畳み込みHMMに基づく歌声の基本周波数制御モデルの提案とそのパラメータ学習方法

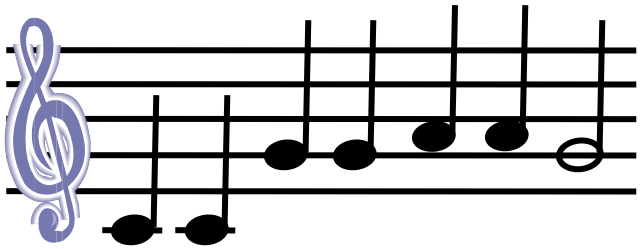
大石 康智[1]，亀岡 弘和[2]
柏野 邦夫[2]，武田 一哉[1]

[1]名古屋大学大学院情報科学研究科
[2]NTTコミュニケーション科学基礎研究所

研究の根底にある興味

- 歌声に含まれる歌唱者の個性の分析

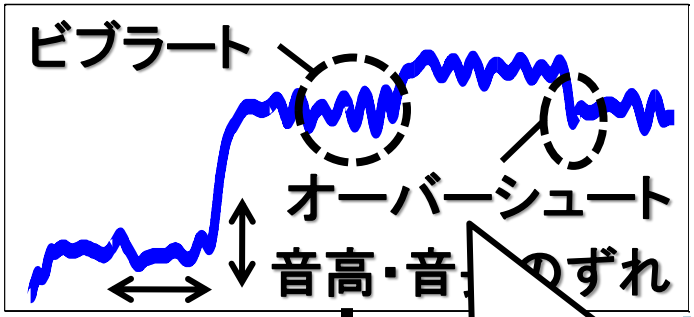
譜面



歌唱者の個性

意図・スタイル
物理的，生理的な制約

歌声のF0軌跡

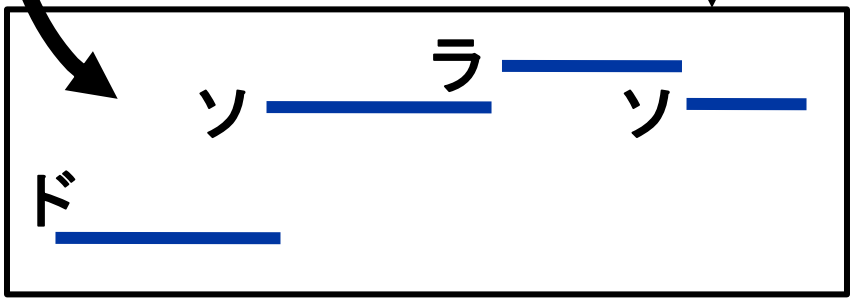


個人性知覚に影響を与える [Saitou, 2006]

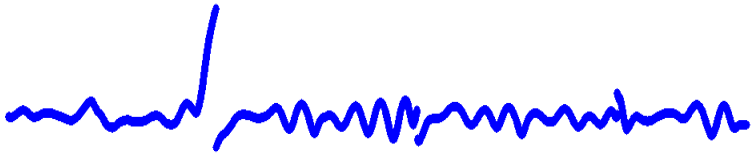
音高・音長のずれ

離散的な音高列と動的な変動成分への分解

離散的な音高列



動的変動成分



ビブラートやオーバーシュートの特性

応用 楽曲検索・採譜，歌唱者間の歌唱スタイルの転写

歌声のF0生成過程に基づく制御モデル

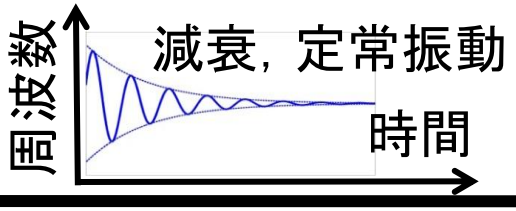
- アイデアの要点

歌声合成への有効性 [Saitou, 2006]

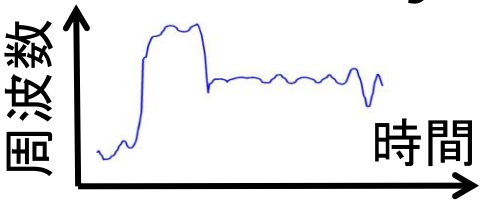
音高列（入力 u）

＊

動的変動因子



F0軌跡（出力 y）



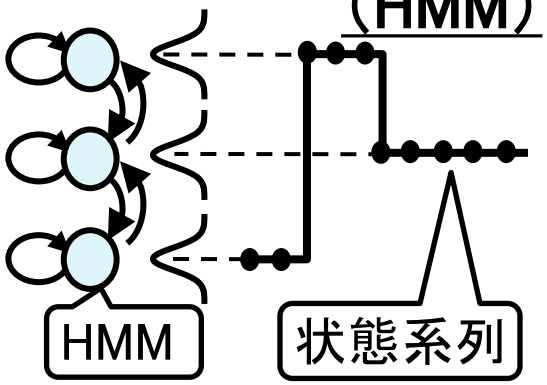
⇩ 階段状の信号?

⇩ 線形システム?

⇩ 観測されるのは F0軌跡のみ!

アイデア1

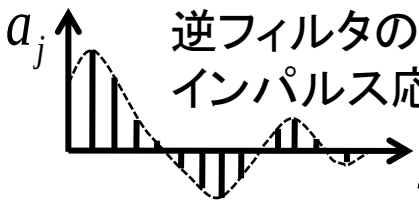
隠れマルコフモデル（HMM）



アイデア2

全極モデル

$$\frac{Y(z)}{U(z)} = \frac{1}{a_0 + \sum_{j=1}^{M} a_j z^{-j}}$$



アイデア3

パラメータ最尤推定

ステップ1　Viterbi学習

階段状入力信号の推定

反復推定

ステップ2　LPC的解法

$(a_0, \ldots, a_M)$ の推定

全極モデルに基づく伝達関数表現

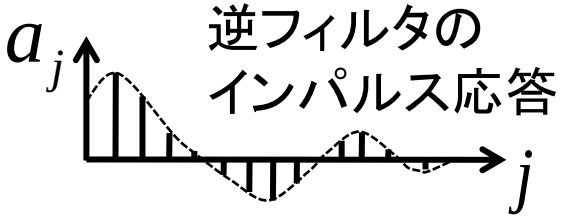
- 入力信号u_nとF0軌跡y_nのz変換を$U(z), Y(z)$とすると

$$\frac{Y(z)}{U(z)} = \frac{1}{a_0 + \sum_{j=1}^{M} a_j z^{-j}}$$

（M：伝達関数の次数）

  - 全極モデルの妥当性：2階の微分方程式を差分近似して解く
- 逆z変換して導かれる差分方程式

$$a_0 y_n + \sum_{j=1}^{M} a_j y_{n-j} = u_n \quad \Rightarrow$$



- 微細変動成分
  - 平均0，分散σ^2の正規分布に従うGauss性白色雑音ε

$$\hat{u}_n - u_n = a_0 y_n + \sum_{j=1}^{M} a_j y_{n-j} - u_n = \varepsilon \quad \left(\varepsilon \sim N\left(0, \sigma^2\right)\right)$$

