

Vocal Dynamics Controller: 歌声のF0動特性をノート単位で編集し、 合成できるインタフェース

大石 康智^{†1} 亀岡 弘和^{†1} 持橋 大地^{†1}
永野 秀尚^{†1} 柏野 邦夫^{†1}

本報告では、歌声のF0動特性をノート単位で編集し、歌い方を多様に変形できる歌声合成インタフェースの実現を目指し、その動特性のモデリングとモデルパラメータ推定に関する新しい解法を提案する。F0動特性は線形2次系に従うと仮定し、その生成過程を完全に確率モデルとして表現する。そして、EM法に基づいて、効率的なモデルパラメータ最適化アルゴリズムを導出する。最終的に、推定された2次系の振動を制御するパラメータと各ノートの音高を表すパラメータを個別に操作し、生成されたF0系列に基づいて歌声音響信号を変形して合成する“Vocal Dynamics Controller”を実装する。

Vocal Dynamics Controller: A note-by-note editing and synthesizing interface for F0 dynamics in singing voices

YASUNORI OHISHI,^{†1} HIROKAZU KAMEOKA,^{†1}
DAICHI MOCHIHASHI,^{†1} HIDEHISA NAGANO^{†1}
and KUNIO KASHINO^{†1}

We present a novel statistical model for dynamics of various singing behaviors, such as vibrato and overshoot, in a fundamental frequency (F0) sequence and develop a note-by-note editing and synthesizing interface for F0 dynamics. We develop a complete stochastic representation of the F0 dynamics based on a second-order linear system and propose a complete, efficient scheme for parameter estimation using the Expectation-Maximization (EM) algorithm. Finally, we synthesize the singing voice using the F0 sequence generated by manipulating model parameters individually which control the oscillation based on the second-order system and the pitch of each note.

1. はじめに

歌声の声の高さを表す基本周波数 (F0 と呼ぶ) 系列には、階段状に変化する旋律成分と、ビブラートやオーバーシュートのような歌唱者の歌唱動作を表す動的な変動成分が複雑に重ね合わされている (図 1)。この F0 動的特徴成分は、歌声知覚および個人性知覚に寄与する特徴であることが知られており^{1),2)}、歌唱スタイルの記述やそれを利用した歌声検索、歌声評価³⁾⁻⁵⁾ のための有用な尺度となる。また、表情豊かかつ多様な歌声合成⁶⁾⁻⁹⁾ を実現するためにも必要不可欠な成分と言える。我々は、この F0 の動特性 (ダイナミクス) のモデリングに関する基礎検討を行っており^{10),11)}、本報告ではこの動特性をノート単位で編集し、歌い方を多様に変形できる歌声合成インタフェースの実現を目指す。

では、これまで F0 の動特性はどのようにモデル化されてきたか。従来は、この F0 動的特徴成分を表現するために 2 次系モデルを利用した^{1),12),13)}。これらの研究では、日本語の話し声の F0 パターンを表現する藤崎モデル¹⁴⁾ がベースとなっている。藤崎モデルは、臨界制動 2 次系のインパルス応答とステップ応答を利用して、日本語の句頭から句末に向けて緩やかに下降するフレーズ成分と、語句に対応して上昇・下降するアクセント成分を表現し、これらを重畳することで、F0 系列を記述する。ただし、歌声の旋律に伴った急激な F0 の上昇・下降の制御及び、ビブラートのような周期的な振動は、臨界制動系では表現できない。そのため、歌声の F0 制御モデルでは 2 次系の伝達関数

$$H(s) = \frac{\Omega^2}{s^2 + 2\zeta\Omega s + \Omega^2} \quad (1)$$

の減衰率 ζ によって表現される、指数減衰 ($\zeta > 1$)、減衰振動 ($0 < \zeta < 1$, オーバーシュートに相当する)、臨界制動 ($\zeta = 1$)、定常振動 ($\zeta = 0$, ビブラートに相当する) を利用する¹⁾。齋藤らは、減衰率 ζ と固有周波数 Ω を聴取実験に基づいて手で調整し、それらによって得られる式 (1) のインパルス応答を、階段状に変化する信号に畳み込んで生成される F0 系列を利用して、表情豊かな歌声合成音を実現した¹⁾。

しかしながら、マイクから入力された歌声の F0 系列を自由に編集 (制御) して、歌い方を調整するためには、先に述べた旋律成分と動的変動成分を F0 系列から自動的に特徴抽出する技術が必要である。我々はこれまで、旋律成分および 2 次系の制御パラメータ ζ, Ω が

^{†1} 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
NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation

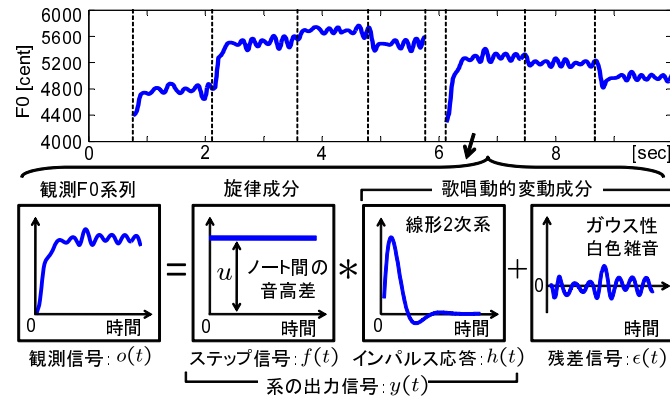


図 1 提案する歌声の F0 生成過程の概略図

いずれも未知の下で、観測される F0 系列だけから、これらを同時に学習する統計的手法の枠組を提唱した¹¹⁾。これは、旋律成分を表す隠れマルコフモデル (HMM) と、差分近似に基づく式 (1) のパラメトリックな表現によって、最尤のモデルパラメータを反復推定する学習アルゴリズムであった。しかし、この学習アルゴリズムでは、モデルパラメータの推定性能が悪かった。そのため、観測 F0 系列と、パラメータによって再合成される F0 系列との誤差が大きくなった。この理由は、図 1 の上部に示すように F0 系列には各ノートの切り替わりやオーバーシュート、ビブラートのような様々な動的変動成分が所々で混在するのに対し、従来法の自由度が高すぎるモデルで、これらの動的変動成分を学習しようとしていたためであると考えている。さらに、F0 系列をフレーム分割して、フレームごとにモデルパラメータを推定する手法¹⁵⁾も提案したが、各動的変動成分を生み出す 2 次系の影響区間が系列上で不明確なため、結局モデルパラメータを適切に推定できなかった。

