2008－MUS－76 (16)
2008／8／7

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† 亀岡 弘和†† 柏野 邦夫†† 武田 一哉†

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

†ohishi[at]sp.m.is.nagoya-u.ac.jp, kazuya.takeda[at]nagoya-u.jp
††kameoka[at]eye.brl.ntt.co.jp, kunio[at]eye.brl.ntt.co.jp

あらまし 歌声の基本周波数 (F0) 軌跡から，歌唱者が意図する旋律概形と歌声の動的変動成分を同時推定する手法を提案する．これまで，旋律概形を表す区分的に一定な階段状の入力信号に，ビブラートやオーバーシュートなどの動的変動因子を表す2次系のインパルス応答を畳み込むことによって，F0軌跡を生成するための制御モデルが提案された．しかし，観測されるF0軌跡だけから，それぞれの信号を推定する逆問題は不良設定問題であるため，従来のモデルではこの問題を解くことができなかった．そこで，我々は階段状の拘束をもつ特殊な入力信号を隠れマルコフモデル (HMM) でモデル化し，2次系を含むシステムの伝達関数を全極モデルで表現することで，Viterbi学習と線形予測分析 (LPC) 的な解法の反復により，モデルパラメータを効率的に推定するアルゴリズムを提案する．本稿ではその定式化と実装を行い，観測されるF0軌跡から旋律概形と動的変動成分をともに推定できること，さらに推定されたパラメータによってF0軌跡を生成可能であることを確認する．

Parameter Estimation Method of F0 Control Model for Singing Voices based on Convolutive HMM

Yasunori Ohishi† Hirokazu Kameoka†† Kunio Kashino†† Kazuya Takeda†

†Graduate School of Information Science, Nagoya University
††NTT Communication Science Laboratories, NTT Corporation

Abstract In this paper, we propose a novel representation of F0 contours that provides a computationally efficient algorithm for automatically estimating the parameters of a F0 control model for singing voices. Although the best known F0 control model, based on a second-order system with a piece-wise constant function as its input, can generate F0 contours of natural singing voices, this model has no means of learning the model parameters from observed F0 contours automatically. Therefore, by modeling the piece-wise constant function by Hidden Markov Models (HMM) and approximating the transfer function of the system by the all-pole model, we estimate model parameters optimally based on iteration of Viterbi training and an LPC-like solver. Our representation is a generative model and can identify both the target musical note sequence and the dynamics of singing behaviors included in the F0 contours. Our experimental results show that the proposed method can separate the dynamics from the target musical note sequence and generate the F0 contours using estimated model parameters.

1 はじめに

本研究では，歌声のF0軌跡から，歌唱者が意図する旋律概形とビブラートやオーバーシュートのような歌声特有の動的変動成分を特徴付ける信号モデルの構築を目指す．歌声のF0軌跡は，階段状の旋律概形に様々な動的変動成分が複雑に重ね合わさった状態で観測され，この変動成分は，歌声の個人性知覚に影響を与えるという知見が報告されている [1, 2]．したがって，旋律に相当する譜面情報に加え，譜面には載っていない歌唱者の個性や感情のような非譜面情報が，F0軌跡に含まれていると考えられる．旋律概形は，ハミング検索や歌声による採譜において必要となる情報である．従来は，F0軌跡を音高と音長を表すシンボル列に変換し，n-gramモデルのような離散的な確率表現を利用してモデル化することが一般的であった [3, 4, 5, 6]．しかし，歌唱者の歌い方や個性による動的変動成分の影響を受けて，旋律を正確にシンボル列で表現することが難しく，性能が低下する問題があった．旋律を正しくシンボル列に変換するためには，動的変動成分を無視することなく，変動成分自身を適切にモデリングする必要がある．一方，この動的変動成分は，人間らしい歌声を作るための歌声合成において必要不可欠な情

報である．これまで，区分的に一定な階段状の入力信号と動的変動因子を表す 2 次系のインパルス応答との畳み込みによる F0 制御モデルが提案され，人間らしい歌声の合成が可能となりつつある [7]．しかし，制御モデルのパラメータは聴取実験に基づいて手動で決定されるものであった．したがって，歌声から制御パラメータを自動的に決定することが期待される．また，サンプリングした歌声の素片を連結する方式 [8]，隠れマルコフモデル (HMM) を利用して合成する方式 [9]，歌声を入力してそれを近似する合成パラメータを推定する方式 [10, 11] が提案され，表情豊かな歌声合成が実現された．知覚に影響を与える動的変動成分を精緻にモデル化する技術が実現されれば，さらなる性能の向上を期待できる．

そこで，我々はこれまで，歌声の F0 軌跡から旋律情報と動的変動成分を同定するために相平面を利用した．相平面に描かれる F0 の渦軌跡を確率分布で表現することで，様々な変動成分に対して頑健に，旋律情報を推定できることを，ハミング検索性能の観点から確認した [12]．しかし，相平面の確率的表現だけでは，変動成分を精緻にモデル化できたとはいえなかった．

本稿では，文献 [7] の F0 制御モデルを参考にして，F0 軌跡から旋律概形を表す階段状の入力信号と動的変動因子を表すインパルス応答を同時に推定 (分離) することを考える．ただ，譜面情報無しの状態で，観測される F0 軌跡だけから，それぞれの信号を推定する逆問題はそもそも不良設定問題であるため，従来のモデルだけではこの問題を解くことができない．そこで，我々は階段状の拘束をもつ入力信号を HMM でモデル化し，2 次系を含むシステムの伝達関数を全極モデルで表現することによる制約条件を設けた上で，Viterbi 学習と線形予測分析 (LPC) 的な解法の反復により，モデルパラメータを効率的に推定するアルゴリズムを提案する．その定式化と実装を行い，提案手法によって，F0 軌跡から旋律概形と動的変動成分をともに推定できること，さらに推定されたモデルパラメータによって，F0 軌跡を生成可能であることを確認する．また，動的変動因子を表すシステムの伝達関数から歌唱者の個性について検証し，歌唱スタイルの転写や歌唱力評価などの応用への可能性を示す．以下，2 章では歌声の F0 制御モデルを提案する．3 章では，F0 制御モデルのパラメータを最尤推定するためのアルゴリズムを提案する．4 章では，新たに作成した歌声データベースを用いて評価実験を行い，実験結果を考察する．5 章でまとめと今後の展開について述べる．

