東京大学
THE UNIVERSITY OF TOKYO

NTT

奈良先端科学技術大学院大学　ゼミナール (2012/12/18)

生成モデルアプローチによる音声音響信号処理

亀岡 弘和

東京大学 大学院情報理工学系研究科
NTT コミュニケーション科学基礎研究所

kameoka@hil.t.u-tokyo.ac.jp / kameoka.hirokazu@lab.ntt.co.jp

自己紹介

- 亀岡弘和

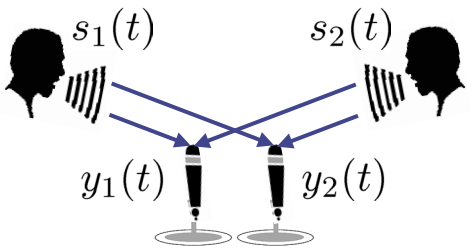
  情報理工学博士
  東京大学大学院情報理工学系研究科　客員准教授
  NTTコミュニケーション科学基礎研究所　研究主任

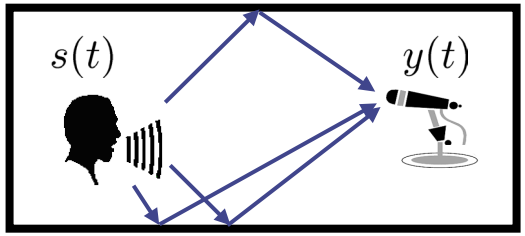
  - 略歴：
    2002.03　東京大学工学部計数工学科卒業
    2004.03　東京大学大学院情報理工学系研究科
    システム情報学専攻修士課程修了.
    2007.03　東京大学大学院情報理工学系研究科
    システム情報学専攻博士課程修了.
    2007.04　日本電信電話株式会社入社.
    NTTコミュニケーション科学基礎研究所配属.
    2011.05　東京大学大学院情報理工学系研究科
    客員准教授に着任.

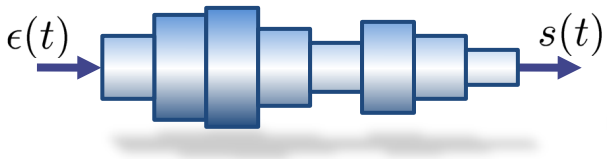
  - 興味：音声や音楽などの音響信号を認識･解析したり合成･生成したりするための統計的信号処理･機械学習の研究

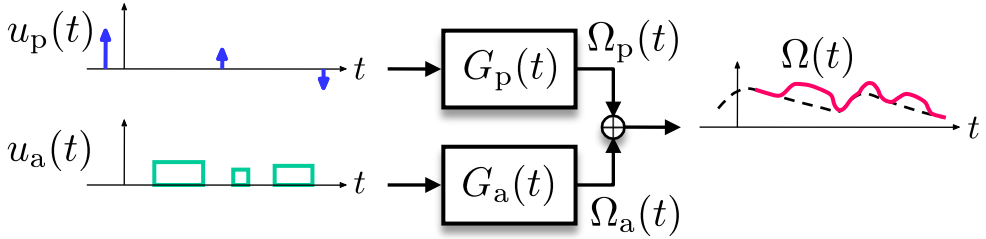
音声音響信号処理問題の多くは逆問題

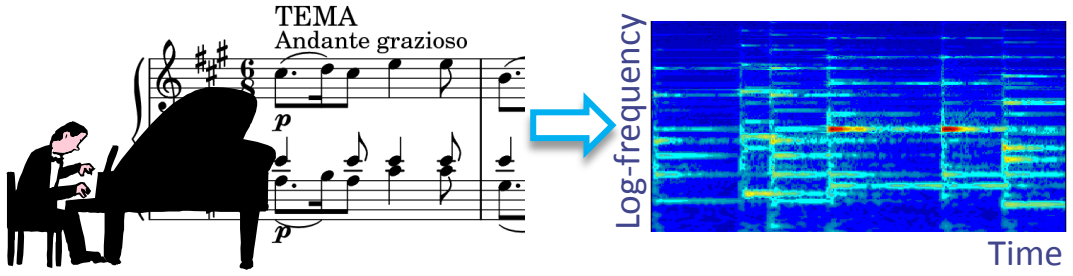
- 音響信号処理
  - ブラインド音源分離
  - 残響除去
  - ・・・
- 音声情報処理
  - 音素特徴抽出
  - 音声認識
  - イントネーション解析
- 音楽情報処理
  - 多重音解析
  - 自動採譜
  - ・・・





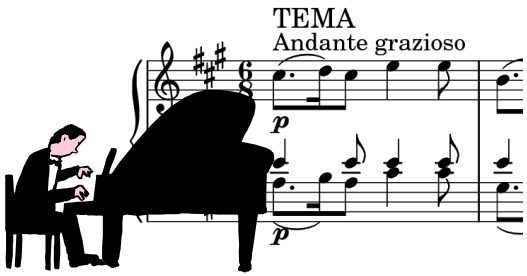


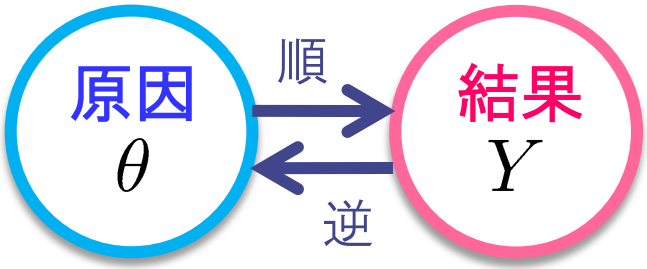


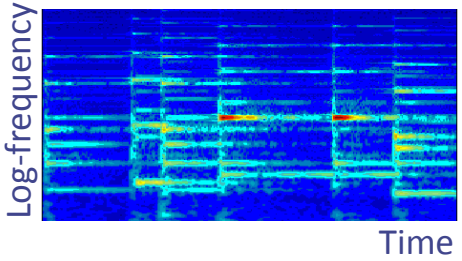


逆問題へのアプローチ

- 逆問題：『「結果」からその「原因」を推定する問題』







（例）
- 音源信号＆混合過程 →混合信号
- 発話内容 →音素特徴系列
- 楽譜＆演奏表情 →音楽音響信号

  - 多くは数理的に不良設定
  - ･･･のはずなのに人間には難なく出来ていることばかり！

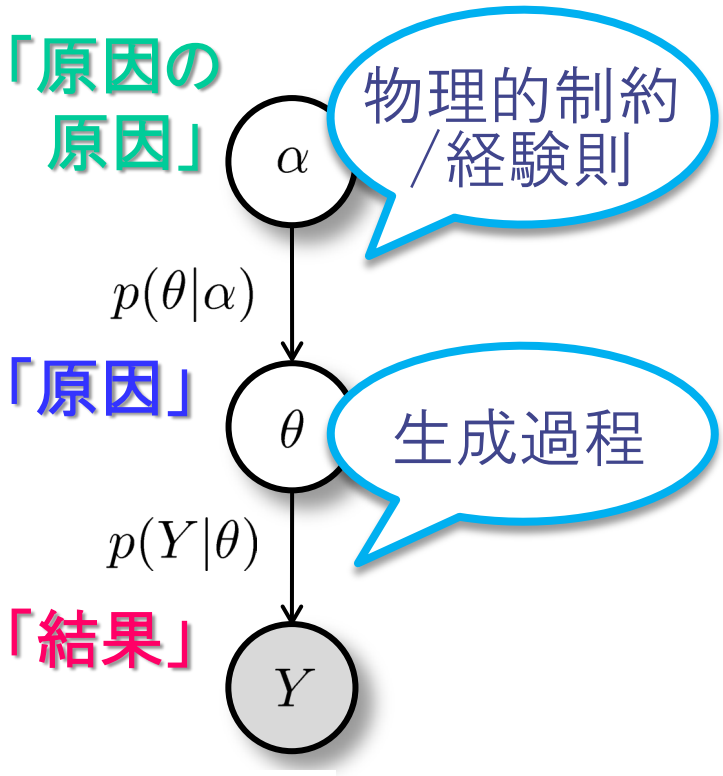
- なぜ？

原因 θ について
「常識的／経験的／物理的にこうでろう」
ということが推定のヒントになっている

生成モデルアプローチ

① 尤度関数の仮定
- 観測データ Y を生成する確率的なプロセス $p(Y|\theta)$ をモデル化

② 事前分布の仮定
- 生成モデルのパラメータ θ の生成プロセス $p(\theta)$ をモデル化



生成モデル（順問題）

③ 推論（逆問題）
- データ Y から θ と α を推論
- 最尤推定量 $\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, p(Y|\theta)$，MAP推定量 $\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, p(\theta, \alpha|Y)$
  MMSE推定量 $\hat{\theta} = \mathbb{E}[\theta, \alpha|Y]$，ベイズ事後分布 $p(\theta, \alpha|Y)$

トークのアウトライン

- 音響信号処理
  - 多重音解析
  - ブラインド音声分離
- 音声情報処理
  - 音声イントネーション解析

各々の構成：

- 生成モデルの設計
- （推論アルゴリズム）
- 実験結果例

トークのアウトライン

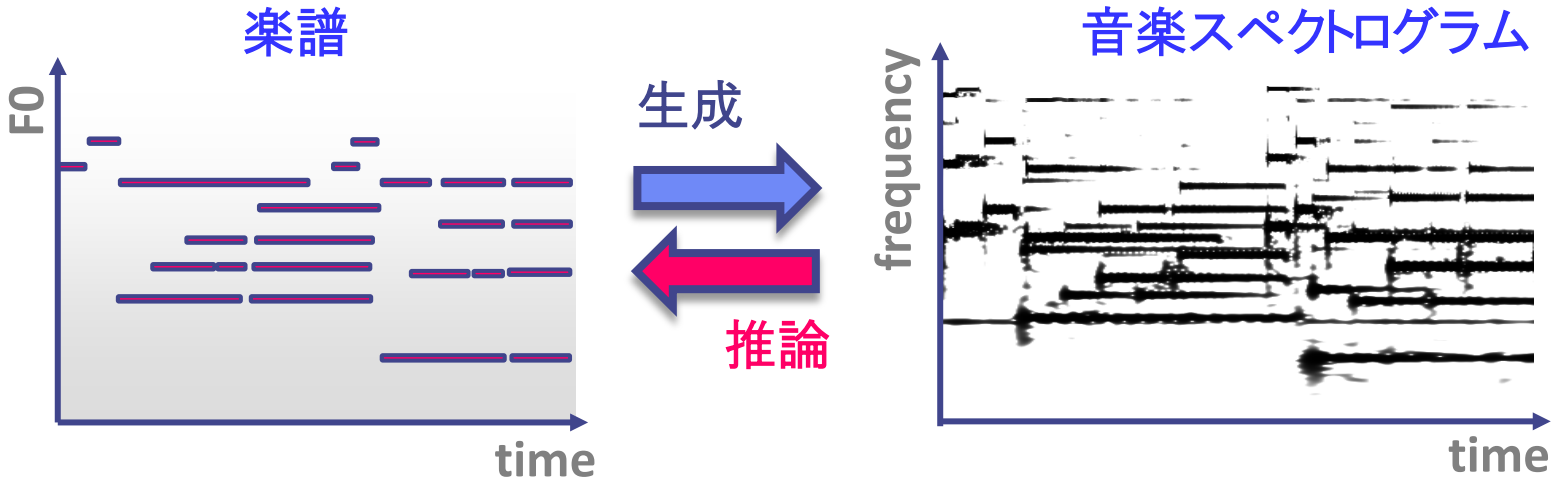
- 音響信号処理
  - 多重音解析
  - ブラインド音声分離
- 音声情報処理
  - 音声イントネーション解析

各々の構成：

- 生成モデルの設計
- （推論アルゴリズム）
- 実験結果例

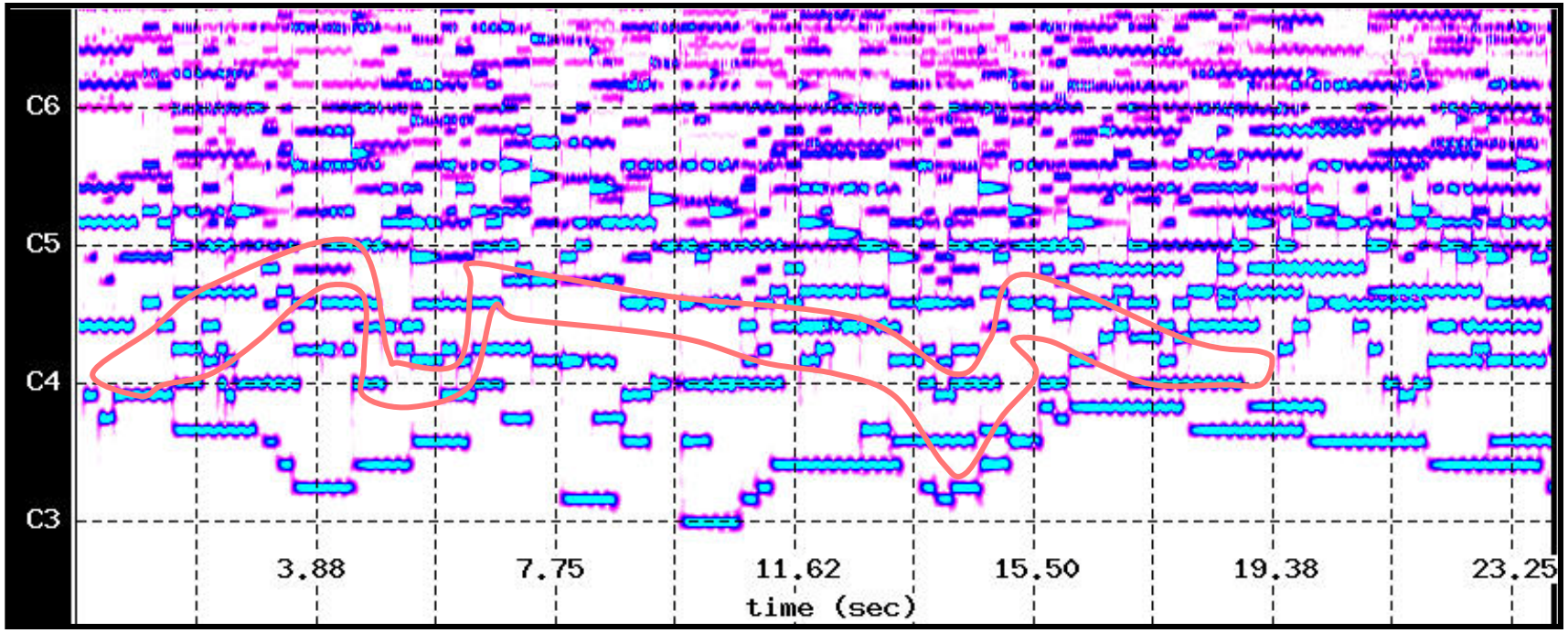
多重音解析

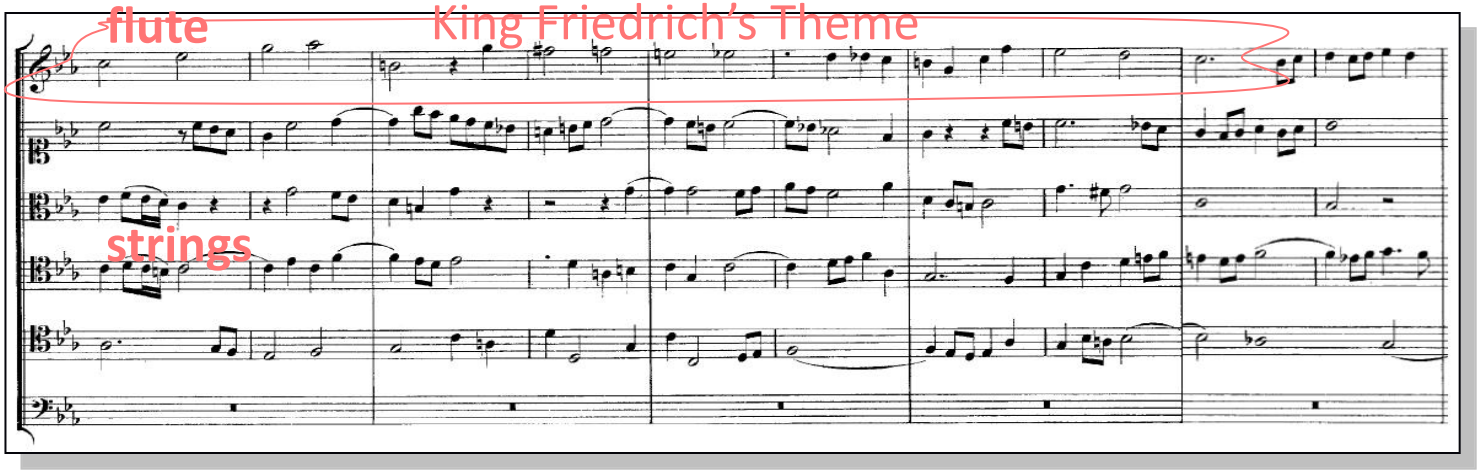
- モノラル多重音信号を個々の音源成分に分解する問題



多重音解析の難しさ

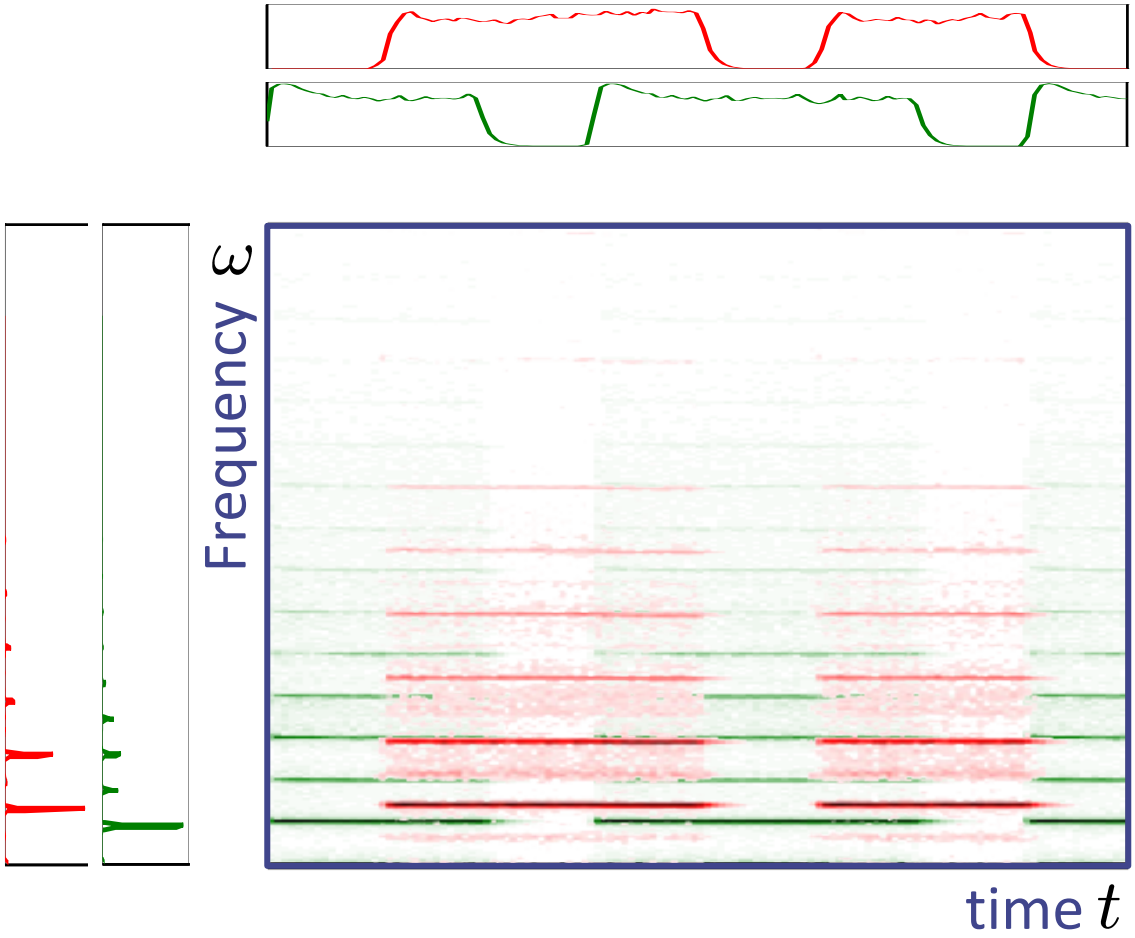
J.S.Bach: Ricercare a 6 aus “Das Musikalische Opfer” BWV1079