本報告は、このようなオーバーフィッティングの問題を解消するために、以下の 2 つの方策を導入した歌声の F0 生成過程の新しい確率モデルを提案する。

- (1) 2 次系のインパルス応答 $h(t)$ を、差分近似¹¹⁾ や全極モデル近似¹⁵⁾ に基づいて構成するのではなく、今回はモデルの自由度を効果的に下げる目的で、あらかじめ用意したいくつかの振動基底の疎 (スパース) な線形和によって構成する。
- (2) 図 1 の上部に示すように、ノートが切り替わる時点を始点終点と考え、F0 系列をいくつかのセグメント (ノートに対応すると考える) に分割する。そして、セグメント

ごとに、図 1 の下部に示すような、信号の生成過程を仮定する。すなわち、各セグメントの観測 F0 系列 $o(t)$ は、ノート間の音高差を表す入力ステップ信号 $f(t)$ と 2 次系のインパルス応答 $h(t)$ との畳みこみによって得られる系の出力信号 $y(t)$ と、残差信号 $e(t)$ との和で構成されたと考える。 $h(t)$ と $e(t)$ はどちらも F0 系列の動的変動成分を表す信号であるが、 $h(t)$ はノートの切り替わり方 (オーバーシュートも含む) を表現する信号の大局的な動特性を、 $e(t)$ はビブラートのような音高が安定するときの局所的な動特性を表現すると考える。このようにセグメント内に混在する動的変動成分を分離して信号の生成過程を仮定し、その確率モデルを考える。

このような信号表現方法に基づき、その最適化アルゴリズムを導く。そして、この手法をもとにして、入力した歌声の F0 系列をノート単位で編集できる新しいタイプの歌声合成インタフェース “Vocal Dynamics Controller” を実現する。

2. 複数の振動基底を利用した線形 2 次系のインパルス応答表現

式 (1) の新しい離散時間表現方法を提案する。式 (1) のラプラス逆変換によって得られるインパルス応答は、 ζ の値によって、以下のように場合分けされる。

$$h(t) = \begin{cases} \frac{\Omega e^{-\zeta\Omega t}}{2\sqrt{\zeta^2-1}} (e^{\sqrt{\zeta^2-1}\Omega t} - e^{-\sqrt{\zeta^2-1}\Omega t}) & (\zeta > 1) \\ \frac{\Omega e^{-\zeta\Omega t}}{\sqrt{1-\zeta^2}} (\sin(\sqrt{1-\zeta^2}\Omega t)) & (0 < \zeta < 1) \\ \Omega^2 t e^{-\Omega t} & (\zeta = 1) \\ \Omega \sin(\Omega t) & (\zeta = 0) \end{cases} \quad (2)$$

これらのインパルス応答をサンプリング周期 Δ で離散化すると、系の入出力関係は $\mathbf{y} = \Phi \mathbf{f}$ と記述できる。ここで、 $\mathbf{y} = [y_1, y_2, \dots, y_N]^T$ 、 $\mathbf{f} = [f_1, f_2, \dots, f_N]^T$ は、出力信号 $y(t)$ と入力信号 $f(t)$ をサンプリング周期 Δ で離散化した時系列信号のベクトルを表す (N は信号の長さ)。この Φ が、系のインパルス応答を表し、例えば、 $\zeta = 1$ の場合、 Φ は下三角行列

$$\Phi = \begin{bmatrix} \Omega^2 \Delta e^{-\Omega \Delta} & & & 0 \\ 2\Omega^2 \Delta e^{-2\Omega \Delta} & \Omega^2 \Delta e^{-\Omega \Delta} & & \\ \vdots & \ddots & \ddots & \\ N\Omega^2 \Delta e^{-N\Omega \Delta} & \dots & 2\Omega^2 \Delta e^{-2\Omega \Delta} & \Omega^2 \Delta e^{-\Omega \Delta} \end{bmatrix}$$

となる。しかし、 $h(t)$ は、式 (2) のように複数の場合からなるので、行列 Φ を以下の方

に構成する.

$$\Phi^{-1} \simeq w_1 \Upsilon^{(1)} + w_2 \Upsilon^{(2)} + \dots + w_I \Upsilon^{(I)} \quad (3)$$

ここでは、予め手動で ζ , Ω を決定し、 I 個の振動現象を表すインパルス応答 $\{\Phi^{(1)}, \Phi^{(2)}, \dots, \Phi^{(I)}\}$ を計算する. そして、これらの逆行列 $\Upsilon^{(i)} := (\Phi^{(i)})^{-1}$ (逆フィルタのインパルス応答を表す. 以後、振動基底と呼ぶ) の重み付き和で Φ^{-1} を近似する. ただし、この重みパラメータ $\mathbf{w} := \{w_1, w_2, \dots, w_I\}$ は疎 (スパース) となるように正則化する. これは、 Φ^{-1} が、ある限られた種類の振動基底のみによって表現されることを意味し、モデルの自由度を効果的に下げる目的として、このような操作を行う. 後に説明するが、これは $\mathbf{w}$ の事前確率を課することで実現される. 系のインパルス応答 Φ を $\Phi^{(i)}$ の重み付き和で表現してもよいが、後に説明するパラメータ最適化アルゴリズムの導出の複雑さを解消するため、逆フィルタのインパルス応答 Φ^{-1} を $\Upsilon^{(i)}$ の重み付き和で表現した. それゆえに、系の入出力関係は、以下のように表現される.

$$(w_1 \Upsilon^{(1)} + w_2 \Upsilon^{(2)} + \dots + w_I \Upsilon^{(I)}) \mathbf{y} = \mathbf{f} \quad (4)$$

ここで、便宜上、 $\Psi := w_1 \Upsilon^{(1)} + w_2 \Upsilon^{(2)} + \dots + w_I \Upsilon^{(I)}$ とおく.