2　歌声の F0 制御モデルの提案

歌声の F0 軌跡は，図 1 に示すように，旋律概形を表す入力信号にフィルタのインパルス応答が畳み込まれた状態で観測されると想定する．以下にそのモデル化のための 3 つのアイデアを述べる．

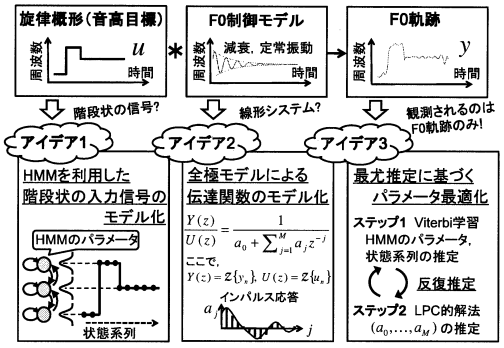

図 1　提案する歌声の F0 制御モデルの概略図

アイデア 1　旋律概形を表す階段状の入力信号は，HMM によってモデル化される．

アイデア 2　F0 制御モデルは，全極モデルで表される伝達関数 (式 (6)) によってモデル化される．

アイデア 3　各モデルパラメータは，最尤推定に基づく以下の 2 つのステップの反復推定 によって最適化される．

- **ステップ 1**　状態系列と HMM のパラメータは Viterbi 学習によって推定される．
- **ステップ 2**　インパルス応答 $a_0, \ldots, a_M$ は，LPC 的な解法によって推定される．

以下の節で各アイデアの詳細を述べる．

2.1　全極モデルによるシステムの伝達関数

階段状の拘束をもつ対数 F0 軌跡を入力信号 u_n とし，F0 制御モデルの動的特性を表すインパルス応答を a_n とすると，フィルタ出力は $u_n * a_n$ ($*$ は畳み込みを表す．) であり，このフィルタ出力から入力を差し引いた $u_n * a_n - u_n$ は，フィルタを介して付加されたビブラートやオーバーシュートなどを含む動的変動成分を表す．u_n，a_n の z 変換を $U(z)$，$A(z)$ とすると，

$$\begin{array}{ll} \textbf{入力信号} & U(z) \\ \textbf{動的変動成分} & U(z)A(z) - U(z) = U(z)(A(z) - 1) \end{array} \tag{1}$$

と書ける．さて，まずここで，$A(z)$ が，音声分析などの用途で広く用いられる自己回帰 (AR) モデル ($a_0 = 1$ という拘束をもつ全極モデル)

$$A(z) = \frac{1}{1 + \sum_{j=1}^{M} a_j z^{-j}} \tag{2}$$

で与えられる場合について考える．このとき，式 (1) から明らかなように，動的変動成分のスケールは入力の大きさに比例して決定される．すなわち，この場合

の F0 制御モデルは，入力される対数 F0 値の大きさに比例した振れ幅のビブラートやオーバーシュートしか付与することができないという不適当な制約を必然的にもってしまうことになる．音声分析の用途においては，零平均の Gauss 性雑音の入力信号を仮定し，フィルタと入力のスケールの任意性を除く目的で通常 $a_0 = 1$ と置くわけであるが，これに対し，ここで我々が考えたい F0 制御モデルには，動的変動成分のスケールを入力音高信号に依らず自由に調節できる機構が備わっていることが望ましい．すなわち，式 (1) のうち，動的変動成分のみを c 倍したもの (c が動的変動成分のスケールを決定するパラメータ) と入力信号とを足したものが出力されるフィルタの伝達関数を新たに考えたい．入力信号と c 倍された動的変動成分の和は，

$$U(z)+cU(z)(A(z)-1) = (cA(z)+(1-c))U(z) \quad (3)$$

と書けるが，式 (3) を $U(z)$ を入力とした伝達関数 $B(z)$ のフィルタ出力と考えると，

$$B(z)U(z) = (cA(z) + (1-c))U(z) \quad (4)$$

なので，

$$B(z) = cA(z) + (1-c) \quad (5)$$

となり，$A(z)$ に $(1-c)/c$ という相対的なバイアス成分が付加されたものとなる．z 領域でのバイアスとは，時間領域の $n = 0$ の振幅値に他ならないので，$B(z)$ の時間領域表現を b_n とすると，b_0 には c をパラメータ化した分の自由度が生まれる．以上のことを逆に考えれば，a_0 をパラメータ化することで，ビブラート等の動的変動成分のスケールを入力信号とは独立に調節できるようになる．すなわち，音声分析の用途では通常定数として扱われる a_0 を，自由パラメータとして扱うことがここでは本質的に重要な意味をもつのである．以上より，F0 制御モデルを，以下の全極モデルで表される伝達関数によってモデル化する．

$$\frac{Y(z)}{U(z)} = \frac{1}{a_0 + \sum_{j=1}^{M} a_j z^{-j}} \quad (6)$$

ここで，観測される F0 軌跡 y_n の z 変換を $Y(z)$，伝達関数の次数を M と表す．式 (6) を逆 z 変換すると，

$$a_0 y_n + \sum_{j=1}^{M} a_j y_{n-j} = u_n \quad (7)$$

となり，入出力関係を表す差分方程式が導かれる．

いま，F0 軌跡と入力信号の関係は式 (7) の差分方程式に従うと想定するが，実際にはこれにさらに微細変動成分 [7] が加わる．この微細変動成分を平均 0，分散 σ^2 の正規分布に従う Gauss 性白色雑音 ε とし，$\hat{u}_n = \sum_{j=0}^{M} a_j y_{n-j}$ と u_n との間に，

$$\hat{u}_n - u_n = \varepsilon, \quad (\varepsilon \sim \mathcal{N}(0, \sigma^2)) \quad (8)$$

