymbol{s}(\omega) = \begin{bmatrix} S_1(\omega) \\ S_2(\omega) \\ \vdots \\ S_M(\omega) \end{bmatrix} \quad \boldsymbol{z}(\omega) = \begin{bmatrix} Z_1(\omega) \\ Z_2(\omega) \\ \vdots \\ Z_N(\omega) \end{bmatrix}$$

線形モデル

- 観測信号は N 個の音源"イメージ"の線形和
  - 音源信号はモノラルでも、複数マイクで観測するとマルチチャネル
  - マイクアレイの地点における音源のイメージとよぶ

単一音源

$$\boldsymbol{z}_S(\omega) = \boldsymbol{a}(\omega)S(\omega)$$

複数音源

$$\boldsymbol{z}_S(\omega) = \sum_{i=1}^{N} \boldsymbol{a}_i(\omega)S_i(\omega) = \boldsymbol{A}(\omega)\boldsymbol{s}(\omega)$$

$$\boldsymbol{z}_S(\omega) = \begin{bmatrix} Z_{S1}(\omega) \\ Z_{S2}(\omega) \\ \vdots \\ Z_{SM}(\omega) \end{bmatrix}$$

$M \times N$の混合行列

$$\boldsymbol{A}(\omega) = [\boldsymbol{a}_1(\omega), \cdots, \boldsymbol{a}_N(\omega)]$$

$$\boldsymbol{s}(\omega) = \begin{bmatrix} S_1(\omega) \\ S_2(\omega) \\ \vdots \\ S_N(\omega) \end{bmatrix}$$

$\boldsymbol{a}_n(\omega)$：音源 n に対する
ステアリングベクトル

確率モデルの定式化

- 物理的現象 $\boldsymbol{z} = \boldsymbol{As} + \boldsymbol{v}$ に対する確率モデルをつくりたい
  - 決定的な信号モデル

$$p(\boldsymbol{v}) = N(\boldsymbol{v}|\boldsymbol{0}, \boldsymbol{K}) \xrightarrow[\boldsymbol{z} = \boldsymbol{As} + \boldsymbol{v}]{\text{線形変換}} \text{尤度関数：} p(\boldsymbol{z}; \boldsymbol{\Theta}) = N(\boldsymbol{z}|\boldsymbol{As}, \boldsymbol{K})$$

Find $\boldsymbol{\Theta} = \{\boldsymbol{A}, \boldsymbol{s}, \boldsymbol{K}\}$ that maximizes $p(\boldsymbol{z}; \boldsymbol{\Theta})$

  - 確率的な信号モデル

$\boldsymbol{A}$ はN個の音源の方向 $\{\theta_1, \cdots, \theta_N\}$ に依存
$\boldsymbol{\Gamma}$ はN個の音源のパワー $\{\gamma_1, \cdots, \gamma_N\}$ に依存

$$p(\boldsymbol{v}) = N(\boldsymbol{v}|\boldsymbol{0}, \boldsymbol{K}) \quad p(\boldsymbol{s}) = N(\boldsymbol{s}|\boldsymbol{0}, \boldsymbol{\Gamma}) \xrightarrow[\boldsymbol{z} = \boldsymbol{As} + \boldsymbol{v}]{\text{線形変換}} \text{尤度関数：} p(\boldsymbol{z}; \boldsymbol{\Theta}) = N(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{A\Gamma A}^H + \boldsymbol{K})$$

Find $\boldsymbol{\Theta} = \{\boldsymbol{A}, \boldsymbol{\Gamma}, \boldsymbol{K}\}$ that maximizes $p(\boldsymbol{z}; \boldsymbol{\Theta})$

$$\boldsymbol{\Gamma} = E[\boldsymbol{ss}^H] \ (= \mathrm{diag}(\gamma_1, \cdots, \gamma_N))$$

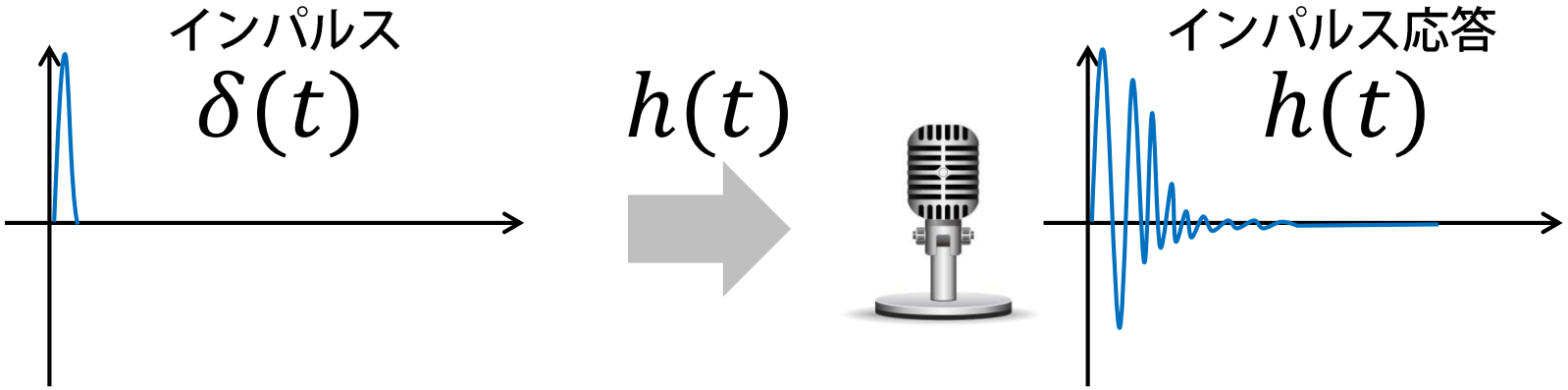
事前分布 $p(\boldsymbol{\Theta})$ を導入すれば
ベイズ的な取り扱いも可能

$$p(\boldsymbol{\Theta}|\boldsymbol{z}) = \frac{p(\boldsymbol{z}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})}{p(z)}$$

インパルス応答

インパルス応答

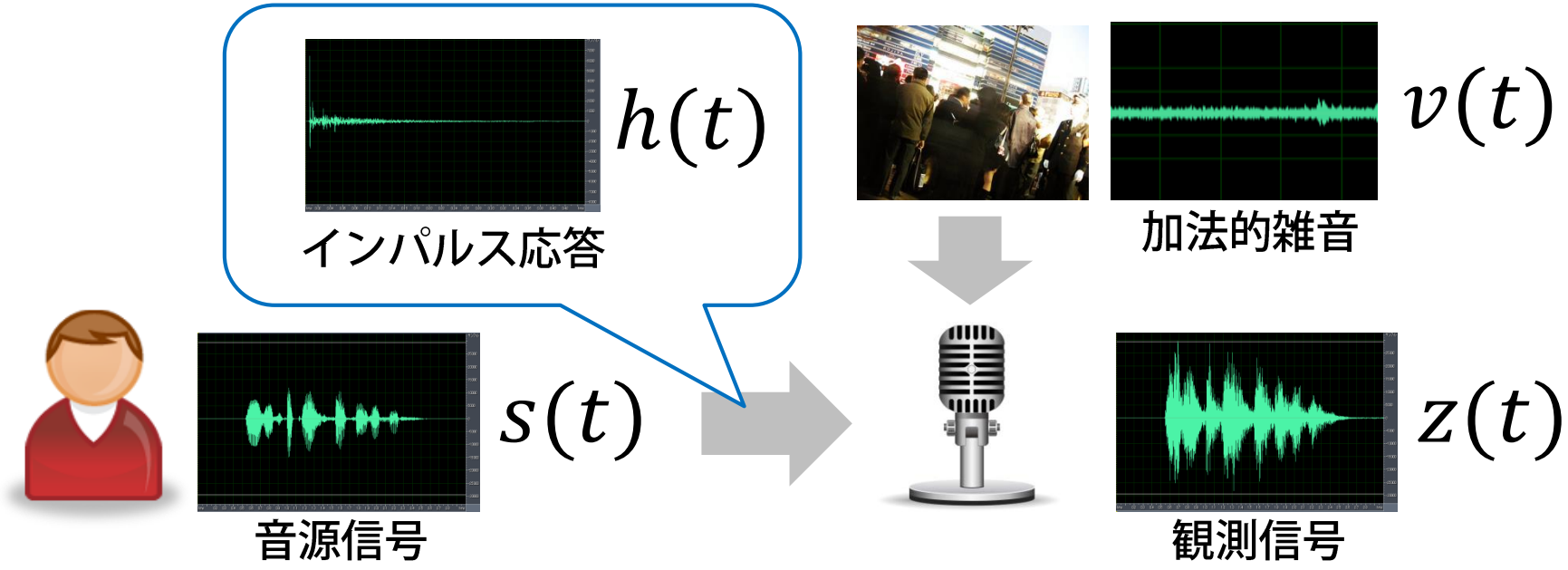
- 音源からインパルス信号が生成された場合に、各マイクで観測される信号のこと
  - 音源信号が伝播するまで、反射やまわりこみなど様々な影響を受ける
  - インパルス応答 (時間領域) = 周波数応答/伝達関数 (周波数/z領域)
  - 部屋の音響特性を表しているので、シミュレーションに有用