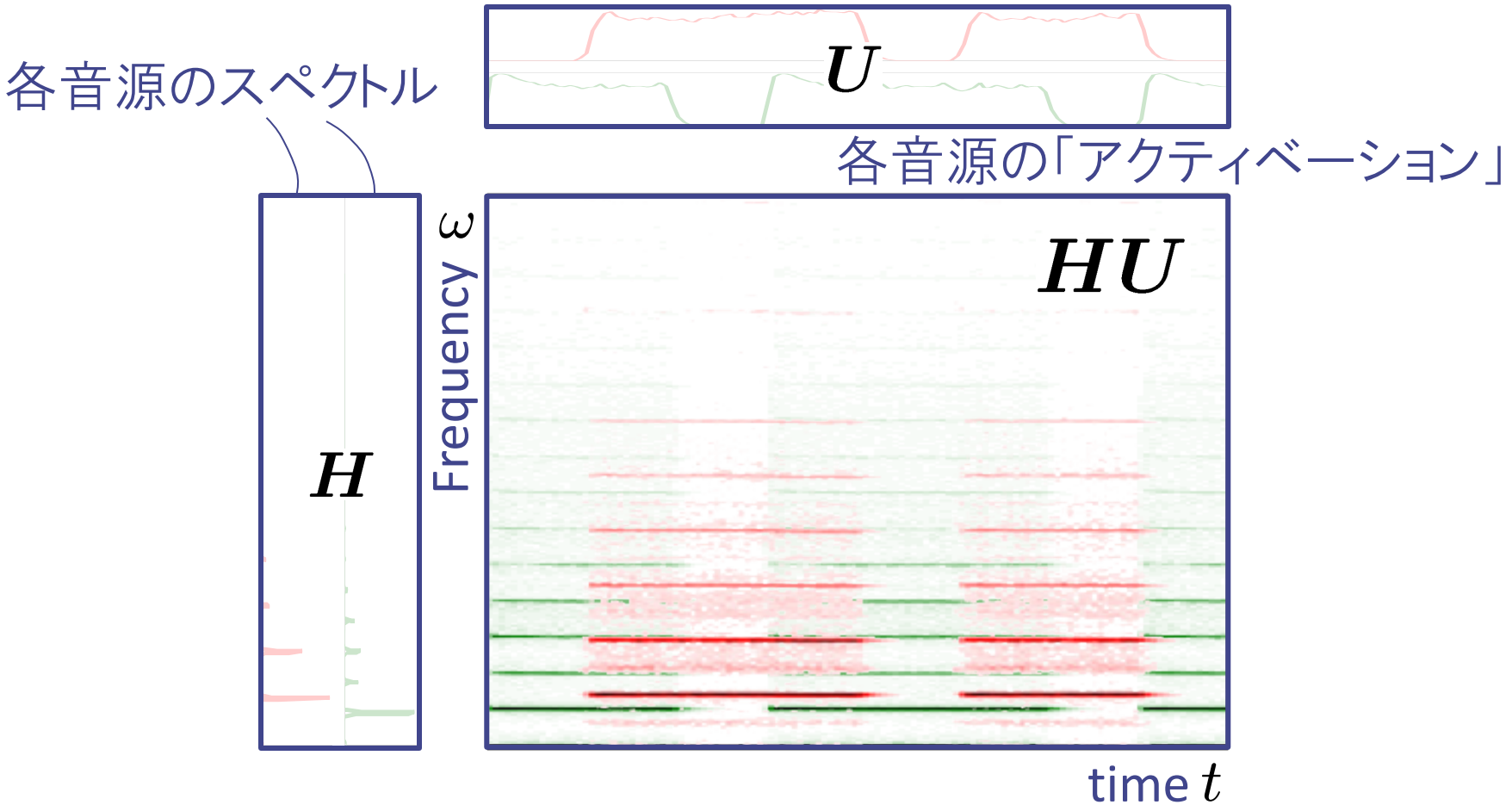
生成モデル化の方針

- 「行列積」としてのスペクトログラム



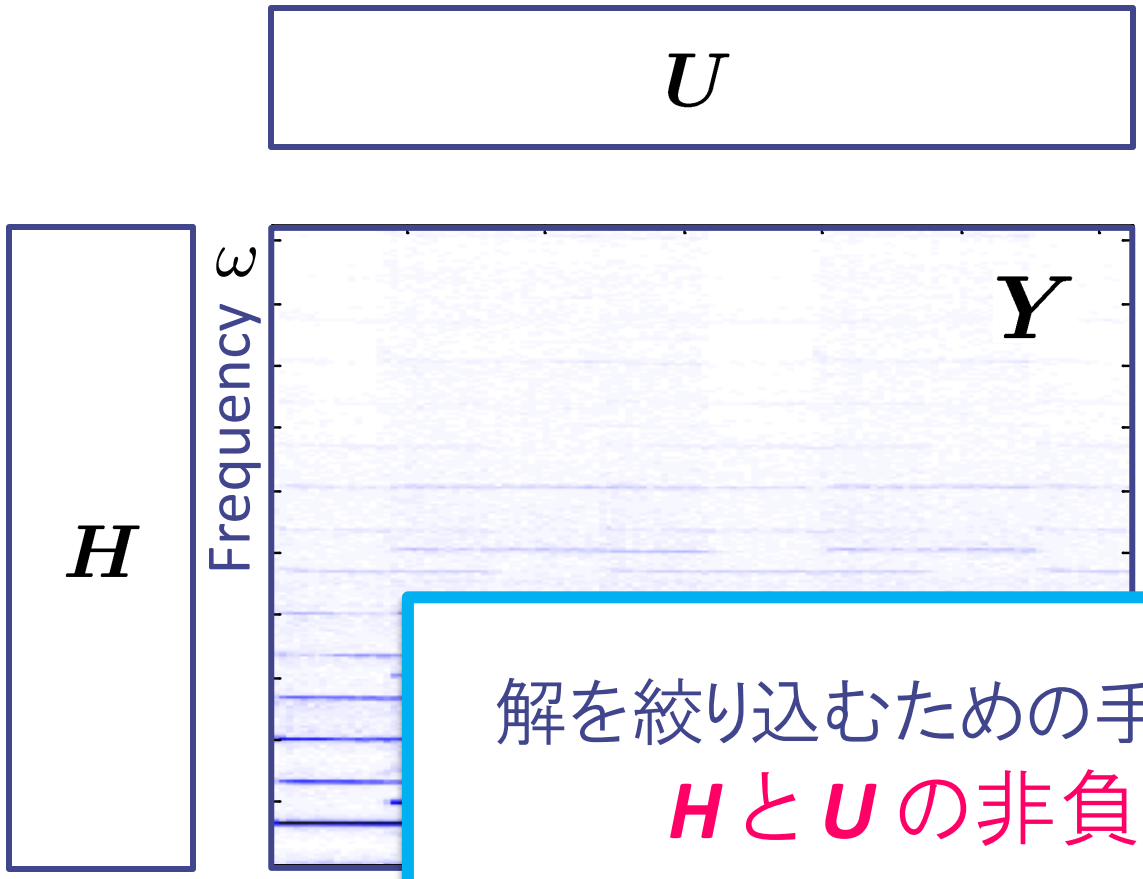
生成モデル化の方針

- 「行列積」としてのスペクトログラム



生成モデル化の方針

- 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

非負値行列因子分解(NMF)

[Lee 1999], [Smaragdis 2003], [Févotte 2008], etc...

- 多重音スペクトログラムの生成モデル化

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

→ 音源 k の複素スペクトログラム

$$Y_{\omega,t} = \sum_k C_{k,\omega,t}$$

→ 多重音の複素スペクトログラム

⇕

→ $C_{k,\omega,t}$ と $C_{k',\omega,t}$ の独立性を仮定

$$Y_{\omega,t} \sim \mathcal{N}_{\mathbb{C}}\left(0, \sum_k H_{\omega,k}U_{k,t}\right)$$

→ $p(Y|\theta)$ $\theta = \{H, U\}$

$p(\theta)$: $\boldsymbol{H}$ と $\boldsymbol{U}$ が負値の場合0になる分布

パラメータ推定

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

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

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

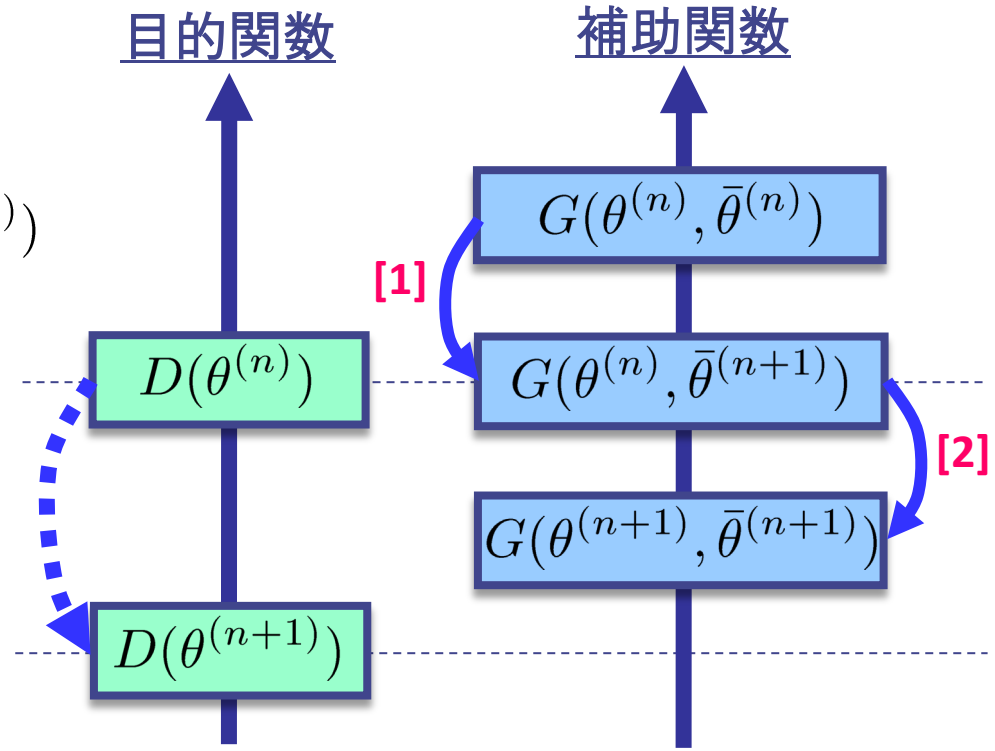
パラメータ推定アルゴリズムの導出 [Kameoka 2006]

- 補助関数法
  - 目的関数 $D(\theta)$ を反復的に降下させる数値解析手法
  - $D(\theta) = \min_{\bar{\theta}} G(\theta, \bar{\theta})$ を満たす$G(\theta, \bar{\theta})$ を補助関数と定義

⇒反復アルゴリズム

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

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



- 収束性

$$D(\theta^{(n)}) = G(\theta^{(n)}, \bar{\theta}^{(n+1)}) \geq G(\theta^{(n+1)}, \bar{\theta}^{(n+1)}) \geq D(\theta^{(n+1)})$$

どうやって上限関数を設計する？

パラメータ推定アルゴリズムの導出 [Kameoka 2006]

- 目的関数

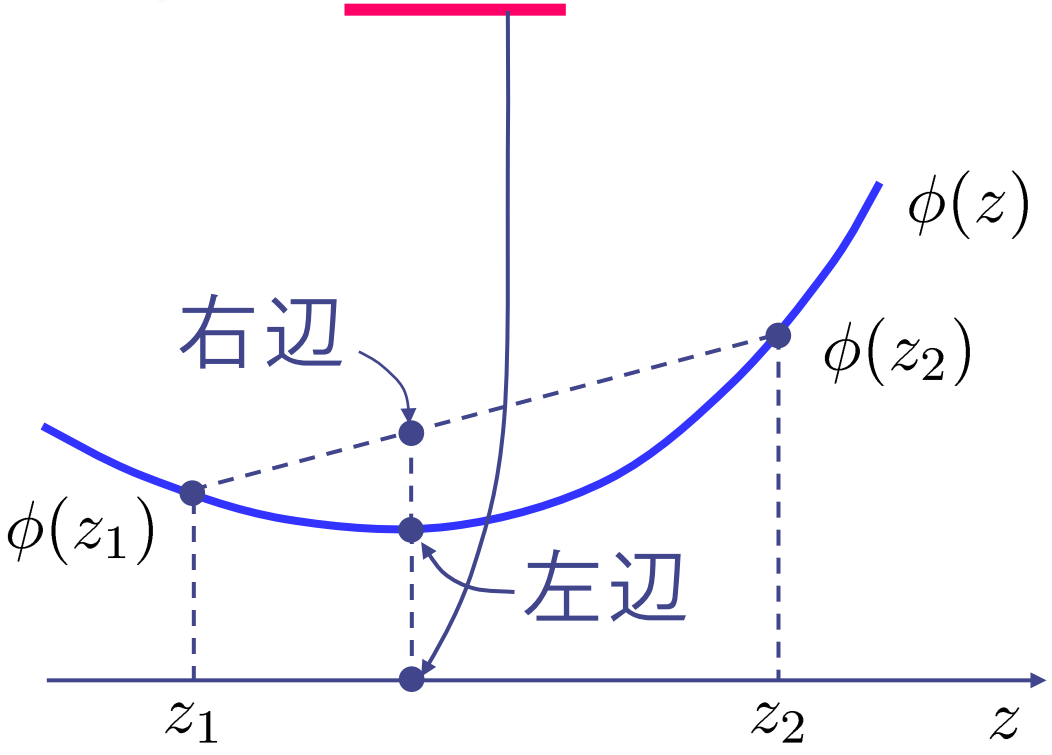
$$\mathcal{D}_{\mathrm{IS}}(\boldsymbol{H},\boldsymbol{U}) = \sum_{\omega,t}\left(\frac{|Y_{\omega,t}|^2}{\sum_k H_{\omega,k}U_{k,t}} + \log\sum_k H_{\omega,k}U_{k,t} - \cdots\right)$$

- 逆数関数は凸関数　⇒　Jensenの不等式で上限設計

パラメータ推定アルゴリズムの導出 [Kameoka 2006]

- Jensenの不等式とは？
  - $\phi(\cdot)$ ：凸関数
  - $\lambda_i \geq 0,\ \sum_i \lambda_i = 1$

$$\phi\left(\sum_i \lambda_i z_i\right) \leq \sum_i \lambda_i \phi(z_i)$$



パラメータ推定アルゴリズムの導出 [Kameoka 2006]

- 目的関数

$$\mathcal{D}_{\mathrm{IS}}(\boldsymbol{H},\boldsymbol{U}) = \sum_{\omega,t}\left(\frac{|Y_{\omega,t}|^2}{\sum_k H_{\omega,k}U_{k,t}} + \log\sum_k H_{\omega,k}U_{k,t} - \cdots\right)$$

- 逆数関数は凸関数 ⇒ Jensenの不等式で上限設計

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

適用

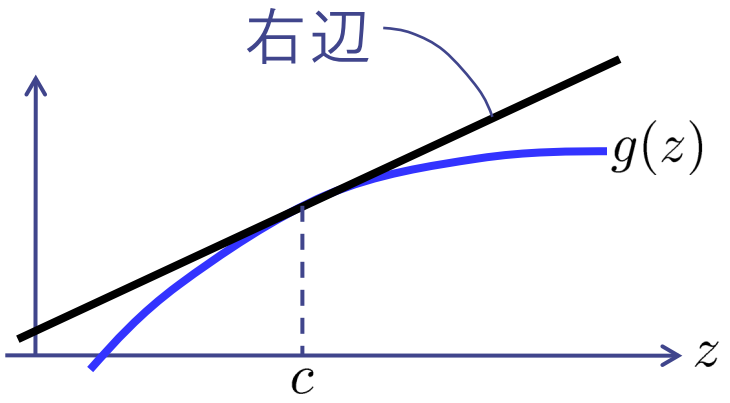
$$\frac{1}{\sum_k H_{\omega,k}U_{k,t}} \le \sum_k \frac{\lambda_{k,\omega,t}^2}{H_{\omega,k}U_{k,t}}$$