を仮定する．

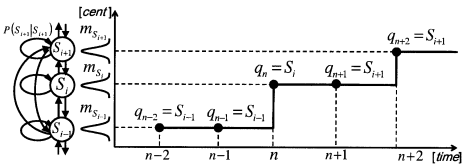

図 2　階段状の入力信号のモデル化

2.2　階段状の入力信号のモデル化

階段状の拘束をもつ入力信号 u_n は，図 2 に示す状態集合 $S = \{S_1, \ldots, S_I\}$ からなる HMM によってモデル化される．1 回の状態遷移により実数が 1 個生成されるモデルであり，観測される F0 軌跡から状態遷移 (どの音高間でどのタイミングで遷移したか) を一意に決定できない．つまり，

$$u_n = m_{q_n}, \ (m_{S_i} \in \mathbb{R}, \ q_n \in \{S_1, \ldots, S_I\}) \quad (9)$$

と表し，ここで m_{S_i} は状態 S_i における出力確率分布 (式 (8) で仮定する正規分布に相当する．) の平均を表す．一様な Markov 連鎖を想定し，状態 S_j から S_i への遷移確率は $\mathbb{P}(S_i|S_j)$ と表す．q_n は状態集合 S の要素の中のいずれかの値をとる．状態系列 $q_1, \ldots, q_N$ と，各状態における出力確率分布の平均 m_{q_n} によって，階段状の入力信号が生成される．

2.3　各モデルパラメータの解釈

インパルス応答 $a_0, \ldots, a_M$ は，歌声の動的変動因子を表すパラメータである．音高が安定するときの振動や，音高目標に到るまでの連続的な音高遷移が表現される．状態系列 $q_1, \ldots, q_N$ は，音長を決定するパラメータである．これは，必ずしも譜面に記される音符の音長に対応するわけではなく，歌唱者の意図やスタイルに基づいて生成される運動指令の長さを表現したものであるとここでは想定する．最後に，HMM の各状態の出力確率分布の平均 $m_{S_1}, \ldots, m_{S_I}$ は，歌唱者が意図する旋律の音高 (音高目標値) に対応するパラメータである．これは，必ずしも譜面に記される音符の音高 (客観的に定まっている音高値) には対応しない．

2.4　時変な F0 制御モデルへの拡張

図 3 より，旋律の変化とともに，ビブラートやオーバーシュートのような動的変動成分が次々と変化することがわかる．これは，図 1 の F0 制御モデルのインパルス応答が，時々刻々と変化するためであると考えられる．そこで，長さ N の F0 軌跡 $\mathbf{y} = \{y_1, \ldots, y_N\}$ と入力信号 $\mathbf{u} = \{m_{q_1}, \ldots, m_{q_N}\}$ をフレームに分割し，フレームごとにインパルス応答を推定する．長さ L のフレーム t における F0 軌跡と入力信号をそれぞれ $\mathbf{y}^{(t)} = \{y_1^{(t)}, \ldots, y_L^{(t)}\}$，$\mathbf{u}^{(t)} = \{m_{q_1^{(t)}}, \ldots, m_{q_L^{(t)}}\}$，インパルス応答を $\boldsymbol{\theta}^{(t)} = \{a_0^{(t)}, \ldots, a_M^{(t)}\}$ と表す．

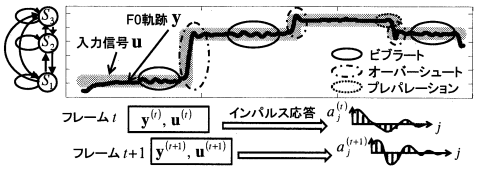

図 3　時変な F0 制御モデルへの拡張

ここで入力信号を生成する HMM のパラメータ $m_{S_1},\ldots,m_{S_I}$ は T 個のフレームに関して共通するものと考える．この HMM のパラメータと状態系列を合わせて，$\boldsymbol{\omega}=\{q_1,\ldots,q_N,m_{S_1},\ldots,m_{S_I}\}$ と表すと，推定しなければならないパラメータの集合は，$\boldsymbol{\Theta}=\{\boldsymbol{\theta}^{(1)},\ldots,\boldsymbol{\theta}^{(T)},\boldsymbol{\omega}\}$ となる．

3　F0 制御モデルのパラメータ最尤推定

F0 軌跡の集合 $\{\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(T)}\}$ が与えられたとき，パラメータ集合 $\boldsymbol{\Theta}$ を最尤推定する方法を述べる．集合 $\{\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(T)}\}$ の対数尤度は，式 (8) より，

$$\begin{aligned}\log\mathbb{P}(\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(T)}|\boldsymbol{\Theta}) &= \sum_{t=1}^{T}\log\mathbb{P}(\mathbf{y}^{(t)}|\boldsymbol{\theta}^{(t)},\boldsymbol{\omega})\\ &= \sum_{t=1}^{T}\Big\{-(L-M)\log(\sqrt{2\pi}\sigma^{(t)}a_0^{(t)})\\ &\quad -\frac{1}{2\sigma^{(t)2}}\sum_{l=M+1}^{L}\Big(\sum_{j=0}^{M}a_j^{(t)}y_{l-j}^{(t)}-m_{q_l^{(t)}}\Big)^2\Big\}\end{aligned}\tag{10}$$

と記述される．ここで各フレームの F0 軌跡は独立に観測されるものと想定する．また，$\sigma^{(t)2}$ はフレーム t において予測誤差が従う正規分布の分散値を表す．パラメータの集合 $\boldsymbol{\Theta}$ の事後確率は，$\mathbb{P}(\boldsymbol{\Theta}|\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(T)})\propto\mathbb{P}(\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(T)}|\boldsymbol{\Theta})\mathbb{P}(\boldsymbol{\Theta})$ なので，$U(\boldsymbol{\Theta})\triangleq\mathbb{P}(\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(T)}|\boldsymbol{\Theta})\mathbb{P}(\boldsymbol{\Theta})$ とおくと，その対数値は，

$$\begin{aligned}\log U(\boldsymbol{\Theta}) &= \sum_{t=1}^{T}\log\mathbb{P}(\mathbf{y}^{(t)}|\boldsymbol{\theta}^{(t)},\boldsymbol{\omega})+\sum_{t=1}^{T}\log\mathbb{P}(\boldsymbol{\theta}^{(t)})\\ &\quad+\log\mathbb{P}(m_{S_1},\ldots,m_{S_I})+\log\mathbb{P}(q_1,\ldots,q_N)\end{aligned}\tag{11}$$

