特別企画 MIRU×KIKU
「音学シンポジウム」連携オーガナイズドセッション

音響信号の分解と再構成

亀岡弘和

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

自己紹介

亀岡弘和(かめおかひろかず)

- 略歴:
  - 2007 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 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 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　客員准教授
  - 2015 NTTコミュニケーション科学基礎研究所
特別研究員
  - 2016 国立情報学研究所　客員准教授
- 専門:
  - 音声・音楽などの音響信号を対象とした信号処理や機械学習
  - 『音や声に含まれる潜在情報の分析合成変換』,『時空間音響情景分析』
- 学会活動:
  - 2013年にMIRUをヒントにした「音学シンポジウム」を企画

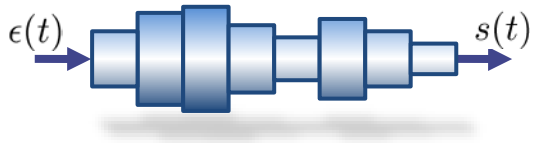
自己紹介

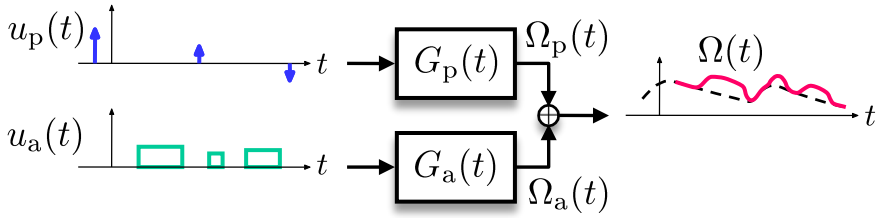
亀岡弘和(かめおかひろかず)

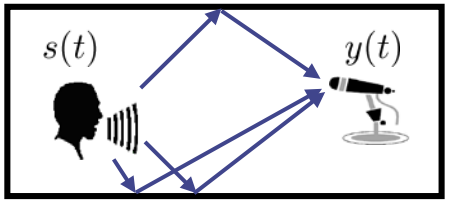
- 略歴:
  - 2007 東京大学大学院情報理工学系研究科 システム情報学専攻　博士課程修了
  - 2007 日本電信電話株式会社入社 NTTコミュニケーション科学基礎研究所配属
  - 2011 東京大学大学院情報理工学系研究科 システム情報学専攻　客員准教授
  - 2015 NTTコミュニケーション科学基礎研究所 特別研究員
  - 2016 国立情報学研究所　客員准教授
- 専門:
  - 音声・音楽などの音響信号を対象とした信号処理や機械学習
  - 『音や声に含まれる潜在情報の分析合成変換』,『時空間音響情景分析』
- 学会活動:
  - 2013年にMIRUをヒントにした「音学シンポジウム」を企画

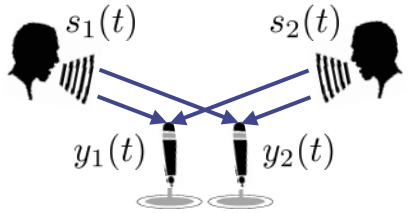
音響信号の分解と再構成

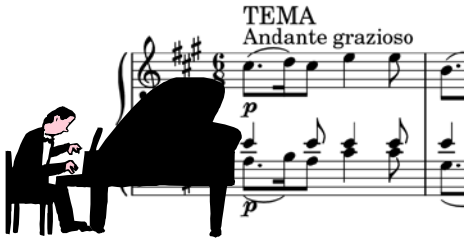
- 音の分解問題
  - 音声のソース･フィルタ分解
  - 音声基本周波数パターンの
    フレーズ･アクセント分解
  - 雑音･残響除去
  - 音源分離

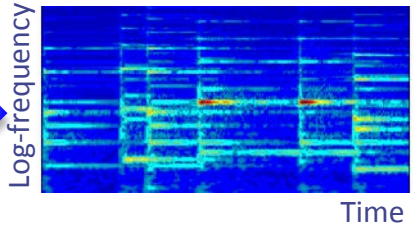












- 分解再構成
  - 不要な成分を除いて再構成　⇒　**目的音の強調**
  - 分解した成分の一部を加工して再構成　⇒　**音の加工**

アウトライン

- 物理モデリング×時系列モデリング×確率モデリングによる音響信号の分解再構成
  - 音声のソース･フィルタ分解
  - 残響除去
  - 音源分離
  - 基本周波数パターンのフレーズ･アクセント分解

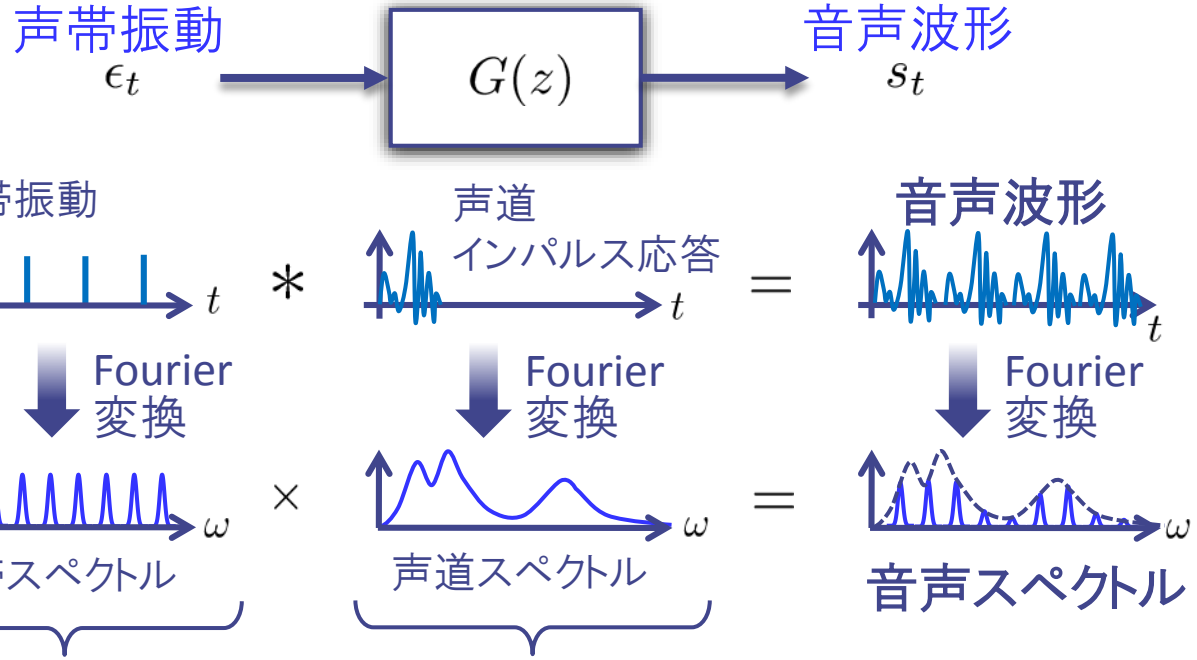
アウトライン

- 物理モデリング×時系列モデリング×確率モデリングによる音響信号の分解再構成
  - 音声のソース･フィルタ分解
  - 残響除去
  - 音源分離
  - 基本周波数パターンのフレーズ･アクセント分解

線形予測分析によるソース･フィルタ分解 [Itakura+1968]