- 対数関数は凹関数 ⇒ この場合はどうしたら？？

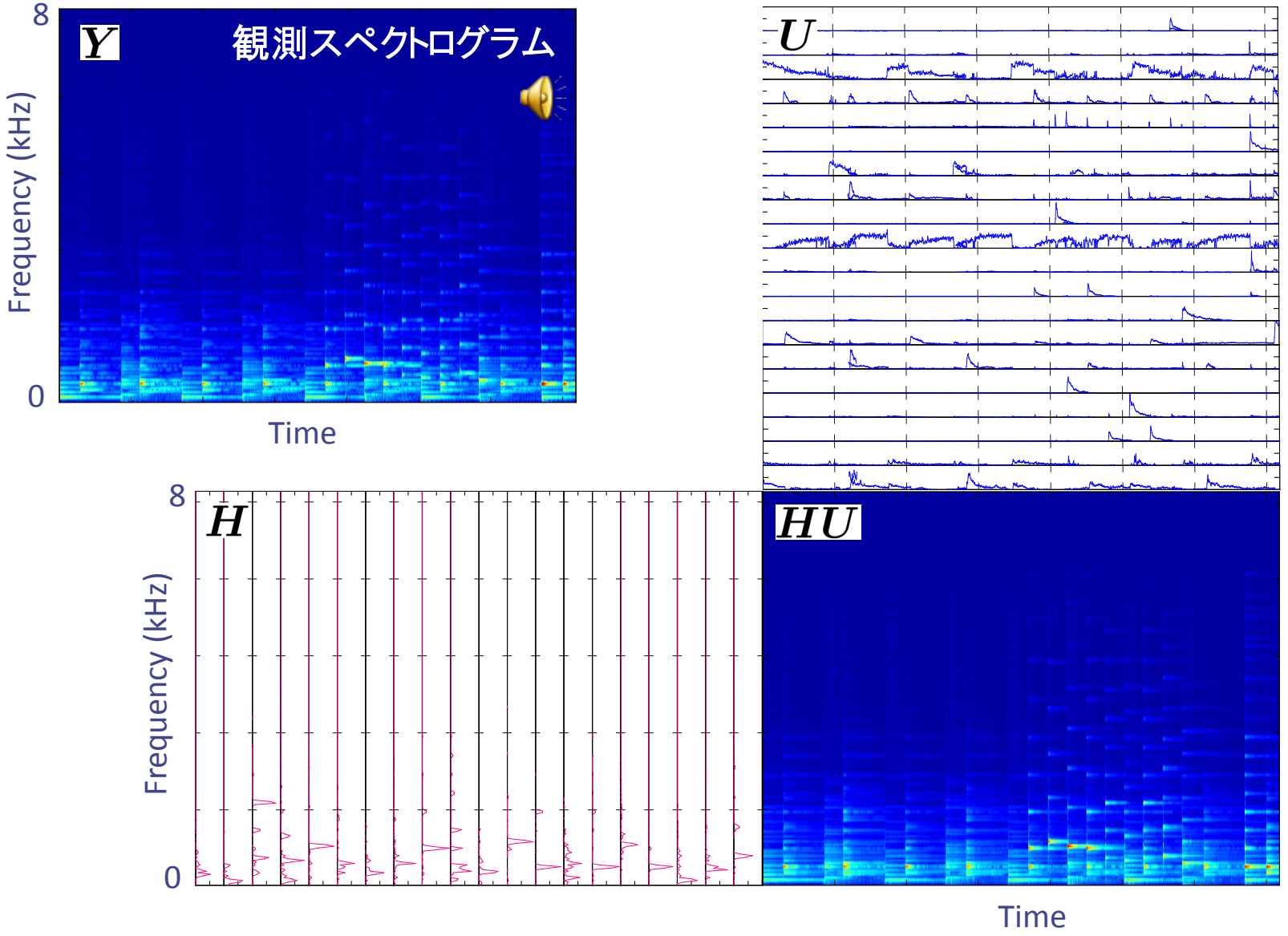
$$g(z) \le g'(c)(z-c) + g(c)$$

適用

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



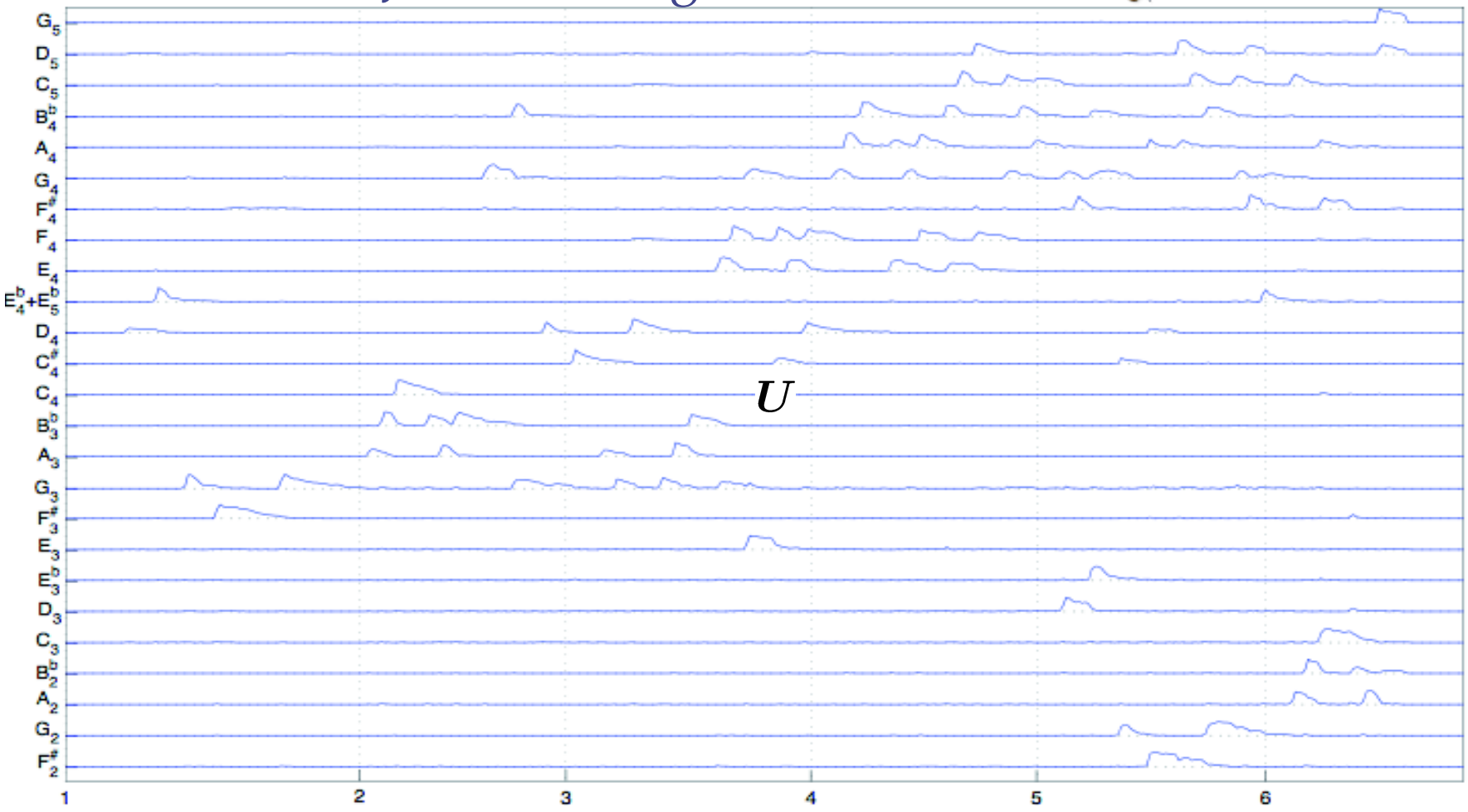
適用例


8
Y
観測スペクトログラム
Frequency (kHz)
0
Time
U
8
H
HU
Frequency (kHz)
0
Time

応用例 (1/5)

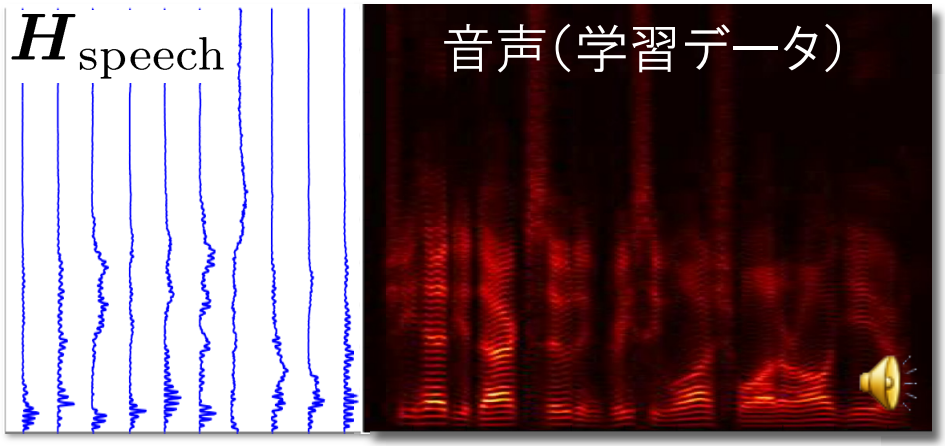
- 自動採譜 [Smaragdis 2003]

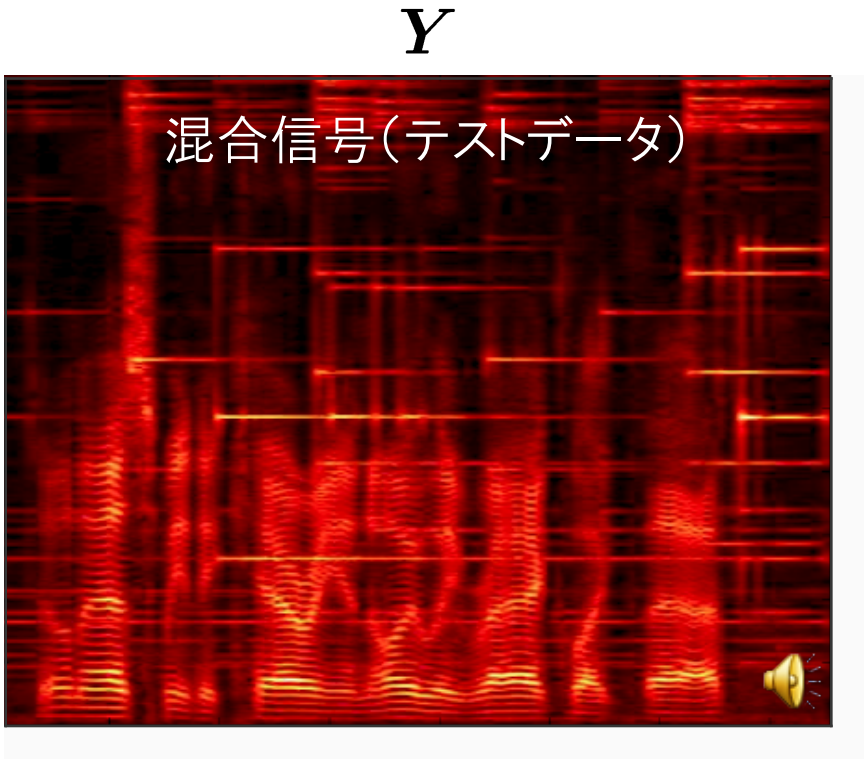
J.S. Bach: Fuge #16 in G minor



応用例 (2/5)

- 教師ありモノラル音源分離 [Smaragdis 2007]





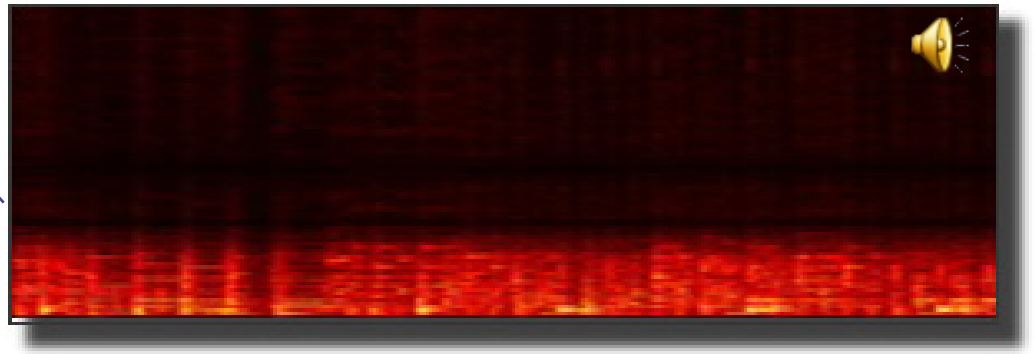
$$Y \simeq \underbrace{[H_{\text{speech}} \; H_{\text{chime}}]}_{\text{固定}} \begin{bmatrix} U_{\text{speech}} \\ U_{\text{chime}} \end{bmatrix}$$

応用例 (3/5)

■音の「超解像」

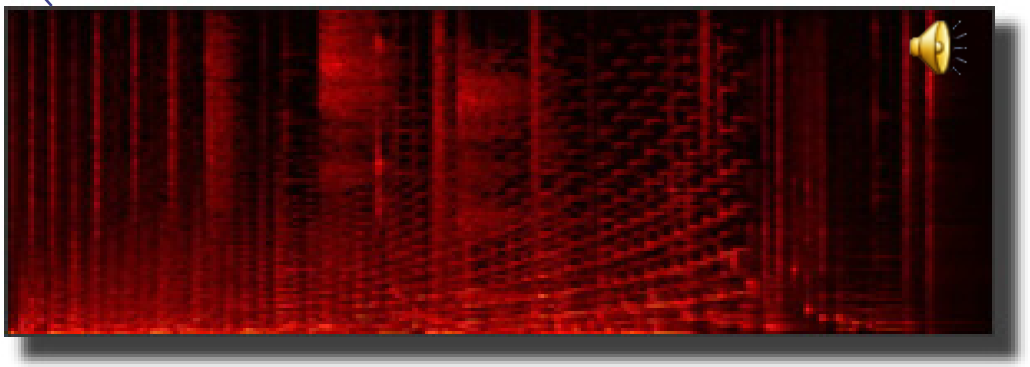
[Smaragdis & Raj 2007]

低サンプリングレートの信号



$\boldsymbol{Y}_{\mathrm{l}}$ D

高サンプリングレート信号の学習データ



$\boldsymbol{H}_{\mathrm{h}}$

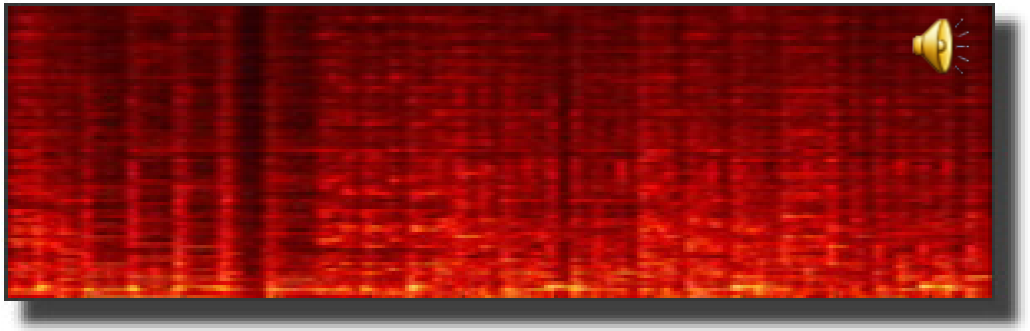
$\boldsymbol{H}_{\mathrm{l}}$ D

$\boldsymbol{H}$

$\boldsymbol{Y}_{\mathrm{l}} \simeq \boldsymbol{H}_{\mathrm{l}}\boldsymbol{U}$

（固定）

高周波帯域が復元された信号

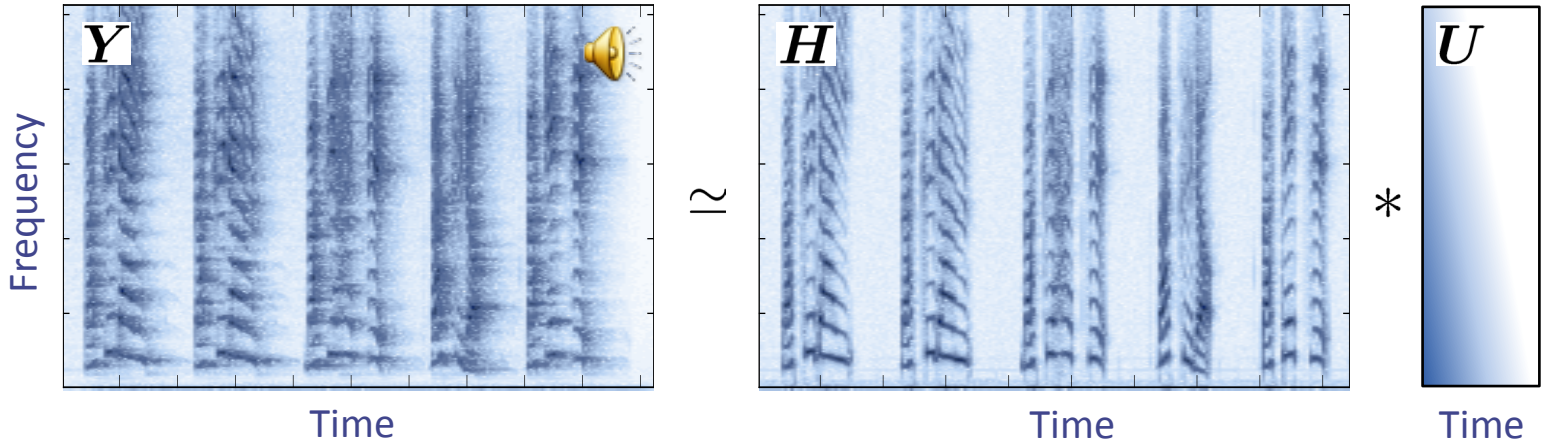


$\boldsymbol{H}_{\mathrm{h}}\boldsymbol{U} = \boldsymbol{Y}_{\mathrm{h}}$

応用例 (4/5)

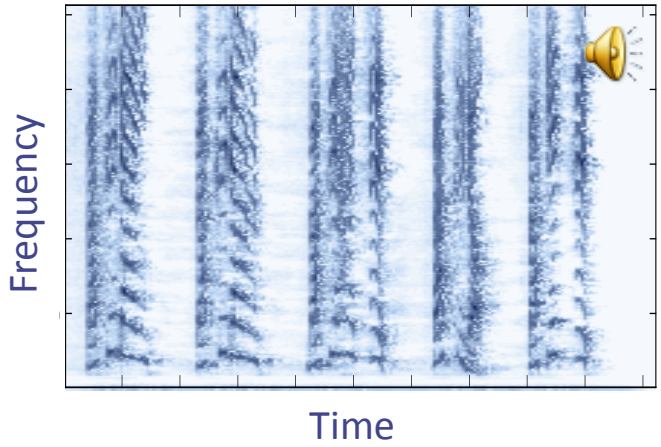
- ブラインド残響除去 [Kameoka 2008]