と記述できる．ここで事前確率 $\mathbb{P}(\boldsymbol{\theta}^{(t)})$ と $\mathbb{P}(m_{S_1},\ldots,m_{S_I})$ は一様分布を想定し，$\mathbb{P}(q_1,\ldots,q_N)$ は，HMM による一様な Markov 連鎖を想定するため，

$$\log\mathbb{P}(q_1,\ldots,q_N)=\log\mathbb{P}(q_1)\mathbb{P}(q_2|q_1)\ldots\mathbb{P}(q_N|q_{N-1})\tag{12}$$

となる．遷移確率 $\mathbb{P}(S_i|S_j)$ は事前に決定する定数である．簡単のため，$P_{S_i,S_j}\equiv\log\mathbb{P}(S_i|S_j)$ とおく．以上より，式 (10) と式 (12) を式 (11) に代入し，定数項を除いた以下の式

$$\begin{aligned}J &\equiv -(L-M)\sum_{t=1}^{T}\log\sigma^{(t)}a_0^{(t)}\\ &\quad-\sum_{t=1}^{T}\frac{1}{2\sigma^{(t)2}}\sum_{l=M+1}^{L}\Big(\sum_{j=0}^{M}a_j^{(t)}y_{l-j}^{(t)}-m_{q_l^{(t)}}\Big)^2+\sum_{n=1}^{N}P_{q_n,q_{n-1}}\end{aligned}\tag{13}$$

がパラメータの集合 $\boldsymbol{\Theta}$ に関して最大化したい目的関数である．しかし，式 (13) を最大化する $\boldsymbol{\Theta}$ は解析的に求めることができない．そこで，$\boldsymbol{\Theta}$ の各要素に関して，他の要素を固定したもとで式 (13) を最大化するステップを式 (13) の値が収束するまで繰り返す (図 4)．

3.1　インパルス応答 $\boldsymbol{\theta}^{(t)}$ の推定

パラメータの集合 $\boldsymbol{\omega}$ を固定して，式 (13) を最大にする $\boldsymbol{\theta}^{(t)}$ を求める．J を $a_{j'}^{(t)}$ に関して偏微分すると，

$$\frac{\partial J}{\partial a_{j'}^{(t)}}=\begin{cases}-\dfrac{L-M}{a_0^{(t)}}-\dfrac{1}{\sigma^{(t)2}}\displaystyle\sum_{l=M+1}^{L}\Big(\sum_{j=0}^{M}a_j^{(t)}y_{l-j}^{(t)}-m_{q_l^{(t)}}\Big)y_l^{(t)}\\ \qquad(j'=0\text{ のとき})\\ -\dfrac{1}{\sigma^{(t)2}}\displaystyle\sum_{l=M+1}^{L}\Big(\sum_{j=0}^{M}a_j^{(t)}y_{l-j}^{(t)}-m_{q_l^{(t)}}\Big)y_{l-j'}^{(t)}\\ \qquad(j'=1,\ldots,M\text{ のとき})\end{cases}\tag{14}$$

となる．式 (10) の正規分布の正規化項に動的変動成分のスケール $a_0^{(t)}$ を含むため，j' が 0 とそれ以外で，偏微分の式が異なる．J を $\sigma^{(t)}$ に関して偏微分すると，

$$\frac{\partial J}{\partial\sigma^{(t)}}=-\frac{L-M}{\sigma^{(t)}}+\frac{1}{\sigma^{(t)3}}\sum_{l=M+1}^{L}\Big(\sum_{j=0}^{M}a_j^{(t)}y_{l-j}^{(t)}-m_{q_l^{(t)}}\Big)^2\tag{15}$$

となる．式 (14) と式 (15) を 0 とおくと，

$$a_0^{(t)}\sum_{l=M+1}^{L}\Big(\sum_{j=0}^{M}a_j^{(t)}y_{l-j}^{(t)}-m_{q_l^{(t)}}\Big)y_l^{(t)}+\sigma^{(t)2}(L-M)=0\qquad(j'=0\text{ のとき})\tag{16}$$

$$\sum_{l=M+1}^{L}\Big(\sum_{j=0}^{M}a_j^{(t)}y_{l-j}^{(t)}-m_{q_l^{(t)}}\Big)y_{l-j'}^{(t)}=0\qquad(j'=1,\ldots,M\text{ のとき})\tag{17}$$

$$\sigma^{(t)2}=\frac{1}{L-M}\sum_{l=M+1}^{L}\Big(\sum_{j=0}^{M}a_j^{(t)}y_{l-j}^{(t)}-m_{q_l^{(t)}}\Big)^2\tag{18}$$

となり，これらの連立方程式を解析的に解くことはできない．そこで，まず，式 (16) の $a_1^{(t)},\ldots,a_M^{(t)},\sigma^{(t)}$ を固定して，$a_0^{(t)}$ に関する 2 次方程式を解析的に解く．次に，式 (17) の $a_0^{(t)}$ を固定して，$a_1^{(t)},\ldots,a_M^{(t)}$ に関する連立方程式を解析的に解く．最後に式 (18) の $\sigma^{(t)2}$ を計算する．この計算過程を目的関数 J が収束するまで順番に繰り返して，$a_0^{(t)},\ldots,a_M^{(t)},\sigma^{(t)}$ を更新させる．