■動機：符号化への応用

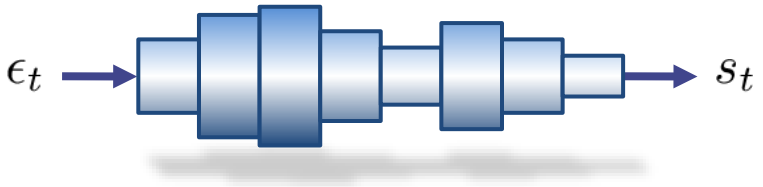
- 音声信号を少ないパラメータで表現したい



白色ノイズと仮定

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化

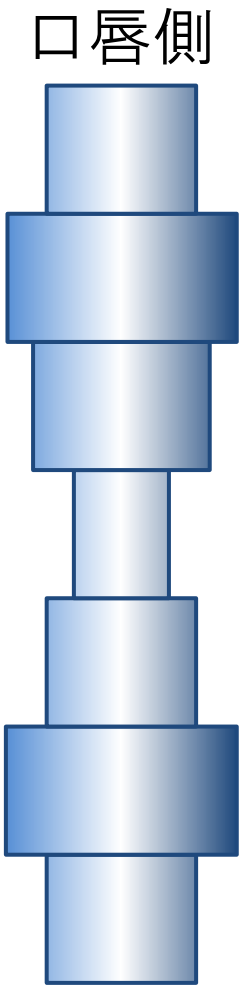


この下でのソース･フィルタ分解

⇒**線形予測分析**

線形予測分析の物理的解釈

- Levinson-Durbin再帰式
  - *m*次の最適予測係数 から *m*+1次の最適予測係数を導く再帰式

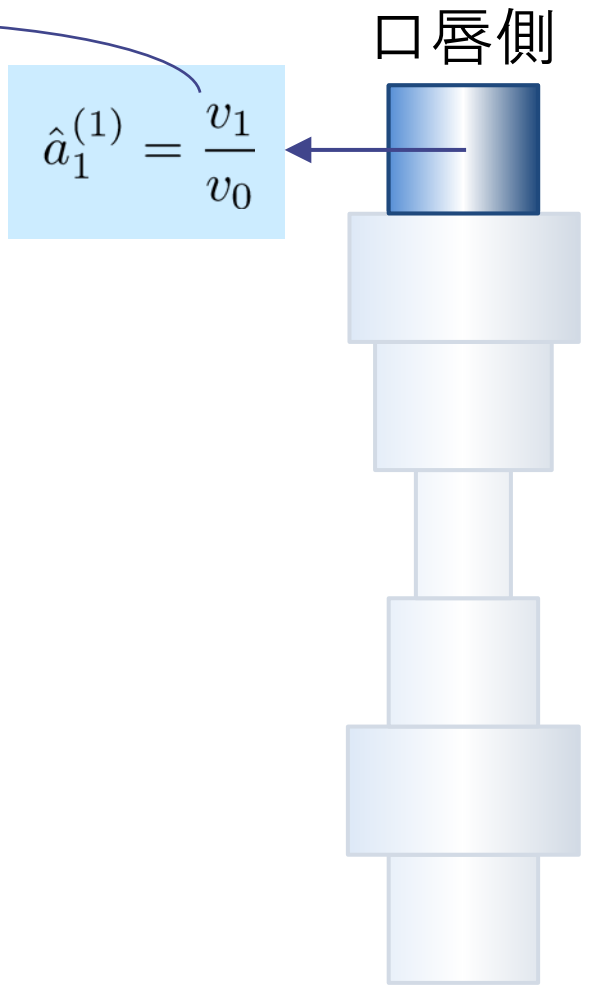


線形予測分析の物理的解釈

- Levinson-Durbin再帰式
  - *m*次の最適予測係数 から *m*+1次の最適予測係数を導く再帰式

自己相関関数 $v_q = \sum_t s_t s_{t-q}$

$\hat{a}_1^{(1)} = \frac{v_1}{v_0}$



線形予測分析の物理的解釈

- Levinson-Durbin再帰式
  - *m*次の最適予測係数 から *m*+1次の最適予測係数を導く再帰式

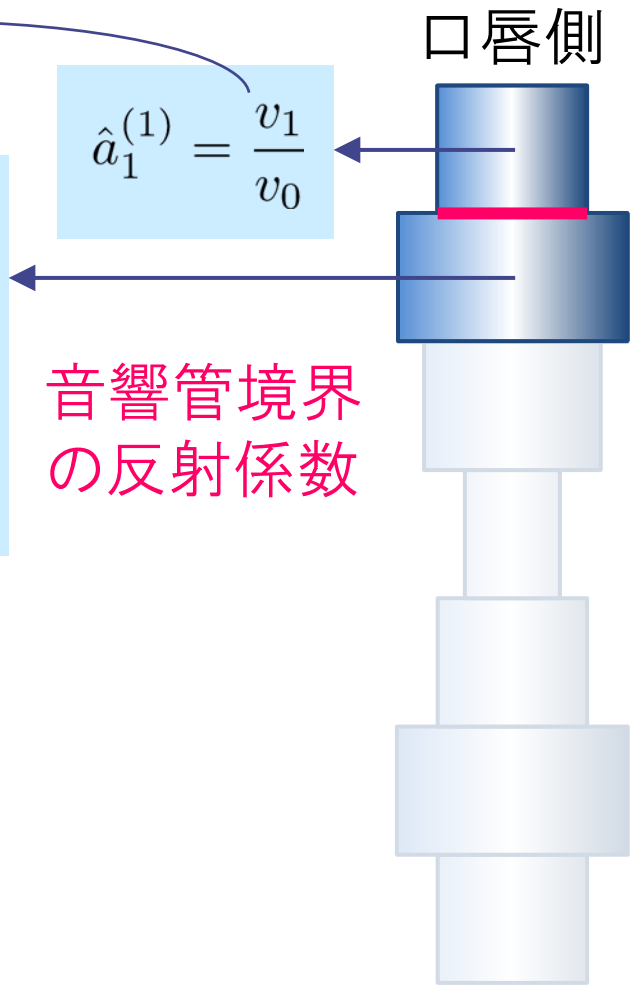
自己相関関数 $v_q = \sum_t s_t s_{t-q}$

$$\hat{a}_1^{(1)} = \frac{v_1}{v_0}$$

$$\hat{a}_1^{(2)} = \hat{a}_1^{(1)} - K_1 \hat{a}_1^{(1)}$$

$$\hat{a}_2^{(2)} = K_1$$

$$K_1 = \frac{v_2 - \hat{a}_1^{(1)} v_1}{v_0 - \hat{a}_1^{(1)} v_1}$$



線形予測分析の物理的解釈

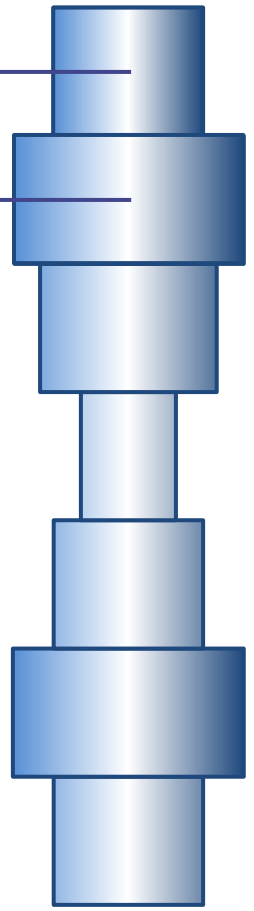
- Levinson-Durbin再帰式
  - *m*次の最適予測係数 から *m*+1次の最適予測係数を導く再帰式

自己相関関数 $v_q = \sum_t s_t s_{t-q}$

$$\hat{a}_1^{(1)} = \frac{v_1}{v_0}$$

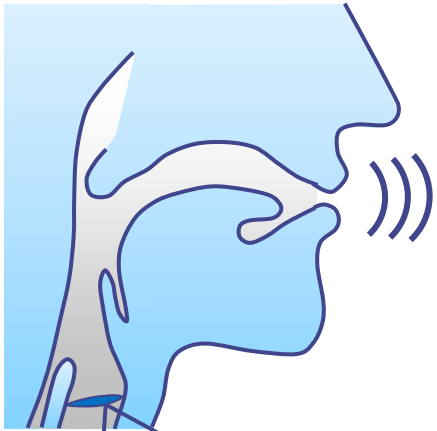
$$\hat{a}_1^{(2)} = \hat{a}_1^{(1)} - K_1 \hat{a}_1^{(1)}$$
$$\hat{a}_2^{(2)} = K_1$$
$$K_1 = \frac{v_2 - \hat{a}_1^{(1)} v_1}{v_0 - \hat{a}_1^{(1)} v_1}$$

$$\hat{a}_p^{(m+1)} = \hat{a}_p^{(m)} - K_m \hat{a}_{m-p+1}^{(m)} \quad (p = 1, \ldots, m)$$
$$K_m = \frac{v_{m+1} - \sum_{p=1}^{m} \hat{a}_p^{(m)} v_{m-p+1}}{v_0 - \sum_{p=1}^{m} \hat{a}_p^{(m)} v_p}$$

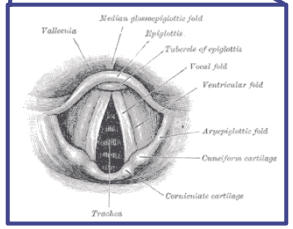


線形予測分析によるソース･フィルタ分解

■音声の分析と合成

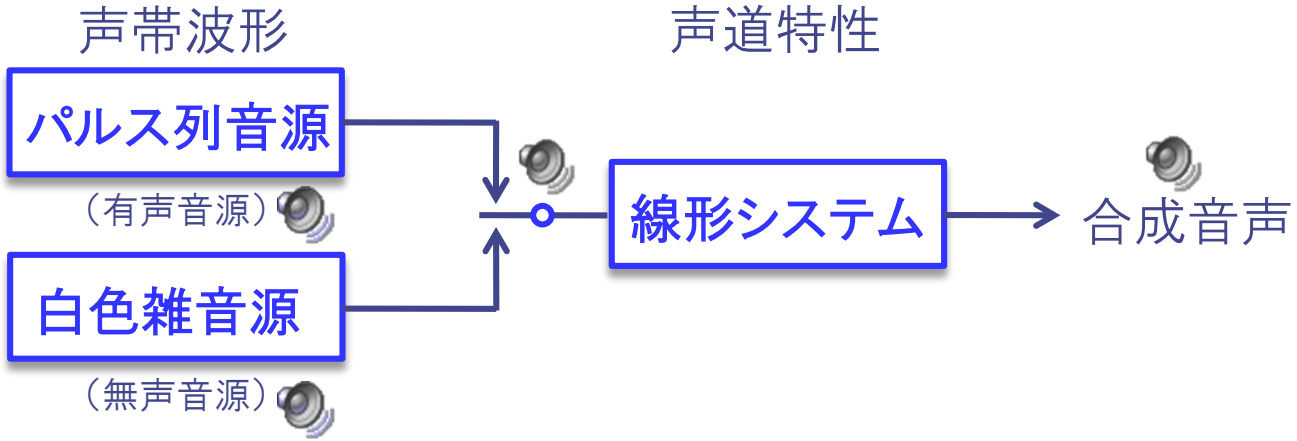


声帯



- ソースフィルタモデル（音声生成過程モデル）

分析：音声波形 を声帯波形と声道特性に分解
合成：推定した声帯情報と声道情報から元音声を再現

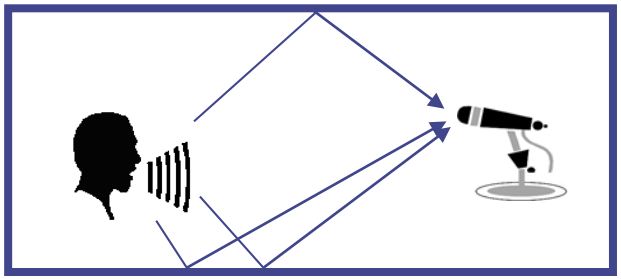


ソース･フィルタ分解の応用

- 音声符号化
  - 線形予測分析は現在も携帯電話やVoIPの標準圧縮方式
- 音声認識／話者認識
  - 「声道フィルタ」は音韻情報(「あ」、「い」、・・・)を担う
  - MFCC(メル周波数ケプストラム係数)
- 音声合成／音声変換
  - 声帯音源と声道フィルタを特徴量とした回帰問題
  - 線形予測分析からSTRAIGHT分析、メル一般化ケプストラム分析へ

ブラインド残響除去

- 室内で収音した音声信号には残響が不可避的に混入
  ⇒音声認識などの音声処理システムの性能劣化の要因
- 話者が移動しないなら残響重畳プロセスは線形時不変システムで記述される



$$s(t) * h(t) = y(t)$$

原音声信号（未知）

室内インパルス応答（未知）

収音信号

➡ 原音声信号も室内インパルス応答も未知の状況で、収音信号のみから原音声信号をいかに推定する問題

アウトライン

- 物理モデリング×時系列モデリング×確率モデリングによる音響信号の分解再構成
  - 音声のソース･フィルタ分解
  - 残響除去
  - 音源分離
  - 基本周波数パターンのフレーズ･アクセント分解

アウトライン

- 物理モデリング×時系列モデリング×確率モデリングによる音響信号の分解再構成
  - 音声のソース･フィルタ分解
  - 残響除去
  - 音源分離
  - 基本周波数パターンのフレーズ･アクセント分解

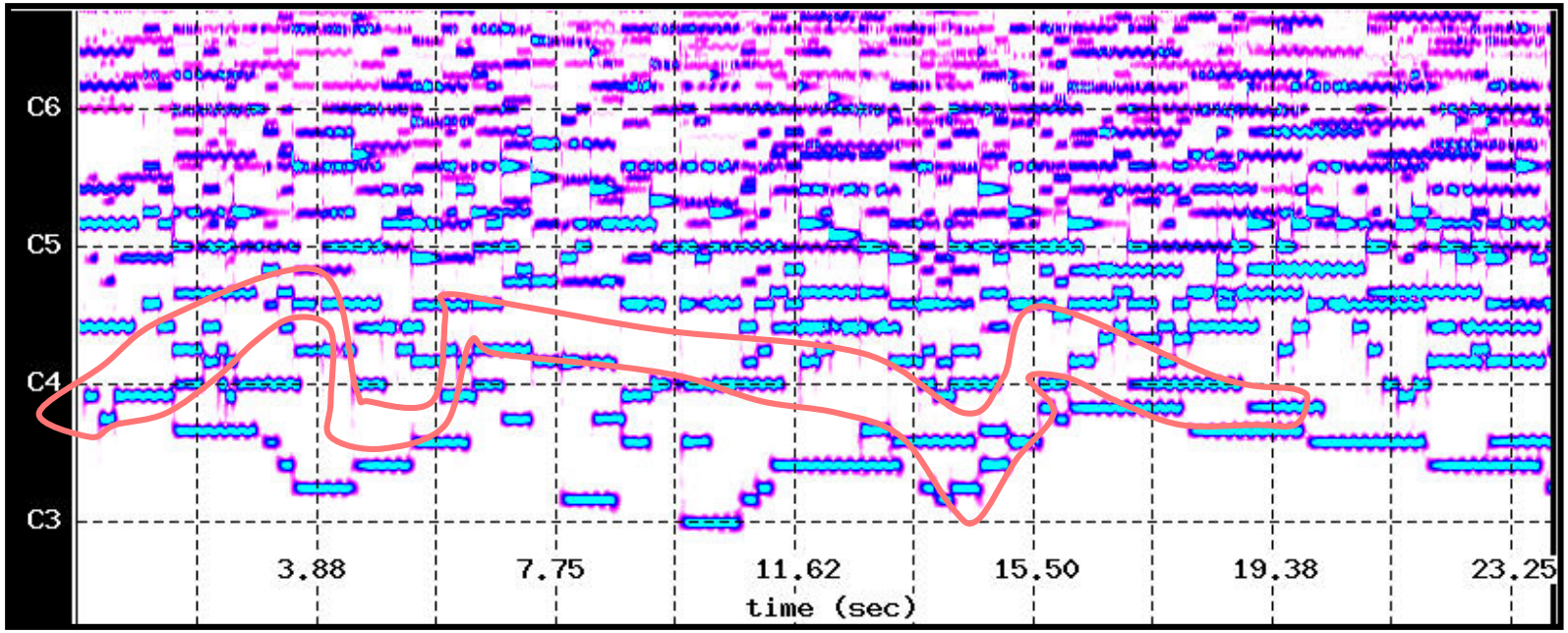
音源分離

- 複数の音が混ざった混合信号からそれぞれの音を分離する問題
  - 音声認識の前処理，ロボット聴覚，楽音信号加工，音声強調，・・・
- 一般に混ざったものから元の成分を復元するのは困難
  - 7 + 3 = 10 を計算するのは簡単だが、X+Y=10のみからXとYを一意に決めることはできない
  - 何らかの仮定が必要
- 問題設定
  - モノラル音源分離： 音源の性質や傾向が手がかりとなる
  - 多チャンネル音源分離： 上記に加え空間情報も手がかりとして使える