3. 2 次系 F0 生成過程の統計的モデリング

式 (4) の 2 次系の入出力関係を統計的にモデル化する.

3.1 入力信号と出力信号の確率モデル

入力信号 $\mathbf{f}$ はノート間の音高差を表すよう、ステップ信号を想定する. そのために、常に同じ値をもつベクトル $\mathbf{u} = [u_1, \dots, u_N]^T = u[1, 1, \dots, 1]^T = u\mathbf{1}$ を用意する. ここで、スカラー値 u は音高差を表すパラメータ、 $\mathbf{1}$ は N 個の 1 の値が並ぶベクトルとする. このベクトル $\mathbf{u}$ を平均とする多次元ガウス分布 $\mathcal{N}(\mathbf{u}, \alpha \mathbf{I}_N)$ から生成される確率変数として、入力信号 $\mathbf{f}$ を表現する. α は分散を表す超パラメータであり、あらかじめ手動で値を設定する. $\mathbf{I}_N$ は $N \times N$ の単位行列を示す.

系の出力信号 $\mathbf{y}$ は、ガウス分布に従う変数集合 $\mathbf{f}$ の線形結合 ($\mathbf{y} = \Psi^{-1} \mathbf{f}$) であるから、 $\mathbf{y}$ 自身もガウス分布に従う. その平均と共分散は、

$$\mathbb{E}[\mathbf{y}] = \Psi^{-1} \mathbb{E}[\mathbf{f}] = \Psi^{-1} \mathbf{u} \quad (5)$$

$$\text{cov}[\mathbf{y}] = \Psi^{-1} \mathbb{E}[\mathbf{f} \mathbf{f}^T] (\Psi^{-1})^T - \Psi^{-1} \mathbb{E}[\mathbf{f}] \mathbb{E}[\mathbf{f}]^T (\Psi^{-1})^T = \alpha \Psi^{-1} (\Psi^{-1})^T \quad (6)$$

となるため、 $\mathbf{y}$ が従う確率分布は、以下のように表現される.

$$\mathbf{y} \sim \mathcal{N}(\Psi^{-1} \mathbf{u}, \alpha \Psi^{-1} (\Psi^{-1})^T) \quad (7)$$

3.2 尤度関数と事前確率

ガウス性白色雑音に従う残差信号 $\boldsymbol{\epsilon} = [\epsilon_1, \epsilon_2, \dots, \epsilon_N]^T \sim \mathcal{N}(\mathbf{0}, \beta \mathbf{I}_N)$ を導入し、観測 F0 系列 $\mathbf{o} = [o_1, o_2, \dots, o_N]^T$ は、系の出力信号 $\mathbf{y}$ に残差信号 $\boldsymbol{\epsilon}$ が加わった信号

$$\mathbf{o} = \mathbf{y} + \boldsymbol{\epsilon} \quad (8)$$

と仮定する. ここで、 β は残差信号の分散を表す超パラメータである. $\mathbf{y}$ と $\boldsymbol{\epsilon}$ は互いに独立であると仮定すると、観測信号 $\mathbf{o}$ が与えられたときのモデルパラメータ $\Theta := \{\mathbf{w}, u, \beta\}$ の尤度関数は、

$$P(\mathbf{o}|\Theta) = \frac{1}{(2\pi)^{N/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{o} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{o} - \boldsymbol{\mu}) \right\} \quad (9)$$

と書ける. ここで、 $\boldsymbol{\mu} = \Psi^{-1} \mathbf{u}$, $\Sigma = \alpha \Psi^{-1} (\Psi^{-1})^T + \beta \mathbf{I}_N$ とする.

Θ の事前確率 $P(\Theta)$ は、各要素の独立性 $P(\Theta) = P(\mathbf{w})P(u)P(\beta)$ を仮定し、 u と β はそれぞれ一様分布に従うものとする. 一方、パラメータ $\mathbf{w}$ の要素には疎 (スパース) な制約をもたせるため、その事前確率は一般化正規分布

$$P(\mathbf{w}) = \prod_{i=1}^I \frac{\lambda p}{2\Gamma(1/p)} \exp^{-\lambda p |w_i|^p} \quad (10)$$

に従うものとする. ただし、 p , λ は一般化正規分布の形状を規定する定数であり、 $0 < p < 2$ のとき $p(\mathbf{w})$ は優ガウスのとなり、スパースネスを測るための尺度となる.

4. EM 法に基づくパラメータ最適化アルゴリズム

観測 F0 系列 $\mathbf{o}$ が与えられたとき、事後確率 $P(\Theta|\mathbf{o}) \propto P(\mathbf{o}|\Theta)P(\Theta)$ を最大化するパラメータ Θ の推定値を決定したい. しかしながら、 Θ の事後 (MAP) 推定値に関する最適解を解析的に求めることは難しい. その理由は、

(1) 観測 F0 系列 $\mathbf{o}$ が出力信号 $\mathbf{y}$ と残差信号 $\boldsymbol{\epsilon}$ の和で構成される.

(2) 尤度関数が $\mathbf{w}$ に関して非線形となる.

である. それぞれの問題を対処するために、

(1) EM 法¹⁶⁾ を適用して、その E-step で、 $\mathbf{o}$ を $\mathbf{y}$ と $\boldsymbol{\epsilon}$ に分離する.

(2) EM 法の M-step に補助関数法¹⁷⁾ を適用して、 Q 関数の補助関数を設計する.

