

第17回情報論的学習理論ワークショップ(IBIS2014)

【第19回情報論的学習理論と機械学習(IBISML)研究会】

第3回IBISMLチュートリアル

音声音響信号処理のための 確率モデルと学習アルゴリズム

亀岡弘和

東京大学大学院情報理工学系研究科

日本電信電話株式会社

NTTコミュニケーション科学基礎研究所

アウトライン

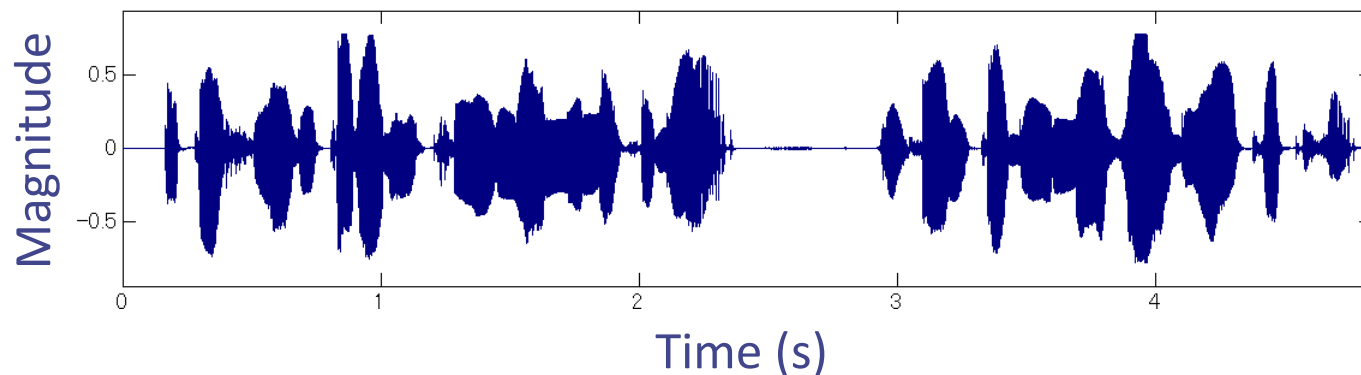
- 音声・音楽・音響信号の認識・分析・合成・変換・分離問題
- 音響信号データの特殊性を意識して設計された様々な確率モデルを紹介
 - 線形予測モデルによる音声分析
 - トラジェクトリHMMによるテキスト音声合成
 - 最尤系列変換による声質変換
 - NMFによる多重音解析
 - 多チャンネルNMFによるブラインド音源分離
 - F_0 パターン生成過程モデルによる韻律分析・合成
- 補助関数法に基づく学習アルゴリズムの導出方法の解説
 - EMアルゴリズムを一般化した最適化手法の原理
 - 補助関数設計クックブック
 - 適用例

アウトライン

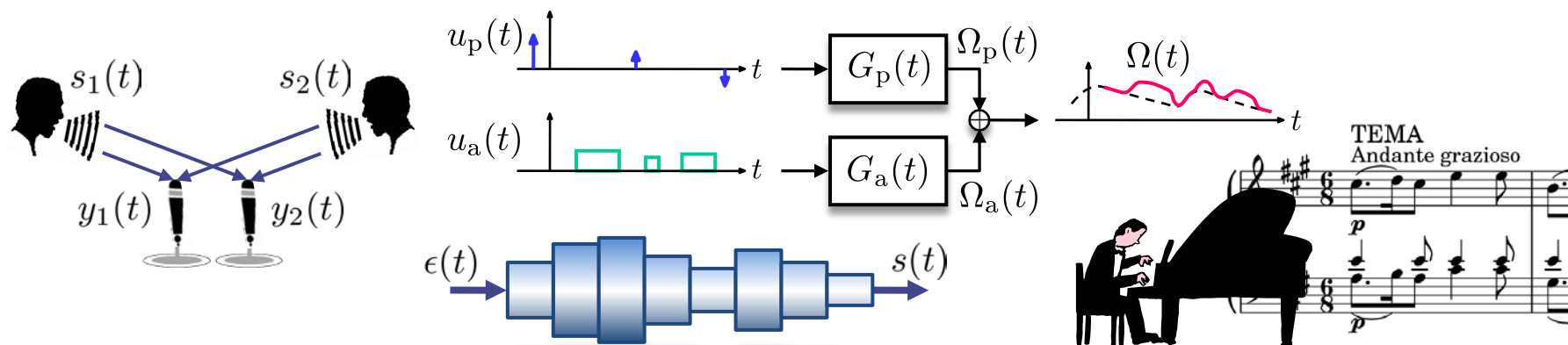
- 音声・音楽・音響信号の認識・分析・合成・変換・分離問題
- 音響信号データの特異性を意識して設計された様々な確率モデルを紹介
 - 線形予測モデルによる音声分析
 - トラジェクトリHMMによるテキスト音声合成
 - 最尤系列変換による声質変換
 - NMFによる多重音解析
 - 多チャンネルNMFによるブラインド音源分離
 - F_0 パターン生成過程モデルによる韻律分析・合成
- 補助関数法に基づく学習アルゴリズムの導出方法の解説
 - EMアルゴリズムを一般化した最適化手法の原理
 - 補助関数設計クックブック
 - 適用例

音響信号データの特殊性

1. 時系列データである



2. 物理的なメカニズムに従って生成されるデータである



音声・音楽・音響信号に関する諸問題

■ 音声・音楽・音響信号の認識・分析・合成・変換・分離

	認識	分析	合成	変換	分離
音声					
音楽			—	—	
音響					

⇒「識別」、「回帰」、「逆問題」のいずれかに相当

音声・音楽・音響信号に関する諸問題

■ 音声・音楽・音響信号の認識・分析・合成・変換・分離

	認識	分析	合成	変換	分離
音声					
音楽			—	—	
音響					

⇒「識別」、「回帰」、「逆問題」のいずれかに相当

音声・音楽・音響の認識

■ 音声の認識

- 音声認識, 話者認識, 音声区間検出, 言語認識, ...

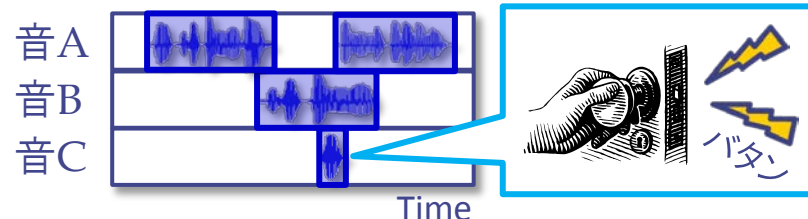


■ 音楽の認識

- 調認識, 和音認識, 構造解析, タギング, 楽譜整合, ジャンル認識, ...

■ 音響の認識

- 音響イベント検出, ...

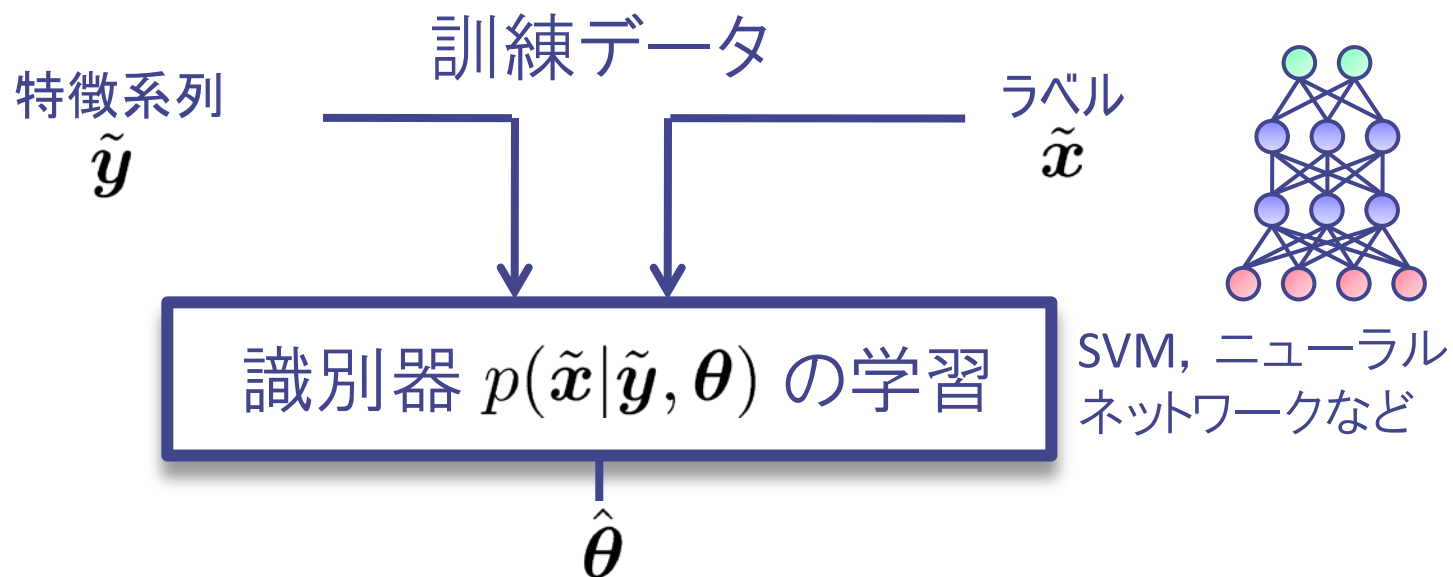


音響信号データ(時系列) y を入力として,
ラベルまたはラベル系列 x を割り当てる問題

⇒ 識別問題

識別問題

学習



入力データ

特徴系列
 y

識別
 $p(x|y, \hat{\theta})$

ラベル
 $\hat{x}$

識別

音声・音楽・音響信号に関する諸問題

■ 音声・音楽・音響信号の認識・分析・合成・変換・分離

	認識	分析	合成	変換	分離
音声					
音楽			—	—	
音響					

⇒「識別」、「回帰」、「逆問題」のいずれかに相当

テキスト音声合成

■ 入力テキストから音声特徴量系列を予測して信号を合成

- 声道特徴量：音韻性や声質に関係
- 声帯特徴量：声の高さや声のかすれに関係

■ 応用例

- 音声自動翻訳
- 医療応用
 - cf) “Voice Bank Project” (ALS等で発声能力を失った個人の声を再現)
- 外国語学習支援
- 擬人化エージェント



[Yamagishi+2012]

テキストラベル y を
音声特徴系列 x に変換する問題

⇒ 回帰問題

HMMによるテキスト音声合成 [Yoshimura+1999]

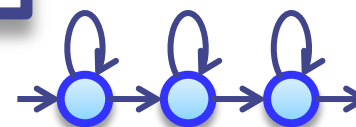
学習

訓練データ
(言語情報付き音声データ)

テキストラベル $\tilde{y}$ $\tilde{x}$ 音声特徴系列

回帰モデル $p(\tilde{x}|\tilde{y}, \theta)$ の学習

HMMで
モデル化



$\hat{\theta}$

入力データ

テキストラベル y

回帰モデル
 $p(x|y, \hat{\theta})$

$\hat{x}$ 合成音声の
特徴系列

予測

音声変換

■ 音声特徴量を異なる話者や発話様式のものに変換

- 声道特徴量: 音韻性や声質に関係
- 声帯特徴量: 声の高さや声のかすれに関係

■ 応用例

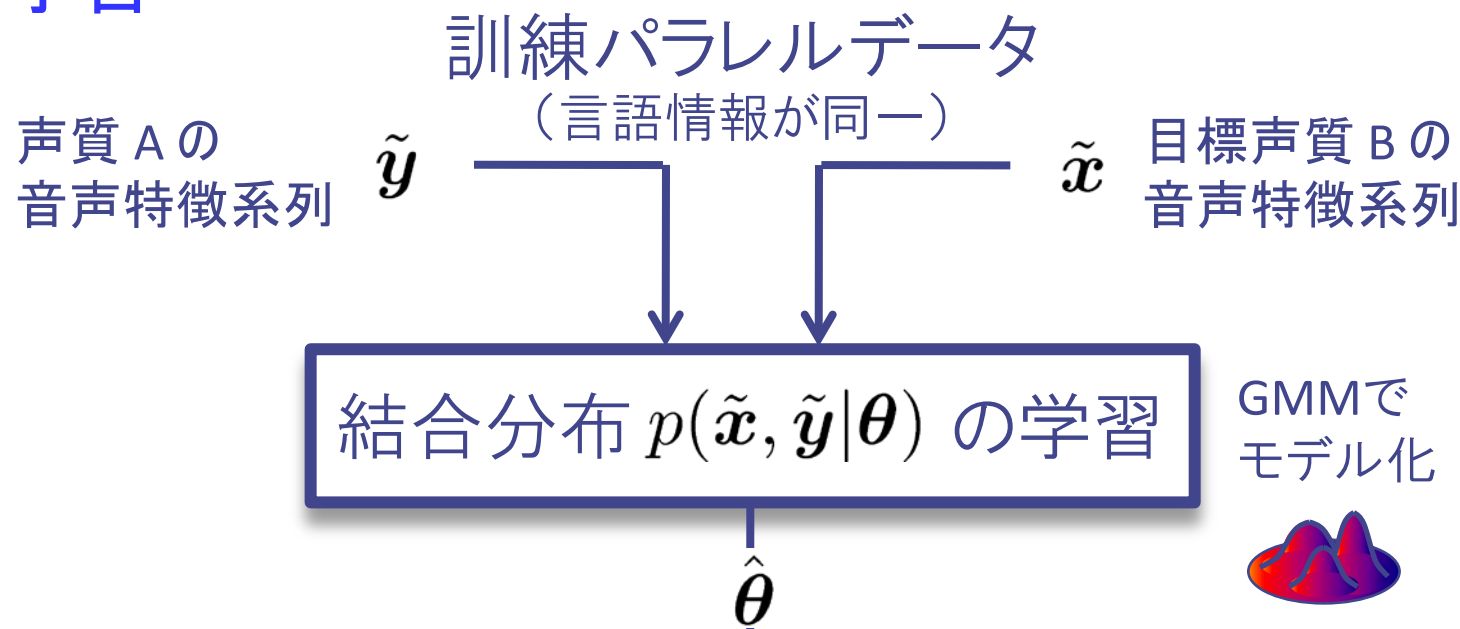
- 体内伝導音声通話 [Toda+2012]
- 発声障害補助 [Nakamura+2012]
- ボイスチェンジャー [Doi+2014]
- 帯域拡張 [Jax & Vary 2003]

音声特徴系列 y を
音声特徴系列 x に変換する問題

⇒ 回帰問題

統計的手法による声質変換 [Stylianou+1998]

学習



入力データ

声質 A の入力
音声特徴系列

y

回帰モデル
 $p(x | y, \hat{\theta})$

$\hat{x}$

変換音声の
特徴系列

変換

音声・音楽・音響信号に関する諸問題

■ 音声・音楽・音響信号の認識・分析・合成・変換・分離

	認識	分析	合成	変換	分離
音声					
音楽			—	—	
音響					

⇒「識別」、「回帰」、「逆問題」のいずれかに相当

音声・音楽信号の分析

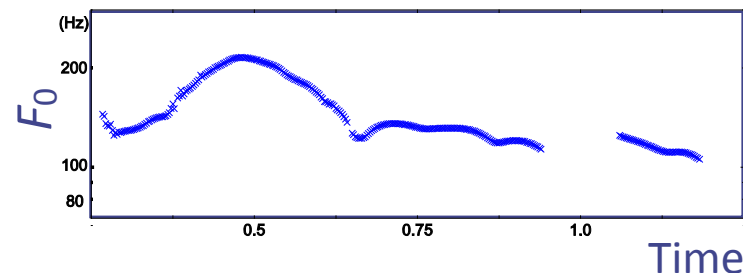
■ 音声分析

■ 声道パラメータ推定

(※“STRAIGHT” [Kawahara+1998-2014] が有名)

基本周波数(F_0)推定, ...

応用例: 音声符号化, 音声認識・合成の特徴量抽出



■ 音楽信号の分析

■ 多重ピッチ推定, ビート推定, オンセット推定, ...

(※MIREXという各種分析手法の国際コンテストが毎年行われている)

応用例: 自動採譜, 音楽検索・推薦システムの特徴量抽出



音響信号 y の生成過程パラメータ θ
を推定する問題

⇒ 逆問題

音声・音楽・音響信号の分離

■ 音声音響信号の分離

- 異なる位置に配置された複数マイクの観測信号から各音源信号を分離(詳細は後述)

- 残響除去 = 直接音と残響音の分離

応用例: ハンズフリー音声認識, 音声強調

■ 音楽信号の分離

- モノラルまたはステレオの観測信号から各楽音信号を分離

応用例: 自動採譜, 音響信号加工



音響信号 y の生成過程パラメータ θ
を推定する問題

⇒ 逆問題

音声・音楽・音響信号に関する諸問題

■ 音声・音楽・音響信号の認識・分析・合成・変換・分離

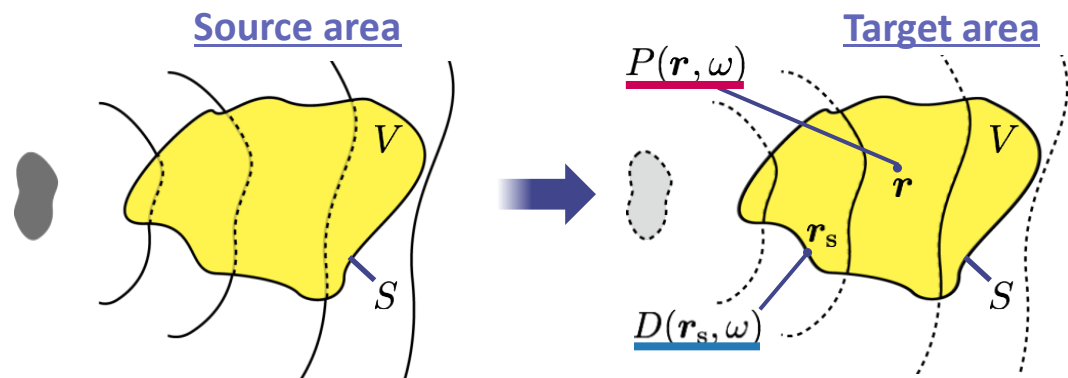
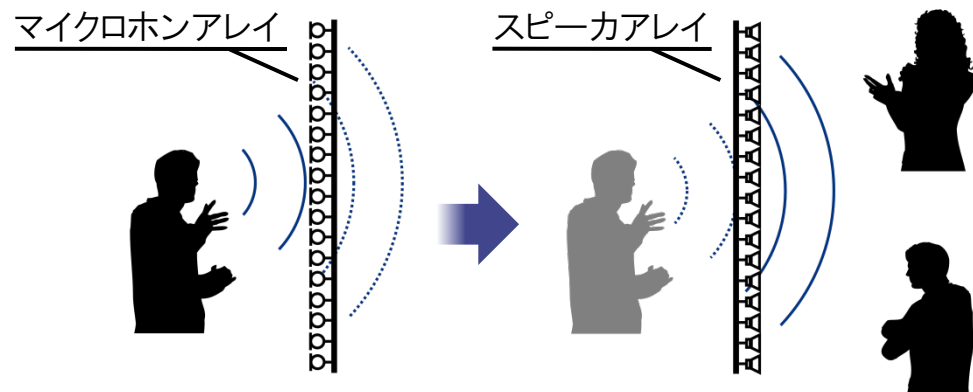
	認識	分析	合成	変換	分離
音声					
音楽			—	—	
音響					

⇒「識別」、「回帰」、「逆問題」のいずれかに相当

音場の分析・合成・変換(音場再現)

■波面合成 cf) [Koyama2012]

- マイクロホン＆スピーカアレイを用いて空間的な音場を再現
- 音場の物理式に基づきスピーカで駆動すべき信号を推定する**逆問題**



Kirchhoff-Helmholtz積分方程式: 境界面 S から V 内の音場を決める物理式

$$\underbrace{P(\mathbf{r}, \omega)}_{\text{内部音場}} = \iint_S \left(\underbrace{G_{3D}(\mathbf{r} - \mathbf{r}_s, \omega)}_{\substack{\text{3次元自由空間} \\ \text{Green関数}}} \underbrace{\frac{\partial P(\mathbf{r}_s, \omega)}{\partial n}}_{\text{音圧勾配}} - \underbrace{P(\mathbf{r}_s, \omega)}_{\text{音圧}} \underbrace{\frac{\partial G_{3D}(\mathbf{r} - \mathbf{r}_s, \omega)}{\partial n}}_{\text{音圧勾配}} \right) dS$$

パラメータ学習問題としての逆問題

学習

観測データ y



事後分布 $p(\theta|y)$ の学習



生成過程
パラメータ $\hat{\theta}$

解きたい逆問題
に応じて y と θ
の生成過程を
モデル化

$$p(\theta|y) \\ \propto p(y|\theta)p(\theta)$$

逆問題 = 観測データをただ一度だけ観測された
学習データと捉えた学習問題

アウトライン

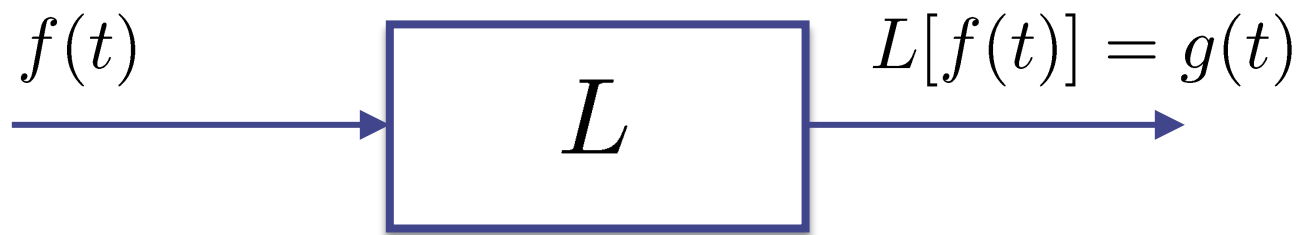
- 音声・音楽・音響信号の認識・分析・合成・変換・分離問題
- 音響信号データの特異性を意識して設計された様々な確率モデルを紹介
 - 線形予測モデルによる音声分析
 - トラジェクトリHMMによるテキスト音声合成
 - 最尤系列変換による声質変換
 - NMFによる多重音解析
 - 多チャンネルNMFによるブラインド音源分離
 - F_0 パターン生成過程モデルによる韻律分析・合成
- 補助関数法に基づく学習アルゴリズムの導出方法の解説
 - EMアルゴリズムを一般化した最適化手法の原理
 - 補助関数設計クックブック
 - 適用例

アウトライン

- 音声・音楽・音響信号の認識・分析・合成・変換・分離問題
- 音響信号データの特殊性を意識して設計された様々な確率モデルを紹介
 - 線形予測モデルによる音声分析
 - トラジェクトリHMMによるテキスト音声合成
 - 最尤系列変換による声質変換
 - NMFによる多重音解析
 - 多チャンネルNMFによるブラインド音源分離
 - F_0 パターン生成過程モデルによる韻律分析・合成
- 補助関数法に基づく学習アルゴリズムの導出方法の解説
 - EMアルゴリズムを一般化した最適化手法の原理
 - 補助関数設計クックブック
 - 適用例

参考資料：信号処理の基礎

■線形時不変システム



■線形性

$$L[C_1 f_1(t) + C_2 f_2(t)] = C_1 L[f_1(t)] + C_2 L[f_2(t)]$$

■時不変性

$$L[f(t - \tau)] = g(t - \tau)$$

■電気回路，力学系，音響系など多くの物理システムのモデル

参考資料: 信号処理の基礎

■ 線形時不変システムの基本演算は「畳み込み」

$$g(t) = L[f(t)] = \int_{-\infty}^{\infty} f(\tau)h(t - \tau)d\tau = f(t) * h(t)$$

畳み込みを
表す記号

なぜ? $\Rightarrow$ 前述の線形性と時不変性から導ける

■ 略証

- デルタ関数に対する線形システムの出力 $\Rightarrow h(t; \tau) := L[\delta(t - \tau)]$
- $f(t)$ を無数のデルタ関数の和で表現 $\Rightarrow f(t) = \int_{-\infty}^{\infty} f(\tau)\delta(t - \tau)d\tau$
- 線形性より $L[f(t)] = \int_{-\infty}^{\infty} f(\tau)L[\delta(t - \tau)]d\tau$
- $h(t) := h(t; 0) = L[\delta(t)]$ と置くと, 時不変性より
$$h(t; \tau) = L[\delta(t - \tau)] = h(t - \tau)$$
- よって一般に $L[f(t)] = \int_{-\infty}^{\infty} f(\tau)h(t - \tau)d\tau$

参考資料: 信号処理の基礎

■ 畳み込み定理

- $f_1(t) * f_2(t)$ のFourier変換(またはZ変換)は $f_1(t)$ と $f_2(t)$ のFourier変換(またはZ変換)の積に等しい

Fourier変換対

$$\begin{array}{l} f_1(t) \longleftrightarrow F_1(\omega) \\ f_2(t) \longleftrightarrow F_2(\omega) \end{array} \quad \Rightarrow \quad f_1(t) * f_2(t) \longleftrightarrow F_1(\omega)F_2(\omega)$$

■ 略証

$$\begin{aligned} F(\omega) &= \int_{-\infty}^{\infty} e^{-j\omega t} \left(\int_{-\infty}^{\infty} f_1(\tau) f_2(t - \tau) d\tau \right) dt \\ &= \int_{-\infty}^{\infty} f_1(\tau) \left(\int_{-\infty}^{\infty} e^{-j\omega t} f_2(t - \tau) dt \right) d\tau \\ &= \int_{-\infty}^{\infty} f_1(\tau) e^{-j\omega \tau} \left(\int_{-\infty}^{\infty} e^{-j\omega(t-\tau)} f_2(t - \tau) dt \right) d\tau \\ &= F_1(\omega) F_2(\omega) \end{aligned}$$

音声・音楽・音響信号に関する諸問題

■ 音声・音楽・音響信号の認識・分析・合成・変換・分離

	認識	分析	合成	変換	分離
音声					
音楽			—	—	
音響					

⇒「識別」、「回帰」、「逆問題」のいずれかに相当

音声・音楽・音響信号に関する諸問題

■ 音声・音楽・音響信号の認識・分析・合成・変換・分離

	認識	分析	合成	変換	分離
音声					
音楽			—	—	
音響					

⇒「識別」、「回帰」、「逆問題」のいずれかに相当

線形予測分析 (LPC) による音声分析 [Itakura+1968]

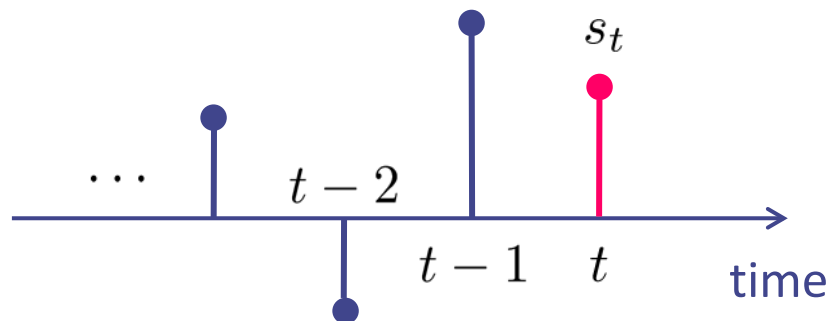
■ 動機: 符号化への応用

- 少ないパラメータで音声信号を表現したい

■ 問題: 線形予測誤差の最小化

- 時刻 t の信号のサンプル値 s_t を、過去のサンプル値 $s_{t-1}, s_{t-2}, \dots, s_{t-P}$ の線形結合で「予測」
- 「予測」の誤差を最小にするには
結合係数(予測係数という)をどう置けば良い？

$$s_t \simeq \sum_{p=1}^P \underline{a_p} s_{t-p}$$



最小二乗誤差推定問題としての定式化

■すべての t で $s_t \simeq \sum_{p=1}^P a_p s_{t-p}$ となる $a_1, \dots, a_P$ を求めたい

■目的関数

$$\mathcal{J}(a_1, \dots, a_P) = \sum_t \left(s_t - \sum_{p=1}^P a_p s_{t-p} \right)^2 \rightarrow \text{minimize}$$

■最小解では $\frac{\partial \mathcal{J}}{\partial a_1} = \frac{\partial \mathcal{J}}{\partial a_2} = \dots = \frac{\partial \mathcal{J}}{\partial a_P} = 0$ を満たすため...

$$\begin{aligned} \longrightarrow \frac{\partial \mathcal{J}}{\partial a_q} &= 2 \sum_t \left(s_t - \sum_{p=1}^P a_p s_{t-p} \right) (-s_{t-q}) \\ &= 2 \left(\sum_p a_p \sum_t s_{t-p} s_{t-q} - \sum_t s_t s_{t-q} \right) \\ &= 0 \quad (q = 1, \dots, P) \end{aligned}$$

最小二乗誤差推定問題としての定式化

■ 連立方程式に帰着

$$\left\{ \begin{array}{l} \sum_p a_p \sum_t s_{t-p} s_{t-1} = \sum_t s_t s_{t-1} \\ \sum_p a_p \sum_t s_{t-p} s_{t-2} = \sum_t s_t s_{t-2} \\ \vdots \\ \sum_p a_p \sum_t s_{t-p} s_{t-P} = \sum_t s_t s_{t-P} \end{array} \right. \longrightarrow c_{q,p} = \sum_t s_{t-p} s_{t-q}$$

以上より最適予測係数は
以下の方程式を満たす

$$\begin{bmatrix} c_{1,1} & \cdots & c_{1,P} \\ \vdots & \ddots & \vdots \\ c_{P,1} & \cdots & c_{P,P} \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_P \end{bmatrix} = \begin{bmatrix} c_{0,1} \\ \vdots \\ c_{0,P} \end{bmatrix}$$

この方程式をYule-Walker方程式という

音響符号化への応用

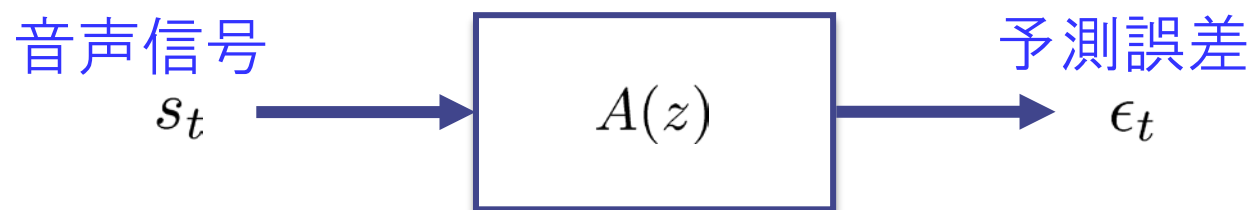
■時系列信号の可逆圧縮符号化



線形システムとしての解釈

- 音声信号から予測誤差を出力する線形システム

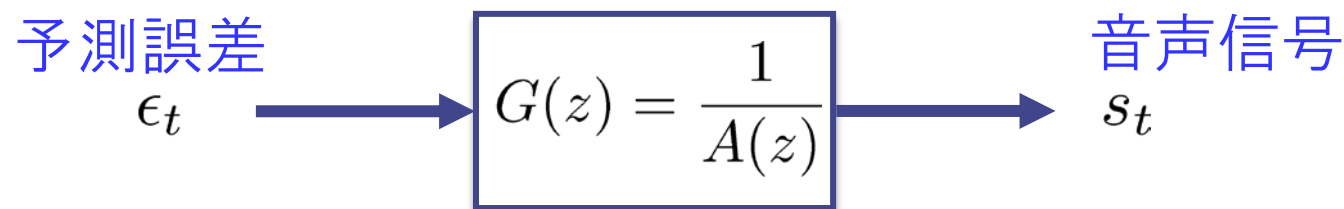
$$\epsilon_t = s_t - \sum_{p=1}^P a_p s_{t-p} = s_t * a_t \quad (a_0 = 1)$$



$$A(z) = 1 - a_1 z^{-1} - a_2 z^{-2} - \dots - a_P z^{-P}$$

移動平均システム(全零モデル)

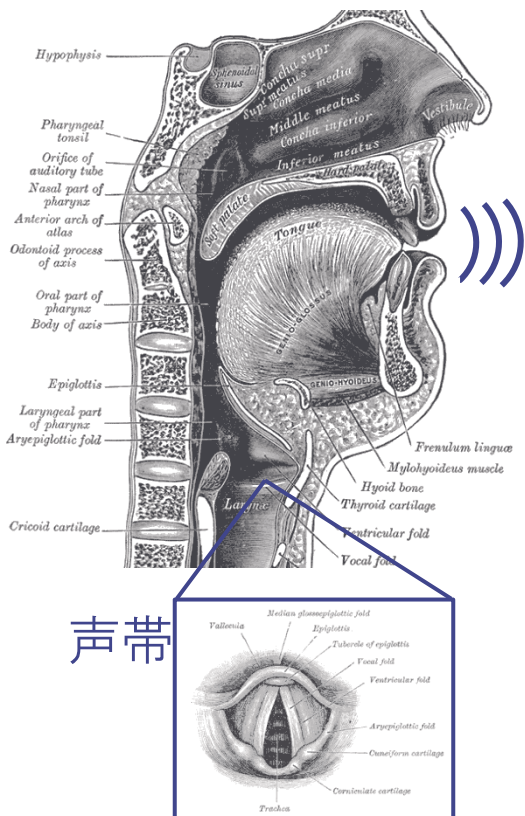
- 予測誤差を入力として音声信号を出力する線形システムは？



自己回帰システム(全極モデル)

ソース・フィルタ理論 [Fant1960]

■ソース・フィルタモデル: 線形な音声生成過程モデル



声帯振動

ϵ_t

$G(z)$

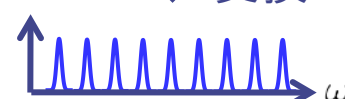
音声波形

s_t

声帯振動



Fourier
変換

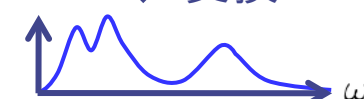


声帯スペクトル

声道

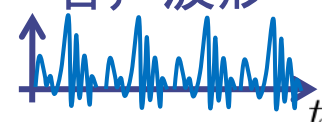


Fourier
変換

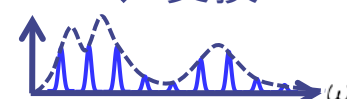


声道スペクトル

音声波形



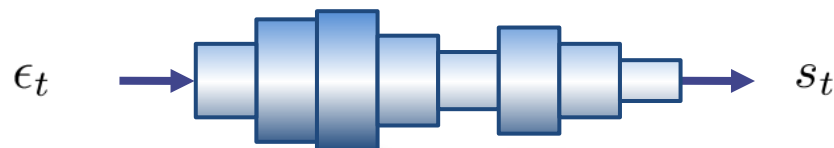
Fourier
変換



音声スペクトル

自己回帰システムの仮定

無損失等長音響管による声道モデルに相当



統計モデルによる音声生成過程の表現

■ 声帯振動に関する仮定

- Gauss性 $\dots$ $\epsilon = (\epsilon_1, \dots, \epsilon_T)^T \sim \mathcal{N}(\epsilon | \mu, \Sigma)$
 - 定常性 $\dots$ $\mu = \mathbf{0}, \quad \Sigma : \text{Toeplitz行列}$
 - 白色性 $\dots$ $\Sigma = \sigma^2 I$
- $\left. \begin{array}{l} \epsilon_t \sim \mathcal{N}(\epsilon_t | 0, \sigma^2) \\ (t = 1, \dots, T) \end{array} \right\}$