$$h(t) = h(t) * \delta(t)$$

基本的な定式化

- 部屋の音響特性はしばしば線形モデルで記述される
  - 音源信号＋音響特性＋加法的雑音 → 観測信号

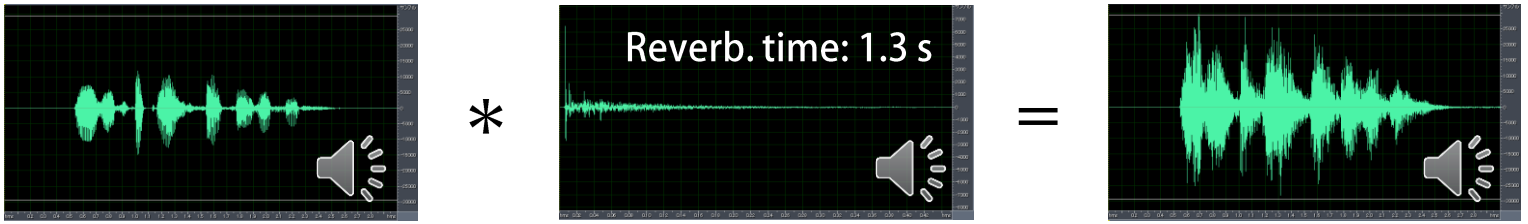


$$z(t) = h(t) * s(t) + v(t)$$

部屋の音響特性のシミュレーション

- インパルス応答を用いれば、音源信号を与えると、
  その部屋で観測されるであろう信号をシミュレートできる

音源信号　　インパルス応答　　音響信号



$*$　　$=$

時間領域の表現

$$s(t) \quad * \quad h(t) \quad = \quad z(t)$$

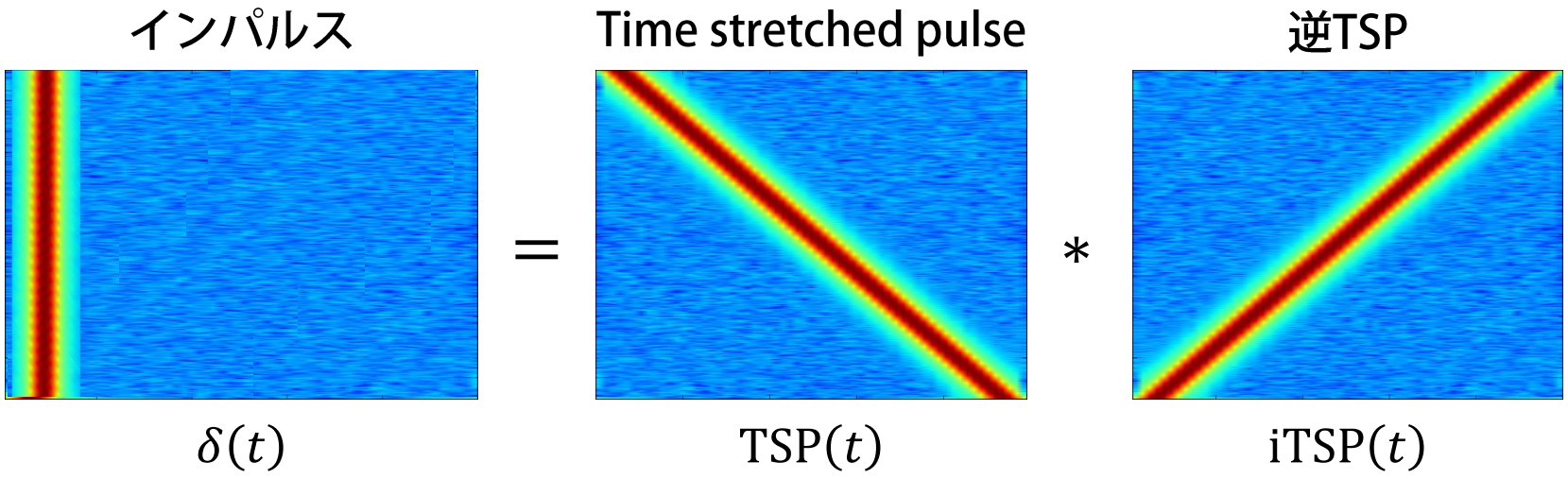
FFT　　iFFT

周波数領域の表現

$$S(\omega) \quad \times \quad H(\omega) \quad = \quad Z(\omega)$$

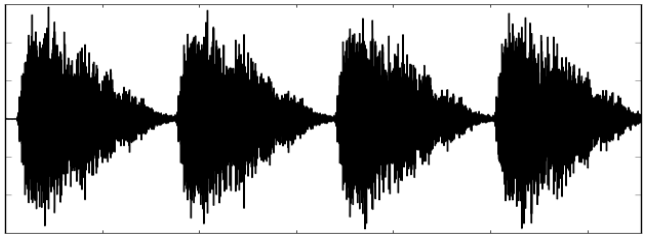
Time Stretched Pulse

- インパルス応答を測定するにはTSP信号を利用する
  - 安直な方法：インパルス信号を出力して、インパルス応答を測定
    - 全周波数帯域に等しいパワーがあるので再生が困難
  - 工夫した方法：TSP応答を測定してからインパルス応答に変換する
    - TSP信号は再生しやすく、逆TSPの畳み込みも容易

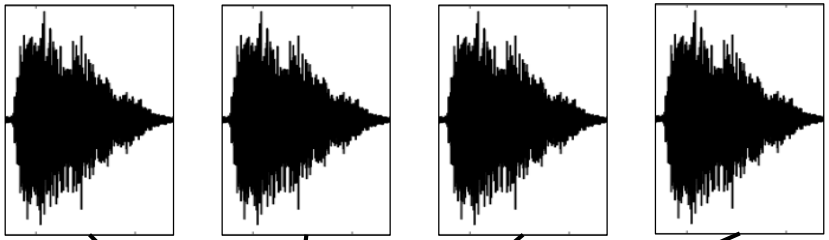


インパルス応答の測定

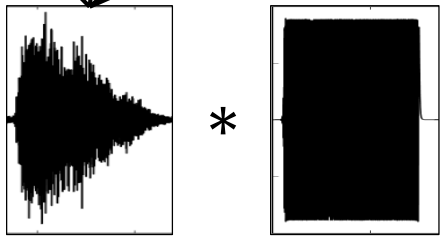
- TSP応答を何度か測定・平均してから、逆TSPを畳み込み
  - TSPとiTSPがキャンセルされてインパルス応答がでてくる