非負値行列「逆畳み込み」



$$Y_{\omega,t} \simeq \sum_{\tau} H_{\omega,\tau} U_{\omega,t-\tau}$$

- 残響除去音声

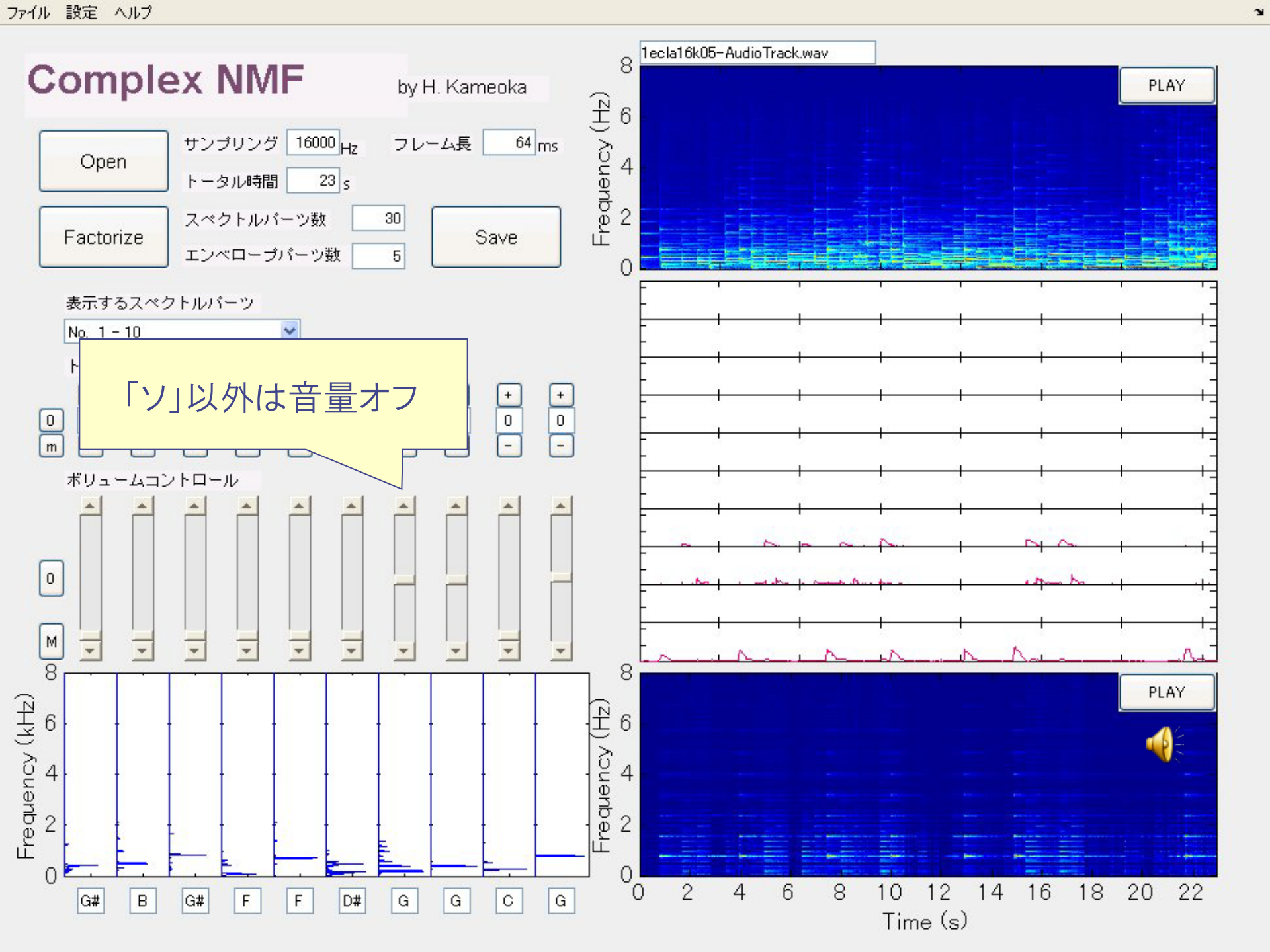


非負値行列逆畳み込み
によって求まった実際の ***H***

応用例 (5/5)

- エコーキャンセラ
  [戸上&川口 2009]
- 音響符号化
  [Nikunen & Virtanen 2010]
- ブラインド音源分離 (NMFの多チャンネル拡張)
  [Ozerov et al. 2010], [Kitano et al. 2010], [Takeda et al. 2011], [Sawada et al. 2011], [Sawada et al. 2012]
- フォルマントトラッキング
  [Durrieu et al. 2011]
- 音素特徴量抽出
  [Hurmalainen et al. 2011]

ファイル 設定 ヘルプ
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)
G# B G# F F D# G G C G
1ecla16k05-AudioTrack.wav
PLAY
Frequency (Hz)
PLAY
Frequency (Hz)
Time (s)


ファイル 設定 ヘルプ
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
1ecla16k05-AudioTrack.wav
PLAY
Frequency (Hz)
Frequency (kHz)
Time (s)
G# B G# F F D# G G C G

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

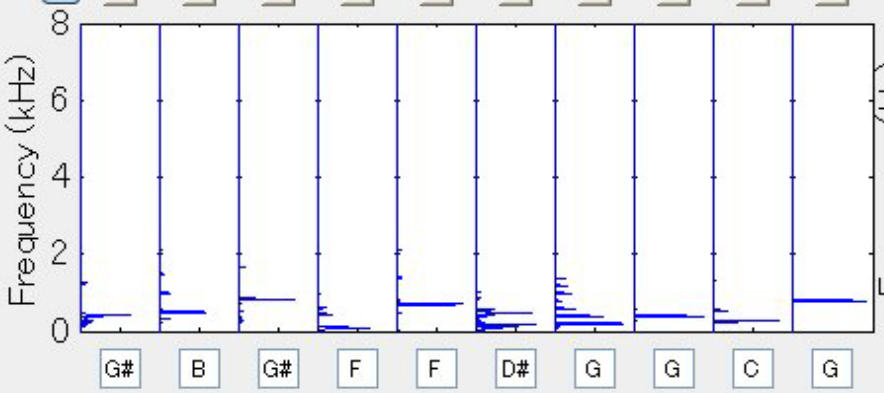
Save

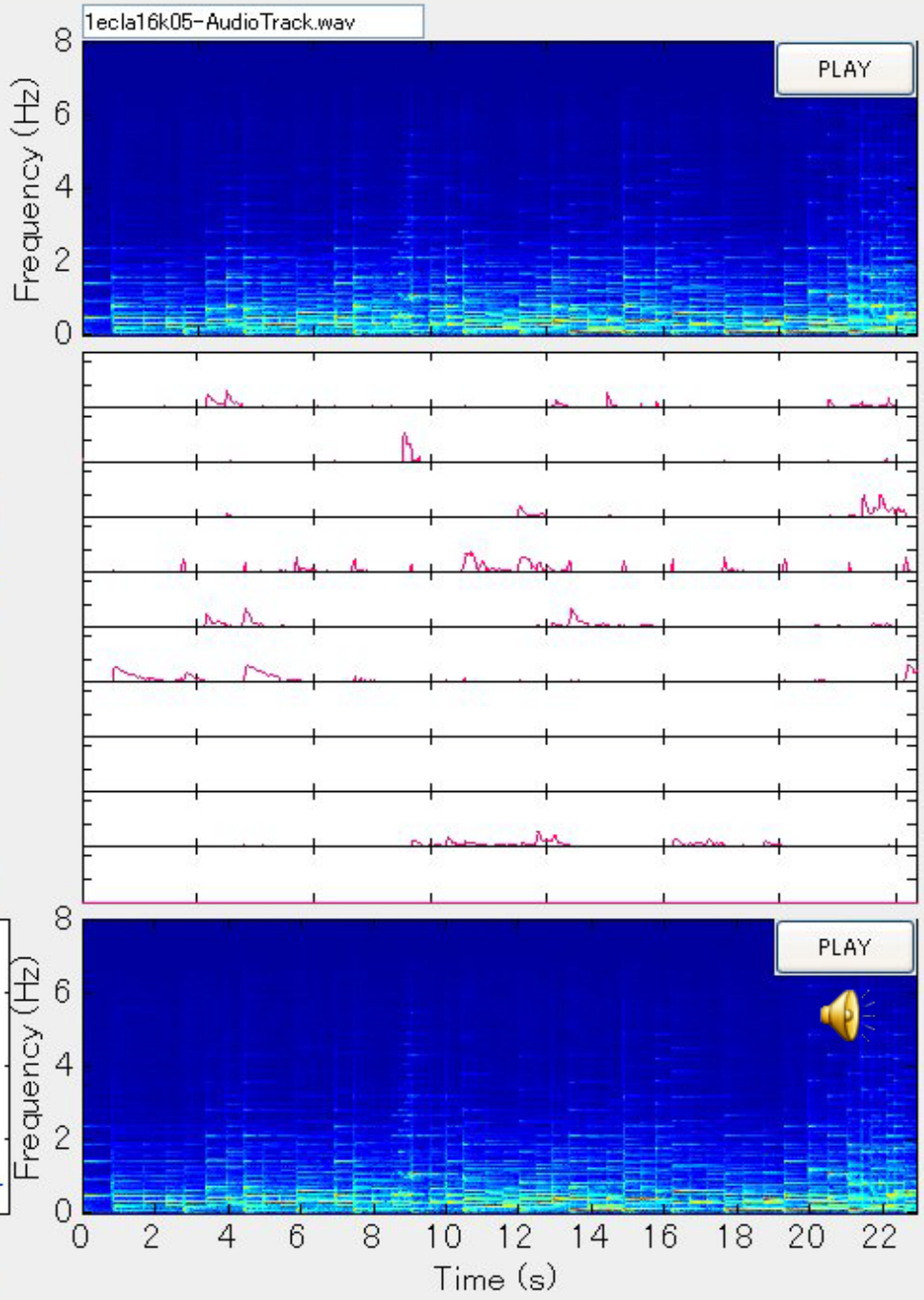
表示するスペクトルパーツ

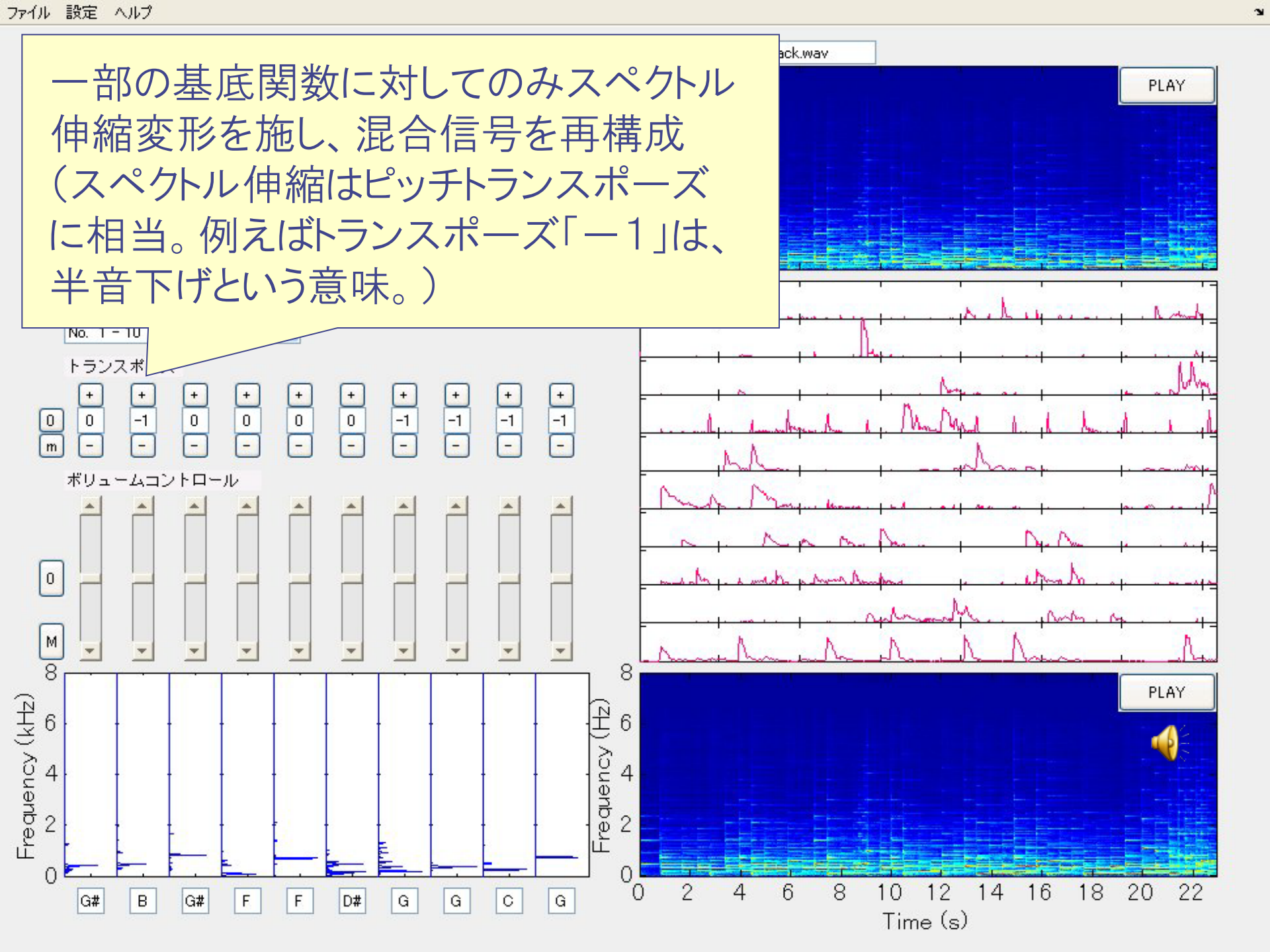
No. 1 - 10

「ソ」だけ音量オフ

ボリュームコントロール






ファイル　設定　ヘルプ
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。）
ack.wav
PLAY
No. 1 - 10
トランスポ
0
m
0 -1 0 0 0 0 -1 -1 -1 -1
ボリュームコントロール
0
M
Frequency (kHz)
8
6
4
2
0
G# B G# F F D# G G C G
Frequency (Hz)
PLAY
0 2 4 6 8 10 12 14 16 18 20 22
Time (s)

トークのアウトライン

- 音響信号処理
  - 多重音解析
  - ブラインド音声分離
- 音声情報処理
  - 音声イントネーション解析

各々の構成：

- 生成モデルの設計
- （推論アルゴリズム）
- 実験結果例

トークのアウトライン

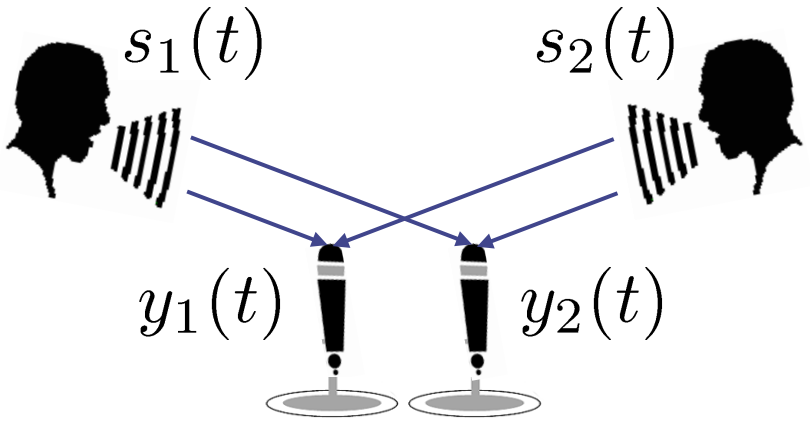
- 音響信号処理
  - 多重音解析
  - ブラインド音声分離
- 音声情報処理
  - 音声イントネーション解析

各々の構成：

- 生成モデルの設計
- （推論アルゴリズム）
- 実験結果例

ブラインド音源分離

- 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」

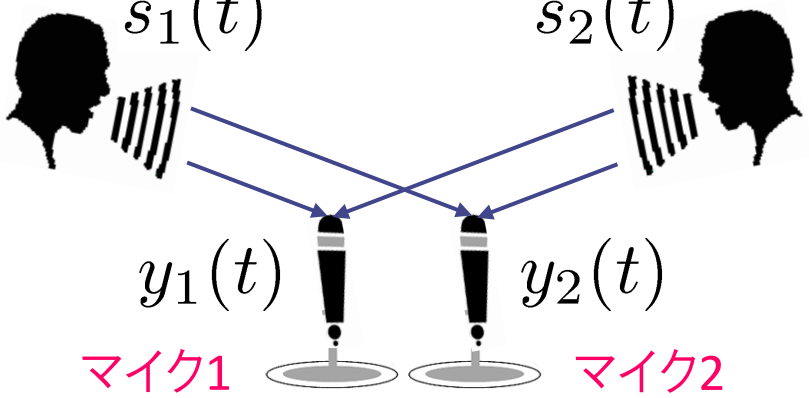


- 実応用場面
  - 音を使った監視システム
    - どこで何が起こっているかを検知
    - 介護や防犯のための安全モニタリングとしての応用
      Cf. ShotSpotter（米国で開発されている発砲事件の検知システム）
  - 人間のカクテルパーティー効果の模倣
    - ロボット聴覚，補聴器への応用

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 $s_1(t)$　　音源2 $s_2(t)$

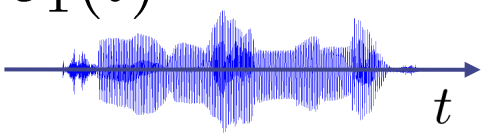


$y_1(t)$　　$y_2(t)$

マイク1　　マイク2

- 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: $a_{2,2}s_2(t+\tau_{2,2})$

$s_1(t)$



t

$*$

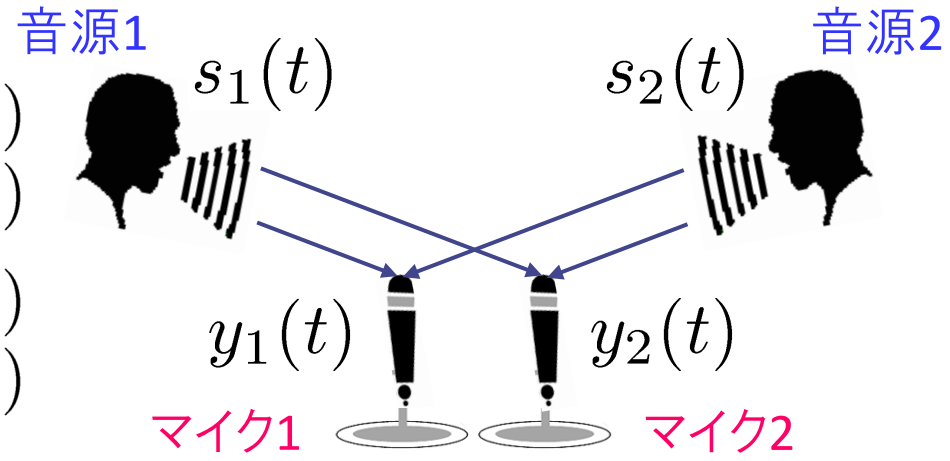
$a_{1,1}$

$\tau_{1,1}$ マイクまでの到達時間

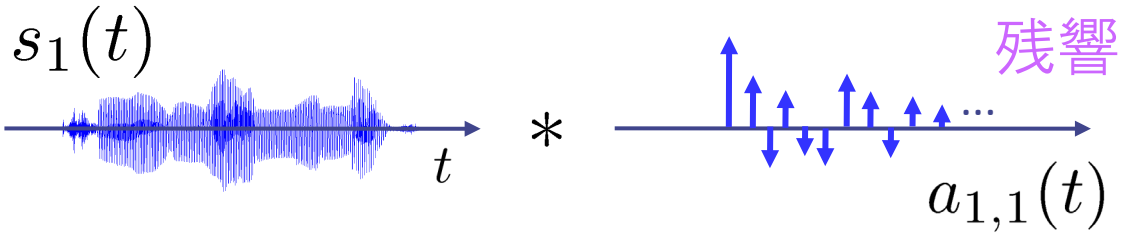
残響があると・・・

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”