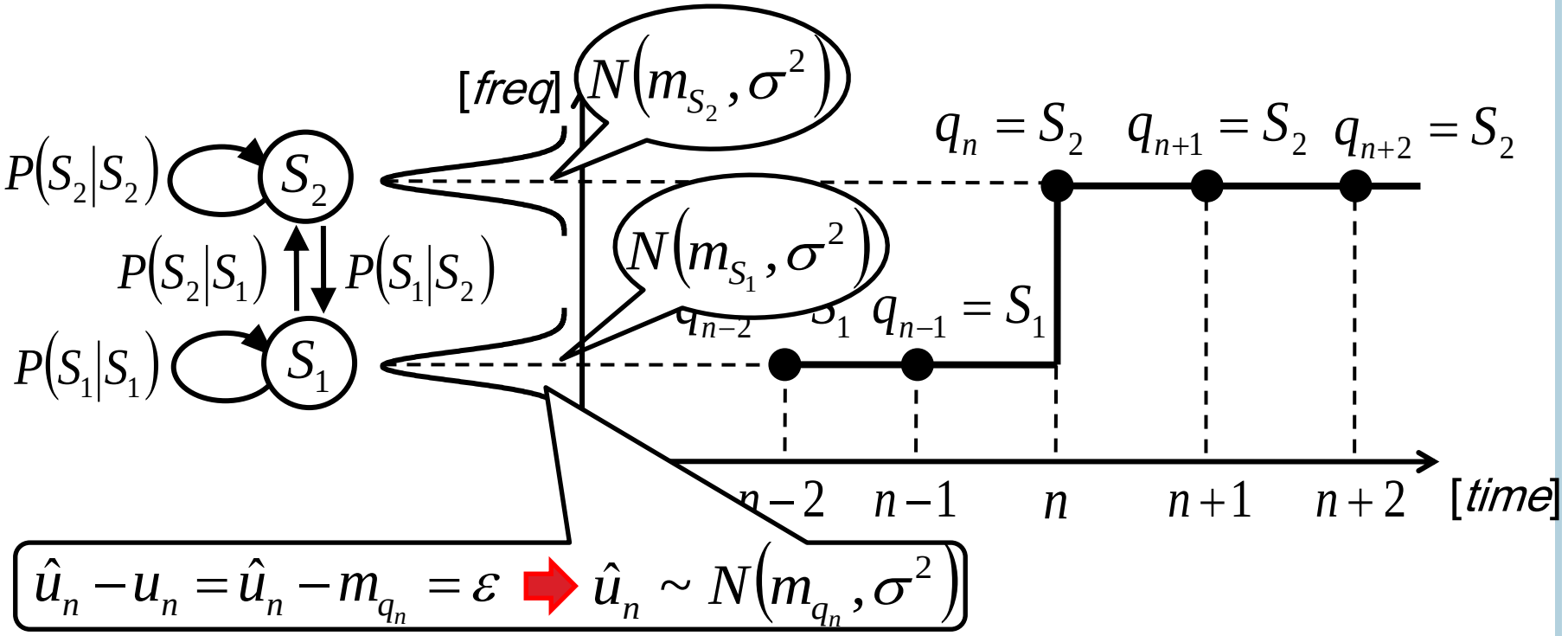
階段状の信号に微細変動成分が付加された信号

階段状の入力信号u_nのモデル化

- I 個の状態集合$S=\{S_1,\ldots,S_I\}$からなるHMMの利用

$$u_n = m_{q_n} \quad (q_n \in \{S_1,\ldots,S_I\})$$

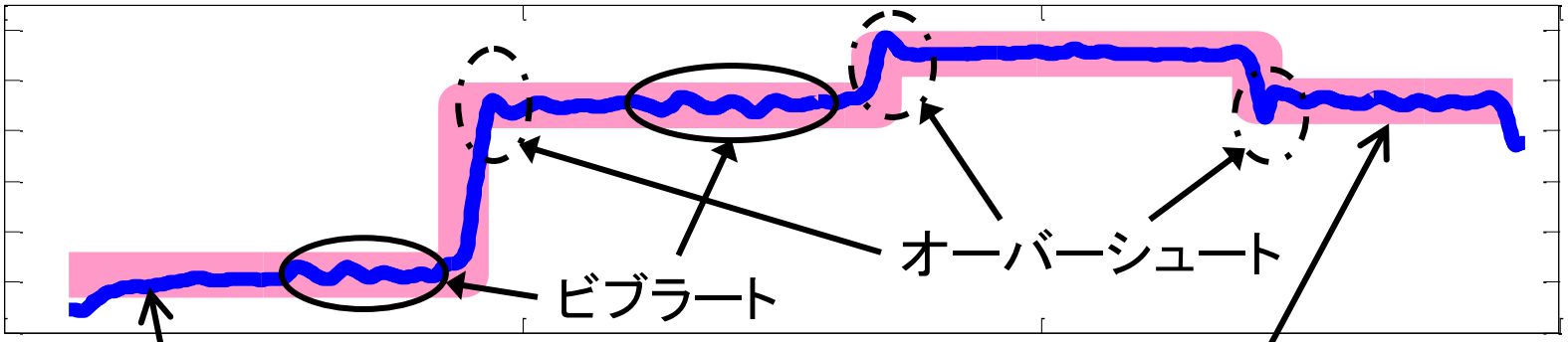
- m_{S_i}：状態 S_i の出力確率分布（正規分布）の平均
- $P(S_i|S_j)$：状態 S_j から S_i への遷移確率



$$\hat{u}_n - u_n = \hat{u}_n - m_{q_n} = \varepsilon \Rightarrow \hat{u}_n \sim N(m_{q_n},\sigma^2)$$

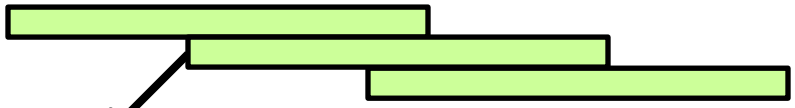
時変なF0制御モデルへの拡張

- 時々刻々と変化する動的変動成分



F0軌跡 : $\mathbf{y} = \{y_1, y_2, \ldots, y_N\}$　入力信号 : $\mathbf{u} = \{m_{q_1}, m_{q_2}, \ldots, m_{q_N}\}$