■ 声道特性に関する仮定

- 自己回帰システム(全極モデル)

$$\underbrace{\begin{bmatrix} 1 & & & & 0 \\ -a_1 & 1 & & & \\ \vdots & \ddots & \ddots & & \\ -a_P & & \ddots & \ddots & \\ 0 & & \ddots & \ddots & \ddots \\ & & & -a_P & \cdots & -a_1 & 1 \end{bmatrix}}_{\Psi} \underbrace{\begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_T \end{bmatrix}}_s = \underbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_T \end{bmatrix}}_{\epsilon}$$

$$\epsilon_t = s_t - \sum_{p=1}^P a_p s_{t-p}$$

最尤推定問題としての定式化

■ 以上の仮定をまとめると・・・

- $\epsilon \sim \mathcal{N}(\epsilon|\mathbf{0}, \sigma^2 \mathbf{I})$

- $\Psi \mathbf{s} = \epsilon$

未知パラメータは $\{a_p\}_{1 \leq p \leq P}, \sigma^2 \longrightarrow \theta$

観測されるのは $\{s_t\}_{1 \leq t \leq T} \longrightarrow \mathbf{s}$

■ 観測信号 $\mathbf{s}$ の確率密度関数(尤度関数)

- $\mathbf{s} = \Psi^{-1} \epsilon \sim \mathcal{N}(\mathbf{s}|\mathbf{0}, \sigma^2 \Psi^{-1} (\Psi^{-1})^\top) \longrightarrow p(\mathbf{s}|\theta)$

- 対数尤度 $\log p(\mathbf{s}|\theta)$ は

$$-\frac{T}{2} \log(2\pi) - \frac{1}{2} \log |\sigma^2 \Psi^{-1} (\Psi^{-1})^\top| - \frac{1}{2\sigma^2} \mathbf{s}^\top \Psi^\top \Psi \mathbf{s}$$

◆ logdet項: $\frac{\log |\sigma^2 \Psi^{-1} (\Psi^{-1})^\top|}{2}$

$$= \log |\sigma^2 \mathbf{I}| - 2 \log |\Psi|$$

$$= T \log \sigma^2 \quad \because |\Psi| = 1$$

$$\begin{aligned} & \|\epsilon\|_2^2 = \\ & \sum_t \left(s_t - \sum_{p=1}^P a_p s_{t-p} \right)^2 \end{aligned}$$

前述の目的関数を得る

周波数領域における対数尤度

■ 音声スペクトル(周波数領域)の確率密度関数

$$x = Fs \quad (F \text{ は離散Fourier変換行列})$$

$$s \xrightarrow{\text{変数変換}} x \sim \mathcal{N}_{\mathbb{C}}(0, \underbrace{\sigma^2 F \Psi^{-1} (\Psi^{-1})^T F^H}_{\text{diag}(\lambda_1, \dots, \lambda_T)})$$

全極型スペクトルモデル

$$\begin{cases} \lambda_k = \frac{\sigma^2}{|A(e^{2\pi j(k-1)/T})|^2} \\ A(z) = 1 - a_1 z^{-1} - \dots - a_P z^{-P} \end{cases}$$



■ 対数尤度

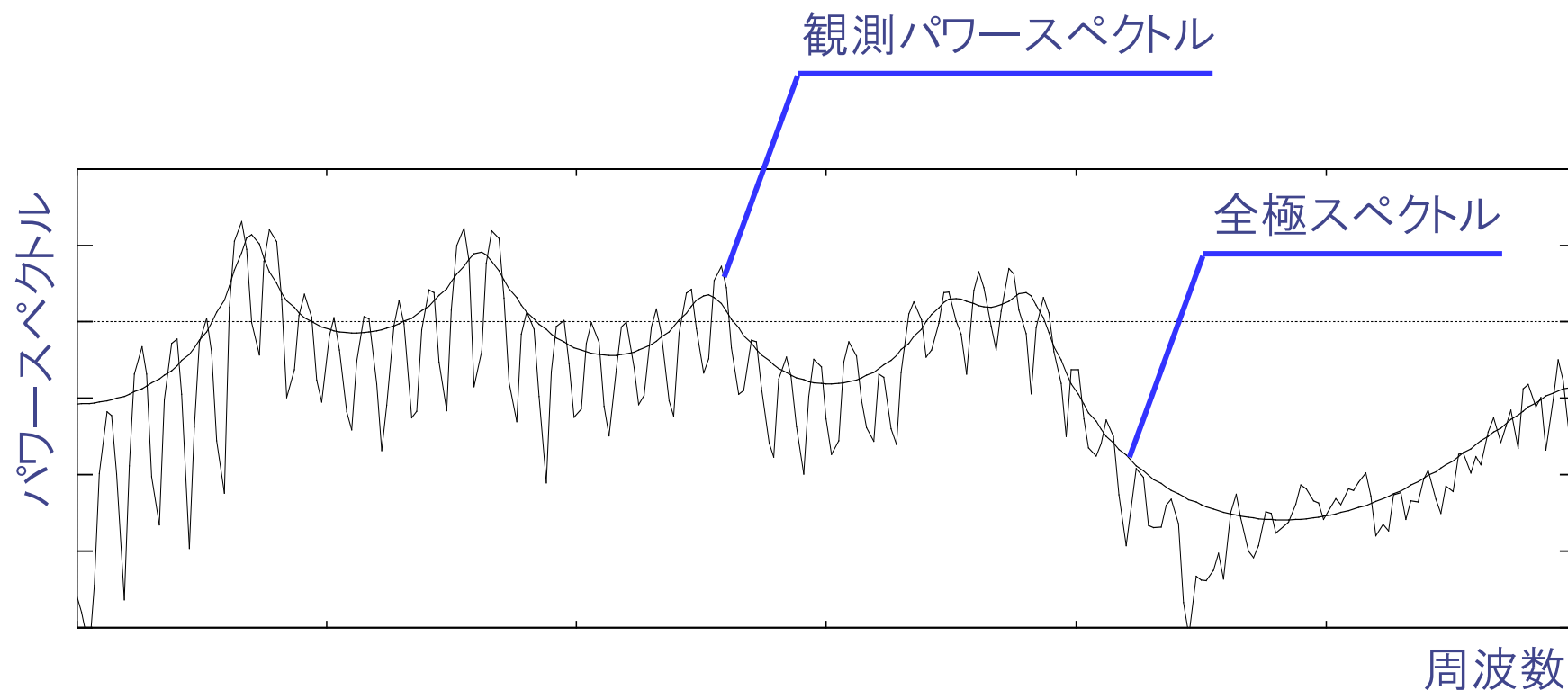
観測パワースペクトル

$$\begin{aligned} \mathcal{L}(a_1, \dots, a_P) &= - \sum_k \left(\log \pi \lambda_k + \frac{|x_k|^2}{\lambda_k} \right) \\ &\stackrel{c}{=} - \sum_k \left(\frac{|x_k|^2}{\lambda_k} - \log \frac{|x_k|^2}{\lambda_k} - 1 \right) \end{aligned}$$

$|x_k|^2$ と λ_k の板倉齋藤距離

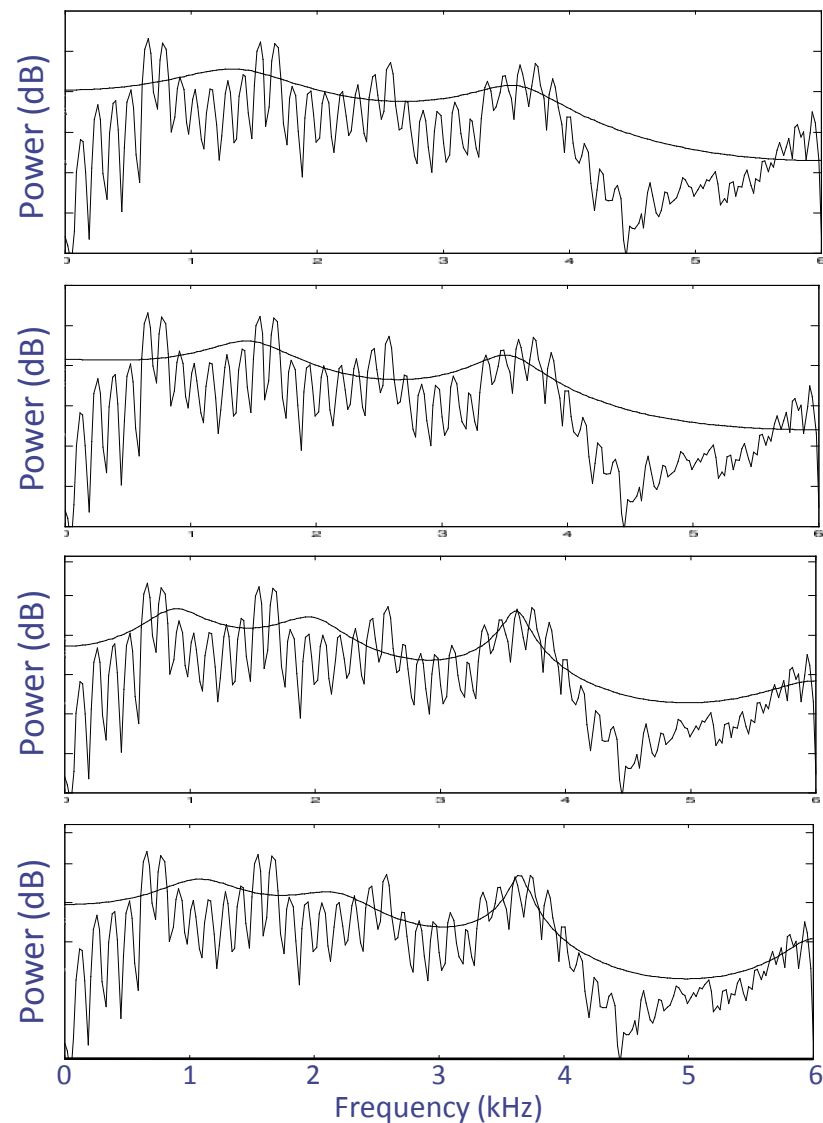
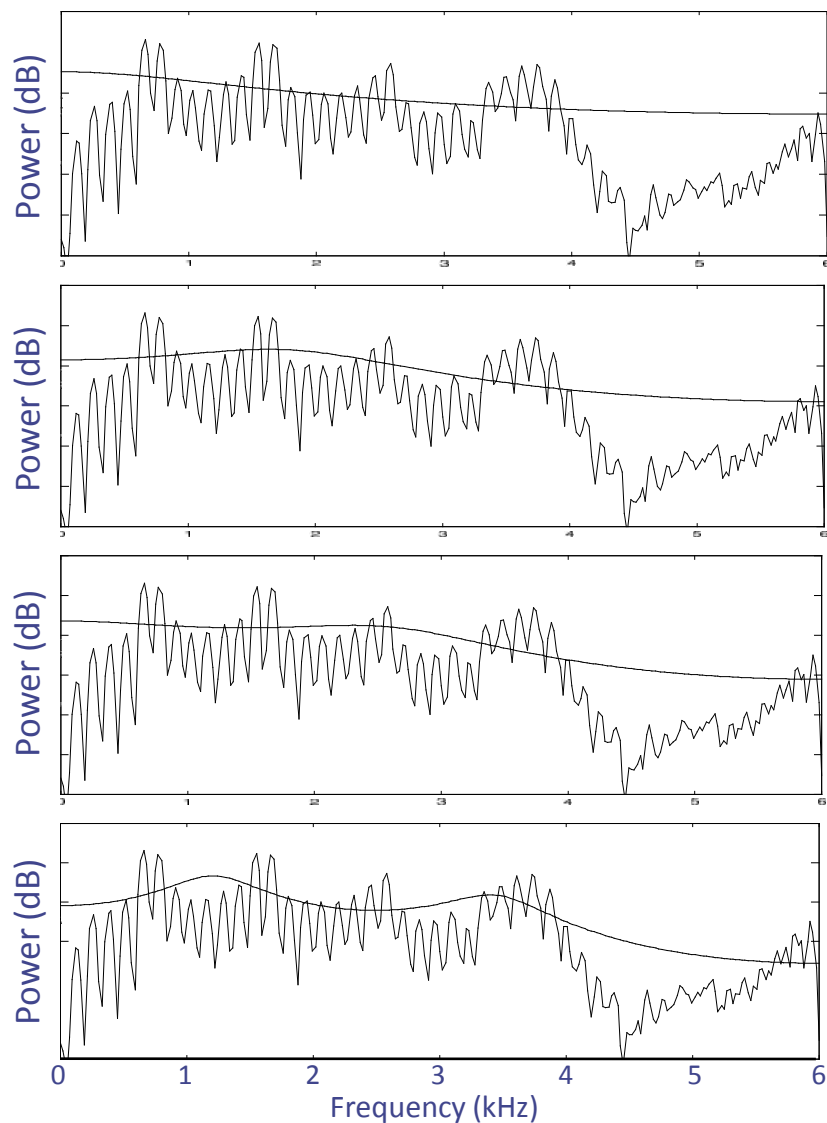
LPCは「スペクトル包絡」推定に相当

- 観測パワースペクトルと全極スペクトルとの板倉斎藤距離最小化



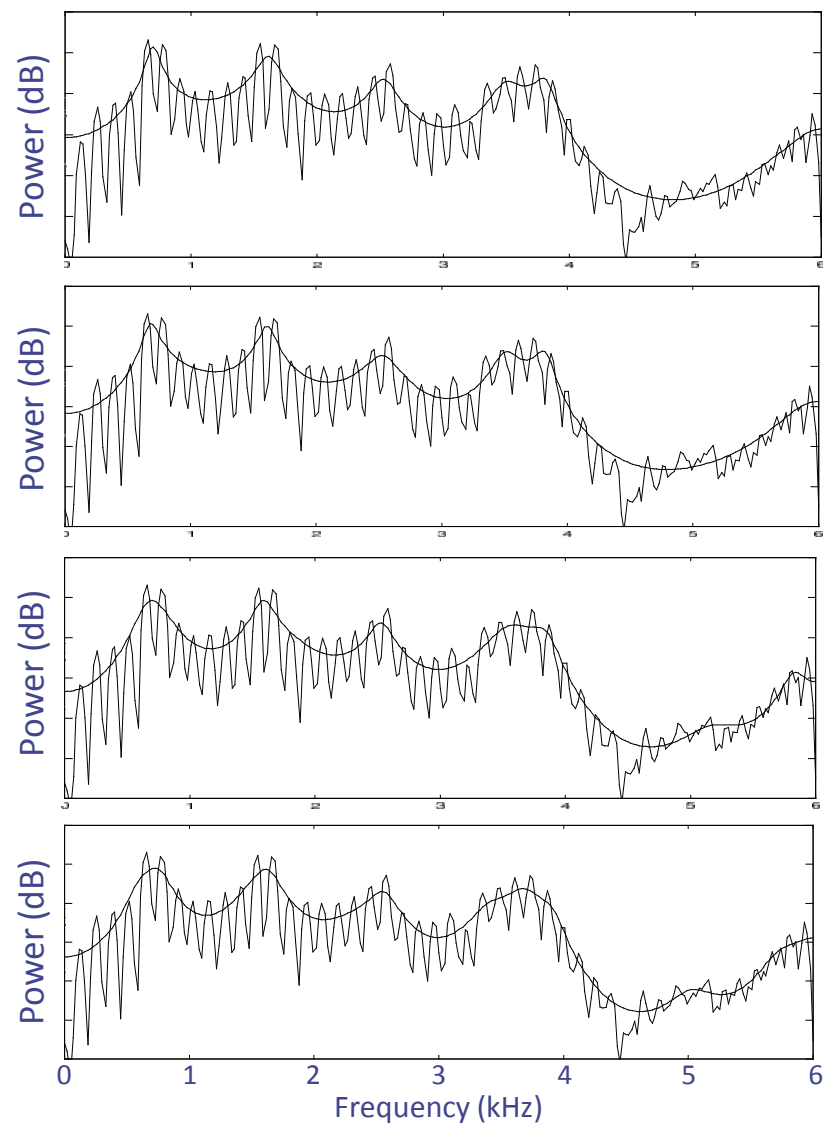
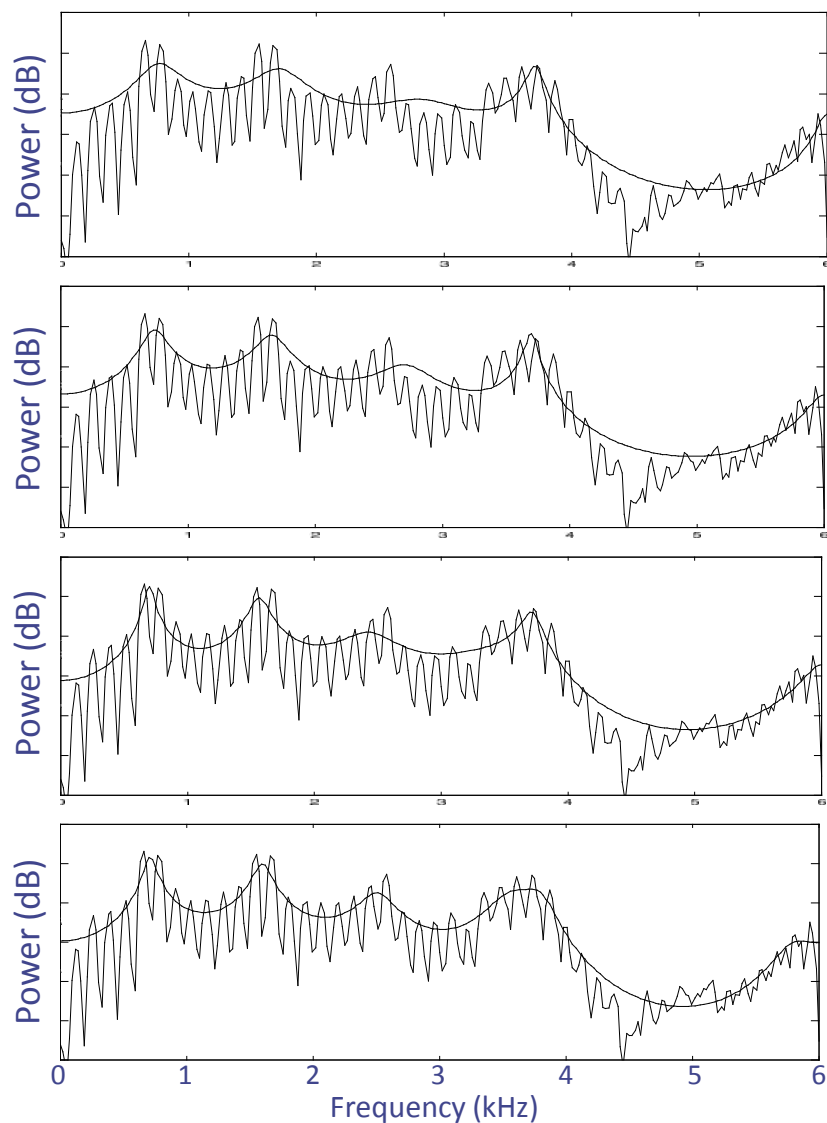
LPCによる音声スペクトル推定の例

■ 1~8次



LPCによる音声スペクトル推定の例

■ 9, 10, 12, 14, 16, 18, 22, 26次



音声・音楽・音響信号に関する諸問題

■ 音声・音楽・音響信号の認識・分析・合成・変換・分離

	認識	分析	合成	変換	分離
音声					
音楽			—	—	
音響					

⇒「識別」、「回帰」、「逆問題」のいずれかに相当

音声・音楽・音響信号に関する諸問題

■ 音声・音楽・音響信号の認識・分析・合成・変換・分離

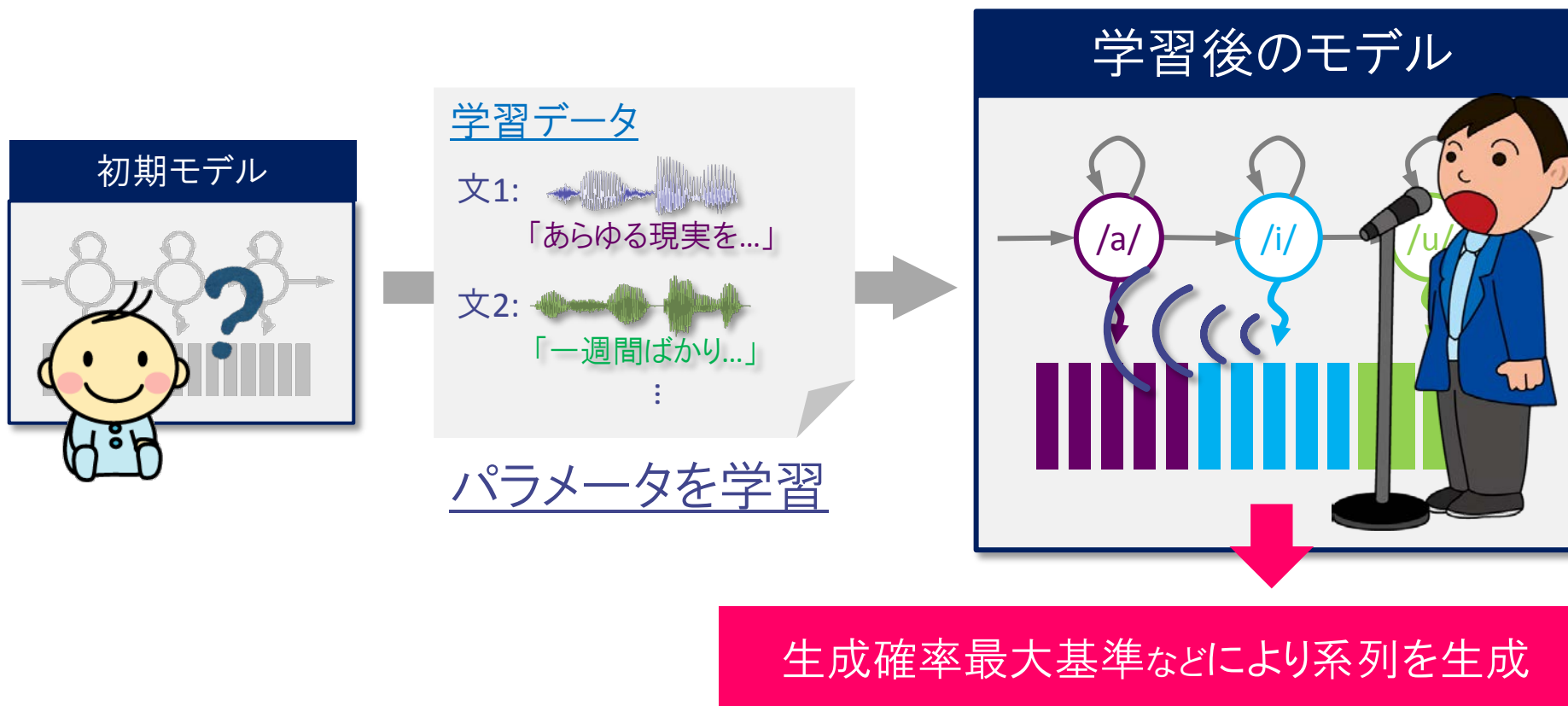
	認識	分析	合成	変換	分離
音声					
音楽			—	—	
音響					

⇒「識別」、「回帰」、「逆問題」のいずれかに相当

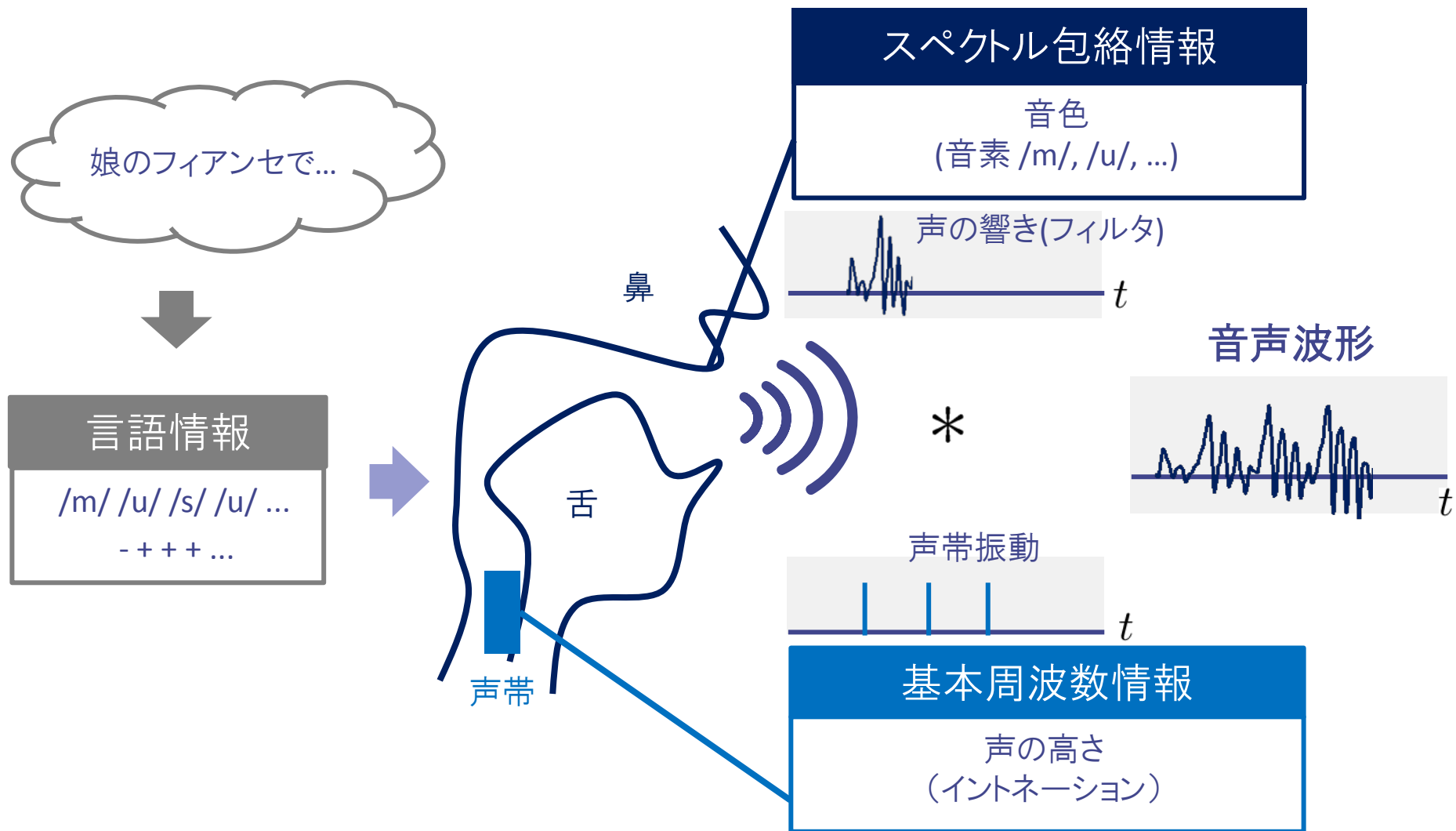
パラメトリック方式のテキスト音声合成

■ 確率的な音声ジェネレータのパラメータ学習

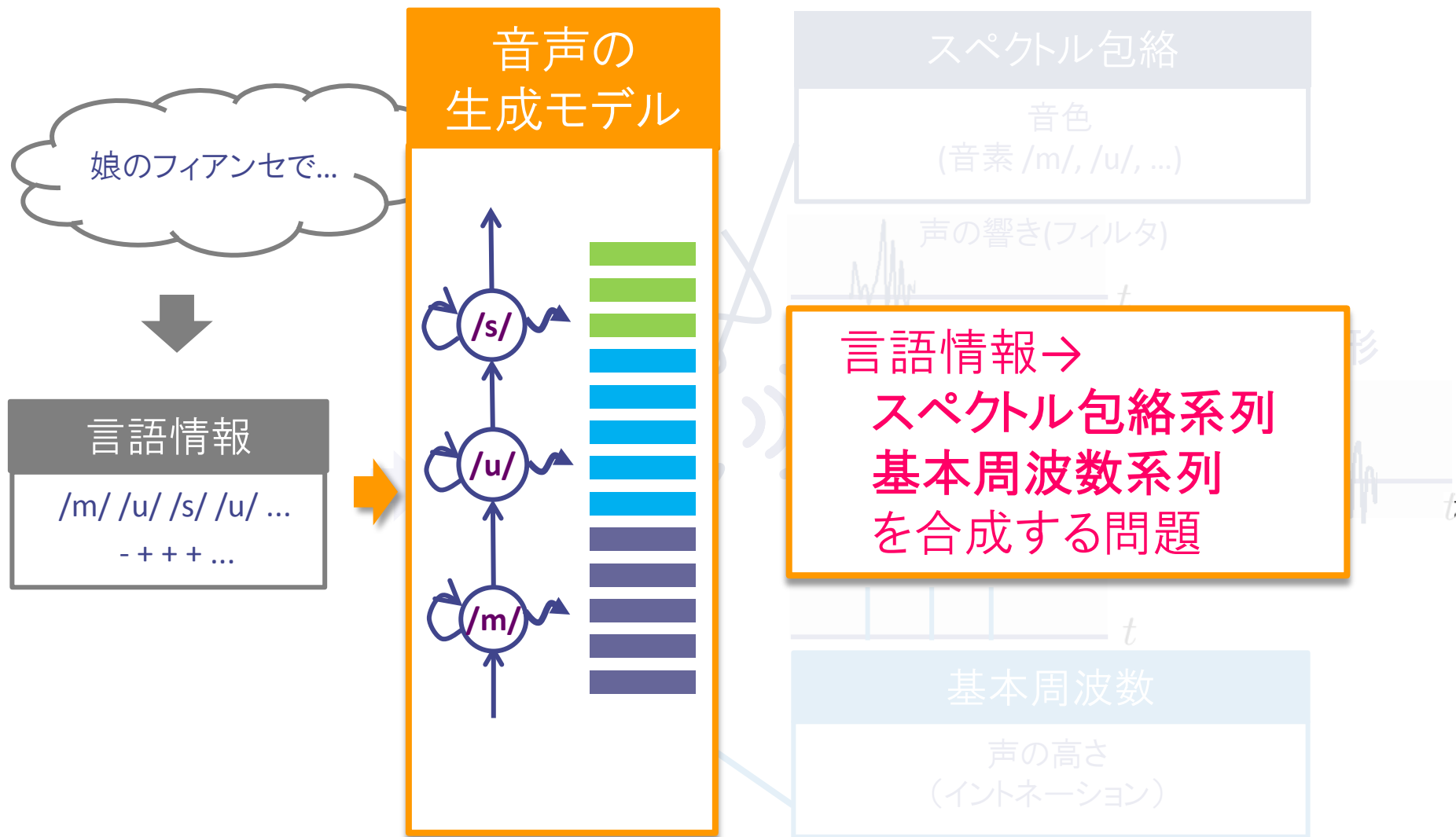
音声の性質を上手く反映したモデルを立てることが重要



音声の生成過程

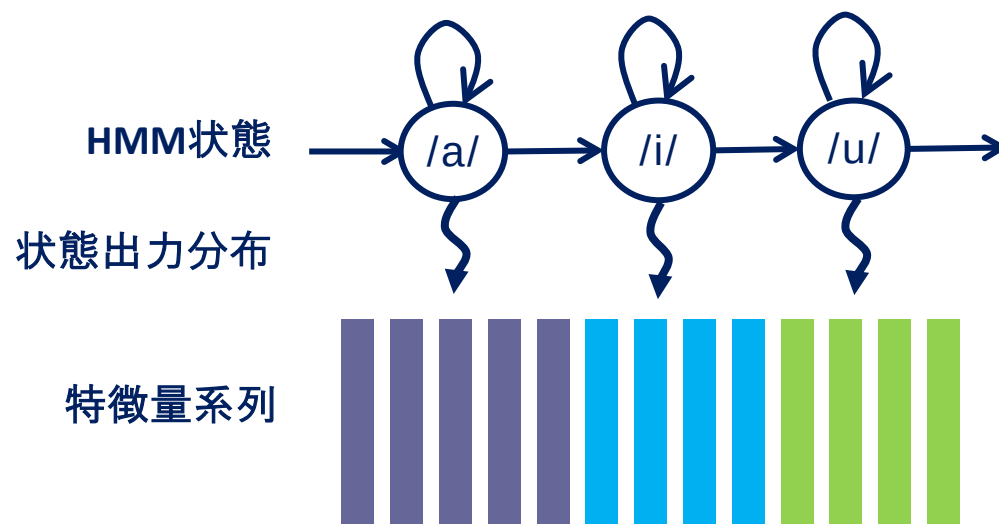


音声の生成過程



HMM音声合成 [Yoshimura+1999]

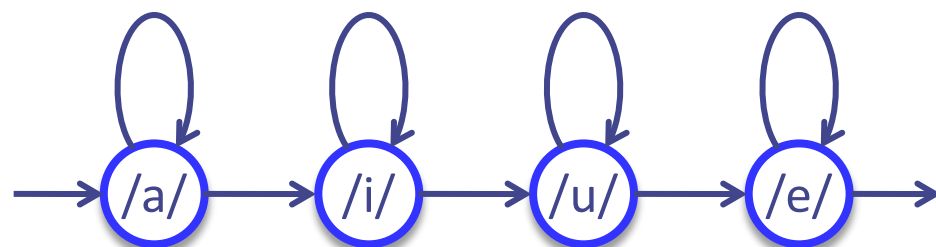
- 隠れマルコフモデル(HMM)による音声特徴量(スペクトル包絡特徴量 & 基本周波数)系列のモデル化