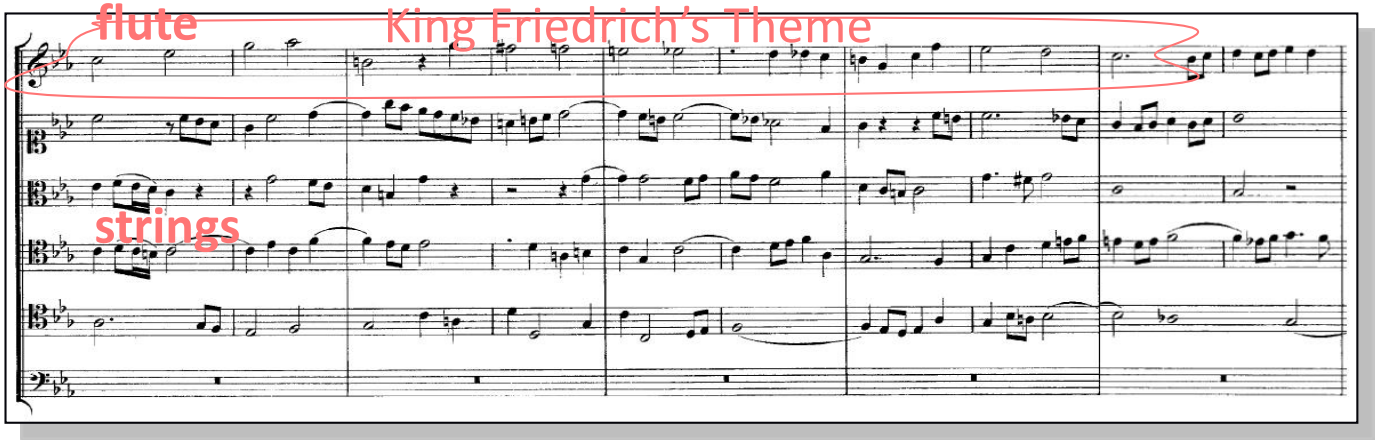
モノラル音源分離の難しさ

スペクトログラム（時間周波数表現）

周波数

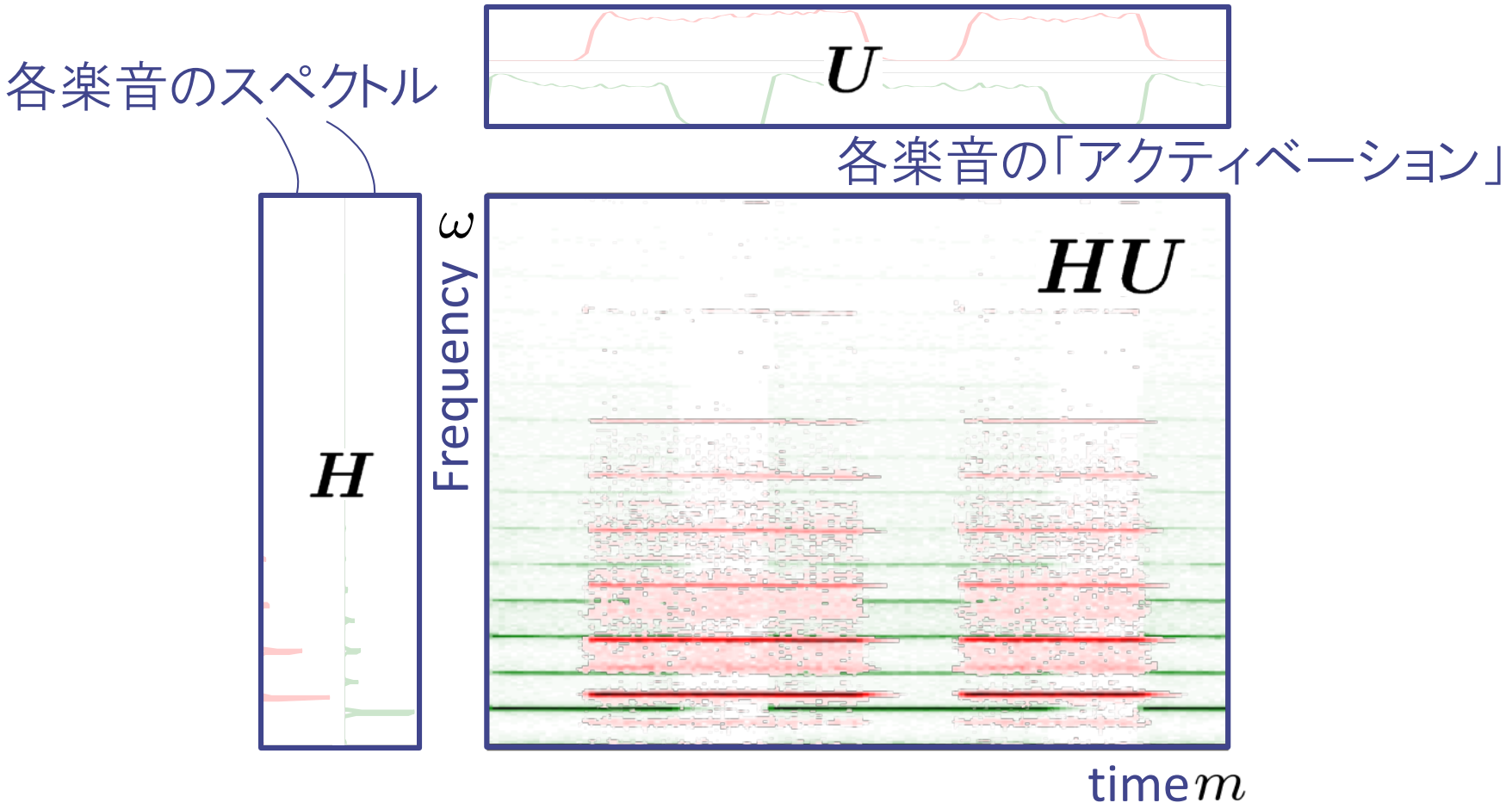


時間



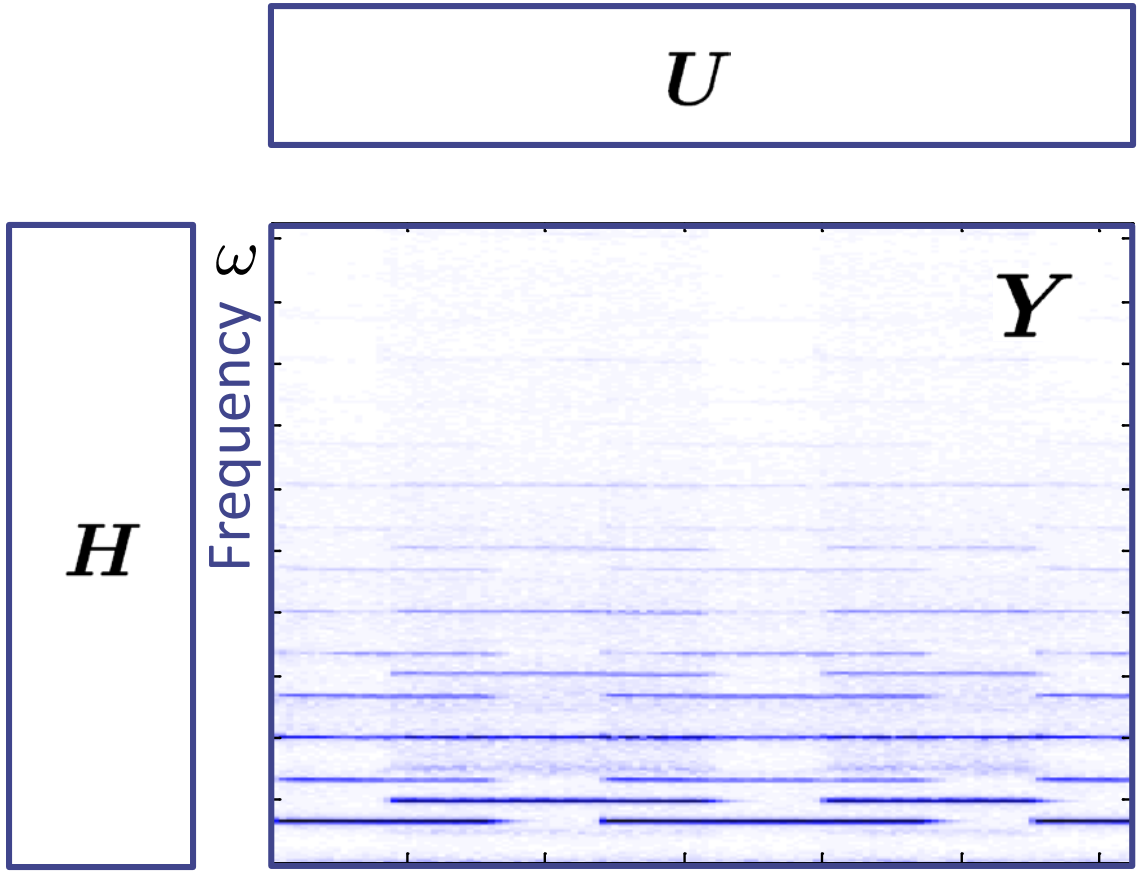
非負値行列因子分解 (NMF) 法 [Smaragdis+2003]

- 「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) 法 [Smaragdis+2003]

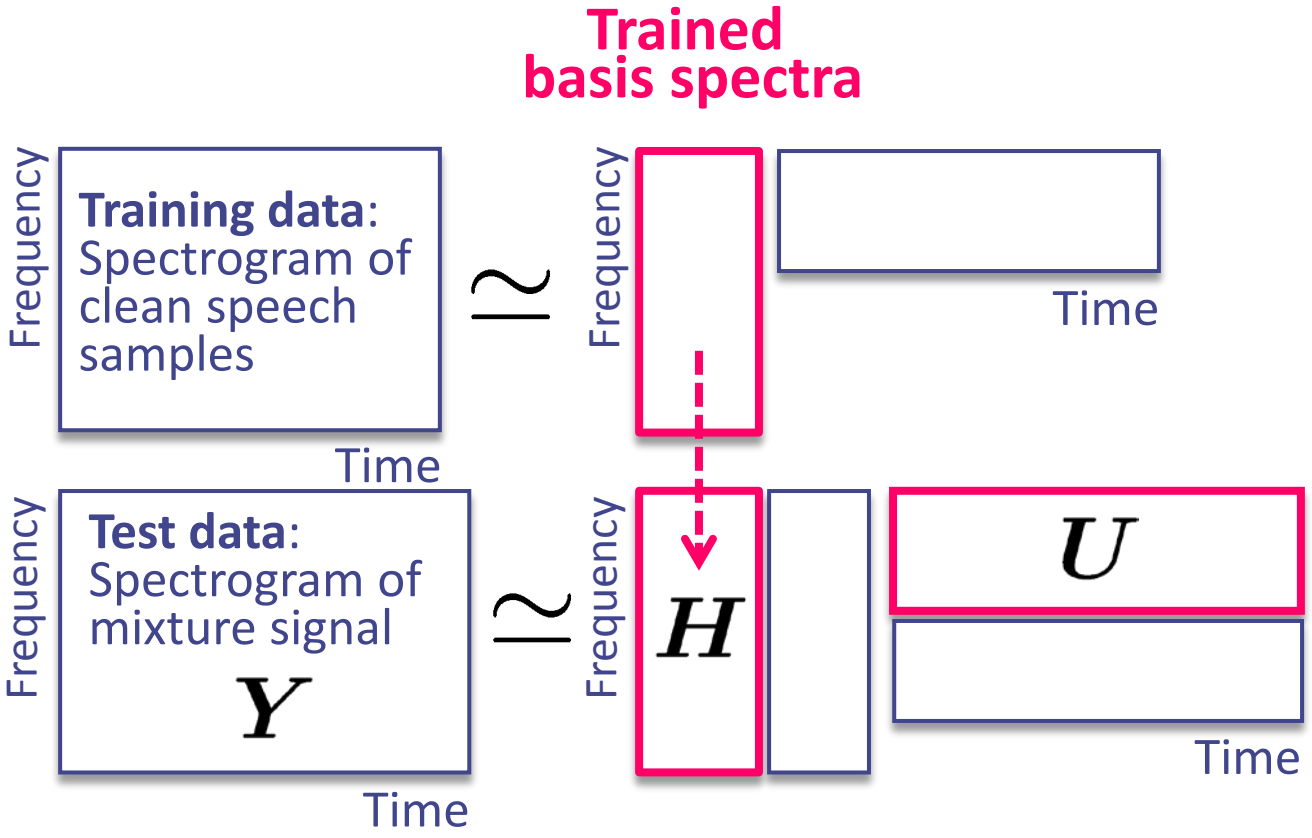
- 行列分解（逆問題）は音源分離に相当



※NMFは画像処理分野からの輸入技術

NMF法による音源分離

- 半教師ありNMF [Smaragdis+2007]
  - クリーンな音声信号のサンプルから基底スペクトルを事前学習
  - 事前学習した基底スペクトルを用いて混合音のスペクトログラムに対しNMFを適用