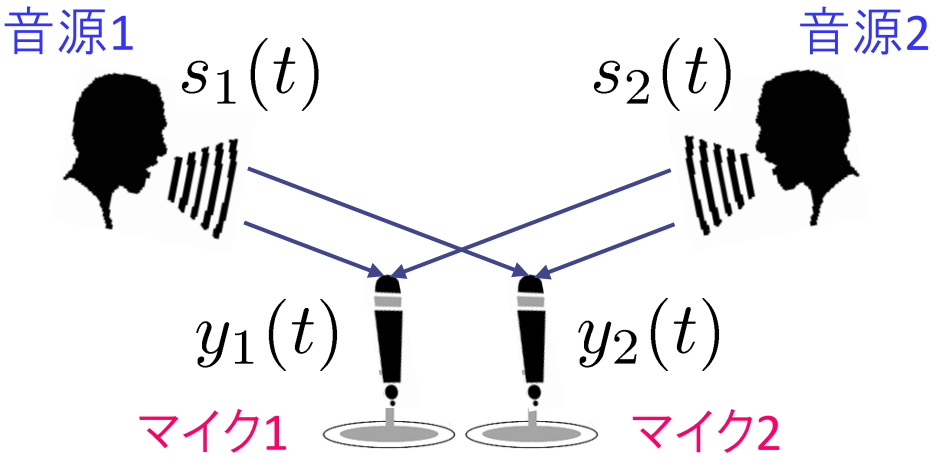
- 音源1→マイク1: $a_{1,1}s_1(t+\tau_{1,1})$
- 音源1→マイク2: $a_{2,1}s_1(t+\tau_{2,1})$
- 音源2→マイク1: $a_{1,2}s_2(t+\tau_{1,2})$
- 音源2→マイク2: $a_{2,2}s_2(t+\tau_{2,2})$



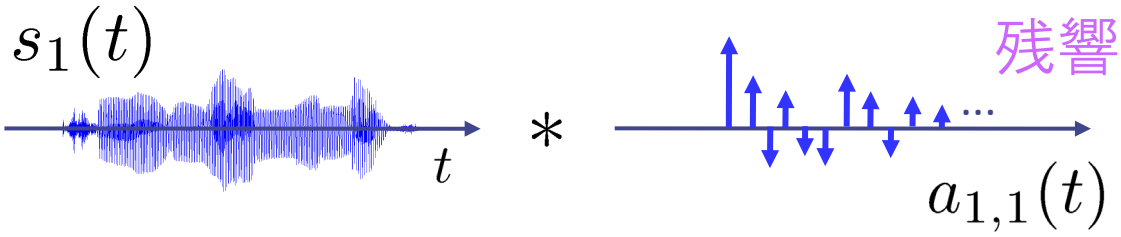
残響があると・・・

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: $y_2(t) = a_{2,1}(t) * s_1(t) + a_{2,2}(t) * s_2(t)$

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

マイクロホンで観測される信号の生成プロセス

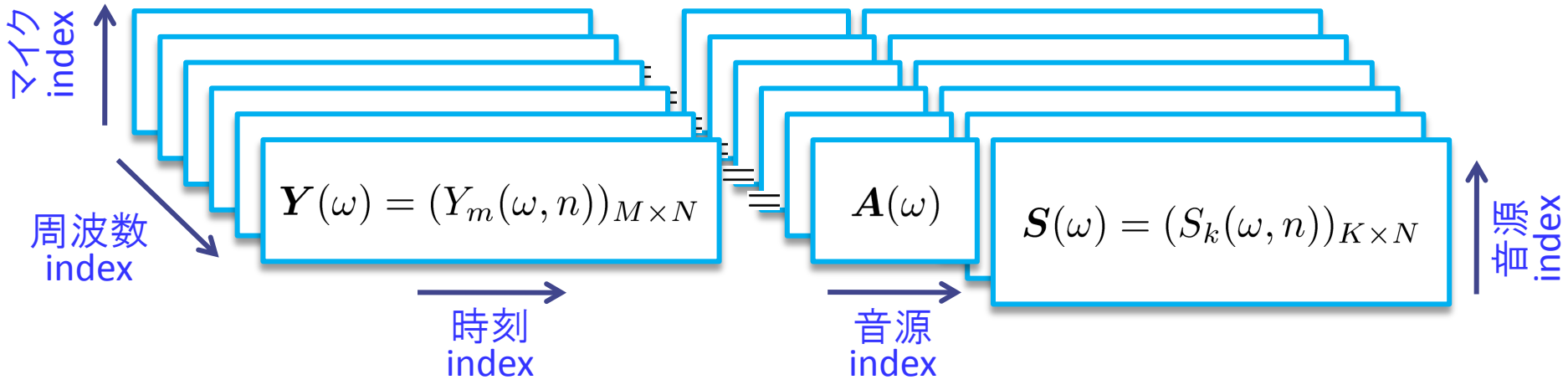
- Fourier変換領域で畳み込みは積になることを利用して畳み込み混合モデルを時間周波数領域に展開：

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

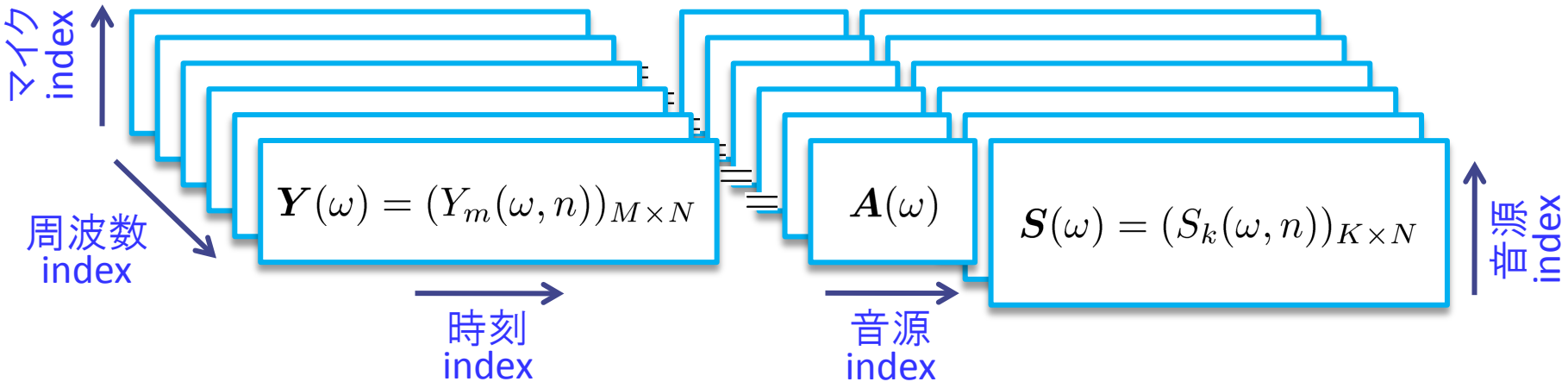
「時間領域の畳み込み混合モデル」

「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



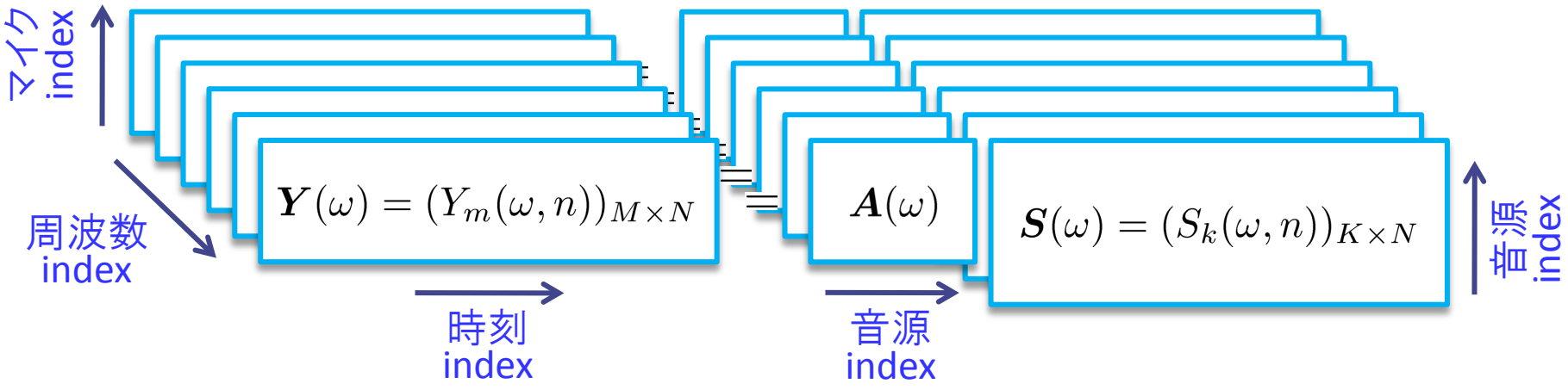
逆問題への手がかり



- 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 ⇒ $\boldsymbol{Y}(\omega) = \boldsymbol{A}(\omega)\boldsymbol{S}(\omega)$ は劣決定
  - パーミュテーションの任意性 $\boldsymbol{Y}(\omega) = \boldsymbol{A}(\omega)\boldsymbol{P}^{\mathsf{T}}(\omega)\boldsymbol{P}(\omega)\boldsymbol{S}(\omega)$

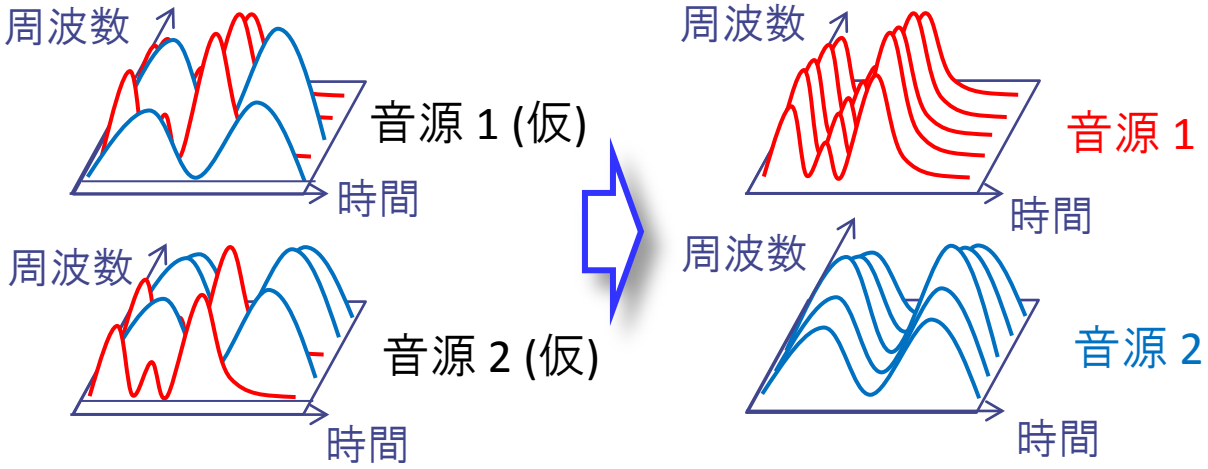
置換行列

逆問題への手がかり

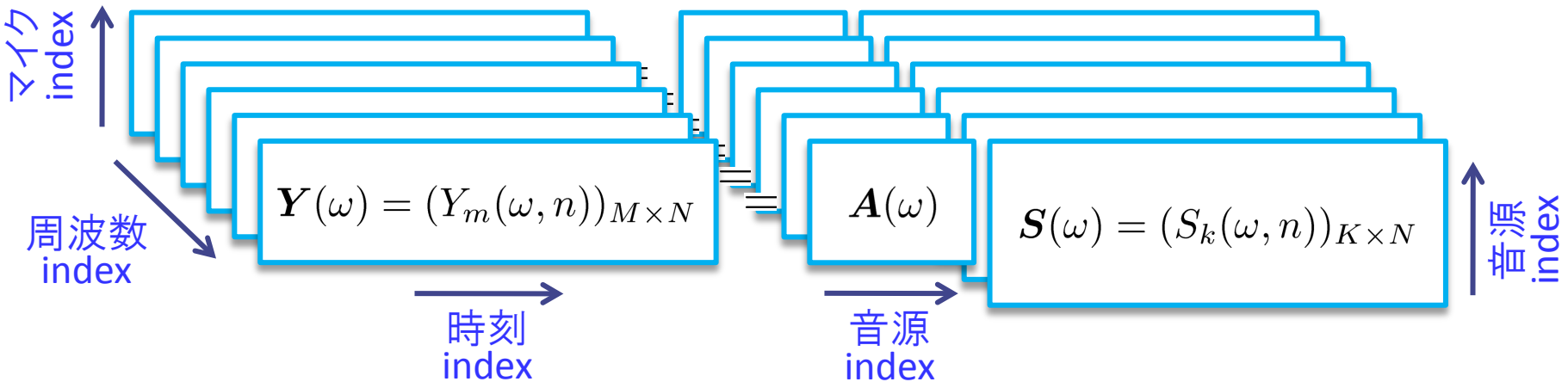


周波数ごとに
個別に分離が
得られても…

パーミュテーション整合

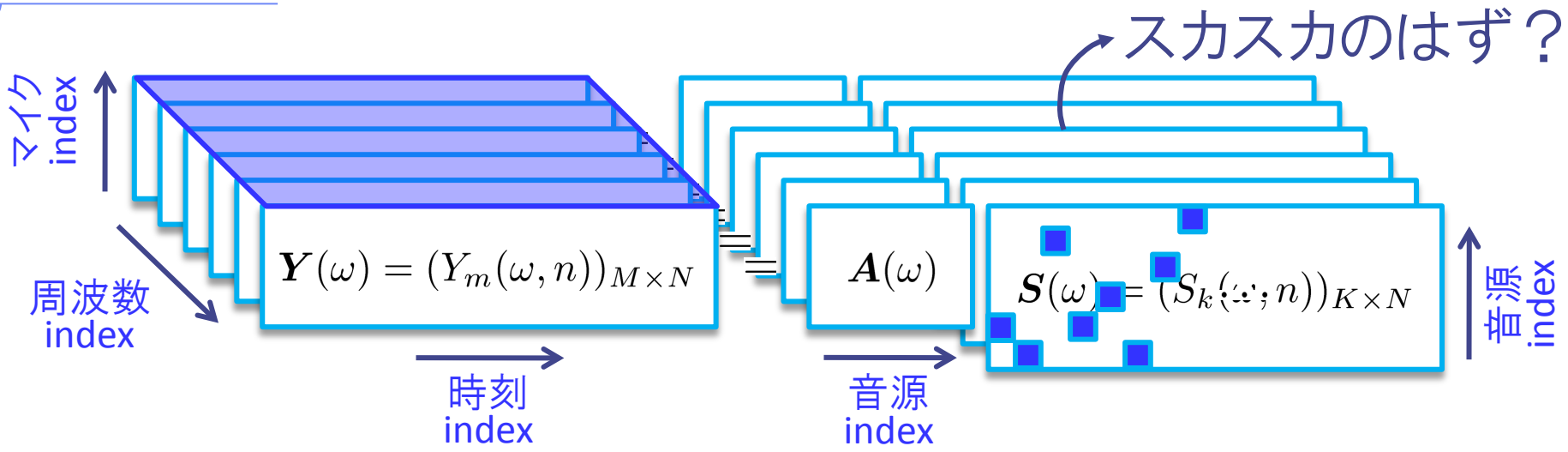


逆問題への手がかり

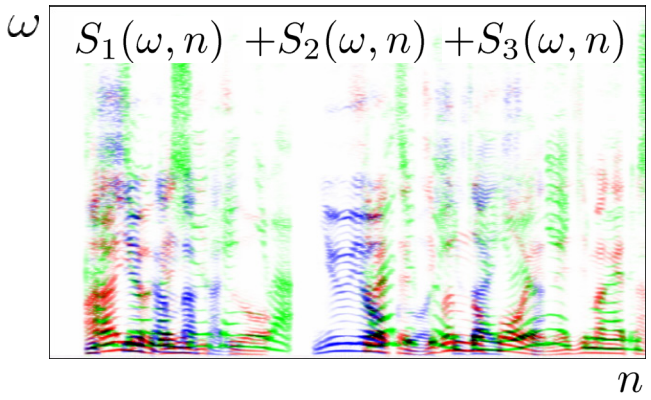


- 以上の生成プロセスの逆問題（音源分離）は不良設定
  - 音源数＞マイク数 ⇒ $\boldsymbol{Y}(\omega) = \boldsymbol{A}(\omega)\boldsymbol{S}(\omega)$ は劣決定
  - パーミュテーションの任意性 $\boldsymbol{Y}(\omega) = \boldsymbol{A}(\omega)\boldsymbol{P}^{\mathsf{T}}(\omega)\boldsymbol{P}(\omega)\boldsymbol{S}(\omega)$
    置換行列
- 解を絞り込むための手がかりが必要
  - （1）音声の性質と（2）音波伝播の性質に着目

(1) 音声のスパース性を仮定して尤度関数設計



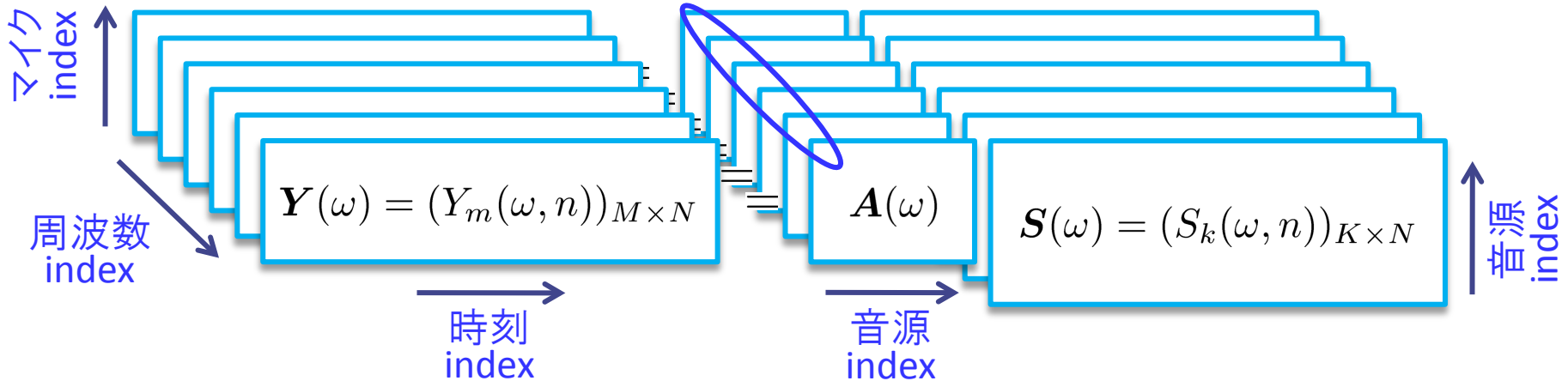
- 音声の時間周波数成分は極めてスパース！



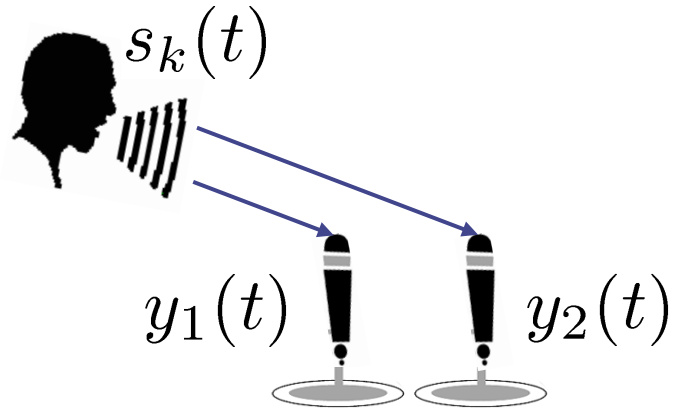
- 各時間周波数点で一つの音源のみがアクティブ（それ以外の音源の成分は0）と仮定できるならば，
  ωごとの行列分解の劣決定性は解消できるのではないか？
  [Yilmaz2004, Izumi2007, Araki2007, Sawada2010]
  [Kameoka 2012]
- アクティブな音源indexを表すインジケータ $z_{\omega,n}$ をパラメータ化