- 言語情報を隠れ状態で表現
- 特徴量の確率的な揺らぎを状態出力分布で表現
- 状態継続長の確率的な揺らぎを状態遷移確率で表現

デルタ特徴量

- HMMに基づく音声認識・合成ではデルタ(時間微分)特徴量が有効
 - 静的特徴量とその時間微分を組にしたベクトルを観測量と扱う場合が多い



状態系列 $\mathbf{q} = (q_1, \dots, q_T)$

状態出力

$$\mathbf{o}_t = \begin{bmatrix} c_t \\ \Delta c_t \\ \Delta^2 c_t \end{bmatrix} \sim \mathcal{N}(\mathbf{o}_t; \boldsymbol{\mu}_{q_t}, \mathbf{U}_{q_t})$$

静的特徴量 (points to c_t)

デルタ特徴量
(c_t の時間差分) (points to Δc_t and $\Delta^2 c_t$)

- 従来方式では静的特徴量とデルタ特徴量を別変数として扱っていたが、実際は両者は線形の関係で結ばれている

$$(\mathbf{o}_1^T, \dots, \mathbf{o}_T^T)^T = \mathbf{W}(c_1^T, \dots, c_T^T)^T \quad (\text{後述})$$

- 静的特徴量とデルタ特徴量の関係を陽に組み込んだHMMの拡張モデル $\Rightarrow$ トラジェクトリHMM [Zen+2007]

トラジェクトリHMM [Zen+2007]

■ 通常のHMM

$$p(\mathbf{o}|\lambda) = \sum_{\mathbf{q}} \underbrace{p(\mathbf{o}|\mathbf{q}, \lambda)}_{\substack{\text{状態出力分布} \\ \text{(ガウス分布の場合)}}} p(\mathbf{q}|\lambda) = \prod_{t=1}^T \mathcal{N}(\mathbf{o}_t | \boldsymbol{\mu}_{q_t}, \mathbf{U}_{q_t})$$

■ $\mathbf{o}$: 観測系列, $\mathbf{q}$: 状態系列, λ : HMMパラメータ

■ トラジェクトリHMM

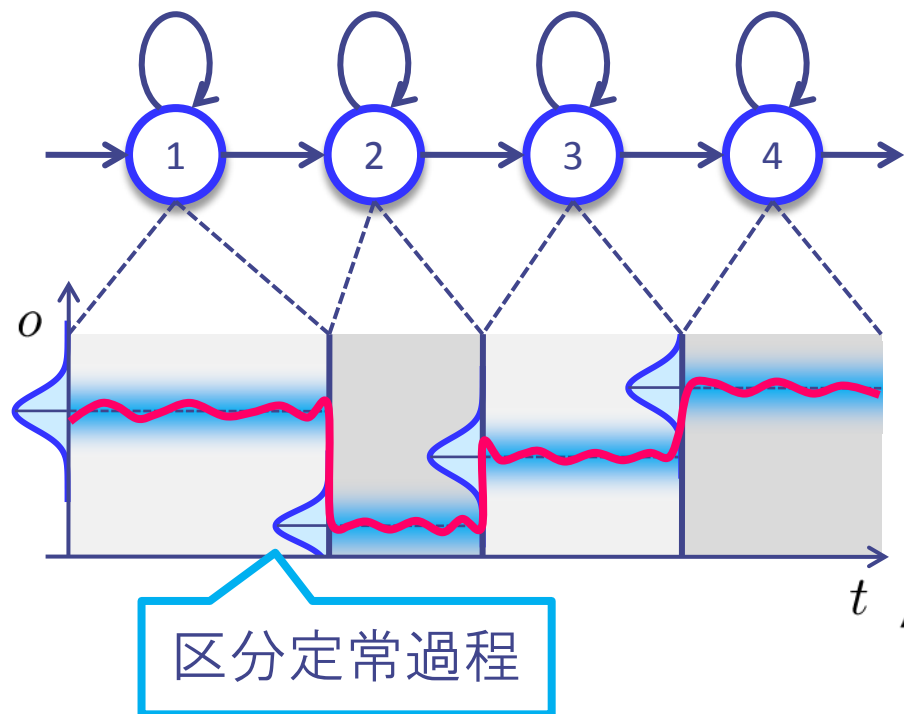
■ $\mathbf{o} = \mathbf{W}\mathbf{c}$ の関係を用いて $\mathbf{c}$ の生成確率に書き換え

$$\begin{aligned} p(\mathbf{c}|\mathbf{q}, \lambda) &\propto \mathcal{N}(\mathbf{W}\mathbf{c} | \boldsymbol{\mu}_q, \mathbf{U}_q) \\ &\propto \exp\left(-\frac{1}{2}(\mathbf{W}\mathbf{c} - \boldsymbol{\mu}_q)^\top \mathbf{U}_q^{-1}(\mathbf{W}\mathbf{c} - \boldsymbol{\mu}_q)\right) \\ &\propto \exp\left(-\frac{1}{2}(\mathbf{c}^\top \mathbf{W}^\top \mathbf{U}_q^{-1} \mathbf{W} \mathbf{c} - 2\mathbf{c}^\top \mathbf{W}^\top \mathbf{U}_q^{-1} \boldsymbol{\mu}_q)\right) \\ &= \mathcal{N}(\mathbf{c} | \bar{\mathbf{c}}_q, \mathbf{P}_q) \\ &\quad \begin{cases} \bar{\mathbf{c}}_q = (\mathbf{W}^\top \mathbf{U}_q^{-1} \mathbf{W})^{-1} \mathbf{W}^\top \mathbf{U}_q^{-1} \boldsymbol{\mu}_q, \\ \mathbf{P}_q = (\mathbf{W}^\top \mathbf{U}_q^{-1} \mathbf{W})^{-1} \end{cases} \end{aligned}$$

$$\begin{bmatrix} \mathbf{o} \\ \vdots \\ \mathbf{o} \end{bmatrix} = \begin{bmatrix} \mathbf{c}_1 \\ \Delta \mathbf{c}_1 \\ \Delta^2 \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_T \\ \Delta \mathbf{c}_T \\ \Delta^2 \mathbf{c}_T \end{bmatrix} \begin{bmatrix} \mathbf{W} \\ \vdots \\ \mathbf{W} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ -1/2 & 0 & 1/2 \\ -1 & 2 & -1 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \vdots \\ \mathbf{c} \end{bmatrix}$$

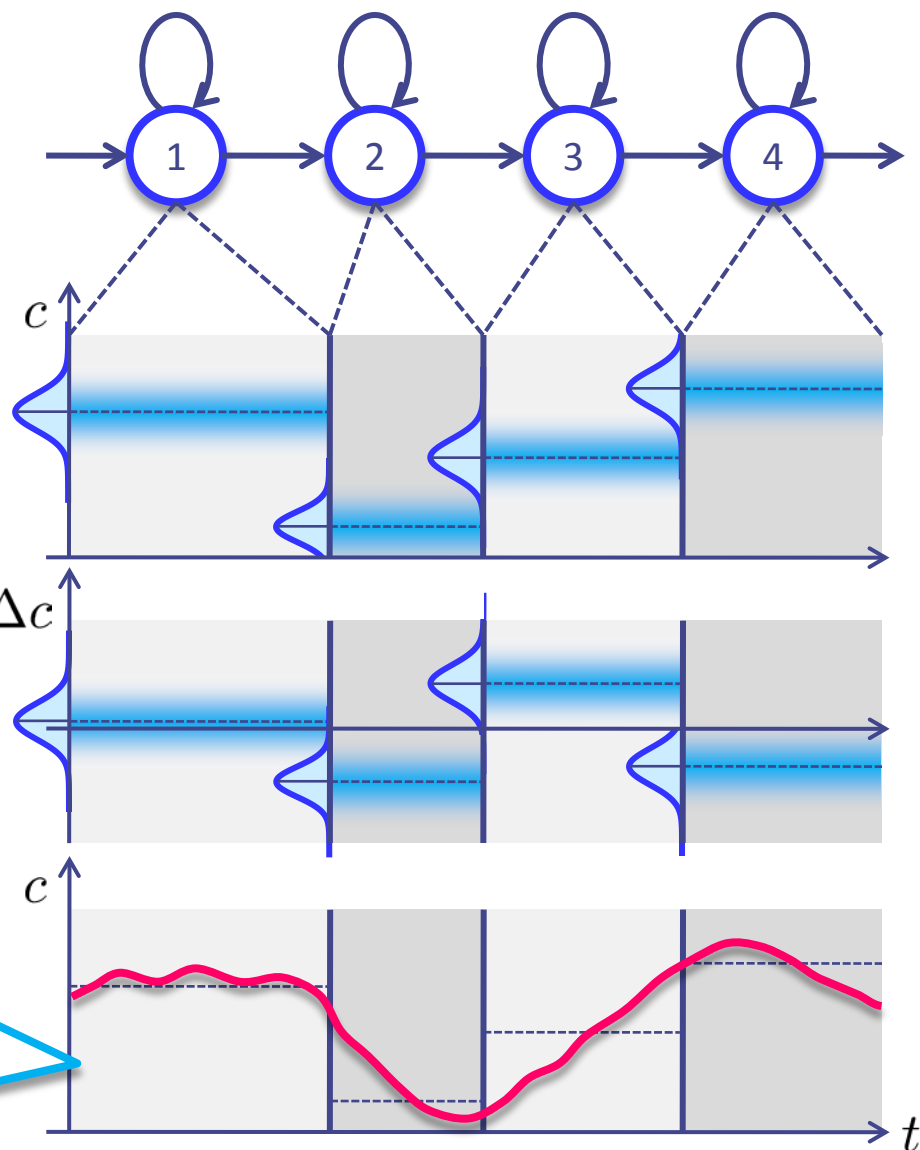
HMMとトラジェクトリHMMの比較

■ HMM



状態遷移に応じて
連続変化する時系列を
表現する確率モデル

■ トラジェクトリHMM



トランジェクトHMMのパラメータ学習 [Zen+2007]

■ 特徴量系列の生成確率

$$p(\mathbf{c}|\lambda) = \sum_{\mathbf{q}} p(\mathbf{c}|\mathbf{q}, \lambda) p(\mathbf{q}|\lambda)$$

■ パラメータ λ の学習アルゴリズム

- $\log p(\mathbf{c}, \mathbf{q}|\lambda) \rightarrow \text{maximize}$

出力分布パラメータ、遷移確率

(Step 1) $\lambda \leftarrow \operatorname{argmax}_{\lambda} \log p(\mathbf{c}, \mathbf{q}|\lambda)$... 解析的に解けないため
最急降下更新

状態系列

(Step 2) $\mathbf{q} \leftarrow \operatorname{argmax}_{\mathbf{q}} \log p(\mathbf{c}, \mathbf{q}|\lambda)$... 動的計画法で解けないため
近似計算により探索

HMM音声合成による合成音声例

		Mel-cepstrum (スペクトル包絡特徴量)		
		static	static + Δ	static + $\Delta + \Delta^2$
log F0 (対数基本 周波数)	static			
	static + Δ			
	static + $\Delta + \Delta^2$			

音声・音楽・音響信号に関する諸問題

■ 音声・音楽・音響信号の認識・分析・合成・変換・分離

	認識	分析	合成	変換	分離
音声					
音楽			—	—	
音響				—	

⇒「識別」、「回帰」、「逆問題」のいずれかに相当

音声・音楽・音響信号に関する諸問題

■ 音声・音楽・音響信号の認識・分析・合成・変換・分離

	認識	分析	合成	変換	分離
音声					
音楽			—	—	
音響				—	

⇒「識別」、「回帰」、「逆問題」のいずれかに相当

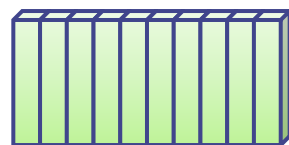
統計的手法による声質変換 [Abe+1990, Stylianou+1998]

同一発話内容の
音声のペア → 音声特徴量
のペア → 時間の
対応付け → 「変換関数」
の統計学習

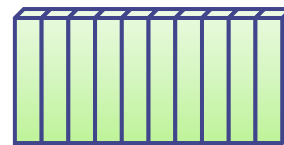
音声 Y



特徴量系列 Y



y_t



音声 X



特徴量系列 X

x_t



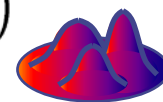
訓練データ

$$\{x_t, y_t\}_{t \in \mathcal{D}}$$



$$p(x, y | \theta)$$

の学習



$$p(x | y, \theta)$$

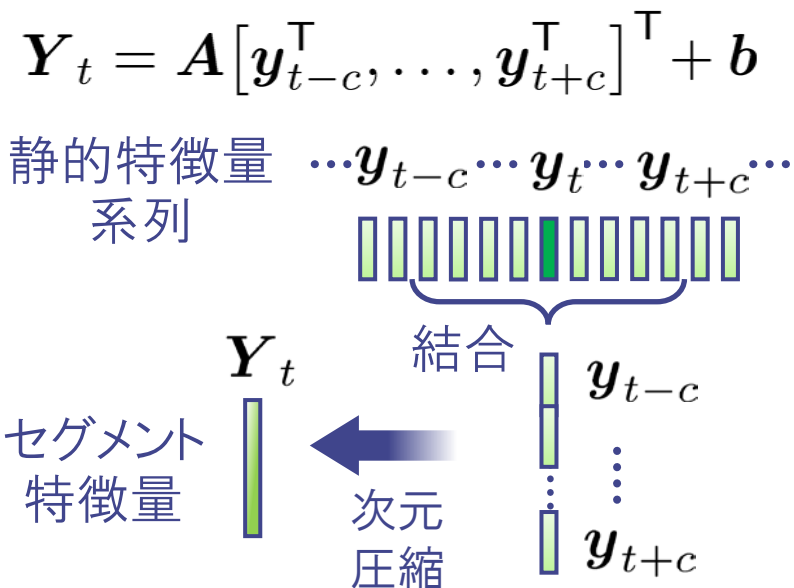
の獲得

※ $p(x | y, \theta)$ を直接学習する
方法も提案されている

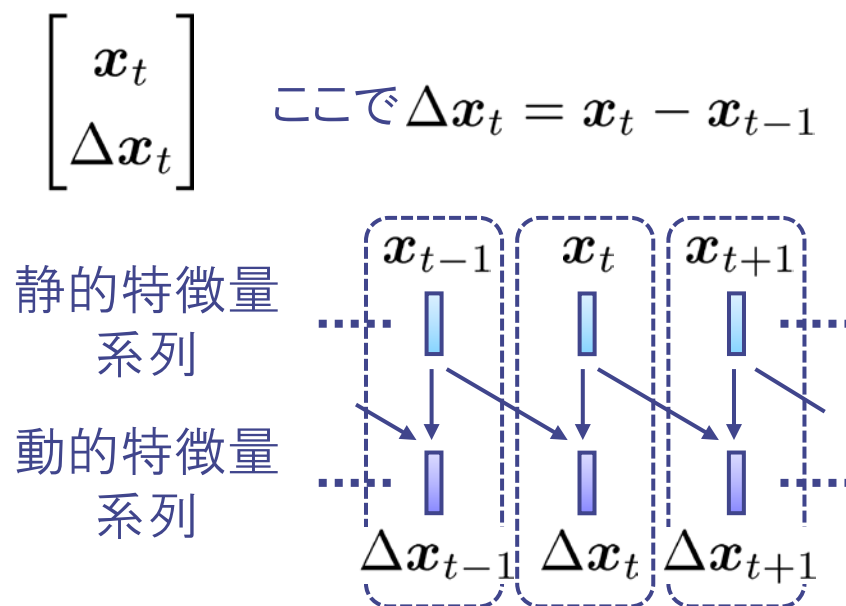
最尤系列変換法 [Toda+2007]

■ 音声特徴量の時系列性を考慮した拡張版

入力: セグメント特徴量

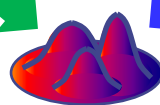


出力: 静的・動的特徴量



学習

混合正規分布
モデル(GMM)



$$\prod_t p(\mathbf{Y}_t, \mathbf{x}_t, \Delta \mathbf{x}_t | \theta)$$

最尤系列変換法 [Toda+2007]

■時系列全体の最尤変換

入力セグメント
特徴量系列

$\mathbf{Y}_1 \mathbf{Y}_2 \dots \mathbf{Y}_t \dots \mathbf{Y}_T$

GMM



$$\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_T = \underset{\mathbf{x}_1, \dots, \mathbf{x}_T}{\operatorname{argmax}} \prod_{t=1}^T p(\mathbf{x}_t | \mathbf{Y}_t, \boldsymbol{\theta}) p(\Delta \mathbf{x}_t | \mathbf{Y}_t, \boldsymbol{\theta})$$

$$\propto p(\mathbf{x}_1, \dots, \mathbf{x}_T | \mathbf{Y}_1, \dots, \mathbf{Y}_T, \boldsymbol{\theta})$$

$\Delta \mathbf{x}_t = \mathbf{x}_t - \mathbf{x}_{t-1}$ という関係を用いて
 $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_T)^\top$ のPDFに書き換え

変換静的
特徴量系列

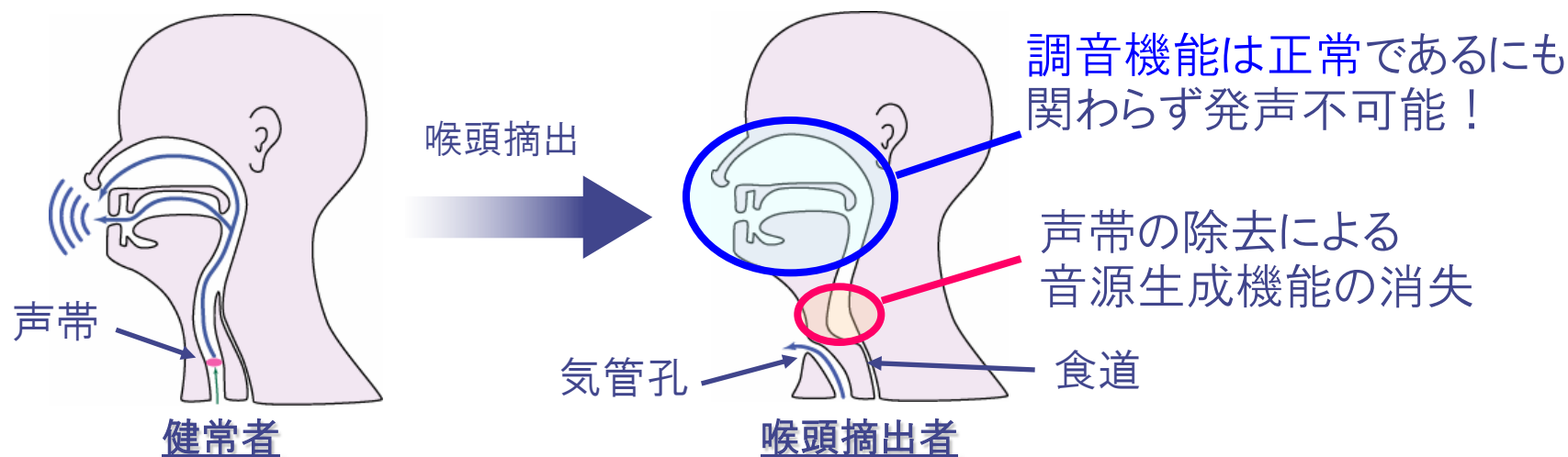
$\hat{\mathbf{x}}_1 \hat{\mathbf{x}}_2 \dots \hat{\mathbf{x}}_t \dots \hat{\mathbf{x}}_T$

音声変換の応用例2: 発声障害者補助 [Toda+2007]

■ 発声障害の一例: 喉頭摘出による発声障害

■ 喉頭摘出者

- 喉頭癌等の理由により喉頭を摘出した人
- 全国に約2万人
- 気管と食道が完全に分離

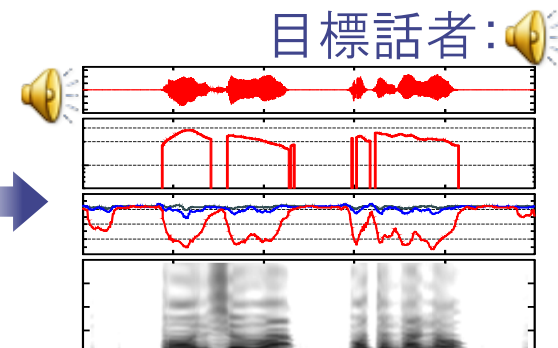
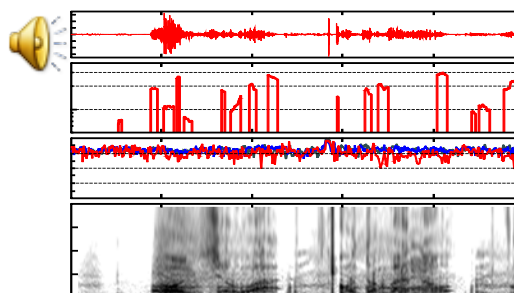
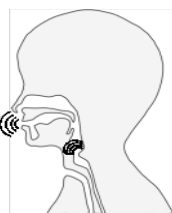


発声能力が失われることによる生活の質 (Quality of Life: QOL) の低下は極めて深刻な問題!

無喉頭音声から通常音声への変換

食道音声の変換

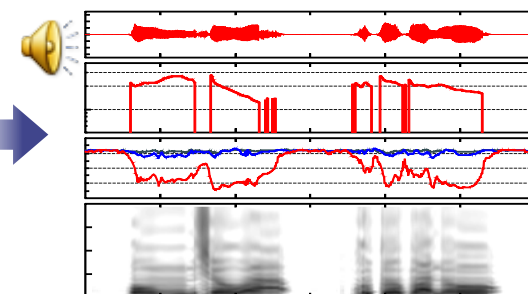
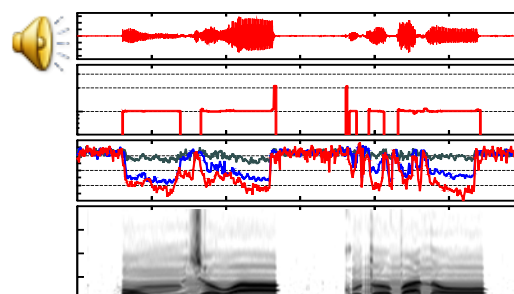
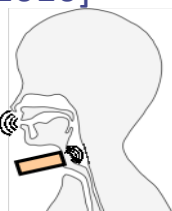
[Doi+2009-2010]



電気音声の変換

[Nakamura+2009-2010]

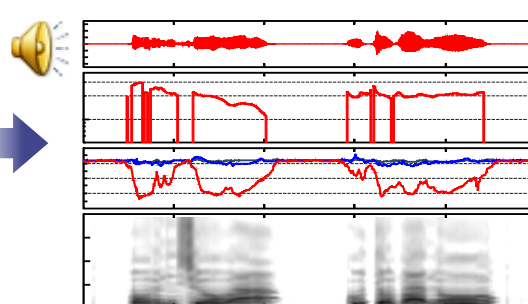
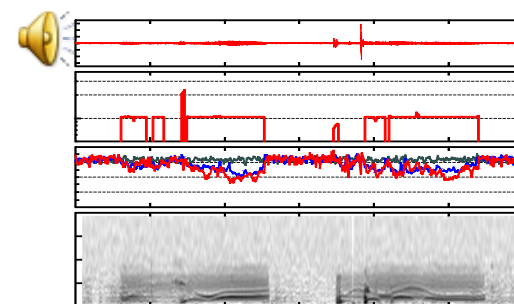
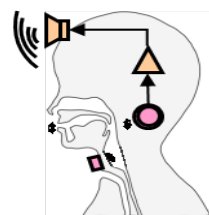
[Doi+2010]



肉伝導微弱電気音声の変換

[Nakamura+2007-2010]

[Doi+2010]



身体的制約を超えた発声(失われた声の回復)を可能とする技術

音声・音楽・音響信号に関する諸問題

■ 音声・音楽・音響信号の認識・分析・合成・変換・分離

	認識	分析	合成	変換	分離
音声					
音楽			—	—	
音響				—	

⇒「識別」、「回帰」、「逆問題」のいずれかに相当

音声・音楽・音響信号に関する諸問題

■ 音声・音楽・音響信号の認識・分析・合成・変換・分離

	認識	分析	合成	変換	分離
音声					
音楽			—	—	
音響				—	

⇒「識別」、「回帰」、「逆問題」のいずれかに相当

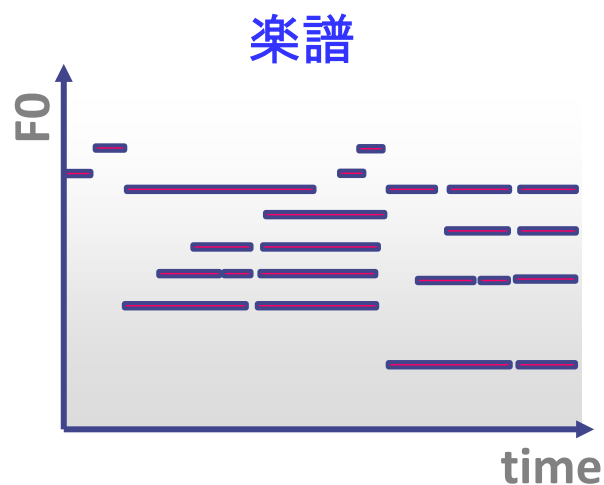
多重音解析

■ モノラル多重音信号を個々の音源成分に分解する問題

多重音信号 $y[t]$

→ $Y_{\omega,m} = \sum_{t=0}^{T-1} W[t]y[t+m]e^{j2\pi\omega t/T}$

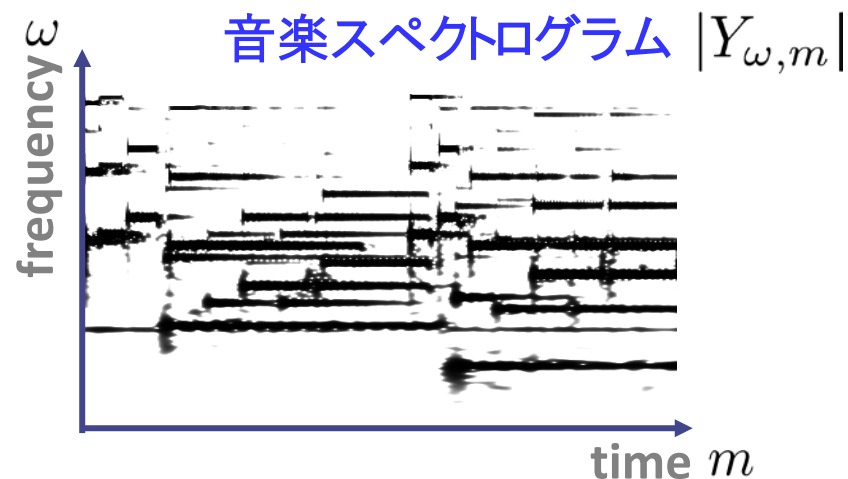
短時間
Fourier変換



生成

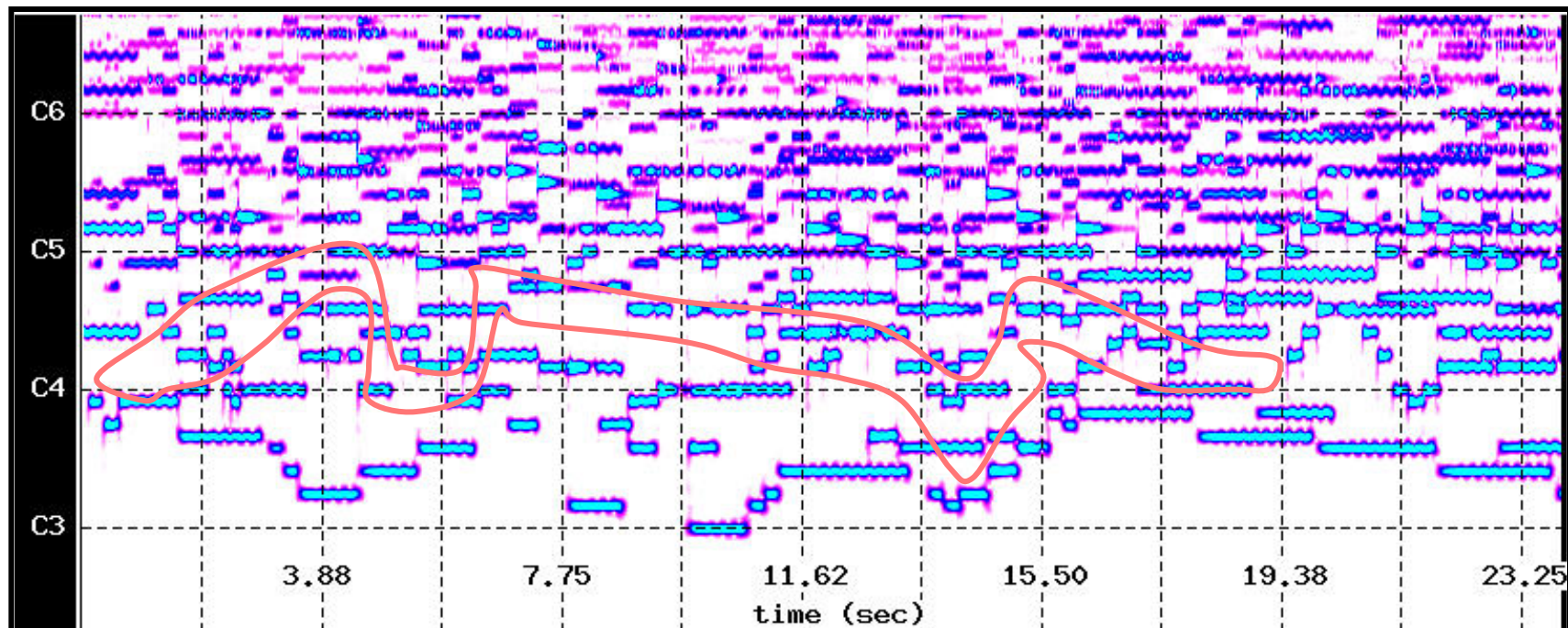


推論



多重音解析の難しさ

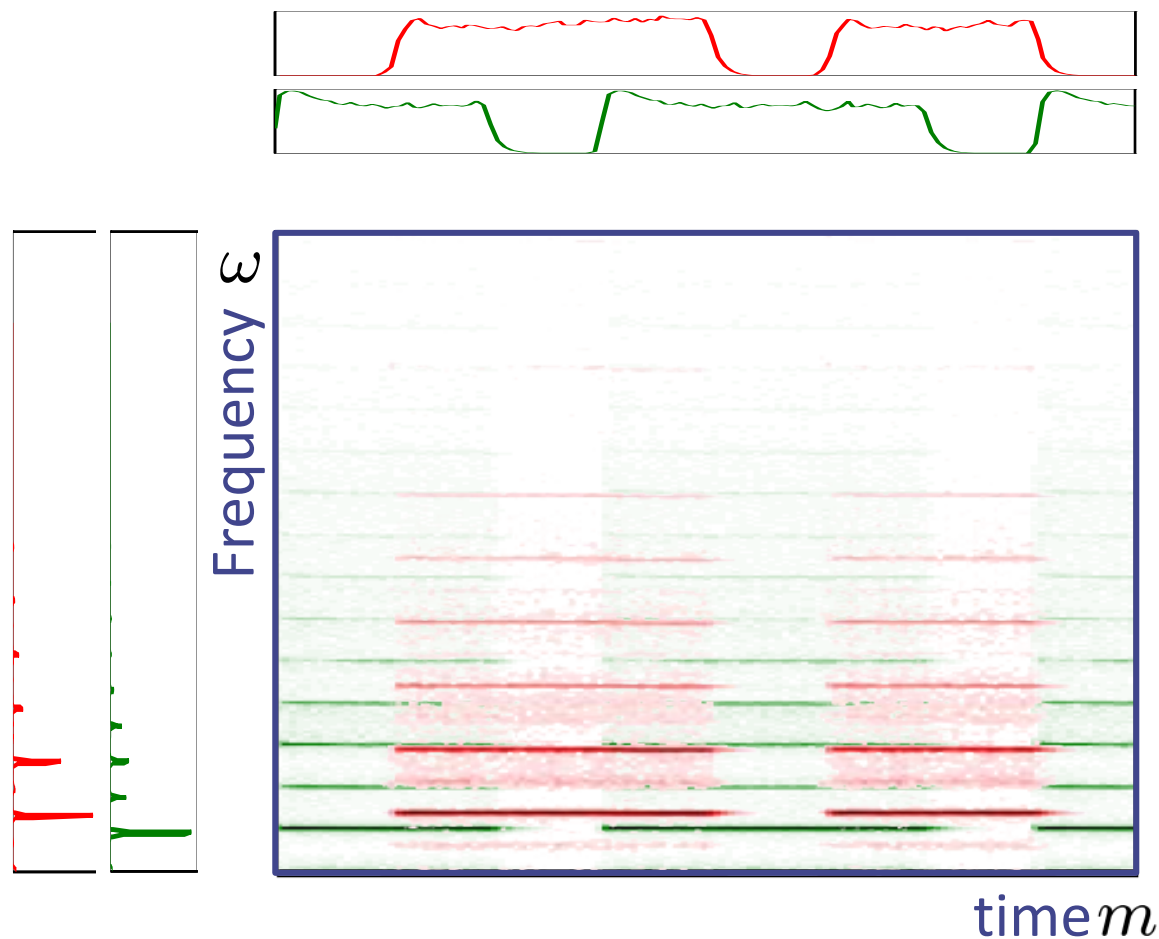
J.S.Bach: Ricercare a 6 aus “Das Musikalische Opfer” BWV1079



A musical score for the piece, showing the flute and strings parts. The flute part is labeled 'flute' and the strings part is labeled 'strings'. The score is in G major and 3/4 time. A red line is drawn across the flute part, highlighting the 'King Friedrich's Theme'.