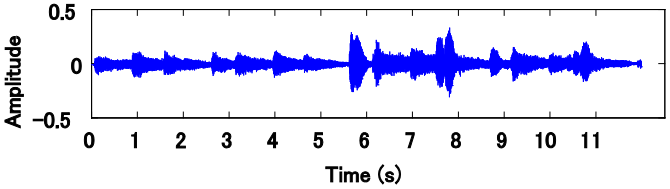
NMF法におけるスペクトログラムモデリングの問題点

- 振幅スペクトルの加法性を仮定（位相の影響を無視）

$y(u)$ 時間領域信号

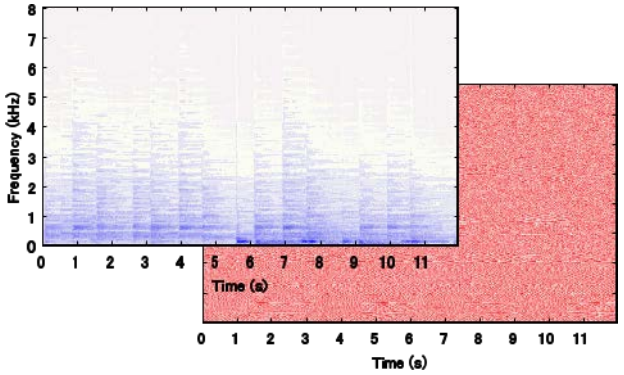
↓ 時間周波数変換 → 線形

$$Y_{\omega,t} = \langle y(u), \psi_{\omega,t}(u) \rangle_{u \in \mathbb{R}}$$



$Y_{\omega,t}$ 複素スペクトログラム

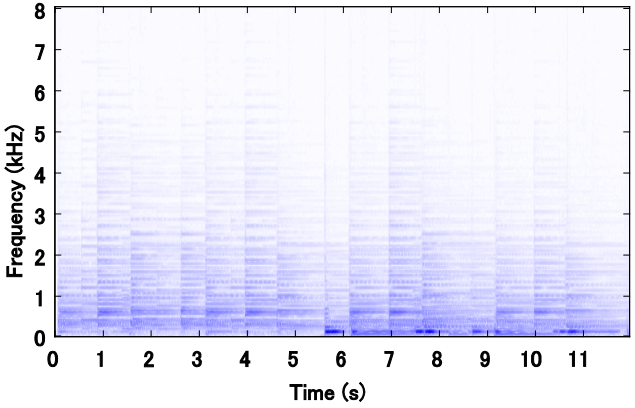
↓ 絶対値計算 → 非線形



$|Y_{\omega,t}|$ 振幅スペクトログラム

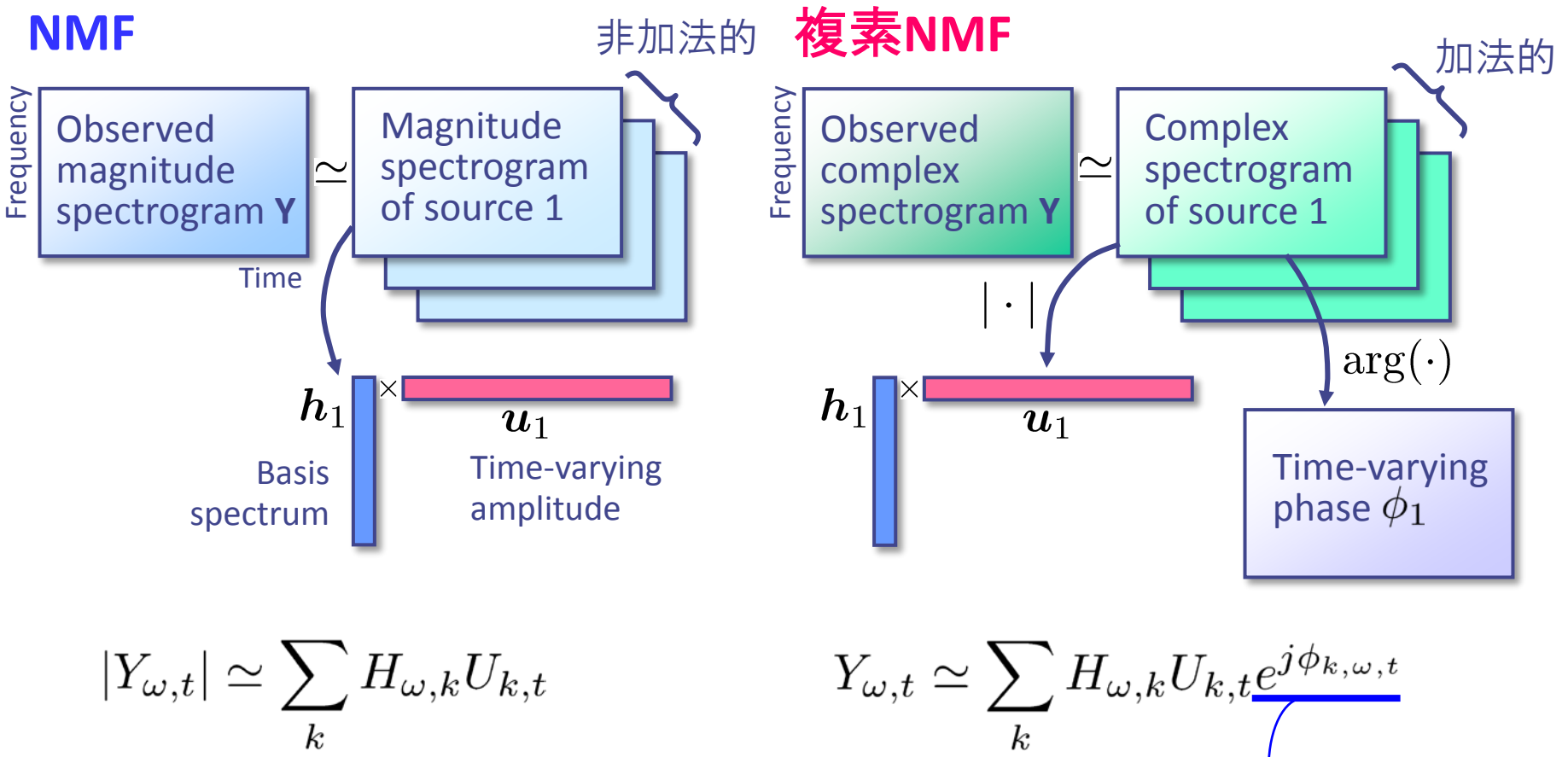
$$|Y_{\omega,t}| = \left| \sum_l X_{l,\omega,t} \right| \neq \sum_l |X_{l,\omega,t}|$$

加法性不成立



複素NMF [Kameoka+2008][Kameoka2015]

■位相を考慮したNMFの複素拡張



$$|Y_{\omega,t}| \simeq \sum_k H_{\omega,k} U_{k,t}$$

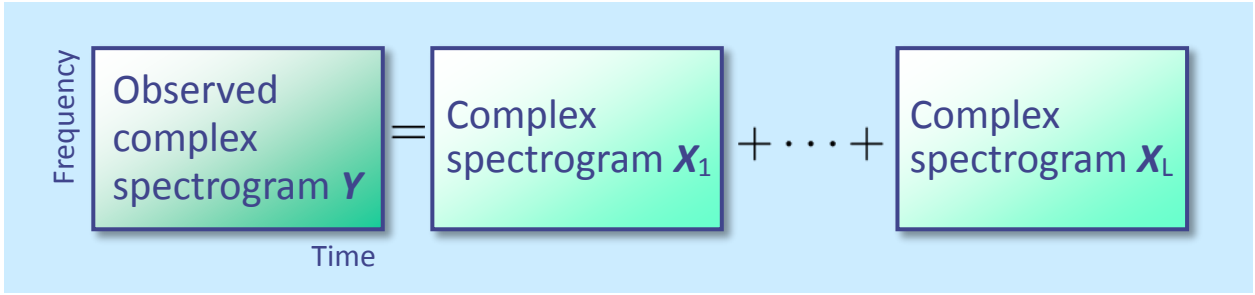
$$Y_{\omega,t} \simeq \sum_k H_{\omega,k} U_{k,t} e^{j\phi_{k,\omega,t}}$$

足す前に
位相スペクトルを付与
（時変な所がポイント）

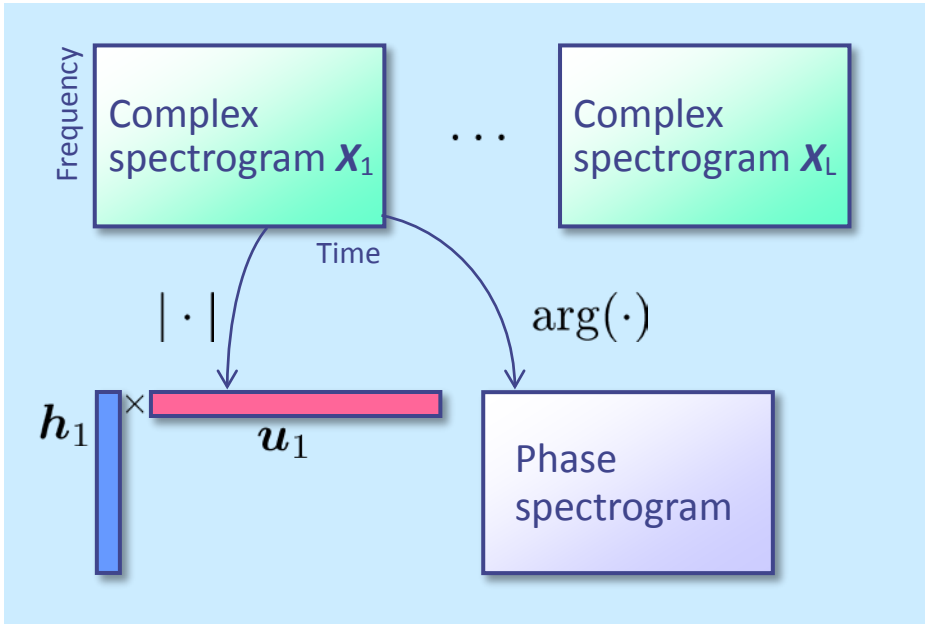
複素NMF [Kameoka+2008][Kameoka2015]

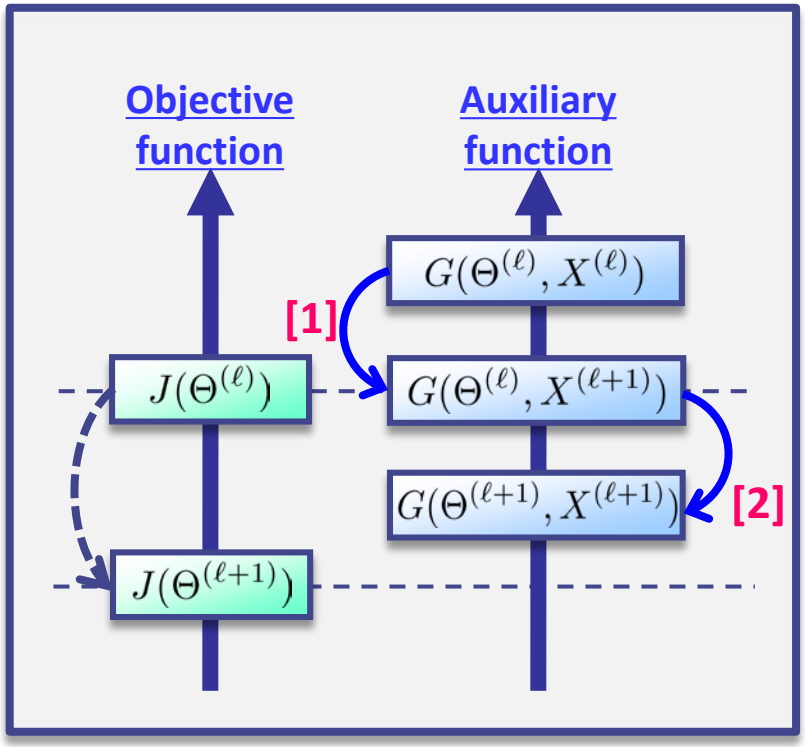
パラメータ推定アルゴリズム

- Step 1) $\boldsymbol{Y}$ の $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_L$ への分解