(2) 音波伝播の性質を活用して事前分布設計

[Kameoka 2012]

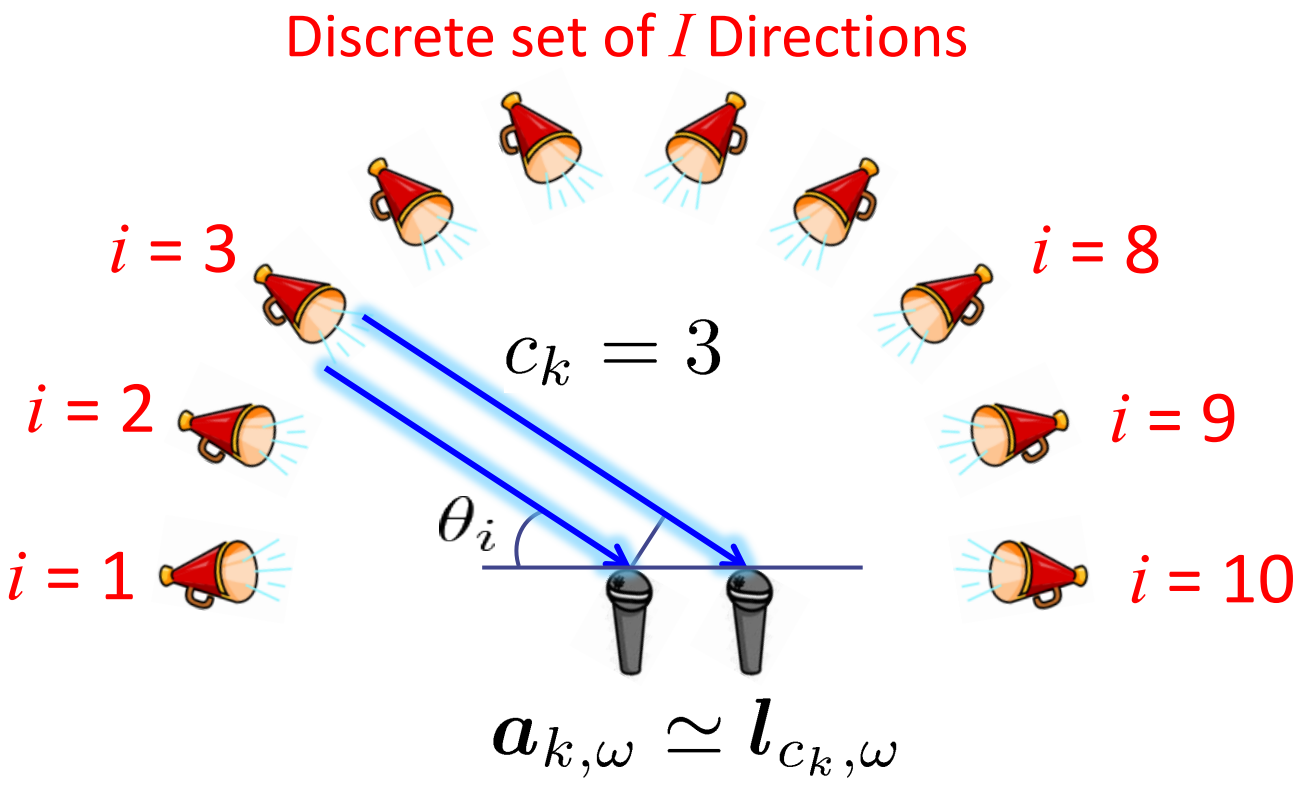


- 正しいパーミュテーション整合が取れているならば $\boldsymbol{A}(\omega)$ はω方向に音波伝播に基づく拘束があるはず
  - 平面波音源を仮定すれば $\boldsymbol{A}(\omega)$ は各音源の到来方向によって陽に表されるはず
  - 音源kの到来方向 c_k をパラメータ化



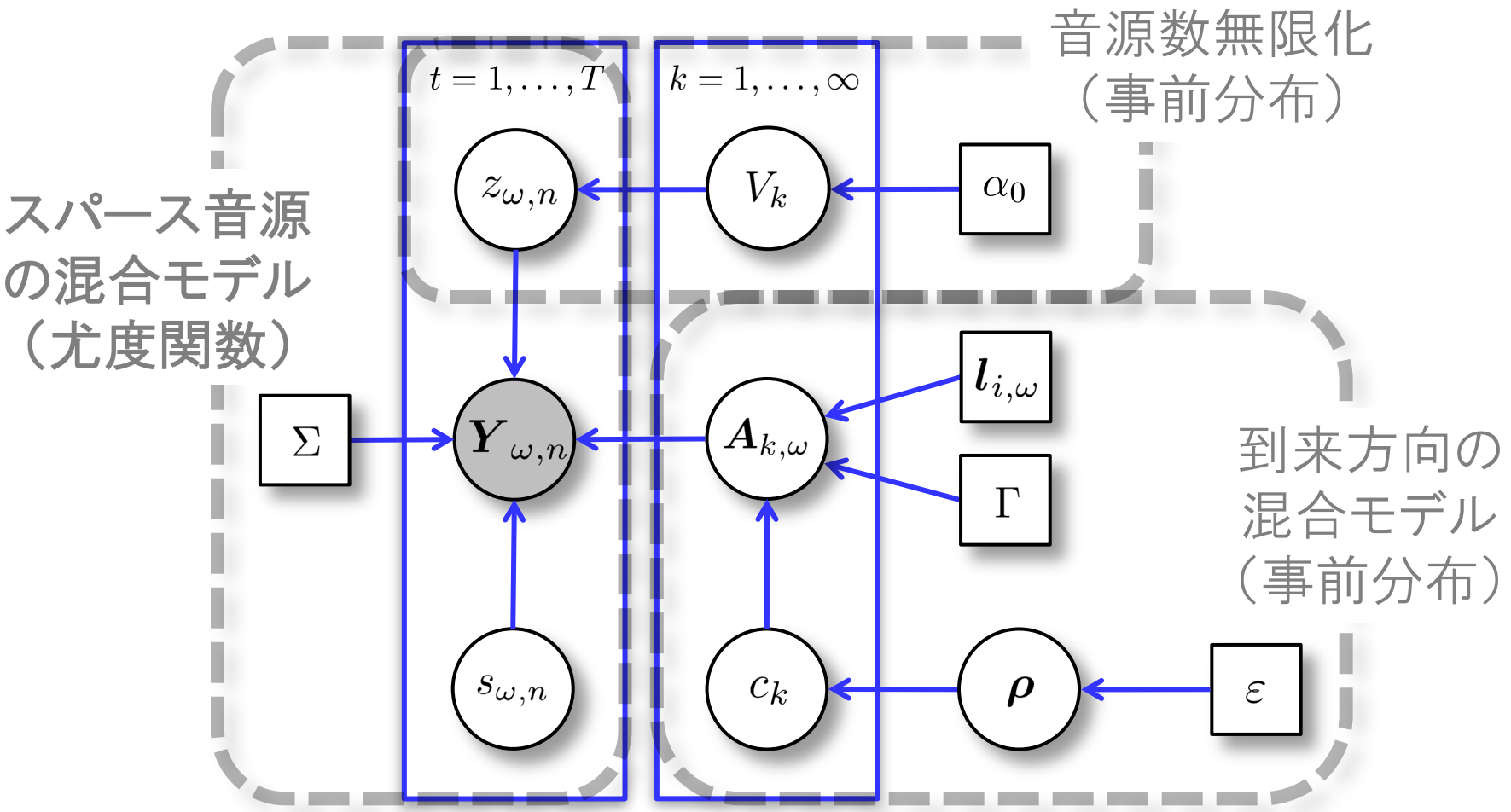
伝達周波数特性の事前分布設計 [Kameoka 2012]

- 伝達周波数特性の生成プロセス
  - I 個の候補値から各音源につき到来方向が1つ選ばれる
  - 選ばれた到来方向に応じて伝達周波数特性が決定される



生成モデルの全体像 [Kameoka 2012]

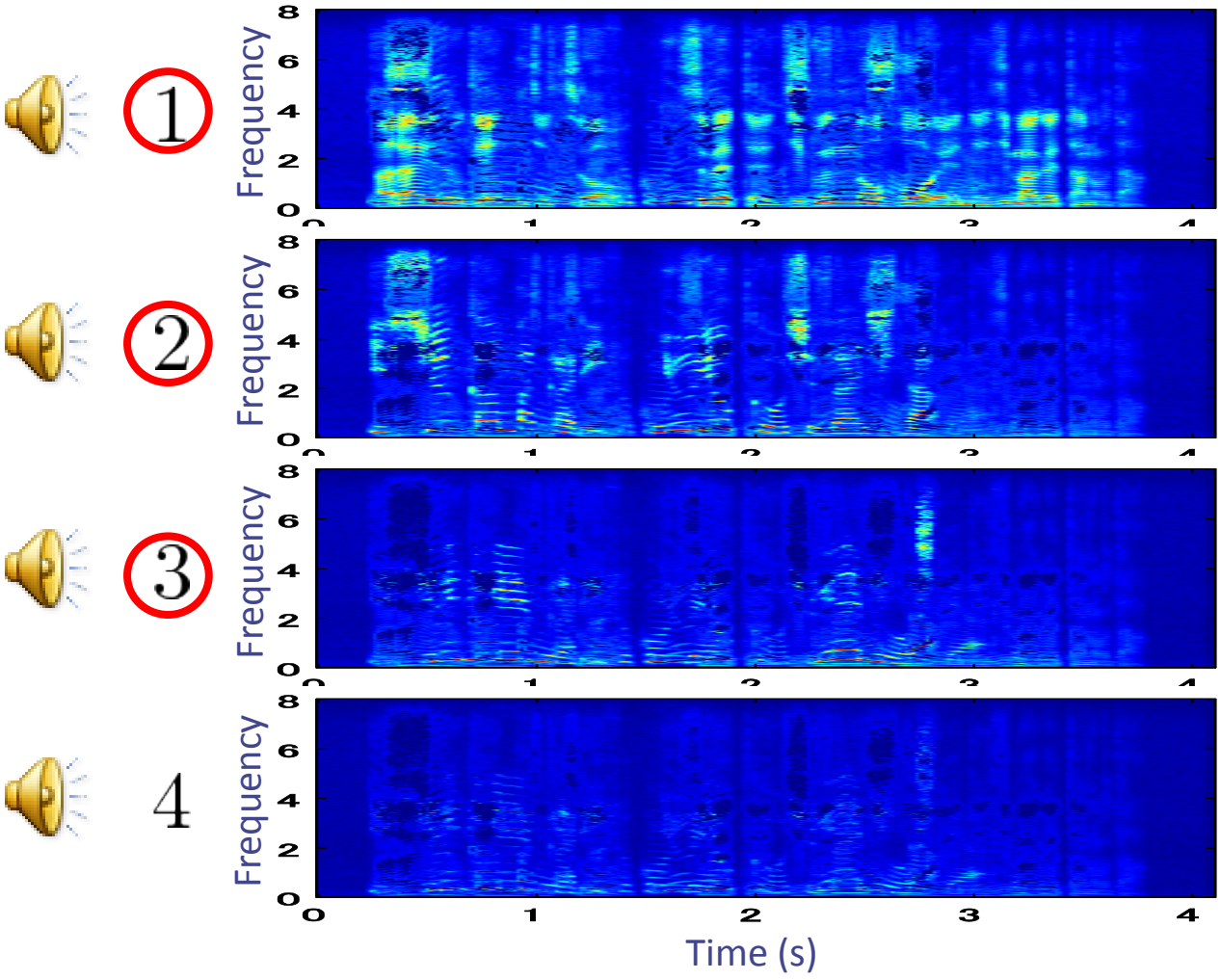
- グラフィカルモデル



- 変分ベイズ法により推論

音源分離結果例 [Kameoka 2012]

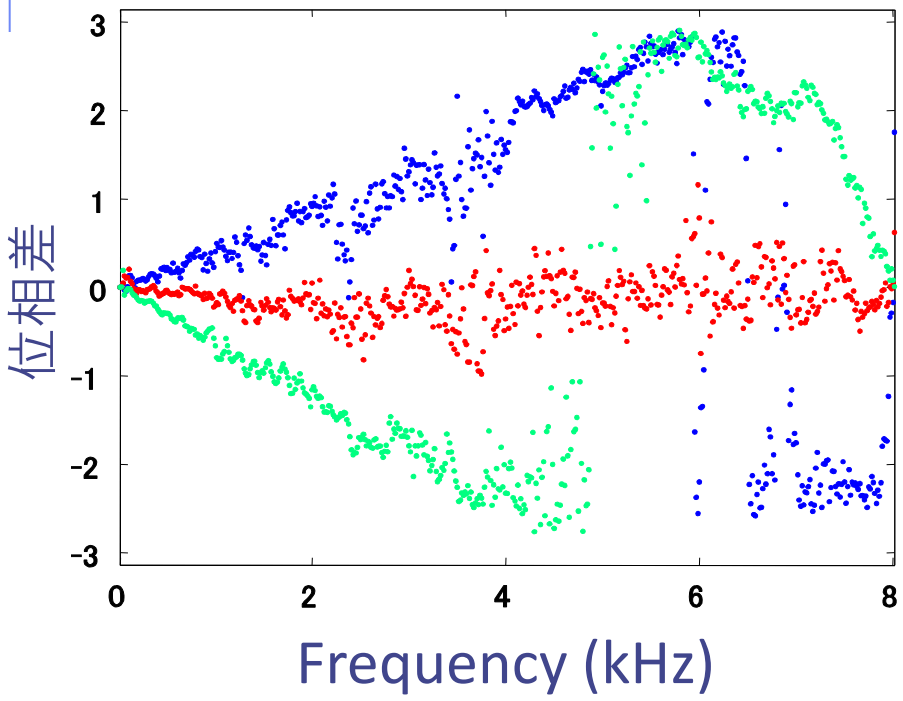
- 観測信号（3音源2マイク）
- 分離結果



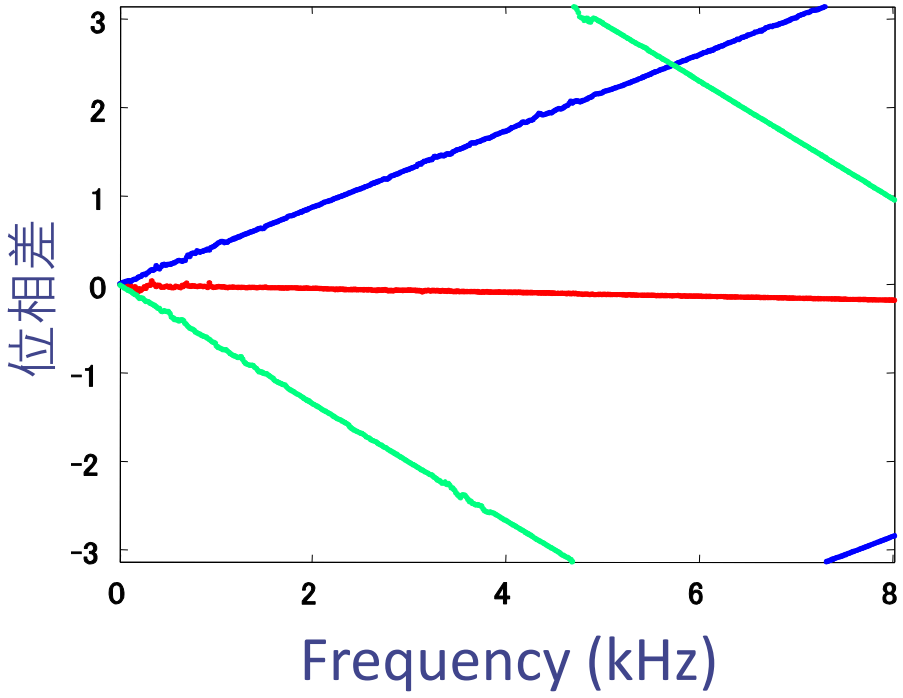
到来方向推定結果 [Kameoka 2012]

- パーミュテーションがうまく整合されていれば各音源の伝達周波数特性のチャンネル間位相差と周波数は直線関係になるはず

チャンネル間位相差（正解）



$\arg([\boldsymbol{m}_{k,\omega}]_2/[\boldsymbol{m}_{k,\omega}]_1)$



トークのアウトライン

- 音響信号処理
  - 多重音解析
  - ブラインド音声分離
- 音声情報処理
  - 音声イントネーション解析

各々の構成：

- 生成モデルの設計
- （推論アルゴリズム）
- 実験結果例

トークのアウトライン

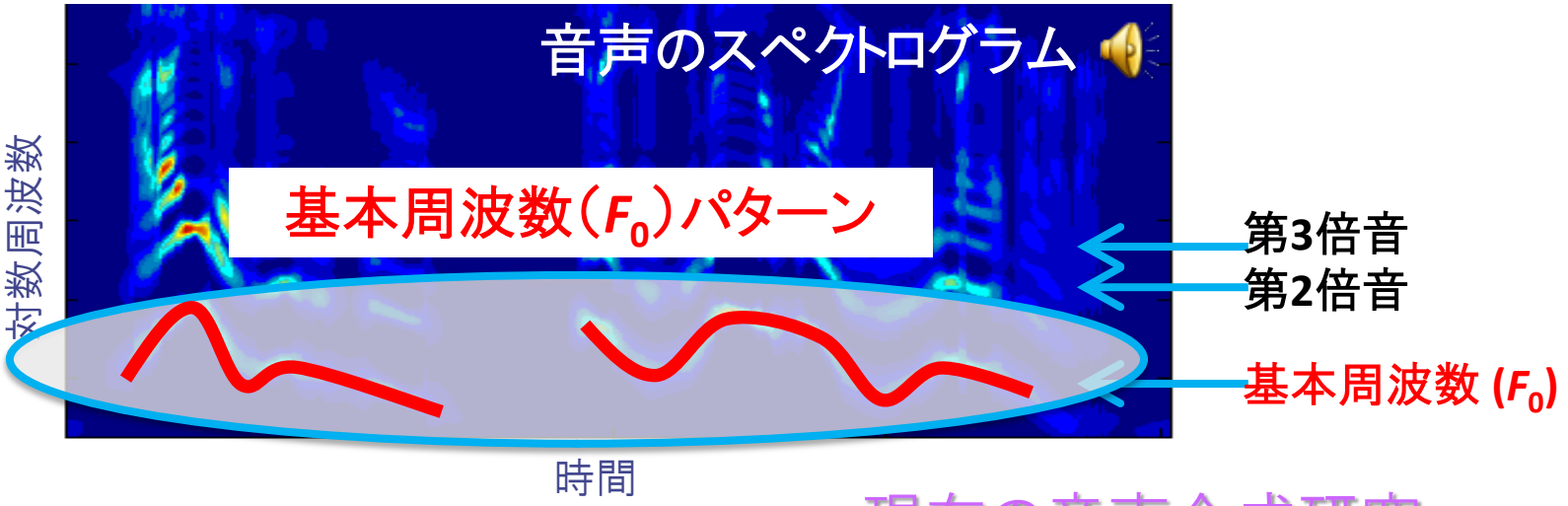
- 音響信号処理
  - 多重音解析
  - ブラインド音声分離
- 音声情報処理
  - 音声イントネーション解析