フレームごとに分析する



フレーム t（長さL）

$$\mathbf{y}^{(t)} = \{y_1^{(t)}, y_2^{(t)}, \ldots, y_L^{(t)}\}$$
$$\mathbf{u}^{(t)} = \{m_{q_1^{(t)}}, m_{q_2^{(t)}}, \ldots, m_{q_L^{(t)}}\}$$

インパルス応答

$$\boldsymbol{\theta}^{(t)} = \{a_0^{(t)}, \ldots, a_M^{(t)}\}$$

仮定する関係式

$$\hat{u}_l^{(t)} - u_l^{(t)} = \sum_{j=0}^{M} a_j^{(t)} y_{l-j}^{(t)} - m_{q_l^{(t)}} = \varepsilon$$

$$\{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(T)}\} \Rightarrow \text{推定} \Rightarrow \boldsymbol{\Theta} = \{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}, \boldsymbol{\omega}\} \quad (\boldsymbol{\omega} = \{q_1, \ldots, q_N, m_{S_1}, \ldots, m_{S_I}\})$$

F0制御モデルのパラメータ最尤推定(1)

- F0軌跡集合 $\{\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(T)}\}$ からパラメータ集合 Θ の推定
  - $\{\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(T)}\}$ の対数尤度 $\log P(\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(T)}\,|\,\Theta)$
    - 各フレームのF0軌跡が独立に観測されると想定すると

$$\log P(\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(T)}\,|\,\Theta)=\sum_{t=1}^{T}\log P(\mathbf{y}^{(t)}\,|\,\boldsymbol{\theta}^{(t)},\boldsymbol{\omega})$$

    - フレーム t において，以下を仮定する

$$\hat{u}_l^{(t)}-u_l^{(t)}=\sum_{j=0}^{M}a_j^{(t)}y_{l-j}^{(t)}-m_{q_l^{(t)}}=\varepsilon \quad \left(\varepsilon\sim N\left(0,\sigma^{(t)2}\right)\right)$$

$$\log P(\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(T)}\,|\,\Theta)=$$

$$\sum_{t=1}^{T}\left\{-(L-M)\log\left(\sqrt{2\pi}\sigma^{(t)}a_0^{(t)}\right)-\frac{1}{2\sigma^{(t)2}}\sum_{l=M+1}^{L}\left(\sum_{j=0}^{M}a_j^{(t)}y_{l-j}^{(t)}-m_{q_l^{(t)}}\right)^2\right\}$$

F0制御モデルのパラメータ最尤推定(2)

- パラメータ Θ の事後確率

$$P(\Theta \mid \mathbf{y}^{(1)},\ldots,\mathbf{y}^{(T)}) \propto P(\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(T)} \mid \Theta)P(\Theta)$$

  - 事前確率 $P(\Theta)$

$$P(\Theta) = \underbrace{P(q_1,\ldots,q_N)}_{\text{一様なマルコフ連鎖}P(q_1)P(q_2|q_1)\ldots P(q_N|q_{N-1})}\underbrace{P(\theta^{(1)})P(\theta^{(2)})\ldots P(\theta^{(T)})}_{\text{一様分布}}P(m_{S_1},\ldots,m_{S_I})$$

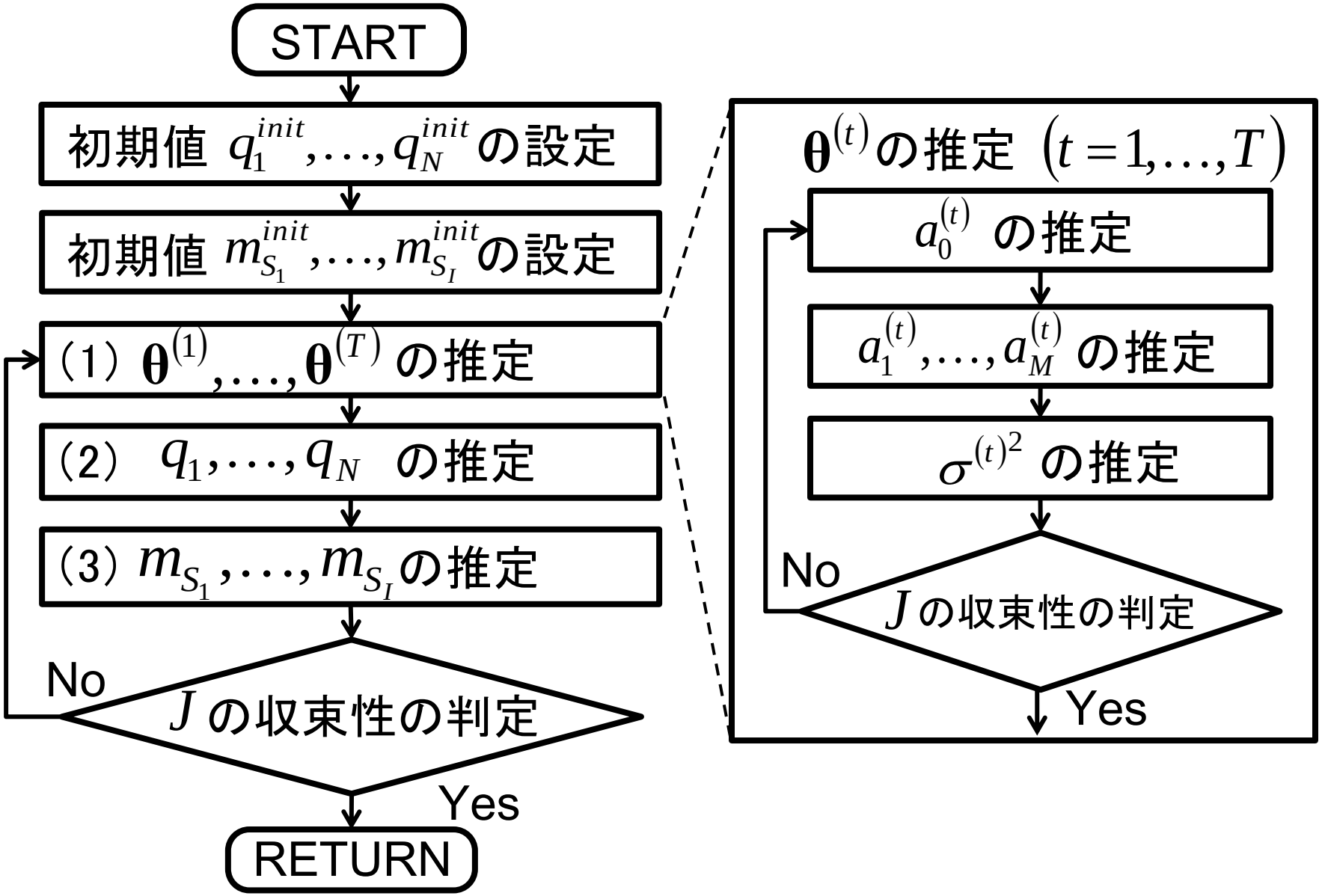
- 目的関数 J

$$J \equiv \sum_{t=1}^{T} \log P(\mathbf{y}^{(t)} \mid \theta^{(t)},\omega) + \log P(q_1)P(q_2|q_1)\ldots P(q_N|q_{N-1})$$