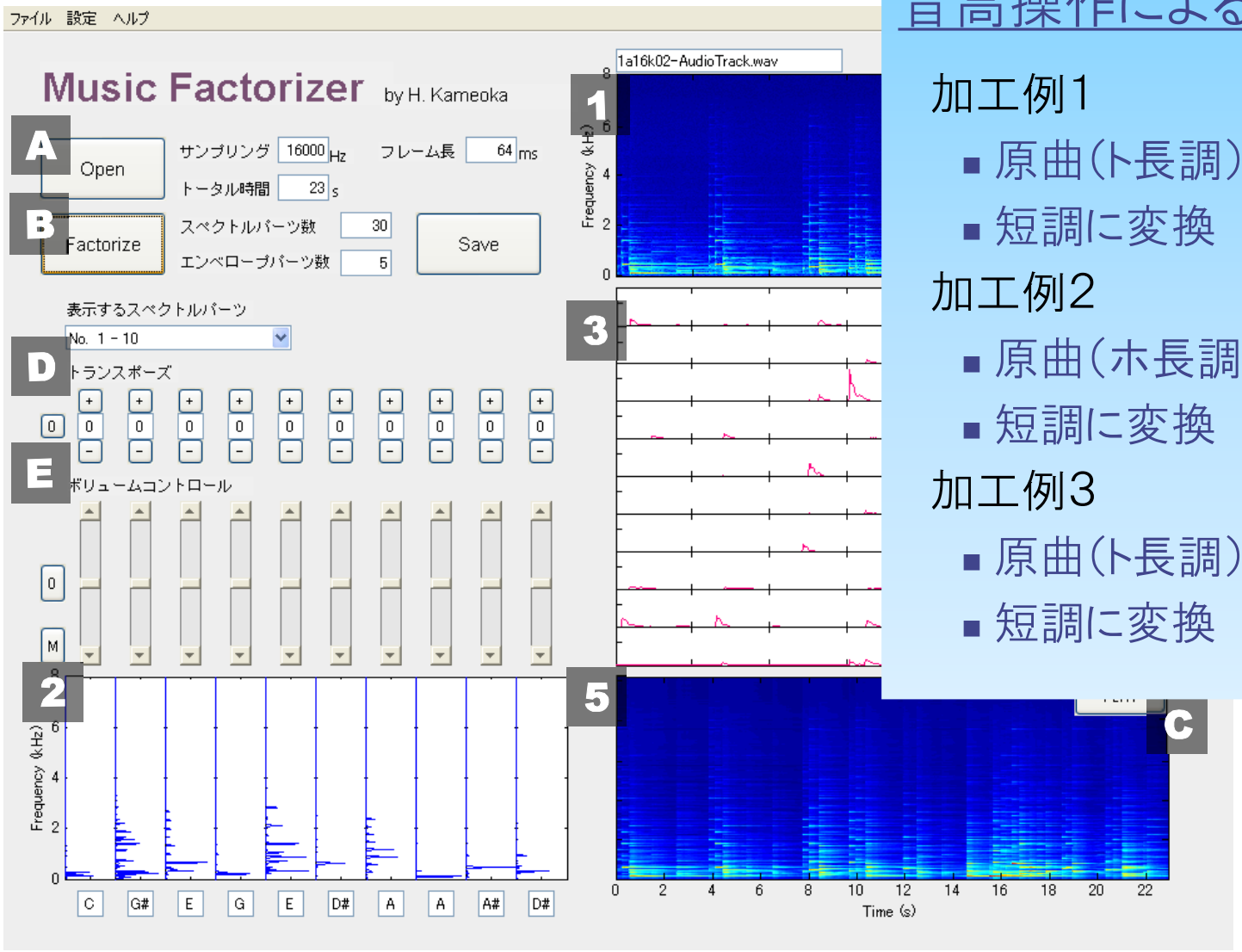


- Step 2) $\boldsymbol{\Theta}=\{\boldsymbol{H}, \boldsymbol{U}, \boldsymbol{\varphi}\}$ の更新





複素NMFによる楽音（曲調）加工



音高操作による調変換

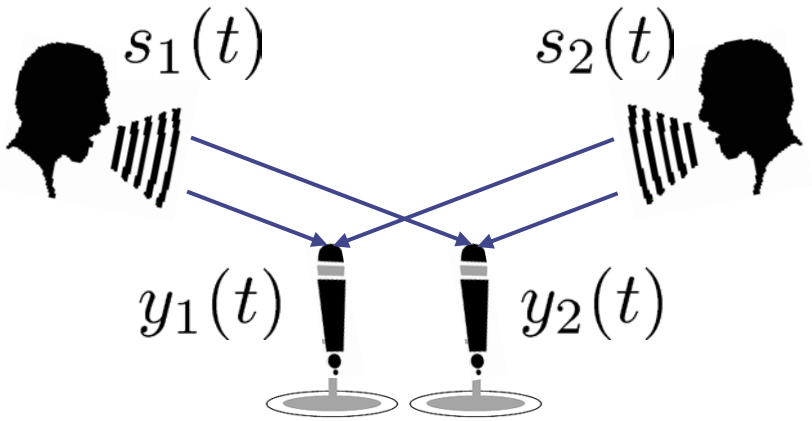
加工例1
- 原曲（ト長調）
- 短調に変換

加工例2
- 原曲（ホ長調）
- 短調に変換

加工例3
- 原曲（ト長調）
- 短調に変換

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」

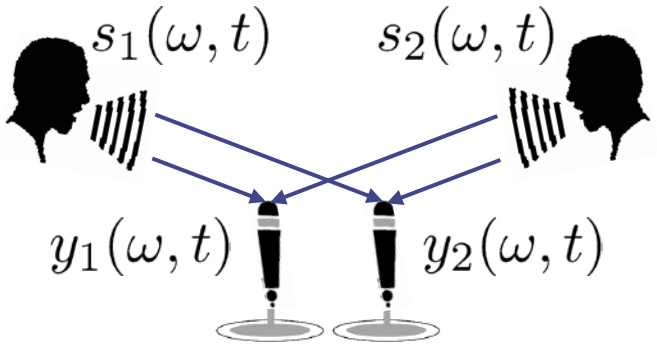


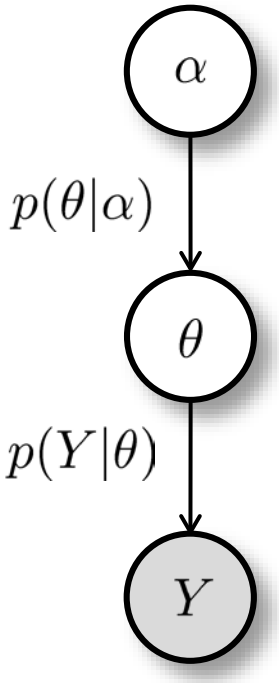
- 応用場面
  - 音を使った監視システム
    - どこで何が起こっているかを検知
    - 介護や防犯のための安全モニタリングとしての応用
      Cf. ShotSpotter（米国で開発されている発砲事件の検知システム）
  - 聴覚障害補助
  - ロボット聴覚
  - テレビ会議システム

ブラインド音源分離 (BlindSourceSeparation)

- 独立ベクトル分析／多チャンネルNMF

[Ono 2011, Ozerov 2010, Kameoka+2010,
Sawada+2012, Kitamura 2015, ...]





$\sigma_k^2(\omega,t)$ パワースペクトログラム

$p(S|\sigma^2)$

$s_k(\omega,t)$ 音源信号

$a_k(\omega)$ アレー応答

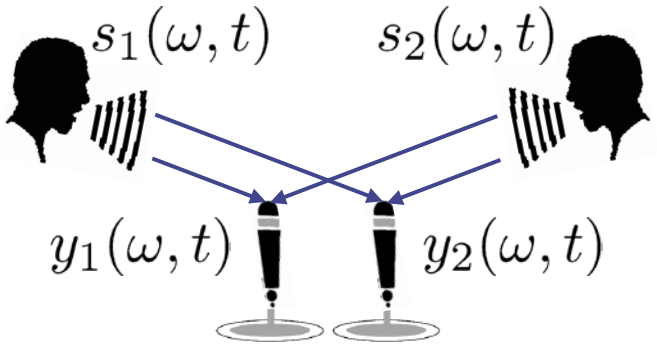
$p(Y|S,A)$

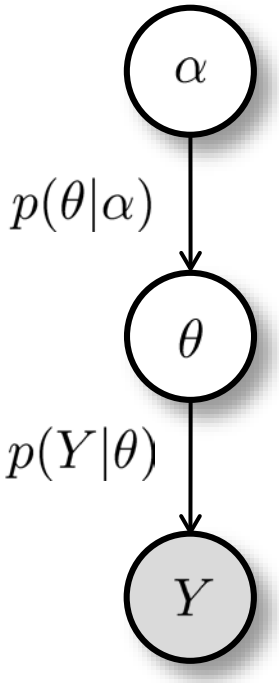
$y_j(\omega,t)$ マイクロホンアレー信号

ブラインド音源分離 (BlindSourceSeparation)

- 独立ベクトル分析／多チャンネルNMF

[Ono 2011, Ozerov 2010, Kameoka+2010,
Sawada+2012, Kitamura 2015, ...]