各々の構成：

- 生成モデルの設計
- （推論アルゴリズム）
- 実験結果例

音声イントネーション解析

- 音声イントネーション（基本周波数(F_0)パターン）に含まれる非言語情報（意図，構文，感情，焦点，言い回しの個性）を抽出するのが目的

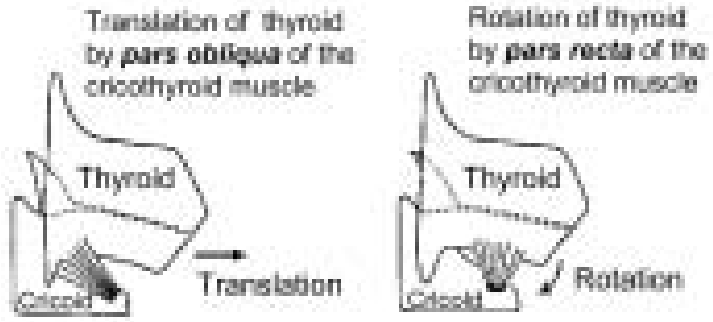


- 実応用場面
  - 自然／感情豊かな音声合成
  - 言い回しの分析･加工･再合成
  - 音声認識･感情認識･言語識別への利用

現在の音声合成研究では自然なイントネーションをいかに生成するかが重要課題の一つ

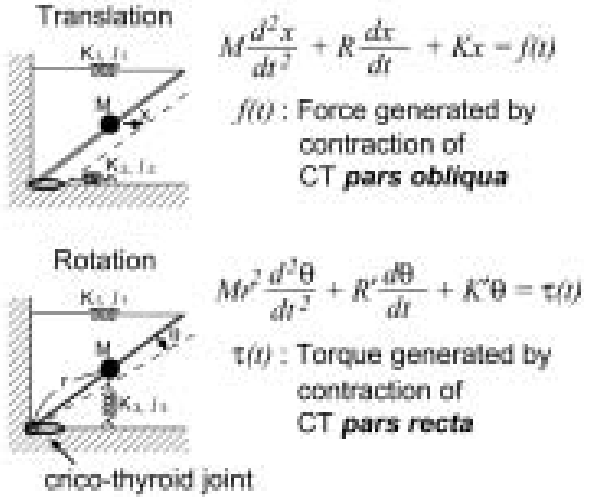
音声F_0パターンについて

- F_0パターンを構成する主要素
  - フレーズ：　句単位の比較的穏やかな音調変化
  - アクセント：　語または音節単位での比較的急激な音調変化
- 藤崎モデル [Fujisaki 1969]
  - F_0パターンを甲状軟骨の運動方程式に基づいて表現したモデル
  - フレーズ･アクセントが，直接パラメータとして表現される
  - 実測のF_0パターンに非常に良く合致 → 韻律研究の分野では有名



平行移動運動
（フレーズに対応）

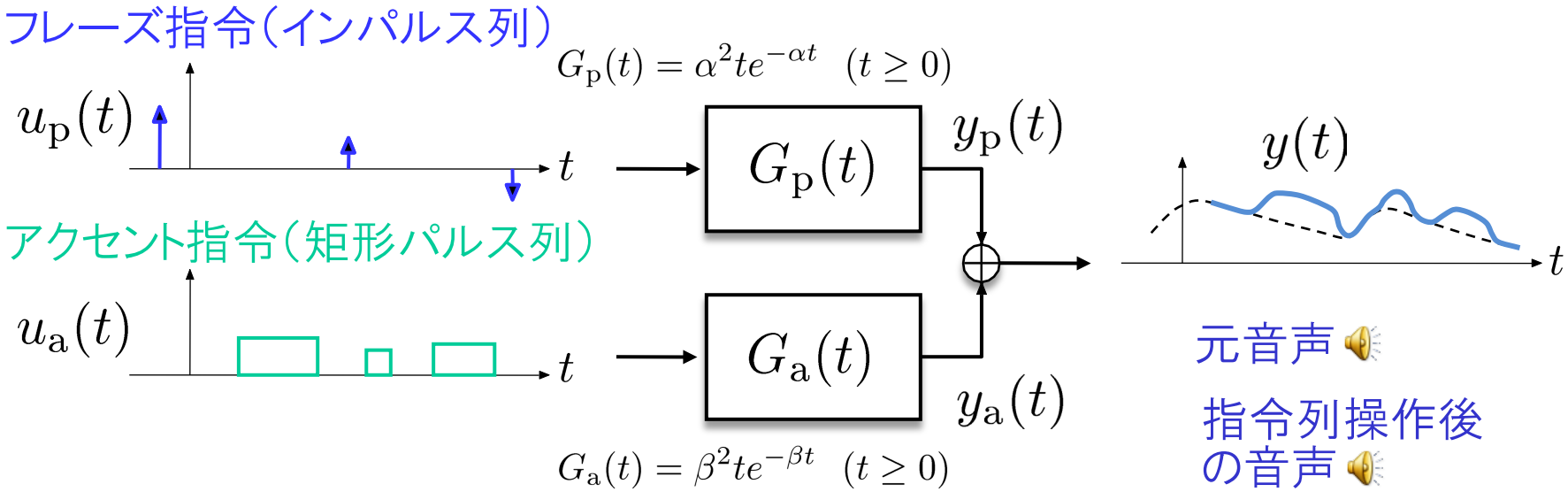
回転運動
（アクセントに対応）



“音声の音調的特徴のモデル化とその応用” [藤崎2005]

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



- 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）

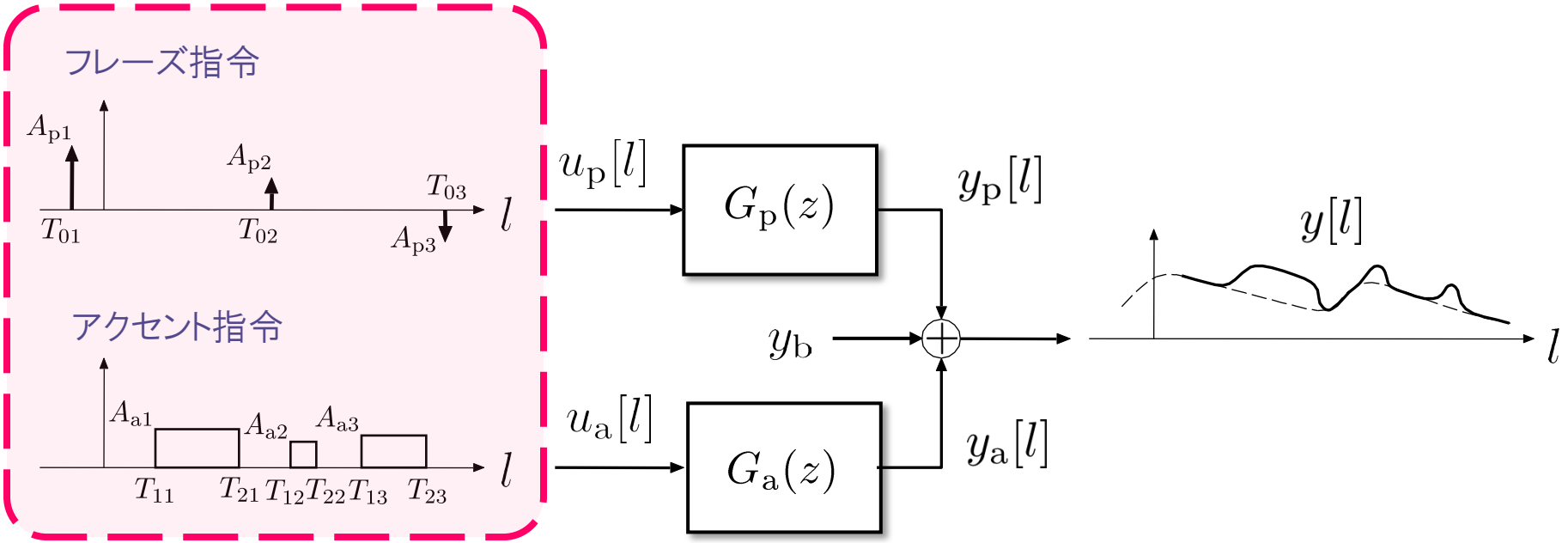
藤崎モデルをベースにF0パターンを生成モデル化

[Kameoka 2010]

- 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

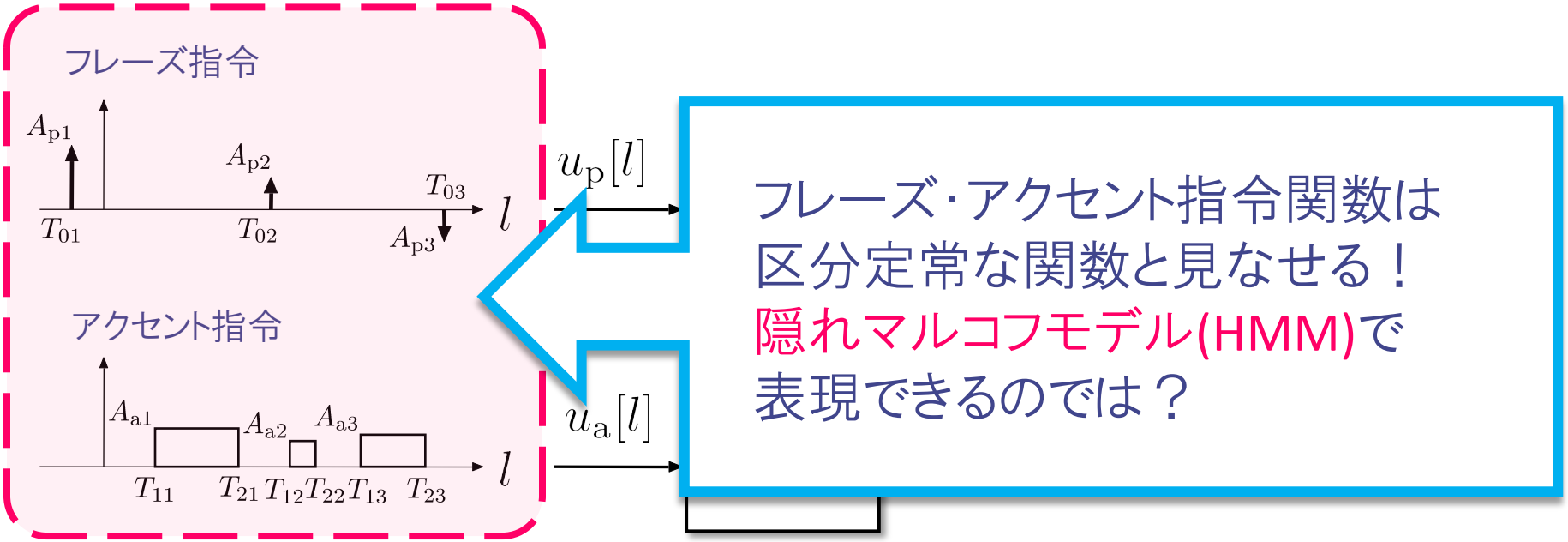
フレーズ・アクセント指令列の生成過程 [Kameoka 2010]

- 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・アクセント指令列の生成過程 [Kameoka 2010]

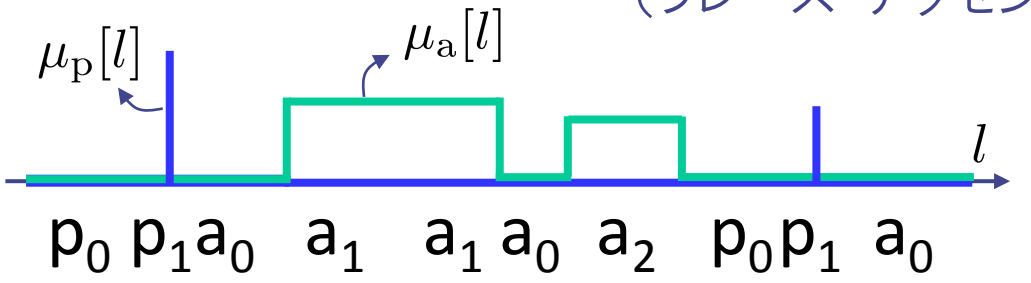
- フレーズ・アクセント指令列の制約
  - フレーズ指令はインパルス
  - アクセント指令は矩形パルス
  - 異なる2つの指令は同時刻に生起しない
  - 指令や指令間の間隔の持続長が「音声らしい」ものである
- フレーズ・アクセント指令列の生成過程の確率モデル化



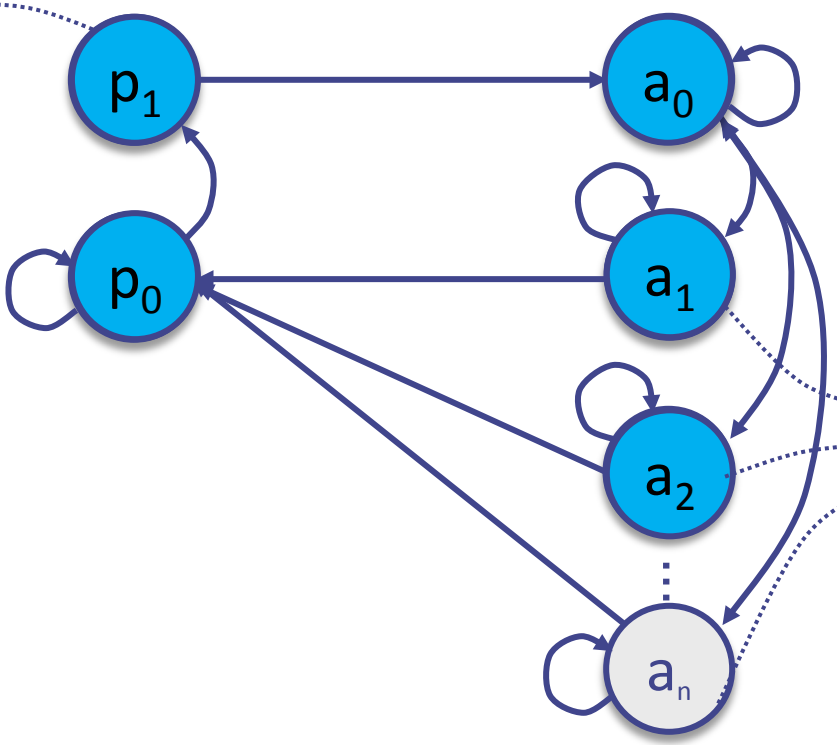
フレーズ･アクセント指令列生成HMM [Kameoka 2010]

状態遷移に伴う平均 $\mu_{\mathrm{p}}[l], \mu_{\mathrm{a}}[l]$ の時間軌跡

（フレーズ･アクセント指令列）



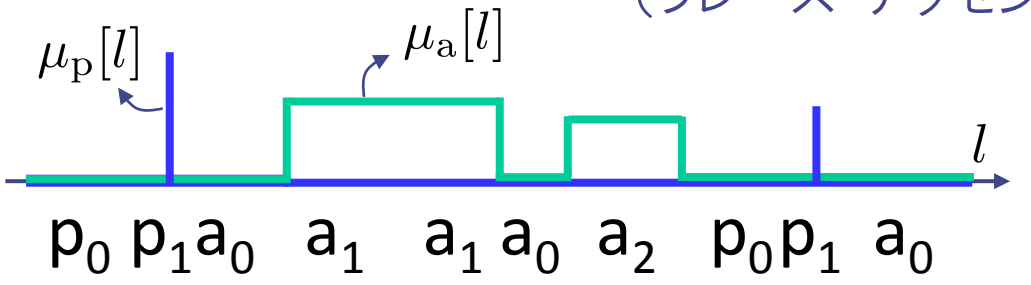
フレーズ指令を
出力する状態



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 [Kameoka 2010]

状態遷移に伴う平均 $\mu_{\mathrm{p}}[l], \mu_{\mathrm{a}}[l]$ の時間軌跡
（フレーズ・アクセント指令列）

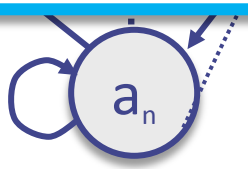


フレーズ指令を
出力する状態

- 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
- フレーズ・アクセント指令列には文法構造があるはず

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！

を



観測F0パターンの確率密度関数 [Kameoka 2010]

- 観測F0パターンは「HMM出力系列の畳み込み混合」

$$y[l] = G_{\mathrm{p}}[l] * u_{\mathrm{p}}[l] + G_{\mathrm{a}}[l] * u_{\mathrm{a}}[l] + y_{\mathrm{b}}$$