非負値行列因子分解 (NMF) による多重音解析 [Smaragdis+2003]

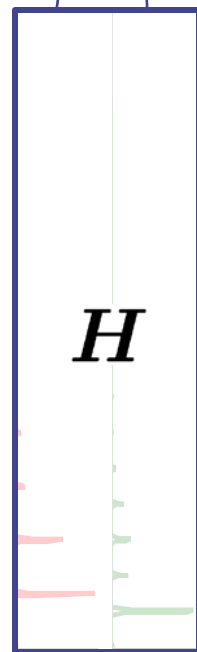
■ 「行列積」としてのスペクトログラム



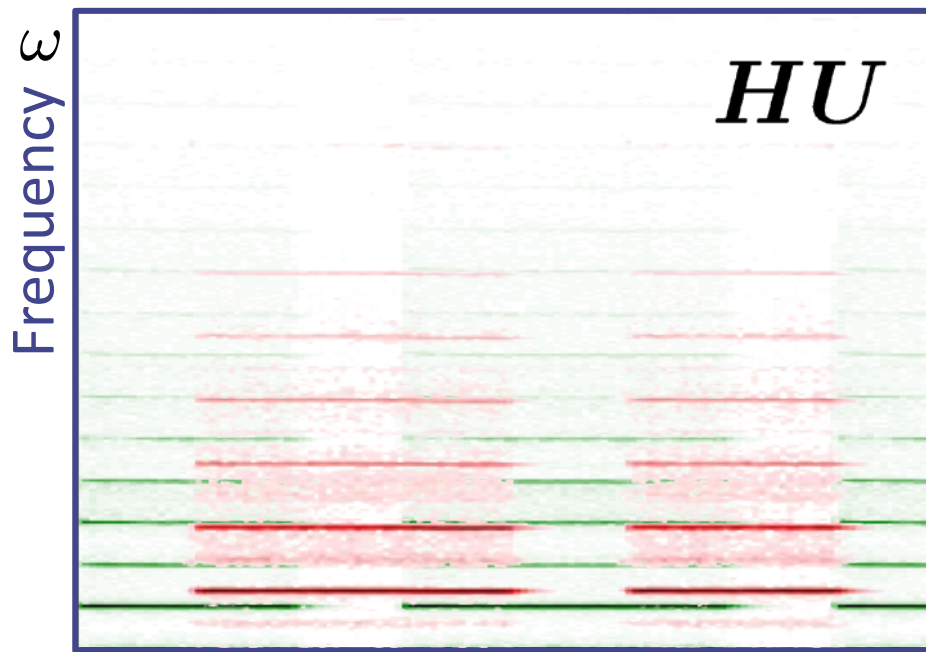
非負値行列因子分解 (NMF) による多重音解析 [Smaragdis+2003]

■ 「行列積」としてのスペクトログラム

各楽音のスペクトル



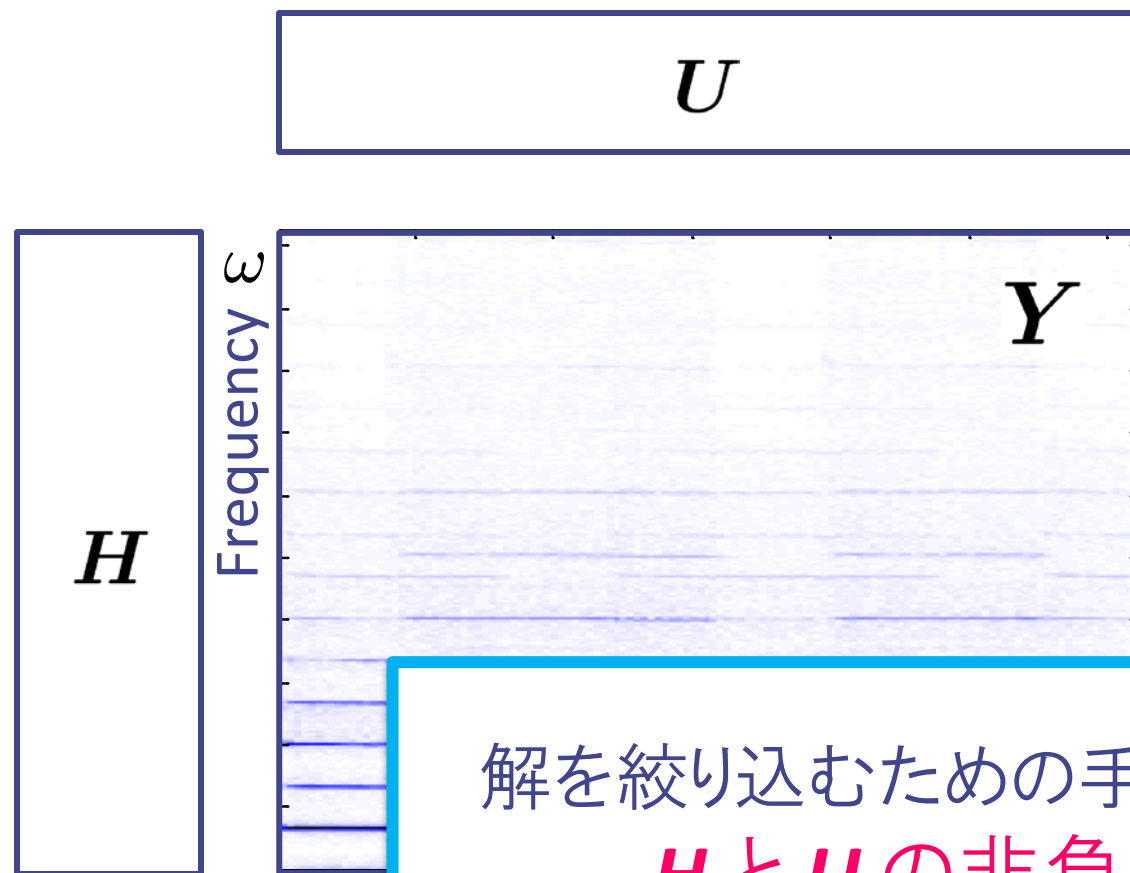
各楽音の「アクティベーション」



time m

非負値行列因子分解 (NMF) による多重音解析 [Smaragdis+2003]

- 行列分解(逆問題)は多重音解析に相当

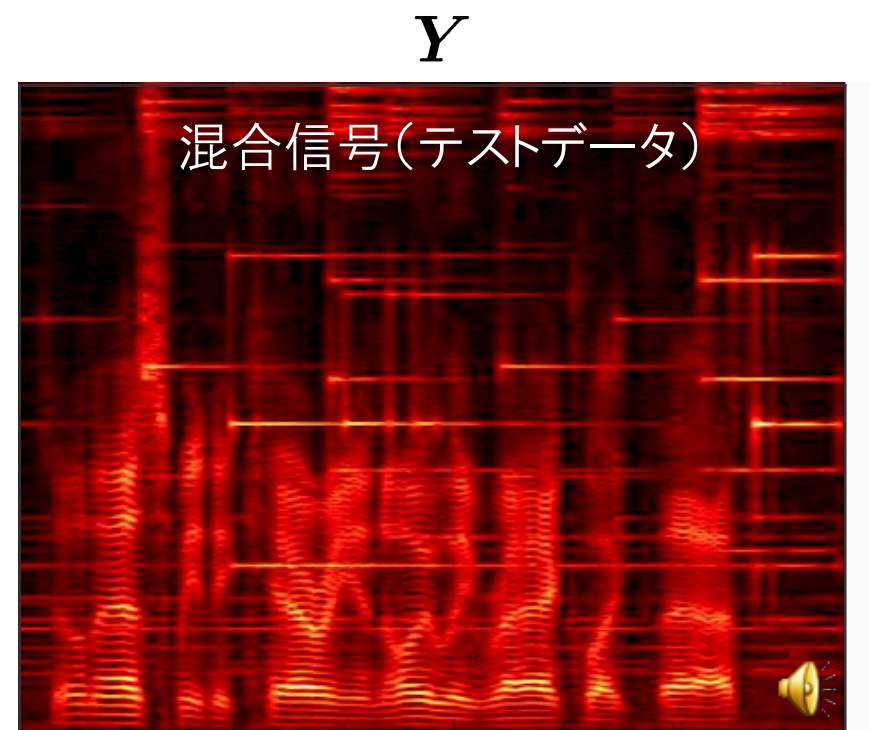
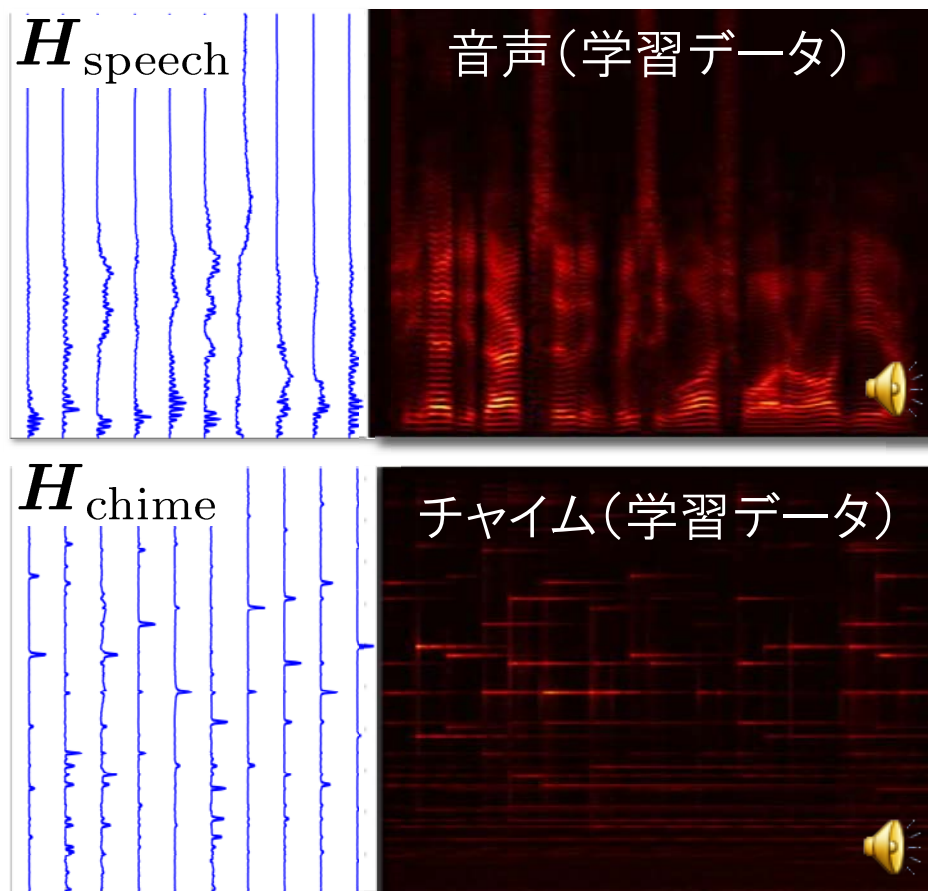


解を絞り込むための手がかり:

H と U の非負性

音声分離への応用例

■「教師あり」モノラル音源分離 [Smaragdis+2007]



$$Y \simeq \underbrace{\begin{bmatrix} H_{\text{speech}} & H_{\text{chime}} \end{bmatrix}}_{\text{固定}} \begin{bmatrix} U_{\text{speech}} \\ U_{\text{chime}} \end{bmatrix}$$

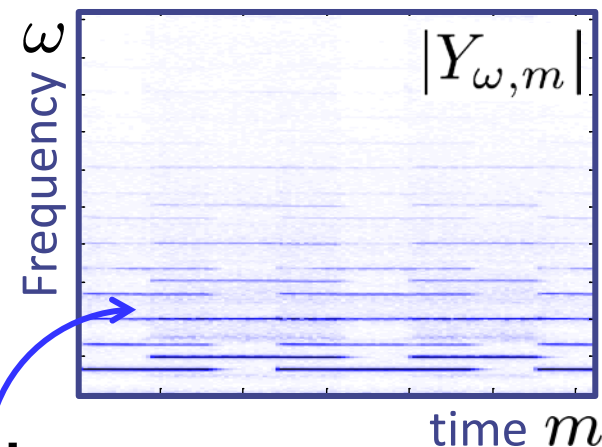
The equation shows the mixed signal Y as a product of a fixed matrix of source filters $\begin{bmatrix} H_{\text{speech}} & H_{\text{chime}} \end{bmatrix}$ and a vector of source signals $\begin{bmatrix} U_{\text{speech}} \\ U_{\text{chime}} \end{bmatrix}$. The matrix is labeled '固定' (fixed). Small speaker icons are placed next to the source signal vector.

スペクトログラムに関する基本事項

- 信号 $y[t]$ の時間周波数展開は線形変換

$$Y_{\omega,m} = \sum_{t=0}^{T-1} W[t]y[t+m]e^{j2\pi\omega t/T}$$

- 信号の時間周波数成分の絶対値成分 $|Y_{\omega,m}|$ を振幅スペクトログラム, 偏角成分 $\arg(Y_{\omega,m})$ を位相スペクトログラムという



- 振幅スペクトログラムは加法的ではない
(多重音信号の振幅スペクトログラムは各音源の振幅スペクトログラムの和にはならない)

⇒ スペクトログラムを行列積で表現するのは厳密には不適当
以上の点を意識したモデル化が必要

NMF

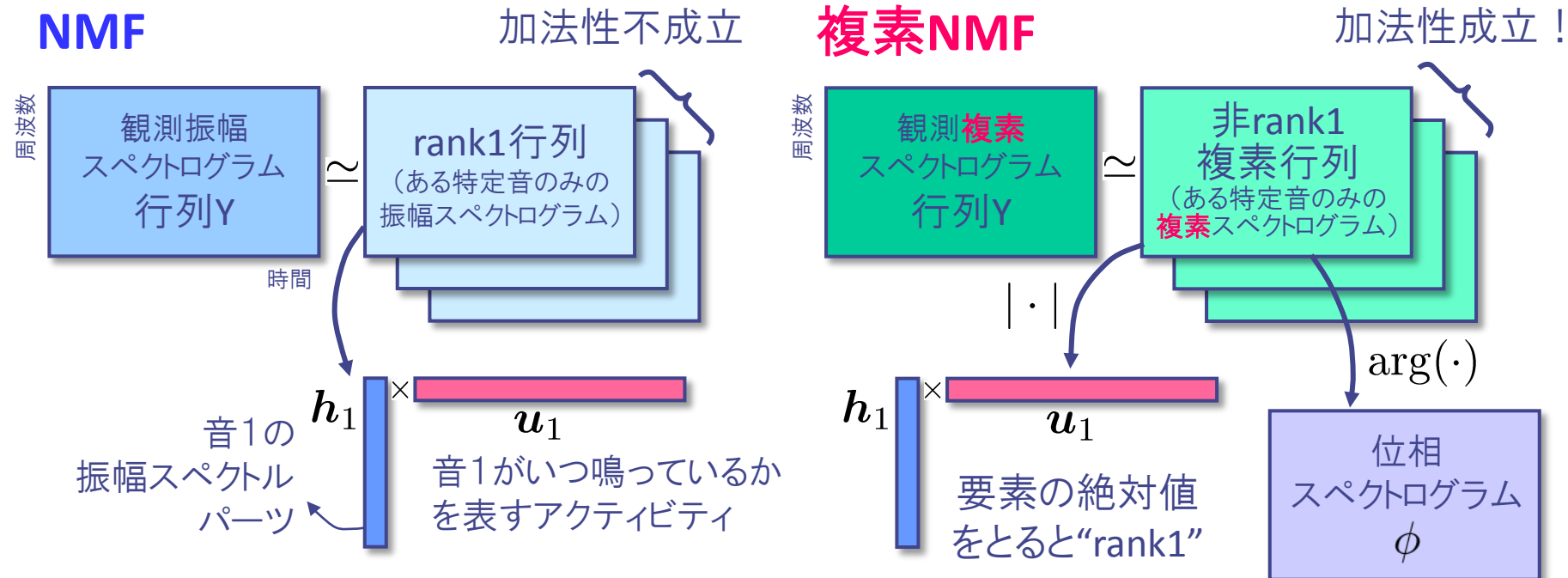
The diagram illustrates the proposed model architecture. It starts with an input box labeled "観測振幅スペクトログラム 行列Y" (Observed Amplitude Spectrogram Matrix Y). This matrix is processed by a "rank1行列" (rank1 matrix), which is noted as being for a specific sound ("ある特定音のみの振幅スペクトログラム"). The output of this step is a vertical blue bar labeled "音1の振幅スペクトル パーツ" (Part of the amplitude spectrum of sound 1) and a horizontal pink bar labeled "音1がいつ鳴っているかを表すアクティビティ" (Activity representing when sound 1 is ringing). The vertical bar is also labeled h_1 and the horizontal bar is labeled u_1 . A curved arrow labeled "時間" (Time) points from the input matrix to the vertical bar. A bracket on the right side of the rank1 matrix indicates it is a stack of matrices.

複素NMF

The diagram illustrates the Rank1 Spectrogram Extraction process. It starts with a green box labeled "観測複素スペクトログラム 行列Y" (Observed Complex Spectrogram Matrix Y). This matrix is compared (indicated by a tilde symbol $\sim$) to a stack of cyan boxes labeled "非rank1 複素行列" (Non-rank1 Complex Matrix), which is further described as "(ある特定音のみの複素スペクトログラム)" (Complex Spectrogram of a specific sound only). The process involves two main steps: 1) Element-wise multiplication ($\cdot$) of matrix Y with a vector h_1 (represented by a blue vertical bar) to produce a vector u_1 (represented by a pink horizontal bar). This step is labeled "要素の絶対値をとると'rank1'" (Taking the absolute value of the elements gives 'rank1'). 2) Applying the $\arg(\cdot)$ function to the result of the multiplication to extract the "位相スペクトログラム ϕ " (Phase Spectrogram ϕ), shown in a purple box.

複素NMF [Kameoka+2008]

■ NMFによる音源分離のコンセプトを複素領域で実現するアイディア



■ 複素スペクトログラムモデル

- 複素スペクトログラム $Y_{\omega,m} \in \mathbb{C}$ をモデル化

$$\begin{aligned}
 Y_{\omega,m} &= \sum_l X_{l,\omega,m} \quad |X_{l,\omega,m}| = H_{\omega,l} U_{l,m} \\
 &= \sum_l |X_{l,\omega,m}| e^{j\phi_{l,\omega,m}} = \sum_l H_{\omega,l} U_{l,m} e^{j\phi_{l,\omega,m}} \quad H_{\omega,l} \geq 0, U_{l,m} \geq 0
 \end{aligned}$$

行列積の形にならない！

複素NMF [Kameoka+2008]

■ 変数の定義

- 複素スペクトログラム $Y_{\omega,m} \in \mathbb{C}$
- 基底スペクトル $\mathcal{H} \equiv \{H_{\omega,l}\}$
- 位相スペクトログラム $\phi \equiv \{\phi_{l,\omega,m}\}$
- 基底アクティビティ $\mathcal{U} \equiv \{U_{l,m}\}$

複素スペクトログラムモデル

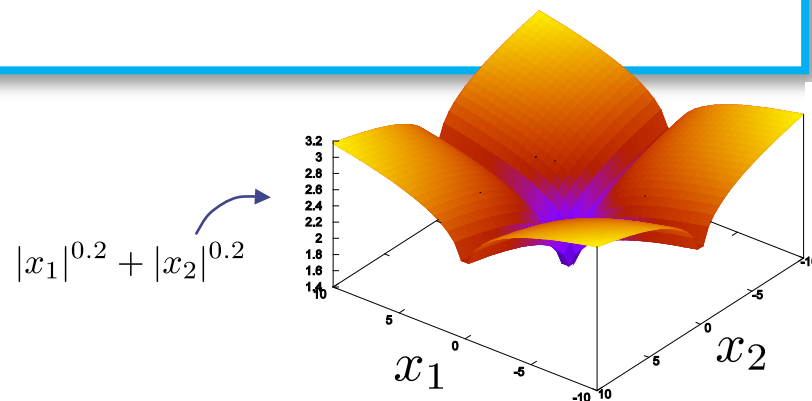
$$F_{\omega,m} = \sum_{l=1}^L H_{\omega,l} U_{l,m} e^{j\phi_{l,\omega,m}}$$

■ 最適化問題の定式化

$$\text{minimize } f(\mathcal{H}, \mathcal{U}, \phi) := \sum_{\omega,m} |Y_{\omega,m} - F_{\omega,m}|^2 + 2\lambda \sum_{l,m} |U_{l,m}|^p$$

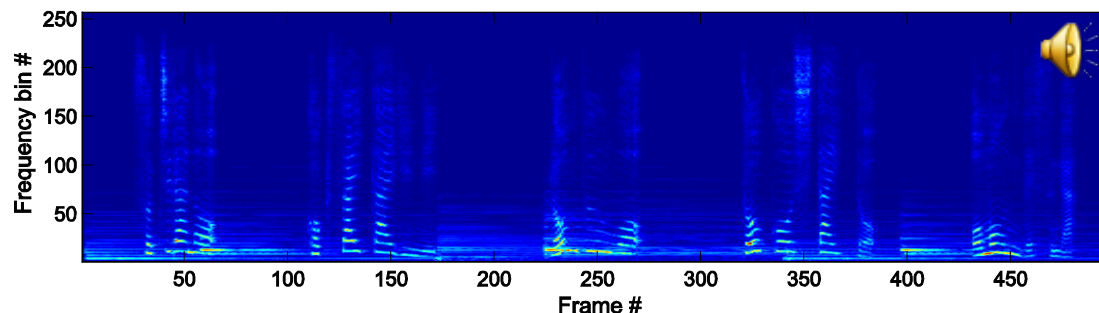
$$\text{subject to } \sum_k H_{\omega,k} = 1 \quad (l = 1, \dots, L)$$

スパース正則化項



複素NMFによる音声と音楽の混合信号の分離

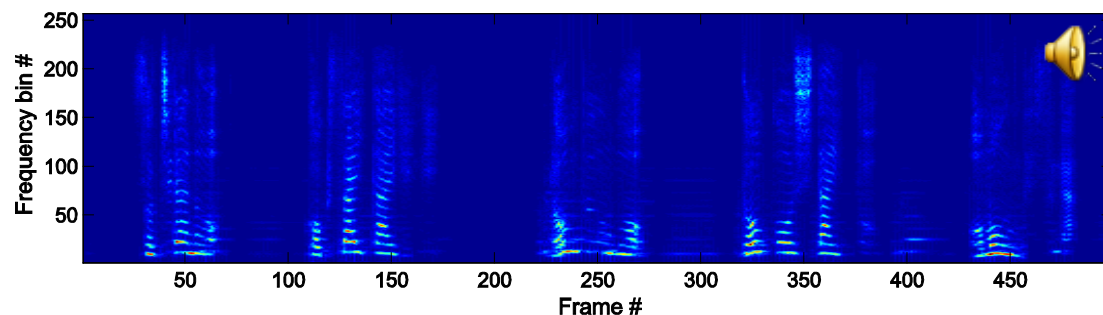
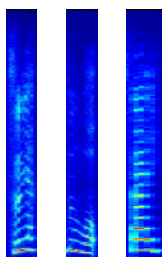
混合信号(音声+音楽)



I

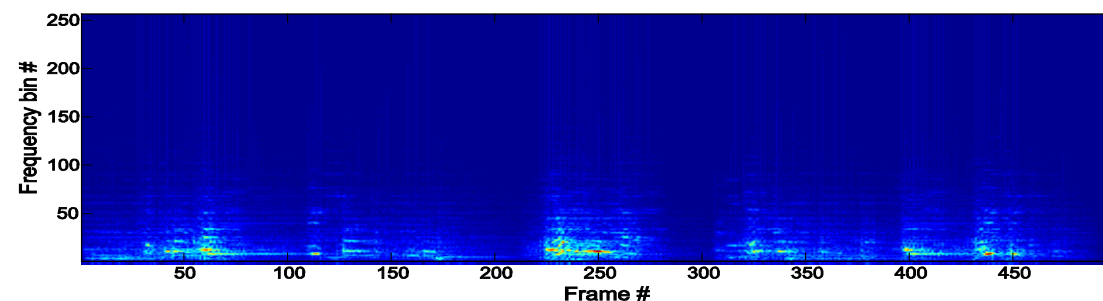
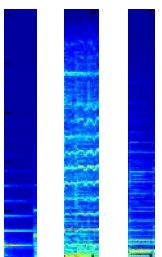
抽出音声信号

音声パーツ



+

音楽パーツ



Complex NMF

by H. Kameoka

Open

サンプリング 16000 Hz

フレーム長 64 ms

トータル時間 23 s

Factorize

スペクトルパーツ数 30

エンベロープパーツ数 5

Save

表示するスペクトルパーツ

No. 1 - 10

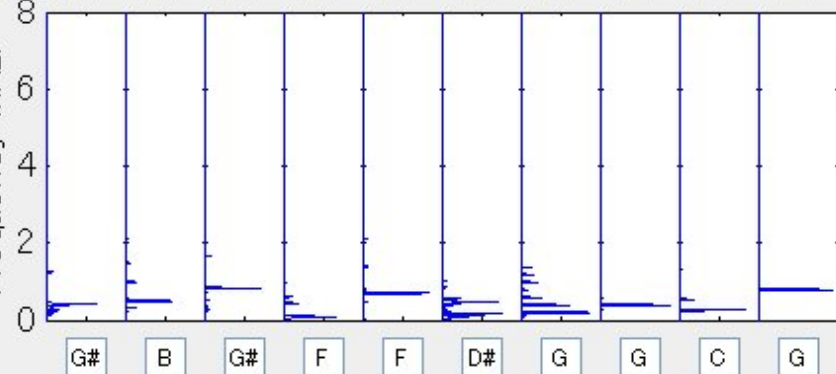
トランスポーズ

0	+	+	+	+	+	+	+	+	+	+
m	-	-	-	-	-	-	-	-	-	-

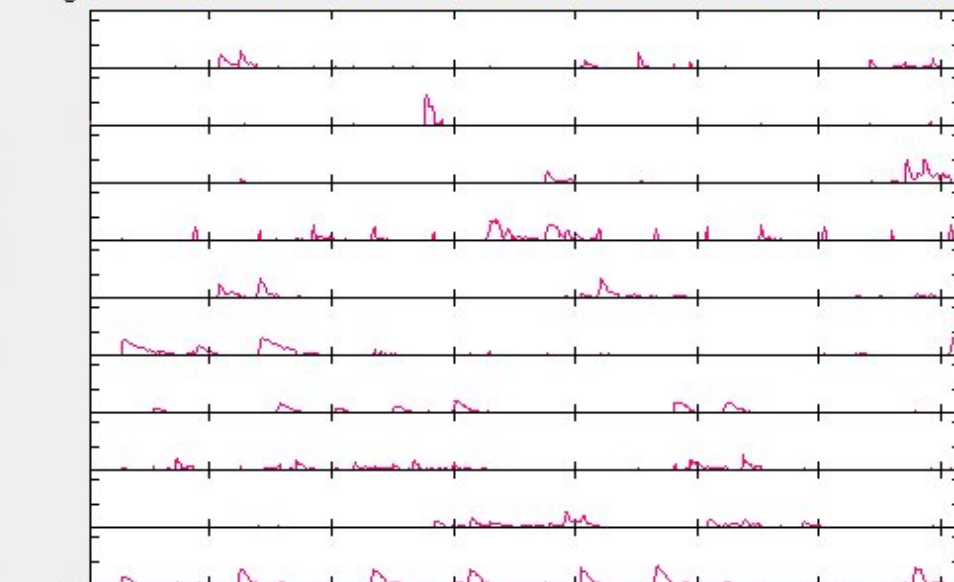
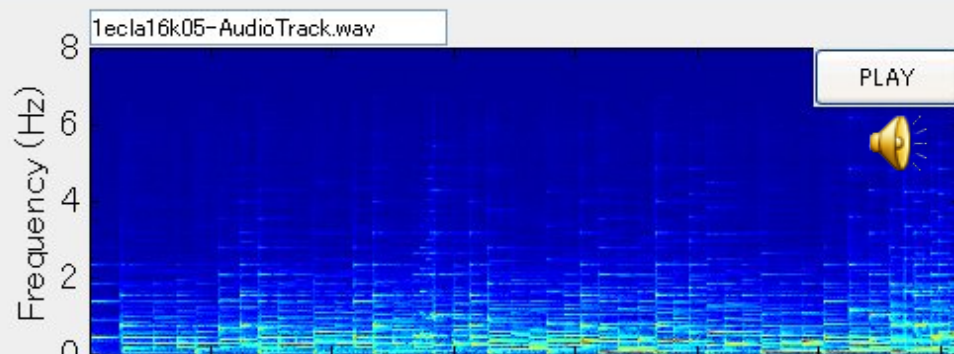
ボリュームコントロール

0										
M										

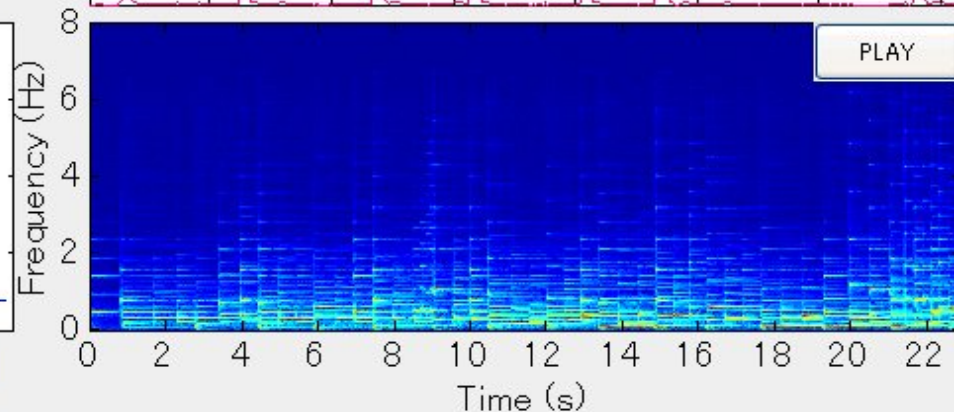
Frequency (kHz)



Frequency (Hz)



Frequency (Hz)



Complex NMF

by H. Kameoka

Open

サンプリング 16000 Hz

フレーム長 64 ms

トータル時間 23 s

Factorize

スペクトルパーツ数 30

エンベロープパーツ数 5

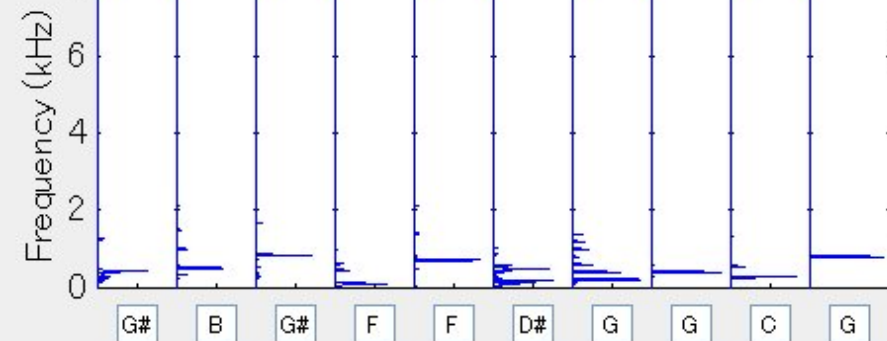
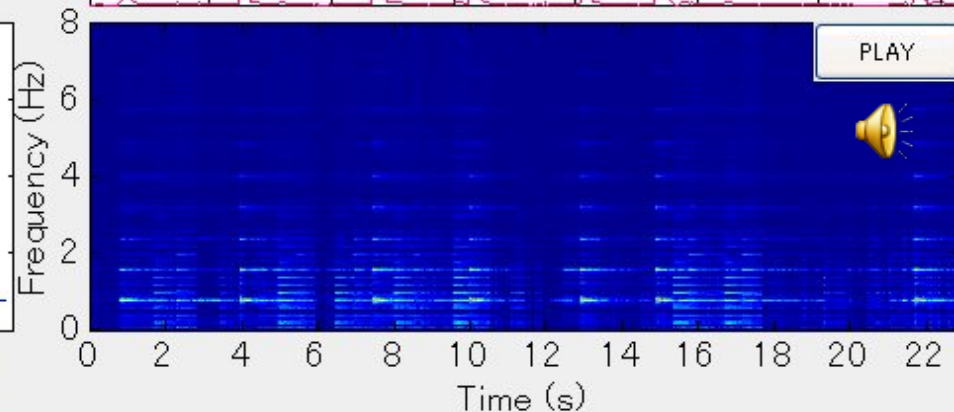
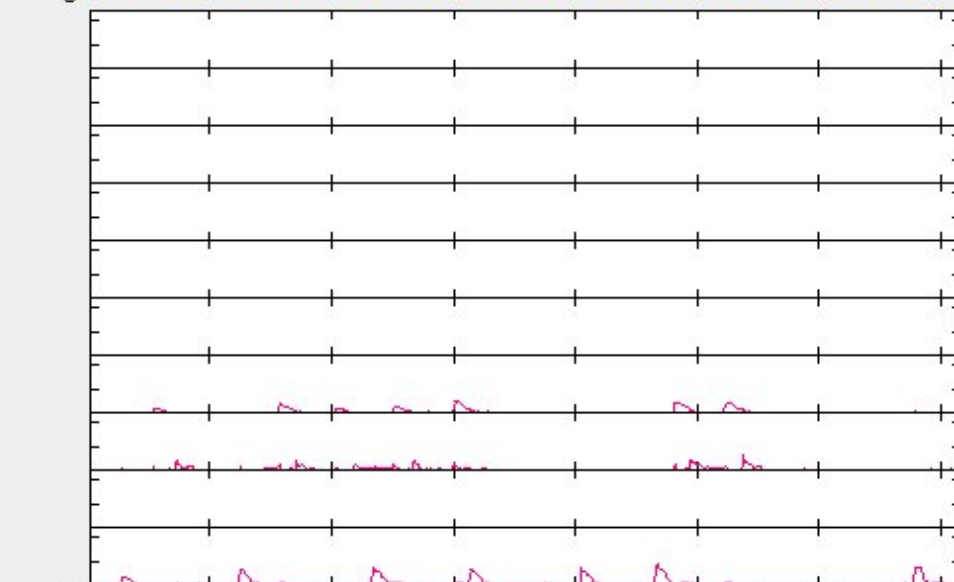
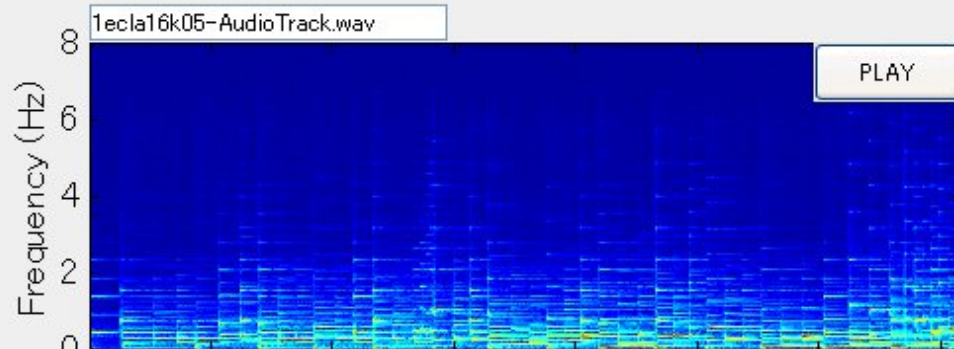
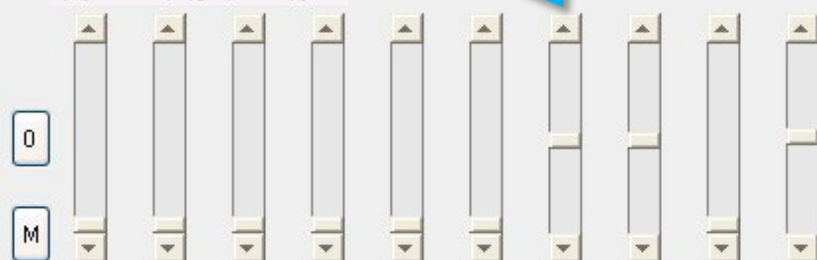
Save