$h(t) * \mathrm{TSP}(t)$

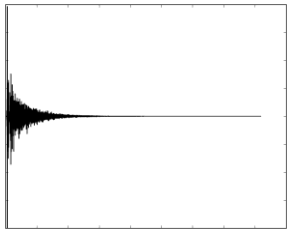


Average



$h(t) * \mathrm{TSP}(t) \quad \mathrm{iTSP}(t)$

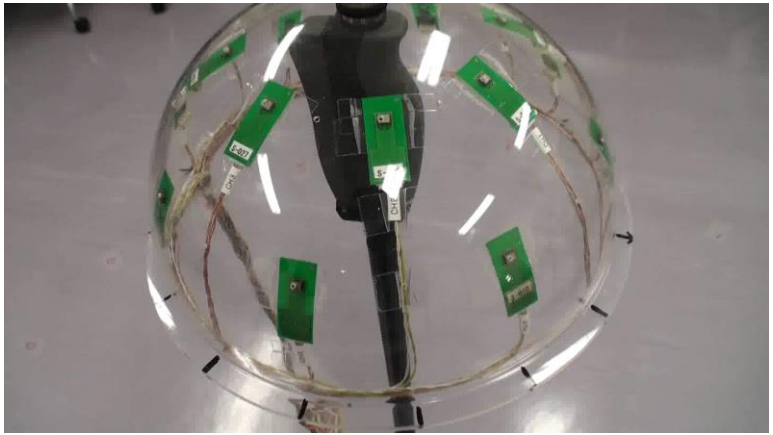
Impulse response

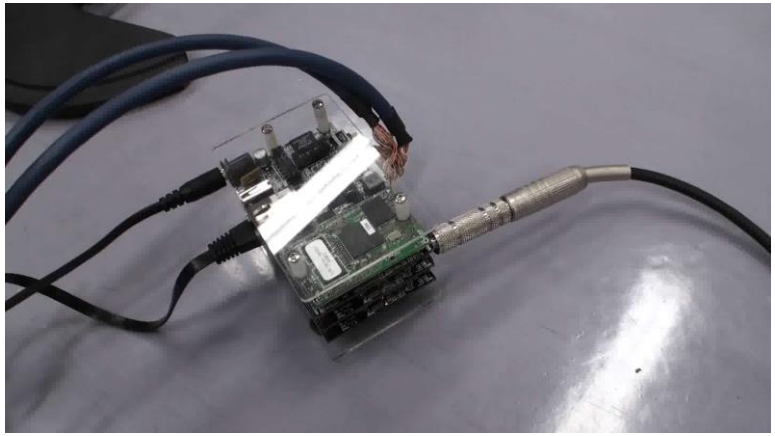


$h(t) * \cancel{\mathrm{TSP}(t)} * \cancel{\mathrm{iTSP}(t)} = h(t)$

TSP応答計測の方法

- TSP信号の生成と録音を行うための機材を準備する





スピーカーと耳栓
　TSP信号でも音量がかなり大きい

マイクアレイ
　各マイクでTSP応答を録音する

録音機材
　すべてのマイクを同期させておく

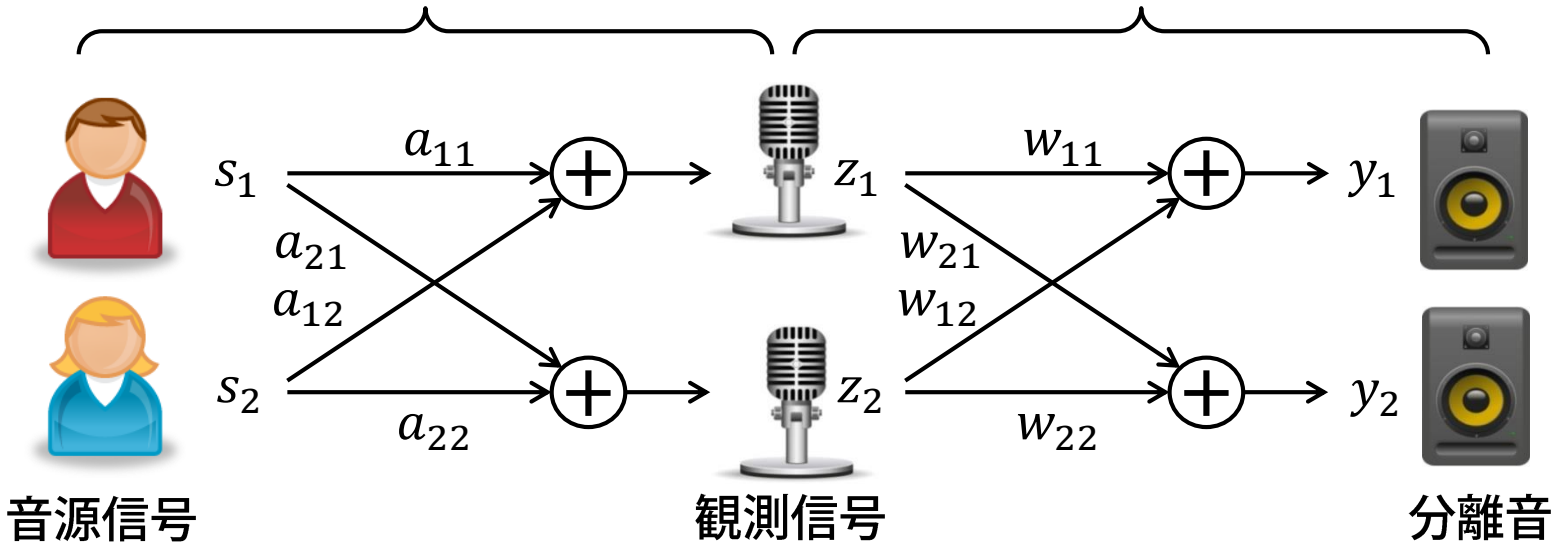
TSP Recording

独立成分分析

基本的な定式化

- 周波数領域において瞬時混合過程をモデル化する
  - N 個の音源を M 個のマイクで観測する　$M = N$ を仮定

$$\boldsymbol{z} = \boldsymbol{A}\boldsymbol{s} = \sum_{i=1}^{N} \boldsymbol{a}_i s_i \qquad \boldsymbol{y} = \boldsymbol{W}\boldsymbol{z} = \boldsymbol{W}\boldsymbol{A}\boldsymbol{s} \quad \Rightarrow \quad \text{if } \boldsymbol{W} = \boldsymbol{A}^{-1},\ \boldsymbol{y} \approx \boldsymbol{s}$$



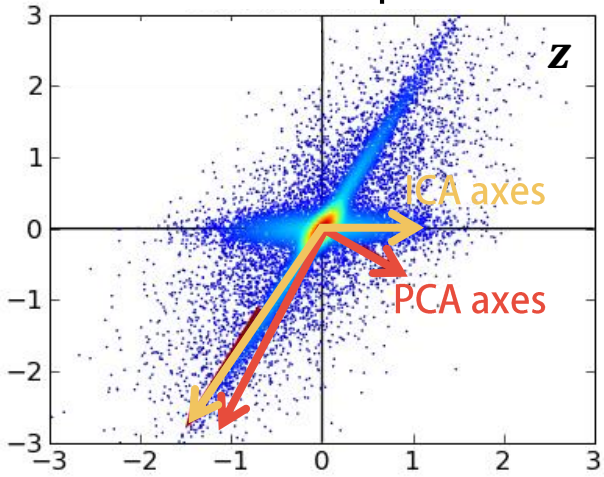
主成分分析 (PCA) と独立成分分析 (ICA)

- PCAは互いに無相関となるような次元に変換
  - 軸が直交していればOK
- ICAは互いに独立となるような次元に変換
  - 軸同士が独立でないといけない

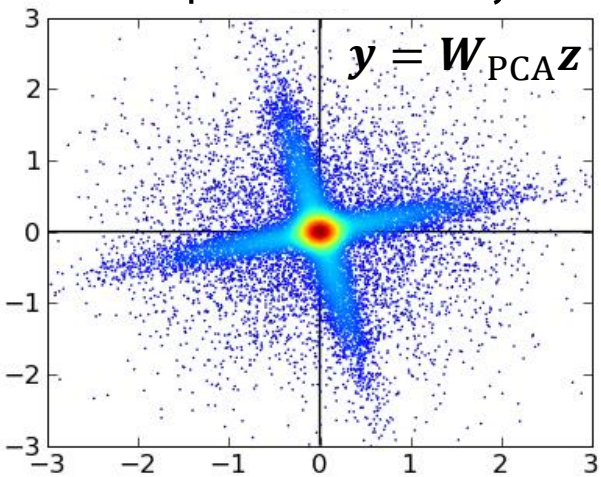
十分条件

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \boldsymbol{w}_1 z_1 + \boldsymbol{w}_2 z_2$$

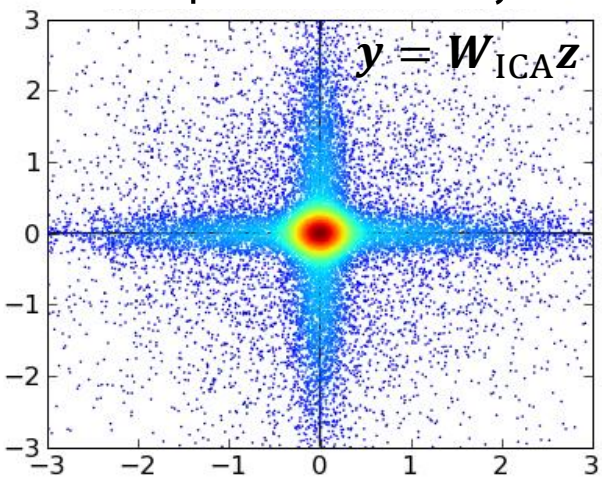
Observed space



Latent space discovered by PCA

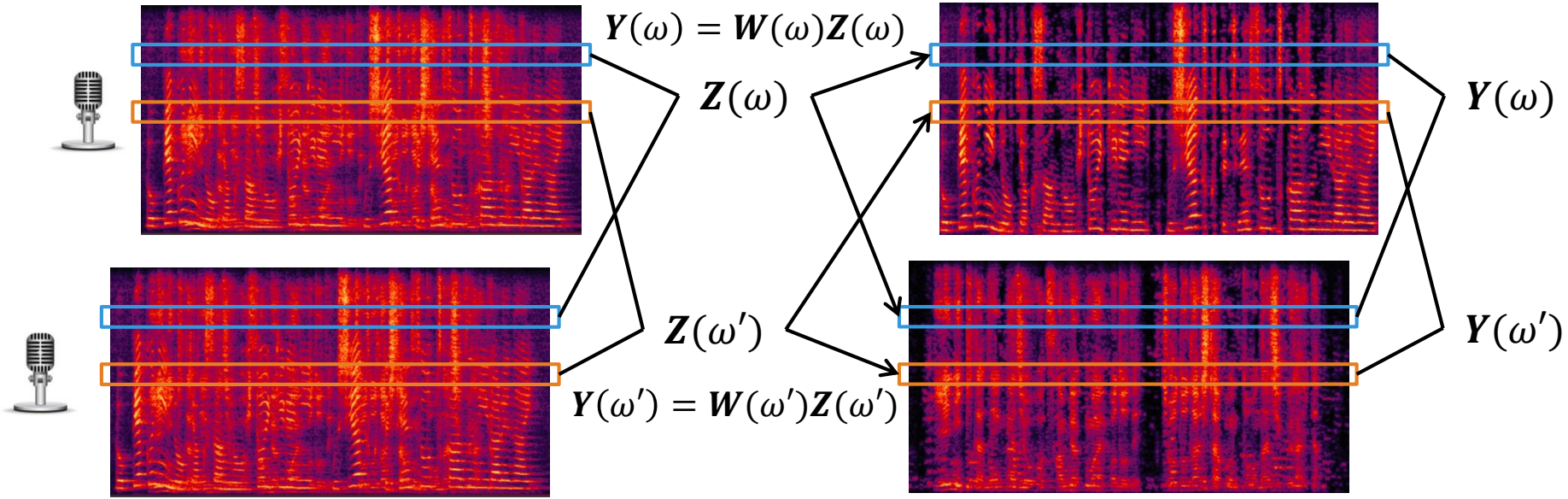


Latent space discovered by ICA



ICAの問題

- パーミュテーション問題
  - 各周波数で別々にICAを行うので、音源ごとにまとめる必要がある
- 振幅の曖昧性の問題
  - 各周波数ごとの振幅のスケールもそろえる必要がある