$\sigma_k^2(\omega,t)$ パワースペクトログラム

⇒手法名の違いはスペクトログラム構造に関する仮定の違い

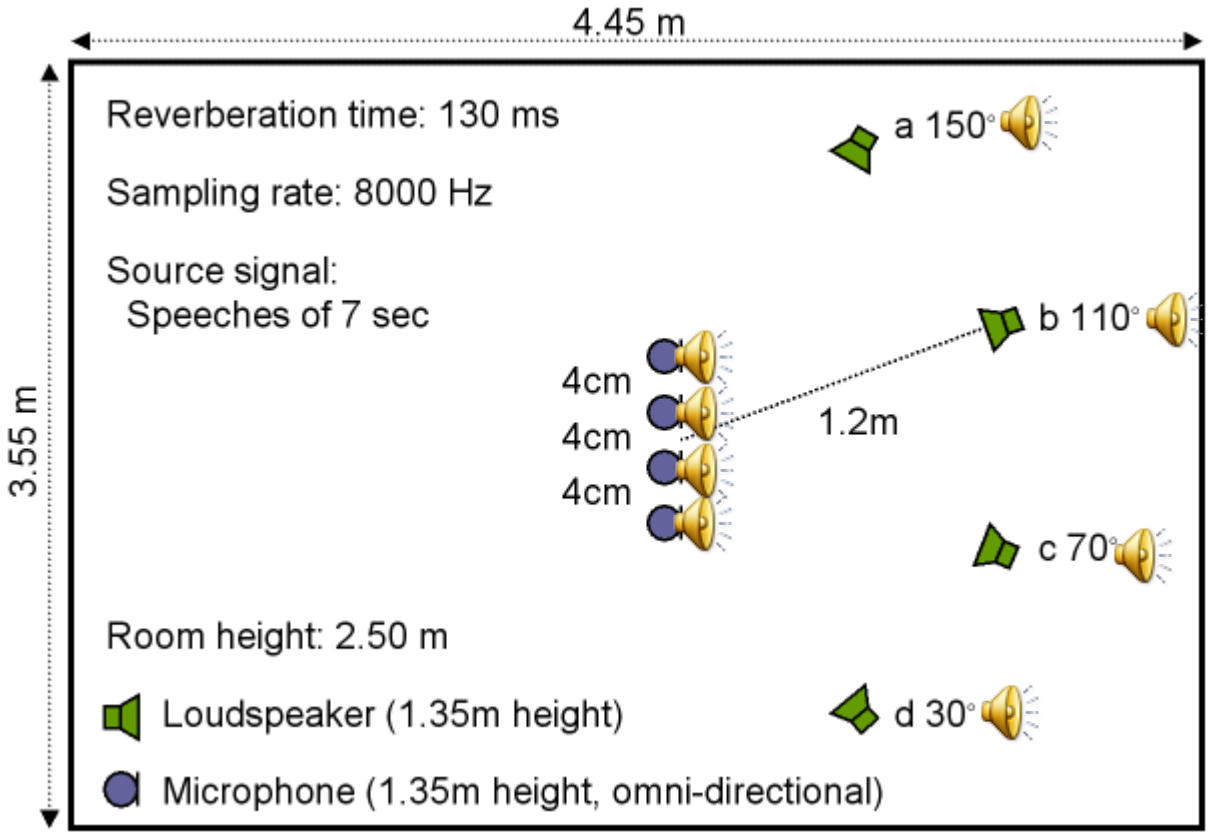
$a_k(\omega)$ アレー応答

$p(Y|A,\sigma^2)$

$y_j(\omega,t)$ マイクロホンアレー信号

$$\log p(Y|A,\sigma^2) \rightarrow \text{maximize}$$

ブラインド音源分離の適用例



分離信号 **y1** **y2** **y3** **y4**

http://www.kecl.ntt.co.jp/icl/signal/sawada/demo/bss2to4/index.html

アウトライン

- 物理モデリング×時系列モデリング×確率モデリングによる音響信号の分解再構成
  - 音声のソース･フィルタ分解
  - 残響除去
  - 音源分離
  - 基本周波数パターンのフレーズ･アクセント分解

アウトライン

- 物理モデリング×時系列モデリング×確率モデリングによる音響信号の分解再構成
  - 音声のソース･フィルタ分解
  - 残響除去
  - 音源分離
  - 基本周波数パターンのフレーズ･アクセント分解

基本周波数パターンのフレーズアクセント分解

- 基本周波数(F_0)パターン
  - 声の高低の時間変化(抑揚)
  - 声の表情，調子，意図，個人性，方言などの非言語的役割を担う
  - 話し方の自然性に大きく影響を与える

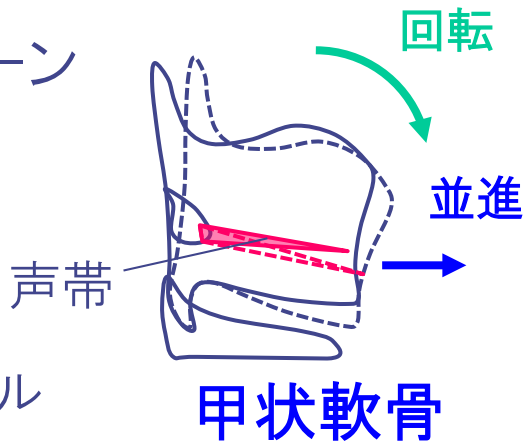
⇒ F_0パターンの自然性を保ったまま音声を合成／変換したい

- F_0パターンの生成過程モデル [Fujisaki+1988]
  - 音声のF_0パターンは2つの成分からなる
    - **フレーズ成分**： 文章全体に及ぶ緩やかな高低パターン
    - **アクセント成分**： 単語内の急峻な高低パターン

⇒甲状軟骨の並進運動、回転運動により制御

⇒ **藤崎モデル**:

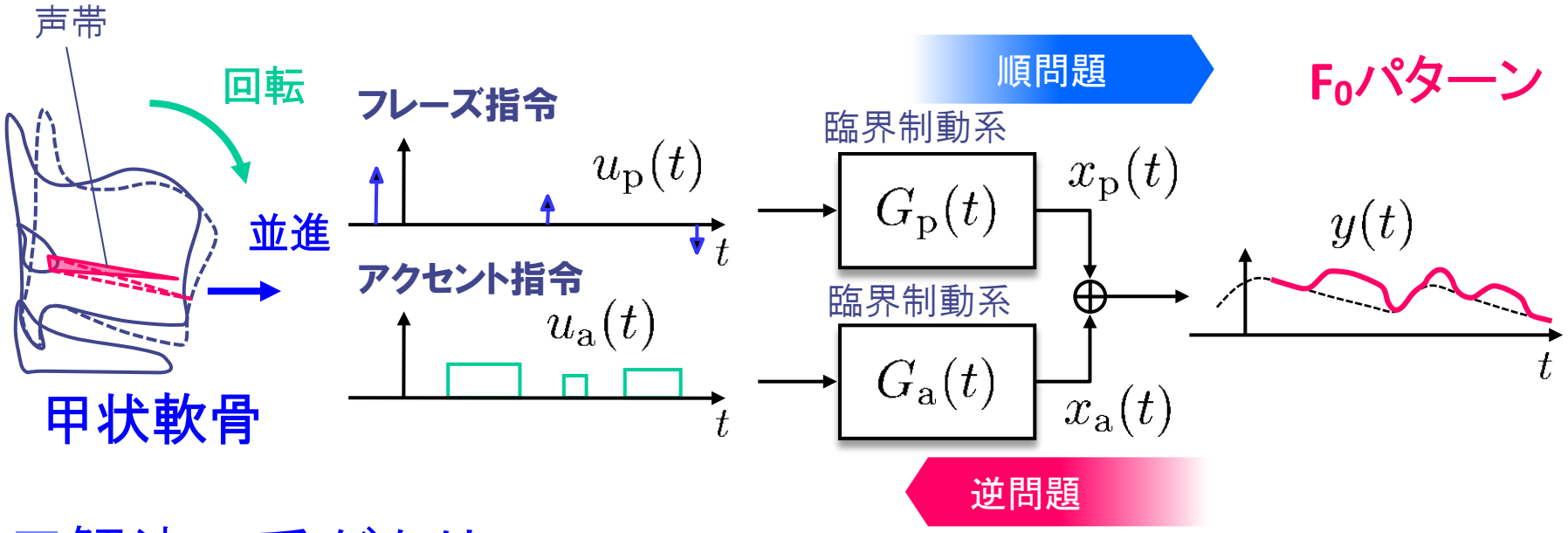
甲状軟骨によるF_0パターンの生成過程を模擬したモデル



藤崎モデルのパラメータ推定

藤崎モデルのパラメータ推定の難しさ

- 藤崎モデルは F_0 パターンをフレーズ･アクセントに対応する運動成分の**和**で表す為，パラメータ推定は不良設定の逆問題
  （7+3 を解くのは簡単でも $x + y = 10$ となる x, y の解は無数に存在）



解決の手がかり

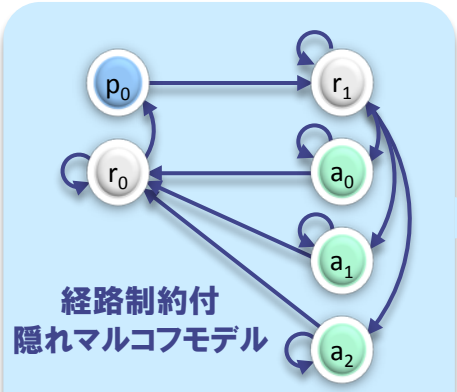
- 自然音声におけるフレーズ･アクセント成分には**統計的な偏り**が存在
- フレーズ成分は**パルス列**、アクセント成分は**矩形パルス列**によって駆動

Statistical Phrase/Accent Command Estimation法 [Kameoka+2010, Yoshizato+2012]

■離散時間確率過程による藤崎モデルの確率モデル版

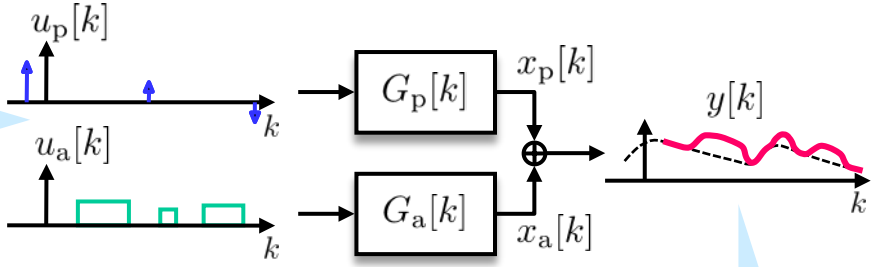
- 統計的手法（EMアルゴリズム，動的計画法）を駆使した効率的なパラメータ推定アルゴリズムを導出できる
- フレーズ・アクセント成分の統計的な偏りを手がかりにすることが可能（不良設定の逆問題を統計的アプローチにより解決）

$p(\theta)$



経路制約付
隠れマルコフモデル

基本周波数パターン生成過程モデル(藤崎モデル)



For $k = 1, \ldots, K$:

$s_k | s_{k-1} \sim \pi_{s_{k-1}, s_k}$（状態系列）

$$\begin{bmatrix} u_p[k] \\ u_a[k] \end{bmatrix} | s_k \sim \mathcal{N}\left(\begin{bmatrix} \mu_{p,s_k} \\ \mu_{a,s_k} \end{bmatrix}, \begin{bmatrix} \sigma_p^2 & 0 \\ 0 & \sigma_a^2 \end{bmatrix} \right)$$

変数変換

基本周波数パターンの確率モデル

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\begin{cases} \boldsymbol{\mu} = \boldsymbol{G}_p \boldsymbol{\mu}_p + \boldsymbol{G}_a \boldsymbol{\mu}_a + \mu_b \mathbf{1} \\ \boldsymbol{\Sigma} = \sigma_p^2 \boldsymbol{G}_p \boldsymbol{G}_p^\top + \sigma_a^2 \boldsymbol{G}_a \boldsymbol{G}_a^\top + \boldsymbol{\Sigma}_b \end{cases}$$