$$= \sum_{t=1}^{T}\left\{-(L-M)\log\left(\sqrt{2\pi}\sigma^{(t)}a_0^{(t)}\right) - \frac{1}{2\sigma^{(t)2}}\sum_{l=M+1}^{L}\left(\sum_{j=0}^{M}a_j^{(t)}y_{l-j}^{(t)} - m_{q_l^{(t)}}\right)^2\right\}$$

$$+ \log P(q_1)P(q_2|q_1)\ldots P(q_N|q_{N-1})$$

パラメータ学習方法のフローチャート



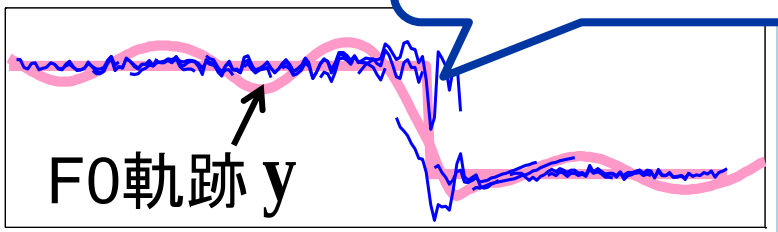
各パラメータの推定

(1) $\boldsymbol{\theta}^{(t)}\left(a_0^{(t)},\ldots,a_M^{(t)}\right)$ の推定（固定：$q_1,\ldots,q_N,m$

$$\hat{u}_l^{(t)}=\sum_{j=0}^{M}a_j^{(t)}y_{l-j}^{(t)}$$

$$\frac{\partial J}{\partial a_{j'}^{(t)}}=0,\ (j'=0,\ldots,M)$$

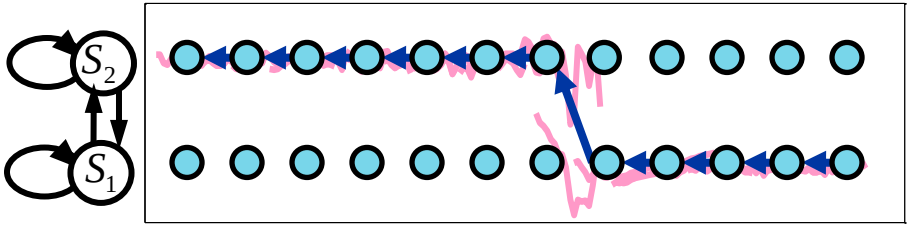
の連立方程式を解く



(2) $q_1,\ldots,q_N$ の推定（固定：$\boldsymbol{\theta}^{(1)},\ldots,\boldsymbol{\theta}^{(T)},m_{S_1},\ldots,m_{S_I}$ ）

Viterbiアルゴリズムに基づく
最適状態系列の推定

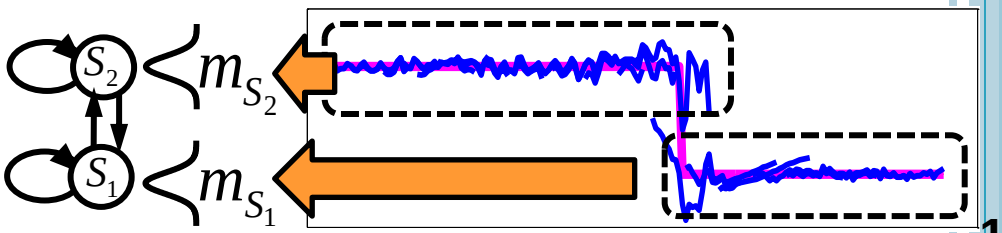
トレリス構造



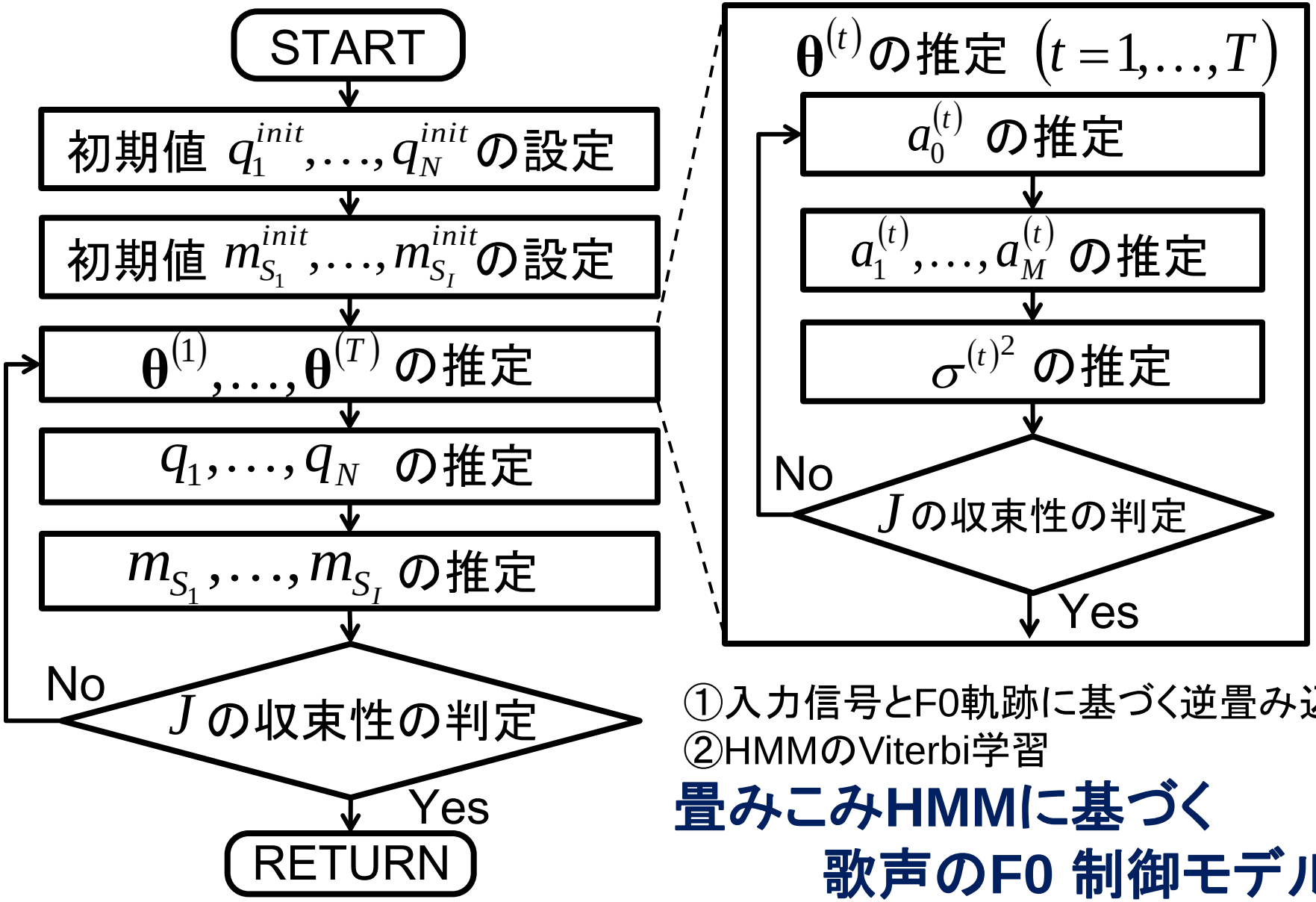
(3) $m_{S_1},\ldots,m_{S_I}$ の推定（固定：$\boldsymbol{\theta}^{(1)},\ldots,\boldsymbol{\theta}^{(T)},q_1,\ldots,q_N$ ）