からなる 2 つの方策に基づいて、最適化アルゴリズムを導出する。

4.1 完全データの定義

この MAP 推定問題に EM 法を適用する際の最初のステップは完全データを定義することである。ここでは、 $\mathbf{y}$ と ϵ を完全データ $\mathbf{x}$ と見なして、EM 法を適用する。不完全データと完全データの関係は、

$$\mathbf{o} = \mathbf{H}\mathbf{x}, \quad \left(\mathbf{H} := \begin{bmatrix} \mathbf{I}_N & \mathbf{I}_N \end{bmatrix}, \quad \mathbf{x} := \begin{bmatrix} \mathbf{y} \\ \epsilon \end{bmatrix} \right) \quad (11)$$

となる。 $\mathbf{x}$ と現在のパラメータ Θ' が与えられたときの、対数尤度関数の条件付き期待値を計算し、さらに $\log P(\Theta)$ を加算すると、以下のような Q 関数を得る。

$$Q(\Theta, \Theta') \stackrel{c}{=} \frac{1}{2} \left[\log |\Lambda^{-1}| - \text{tr} \left(\Lambda^{-1} \mathbb{E}[\mathbf{x}\mathbf{x}^T | \mathbf{o}; \Theta'] \right) + 2\mathbf{m}^T \Lambda^{-1} \mathbb{E}[\mathbf{x} | \mathbf{o}; \Theta'] - \mathbf{m}^T \Lambda^{-1} \mathbf{m} \right] + \log P(\Theta) \\ \left(\mathbf{m} := \begin{bmatrix} \Psi^{-1} \mathbf{u} \\ \mathbf{0} \end{bmatrix}, \quad \Lambda^{-1} := \begin{bmatrix} \frac{1}{\alpha} \Psi^T \Psi & \mathbf{0} \\ \mathbf{0} & \frac{1}{\beta} \mathbf{I}_N \end{bmatrix} \right) \quad (12)$$

となる。ここで、 $\text{tr}(\cdot)$ は行列のトレースを表し、 $\mathbb{E}[\mathbf{x} | \mathbf{o}; \Theta']$ と $\mathbb{E}[\mathbf{x}\mathbf{x}^T | \mathbf{o}; \Theta']$ は、条件付きガウス分布の性質より、

$$\mathbb{E}[\mathbf{x} | \mathbf{o}; \Theta'] = \mathbf{m} + \Lambda \mathbf{H}^T (\mathbf{H} \Lambda \mathbf{H}^T)^{-1} (\mathbf{o} - \mathbf{H} \mathbf{m}) \quad (13)$$

$$\mathbb{E}[\mathbf{x}\mathbf{x}^T | \mathbf{o}; \Theta'] = \Lambda - \Lambda \mathbf{H}^T (\mathbf{H} \Lambda \mathbf{H}^T)^{-1} \mathbf{H} \Lambda + \mathbb{E}[\mathbf{x} | \mathbf{o}; \Theta'] \mathbb{E}[\mathbf{x} | \mathbf{o}; \Theta']^T \quad (14)$$

と書ける。EM 法の E-step では、直前に更新されたモデルパラメータを Θ' に代入し、 $\mathbb{E}[\mathbf{x} | \mathbf{o}; \Theta']$ と $\mathbb{E}[\mathbf{x}\mathbf{x}^T | \mathbf{o}; \Theta']$ を計算する。後の計算のため、 $\mathbf{y}$ 、 ϵ に対応するように $\mathbb{E}[\mathbf{x} | \mathbf{o}; \Theta']$ と $\mathbb{E}[\mathbf{x}\mathbf{x}^T | \mathbf{o}; \Theta']$ を

$$\mathbb{E}[\mathbf{x} | \mathbf{o}; \Theta'] = \begin{bmatrix} \bar{\mathbf{x}}_y \\ \bar{\mathbf{x}}_\epsilon \end{bmatrix}, \quad \mathbb{E}[\mathbf{x}\mathbf{x}^T | \mathbf{o}; \Theta'] = \begin{bmatrix} \mathbf{R}_y & * \\ * & \mathbf{R}_\epsilon \end{bmatrix} \quad (15)$$

のように区分表現する。 $\bar{\mathbf{x}}_y$ と $\bar{\mathbf{x}}_\epsilon$ はどちらも長さ N のベクトルであり、 $\mathbf{R}_y$ と $\mathbf{R}_\epsilon$ はどちらも $N \times N$ の正方行列を表す。

4.2 M-step 更新式

式 (12) からパラメータ集合 Θ に関連する項を取り出し、最大化する目的関数を、

$$f(\mathbf{w}, u, \beta) := -\frac{N}{2} \log \alpha \beta + \sum_{n=1}^N \log \left(\sum_{i=1}^I w_i \Upsilon_{n,n}^{(i)} \right) + \frac{1}{\alpha} \mathbf{u}^T \Psi \bar{\mathbf{x}}_y - \frac{1}{2\alpha} \text{tr}(\Psi^T \Psi \mathbf{R}_y) \\ - \frac{1}{2\beta} \text{tr}(\mathbf{R}_\epsilon) - \frac{1}{2\alpha} \mathbf{u}^T \mathbf{u} - \lambda^p \sum_{i=1}^I |w_i|^p \quad (16)$$

と改めて定義する。ここで、 $\Upsilon_{n,n}^{(i)}$ は $\Upsilon^{(i)}$ の n 行 n 列目の対角要素を表す。この最大化問題を解く更新式を補助関数法¹⁷⁾により導く。式 (16) で与えられる目的関数の補助関数を以下の 2 つの不等式（証明略）を用いて導く。

$$\sum_{n=1}^N \log \left(\sum_i w_i \Upsilon_{n,n}^{(i)} \right) \geq \sum_{n=1}^N \sum_{i=1}^I \gamma_{i,n} \log \frac{w_i \Upsilon_{n,n}^{(i)}}{\gamma_{i,n}} \quad (17)$$

$$|w_i|^p \leq p |\bar{w}_i|^{p-1} w_i + |\bar{w}_i|^p - p |\bar{w}_i|^p, \quad (0 < p \leq 1) \quad (18)$$

ここで、補助変数 $\bar{\mathbf{w}} := \{\bar{w}_1, \bar{w}_2, \dots, \bar{w}_I\}$, $\gamma := \{\gamma_{1,1}, \dots, \gamma_{I,N}\}$ を定義する。式 (17), (18) を式 (16) に代入すると、

$$f^+(\mathbf{w}, u, \beta, \bar{\mathbf{w}}, \gamma) := -\frac{N}{2} \log \alpha \beta + \sum_{n=1}^N \sum_{i=1}^I \gamma_{i,n} \log \frac{w_i \Upsilon_{n,n}^{(i)}}{\gamma_{i,n}} + \frac{1}{\alpha} \mathbf{u}^T \Psi \bar{\mathbf{x}}_y - \frac{1}{2\alpha} \text{tr}(\Psi^T \Psi \mathbf{R}_y) \\ - \frac{1}{2\beta} \text{tr}(\mathbf{R}_\epsilon) - \frac{1}{2\alpha} \mathbf{u}^T \mathbf{u} - \lambda^p \sum_{i=1}^I \left(p |\bar{w}_i|^{p-1} w_i + |\bar{w}_i|^p - p |\bar{w}_i|^p \right) \quad (19)$$