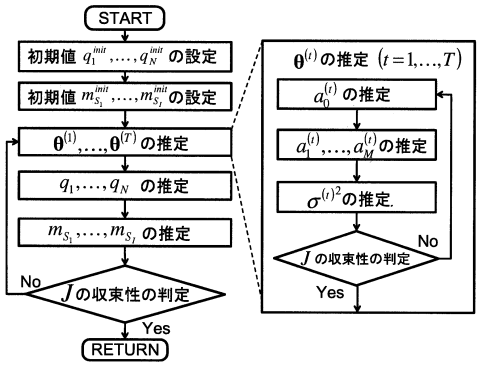

図 4　反復法によるモデルパラメータの推定手順

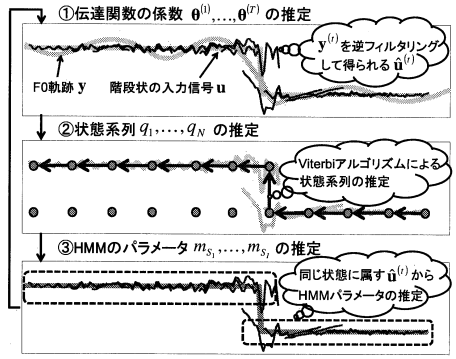

図 5　パラメータ推定のイメージ図

以上の推定過程のイメージを図 5 の上段に示す．$\boldsymbol{\theta}^{(t)}$ を推定することによって，F0 軌跡 $\mathbf{y}^{(t)}$ を逆フィルタリングした $\hat{\mathbf{u}}^{(t)}$ が推定される．

3.2　状態系列 $q_1,\ldots,q_N$ の推定

インパルス応答の集合 $\{\boldsymbol{\theta}^{(1)},\ldots,\boldsymbol{\theta}^{(T)}\}$ と HMM のパラメータ $m_{S_1},\ldots,m_{S_I}$ を固定して，状態系列 $q_1,\ldots,q_N$ に関して，式 (13) の最大化を考える．この問題は，Viterbi アルゴリズム (動的計画法) により効率的に解ける．最初から時刻 k に状態 S_i に至るまでの最適な状態系列は，漸化式を利用して，

$$\delta_k(S_i)=\max_{S_j}\Big[\delta_{k-1}(S_j) - \sum_{(t,l)\in\mathcal{C}_k}\frac{1}{2\sigma^{(t)2}}(\hat{u}_l^{(t)}-m_{S_i})^2+P_{S_i,S_j}\Big] \quad (\mathcal{C}_k=\{(t,l)|y_k\in\mathbf{y}^{(t)},1\le t\le T,1\le l\le L\}) \tag{19}$$

と記述できる．ここで，$\mathcal{C}_k$ は，1 から T までのフレームの中で，y_k に相当する $y_l^{(t)}$ のフレーム番号 t とインデクス l の組み合わせからなる集合である．式 (19) を $k=N$ まで計算すれば，最適な経路が求められる．図 5 の中段に示すように，前節で推定された $\hat{\mathbf{u}}^{(t)}$ とトレリス構造に基づいて，最適な状態系列が求められる．

3.3　HMM のパラメータの推定

インパルス応答の集合 $\{\boldsymbol{\theta}^{(1)},\ldots,\boldsymbol{\theta}^{(T)}\}$ と状態系列 $q_1,\ldots,q_N$ を固定して，式 (13) が最大になるように HMM のパラメータ $m_{S_1},\ldots,m_{S_I}$ を推定する．J を m_{S_i} に関して偏微分すると，

$$\frac{\partial J}{\partial m_{S_i}}=-\sum_{n\in\mathcal{D}_i}\sum_{(t,l)\in\mathcal{C}_n}\frac{1}{2\sigma^{(t)2}}\left(\hat{u}_l^{(t)}-m_{S_i}\right)=0 \tag{20}$$

となり，

$$m_{S_i}=\frac{1}{|\mathcal{D}_i|}\sum_{n\in\mathcal{D}_i}\frac{1}{|\mathcal{C}_n|}\sum_{(t,l)\in\mathcal{C}_n}\hat{u}_l^{(t)} \tag{21}$$

を得る．ここで，集合 $\mathcal{D}_i=\{n|q_n=S_i\}$ とし，$|\mathcal{D}_i|$ をその要素数とする．図 5 の下段に示すように，3.2 節で推定された状態系列に基づいて，同じ状態 S_i に属する $\hat{u}_l^{(t)}$ を集めて，その状態の HMM のパラメータ m_{S_i} を推定する．最終的に，3.2 節と 3.3 節で推定された状態系列 $q_1,\ldots,q_N$ と HMM のパラメータ $m_{S_1},\ldots,m_{S_I}$ から，階段状の入力信号 $u_n=m_{q_n},(n=1,2,\ldots,N)$ が決定される．

3.4　初期値の設定

3.1 節，3.2 節，3.3 節で述べた 3 段階のパラメータ推定を，目的関数 J が収束するまで順番に繰り返す (図 4)．ただ，局所解に収束することを防ぐために，以下の 2 つの初期値設定を行う．まず，観測系列 $\mathbf{y}$ に 3.2 節で述べた Viterbi アルゴリズムを適用して初期状態系列を決定する．目的関数 J_{init}

$$J_{\text{init}}=-\frac{1}{2\sigma_{\text{init}}^2}\sum_{n=1}^{N}(y_n-m_{q_n})^2+\sum_{n=1}^{N}P_{q_n,q_{n-1}} \tag{22}$$

を最大化する状態系列を初期状態系列 $q_1^{\text{init}},\ldots,q_N^{\text{init}}$ とする．次に，3.3 節と同様に，初期状態系列から HMM のパラメータを決定する．すなわち，

$$m_{S_i}=\frac{1}{|\mathcal{D}_i^{\text{init}}|}\sum_{n\in\mathcal{D}_i^{\text{init}}}y_n,\quad(\mathcal{D}_i^{\text{init}}=\{n|q_n^{\text{init}}=S_i\}) \tag{23}$$

を初期の HMM パラメータ $m_{S_1}^{\text{init}},\ldots,m_{S_I}^{\text{init}}$ とする．

3.5　提案手法と類似したモデル

提案手法は F0 軌跡から旋律概形と動的変動成分を同時に推定するという問題設定から導かれたモデルであるが，これと類似するモデルが音声分析という全く異なる目的から生まれていることは興味深い．高基本周波数の音声を分析する場合，推定された声道特性スペクトルのピークが声帯振動に起因する調波構造に偏