$$\frac{\partial J}{\partial m_{S_i}}=0,\ (i=1,\ldots,I)$$

の方程式を解く



パラメータ学習方法のフローチャート



①入力信号とF0軌跡に基づく逆畳み込み
②HMMのViterbi学習

畳みこみHMMに基づく
歌声のF0 制御モデル

F0軌跡の再合成

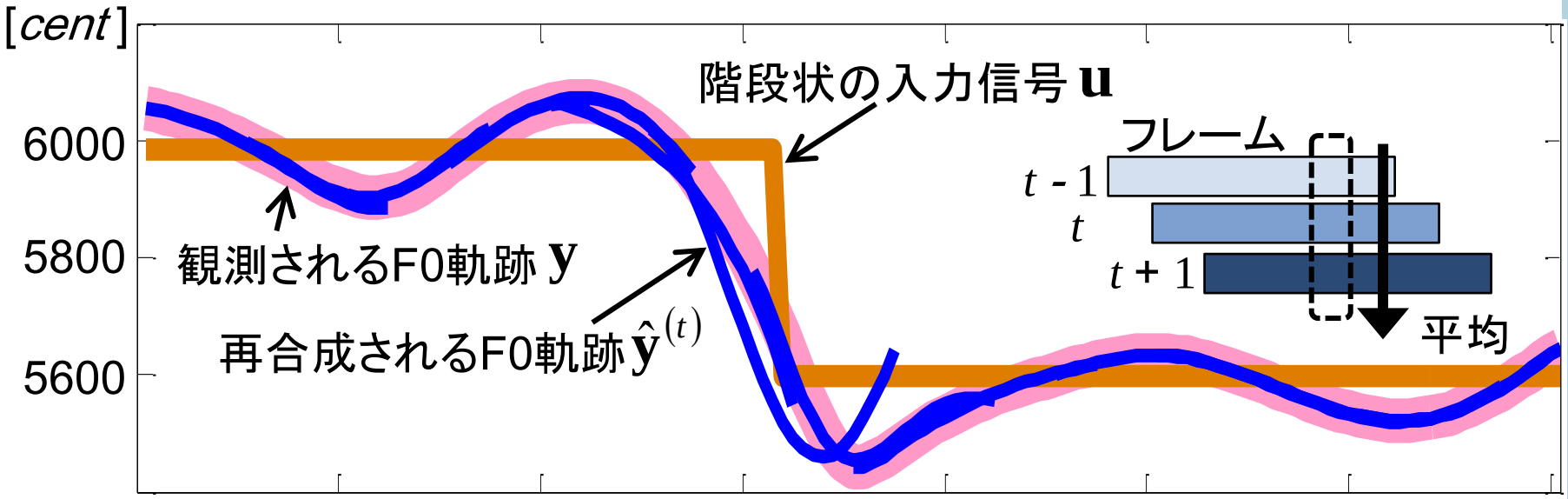
- パラメータ集合$\boldsymbol{\Theta} = \{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}, \boldsymbol{\omega}\}$から$\{\hat{\mathbf{y}}^{(1)}, \ldots, \hat{\mathbf{y}}^{(T)}\}$の生成

（1）$\boldsymbol{\omega} = \{q_1, \ldots, q_N, m_{S_1}, \ldots, m_{S_I}\}$から$\{\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(T)}\}$を求める

（2）フレームごとに，$\boldsymbol{\theta}^{(t)} = \{a_0^{(t)}, \ldots, a_M^{(t)}\}$を利用して

$$\hat{y}_l^{(t)} = \frac{1}{a_0}\left(u_l^{(t)} - \sum_{j=1}^{M} a_j^{(t)} y_{l-j}^{(t)}\right) \quad (l \geq M+1)$$

（ただし，$\hat{y}_l^{(t)} = y_l^{(t)}, (l \leq M)$ ）



評価実験

- パラメータ学習アルゴリズムの収束性
- 信号の推定性能
  - 各フレームにおいて推定される階段状の入力信号$\mathbf{u}^{(t)}$
  - $\mathbf{u}^{(t)}$とインパルス応答$\boldsymbol{\theta}^{(t)}$から再合成されるF0軌跡$\hat{\mathbf{y}}^{(t)}$
- 歌唱者ごとの動的変動成分の特性