を得る。このとき、 $f(\mathbf{w}, u, \beta) \geq f^+(\mathbf{w}, u, \beta, \bar{\mathbf{w}}, \gamma)$ が成り立ち、等号成立は、

$$\bar{w}_i = w_i, \quad \gamma_{i,n} = \frac{\bar{w}_i \Upsilon_{n,n}^{(i)}}{\sum_{i'=1}^I \bar{w}_{i'} \Upsilon_{n,n}^{(i')}}, \quad (i = 1, 2, \dots, I, n = 1, 2, \dots, N) \quad (20)$$

のときであるため、式 (19) は補助関数の定義を満たす。

式 (19) を $w_{i'}$ に関して微分して 0 とおくと、

$$\frac{1}{\alpha} \sum_{i=1}^I \text{tr} \left(\mathbf{R}_y^T \Upsilon^{(i)T} \Upsilon^{(i')} \right) w_i - \frac{1}{\alpha} \mathbf{u}^T \Upsilon^{(i')} \bar{\mathbf{x}}_y + \lambda^p p |\bar{w}_{i'}|^{p-1} - \sum_{n=1}^N \frac{\gamma_{i',n}}{\bar{w}_{i'}} = 0 \\ (i' = 1, 2, \dots, I) \quad (21)$$

を得る。ただし、式 (21) は、 $w_1, w_2, \dots, w_I$ に関して、非線形な連立方程式となるため、

初期化: パラメータ $\Theta = \{w, u, \beta\}$ に初期値を与える.
E-step: 条件付き期待値 $E[x|o; \Theta']$, $E[xx^T|o; \Theta']$, 補助変数 $\bar{w}$, γ の更新.
M-step: 式 (22), (23) から, パラメータ $\Theta = \{w, u, \beta\}$ の更新.
収束判定: 式 (19) の値が収束していなければ, $\Theta' = \Theta$ として **E-step** に戻る.

図 2 EM 法と補助関数法に基づく, 2 次系 F0 生成過程のモデルパラメータ最適化アルゴリズム

Coordinate descent 法¹⁸⁾ を利用して解く. まず, 初期値として $w_1, w_2, \dots, w_I$ をすべて 0 に設定する. そして, 式 (21) を $w_{i'}$ に関して,

$$w_{i'} = \frac{-Y^2 + \sqrt{Y^2 - 4XZ}}{2X} \quad (22)$$

と変形する. ここで,

$$X = \text{tr} \left(R_y^T \mathbf{r}^{(i')}^T \mathbf{r}^{(i')} \right),$$

$$Y = \sum_{i \neq i'} \text{tr} \left(R_y^T \mathbf{r}^{(i)}^T \mathbf{r}^{(i')} \right) w_i - \mathbf{u}^T \mathbf{r}^{(i')} \bar{x}_y + \alpha \lambda^p p |\bar{w}_{i'}|^{p-1}, \quad Z = -\alpha \sum_{n=1}^N \gamma_{i', n}$$

とする. $i' = 1, 2, \dots, I$ に関して, 式 (22) による更新を順番に繰り返し, $w_1, w_2, \dots, w_I$ の値がそれぞれ変化しなくなるまで更新を続ける.

一方, $f^+(w, u, \beta, \bar{w}, \gamma)$ を u, β それぞれに関して微分して 0 とおくと,

$$u = \frac{1}{N} \mathbf{1}^T \Psi \bar{x}_y, \quad \beta = \frac{1}{N} \text{tr}(\mathbf{R}_\epsilon) \quad (23)$$

が得られ, u と β を更新する. パラメータ最適化アルゴリズムの流れを図 2 にまとめる.

5. Vocal Dynamics Controller の実装

以上のアルゴリズムにより各セグメントの w, u, β がひとたび求まれば, F0 系列 o を式 (8) に従って w, u, β の操作を介して加工することができる. 提案法を核として, 入力した歌声の F0 系列をノート単位で編集できる歌声合成インタフェース “Vocal Dynamics Controller” を実装した. 図 3 にグラフィカルユーザインタフェース (GUI) の表示画面を示す. 操作方法および機能は以下のとおりである.

A: 歌声音響信号の読み込みと F0 推定 音響信号を読み込み, F0 を推定する. F0 は, YIN¹⁹⁾ を利用して, 5ms ごとに推定される ($\Delta = 5\text{ms}$). なお, Hz で表される周

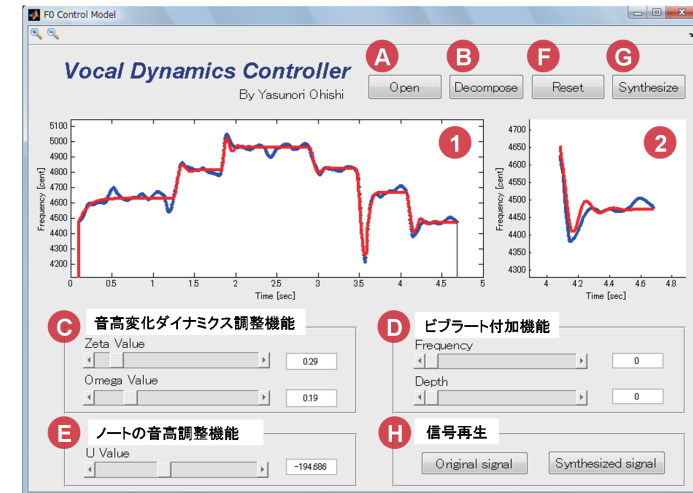


図 3 Vocal Dynamics Controller の表示画面, ウィンドウ①, ②と④～①の操作方法に関する説明は 5 章参照.