$p(Y|\theta)$

隠れマルコフモデルによる区分定常なガウス過程の表現

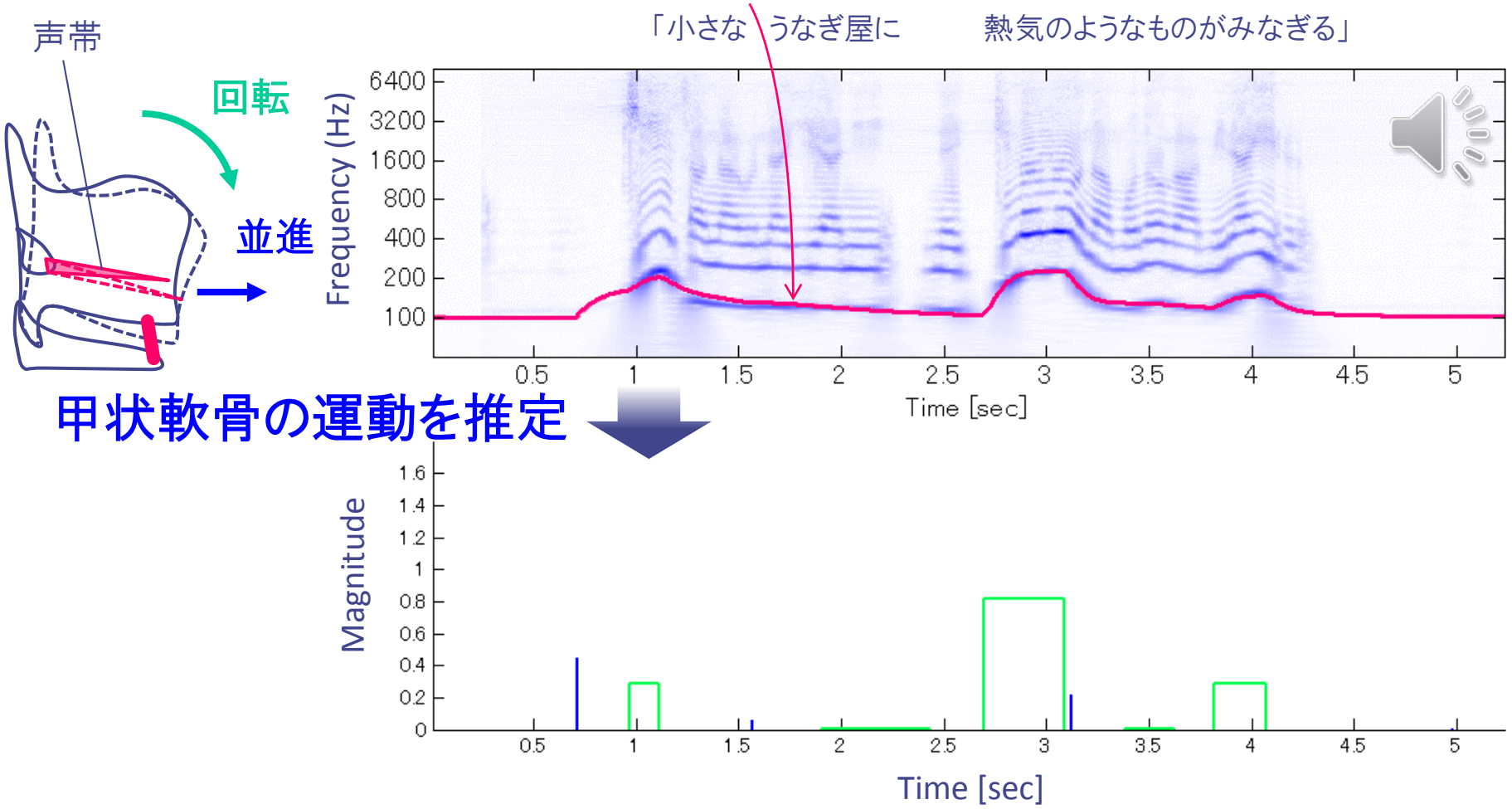
運動方程式に基づく変数変換

EMアルゴリズムと動的計画法による最尤推定

SPACE法の適用例

■入力音声

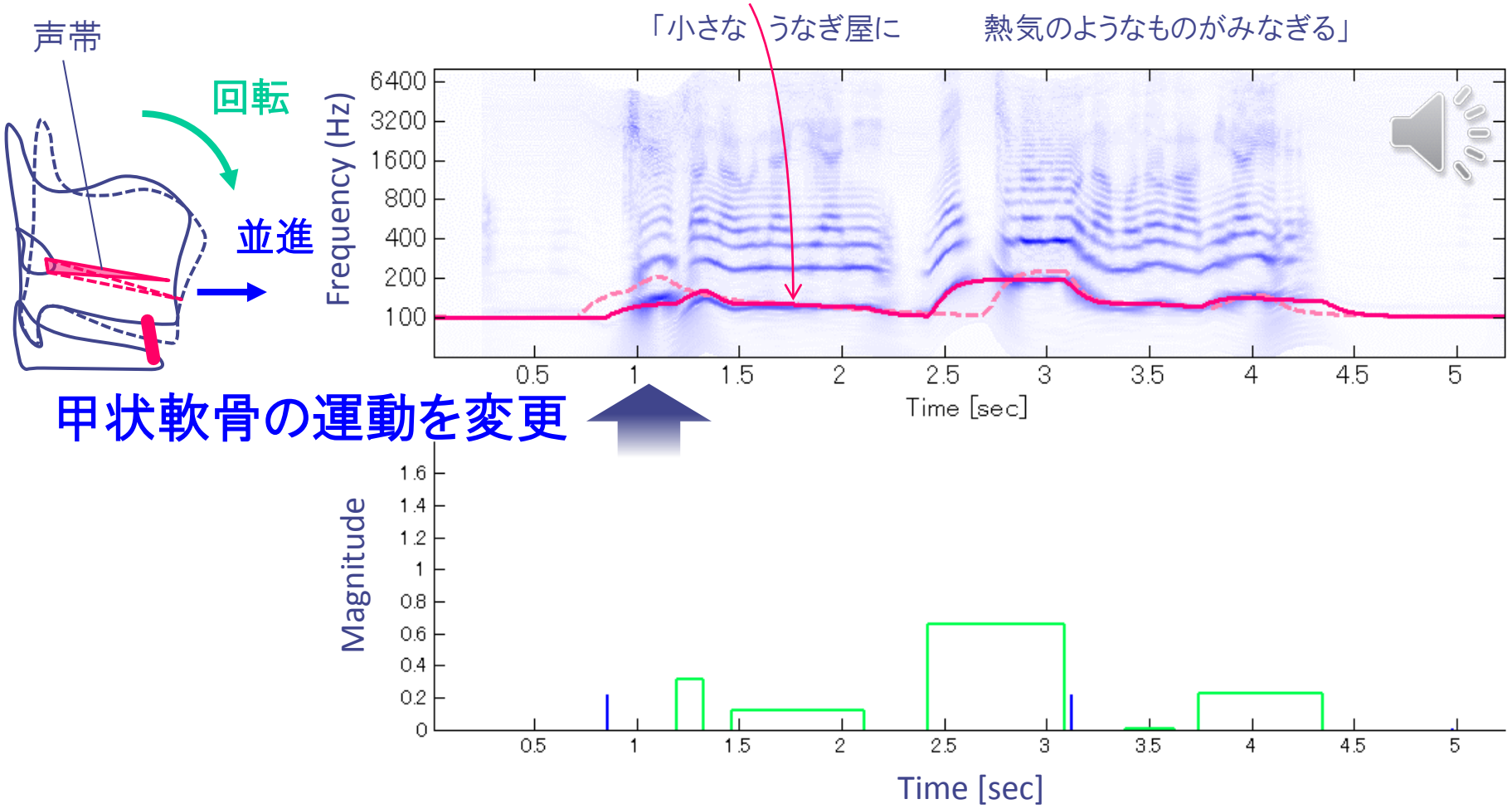
基本周波数パターン（声の高さの時間変化）



SPACE法の適用例

■アクセントのタイミングを変更

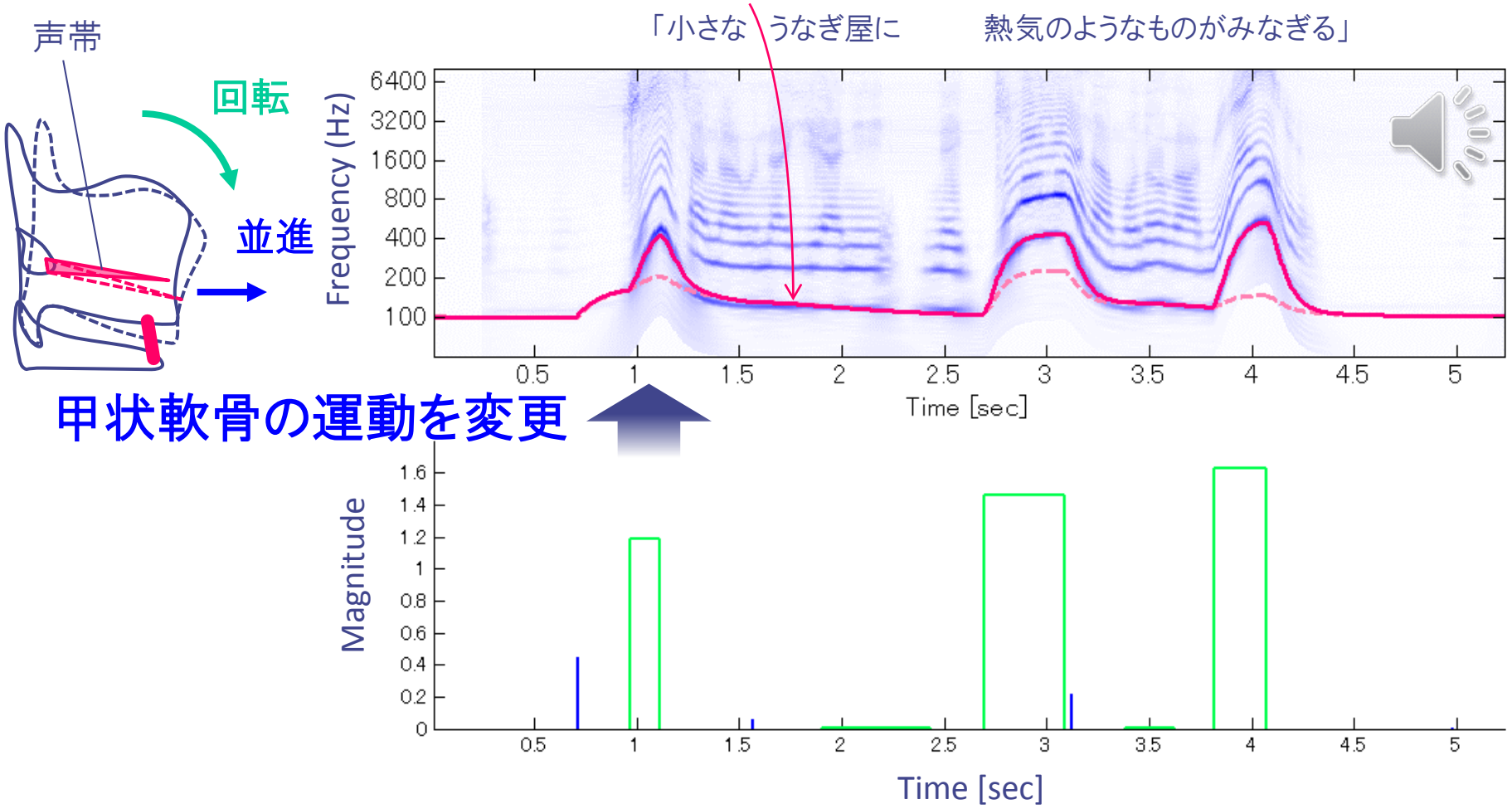
基本周波数パターン（声の高さの時間変化）



SPACE法の適用例

■アクセントの強さを変更

基本周波数パターン（声の高さの時間変化）



F_0パターンのフレーズ・アクセント分解の応用

- 聞き取りやすい音声への変換
  - 非母語話者音声の抑揚を母語話者風に
  - 通常音声をアナウンサーのようにメリハリのある声に
  - 喉頭摘出者の電気喉頭音声を自然音声風に
    ※戸田智基先生（名大）と共同研究中

電気式
人工喉頭

- 音声コンテンツの加工
  - 映画やアニメの俳優・声優の話し方を自分好みに変換

- 音声合成
  - 表情豊かに話すロボット

アウトライン

- 物理モデリング×時系列モデリング×確率モデリングによる音響信号の分解再構成
  - 音声のソース･フィルタ分解
  - 残響除去
  - 音源分離
  - 基本周波数パターンのフレーズ･アクセント分解

アウトライン

- 物理モデリング×時系列モデリング×確率モデリングによる音響信号の分解再構成
  - 音声のソース･フィルタ分解
  - 残響除去
  - 音源分離
  - 基本周波数パターンのフレーズ･アクセント分解

まとめ

- 音響信号の分解と再構成
  - 音声のソースフィルタ分解
  - 残響除去
  - 音源分離
  - 音声の基本周波数パターンのフレーズアクセント分解
- 今後の課題
  - 深層学習をいかに組み込むか？
    - 画像生成の分野で最近注目されているDeep Generative Models (VAE, GAN, ...)
    - DNN音声分離（後述）の例のように分離問題を識別問題にうまく翻訳するアイディア
    - 音や声の物理法則を制約として内部に組み込んだDNN（聴感上の品質を保証、過学習を防止、・・・）