- 歌声データベース
  - プロの声楽家，ポップス歌手，素人（各男女1名ずつ）
  - きらきら星，ベートーベン作曲交響曲第9番（よろこびの歌）

ヘッドフォンでガイドトーン（メロディ）を聴きながら

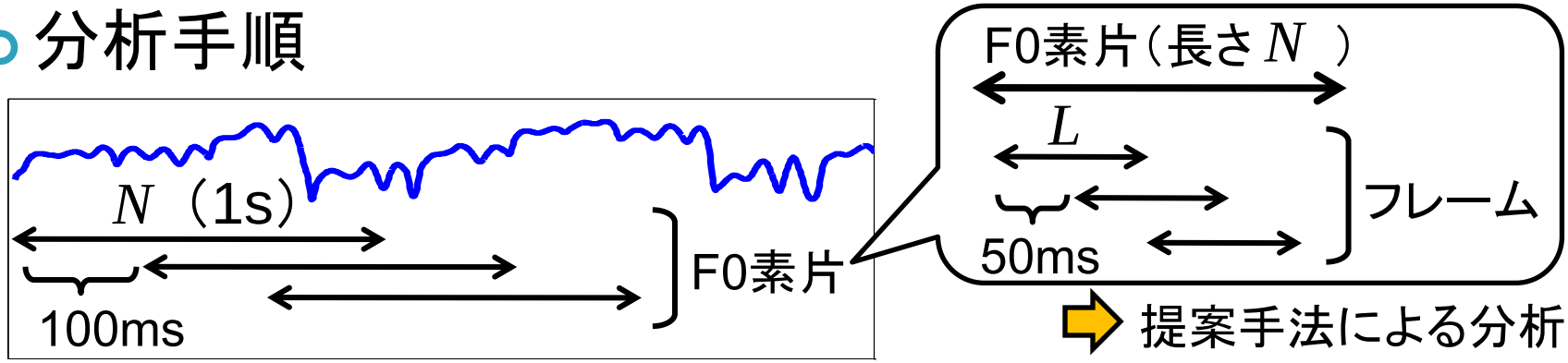
①ラララで歌唱　②歌詞付きで歌唱

↓

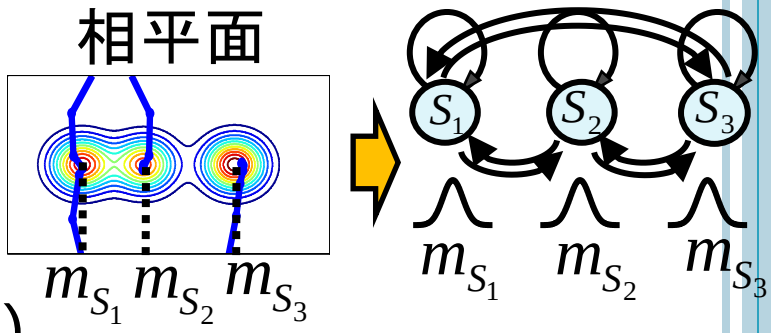
推定される入力信号$\mathbf{u}^{(t)}$との比較

評価実験

- F0推定：YIN（Cheveigne *et.al*, 2002）の利用
  - 歌声，ガイドトーン（$\mathbf{g}$）ともに1msごとに推定
  - Hzをcentで表される対数スケールの周波数へ変換
  - 200ms以内のF0が推定されない区間の線形補間
- 分析手順

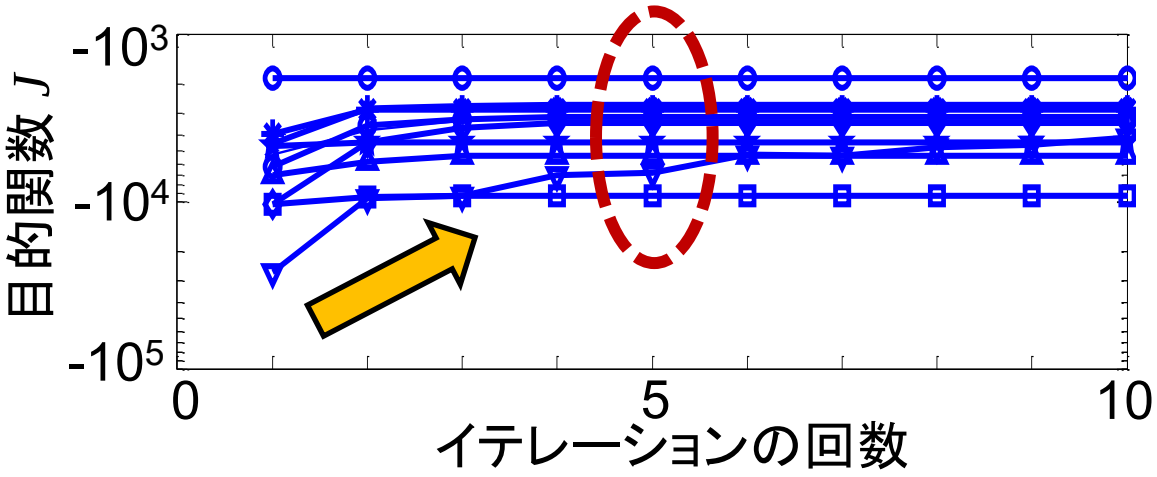


- HMMの初期化
  - 状態数I の決定：相平面の利用
  - 初期状態確率：$1/I$
  - $P(S_i|S_i)$：0.9，$P(S_i|S_j)$：0.1/(I-1)