波数 o_{Hz} を, cent で表される対数スケールの周波数 o_{cent} に変換する.

$$o_{cent} = 1200 \log_2 \frac{o_{Hz}}{440 \times 2^{\frac{3}{12} - 5}} \quad (24)$$

F0 推定結果はウィンドウ①に青線で出力される.

B: 提案モデルのパラメータ最適化アルゴリズムの実行 4 章で導入したパラメータ最適化アルゴリズムを実行する. 処理の流れは以下のとおりである.

- (1) 42 個の状態からなるエルゴディック HMM を利用した, 観測 F0 系列の Viterbi 探索によって, F0 系列を自動的にセグメント分割する. HMM の各状態は半音の間隔で配置される 12 平均律の構成音を表し, 単一ガウス分布による出力確率分布を持つ. cent で表される対数スケールの周波数は半音が 100cent に相当するので, 各状態の出力確率分布の平均値は, 0 (無音), および 3000cent から 7000cent まで 100cent 刻みで変化させた値のいずれかをとるものとする. 出力確率分布の分散値は, すべての状態で同じ 64000 とした. また, 初期状態確率は, どの状態も $1/42$ とした. 自己遷移確率は 0.999999, 他の状態への遷移確率はすべて同じ $0.000001/41$ とした. これらの HMM パラメータは実験的に決定した. 分割結果はウィンドウ①に赤線で出力される (図 4 の上部).

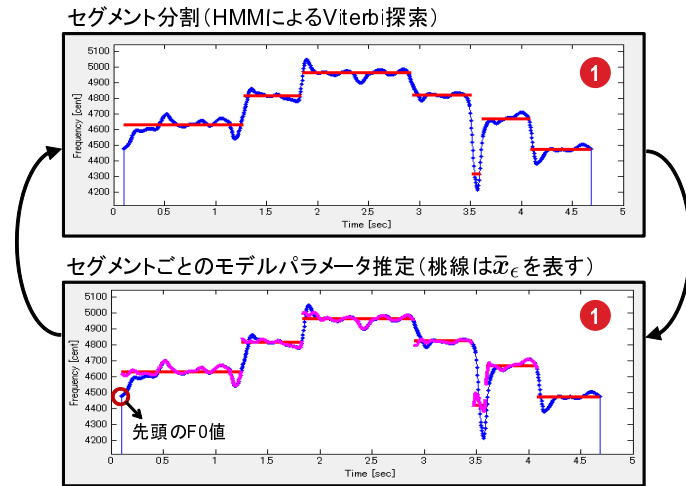


図4 処理Bでは、セグメント分割と各セグメントにおけるモデルパラメータ推定を繰り返す。

- (2) 分割されたセグメントごとに、パラメータ最適化アルゴリズムを適用するために、パラメータの初期値を設定する。まず、 $\{\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \dots, \mathbf{Y}^{(I)}\}$ を作成するために、 ζ は0から2までの間を0.02刻みで、 Ω は0.05から0.3までの間を0.005刻みで変化させた。その結果、 $I = 3100$ となる。 $\mathbf{w} = \{w_1, w_2, \dots, w_I\}$ の初期値はすべて $1/I$ に設定する。 u は、各セグメントの観測 F0 系列 $\mathbf{o}$ の要素の中央値を初期値とする。 β は、 $\beta = 100$ を初期値とする。これらのパラメータの初期値はすべて実験的に決定した。また、 $\alpha = 2$, $\lambda = 10000$, $p = 0.8$ に固定した。
- (3) 最初のセグメントから順番にパラメータ最適化アルゴリズム (図2) を実行する。前処理として、以下のどちらかの観測 F0 系列の正規化を行う。
 - 一つ前のセグメントが無音区間であれば、対象とするセグメントの F0 系列から、その先頭の F0 値 (図4の下部参照) を減算する。
 - 一つ前のセグメントが無音区間でなければ、無音区間直後のセグメントの先頭の F0 値および、それ以降から一つ前までのセグメントにおいて推定されたパラメータ u の総和を、対象とするセグメントの F0 系列から減算する。
- (4) すべてのセグメントに対して処理 (2), (3) を実行したら、各セグメントで最終的に求まる $\bar{x}_e$ をすべてつなぎ合わせて、全セグメントの残差信号の期待値信号

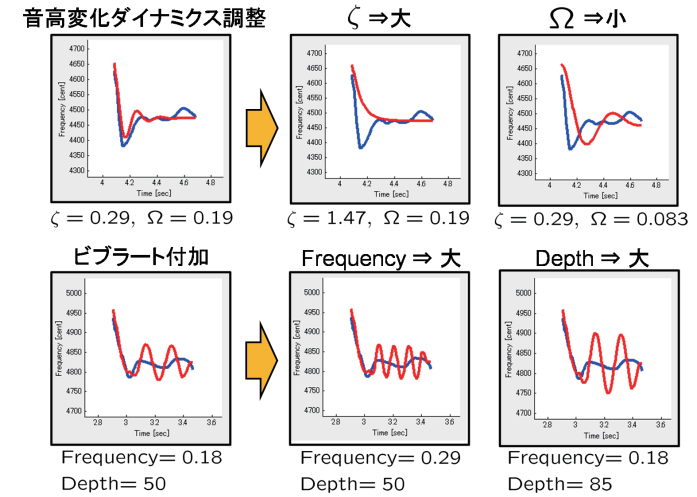


図5 処理C, Dにおける動的変動成分の制御

ϵ_{all} を作成する (図4の下部の桃線)。この信号は、ノートが切り替わる時の大局的な動特性を観測 F0 系列から取り除いた信号と言える。この信号 ϵ_{all} を利用して、再度、処理 (1) の HMM による Viterbi 探索を行い、処理 (2), (3) へと進む。Viterbi 探索によって分割されるセグメントの位置が変化しなくなるまで、この一連の処理を繰り返す。

- (5) 処理 (1)~(4) が終わったら、セグメントごとに推定されたパラメータから式 (9) の $\mu = \Psi^{-1}u$ を計算し、これを生成 F0 系列としてウィンドウ①に赤線で示す。