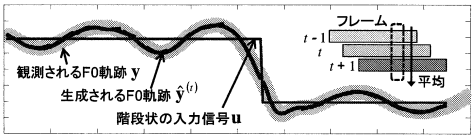

図 6　F0 軌跡の生成

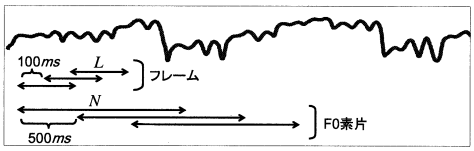

図 7　フレーム化処理の概略図

るため，フォルマント構造を正しく抽出できないという問題があった．佐宗らは，これを解決するために，声道特性は AR モデルを用い，声門音源は状態を環状に接続した HMM を用いて音声信号をモデル化する AR HMM を提案した [13, 14]．両者を比較すると，問題設定が異なるとともに，2.1 節で述べたように対象とする信号が異なるため，そのモデル化手法 (具体的には，スケール可変の AR モデルを導入した点) も異なる．さらに提案手法では，図 3 に示すように時々刻々と変化する動的変動成分をモデル化するために，フレーム化処理を含めて定式化を行った．以上を踏まえて，階段状の入力信号を HMM によってモデル化し，そのパラメータ学習方法が入力信号と F0 軌跡に基づく逆畳み込みと Viterbi 学習から成ることから，提案モデルを先行研究と区別して，**畳み込み HMM(Convolutive HMM)** に基づく歌声の F0 制御モデルと呼ぶ．

3.6　F0 軌跡の生成

フレーム t において推定されたインパルス応答 $\boldsymbol{\theta}^{(t)}$ と状態系列 $q_1^{(t)},\ldots,q_L^{(t)}$，HMM のパラメータ $m_{S_1},\ldots,m_{S_I}$ から，F0 軌跡を生成する．式 (7) より，フレーム t において，生成される F0 軌跡 $\hat{\mathbf{y}}^{(t)}$ は，

$$\hat{y}_l^{(t)} = \frac{1}{a_0^{(t)}}\left(u_l^{(t)} - \sum_{j=1}^{M} a_j^{(t)}\hat{y}_{l-j}^{(t)}\right),\quad (l \geq M+1) \quad (24)$$

から計算される．ただし，$\hat{y}_l^{(t)} = y_l^{(t)}$，$(l \leq M)$ とする．フレーム化処理のため，生成される F0 が時間的に重複する部分は，平均をとることによって平滑化すればよい (図 6)．

4　評価実験

畳み込み HMM に基づく F0 制御モデルのパラメータ学習アルゴリズムの収束性と，各フレームにおいて推定される階段状の入力信号 $\mathbf{u}^{(t)}$ とインパルス応答 $\boldsymbol{\theta}^{(t)}$ から生成される F0 軌跡 $\hat{\mathbf{y}}^{(t)}$ の性能を評価する．

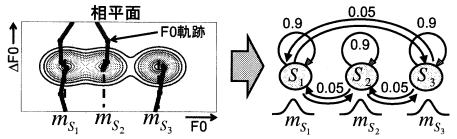

図 8　HMM の状態数の決定方法

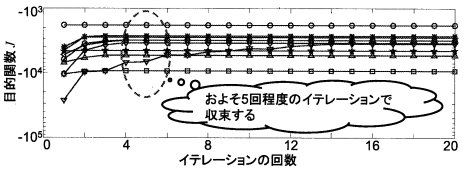

図 9　パラメータ学習アルゴリズムの収束性の評価

さらに，伝達関数の振幅応答から歌唱者ごとの動的変動成分の違いについて検証する．

4.1　歌声データベース

クラシックの声楽家，プロのポップス歌手，音楽的な訓練を受けていない素人 (それぞれ男女 1 名ずつの計 6 名) の歌声からなるデータベースを構築した．歌唱者は，ヘッドフォンで歌唱曲の旋律 (ガイドトーン) を聴きながら伴奏なしで歌唱した．歌唱曲は，「きらきら星」，「喜びの歌 (Beethoven の交響曲第 9 番第 4 楽章の歌の部分を岩佐東一郎氏によって作詞されたもの)」である．日本語歌詞による歌唱，ハミングによる歌唱の 2 パターンを収録した．

4.2　実験条件

F0 は，de Cheveigné らの提案した YIN[15] を利用して 10ms ごとに推定される．なお，Hz で表される周波数 y_{Hz} を，次のように，cent で表される対数スケールの周波数 y_{cent} に変換する．

$$y_{cent} = 1200\log_2 \frac{y_{Hz}}{440 \times 2^{\frac{3}{12}-5}} \quad (25)$$

200ms 以内の F0 が推定されない区間は，無声音の区間とみなし，F0 を線形補間する．また，ガイドトーンの音響信号からも F0 を推定し，提案手法によって推定される階段状の入力信号との比較のために利用する．

図 7 に示すように，F0 軌跡は，まず長さ N の素片に分割される．本実験では $N = 2s$ とする．この素片が 2.4 節で述べた $\mathbf{y}$ に相当し，素片ごとに提案モデルを適用する．素片の中に F0 が推定されない区間 (休符区間) が含まれていた場合，その素片は分析しない．各素片において，フレームは $100ms$ ずつシフトさせる．図 8 に示すように，各素片における HMM の状態数 I は相平面を利用して決定する．相平面に描かれる F0 軌跡のアトラクタを混合ガウス分布 (GMM) に