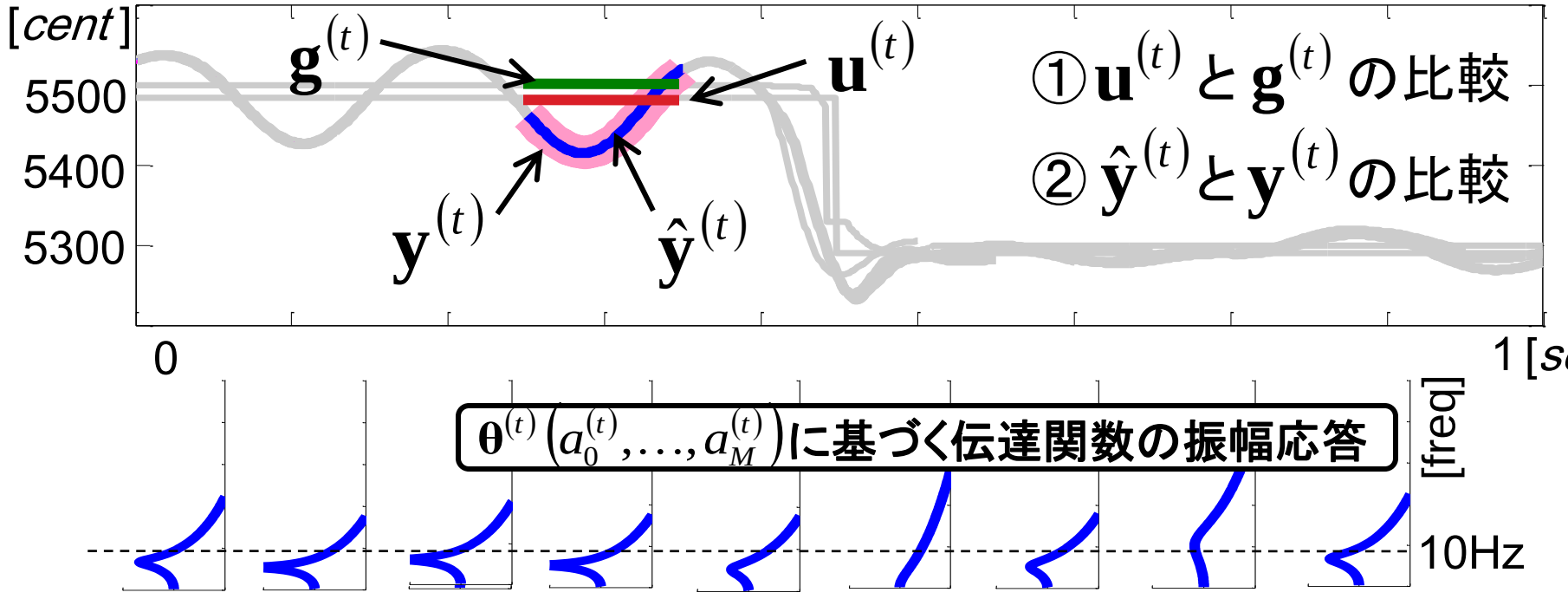


実験結果

- パラメータ学習アルゴリズムの収束性



- あるF0素片の推定結果（$M = 3,\ L = 100\text{ms}$）



実験結果

- 平均二乗誤差(RMS)に基づく正解率

$$\mathrm{RMS}_{\mathbf{u}^{(t)}} = \sqrt{\frac{1}{L-M}\sum_{l=M+1}^{L}\left(u_l^{(t)} - g_l^{(t)}\right)^2} \qquad \mathrm{RMS}_{\hat{\mathbf{y}}^{(t)}} = \sqrt{\frac{1}{L-M}\sum_{l=M+1}^{L}\left(\hat{y}_l^{(t)} - y_l^{(t)}\right)^2}$$

$$\text{正解率}_{\mathbf{u}^{(t)}} = \frac{\mathrm{RMS}_{\mathbf{u}^{(t)}} < \xi \text{ となるフレーム数}}{\text{すべての素片の総フレーム数}} \times 100\ [\%]$$

(ξ : $\mathbf{y}$ と $\mathbf{g}$ のRMS)

$$\text{正解率}_{\hat{\mathbf{y}}^{(t)}} = \frac{\mathrm{RMS}_{\hat{\mathbf{y}}^{(t)}} < 10\text{cent} \text{ となるフレーム数}}{\text{すべての素片の総フレーム数}} \times 100\ [\%]$$

フレーム長Lの評価($M = 5$)

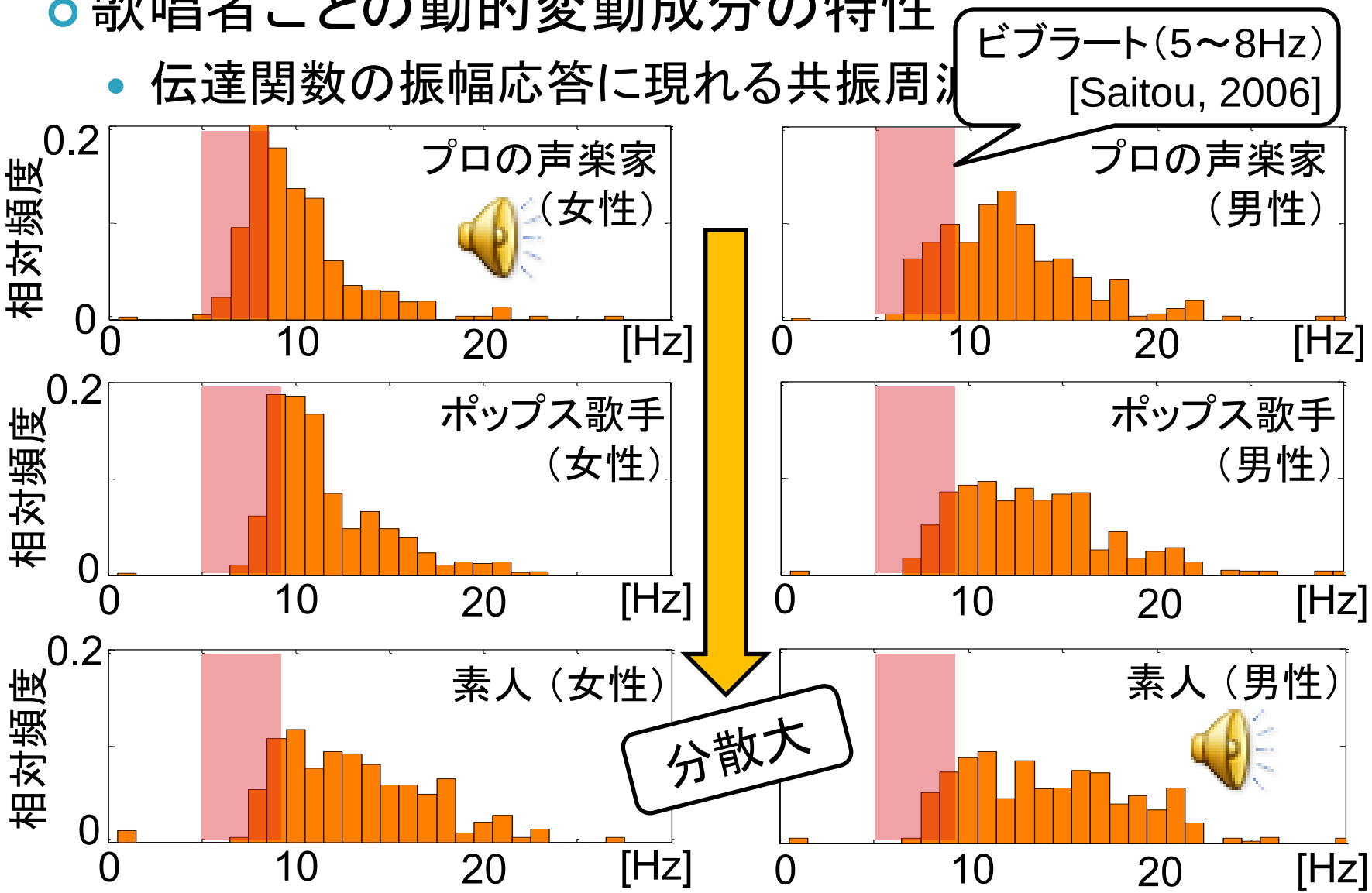
L	正解率 $\mathbf{u}^{(t)}$	正解率 $\hat{\mathbf{y}}^{(t)}$
50ms	82.1	65.9
100ms	80.7	55.8
200ms	79.6	28.9

伝達関数の次数Mの評価(L = 50ms)

M	正解率 $\mathbf{u}^{(t)}$	正解率 $\hat{\mathbf{y}}^{(t)}$
3	82.2	66.5
5	82.1	65.9
10	81.2	52.5