- C: 音高変化ダイナミクス調整機能** ウィンドウ①で、音高変化の動特性を調整したいセグメントをクリックすると、ウィンドウ②にそのセグメントの観測 F0 系列 (青線) と生成 F0 系列 (赤線) が表示される。スライダで操作された ζ と Ω の値から計算される $h(t)$ に基づいて Φ を構成し、 Φu を計算して、そのセグメントの生成 F0 系列としてウィンドウ①, ②に赤線で再描画する。図5の上部に示すように、 ζ が小さくするとオーバーシュートが起り、大きくすると指数的に目標音高に減衰する。一方、 Ω を小さくするとノートの切り替わり時間は長く、大きくすると切り替わり時間は短くなる。
- D: ビブラート付加機能** ウィンドウ①で、ビブラートを付加したいセグメントをクリックすると、ウィンドウ②にそのセグメントの観測 F0 系列 (青線) と生成 F0 系列 (赤線)

が表示される。Depth で設定された値 (単位は cent) に基づくステップ信号と、式 (2) の $\zeta = 0$ の場合の Ω に Frequency の値を代入したインパルス応答との畳み込みによって得られる信号を、選択されたセグメントの生成 F0 系列に足し合わせてウィンドウ①、②に赤線で再描画する。図 5 の下部に示すように、Frequency を大きくすればビブラートの周期が短くなり、Depth を大きくすればビブラートの振幅は大きくなる。

E: ノートの音高調整機能 ウィンドウ①で、ノートの音高を調整したいセグメントをクリックすると、ウィンドウ②にそのセグメントの観測 F0 系列 (青線) と生成 F0 系列 (赤線) が表示される。操作されたパラメータ u に基づいて構成される u と、そのセグメントにおける ζ と Ω から Φ を構成し、 Φu を計算して、そのセグメントの生成 F0 系列としてウィンドウ①、②に赤線で再描画する。

F: リセット機能 C, D, E で操作した内容をすべてリセットして、B で推定されたパラメータに基づいて、生成 F0 系列をウィンドウ①に赤線で再描画する。

G: 歌声音響信号の合成 B~E の操作によって生成された F0 系列に基づいて、入力された歌声音響信号をフレームごとに伸縮し、Griffin-Lim の反復 STFT 法²⁰⁾ に基づいて音響信号を合成する。処理の流れは以下のとおりである。

- (1) 生成 F0 系列と観測 F0 系列の各時刻における F0 値の比率を計算し、これを伸縮率とする。
 - (2) 入力歌声音響信号の短時間フーリエ変換 (STFT) により、時間周波数解析を行って、スペクトログラム $\mathbf{Y} = (Y_{f,t})_{F \times T}$ を得る。STFT では、フレーム長 20ms の Hanning 窓をシフト幅 5ms により分析した。
 - (3) フレームごとに線形予測分析 (LPC)²¹⁾ を行い、パワースペクトルをスペクトル包絡と駆動音源スペクトルに分離する。
 - (4) (1) で求めた伸縮率に基づいて、フレームごとに駆動音源スペクトルを線形伸縮し、スペクトル包絡と掛け合わせる。これにより音韻情報を表すスペクトル包絡を保存したピッチ変換を行うことができる。
 - (5) (4) の処理によって変換されたスペクトログラム $\{X_{\omega,t}\} \in \mathbb{R}^{\geq 0}$ の位相をランダムに設定し、位相が付加された複素数値の時間周波数成分を $V_{\omega,t} \in \mathbb{C}$ とする。
 - (6) $\{V_{f,t}\}_{f \in \{1, \dots, F\}, t \in \{1, \dots, T\}}$ に対し、逆 STFT を行い、実信号 $v[m]_{m=1}^M$ を得る。
 - (7) 実信号 $v[m]_{m=1}^M$ に対し、STFT を行い、 $\{V_{f,t}\}_{f \in \{1, \dots, F\}, t \in \{1, \dots, T\}}$ を更新する。
 - (8) すべての f, t について、 $V_{f,t} \leftarrow X_{f,t} \frac{V_{f,t}}{|V_{f,t}|}$ により $V_{f,t}$ を更新し、(6) に戻る。
- なお、上記の (6)~(8) の Griffin-Lim の反復 STFT 法以外に Le Roux の手法²²⁾ を用

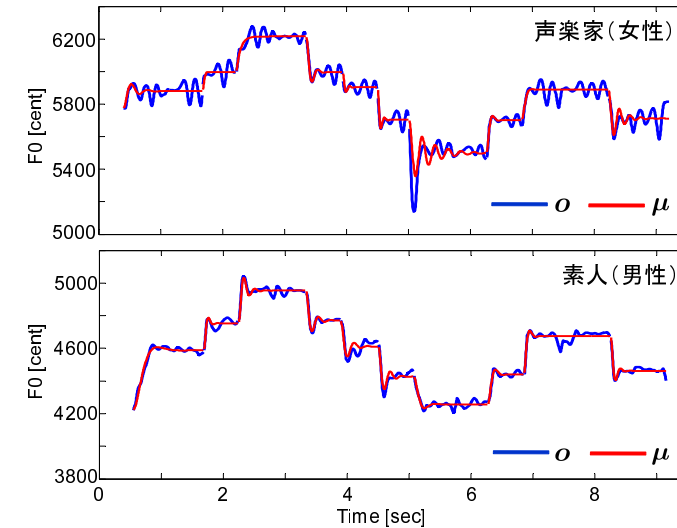


図 6 観測 F0 系列 o と推定されたモデルパラメータによって生成される F0 系列 μ との比較

いることもできる。

H: 信号再生機能 A の処理によって読み込まれた入力歌声音響信号、および F の処理によって合成された歌声音響信号を再生する。

6. モデルパラメータによって生成される F0 系列に関する考察

最適化アルゴリズムによって推定されるモデルパラメータについて定性的に評価する。図 6 は、パラメータ w, u から計算される式 (9) の μ (生成 F0 系列) と観測 F0 系列 o を比較する。音楽訓練を受けた女性の声楽家と男性の素人歌唱者がそれぞれ、「喜びの歌 (Beethoven の交響曲第 9 番第 4 楽章の歌の部分)」を日本語歌詞かつアカペラで歌唱した歌声信号を入力とした。声楽家と素人歌唱者のどちらも、パラメータ w, u によって生成される F0 系列は、オーバーシュートやボルタメントのようなノートが切り替わるときの大局的な動特性を表現できていることがわかる。従来法¹¹⁾ と比べてどれだけ精度よく推定が可能であるか、多くの歌声データを利用した定量的な評価が必要である。素人歌唱者に比べて、声楽家の観測 F0 系列には所々にビブラートが観測される。提案モデルでは、このような変動成分はすべて、ガウス性白色雑音に従うとし、分散パラメータ β によって表現される。この残差信号 ϵ