観測F_0 パターン　　隠れマルコフモデルの出力　　隠れマルコフモデルの出力

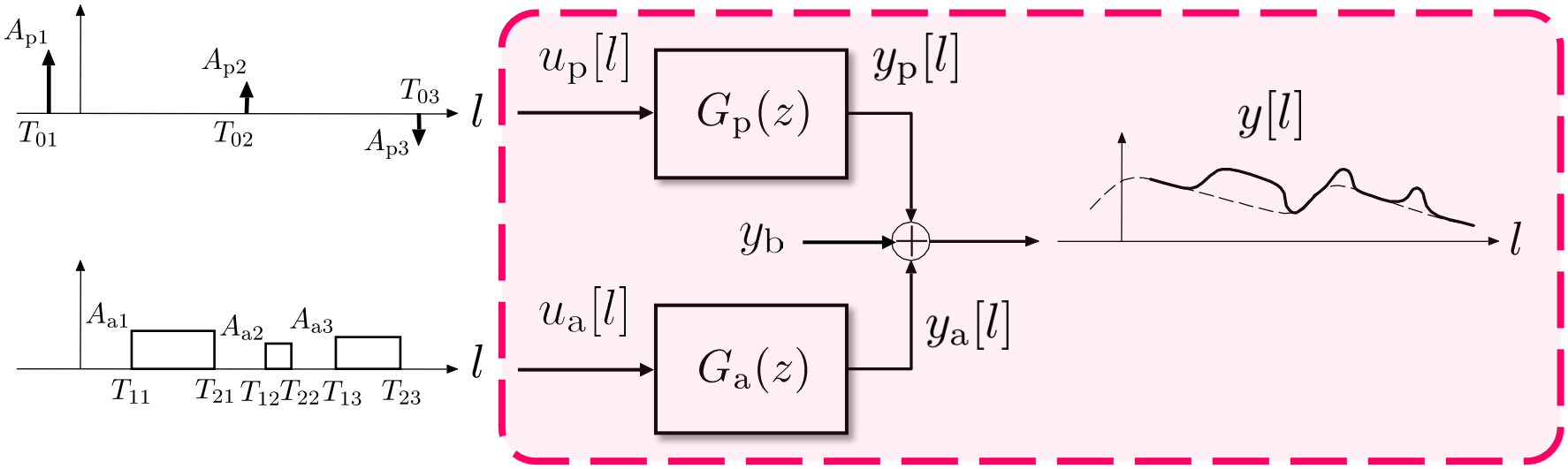
状態系列（＝指令列）

$p(Y|\theta)$

  - HMMの状態出力分布をガウス分布とすれば・・・

$$\boldsymbol{y} = \boldsymbol{G}_{\mathrm{p}}\boldsymbol{u}_{\mathrm{p}} + \boldsymbol{G}_{\mathrm{a}}\boldsymbol{u}_{\mathrm{a}} + y_{\mathrm{b}}\boldsymbol{1}$$

$$\sim \mathcal{N}(\boldsymbol{G}_{\mathrm{p}}\boldsymbol{\mu}_{\mathrm{p}} + \boldsymbol{G}_{\mathrm{a}}\boldsymbol{\mu}_{\mathrm{a}} + \mu_{\mathrm{b}}\boldsymbol{1},\ \boldsymbol{G}_{\mathrm{p}}\boldsymbol{\Sigma}_{\mathrm{p}}\boldsymbol{G}_{\mathrm{p}}^{\mathsf{T}} + \boldsymbol{G}_{\mathrm{a}}\boldsymbol{\Sigma}_{\mathrm{a}}\boldsymbol{G}_{\mathrm{a}}^{\mathsf{T}} + \boldsymbol{\Sigma}_{\mathrm{b}})$$



アクセント・フレーズ指令列の最大事後確率推定

[Kameoka 2010]

- 観測F0パターン Y が与えられた下で $P(\theta|Y) \propto P(Y|\theta)P(\theta)$ 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 $y_{\mathrm{p}}[l]$，アクセント成分 $y_{\mathrm{a}}[l]$，ベースライン成分 $y_{\mathrm{b}}[l]$ を完全データと見なすとEM法が適用できる！

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{y}_{\mathrm{p}} \\ \boldsymbol{y}_{\mathrm{a}} \\ \boldsymbol{y}_{\mathrm{b}} \end{bmatrix} \sim \mathcal{N}\left(\underbrace{\begin{bmatrix} \boldsymbol{G}_{\mathrm{p}}\boldsymbol{\mu}_{\mathrm{p}} \\ \boldsymbol{G}_{\mathrm{a}}\boldsymbol{\mu}_{\mathrm{a}} \\ \mu_{\mathrm{b}}\boldsymbol{1} \end{bmatrix}}_{\boldsymbol{m}}, \underbrace{\begin{bmatrix} \boldsymbol{G}_{\mathrm{p}}\boldsymbol{\Sigma}_{\mathrm{p}}\boldsymbol{G}_{\mathrm{p}}^{\mathsf{T}} & O & O \\ O & \boldsymbol{G}_{\mathrm{a}}\boldsymbol{\Sigma}_{\mathrm{a}}\boldsymbol{G}_{\mathrm{a}}^{\mathsf{T}} & O \\ O & O & \boldsymbol{\Sigma}_{\mathrm{b}} \end{bmatrix}}_{\boldsymbol{\Lambda}} \right)$$

  - “不完全データ” $\boldsymbol{y}$ と完全データの関係： $\boldsymbol{y} = \overbrace{[\boldsymbol{I}\ \boldsymbol{I}\ \boldsymbol{I}]}^{\boldsymbol{H}} \begin{bmatrix} \boldsymbol{y}_{\mathrm{p}} \\ \boldsymbol{y}_{\mathrm{a}} \\ \boldsymbol{y}_{\mathrm{b}} \end{bmatrix}$
  - Q関数

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

EMアルゴリズムによるパラメータ推定 [Kameoka 2010]

- E-step

藤崎モデル

$$\mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\theta}] = \boldsymbol{m} + \boldsymbol{\Lambda}\boldsymbol{H}^{\mathrm{T}}(\boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^{\mathrm{T}})^{-1}(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{m})$$

$$\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}|\boldsymbol{y};\boldsymbol{\theta}] = \boldsymbol{\Lambda} - \boldsymbol{\Lambda}\boldsymbol{H}^{\mathrm{T}}(\boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^{\mathrm{T}})^{-1}\boldsymbol{H}\boldsymbol{\Lambda} + \mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\theta}]\mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\theta}]^{\mathrm{T}}$$

- M-step

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

を Θ に関して最大化

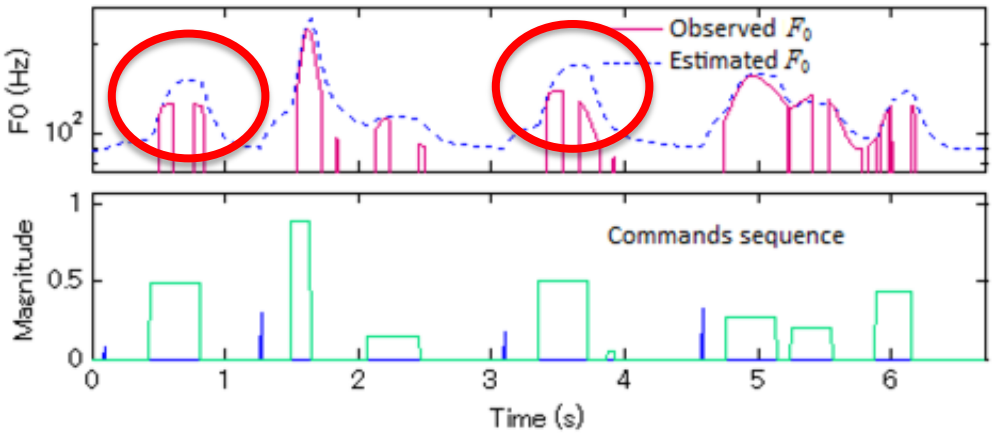
- 状態遷移系列（推定指令列）の更新は Viterbiアルゴリズムで実現可能

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。」

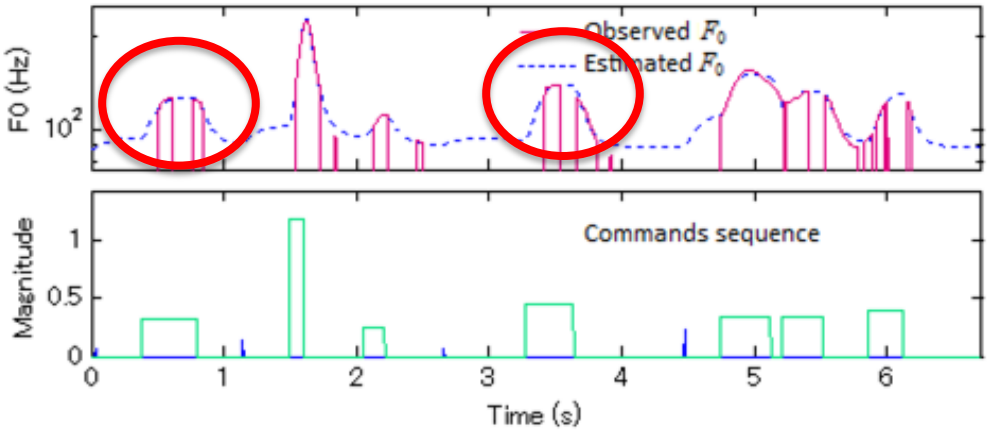
従来法の推定結果

“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

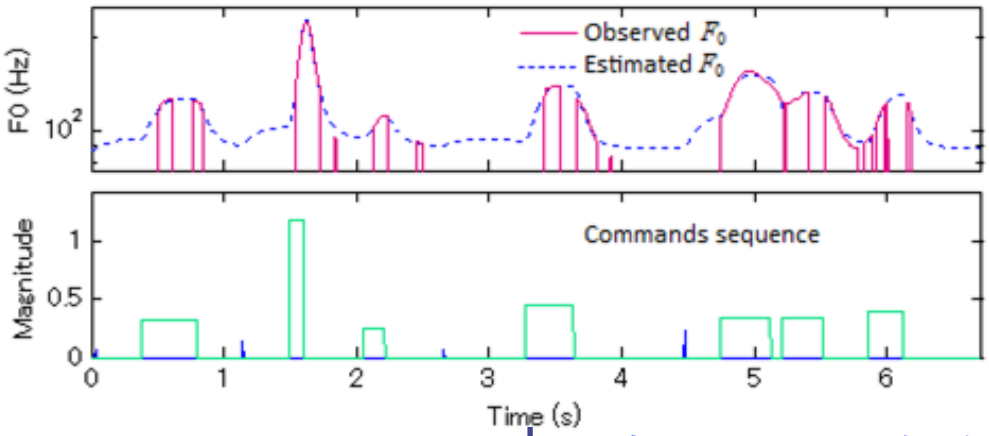
(推定指令列から合成したF_0パターンを用いて再合成した音声)



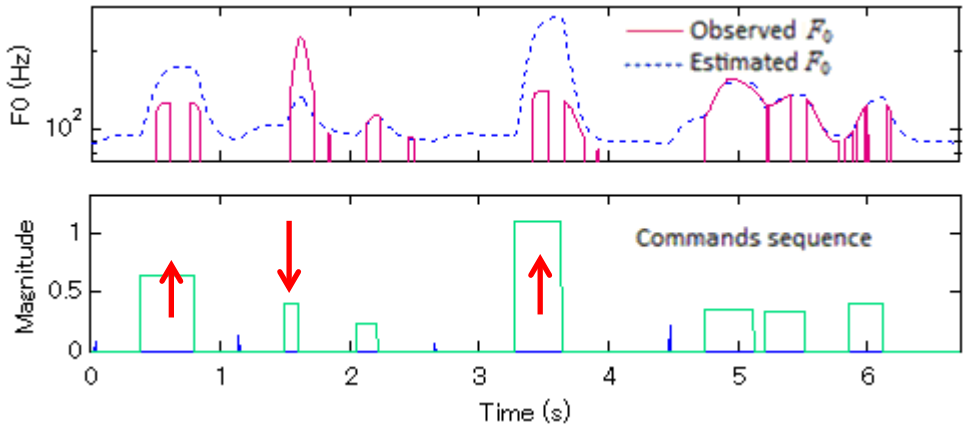
指令列の推定結果例 [Yoshizato 2012]

- 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

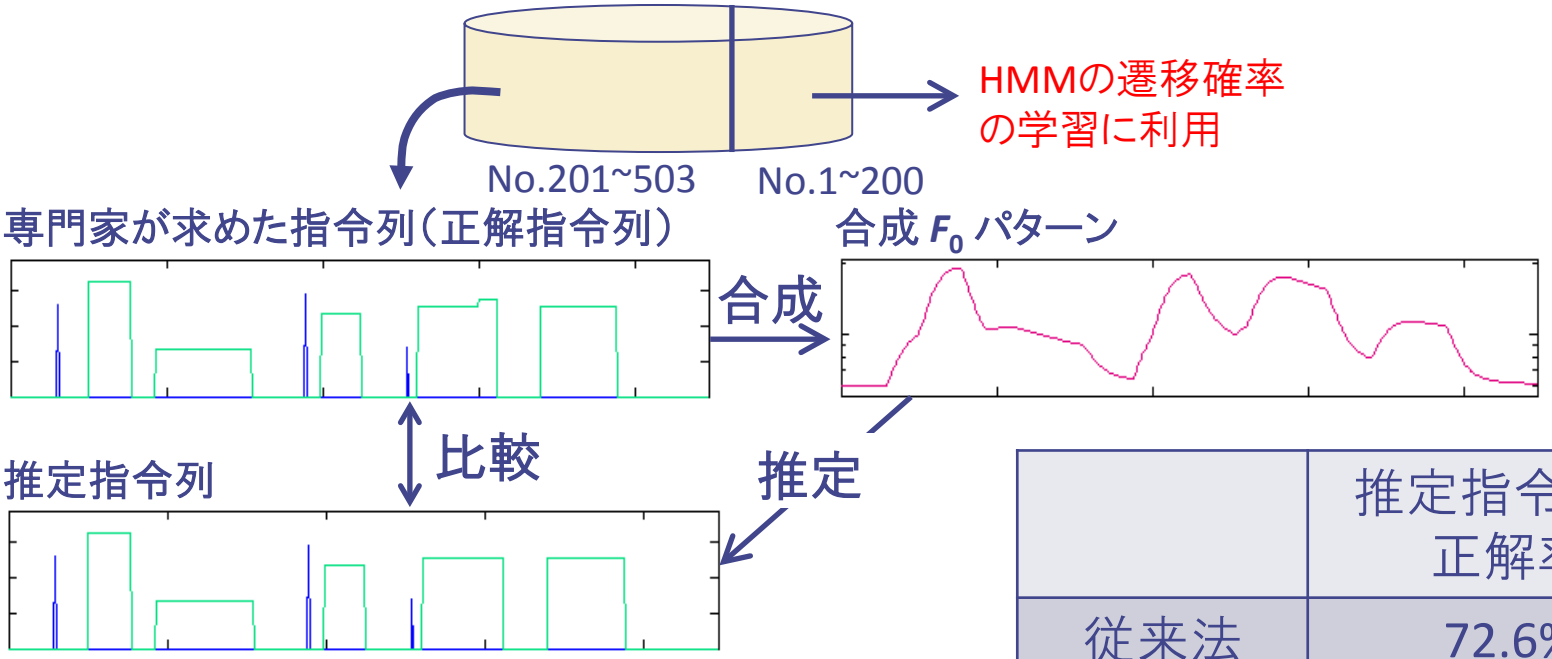


(操作後の指令列から再合成した音声)

合成F_0パターンに対する指令列推定精度評価

[Yoshizato 2012]

ATR日本語音声データベースBセット話者MHT（全503文）
＋ 韻律研究の専門家が求めた指令列のデータ

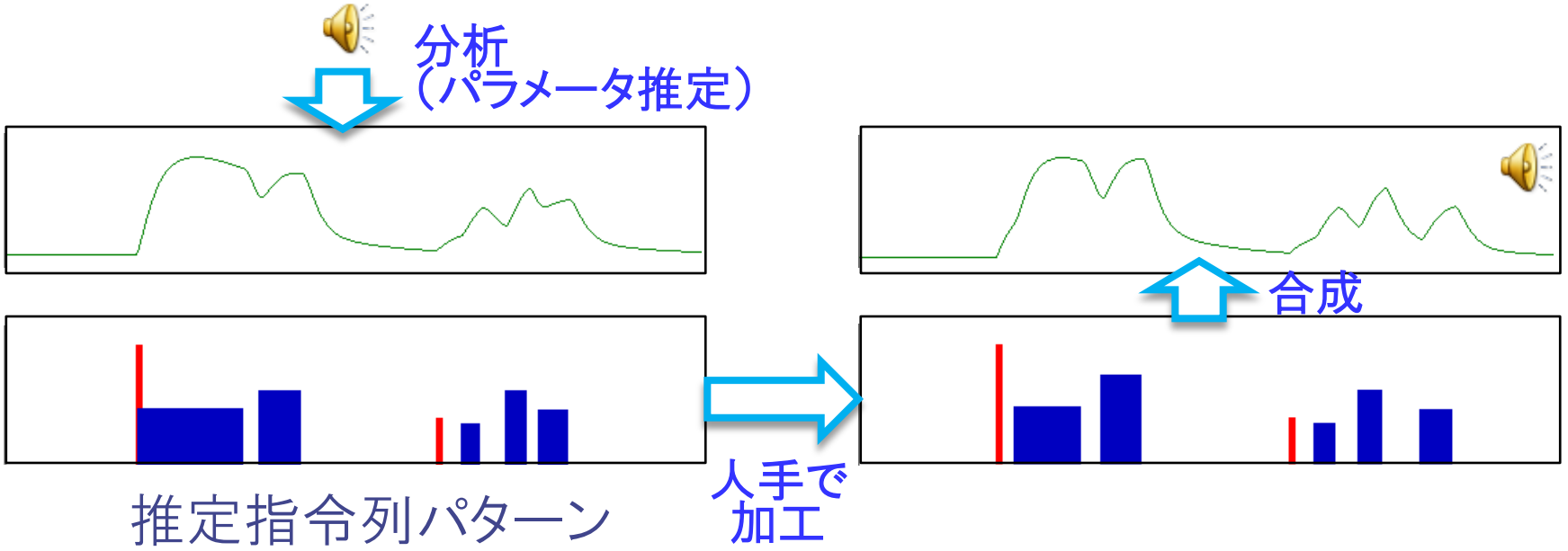


	推定指令列の正解率
従来法	72.6%
提案手法	**83.4%**

- 逆解析の性能で従来法を上回った
  - 推定アルゴリズムがうまく局所解を回避した可能性
  - 指令列の統計的傾向を遷移確率として推定に組み込んだことによって，音声として自然な解が選ばれた可能性

おまけのデモ

- 標準語風イントネーションを関西弁風に変換



まとめ

- 音声・音響信号処理問題の多くは不良設定の逆問題
  - にもかかわらず人間が難なく出来るのは，ボトムアップな推定とトップダウンな推定を統合的に行っているため
  - 生成モデルアプローチが有効なアプローチになりうる
    - 生成過程のモデル化により尤度関数設計
    - 物理制約や経験則により事前分布設計
    - 推論アルゴリズムの導出
- 3つの研究事例を上記の観点で紹介
  - 音響信号処理　→　多重音解析、ブラインド音源分離
  - 音声情報処理　→　音声イントネーション解析

付録：EMアルゴリズムについて

- θに関してp(Y|θ)を最大化する問題

目的関数　　潜在変数

$$\log p(Y|\theta) = \log \int p(Y, X|\theta) \mathrm{d}X$$

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

$$\geq \int \lambda(Y, X) \log \frac{p(Y, X|\theta)}{\lambda(Y, X)} \mathrm{d}X$$

←- - - 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

$$\lambda(Y, X) \leftarrow \frac{p(Y, X|\theta)}{\int p(Y, X'|\theta) \mathrm{d}X'} = p(X|Y, \theta)$$

Mステップ：

θに関して補助関数を最大化するステップ

$$\theta \leftarrow \underset{\theta}{\mathrm{argmax}} \int \lambda_{Y,X} \log \frac{p(Y, X|\theta)}{\lambda(Y, X)} \mathrm{d}X$$