実験結果

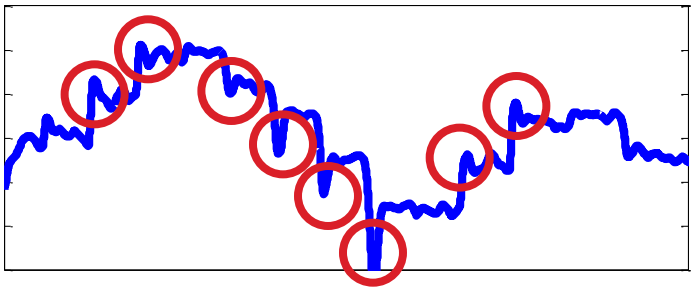
- 歌唱者ごとの動的変動成分の特性
  - 伝達関数の振幅応答に現れる共振周波



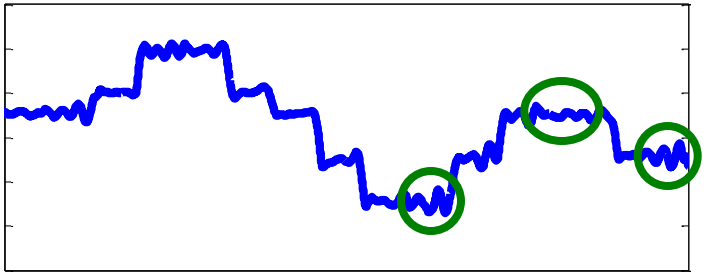
おまけのデモンストレーション

- 声楽家のF0軌跡の動的変動成分を素人に**転写**

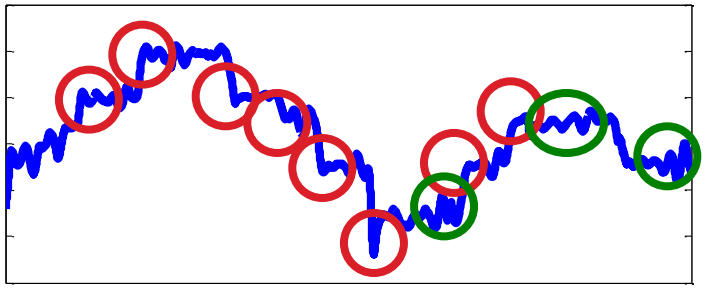
①素人の歌声のF0軌跡

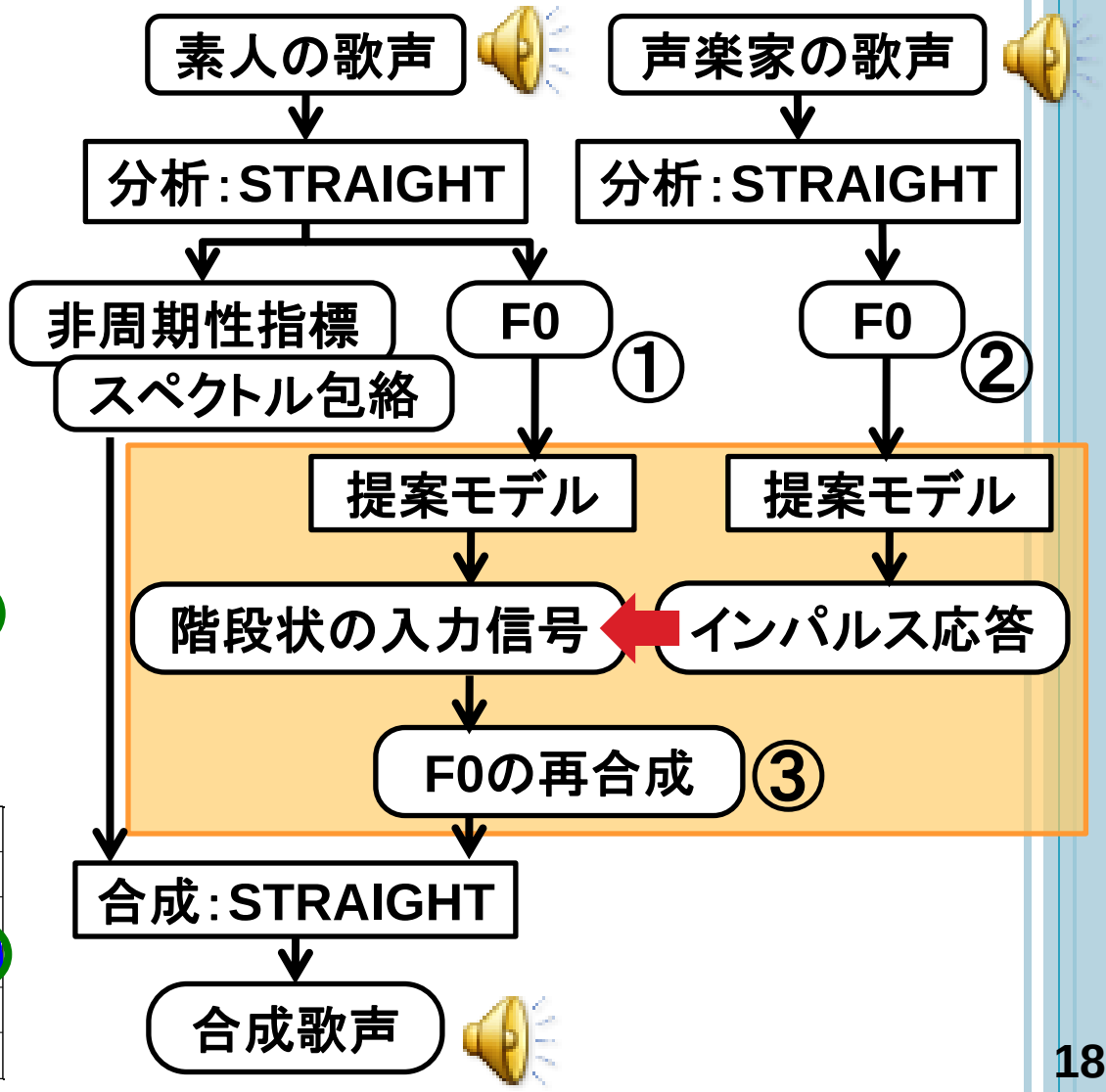


②声楽家の歌声のF0軌跡



③再合成したF0軌跡





まとめと今後の展開

- 歌声のF0制御モデルとパラメータ学習アルゴリズム
  - 階段状の入力信号と動的変動成分への分解
  - 学習アルゴリズムの収束性，信号の推定性能
  - 動的変動成分に基づく歌唱者の個性の調査
- フレーム長や伝達関数の次数を可変にしたパラメータ学習方法
- 歌唱者の歌い方を反映した歌声合成，歌唱力評価などへの応用
- 楽器音や生体信号などの時系列信号への適用
- 提案モデルを多変量化し，MFCCベクトルの時系列などを動的モデリングして音声分析に利用すること