表示するスペクトルパーツ

No. 1 - 10

「ソ」以外は音量オフ

ボリュームコントロール



Complex NMF

by H. Kameoka

Open

サンプリング 16000 Hz

フレーム長 64 ms

トータル時間 23 s

Factorize

スペクトルパーツ数 30

エンベロープパーツ数 5

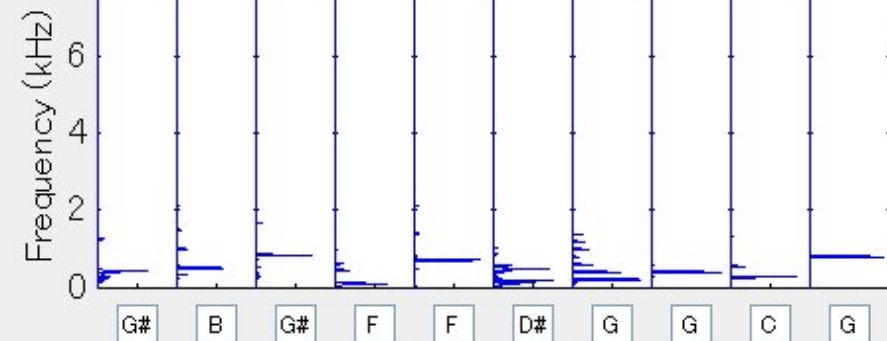
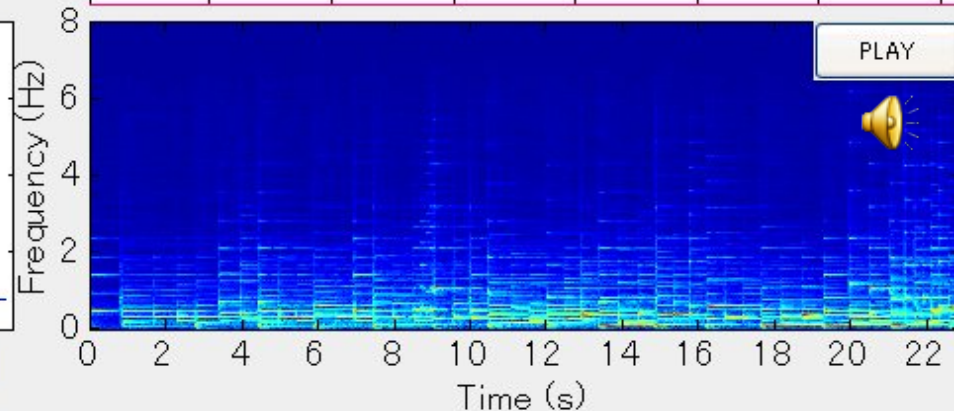
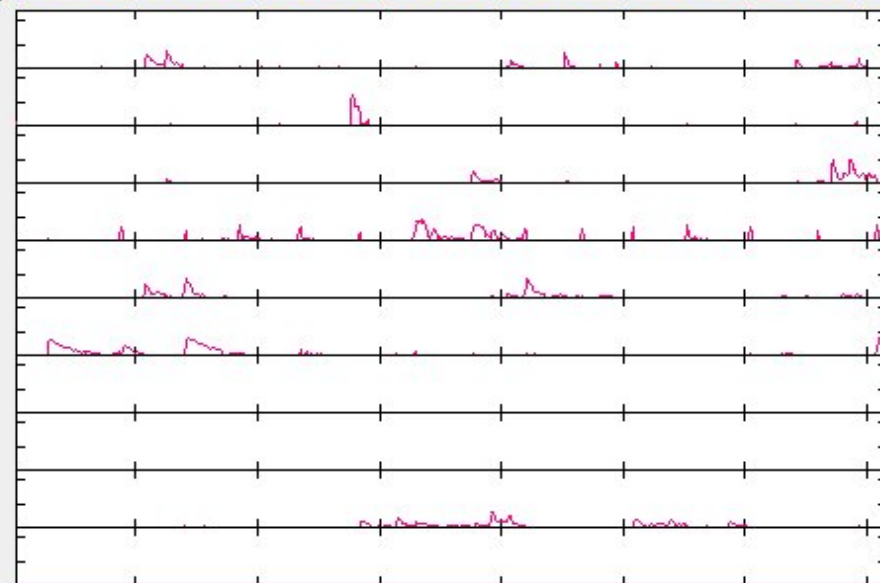
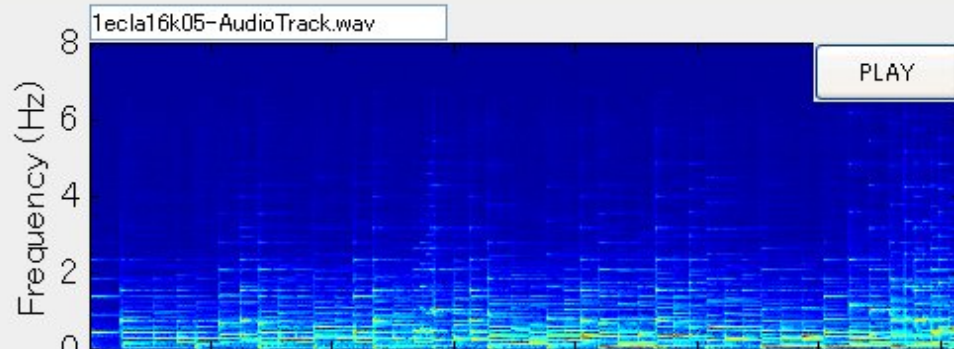
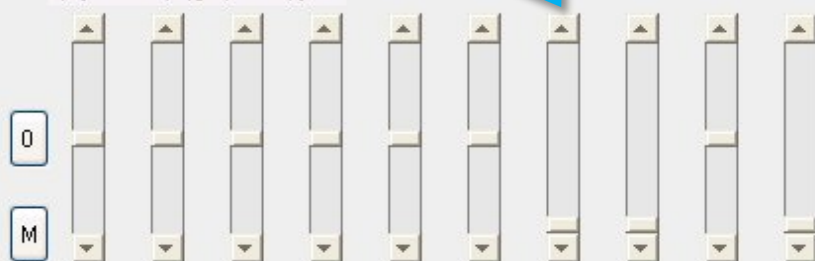
Save

表示するスペクトルパーツ

No. 1 - 10

「ソ」だけ音量オフ

ボリュームコントロール



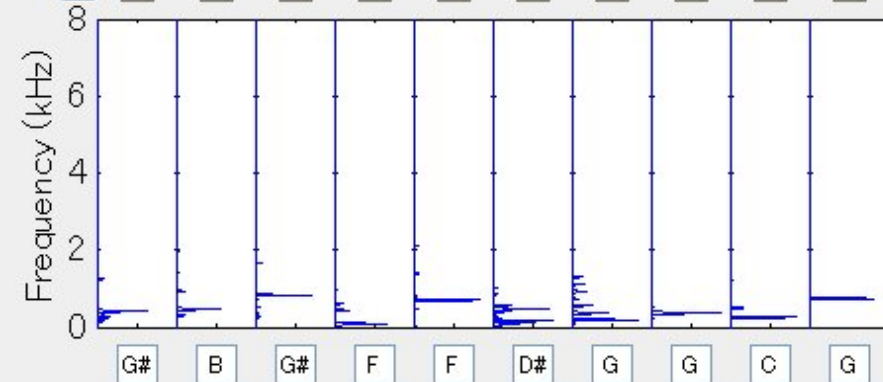
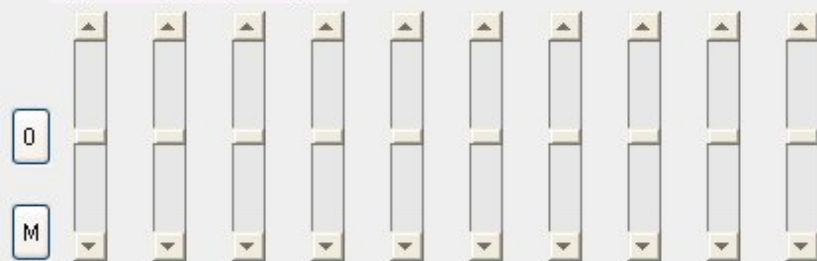
一部の基底関数に対してのみスペクトル伸縮変形を施し、混合信号を再構成（スペクトル伸縮はピッチトランスポーズに相当。例えばトランスポーズ「-1」は、半音下げという意味。）

No. 1-10

トランスポ

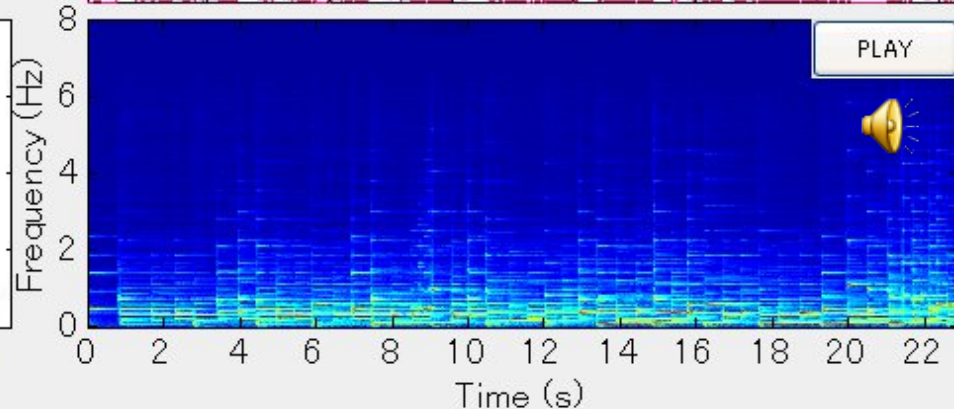
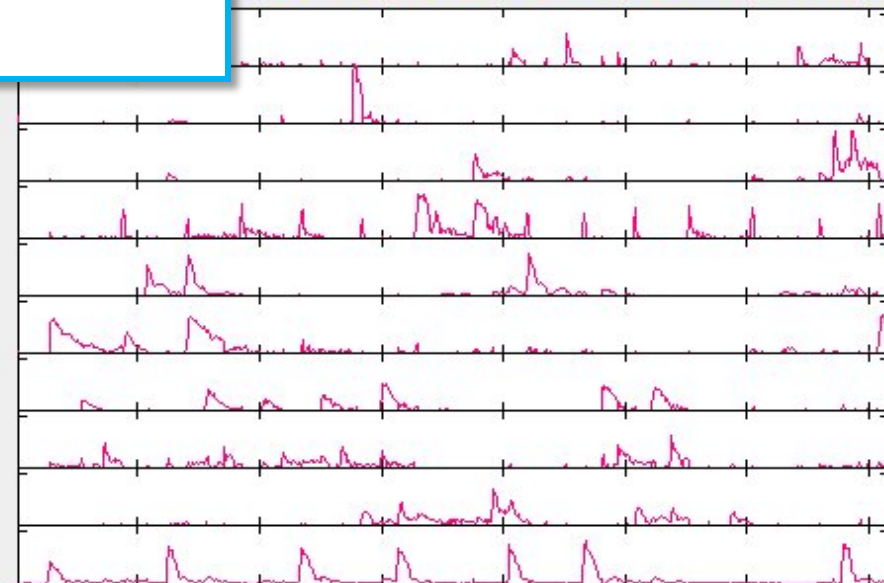


ボリュームコントロール



ack.wav

PLAY



板倉齋藤 NMF [Févotte+2009]

■ 多重音スペクトログラムの生成モデル化

$$C_{l,\omega,m} \sim \mathcal{N}_{\mathbb{C}}(0, H_{\omega,l} U_{l,m})$$

音源 l の複素スペクトログラム

$$Y_{\omega,m} = \sum_l C_{l,\omega,m}$$

多重音の複素スペクトログラム



→ $C_{l,\omega,m}$ と $C_{l',\omega,m}$ の独立性を仮定

$$Y_{\omega,m} \sim \mathcal{N}_{\mathbb{C}}\left(0, \sum_l H_{\omega,l} U_{l,m}\right)$$

$p(Y|\theta)$ $\theta = \{H, U\}$

$p(\theta)$: H と U が負値の場合 0 になる分布

$$\hat{\theta} = \operatorname{argmax}_{\theta} p(Y|\theta)$$

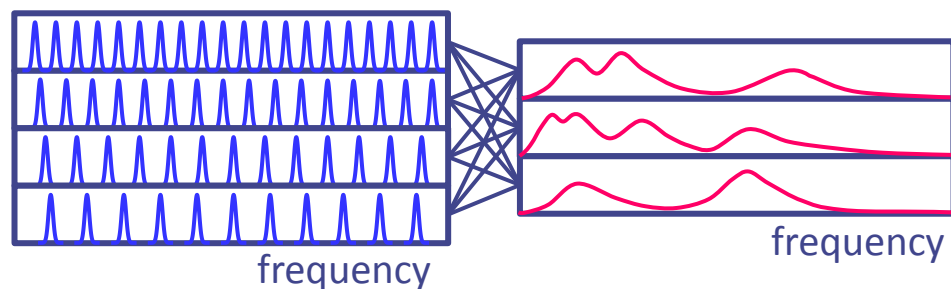
$$= \operatorname{argmin}_{H,U} \sum_{\omega,m} \left(\frac{|Y_{\omega,m}|^2}{\sum_l H_{\omega,l} U_{l,m}} - \log \frac{|Y_{\omega,t}|^2}{\sum_k H_{\omega,l} U_{l,m}} - 1 \right)$$

板倉齋藤距離

H と U の非負制約のもとで $Y \simeq HU$ となる H と U を求める問題

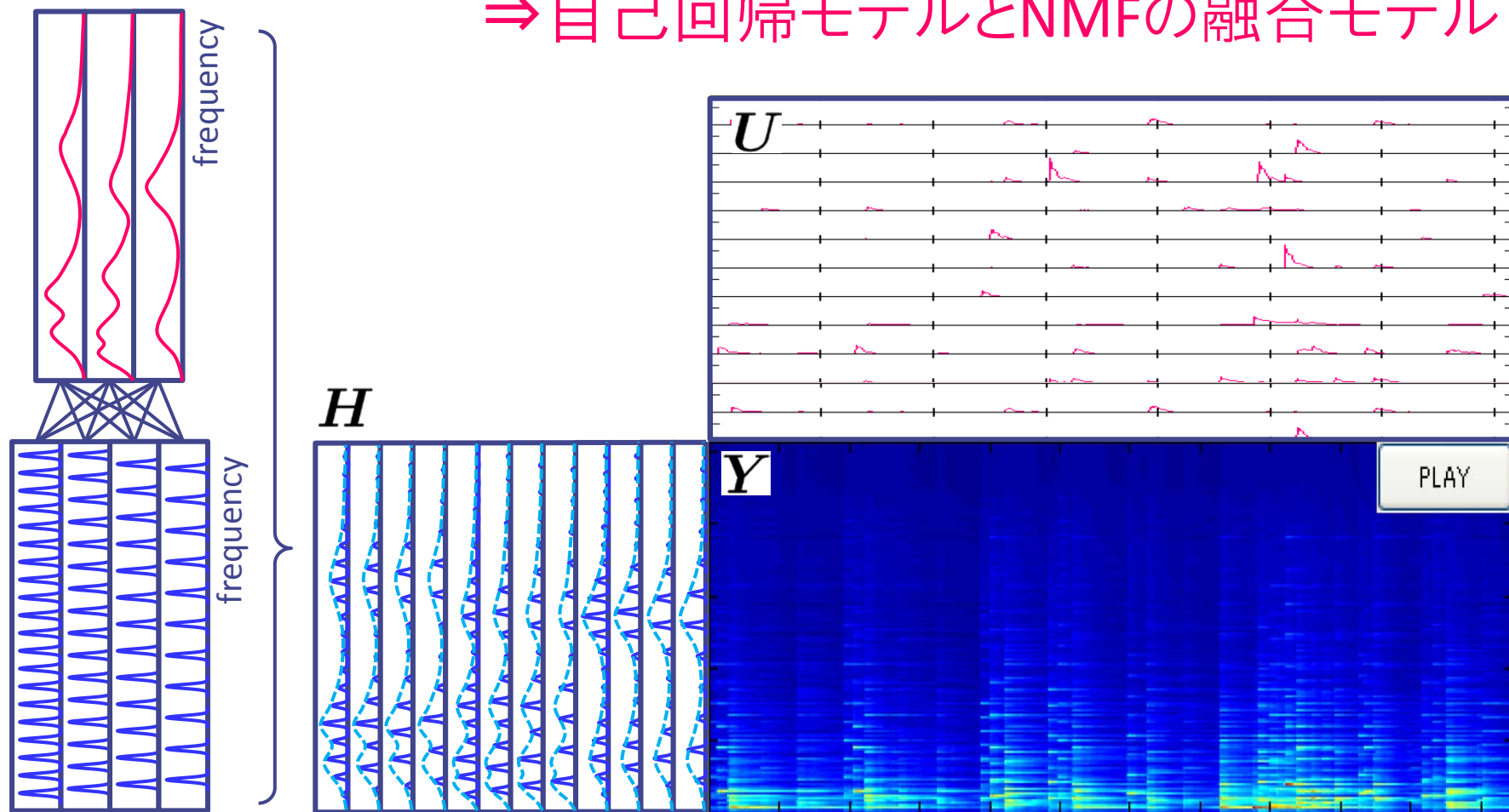
複合自己回帰モデル [Kameoka+2009, Yoshii+2012]

- 混合信号をI個のソースとJ個のフィルタの出力の和で表現



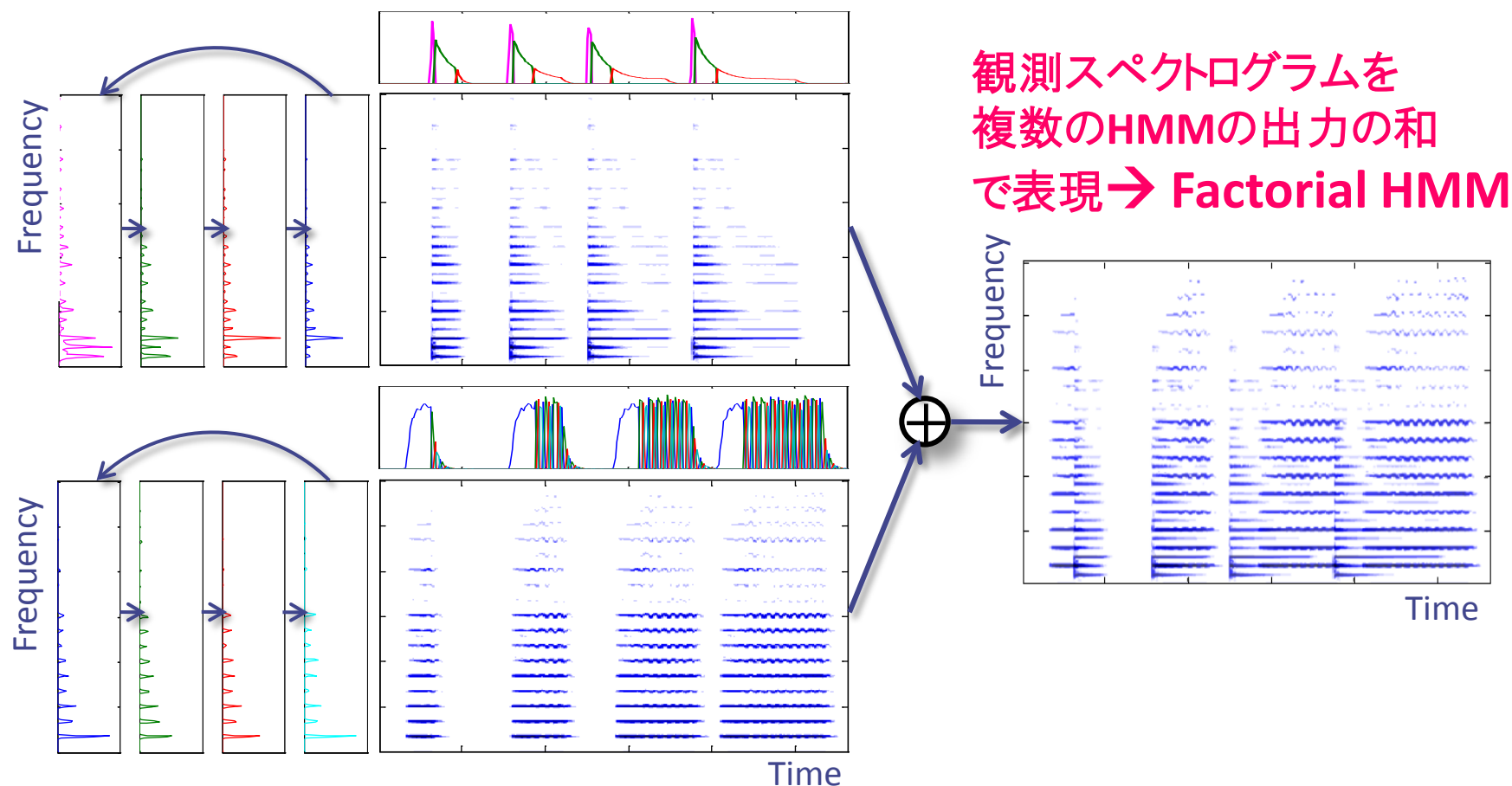
複合自己回帰モデル [Kameoka+2009, Yoshii+2012]

- 混合信号をI個のソースとJ個のフィルタの出力の和で表現
⇒ 自己回帰モデルとNMFの融合モデル



Factorial HMMによる多重音モデル [Nakano+2010]

- 楽音のスペクトルは通常時間変化する
⇒ 状態遷移により各基底スペクトルが時間変化するモデル



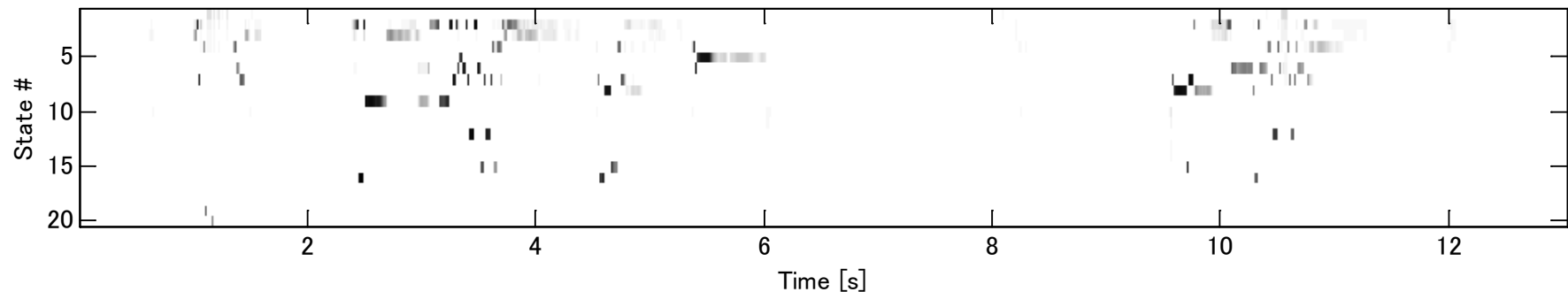
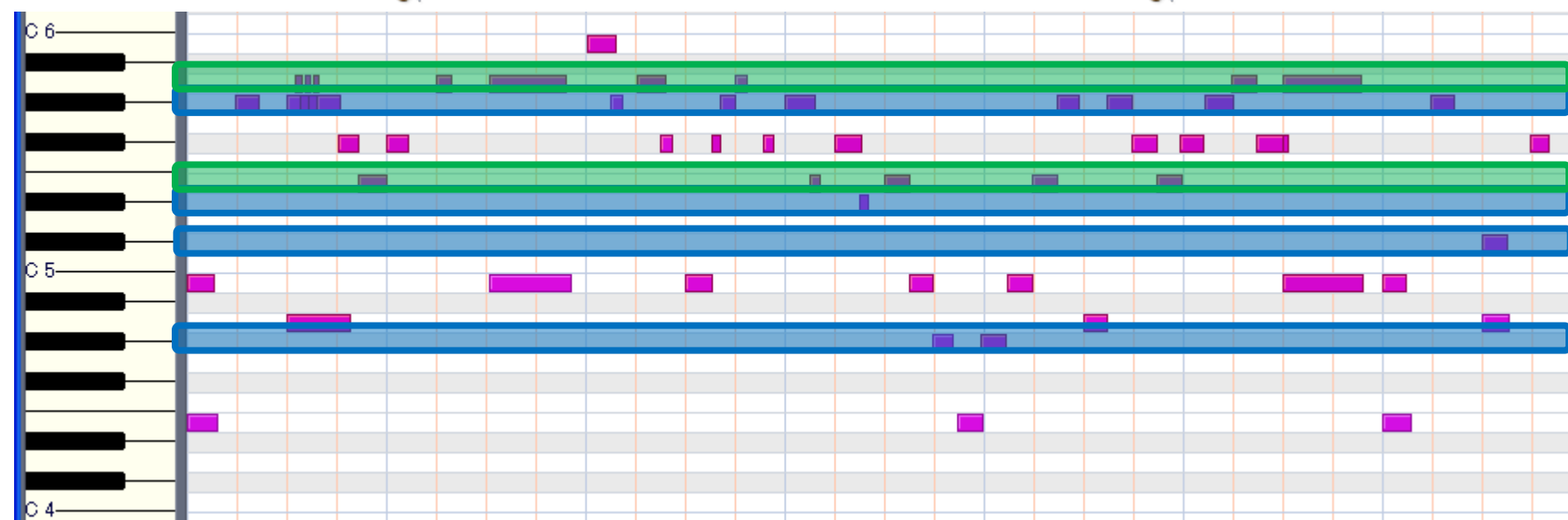
FHMMによる楽音加工例 [Nakano+2010]

■ Note-level audio manipulation (example 1)

Original signal:



Manipulated signal:



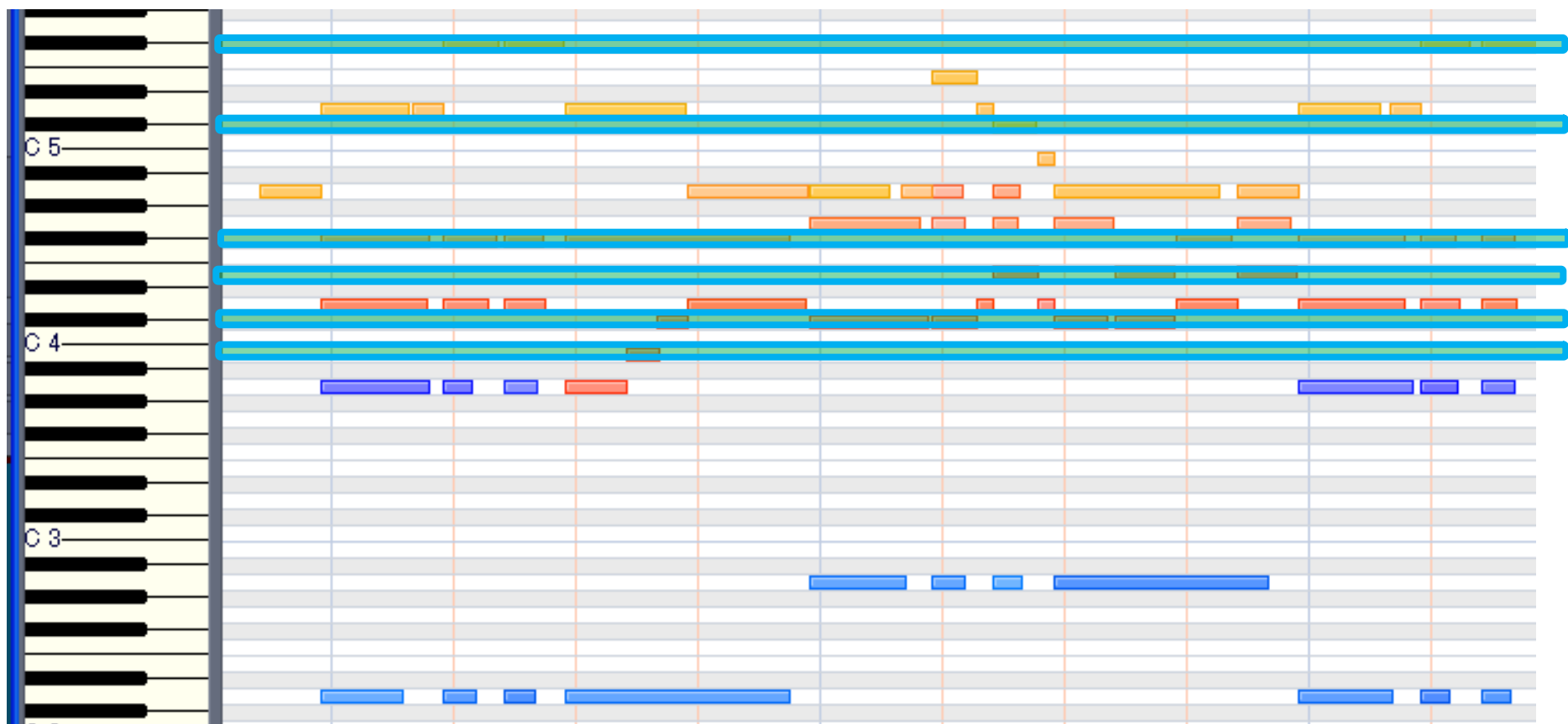
FHMMによる楽音加工例 [Nakano+2010]

■ Note-level audio manipulation (example 2)

Original signal:



Manipulated signal:



音声・音楽・音響信号に関する諸問題

■ 音声・音楽・音響信号の認識・分析・合成・変換・分離

	認識	分析	合成	変換	分離
音声					
音楽			—	—	
音響				—	

⇒「識別」、「回帰」、「逆問題」のいずれかに相当

音声・音楽・音響信号に関する諸問題

■ 音声・音楽・音響信号の認識・分析・合成・変換・分離

	認識	分析	合成	変換	分離
音声					
音楽			—	—	
音響				—	

⇒「識別」、「回帰」、「逆問題」のいずれかに相当

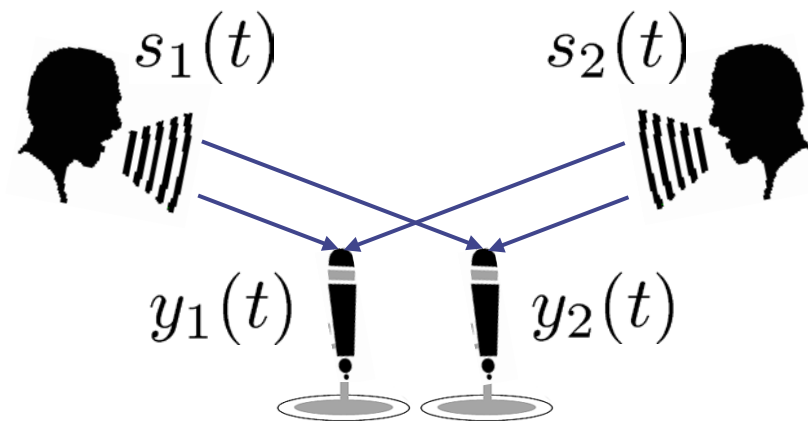
ブラインド音源分離 (BlindSourceSeparation)

■ 複数のマイクで取得した音響信号のみから各音源信号を分離／定位する問題

- 音源信号, 混合過程が
いずれも未知であること
から「ブラインド」

■ 実応用場面

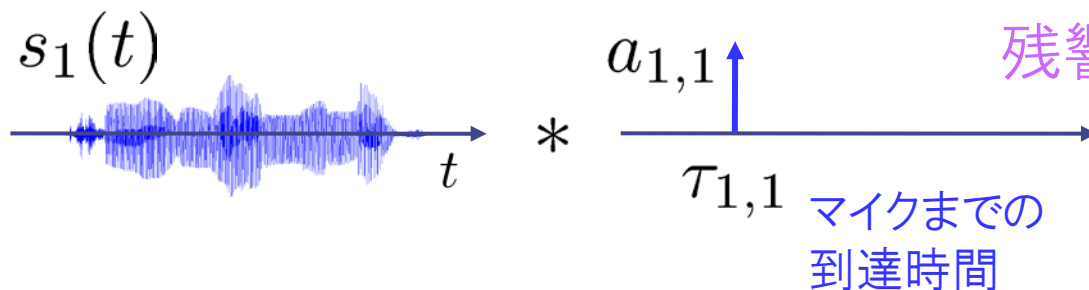
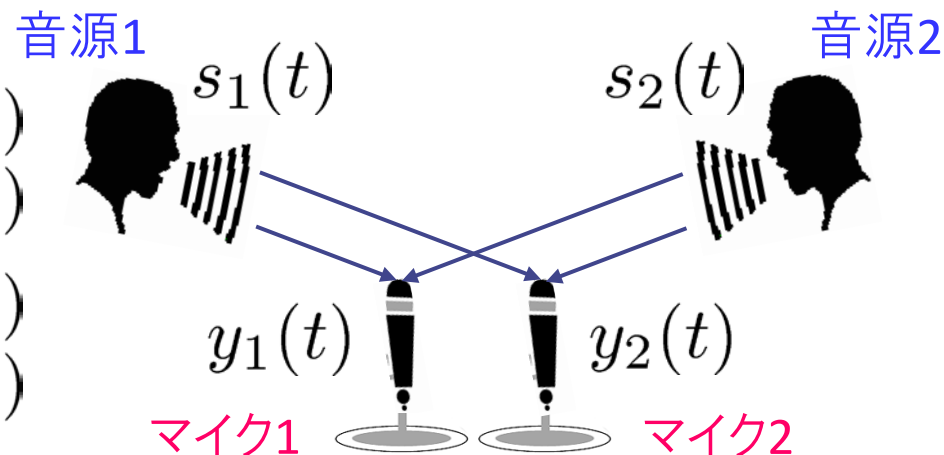
- 音を使った監視システム
 - ◆ どこで何が起こっているかを検知
 - ◆ 介護や防犯のための安全モニタリングとしての応用
Cf. ShotSpotter (米国で開発されている発砲事件の検知システム)
- 人間のカクテルパーティー効果の模倣
 - ◆ ロボット聴覚, 補聴器への応用