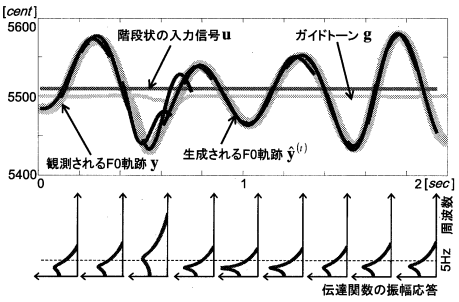

図 10　推定結果の例 1

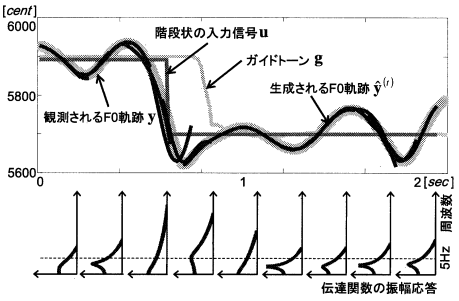

図 11　推定結果の例 2

よって表現し，その極大値の数を HMM の状態数 I とする [12]．各状態の初期状態確率は $1/I$，自己遷移確率は 0.9，異なる状態への遷移確率は $0.1/(I-1)$ と手動で決定した．提案手法の初期値設定における σ^2_{init} は 2500 とした．

4.3　パラメータ学習アルゴリズムの収束性

素片をランダムに 10 個選択し，パラメータ集合 $\boldsymbol{\Theta}$ の反復推定における目的関数 J の収束性を確認した．図 9 より，およそ 5 回程度のイテレーションで目的関数 J が収束することがわかる．イテレーション開始から J の値がほぼ変化しない素片も存在した．

4.4　フレーム長と伝達関数の次数の評価

推定結果の例を図 10，11 に示す．フレーム長 L は $250ms$，伝達関数の次数 M は 3 とした．音高遷移の部分を除いて，入力信号 $\mathbf{u}$ は対応する区間のガイドトーンの F0 軌跡 $\mathbf{g}$ と近いものが推定される．しかし，F0 軌跡 $\mathbf{y}^{(t)}$ と生成される F0 軌跡 $\hat{\mathbf{y}}^{(t)}$ との間に誤差が生じる箇所がいくつか存在する．それでも図 10，11 は比較的小さい誤差で $\hat{\mathbf{y}}^{(t)}$ が生成された結果であり，素片によっては $\hat{\mathbf{y}}^{(t)}$ が発散してしまうフレームも存在した．以上を踏まえて，L と M を変化させたときの各フレームの $\mathbf{u}^{(t)}$ と $\hat{\mathbf{y}}^{(t)}$ の推定性能を評価した．各フレームごとに以下の平均二乗誤差 (RMS)

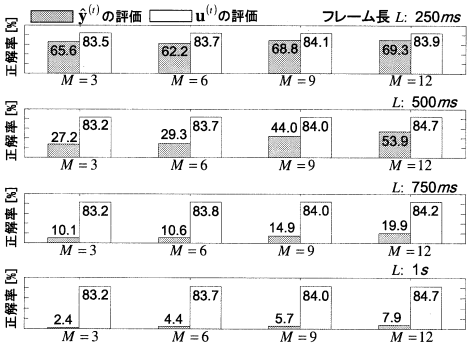

図 12　フレーム長と伝達関数の次数に関する評価

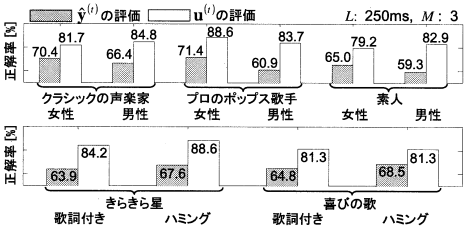

図 13　歌唱者，歌唱曲ごとの推定性能

$$\mathrm{RMS}_{\mathbf{u}^{(t)}} = \sqrt{\frac{1}{L-M}\sum_{l=M+1}^{L}(u_l^{(t)} - g_l^{(t)})^2}$$
$$\mathrm{RMS}_{\hat{\mathbf{y}}^{(t)}} = \sqrt{\frac{1}{L-M}\sum_{l=M+1}^{L}(\hat{y}_l^{(t)} - y_l^{(t)})^2} \tag{26}$$

を計算し，すべてのフレームのうち正しく推定された割合を正解率として以下に定義する．

$$\text{正解率}_{\mathbf{u}^{(t)}} = \frac{\mathrm{RMS}_{\mathbf{u}^{(t)}} < \xi\text{となるフレーム数}}{\text{すべての素片の総フレーム数}} \times 100$$
$$\text{正解率}_{\hat{\mathbf{y}}^{(t)}} = \frac{\mathrm{RMS}_{\hat{\mathbf{y}}^{(t)}} < 10\text{ となるフレーム数}}{\text{すべての素片の総フレーム数}} \times 100 \tag{27}$$

ここで，ξ は各素片の $\mathbf{y}$ と $\mathbf{g}$ の RMS とする．ガイドトーンを聴きながら歌唱しても，歌唱者によってはガイドトーンからずれて歌唱するため，ξ 以内で $\mathbf{u}^{(t)}$ が推定されれば正解とした．$\hat{\mathbf{y}}^{(t)}$ は，$\mathrm{RMS}_{\hat{\mathbf{y}}^{(t)}}$ が $10cent$ 以内で生成されれば正解とした．

推定性能を図 12 に示す．L を短くするにつれて，$\hat{\mathbf{y}}^{(t)}$ の正解率が向上した．つまり，時変な F0 制御モデルとして，素片をフレームに分割し，各フレームごとに伝達関数を推定することの有効性を確認できた．また，M の増加によって伝達関数の自由度が増すため，正解率が向上した．しかし，それでも総フレーム数の 3 割は，伝達関数が適切に推定されず，F0 軌跡が発散して生成されてしまった．また，L を短く，M を大きくしすぎると，素片内で扱うモデルパラメータ