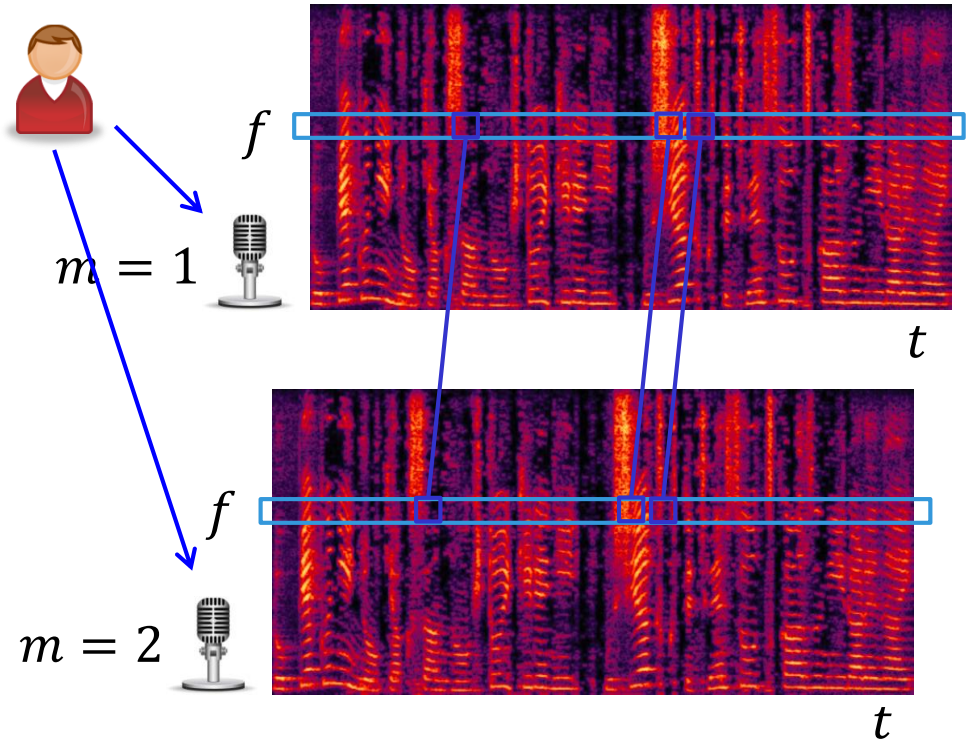


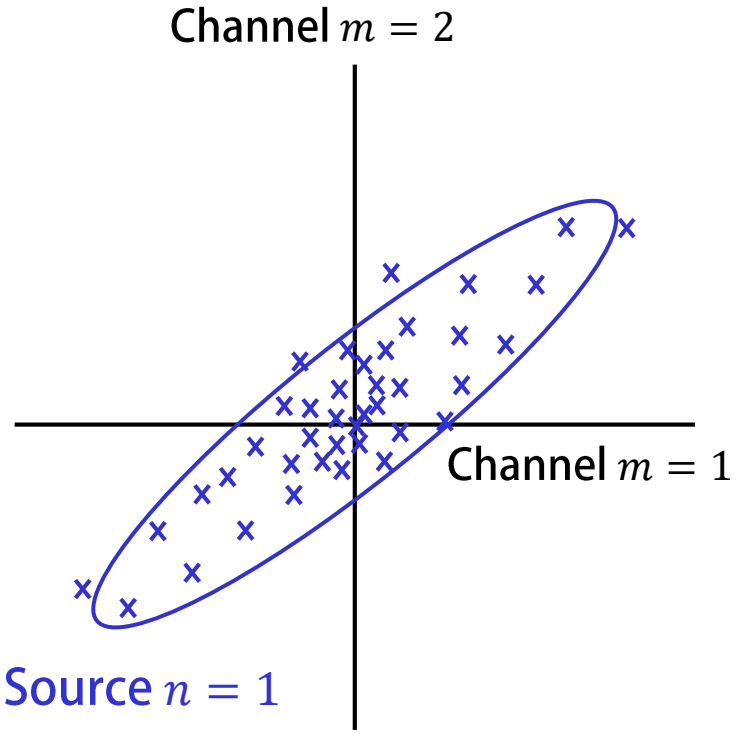
時間周波数クラスタリング

単一音源の観測

- 各音源由来のイメージは明確な空間的な構造を持つ
  - 空間相関行列をもつガウス分布に従う

観測データ：$\boldsymbol{x}_{tf} = [x_{tf1}, x_{tf2}, \cdots, x_{tfM}]$

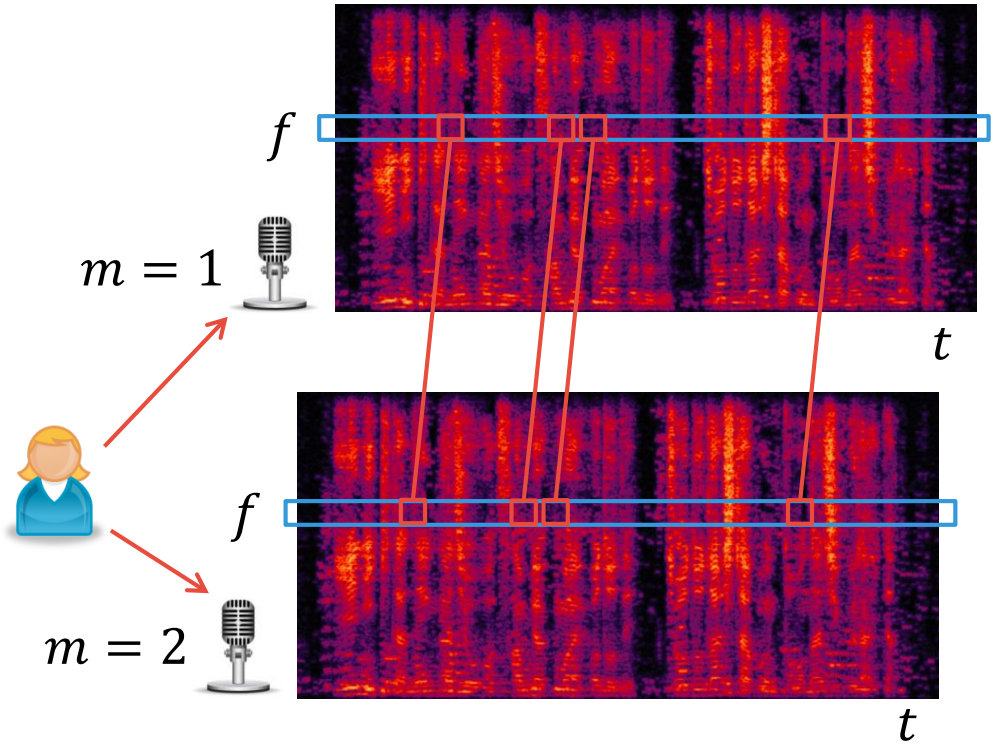


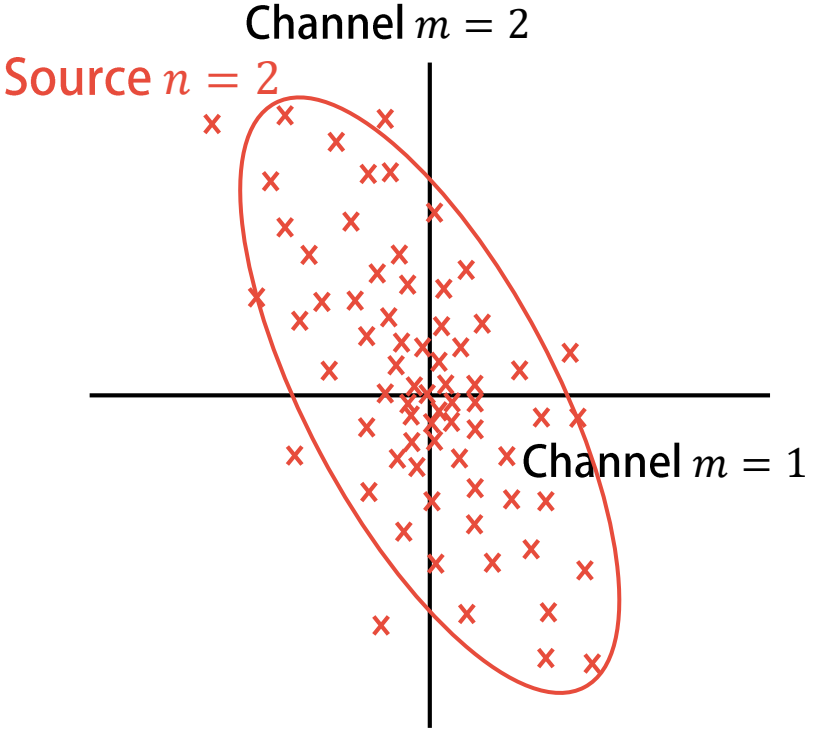


単一音源の観測

- 各音源由来のイメージは明確な空間的な構造を持つ
  - 空間相関行列をもつガウス分布に従う

観測データ：$\boldsymbol{x}_{tf} = [x_{tf1}, x_{tf2}, \cdots, x_{tfM}]$

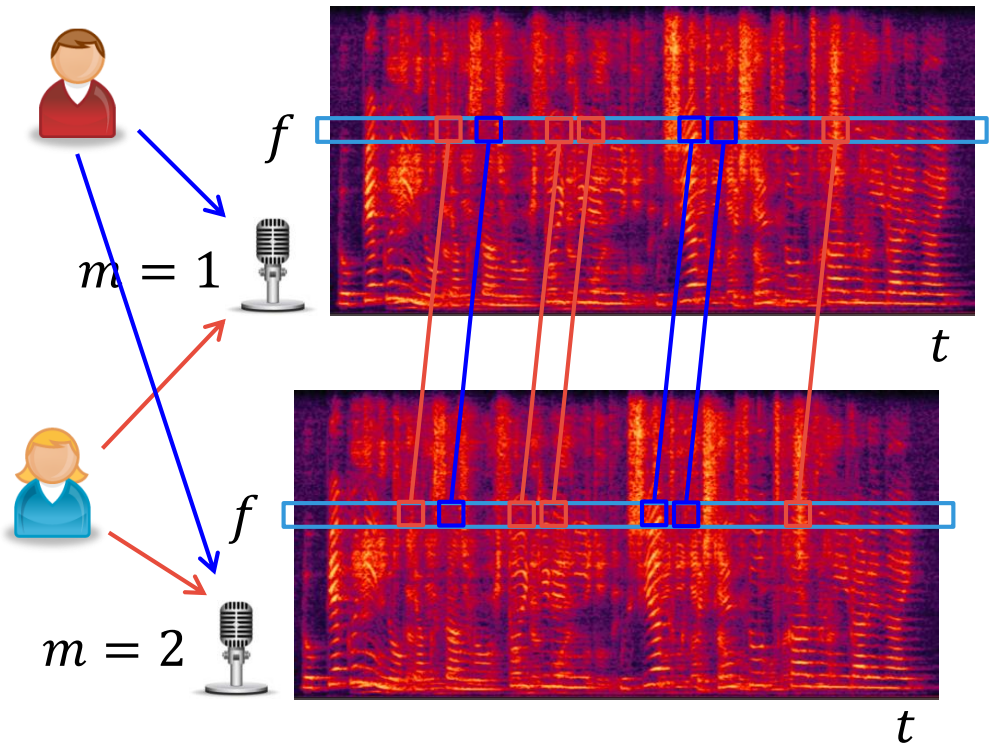


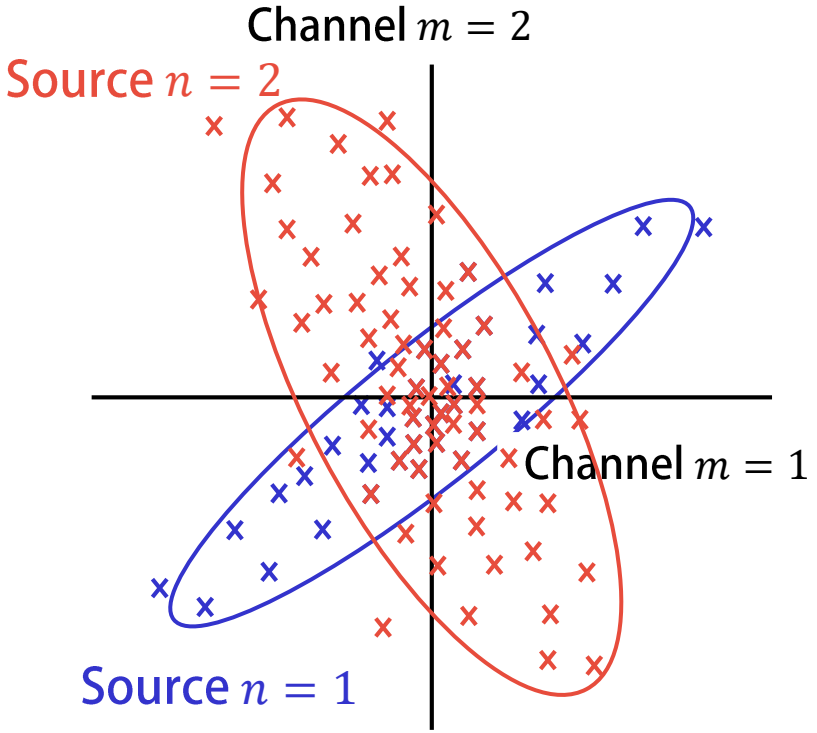