マイクロホンで観測される信号の生成プロセス

■ “畳みこみ混合”

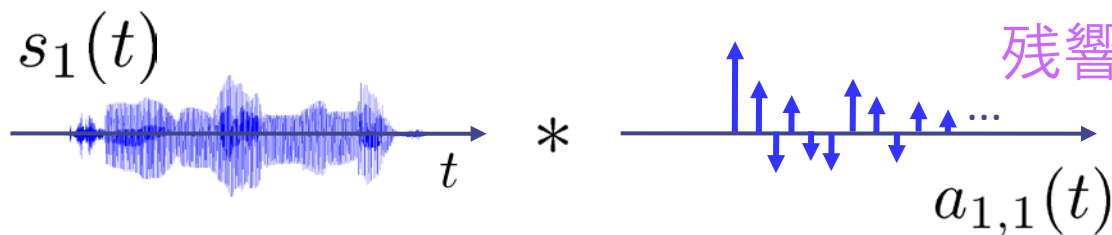
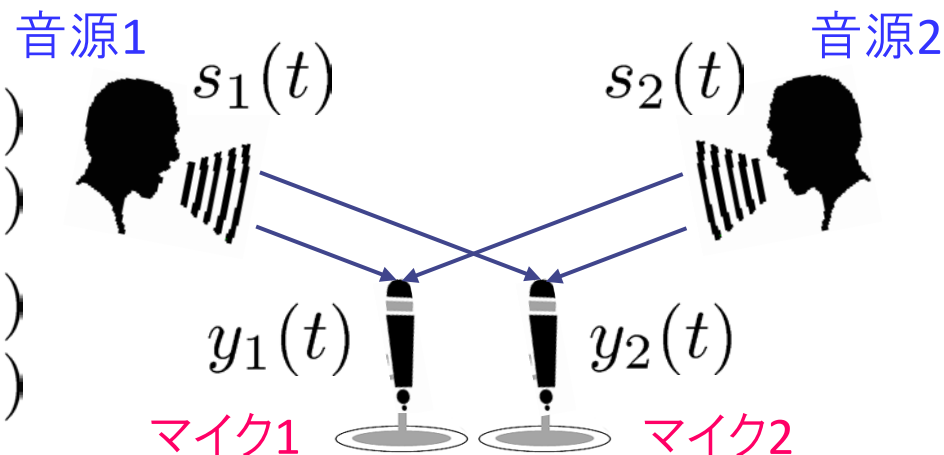
- 音源1 → マイク1: $a_{1,1}s_1(t + \tau_{1,1})$
- 音源1 → マイク2: $a_{2,1}s_1(t + \tau_{2,1})$
- 音源2 → マイク1: $a_{1,2}s_2(t + \tau_{1,2})$
- 音源2 → マイク2: $a_{2,2}s_2(t + \tau_{2,2})$



マイクロホンで観測される信号の生成プロセス

■ “畳みこみ混合”

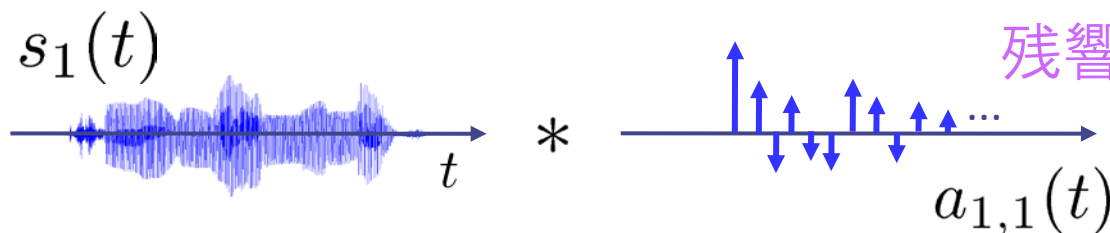
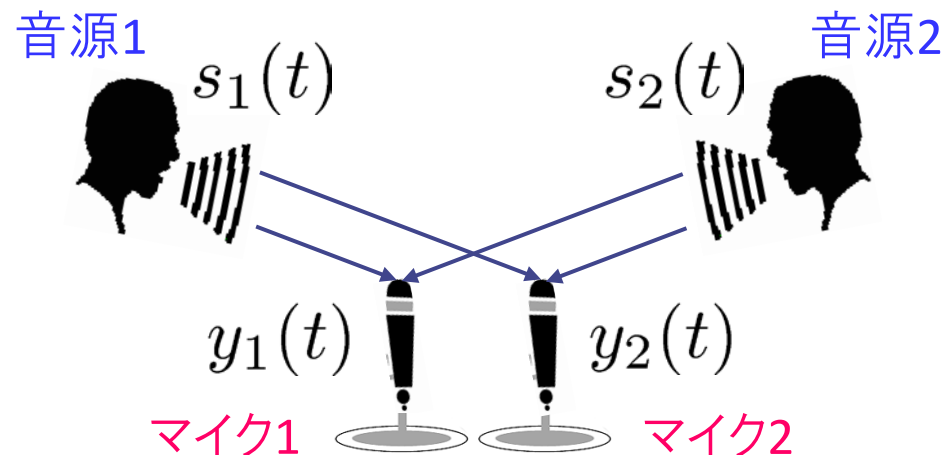
- 音源1 → マイク1: $a_{1,1}s_1(t + \tau_{1,1})$
- 音源1 → マイク2: $a_{2,1}s_1(t + \tau_{2,1})$
- 音源2 → マイク1: $a_{1,2}s_2(t + \tau_{1,2})$
- 音源2 → マイク2: $a_{2,2}s_2(t + \tau_{2,2})$



マイクロホンで観測される信号の生成プロセス

■“畳みこみ混合”

- 音源1→マイク1: $a_{1,1}(t) * s_1(t)$
- 音源1→マイク2: $a_{2,1}(t) * s_1(t)$
- 音源2→マイク1: $a_{1,2}(t) * s_2(t)$
- 音源2→マイク2: $a_{2,2}(t) * s_2(t)$



$\therefore$ マイク1の観測信号: $y_1(t) = a_{1,1}(t) * s_1(t) + a_{1,2}(t) * s_2(t)$
 マイク2の観測信号: $y_2(t) = a_{2,1}(t) * s_1(t) + a_{2,2}(t) * s_2(t)$

$\Rightarrow y_m(t) = \sum_k a_{m,k}(t) * s_k(t)$

マイクロホンで観測される信号の生成プロセス

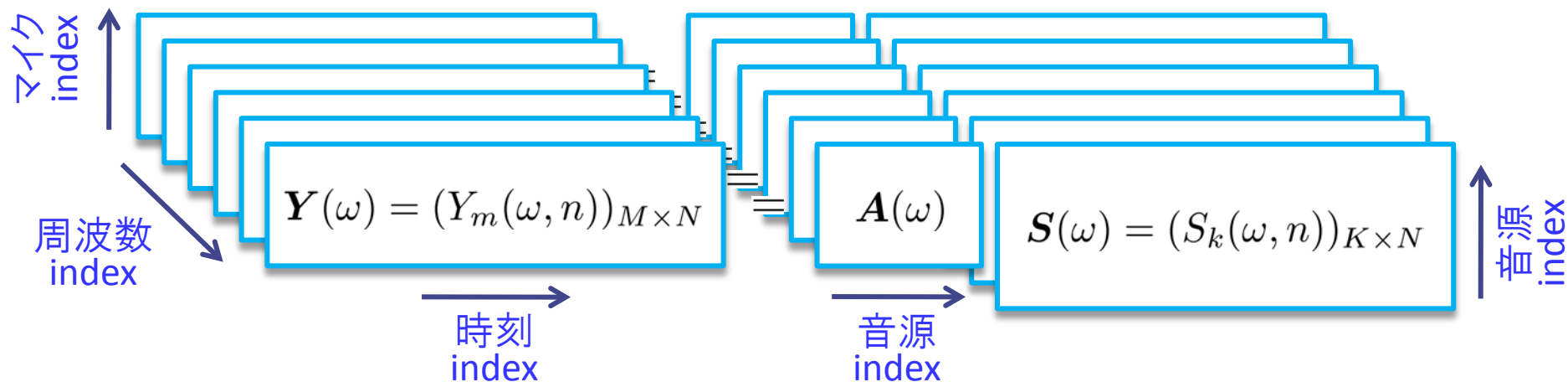
- 畳み込み定理を利用して
畳み込み混合モデルを時間周波数領域に展開：

$$y_m(t) = \sum_k a_{m,k}(t) * s_k(t) \longrightarrow Y_m(\omega, n) = \sum_k A_{m,k}(\omega) S_k(\omega, n)$$

「時間領域の
畳み込み混合モデル」

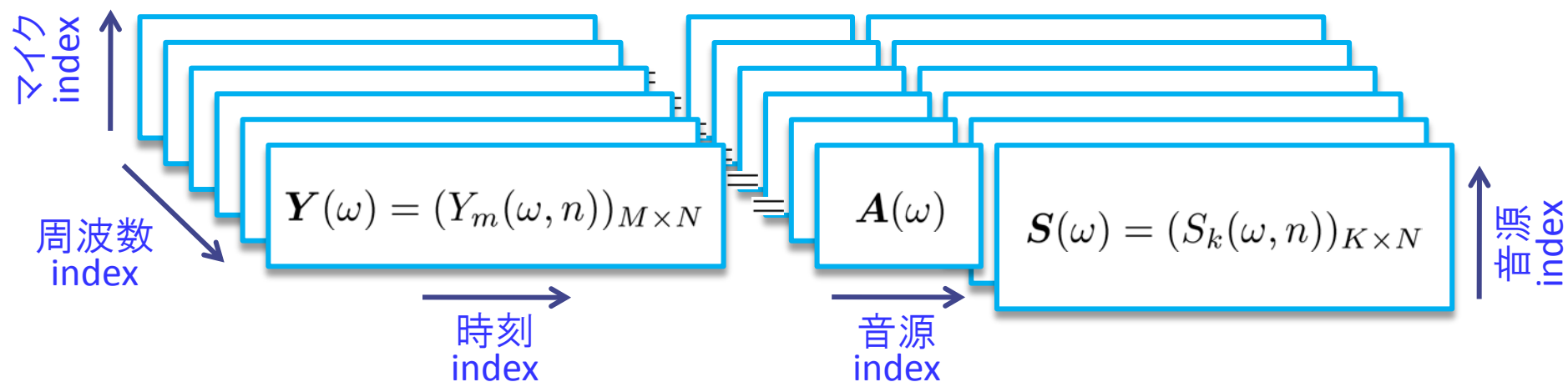
「時間周波数領域の
瞬時混合モデル」

周波数 ω ごとに見れば行列積

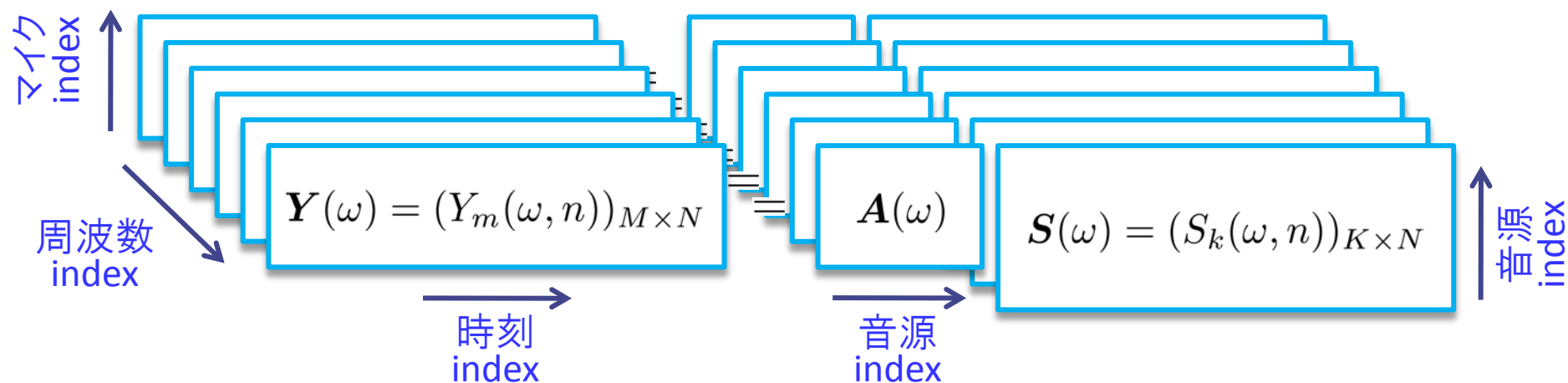


逆問題への手がかり

周波数 ω ごとに見れば行列積



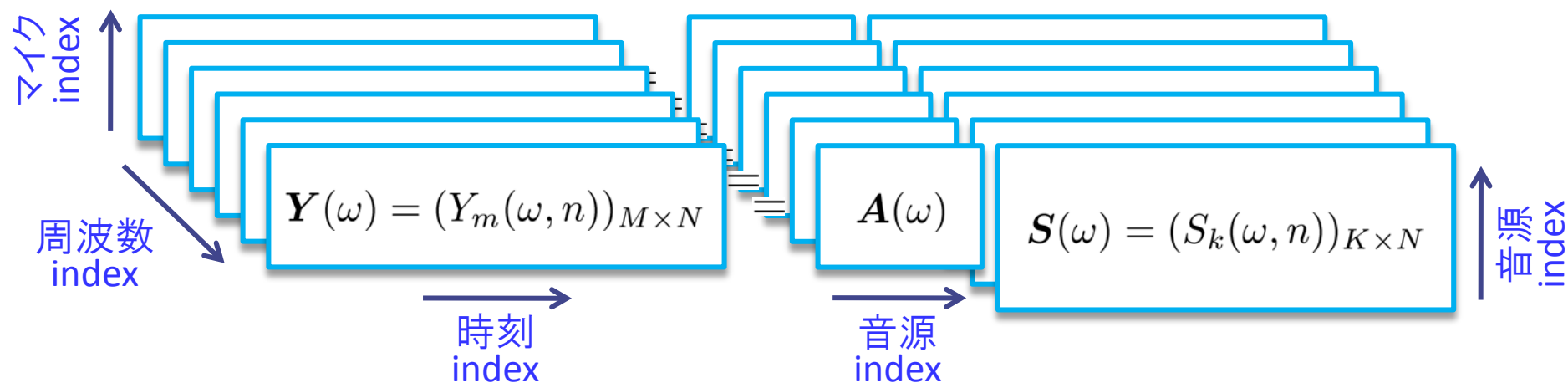
逆問題への手がり



■ 以上の生成プロセスの逆問題(音源分離)は不良設定

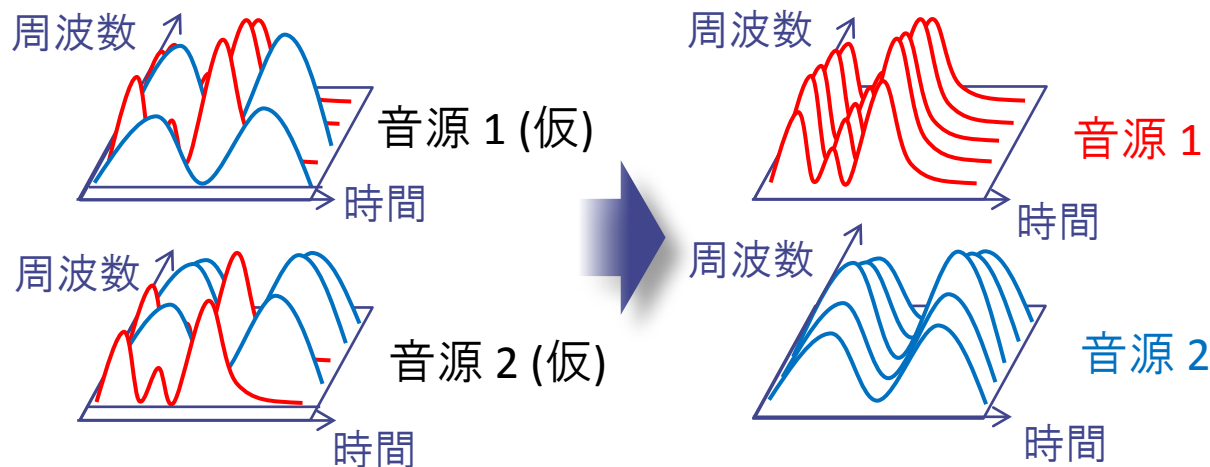
- 音源数 > マイク数 $\longrightarrow Y(\omega) = A(\omega)S(\omega)$ は劣決定
- パーミュテーションの任意性 $Y(\omega) = A(\omega)P^T(\omega)\underline{P(\omega)}S(\omega)$
置換行列

逆問題への手がり



パーミュテーション整合

周波数ごとに
個別に分離が
得られても...



参考資料：BSSの従来研究

■ 時間周波数領域BSSの代表的アプローチ

- 音源数＝マイク数の場合

独立成分分析 (ICA) e.g. [Smaragdis+2003]

音源間の統計的独立性規準 f を最大化するように $\mathbf{A}$ の逆行列 $\mathbf{W}$ を推定

$$\hat{\mathbf{W}}(\omega) = \underset{\mathbf{W}(\omega)}{\operatorname{argmax}} f(\mathbf{W}(\omega)\mathbf{Y}(\omega))$$

- 音源数＞マイク数の場合

時間周波数マスキング e.g. [Yilmaz+2004]

- ・各時間周波数点において単一音源のみが支配的と仮定
- ・各時間周波数点で得られる音源到来方向に関する特徴量をもとに各時間周波数点をクラスタリング

■ パーミュテーション整合のアプローチ

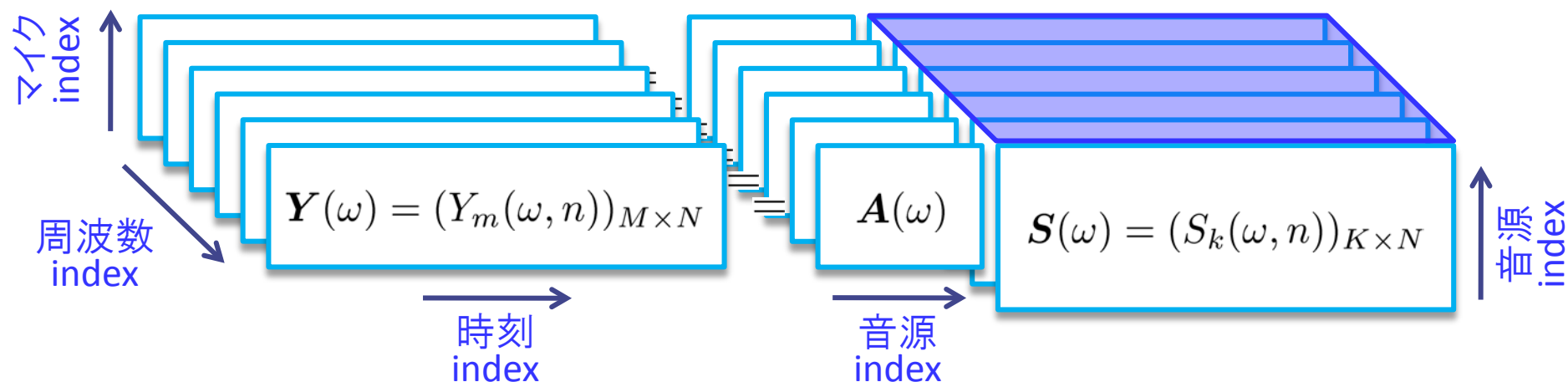
- **音源到来方向** [Kurita+2000], **帯域間の振幅相関** [Murata+2001], **調波性** [Sawada+2004], **デルタ振幅の帯域間の同期性** [Ono+2010]に基づく手法など多数



■ BSSとパーミュテーションの同時解決アプローチ

- マイク数＝音源数の場合 **独立ベクトル分析 (IVA)** [Kim+2006, Hiroe2006]
- 音源数＞マイク数の場合 **多チャンネルNMF** [Sawada+2012]

逆問題への手がかかり



- 以上の生成プロセスの逆問題(音源分離)は不良設定
 - 音源数 > マイク数 $\longrightarrow Y(\omega) = A(\omega)S(\omega)$ は劣決定
 - パーミュテーションの任意性 $Y(\omega) = A(\omega)P^T(\omega)\underline{P(\omega)}S(\omega)$
置換行列
- 解を絞り込むための手がかかりが必要
 - NMF同様, 音源スペクトログラムのrank-1構造を仮定
 - $\Rightarrow$ NMFの多チャンネル拡張

多チャンネルNMF [Sawada+2012]

■ 多チャンネル観測信号の生成モデル化

$$S_k(\omega, n) \sim \mathcal{N}_{\mathbb{C}}(0, H_{\omega,k} U_{k,n})$$

→ 音源 k の複素スペクトログラム

$$Y_m(\omega, n) = \sum_k A_{m,k}(\omega) S_k(\omega, n)$$

→ マイク m の観測複素スペクトログラム

$$\Leftrightarrow \mathbf{y}(\omega, n) = \sum_k \mathbf{a}_k(\omega) S_k(\omega, n)$$



→ $S_k(\omega, n)$ と $S_{k'}(\omega, n)$ の独立性を仮定

$$\mathbf{y}(\omega, n) \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \sum_k H_{\omega,k} U_{k,n} \mathbf{a}_k(\omega) \mathbf{a}_k(\omega)^H\right)$$

→ $p(Y|\theta)$
 $\theta = \{H, U, A\}$

$$\hat{\theta} = \operatorname{argmax}_{\theta} p(Y|\theta)$$

Log-determinant (LD) ダイバージェンス
 (板倉 齋藤 距離の行列版)

$$= \operatorname{argmin}_{H, U, A} \sum_{\omega, n} \left(\operatorname{tr}(\mathbf{y}(\omega, n) \mathbf{y}(\omega, n)^H \mathbf{X}(\omega, n)^{-1}) + \log \det \mathbf{X}(\omega, n) \right)$$

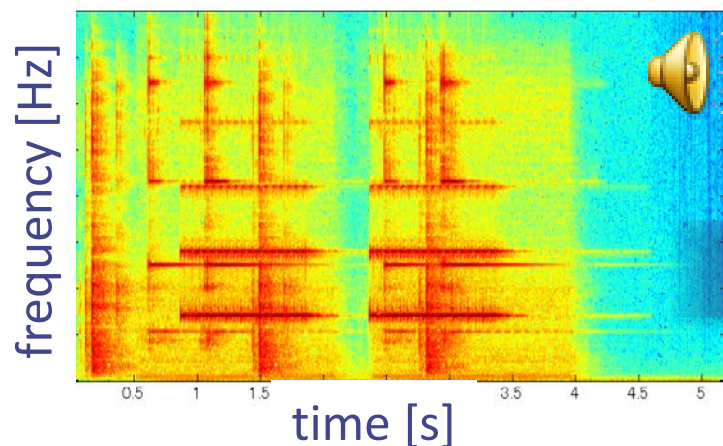
$$\mathbf{X}(\omega, n) = \sum_k H_{\omega,k} U_{k,n} \mathbf{a}_k(\omega) \mathbf{a}_k(\omega)^H$$

多チャンネルFHMMによる音源分離結果例 [Higuchi+2014]

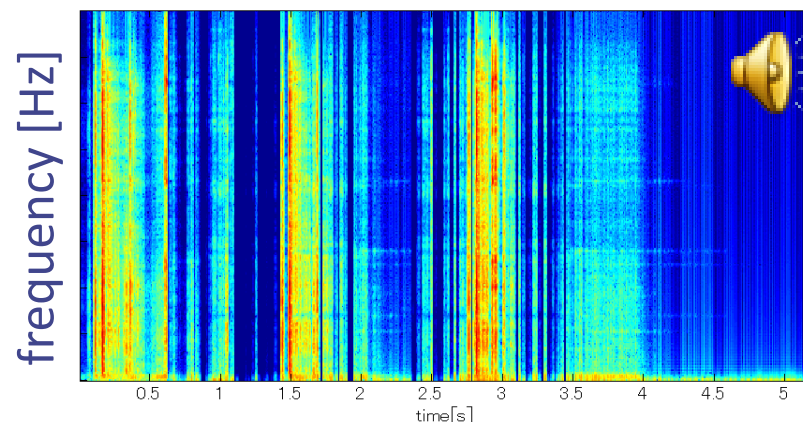
■ 多チャンネルFHMM: 多チャンネルNMFの拡張
(音源分離・残響除去・音響イベント検出の同時解法)

IBISワークショップ
で発表予定

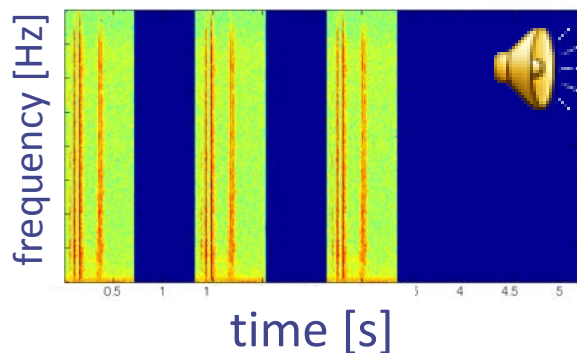
残響下の混合音スペクトログラム



分離音スペクトログラム



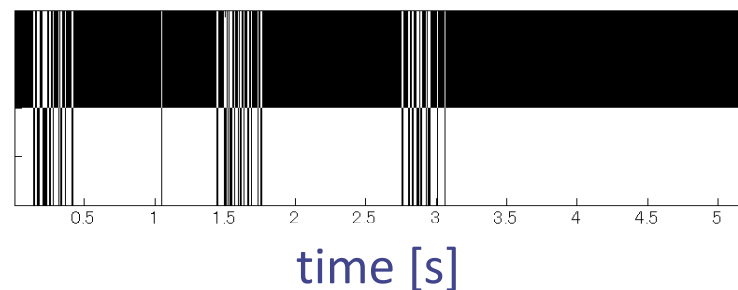
元音源信号のスペクトログラム



音響イベント検出結果

無音状態

有音状態



多チャンネルNMFによる分離結果:



音声・音楽・音響信号に関する諸問題

■ 音声・音楽・音響信号の認識・分析・合成・変換・分離

	認識	分析	合成	変換	分離
音声					
音楽			—	—	
音響					

⇒「識別」、「回帰」、「逆問題」のいずれかに相当

音声・音楽・音響信号に関する諸問題

■ 音声・音楽・音響信号の認識・分析・合成・変換・分離

	認識	分析	合成	変換	分離
音声					
音楽			—	—	
音響					

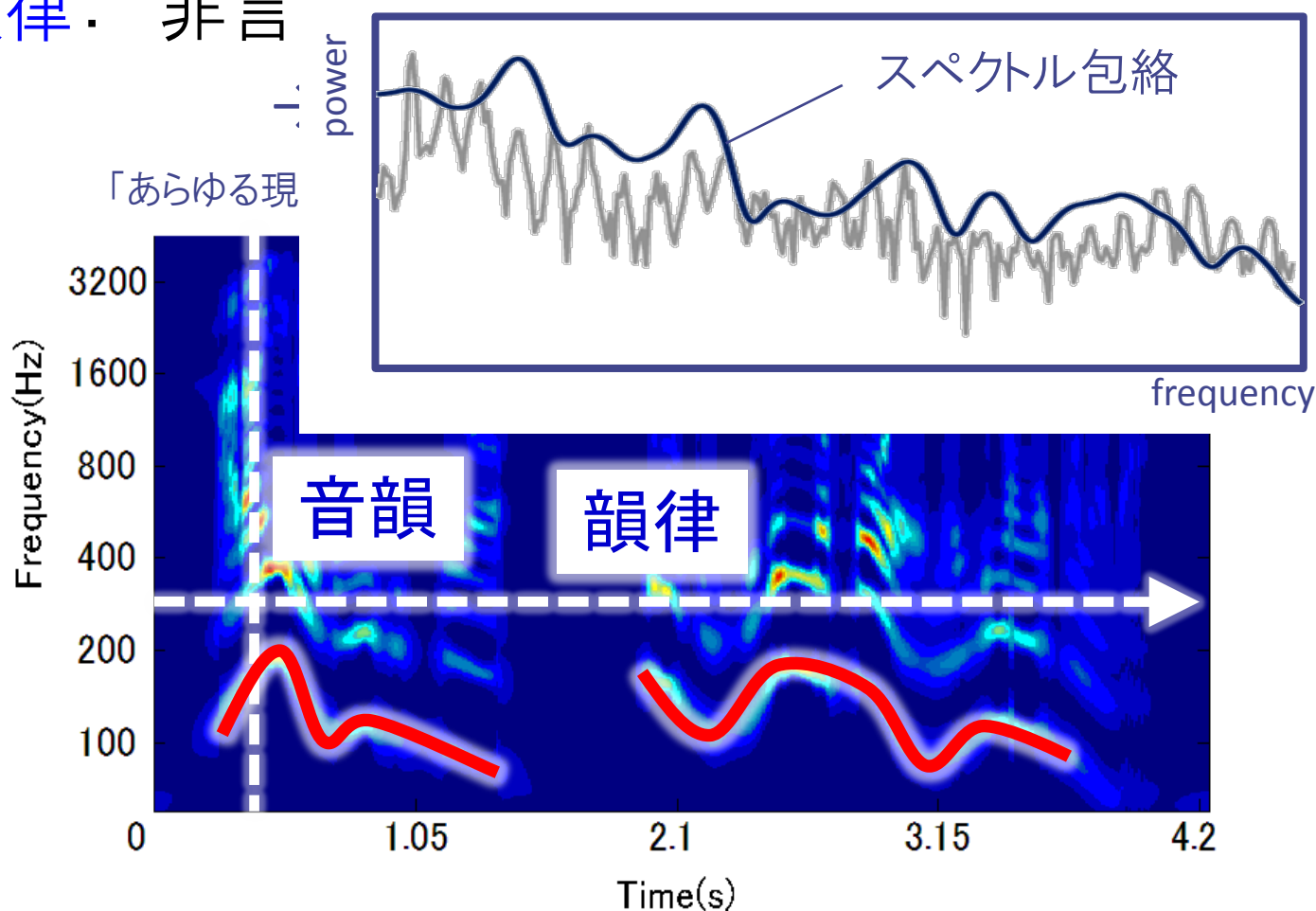
⇒「識別」、「回帰」、「逆問題」のいずれかに相当

音韻と韻律

■ 音韻： 言語情報（文字に書き起こせる情報）

→ 音声の瞬時スペクトルに表れる

■ 韻律： 非言

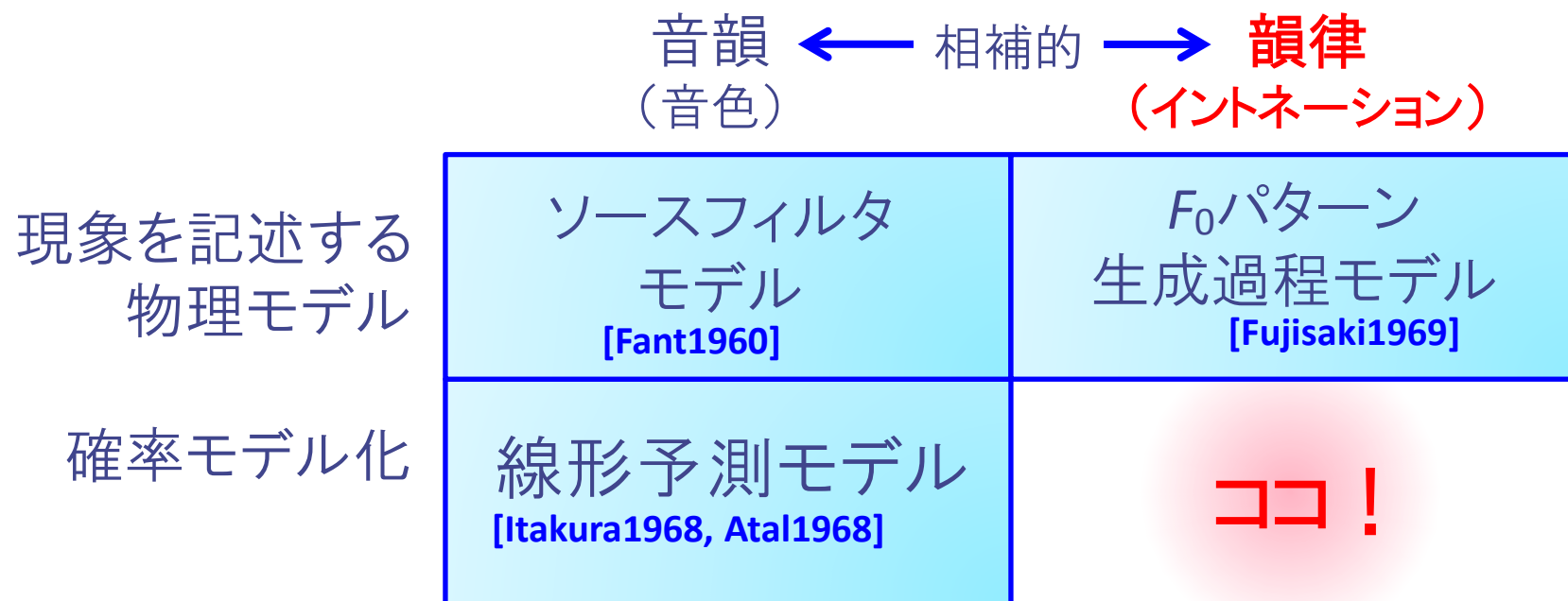


れる

司波数表現)

韻律の確率モデル化

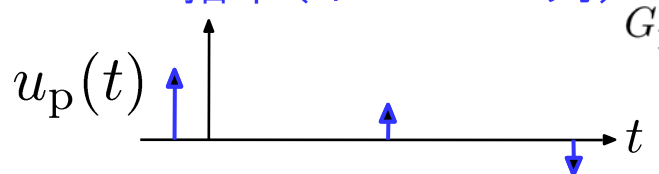
- 韻律は声の表情・表現・個人性などの非言語情報をもつ
- 韻律に着目した音声情報処理を実現するための基盤



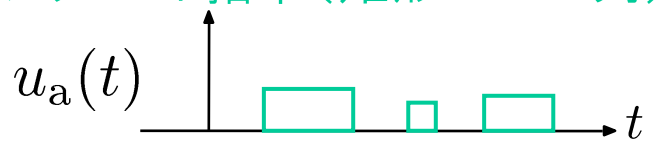
藤崎モデル(F_0 パターンモデル) [Fujisaki+1969]