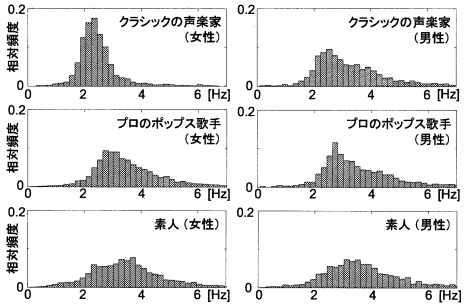

図 14　歌唱者ごとの共振周波数の分布

数が増えてしまう問題もある．したがって，情報量基準を利用して L や M を変化させながら推定を行うなどの改善策が必要とされる．図 13 は歌唱者，歌唱曲ごとの推定性能を示す．フレーム長 L は $250ms$，伝達関数の次数 M は 3 とした．プロの歌手に比べて素人の方が正解率が低い結果が得られた．一方，歌唱曲が変わっても，推定性能には大きく影響しないことがわかった．ただ，歌詞付きの歌唱に比べて，ハミングによる歌唱の方が正解率が高いことがわかる．歌詞付きで歌うと，ビブラートのような 2 次系の動的変動成分に加え，歌詞の発声に基づく非定常な変動成分が含まれるため，推定性能に影響を与えると考えられる．

4.5　伝達関数の振幅応答に基づく歌唱者の個性に関する考察

図 10，11 の下段に式 (6) から計算される伝達関数の振幅応答を示す．音高が安定するフレームでは共振周波数をもち，そのいくつかはビブラート (4〜$7Hz$ 程度の周期的変動 [2]) に対応すると考えられる．そこで，式 (27) で正解とされたフレームの共振周波数の分布を図 14 に示す．歌唱者ごとに分布の形状が異なり，特に女性の声楽家は 2.5Hz あたりに分布が集中した．この歌手は音高が安定するときに顕著に振動させる歌い方をすることを確認した．一方，素人の歌唱者ほど平坦な分布となり，推定された伝達関数の振幅応答から個性や歌唱力の違いを確認することができた．歌い方に基づいた歌唱者識別や歌唱力評価，歌い方を反映した歌声合成などの応用が考えられる．

5　まとめと今後の展開

歌声の F0 軌跡から，階段状の入力信号と動的変動因子を表すインパルス応答を同時推定するための F0 制御モデルとパラメータ学習アルゴリズムを提案した．実験結果より，分析を行った総フレーム数の 69.3% は正確にそれぞれの信号を推定できることを確認した．また，推定される伝達関数の振幅応答から，歌唱者の個性や歌い方を抽出できることを確認した．

さらなる性能向上のためには，フレーム長や伝達関数の次数を可変にしてパラメータを推定することが必要とされる．また，F0 軌跡の動的変動因子は数個のインパルス応答によって表されるものと考え，あらかじめ数個の基底となるインパルス応答を用意し，その重み付き和で各フレームのインパルス応答を表現するなどのモデルの拡張が考えられる．歌唱者の個性や歌い方は基底となるインパルス応答から判断する．本研究の応用先は，F0 軌跡の符号化，歌唱者の歌い方を反映した歌声合成，歌唱力評価などが考えられる．また歌声に限らず，楽器音や生体信号などの時系列信号への適用も考えられる．さらに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なども今後の展開である．

謝辞

本研究は日本学術振興会特別研究員 (DC2) 科研費の補助を受けた．また，これまで本研究に対し，有益なご助言を頂いた後藤真孝氏 (産総研)，伊藤克亘氏 (法政大) に感謝致します．また，F0 制御モデルに関して有益なご意見を頂いた南泰浩氏 (NTTCS 研) に感謝致します．

参考文献

[1] 齋藤　毅ほか：歌声の基本周波数変化に含まれるオーバーシュートの知覚への影響に関する検討，聴覚研資，Vol. 36, No. 7, pp. 611–616 (2006).
[2] 齋藤　毅ほか：歌声の個人性知覚に寄与する音響特徴の検討，音講論集, 2-Q-26，pp. 601–602 (2007).
[3] Song, J. et al.: Mid-Level Music Melody Representation of Polyphonic Audio for Query-by-Humming System, *In Proc. ISMIR 2002* (2002).
[4] Pauws, S.: CubyHum: A fully operational query by humming system, *In Proc. ISMIR 2002* (2002).
[5] Pardo, B. et al.: Name that tune: a pilot study in finding a melody from a sung query, *JASIST*, Vol. 55, No. 4, pp. 283–300 (2004).
[6] Dannenberg, R. B., Birmingham, W. P. et al.: A Comparative Evaluation of Search Techniques for Query-by-Humming Using the MUSART Testbed, *JASIST*, Vol. 58, No. 5, pp. 687–701 (2007).
[7] 齋藤　毅ほか：SingBySpeaking:歌声知覚に重要な音響特徴を制御して話声を歌声に変換するシステム，情処研報音楽情報科学，Vol. 2008, No. 74, pp. 25–32 (2008).
[8] 剣持秀紀ほか:歌声合成システム VOCALOID-現状と課題, 情処研報音楽情報科学，Vol. 2008, No. 74, pp. 51–58 (2008).
[9] 酒向慎司ほか：声質と歌唱スタイルを自動学習可能な歌声合成システム，情処研報音楽情報科学，Vol. 2008, No. 74, pp. 51–58 (2008).
[10] Janer, J. et al.: Performance-Driven Control for Sample-Based Singing Voice Synthesis, *In Proc. DAFx-06*, pp. 42–44 (2006).
[11] 中野倫靖ほか：VocaListener:ユーザ歌唱を真似る歌声合成パラメータを自動推定するシステムの提案，情処研報音楽情報科学，Vol. 2008, No. 75, pp. 49–56 (2008).
[12] 大石康智ほか：相平面を利用した歌声の F0 軌跡の新しい表現方法，電子情報通信学会総合大会，pp. S–51–S–52 (2008).
[13] 佐宗　晃ほか：HMM による音源のモデリングと高基本周波数に頑健な声道特性抽出，電子情報通信学会論文誌 (D-II)，Vol. J84-D-II, No. 9, pp. 1960–1969 (2001).
[14] Sasou, A. et al.: An Auto-Regressive, Non-Stationary Excited Signal Parameter Estimation Method and an Evaluation of a Singing-Voice Recognition, *In Proc. ICASSP 2005*, pp. 237–240 (2005).
[15] de Cheveigné, A. and Kawahara, H.: YIN, a fundamental frequency estimator for speech and music, *Journal of the Acoustical Society of America*, Vol. 111, No. 4, pp. 1917–1930 (2002).