複数音源の観測

- 観測スペクトルは複数の空間的な構造がまじっている
  - 各ビンにおいてはいずれか一つの音源が優勢と仮定してみる

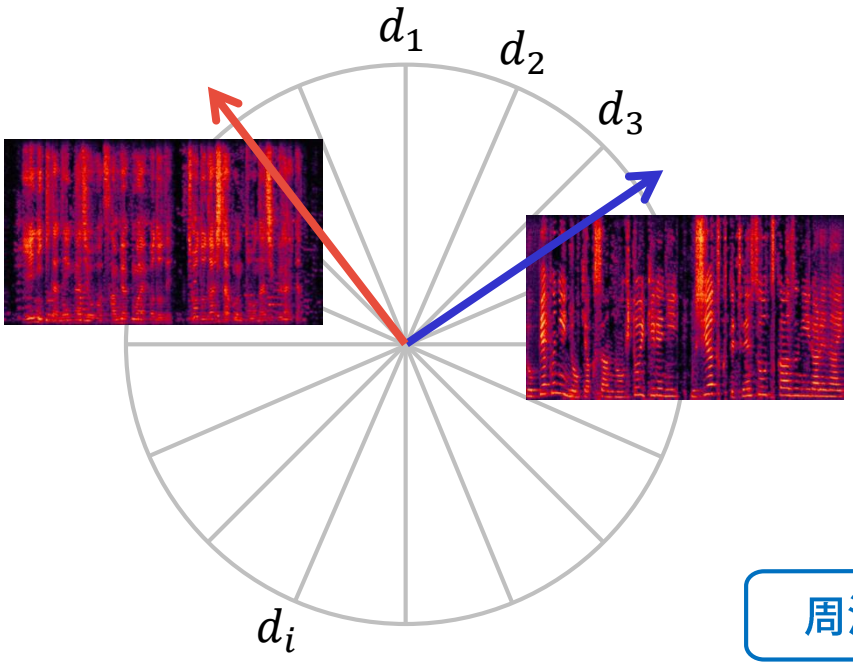
Observed data: $\boldsymbol{x}_{tf} = [x_{tf1}, x_{tf2}, \cdots, x_{tfM}]$





時間・周波数クラスタリング

- 各時間・周波数ビンをいずれかの音源にクラスタリング
  - 時間 t・周波数 f が音源 k に属する：$z_{tf} = k$
  - $\boldsymbol{H}_{fd}$：周波数 f・ 方向 d に関する空間相関行列



観測モデル [Duong 2010]

$$\boldsymbol{x}_{tf} \sim N_c\left(\boldsymbol{x}_{tf} \middle| \boldsymbol{0}, \left(\lambda_{tf}\boldsymbol{H}_{f d_{z_{tf}}}\right)^{-1}\right)$$