■ 甲状軟骨による F_0 パターンの制御機構の物理モデル

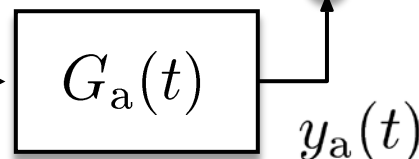
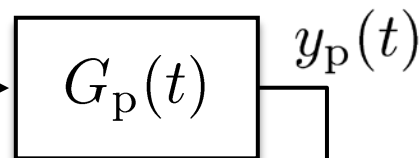
フレーズ指令(インパルス列)



アクセント指令(矩形パルス列)

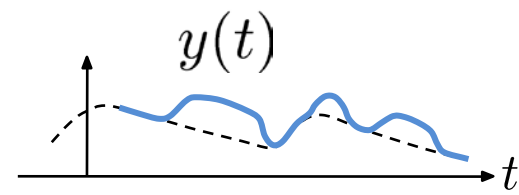


$$G_p(t) = \alpha^2 t e^{-\alpha t} \quad (t \geq 0)$$



$$G_a(t) = \beta^2 t e^{-\beta t} \quad (t \geq 0)$$

y_b



元音声 🗣️

指令列操作後の音声 🗣️

■ 藤崎モデルの指令列推定問題

- 指令列はイントネーション特徴を具現化した物理量
- 重要だが解析的に難しい
(インパルス列・矩形パルス列という拘束条件をどう扱うか)
- そして不良設定(解は無数)

藤崎モデルに基づく F_0 パターンの確率モデル化 [Kameoka2010]

■ 動機:

- 指令列関数(インパルス列・矩形パルス列)をうまく確率モデルで表現し, F_0 パターンを確率モデル化したい → 尤度関数
- 指令列に関する統計的な傾向(音声らしさ)を組み込みたい → 事前分布
- 統計的手法(EM法)を駆使して効率的に指令列を推定したい → 推論アルゴリズムの導出

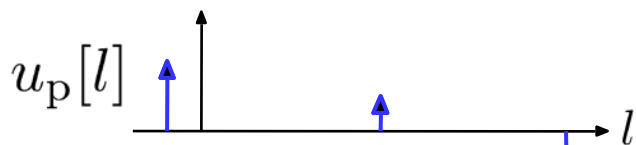
フレーズ・アクセント指令列の生成過程 [Kameoka2010]

■ フレーズ・アクセント指令列の制約

- フレーズ指令はインパルス
- アクセント指令は矩形パルス
- 異なる2つの指令は同時刻に生起しない
- 指令や指令間の間隔の持続長が「音声らしい」ものである

■ フレーズ・アクセント指令列の生成過程の確率モデル化

フレーズ指令(インパルス列)



アクセント指令(矩形パルス列)

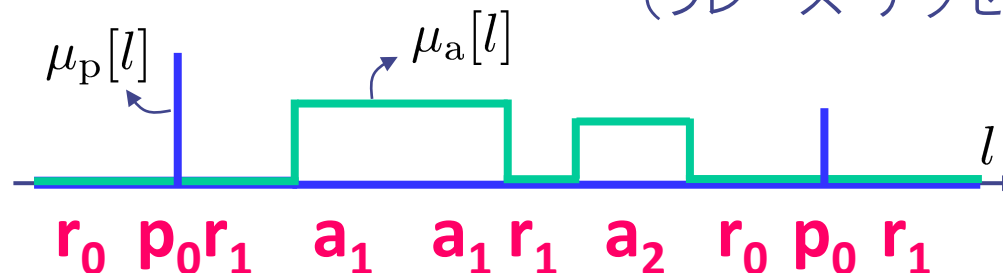


フレーズ・アクセント指令関数は
区分定常な関数と見なせる！
隠れマルコフモデル(HMM)で
表現できる

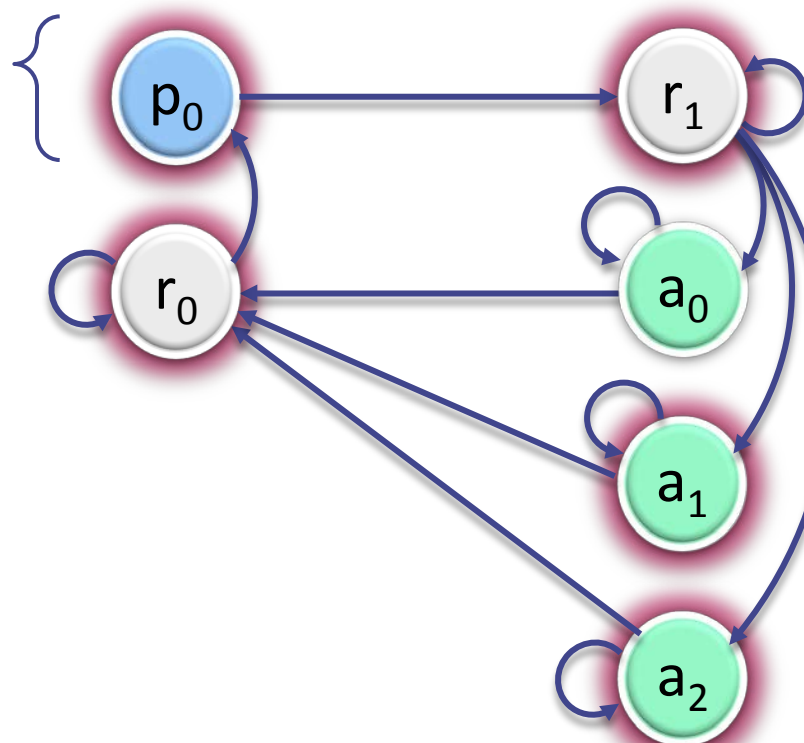
$g_a[l]$

フレーズ・アクセント指令列生成HMM [Kameoka2010]

状態遷移に伴う平均 $\mu_p[l], \mu_a[l]$ の時間軌跡
(フレーズ・アクセント指令列)



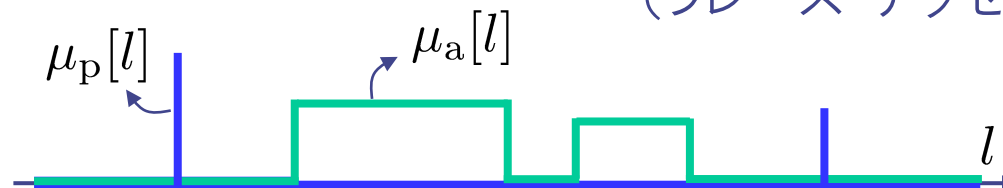
フレーズ指令を
出力する状態



アクセント指令を
出力する状態

フレーズ・アクセント指令列生成HMM [Kameoka2010]

状態遷移に伴う平均 $\mu_p[l], \mu_a[l]$ の時間軌跡
(フレーズ・アクセント指令列)

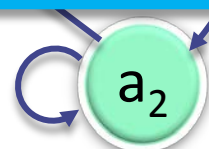


フレーズ指令を
出力する状態



- フレーズ・アクセント指令の長さや、指令間の
間隔は「音声らしい」ものであるはず

HMMだから、こういう先験的知識は
トポロジーや遷移確率(＝事前確率)に反映可能！



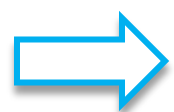
観測 F_0 パターンの確率密度関数 [Kameoka2010]

■ 観測 F_0 パターンは「HMM出力系列の畳み込み混合」

$$\underbrace{y[l]}_{\text{観測 } F_0 \text{ パターン}} = \underbrace{G_p[l]}_{\text{隠れマルコフモデルの出力}} * \underbrace{u_p[l]}_{\text{隠れマルコフモデルの出力}} + G_a[l] * \underbrace{u_a[l]}_{\text{隠れマルコフモデルの出力}} + y_b$$

状態系列
(=指令列)

■ HMMの状態出力分布をガウス分布とすれば...

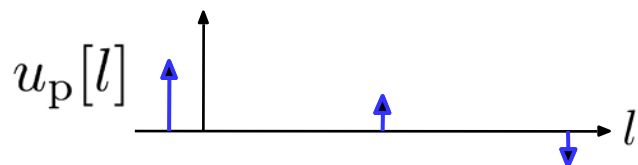


$$\mathbf{y} = \mathbf{G}_p \mathbf{u}_p + \mathbf{G}_a \mathbf{u}_a + y_b \mathbf{1}$$

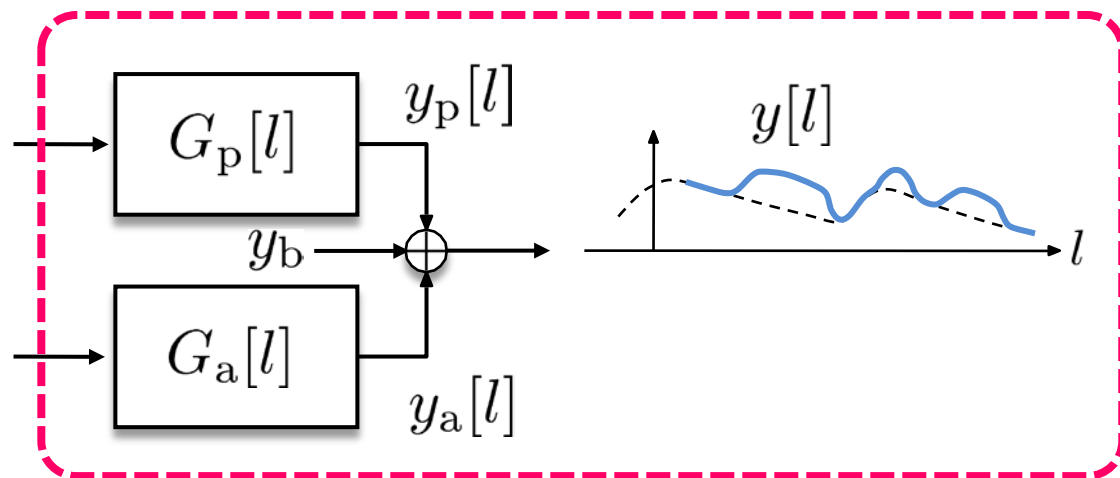
$$\sim \mathcal{N}(\mathbf{G}_p \boldsymbol{\mu}_p + \mathbf{G}_a \boldsymbol{\mu}_a + \mu_b \mathbf{1}, \mathbf{G}_p \boldsymbol{\Sigma}_p \mathbf{G}_p^T + \mathbf{G}_a \boldsymbol{\Sigma}_a \mathbf{G}_a^T + \boldsymbol{\Sigma}_b)$$

$p(\mathbf{y}|\boldsymbol{\theta})$

フレーズ指令(インパルス列)



アクセント指令(矩形パルス列)



アクセント・フレーズ指令列の推定 [Kameoka2010]

- 観測 F_0 パターン $\mathbf{y}$ が与えられた下で

$P(\theta|\mathbf{y}) \propto P(\mathbf{y}|\theta)P(\theta)$ を最大化する状態系列(指令列)

→ 指令列に関する先験的知識がここに反映

- フレーズ成分 $y_p[l]$, アクセント成分 $y_a[l]$, ベースライン成分 $y_b[l]$ を完全データと見なすとEMアルゴリズムが適用できる

$$\mathbf{x} = \begin{bmatrix} \mathbf{y}_p \\ \mathbf{y}_a \\ \mathbf{y}_b \end{bmatrix} \sim \mathcal{N} \left(\underbrace{\begin{bmatrix} \mathbf{G}_p \boldsymbol{\mu}_p \\ \mathbf{G}_a \boldsymbol{\mu}_a \\ \mu_b \mathbf{1} \end{bmatrix}}_{\mathbf{m}}, \underbrace{\begin{bmatrix} \mathbf{G}_p \boldsymbol{\Sigma}_p \mathbf{G}_p^\top & O & O \\ O & \mathbf{G}_a \boldsymbol{\Sigma}_a \mathbf{G}_a^\top & O \\ O & O & \boldsymbol{\Sigma}_b \end{bmatrix}}_{\boldsymbol{\Lambda}} \right)$$

- “不完全データ” $\mathbf{y}$ と完全データの関係: $\mathbf{y} = \overbrace{[\mathbf{I} \ \mathbf{I} \ \mathbf{I}]}^{\mathbf{H}} \begin{bmatrix} \mathbf{y}_p \\ \mathbf{y}_a \\ \mathbf{y}_b \end{bmatrix}$

- Q関数

$$Q(\Theta, \Theta') \stackrel{c}{=} -\frac{1}{2} \left[\text{tr}(\boldsymbol{\Lambda}^{-1} \mathbb{E}[\mathbf{x} \mathbf{x}^\top | \boldsymbol{\Omega}; \Theta]) - 2\mathbf{m}^\top \boldsymbol{\Lambda}^{-1} \mathbb{E}[\mathbf{x} | \boldsymbol{\Omega}; \Theta] + \mathbf{m}^\top \boldsymbol{\Lambda}^{-1} \mathbf{m} \right]$$

EMアルゴリズムによる指令列推定 [Kameoka2010]

■ E-step

藤崎モデル

$$\begin{aligned}\mathbb{E}[x|y; \theta] &= m + \Lambda H^T (H \Lambda H^T)^{-1} (y - \underline{Hm}) \\ \mathbb{E}[xx^T|y; \theta] &= \Lambda - \Lambda H^T (H \Lambda H^T)^{-1} H \Lambda + \mathbb{E}[x|y; \theta] \mathbb{E}[x|y; \theta]^T\end{aligned}$$

■ M-step

$$Q(\Theta, \Theta') \stackrel{c}{=} -\frac{1}{2} \left[\text{tr}(\Lambda^{-1} \mathbb{E}[xx^T | \Omega; \Theta]) - 2m^T \Lambda^{-1} \mathbb{E}[x | \Omega; \Theta] + m^T \Lambda^{-1} m \right]$$

を Θ に関して最大化

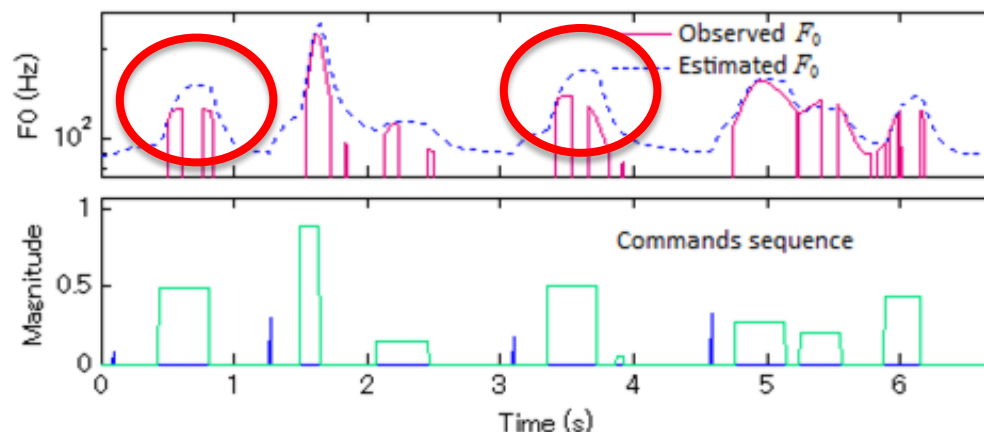
- 状態遷移系列(推定指令列)の更新は Viterbi アルゴリズムで実現可能

指令列の推定結果例 [Yoshizato+2012]

■ ATR No.353 「虫、自動車の排気ガス、台風など、苗木を害するものは多い。」 🗣️

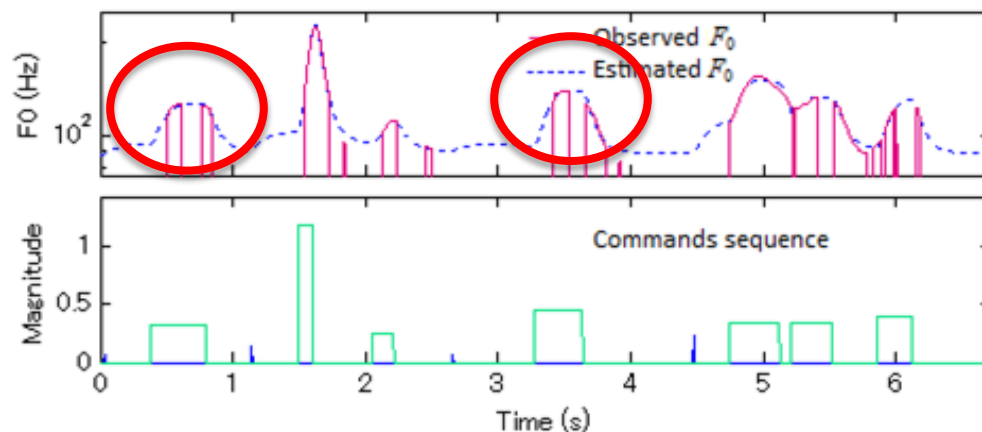
従来法の推定結果

“音声の基本周波数パターン生成過程モデル
のパラメータ自動抽出法” [成澤2002]



提案手法の推定結果

🗣️ (推定指令列から合成した F_0 パターンを
用いて再合成した音声)

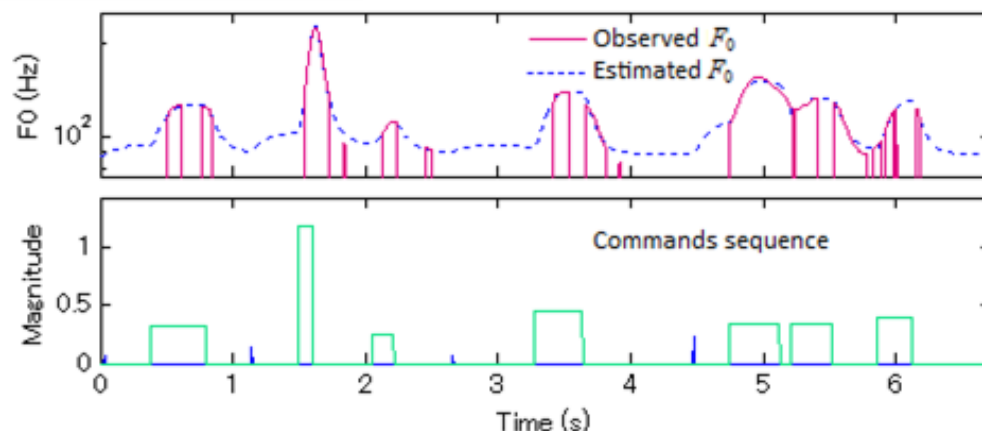


指令列の推定結果例 [Yoshizato+2012]

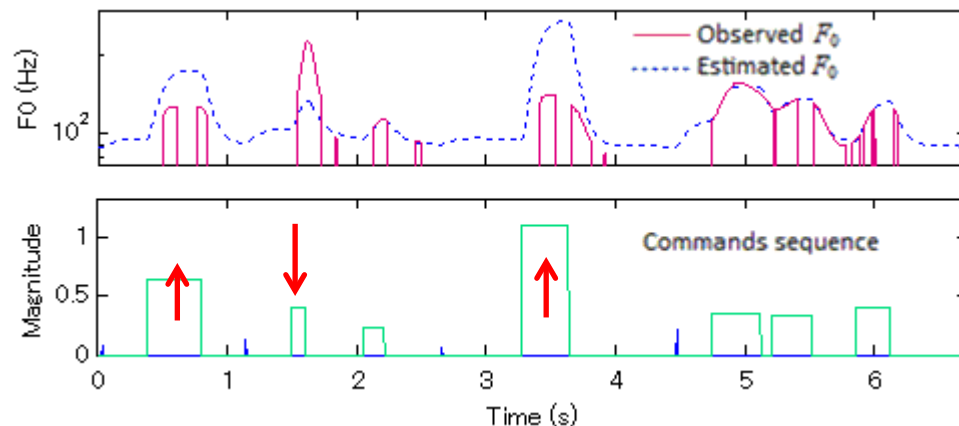
■ ATR No.353 「虫、自動車の排気ガス、台風など、苗木を害するものは多い。」

提案法の推定結果

(推定指令列から合成した F_0 パターンを用いて再合成した音声)



一部のアクセント指令の
振幅を操作



(操作後の指令列から
再合成した音声)

アウトライン

- 音声・音楽・音響信号の認識・分析・合成・変換・分離問題
- 音響信号データの特殊性を意識して設計された様々な確率モデルを紹介
 - 線形予測モデルによる音声分析
 - トラジェクトリHMMによるテキスト音声合成
 - 最尤系列変換による声質変換
 - NMFによる多重音解析
 - 多チャンネルNMFによるブラインド音源分離
 - F_0 パターン生成過程モデルによる韻律分析・合成
- 補助関数法に基づく学習アルゴリズムの導出方法の解説
 - EMアルゴリズムを一般化した最適化手法の原理
 - 補助関数設計クックブック
 - 適用例

アウトライン

- 音声・音楽・音響信号の認識・分析・合成・変換・分離問題
- 音響信号データの特殊性を意識して設計された様々な確率モデルを紹介
 - 線形予測モデルによる音声分析
 - トラジェクトリHMMによるテキスト音声合成
 - 最尤系列変換による声質変換
 - NMFによる多重音解析
 - 多チャンネルNMFによるブラインド音源分離
 - F_0 パターン生成過程モデルによる韻律分析・合成
- 補助関数法に基づく学習アルゴリズムの導出方法の解説
 - EMアルゴリズムを一般化した最適化手法の原理
 - 補助関数設計クックブック
 - 適用例

補助関数法による最適化アルゴリズム

■ 音声信号処理における様々な非線形最適化問題

- 確率モデルの学習, 実測データへのモデルフィッティングなど

■ 最適化問題を難しくする要因の例

- 最適解が解析的に求められない
- 目的関数の不連続性
- 制約つき最適化問題における複雑な制約式の扱い
- 大規模行列の逆行列計算

■ 補助関数法 (a.k.a. Majorization-Minimization) による最適化アルゴリズム

- 目的関数の代わりに「補助関数」を反復的に最大(小)化
- 最適化問題に応じたテイラーメイドなアルゴリズムになるので, 勾配法などの汎用的なアルゴリズムに比べて収束が早い場合が多い

cf) [Kameoka+2006, Kameoka+2007, Kameoka+2008, Kameoka+2008, Nakano+2010, Kameoka+2009, Ono2011, Sawada+2012, Kameoka+2014, Takamune+2014,...]

- 上記の難しさをうまく回避するような補助関数を設計できるかがカギ

補助関数法 (EMアルゴリズムを一般化した原理)

■ 目的関数 $D(\theta)$ を局所最小化する θ を得るための方法論

■ $D(\theta) = \min_{\alpha} G(\theta, \alpha)$ を満たす $G(\theta, \alpha)$ を補助関数と定義

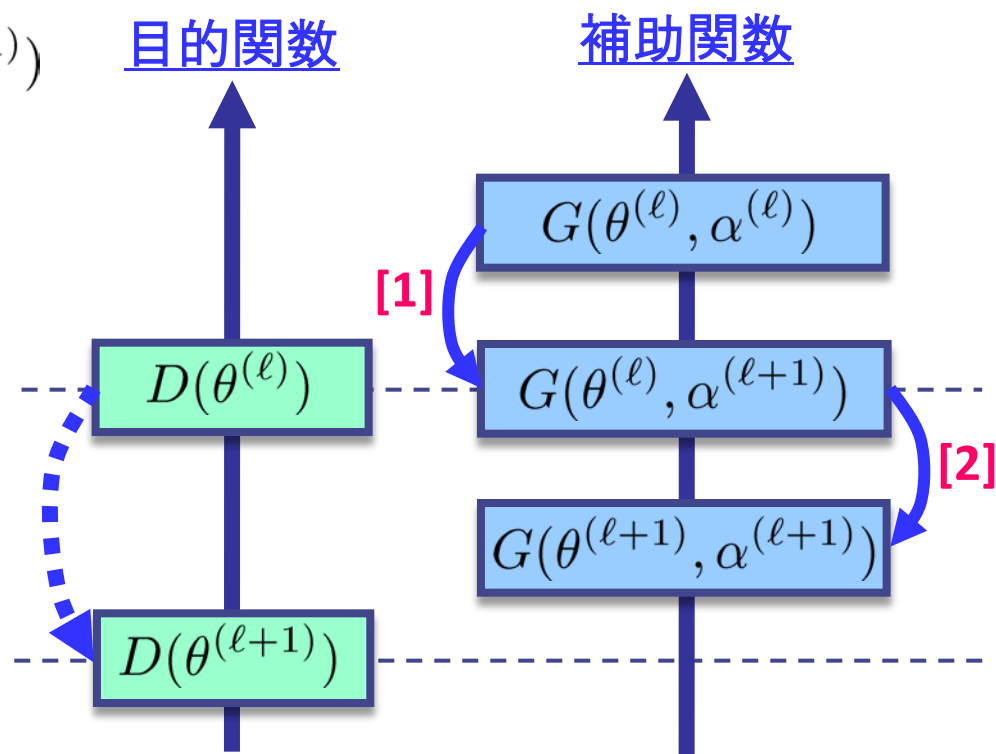
■ パラメータ更新アルゴリズム

[1] $\alpha^{(\ell+1)} = \operatorname{argmin}_{\bar{\alpha}} G(\theta^{(\ell)}, \alpha)$

[2] $\theta^{(\ell+1)} = \operatorname{argmin}_{\theta} G(\theta, \alpha^{(\ell+1)})$

■ 収束性

$$\begin{aligned} D(\theta^{(\ell)}) &= G(\theta^{(\ell)}, \alpha^{(\ell+1)}) \\ &\geq G(\theta^{(\ell+1)}, \alpha^{(\ell+1)}) \\ &\geq D(\theta^{(\ell+1)}) \end{aligned}$$



補助関数法 (EMアルゴリズムを一般化した原理)

■ 目的関数 $D(\theta)$ を局所最小化する θ を得るための方法論

■ $D(\theta) = \min_{\alpha} G(\theta, \alpha)$ を満たす $G(\theta, \alpha)$ を補助関数と定義

■ パラメータ更新アルゴリズム

[1] $\alpha^{(\ell+1)} = \operatorname{argmin}_{\bar{\theta}} G(\theta^{(\ell)}, \alpha)$

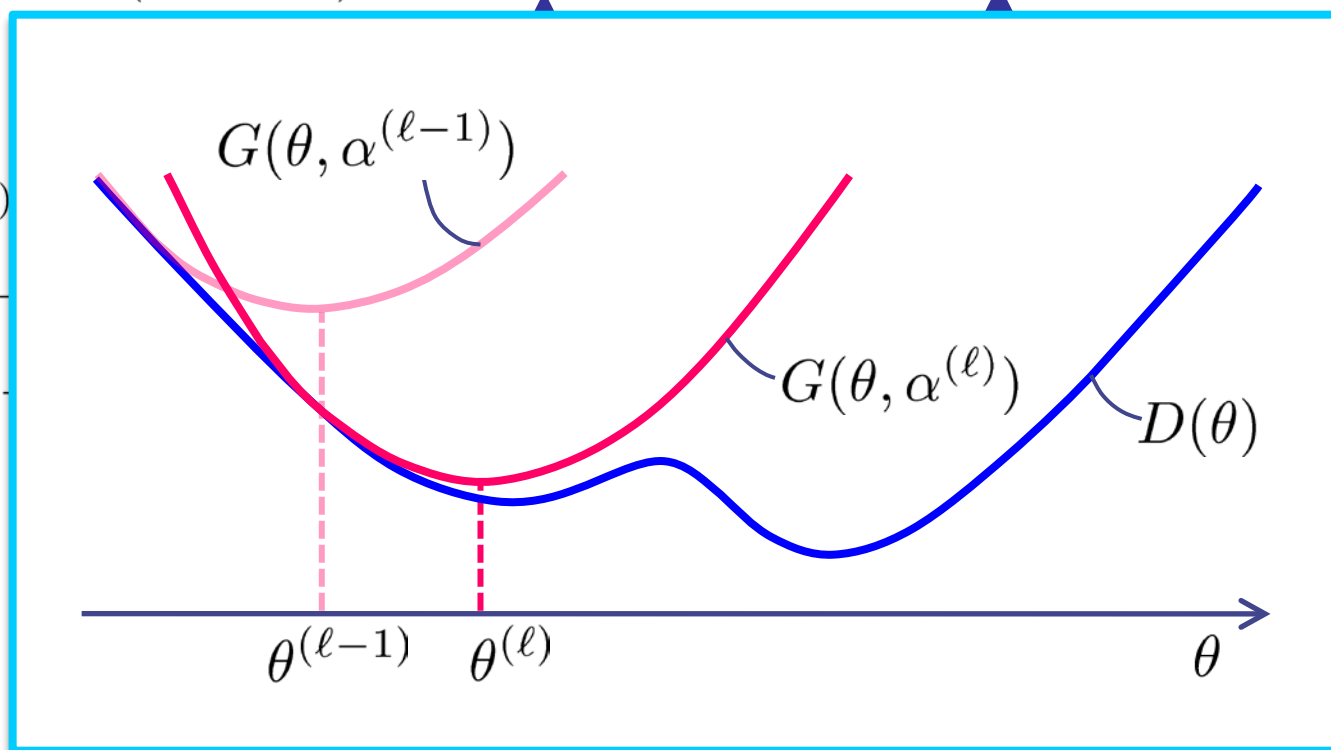
[2] $\theta^{(\ell+1)} = \operatorname{argmin}_{\theta} G(\theta, \alpha^{(\ell+1)})$

目的関数

補助関数

■ 収束性

$$\begin{aligned} D(\theta^{(\ell)}) &= G(\theta^{(\ell)}) \\ &\geq G(\theta^{(\ell-1)}) \\ &\geq D(\theta^{(\ell-1)}) \end{aligned}$$



参考資料: EMアルゴリズム

■ θ に関して $p(Y|\theta)$ を最大化する問題

目的関数

潜在変数

$$\log p(Y|\theta) = \log \int p(Y, X|\theta) dX$$

$$= \log \int \lambda(Y, X) \frac{p(Y, X|\theta)}{\lambda(Y, X)} dX$$

$$\geq \int \lambda(Y, X) \log \frac{p(Y, X|\theta)}{\lambda(Y, X)} dX \quad \leftarrow \text{--- Jensenの不等式}$$

補助関数

Eステップ:

λ に関して補助関数を最大化するステップ

$$\lambda(Y, X) \leftarrow \frac{p(Y, X|\theta)}{\int p(Y, X'|\theta) dX'} \\ \parallel \\ p(X|Y, \theta)$$

Mステップ:

θ に関して補助関数を最大化するステップ

$$\theta \leftarrow \operatorname{argmax}_{\theta} \int \lambda_{Y, X} \log \frac{p(Y, X|\theta)}{\lambda(Y, X)} dX$$

補助関数設計クックブック (1/6)

■ Jensenの不等式 (変数が正の場合)

g : 凸関数,

$$x_i \geq 0,$$

$$\lambda_i \geq 0, \quad \sum_i \lambda_i = 1$$

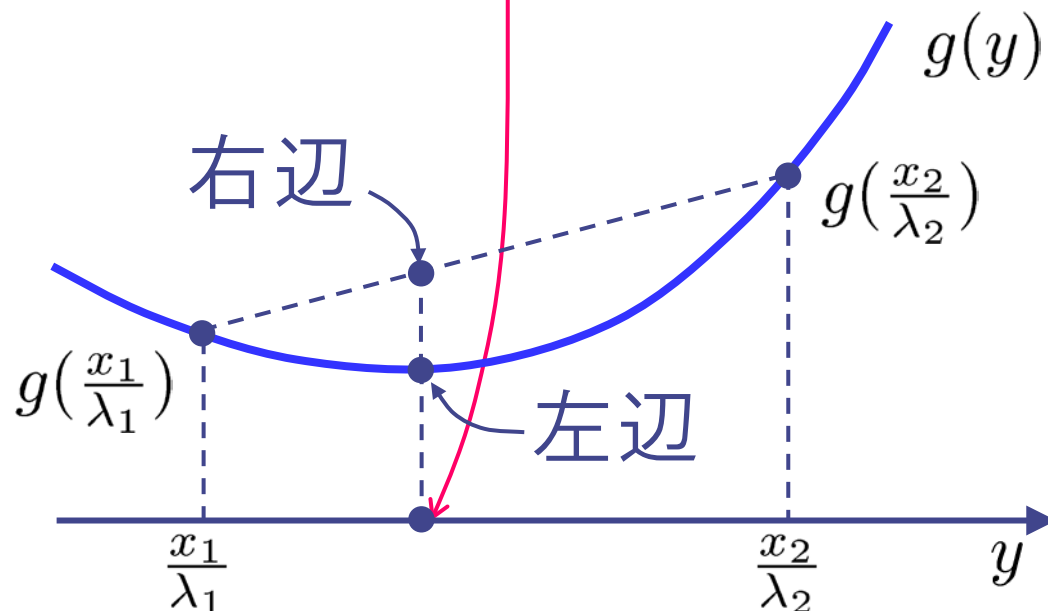
$$g\left(\sum_i x_i\right) \leq \sum_i \lambda_i g\left(\frac{x_i}{\lambda_i}\right)$$

等号成立条件は $\lambda_i = \frac{x_i}{\sum_n x_n}$

例えば,

$g(x) = -\log x$ の場合:

$$\begin{aligned} & -\log\left(\sum_i \lambda_i \frac{x_i}{\lambda_i}\right) \\ & \leq -\sum_i \lambda_i \log \frac{x_i}{\lambda_i} \end{aligned}$$



補助関数設計クックブック (2/6)

■ Jensenの不等式 (変数が実数・複素数の場合)