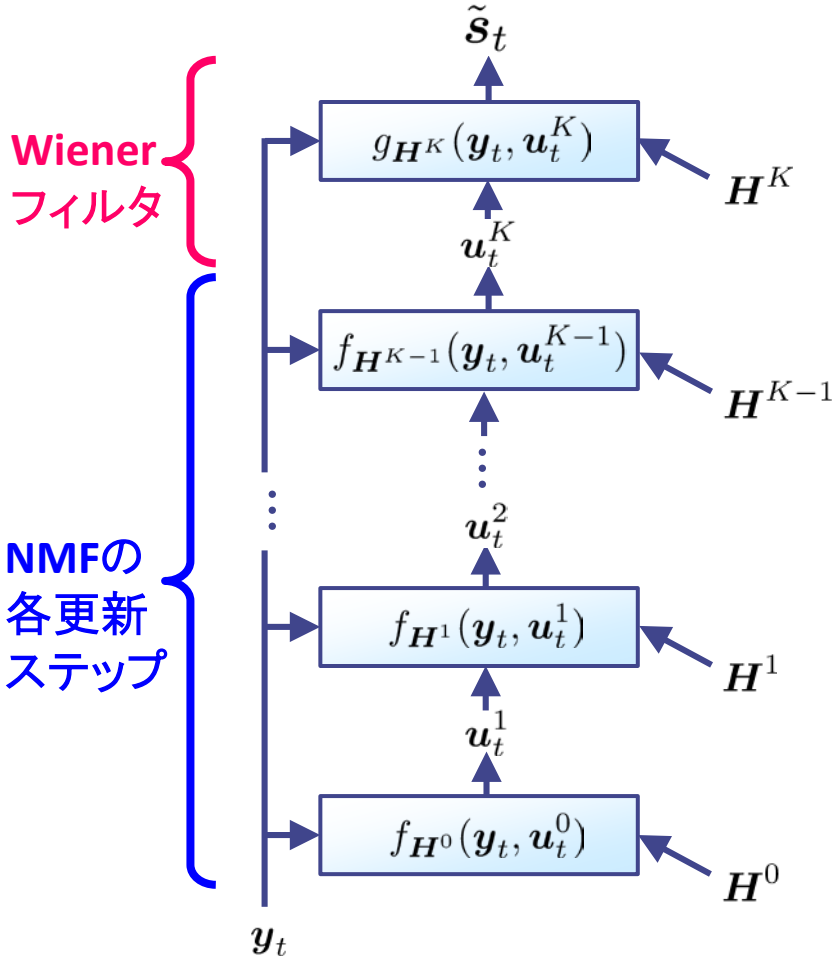
Deep NMF [Le Roux+2015]

Deep Unfolding [Hershey+2014]

- パラメータの反復更新アルゴリズムを"unfold（展開）"してDNNのようなDeep architectureを作る
- Sigmoid型NNは「マルコフ確率場の平均場推論アルゴリズムを"unfold"したもの」と見なせる

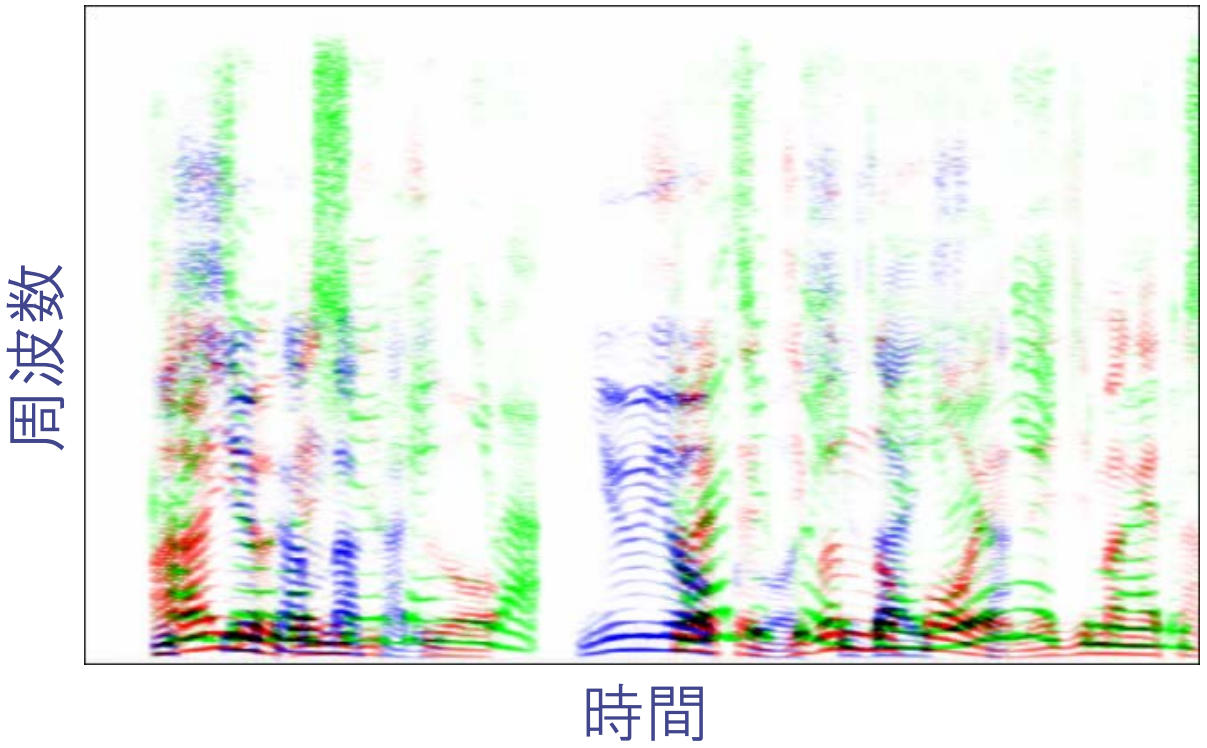
Deep NMF

- DNNとNMFのハイブリッドアプローチ
- NMFの乗法更新則が活性化関数になったDNN
  - NMFの良さとDeep Networkの良さを併せ持つ
  - 分離性能が最大になるように $\boldsymbol{H}$ を学習できる



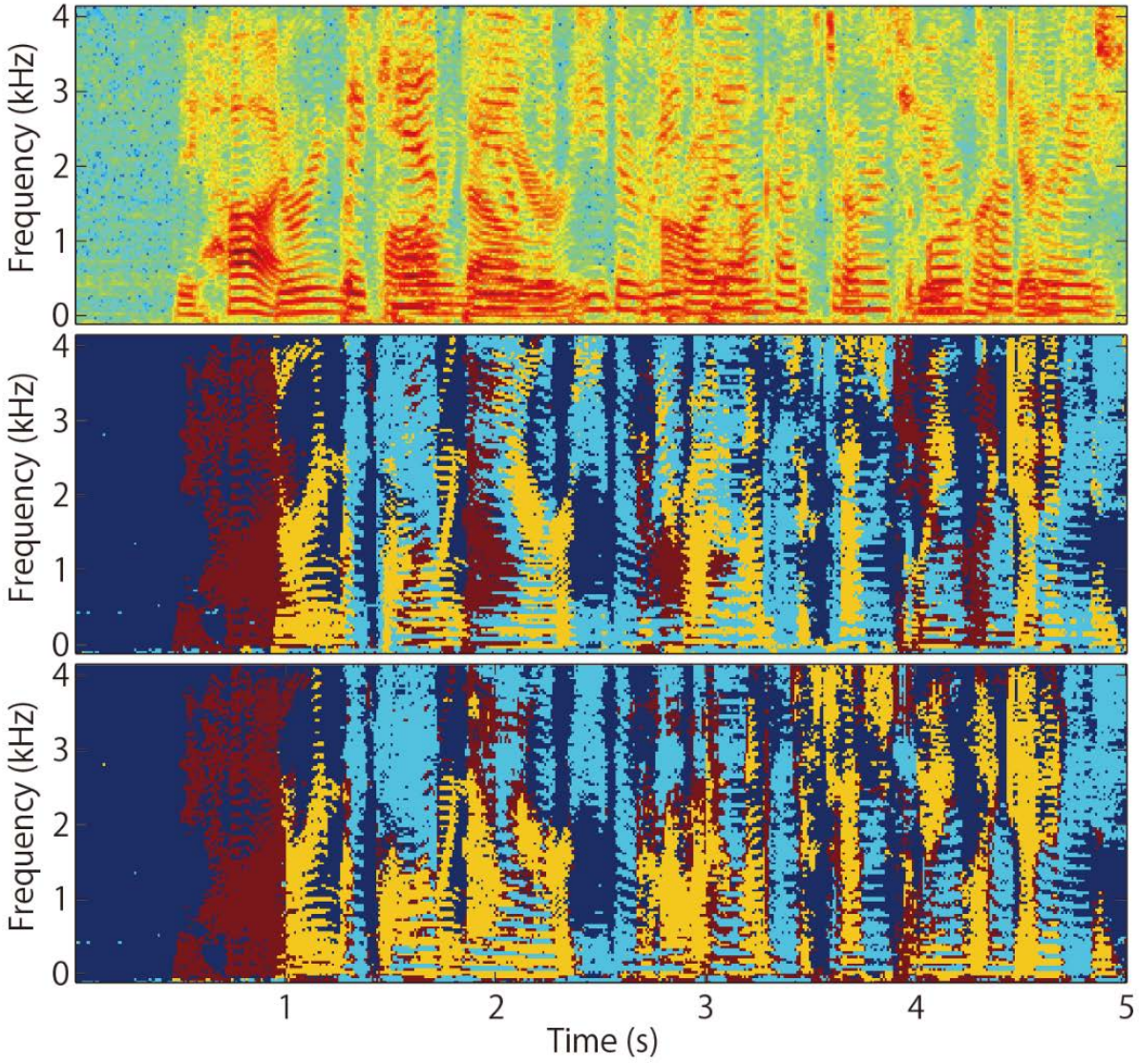
DNN音声分離 [Weninger+2015, Wang+2014, Du+2014, Hershey+2016]

- DNNによる時間周波数マスク推定
  - 音声のスペクトログラムはスパースになる傾向
  - 各時間周波数点が何の音に帰属しているかを識別
    （画像のセグメンテーションに類似）
  - 同一音源に帰属する時間周波数点の成分のみを通過させるマスク



DNN音声分離 [Weninger+2015, Wang+2014, Du+2014, Hershey+2016]

- 混合音声（3話者）のスペクトログラム
- 正解時間周波数マスク
- 推定時間周波数マスク



[Hershey+2016]より転用

宣伝

- 機械学習プロフェッショナルシリーズ（全29巻）
  - 第6期 「統計的音響信号処理」（12月刊行予定）

講談社
サイエンティフィク

講談社サイエンティフィクは科学一般から地球環境科学まで、
多くの自然科学関連書籍を出版しています。
メールマガジン登録
近刊情報（PDF）

書籍情報 | 購入方法 | ダウンロード | 会社概要 | 採用情報 | お問い合わせ

Home ＞ 書籍情報 ＞ 機械学習プロフェッショナルシリーズ

書籍情報
検索
ジャンル
科学一般
科学英語
数学・情報科学
物理学
工学
化学
生物学・生物科学
医学・薬学・臨床検査・看護
生活科学
獣医学・畜産学・農学・水産学
地球環境科学
健康科学・スポーツ医科学
その他
» 実験ノート
シリーズ
穴埋め式らくらくワークブック
栄養科学シリーズＮＥＸＴ
英文書
エキスパート応用化学テキスト
絵でわかる
機械学習プロフェッショナル
基礎物理数学　第４版
今日から使える

機械学習プロフェッショナルシリーズ

［シリーズ編者］
杉山 将　　東京大学大学院新領域創成科学研究科 教授

［本シリーズの特徴］
- 発展著しい機械学習技術の数学的な基礎理論、実用的なアルゴリズム、それらの活用法を解説。
- 手に取りやすいページ数で、大事な点を簡潔丁寧にまとめた。
- ビッグデータ時代を牽引している若手・中堅の現役研究者が一堂に会した最強の執筆陣！

◆リンク　シリーズ一覧

［今後の刊行予定］
［第5期］（2016年8月刊行予定）
バンディット問題の理論とアルゴリズム／ウェブデータの機械学習／データ解析におけるプライバシー保護
［第6期］
機械学習のための連続最適化／関係データ学習／オンライン予測／統計的音響信号処理／画像認識
［第7期］
ガウス過程と機械学習／統計的因果探索／深層学習による自然言語処理／映像認識／脳画像のパターン認識／強化学習／ロボットの運動学習


グラフィカルモデル（機械学習プロフェッショナルシリーズ）　購入する
渡辺有祐・著　　本体2800円(税別)
発行: 2016/04/19　ISBN: 978-4-06-1529168


ヒューマンコンピュテーションとクラウドソーシング（機械学習プロフェッショナルシリーズ）　購入する
鹿島久嗣/小山聡/馬場雪乃・著　　本体2400円(税別)
発行: 2016/04/19　ISBN: 978-4-06-1529137

亀岡弘和（NTT）・
吉井和佳（京大）
著