時間 t・周波数 f における優勢な音源

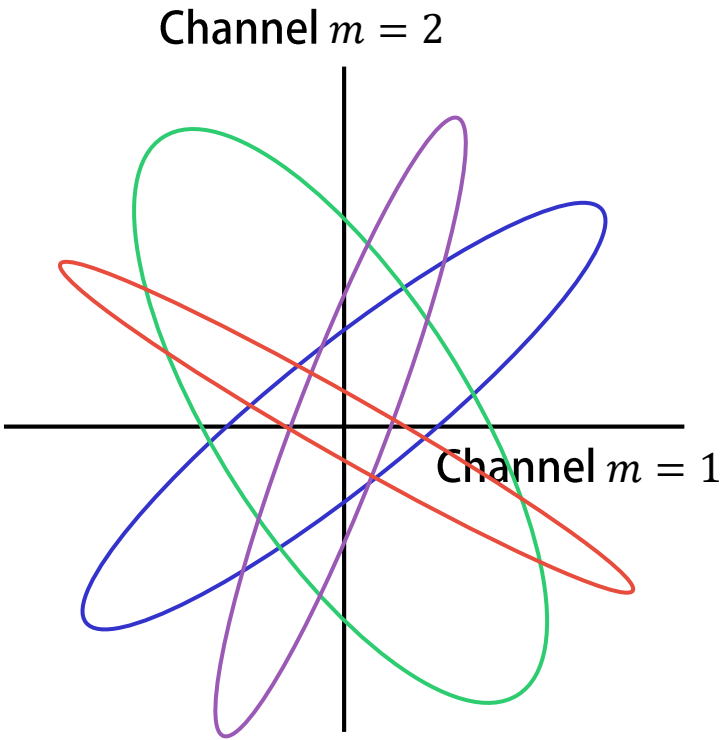
事前分布の導入とベイズ推論 [Otsuka 2014]

$$\boldsymbol{H}_{fd} \sim W_c\left(\left(\boldsymbol{a}_{fd}\boldsymbol{a}_{fd}^H + \epsilon\boldsymbol{I}\right)^{-1}, \nu_0\right)$$

周波数 f ・方向 d におけるステアリングベクトル

ノンパラメトリックベイズ拡張

- 観測信号に合わせて適切な個数の音源信号を推定
  - 理論的には無限個存在するが、そのうち一部だけが実体化される



観測モデル [Duong 2010]

$$\boldsymbol{x}_{tf} \sim N_c\left(\boldsymbol{x}_{tf} \middle| \boldsymbol{0}, \left(\lambda_{tf}\boldsymbol{H}_{f d_{z_{tf}}}\right)^{-1}\right)$$

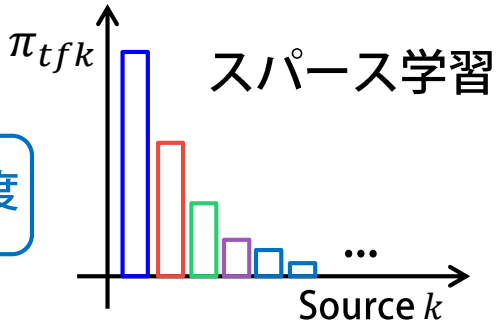
時間 t・周波数 f における優勢な音源

階層ディリクレ過程 $(k \to \infty)$ [Otsuka 2014]

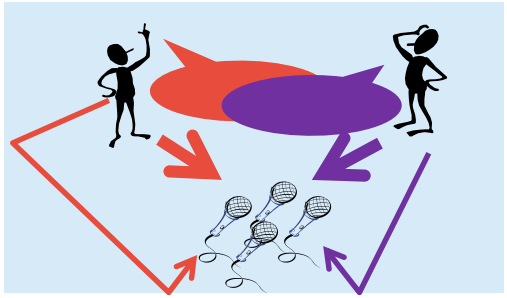
$$\boldsymbol{\pi}_{tf} \sim \mathrm{HDP}(\alpha, \gamma, \boldsymbol{\beta})$$