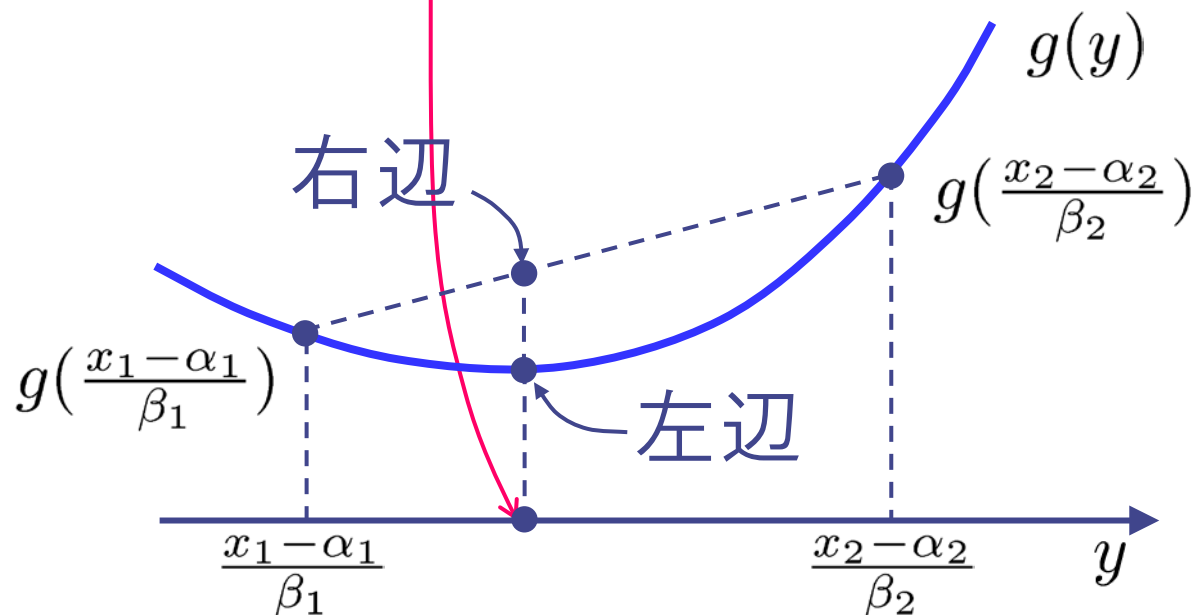
g : 凸関数,

$$\sum_i \alpha_i = 0$$

$$\beta_i \geq 0, \quad \sum_i \beta_i = 1$$

$$g\left(\sum_i x_i\right) \leq \sum_i \beta_i g\left(\frac{x_i - \alpha_i}{\beta_i}\right)$$

等号成立条件は $\alpha_i = x_i - \beta_i \left(\sum_i x_i\right)$



補助関数設計クックブック (3/6)

■ 接線不等式

h : 凹関数

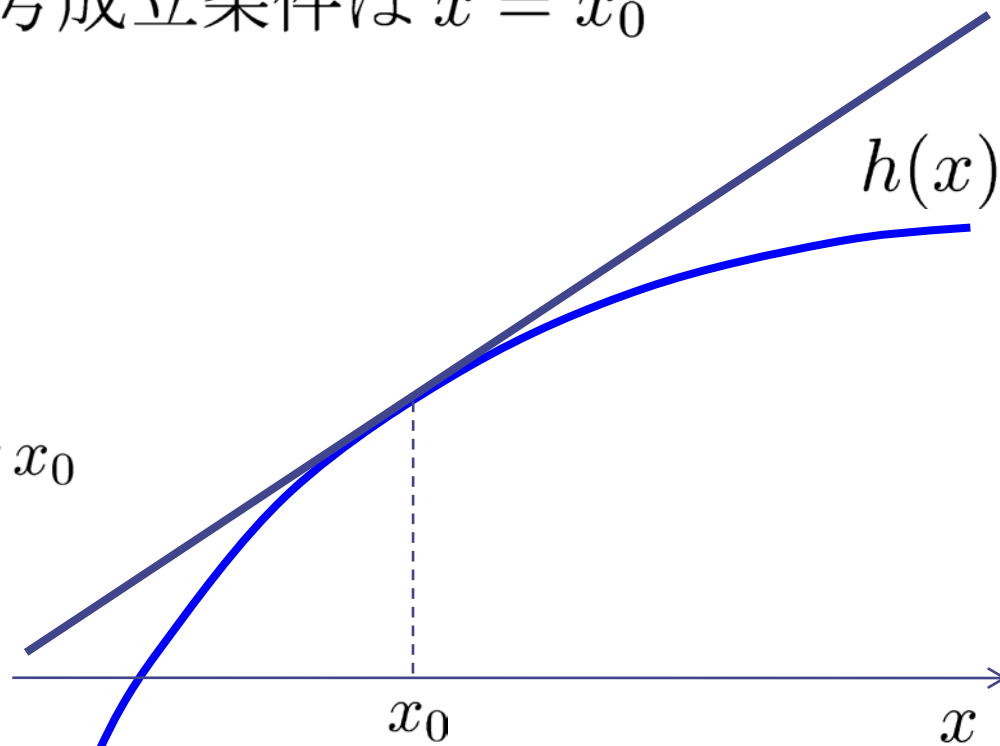
$$h(x) \leq h'(x_0)(x - x_0) + h(x_0)$$

等号成立条件は $x = x_0$

例えば,

$h(x) = \log x$ の場合:

$$\log x \leq \frac{1}{x_0}(x - x_0) + \log x_0$$



補助関数設計クックブック (3/6)

■ 対数関数に対する接線不等式の行列拡張

スカラー変数の場合：

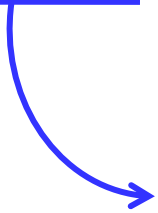
$$\log x \leq \frac{x}{x_0} + \log x_0 - 1$$

等号成立条件は $x = x_0$

行列変数の場合：

$$\underline{\log \det \mathbf{X}} \leq \text{tr}(\mathbf{X}_0^{-1} \mathbf{X}) + \log \det \mathbf{X}_0 - M$$

等号成立条件は $\mathbf{X}_0 = \mathbf{X}$

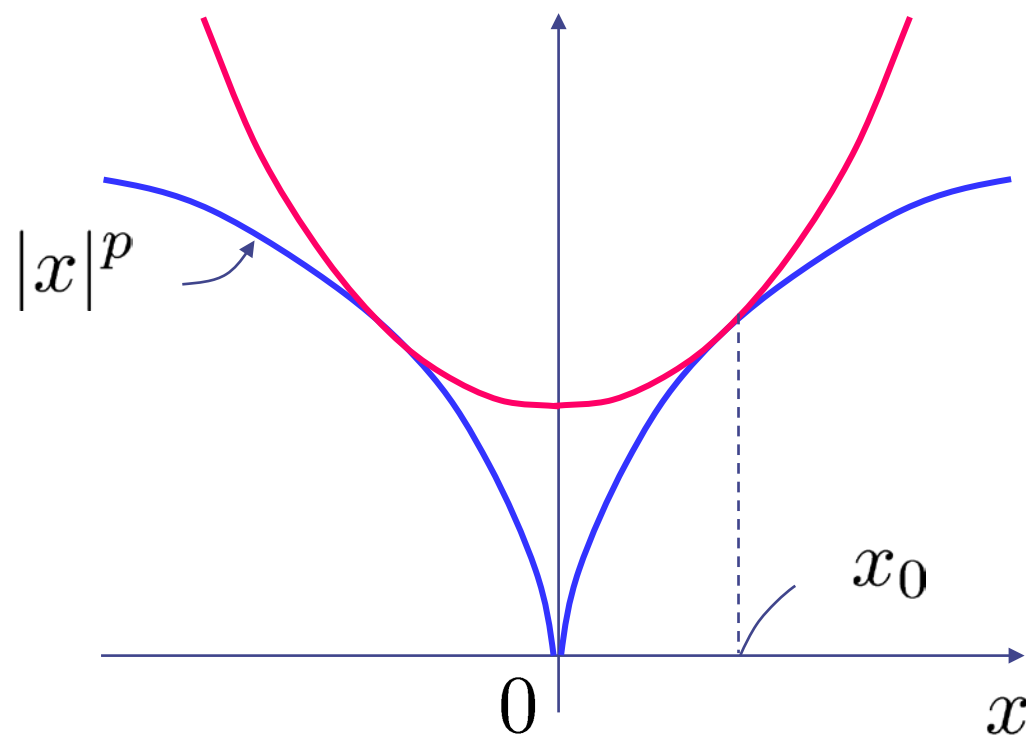
 $\mathbf{X}$ の固有値の対数の和
であることから示せる

補助関数設計クックブック (4/6)

■ 接放物線不等式

$$|x|^p \leq p|x_0|^{p-2}x^2 + 2|x_0|^p - p|x_0|^p$$

等号成立条件は $x = x_0$



補助関数設計クックブック (5/6)

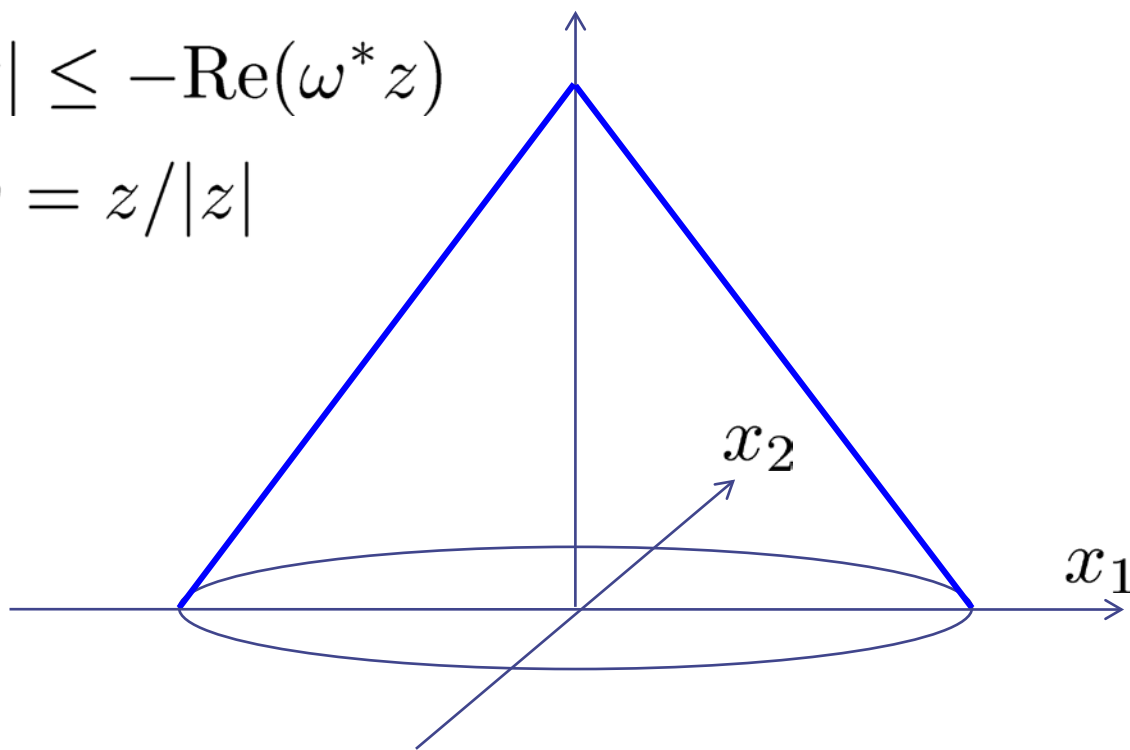
■接平面不等式

$$\|\boldsymbol{w}\|_2 = 1 \longrightarrow -\|\boldsymbol{x}\|_2 \leq -\boldsymbol{w}^T \boldsymbol{x}$$

等号成立条件は $\boldsymbol{w} = \boldsymbol{x} / \|\boldsymbol{x}\|_2$

$$|\omega| = 1 \longrightarrow -|z| \leq -\operatorname{Re}(\omega^* z)$$

等号成立条件は $\omega = z / |z|$



補助関数設計クックブック (6/6)

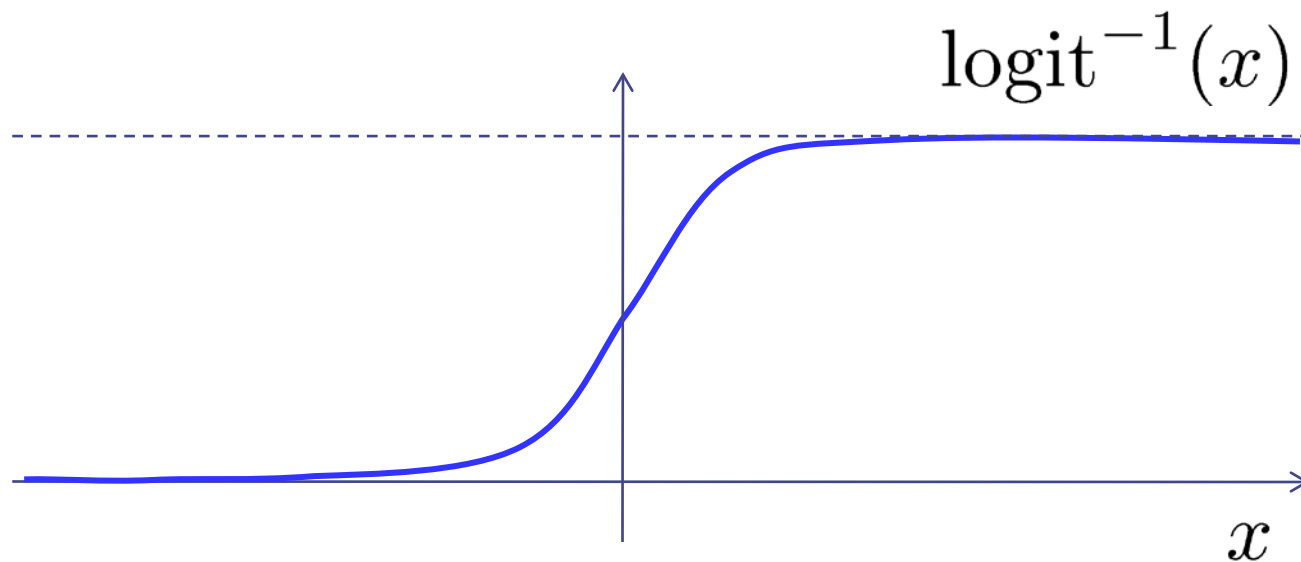
■ Softmax関数

$$\text{logit}^{-1}(x) = \frac{1}{1 + \exp(-x)}$$

$$\text{logit}^{-1}(x) \geq \text{logit}^{-1}(\eta) \exp\left(\frac{x - \eta}{2} - \frac{\tanh(\eta/2)}{4\eta}(x^2 - \eta^2)\right)$$

等号成立条件は $x = \eta$

Gauss分布型の関数



適用例1: 二乗誤差規準のNMFアルゴリズム [Lee+2000]

■ 目的関数

非負値制約がネック

$$\begin{aligned}\|Y - HU\|_F^2 &= \sum_{\omega,t} \left| Y_{\omega,t} - \sum_i H_{\omega,i} U_{i,t} \right|^2 \\ &= \sum_{\omega,t} \left(|Y_{\omega,t}|^2 - 2Y_{\omega,t} \sum_k H_{\omega,k} U_{k,t} + \underbrace{\left| \sum_k H_{\omega,k} U_{k,t} \right|^2}_{\text{non-negative constraint is the bottleneck}} \right)\end{aligned}$$

■ 下線部に対してJensenの不等式(非負値版)を適用

$$\begin{aligned}\left(\sum_i z_i \right)^2 &= \left(\sum_i \lambda_i \frac{z_i}{\lambda_i} \right)^2 \leq \sum_i \lambda_i \left(\frac{z_i}{\lambda_i} \right)^2 = \sum_i \frac{z_i^2}{\lambda_i} \\ &\Rightarrow \left| \sum_k H_{\omega,k} U_{k,t} \right|^2 \leq \sum_k \frac{H_{\omega,k}^2 U_{k,t}^2}{\lambda_{k,\omega,t}}\end{aligned}$$

■ 補助関数

$H_{\omega,k}$ または $U_{k,t}$ ごとの二次関数の和になっている

$$\|Y - HU\|_F^2 \leq \sum_{\omega,t} \left(|Y_{\omega,t}|^2 - 2Y_{\omega,t} \sum_k H_{\omega,k} U_{k,t} + \sum_k \frac{H_{\omega,k}^2 U_{k,t}^2}{\lambda_{k,\omega,t}} \right)$$

⇒パラメータ毎の非負制約付きの二次関数最小化問題に帰着

適用例2: 板倉齋藤NMFアルゴリズム [Kameoka+2006]

■ 目的関数

非負値制約がネック

$$\mathcal{D}_{\text{IS}}(\mathbf{H}, \mathbf{U}) = \sum_{\omega, m} \left(\frac{|Y_{\omega, m}|^2}{\sum_l H_{\omega, l} U_{l, m}} + \log \sum_k H_{\omega, k} U_{k, m} - \dots \right)$$

- 逆数関数は正領域で凸関数 $\Rightarrow$ Jensenの不等式 (非負値版)

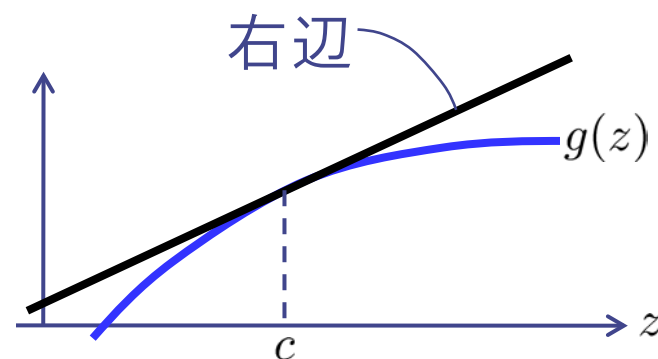
$$\frac{1}{\sum_i z_i} = \frac{1}{\sum_i \lambda_i \frac{z_i}{\lambda_i}} \leq \sum_i \lambda_i \frac{1}{\frac{z_i}{\lambda_i}} = \sum_i \frac{\lambda_i^2}{z_i}$$

$$\Rightarrow \frac{1}{\sum_l H_{\omega, l} U_{l, m}} \leq \sum_l \frac{\lambda_{l, \omega, m}^2}{H_{\omega, l} U_{l, m}}$$

- 対数関数は凹関数 $\Rightarrow$ 接線不等式

$$g(z) \leq g'(c)(z - c) + g(c)$$

$$\Rightarrow \log \sum_k H_{\omega, k} U_{k, t} \leq \frac{1}{c_{\omega, t}} \left(\sum_k H_{\omega, k} U_{k, t} - c_{\omega, t} \right) + \log c_{\omega, t}$$



適用例3: 複素NMFアルゴリズム [Kameoka+2008]

■ 目的関数

$$f(\mathcal{H}, \mathcal{U}, \phi) = \sum_{k,m} \underbrace{|Y_{k,m} - F_{k,m}|^2}_{\text{blue underline}} + 2\lambda \sum_{l,m} \underbrace{|U_{l,m}|^p}_{\text{green underline}}$$

$\nearrow \sum_l H_{k,l} U_{l,m} e^{j\phi_{l,k,m}}$

- 最適解が解析的に解けない
- 目的関数が不連続

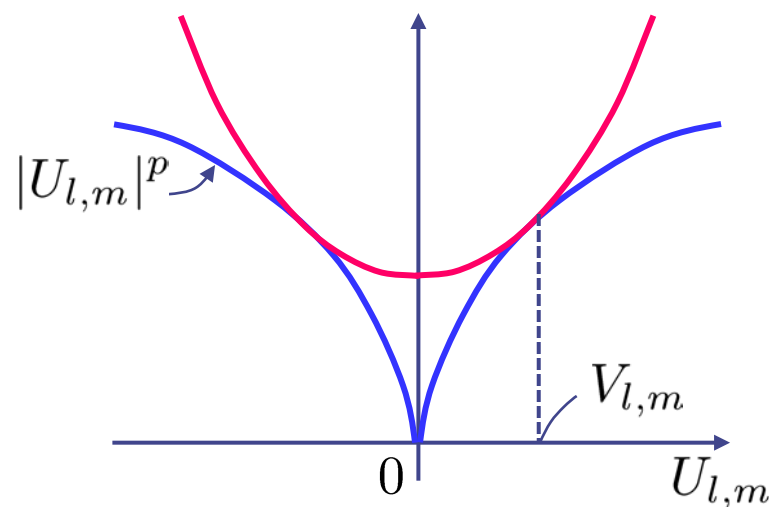
■ 二次関数は凸関数 $\Rightarrow$ Jensenの不等式(複素数版)

$$|Y_{k,m} - F_{k,m}|^2 \leq \sum_l \frac{|X_{l,k,m} - H_{k,l} U_{l,m} e^{j\phi_{l,k,m}}|^2}{\beta_{l,k,m}}$$

$$\text{subject to } \sum_l X_{l,k,m} = Y_{k,m}$$

■ pノルム $\Rightarrow$ 接放物線不等式

$$|U_{l,m}|^p \leq p|V_{l,m}|^{p-2} U_{l,m}^2 + (2-p)|V_{l,m}|^p$$



適用例4: 多チャンネルNMF/FHMMアルゴリズム

[Sawada+2012, Higuchi+2014]

■ 目的関数

$$\mathcal{J}(H, U, A) = \sum_{\omega, n} \left(\underbrace{\text{tr}(\mathbf{y}(\omega, n) \mathbf{y}(\omega, n)^H \mathbf{X}(\omega, n)^{-1})}_{\text{blue}} + \underbrace{\log \det \mathbf{X}(\omega, n)}_{\text{green}} \right)$$

$$\mathbf{X}(\omega, n) = \sum_k H_{\omega, k} U_{k, n} \mathbf{a}_k(\omega) \mathbf{a}_k(\omega)^H$$

※板倉齋藤NMFアルゴリズム [Kameoka+2006] の行列版

■ 逆数関数は正領域で凸関数より

$$\frac{1}{\sum_k x_k} = \frac{1}{\sum_k \frac{r_k}{r_k} x_k} \leq \sum_k \frac{r_k^2}{x_k} \quad (\because \text{変数が正の場合の Jensen 不等式})$$

行列拡張
 $\longrightarrow \text{tr} \left[\left(\sum_k \mathbf{X}_k \right)^{-1} \right] \leq \sum_k \text{tr} [\mathbf{R}_k \mathbf{X}_k^{-1} \mathbf{R}_k]$

■ 対数関数に対する接線不等式の行列版より

$$\log \det \mathbf{X} \leq \text{tr}(\mathbf{X}_0^{-1} \mathbf{X}) + \log \det \mathbf{X}_0 - M$$

IBISワークショップで
発表予定なので、
詳しくはポスターで！

適用例5: 制約付きボルツマンマシン (RBM) の学習

[Kameoka+2014, Takamune+2014]

■ RBM: ニューラルネットワークモデルの一種

- 識別・回帰・教師なし特徴量抽出器として有用
- 深層学習のpre-trainingとして有効

■ 結合分布

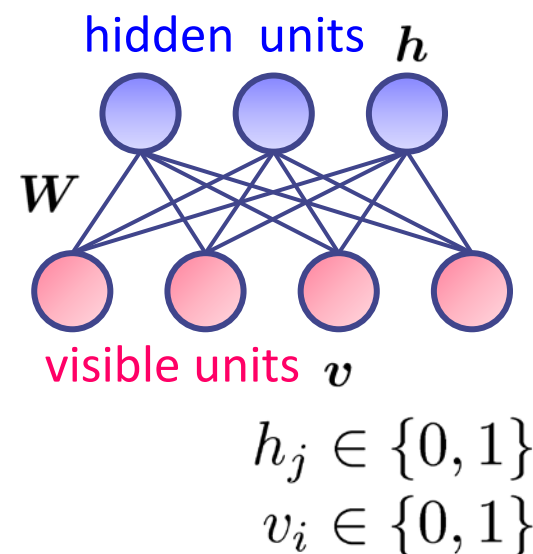
$$p(\mathbf{v}, \mathbf{h} | \Theta) = \frac{\exp(-E(\mathbf{v}, \mathbf{h}; \Theta))}{Z(\Theta)}$$

$$\begin{cases} E(\mathbf{v}, \mathbf{h}; \Theta) = - \sum_i \underline{a_i} v_i - \sum_j \underline{b_j} h_j - \sum_{i,j} v_i \underline{w_{i,j}} h_j \\ Z(\Theta) = \sum_{\mathbf{v}} \sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}; \Theta)) \end{cases}$$

■ 対数尤度

学習サンプル: $\mathbf{v}^{(1)}, \dots, \mathbf{v}^{(N)}$

$$\mathcal{L}(\Theta) = \frac{1}{N} \sum_n \log p(\mathbf{v}^{(n)} | \Theta) = \frac{1}{N} \sum_n \log \sum_{\mathbf{h}} \underline{p(\mathbf{v}^{(n)}, \mathbf{h} | \Theta)}$$



IBISワークショップで
発表予定なので、
詳しくはポスターで！

アウトライン

- 音声・音楽・音響信号の認識・分析・合成・変換・分離問題
- 音響信号データの特殊性を意識して設計された様々な確率モデルを紹介
 - 線形予測モデルによる音声分析
 - トラジェクトリHMMによるテキスト音声合成
 - 最尤系列変換による声質変換
 - NMFによる多重音解析
 - 多チャンネルNMFによるブラインド音源分離
 - F_0 パターン生成過程モデルによる韻律分析・合成
- 補助関数法に基づく学習アルゴリズムの導出方法の解説
 - EMアルゴリズムを一般化した最適化手法の原理
 - 補助関数設計クックブック
 - 適用例

謝辞

- 本資料で紹介した一部の研究は、東京大学・NTT・NIIの共同研究で進展させたものです。

Special thanks to

中野允裕氏, 樋口卓哉氏, 北条伸克氏, 吉里幸太氏

- 本資料の準備の際, 小山翔一氏(東大)と戸田智基氏(NAIST)からスライドをご提供いただき, 参考にさせていただきました。

参考文献 (1/3)

- [1] J. Yamagishi, C. Veaux, S. King and S. Renals, “Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction,” In *Acoustical Science and Technology*, vol. 33, no. 1, pp. 1-5, 2012.
- [2] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, “Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis,” in *Proc. Sixth European Conference on Speech Communication and Technology (Eurospeech 1999)*, pp. 2347–2350, 1999.
- [3] T. Toda, K. Nakamura, H. Sekimoto and K. Shikano, “Voice conversion for various types of body transmitted speech,” *Proc. in Proc. 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009)*, pp. 3601–3604, Apr. 2009.
- [4] K. Nakamura, T. Toda, H. Saruwatari and K. Shikano, “Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech,” *Speech Communication*, vol. 54, no. 1, pp. 134–146, 2012.
- [5] H. Doi, T. Toda, K. Nakamura, H. Saruwatari and K. Shikano, “Alaryngeal speech enhancement based on one-to-many eigenvoice conversion ,” *IEEE Trans. Audio, Speech and Language Process.*, vol. 22, no. 1, pp. 172–183, 2014.
- [6] P. Jax and P. Vary, “On artificial bandwidth extension of telephone speech,” *Signal Processing*, vol. 83, pp. 1707–1719, 2003.
- [7] Y. Stylianou, O. Cappe and E. Moulines, “Continuous probabilistic transform for voice conversion,” *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 131–142, 1998.
- [8] 河原, “音声分析合成技術の動向,” *日本音響学会誌*, Vol. 67, No. 1, pp. 40-45, 2011.
- [9] 小山, “音場再現技術における数値問題: 波面合成・高次アンビソニックスの数値,” *日本音響学会誌*, vol. 68, no. 11, pp. 584-589, 2012.
- [10] F. Itakura and S. Saito, “Analysis synthesis telephony based on the maximum likelihood method,” *Proc. ICA’68*, pp. C-17–C-20, 1968.

参考文献 (2/3)

- [11] G. Fant, *Acoustic Theory of Speech Production*, The Hague, Netherlands: Mouton. (The Old Bible.)
- [12] H. Zen, K. Tokuda, and T. Kitamura, "An introduction of trajectory model into HMM-based speech synthesis," in Proc. of 5th ISCA Speech Synthesis Workshop, pp. 191–196, 2004.
- [13] <http://hts.sp.nitech.ac.jp/>
- [14] M. Abe, S. Nakamura, K. Shikano and H. Kuwabara, "Voice conversion through vector quantization," J. Acoust. Soc. Jpn. (E), vol. 11, no. 2, pp. 71–76, 1990.
- [15] T. Toda, A.W. Black and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," IEEE Trans. Audio, Speech and Language Process., vol. 15, no. 8, pp. 2222–2235, 2007.
- [16] H. Doi, K. Nakamura, T. Toda, H. Saruwatari and K. Shikano, "Esophageal speech enhancement based on statistical voice conversion with Gaussian mixture models," IEICE Trans. on Inf. and Syst., vol. E93-D, no. 9, pp. 2472–2482, 2010.
- [17] 中村, 戸田, 猿渡, 鹿野, "肉伝導人工音声の変換に基づく喉頭全摘出者のための音声コミュニケーション支援システム," 電子情報通信学会論文誌, vol. J90-D, no. 3, pp. 780–787, 2007.
- [18] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for music transcription," in Proc. 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2003), pp. 177–180, 2003.
- [19] P. Smaragdis, B. Raj and M.V. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in Proc. the 7th International Conference on Independent Component Analysis and Blind Source Separation (ICA 2007), pp. 414–421, 2007.
- [20] 亀岡, 小野, 柏野, 嵯峨山, "複素NMF: 新しいスパース信号分解表現と基底系学習アルゴリズム," 日本音響学会2008 年秋季研究発表会講演論文集, 2-8-13, pp. 657–660, 2008.

参考文献 (3/6)

- [21] H. Kameoka, N. Ono, K. Kashino and S. Sagayama, "Complex NMF: A new sparse representation for acoustic signals," in Proc. 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2009), pp. 3437-3440, Apr. 2009.
- [22] C. Févotte, N. Bertin and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," Neural Computation, vol. 21, no. 3, pp. 793–830, 2009.
- [23] H. Kameoka and K. Kashino, "Composite autoregressive system for sparse source-filter representation of speech," in Proc. 2009 IEEE International Symposium on Circuits and Systems (ISCAS2009), pp. 2477-2480, 2009.
- [24] H. Kameoka, T. Yoshioka, M. Hamamura, J. Le Roux and K. Kashino, "Statistical model of speech signals based on composite autoregressive system with application to blind source separation," in Proc. 9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2010), LNCS 6365, pp. 245-253, 2010.
- [25] K. Yoshii and M. Goto, "Infinite composite autoregressive models for music signal analysis," The 13th International Society for Music Information Retrieval Conference (ISMIR), pp. 79–84, 2012.
- [26] M. Nakano, J. Le Roux, H. Kameoka, Y. Kitano, N. Ono and S. Sagayama, "Nonnegative matrix factorization with Markov-chained bases for modeling time-varying patterns in music spectrograms," in Proc. 9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2010), LNCS 6365, pp. 149-156, 2010.
- [27] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," Neural Computation, vol. 22, pp. 21–34, 1998.
- [28] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," IEEE Trans. Signal Process., vol.52, no.7, pp.1830–1847, 2004.

参考文献 (4/6)

- [29] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda and F. Itakura, “Evaluation of blind signal separation method using directivity pattern under reverberant conditions,” in Proc. 2000 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2000), pp. 3140–3143, 2000.
- [30] N. Murata, S. Ikeda and A. Ziehe, “An approach to blind source separation based on temporal structure of speech signals,” *Neurocomputing*, vol. 41, pp. 1-4, 2001.
- [31] H. Sawada, R. Mukai, S. Araki and S. Makino, “A robust and precise method for solving the permutation problem of frequency-domain blind source separation,” *IEEE Trans. on Speech and Audio Processing*, vol. 12, no. 5, pp. 530-538, 2004.
- [32] 小野, “周波数領域ICAにおける Δ スペクトログラムに基づくパーミュテーション解法,” 日本音響学会秋季研究発表会講演集, pp.581-582, Sep. 2010.
- [33] T. Kim, T. Eltoft and T. Lee, “Independent vector analysis: An extension of ICA to multivariate components,” in Proc. 6th International Conference on Independent Component Analysis and Blind Source Separation (ICA 2006), pp. 165–172, 2006.
- [34] A. Hiroe, “Solution of permutation problem in frequency domain ICA, using multivariate probability density functions,” in Proc. 6th International Conference on Independent Component Analysis and Blind Source Separation (ICA 2006), pp. 601–608, 2006.
- [35] H. Sawada, H. Kameoka, S. Araki and N. Ueda, “Efficient algorithms for multichannel extensions of Itakura-Saito nonnegative matrix factorization,” in Proc. 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012), pp. 261–264, 2012.
- [36] T. Higuchi, H. Takeda, T. Nakamura and H. Kameoka, “A unified approach for underdetermined blind signal separation and source activity detection by multichannel factorial hidden Markov models,” in Proc. The 15th Annual Conference of the International Speech Communication Association (Interspeech 2014), pp. 850-854, Sep. 2014.

参考文献 (5/6)

- [37] T. Higuchi and H. Kameoka, “Joint audio source separation and dereverberation based on multichannel factorial hidden Markov model,” in Proc. The 24th IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2014), Sep. 2014.
- [38] H. Fujisaki and S. Nagashima, “A model for synthesis of pitch contours of connected speech,” Annual Report of Engineering Research Institute, University of Tokyo, Vol. 28, pp. 53–60, 1969.
- [39] H. Kameoka, J. Le Roux and Y. Ohishi, “A statistical model of speech F0 contours,” ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition (SAPA 2010), pp. 43-48, 2010.
- [40] K. Yoshizato, H. Kameoka, D. Saito and S. Sagayama, “Hidden Markov convolutive mixture model for pitch contour analysis of speech,” in Proc. The 13th Annual Conference of the International Speech Communication Association (Interspeech 2012), Mon.O2d.06, Sep. 2012.
- [41] 亀岡, 後藤, 嵯峨山, “スペクトル制御エンベロープによる混合音中の周期および非周期成分の選択的イコライザ,” 情報処理学会研究報告, 2006-MUS-66-13, pp. 77-84, Aug. 2006.
- [42] H. Kameoka, *Statistical Approach to Multipitch Analysis*, Ph.D. thesis, The University of Tokyo, 2007.
- [43] H. Kameoka, N. Ono, and S. Sagayama, “Auxiliary function approach to parameter estimation of constrained sinusoidal model for monaural speech separation,” in 2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2008), pp. 29–32, 2008.
- [44] 亀岡, 鎌本, 原田, 守谷, “予測誤差のGolomb-Rice符号量を最小化する線形予測分析,” 電子情報通信学会論文誌, Vol. J91-A, No. 11, pp. 1017-1025, Nov. 2008.
- [45] H. Kameoka, T. Nakatani, and T. Yoshioka, “Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms,” in 2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2009), pp. 45–48, 2009.

参考文献 (6/6)

- [46] N. Ono, “Stable and fast update rules for independent vector analysis based on auxiliary function technique,” in *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA2011)*, pp. 261–264, 2011.
- [47] N. Yasuraoka, H. Kameoka, T. Yoshioka, and H. G. Okuno, “ I -divergence-based dereverberation method with auxiliary function approach,” in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2011)*, pp. 369–372, 2011.
- [48] H. Kameoka and N. Takamune, “Training restricted Boltzmann machines with auxiliary function approach,” in Proc. The 24th IEEE International Workshop on Machine Learning for Signal Processing (MLSP2014), Sep. 2014.
- [49] N. Takamune and H. Kameoka, “Maximum reconstruction probability training of restricted Boltzmann machines with auxiliary function approach,” in Proc. The 24th IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2014), Sep. 2014.