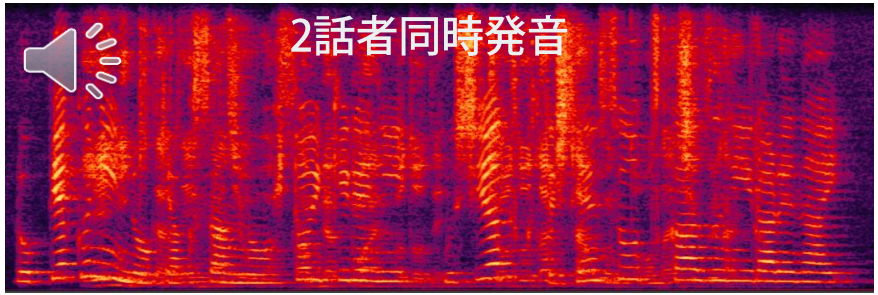
集中度　基底速度

$$z_{tf} \sim \mathrm{Categorical}(\boldsymbol{\pi}_{tf})$$

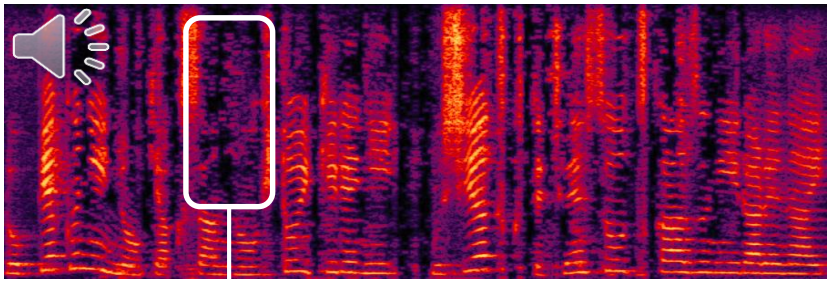


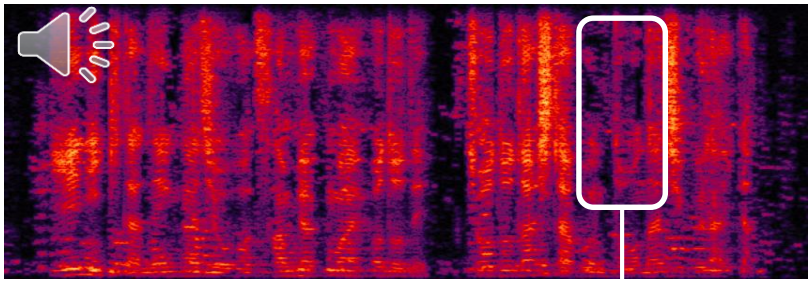
定位＋分離＋残響除去



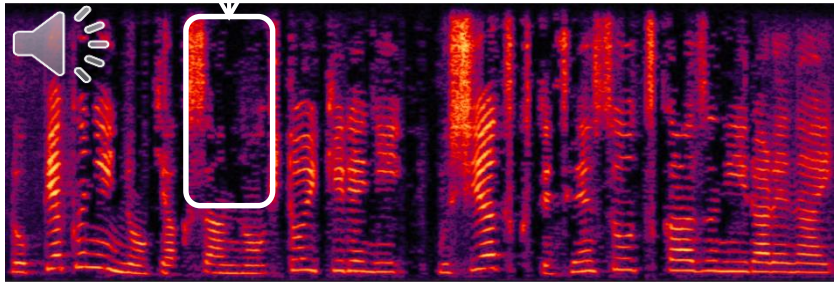


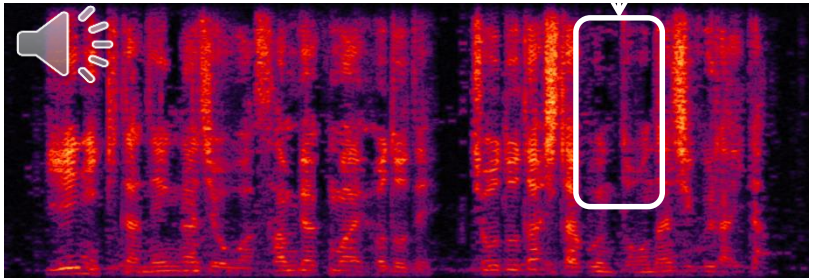
残響除去なし





残響除去あり





ポスター会場で実稼働

京大時計台国際交流ホール

マイクロホンアレイ

観測した混合音

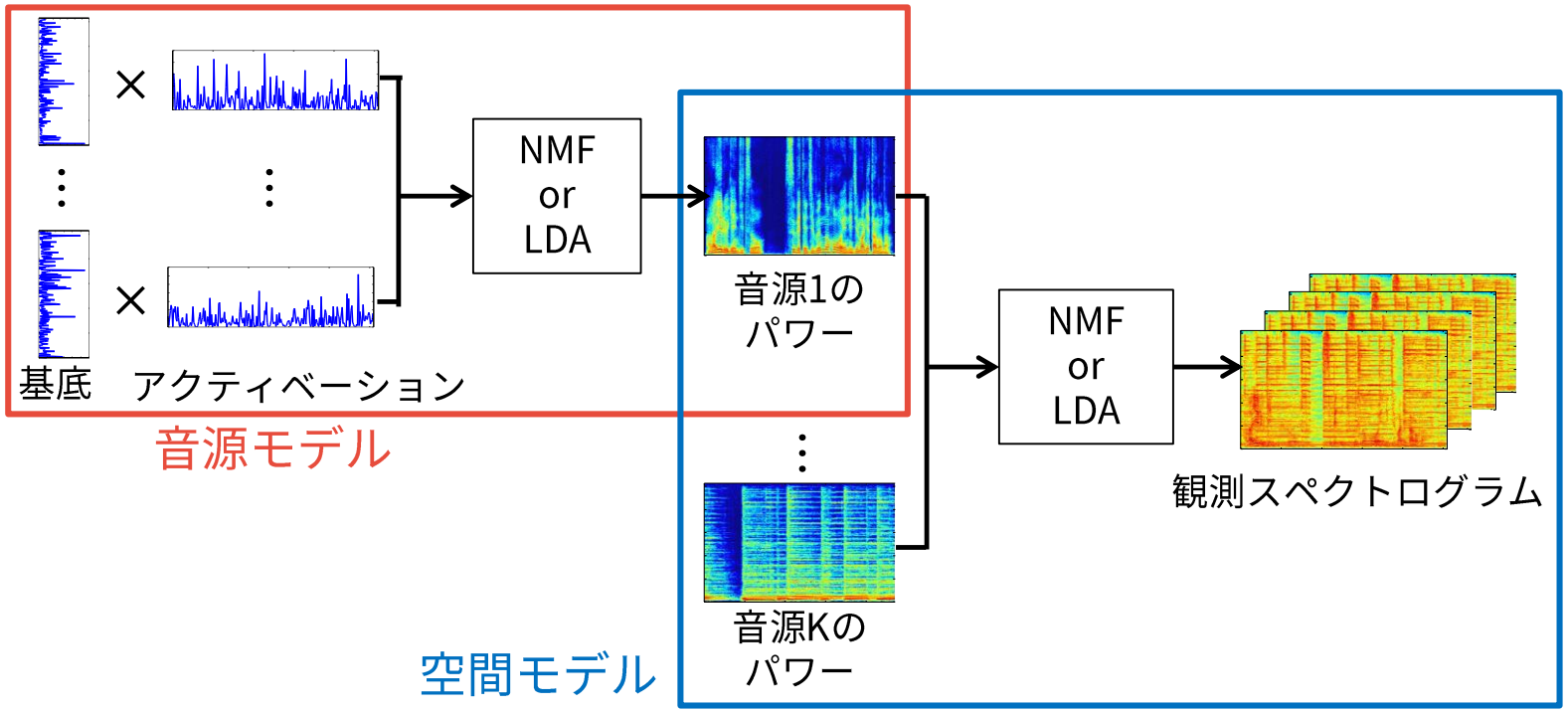
分離音

周囲の人の話し声　背景雑音

音源信号と観測信号の階層ベイズモデリング

観測スペクトログラムの階層的生成モデル

- 音源モデル：音源信号のパワーの生成過程をモデル化
- 空間モデル：音源イメージの位相をモデル化



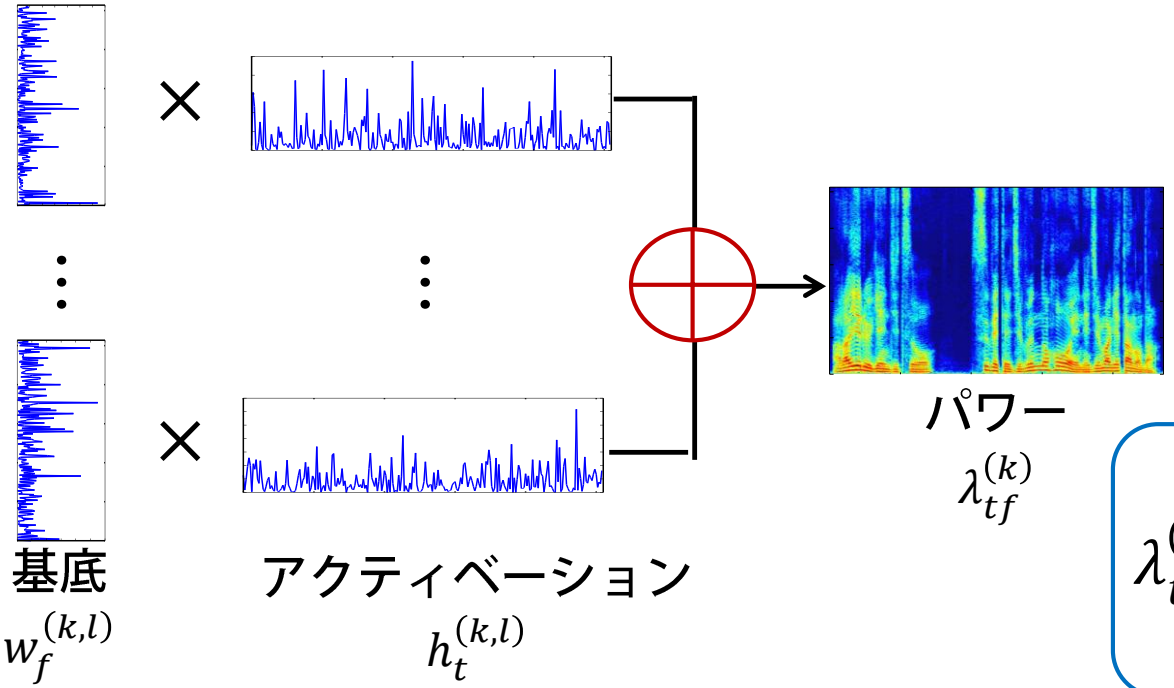
最新の統計的マイクアレイ処理

- 音源信号の生成過程と音源信号の重畳過程を同時に考慮
  - 音源モデル：各音源の音源スペクトログラムの構造を表現
    - 低ランク性が有用 (e.g., NMF, LDA)
  - 空間モデル：複数の音源がどのように混じるかを表現
    - 従来の空間相関行列が手がかり

音源モデル ＼ 空間モデル	因子モデル (factor)	混合モデル (mixture)
因子モデル(factor)	factor-factor	factor-mixture
混合モデル (mixture)	mixture-factor	mixture-mixture

音源モデル：因子モデル (NMF)

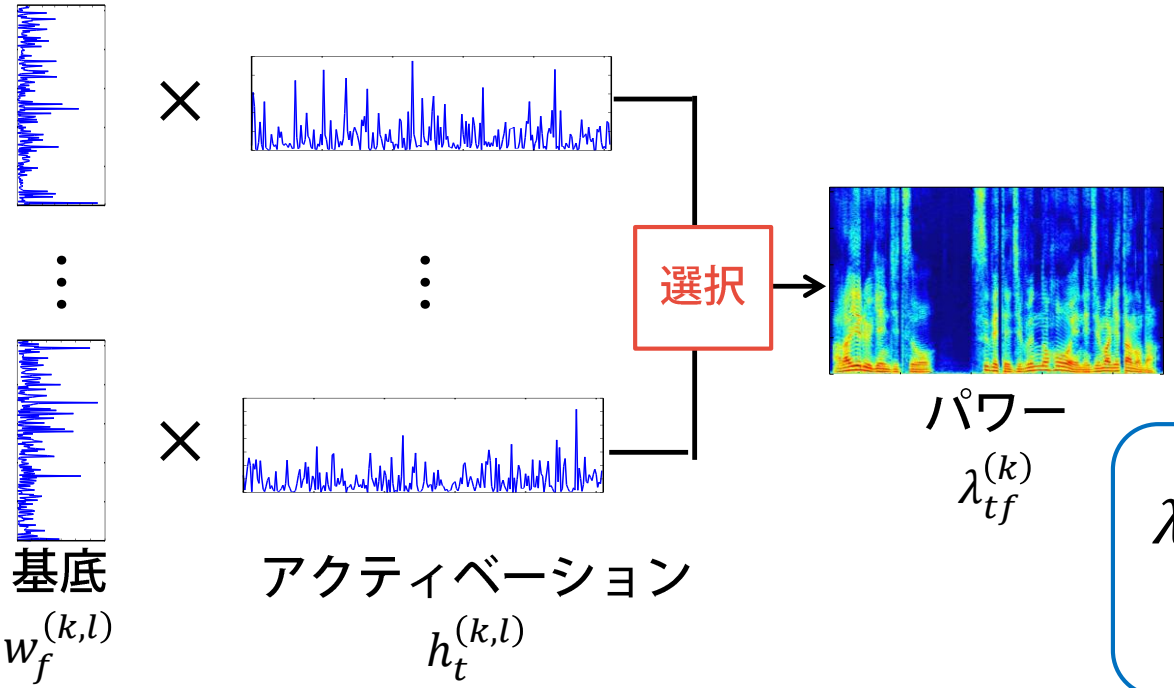
- 音源スペクトログラムを基底スペクトログラムの足し合わせで表現
  - 音源 k の時間 t・周波数 f におけるパワー λ_{tf}^{k} を、
    基底とアクティベーションの積のすべてを使って表現



$$\lambda_{tf}^{(k)} = \sum_{l=1}^{L} w_f^{(k,l)} h_t^{(k,l)}$$

音源モデル：混合モデル (LDA)

- 音源スペクトログラムを基底スペクトログラムのパッチワークで表現
  - 音源 k の時間 t・周波数 f におけるパワー λ_{tf}^{k} を、
    基底とアクティベーションの積のうちのいずれか一つで表現

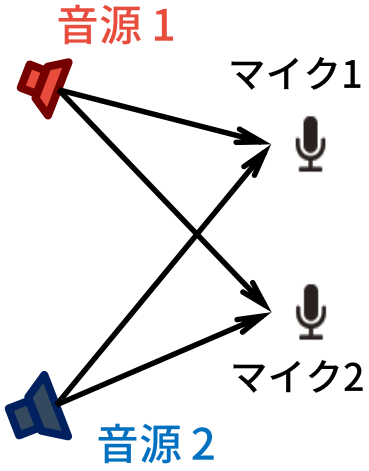


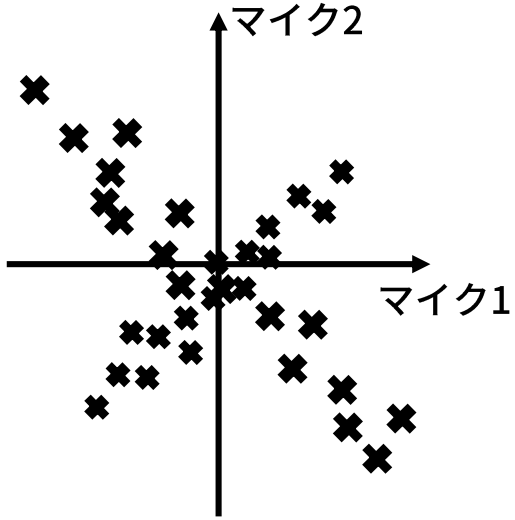
$$\lambda_{tf}^{(k)} = w_f^{(k,l')} h_t^{(k,l')}$$

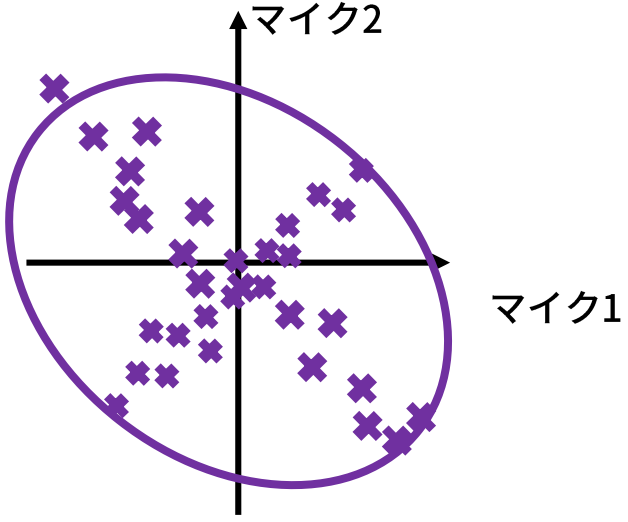
l'：選択された基底

空間モデル：因子モデル

- 観測スペクトルを音源イメージの足し合わせで表現
  - 各時間 t・周波数 f にはすべての音源が貢献







$$\boldsymbol{x}_{tf}^{(1)} \sim N_c\left(\mathbf{0}, \lambda_{tf}^{(1)} \boldsymbol{G}_f^{(1)}\right)$$

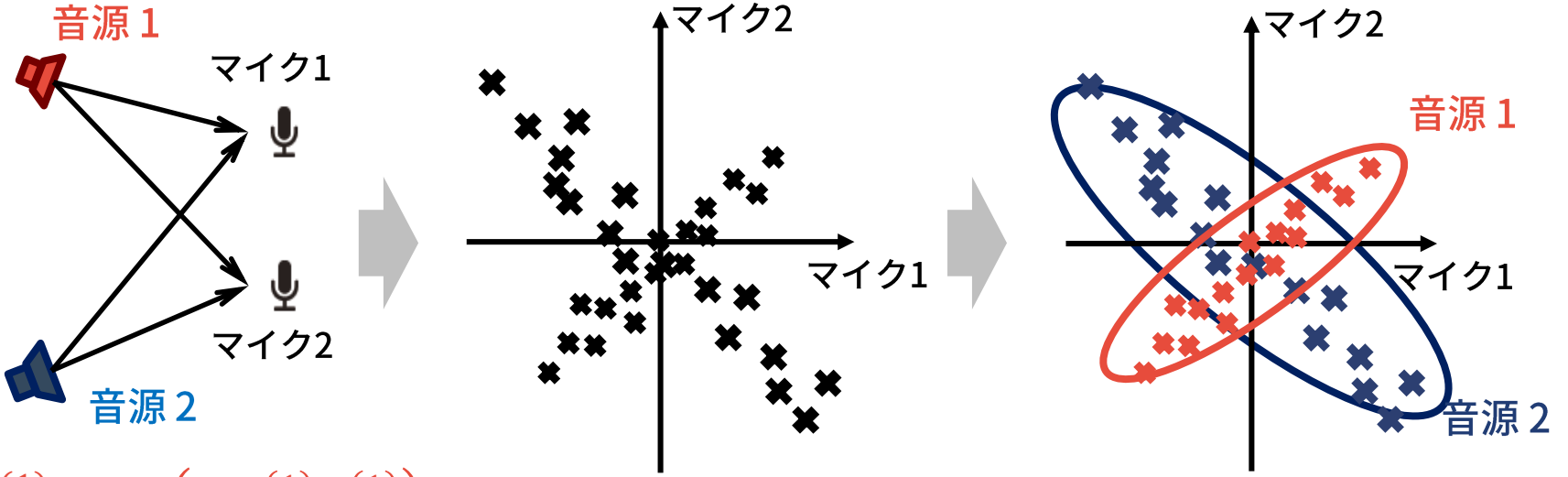
$$\boldsymbol{x}_{tf}^{(2)} \sim N_c\left(\mathbf{0}, \lambda_{tf}^{(2)} \boldsymbol{G}_f^{(2)}\right)$$

$$\boldsymbol{x}_{tf} \sim N_c\left(\mathbf{0}, \sum_k \lambda_{tf}^{(k)} \boldsymbol{G}_f^{(k)}\right)$$

観測　パワー　方向情報

空間モデル：混合モデル

- 観測スペクトルを音源イメージのパッチワークで表現
  - 各時間 t・周波数 f にはいずれか一つのが貢献



$$\boldsymbol{x}_{tf}^{(1)} \sim N_c\left(\mathbf{0}, \lambda_{tf}^{(1)} \boldsymbol{G}_f^{(1)}\right)$$

$$\boldsymbol{x}_{tf}^{(2)} \sim N_c\left(\mathbf{0}, \lambda_{tf}^{(2)} \boldsymbol{G}_f^{(2)}\right)$$

選択

$$\boldsymbol{x}_{tf} \sim \sum_k \pi_{tf}^{(k)} N_c\left(\mathbf{0}, \lambda_{tf}^{(k)} \boldsymbol{G}_f^{(k)}\right)$$

観測　混合率　パワー　方向情報

四種類の統一的な確率モデル

- すべての組み合わせの定式化に成功 [Itakura 2016, 2017]
  - 周辺化ギブスサンプリングを用いたベイズ推論
  - すべての変数を交互にサンプリング

∑の位置に注意！

空間モデル

音源モデル

	因子モデル (factor)	混合モデル (mixture)
因子モデル (factor)	$N_c\left(\mathbf{0}, \sum_k \sum_l w_f^{(k,l)} h_t^{(k,l)} \boldsymbol{G}_f^{(k)}\right)$	$\sum_k \pi_{tf}^{(k)} N_c\left(\mathbf{0}, \sum_l w_f^{(k,l)} h_t^{(k,l)} \boldsymbol{G}_f^{(k)}\right)$
混合モデル (mixture)	$N_c\left(\mathbf{0}, \sum_k \sum_l \psi_{tf}^{(k,l)} w_f^{(k,l)} h_t^{(k,l)} \boldsymbol{G}_f^{(k)}\right)$	$\sum_k \sum_l \pi_{tf}^{(k)} \psi_{tf}^{(k,l)} N_c\left(0, w_f^{(k,l)} h_t^{(k,l)} \boldsymbol{G}_f^{(k)}\right)$

まとめ

- マイクアレイ信号処理技術の基礎と最先端を紹介した
  - 基礎的な事項
    - 音の伝播のための線形システム
    - インパルス応答の意味・測定方法
  - 独立成分分析 (ICA)
    - 主成分分析 (PCA) との違い
    - パーミュテーション問題・スケール問題
  - 時間周波数マスキング
    - ノンパラメトリックベイズ拡張
  - 音源信号と観測信号の階層ベイズモデリング
    - 音源モデル＋空間モデル (それぞれ因子モデル or 混合モデル)

音楽情報処理や音響信号処理で
一緒に研究してくださる方
を探しています。
ぜひお問い合わせください。