をより精緻にモデル化して、ビブラートのような周期的な変動成分をも特徴抽出することは今後の課題とする。例えば、近年機械学習の分野で注目を集める Multiple Kernel Learning を利用したガウス過程^{23)–25)}に基づく信号表現との類推から、 ϵ が従う確率分布の共分散行列を周期カーネルによって表現する方法が考えられる。さらに、パラメータ w , u , β の歌唱者ごとの違いから歌唱スタイルについて分析することも今後の課題である。

7. おわりに

本報告では、歌声の F0 系列の動特性をノート単位で編集し、歌い方を多様に変形できる歌声合成インタフェースの実現を目指し、その動特性のモデリングとモデルパラメータ推定に関する新しい解法を検討した。今後は、インタフェースの機能の充実化を図るとともに、提案モデルを多変量化して MFCC 信号に適用し、声質の動特性の制御についても考えたい。

謝辞 本研究の歌声合成インタフェースにおける音響信号の位相復元の技術に関して、有益なご助言を頂いた Jonathan Le Roux 氏 (NTT CS 研) に感謝致します。

参 考 文 献

- 1) Saitou, T. et al.: Speech-To-Singing Synthesis: Converting Speaking Voices to Singing Voices by Controlling Acoustic Features Unique to Singing Voices, *Proc. WASSPA 2007*, pp.215–218 (2007).
- 2) Saitou, T. et al.: Acoustic and Perceptual Effects of Vocal training in Amateur Male Singing, *Proc. EUROSPEECH 2009*, pp.832–835 (2009).
- 3) Nakano, T. et al.: An Automatic Singing Skill Evaluation Method for Unknown Melodies Using Pitch Interval Accuracy and Vibrato Features, *Proc. ICSLP 2006*, pp.1706–1709 (2006).
- 4) Kako, T. et al.: Automatic Identification for Singing Style Based on Sung Melodic Contour Characterized in Phase Plane, *Proc. ISMIR 2009*, pp.393–397 (2009).
- 5) Proutskova, P. and Casey, M.: You Call That Singing? Ensemble Classification for Multi-Cultural Collections of Music Recordings, *Proc. ISMIR 2009*, pp.759–764 (2009).
- 6) Sundberg, J.: The KTH synthesis of singing, *Advances in Cognitive Psychology. Special issue on Music Performance*, Vol.2, No.2-3, pp.131–143 (2006).
- 7) Bonada, J. and Lascos, A.: Sample-based singing voice synthesizer by spectral concatenation, *Proc. SMAC 2003* (2003).
- 8) Nakano, T. et al.: VocaListener: A Singing-to-Singing Synthesis System Based on Iterative Parameter Estimation, *Proc. SMC 2009*, pp.343–348 (2009).
- 9) Fukayama, S. et al.: Orpheus: Automatic Composition System Considering Prosody of Japanese Lyrics, *Proc. ICEC 2009*, pp.309–310 (2009).
- 10) Ohishi, Y. et al.: A Stochastic Representation of the Dynamics of Sung Melody, *Proc. ISMIR 2007* (2007).
- 11) Ohishi, Y. et al.: Parameter Estimation Method of F0 Control Model for Singing Voices, *Proc. ICSLP 2008*, pp.139–142 (2008).
- 12) 柏野邦夫ほか：パート譜を用いたボーカル音分離システム，音講論集，2-9-1，pp.625–626 (1998).
- 13) Minematsu, N. et al.: Prosodic Modeling of Nagauta Singing and Its Evaluation, *Proc. SpeechProsody 2004*, pp.487–490 (2004).
- 14) Fujisaki, H.: A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour, *Vocal Physiology: Voice Production, Mechanisms and Functions*, (O. Fujimura, ed.), Raven Press, pp. 347–355 (1988).
- 15) 大石康智ほか：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，情処研報音楽情報科学，Vol.2008, No.76, pp.89–96 (2008).
- 16) Feder, M. and Weinstein, E.: Parameter estimation of superimposed signals using the EM algorithm, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol.36, No.4, pp.477–489 (1988).
- 17) Kameoka, H. et al.: Complex NMF: A New Sparse Representation for Acoustic Signals, *Proc. ICASSP 2009*, pp.3437–3440 (2009).
- 18) Meng, X.L. and Rubin, D.B.: Maximum Likelihood Estimation via the ECM Algorithm: A general framework, *Biometrika*, Vol.80, pp.267–278 (1993).
- 19) de Cheveigné, A. and Kawahara, H.: YIN, a fundamental frequency estimator for speech and music, *JASA*, Vol.111, No.4, pp.1917–1930 (2002).
- 20) Griffin, D.W. and Lim, J.S.: Signal estimation from modified short-time Fourier transform, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol.32, No.2, pp.236–243 (1984).
- 21) Itakura, F. and Saito, S.: Digital filtering techniques for speech analysis and synthesis, *Proc. ICA 1971*, Vol.25-C-1, pp.261–264 (1971).
- 22) Le Roux, J. et al.: Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction, *Proc. SAPA 2008* (2008).
- 23) Rasmussen, C.E. and Williams, C. K.I.: *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, Mass, USA (2006).
- 24) Bach, F. et al.: Multiple kernel learning, conic duality, and the smo algorithm, *Proc. ICML 2004*, pp.6–13.
- 25) 亀岡弘和ほか：マルチカーネル線形予測モデルによる音声分析，音講論集，2-Q-24，pp.499–502 (2010).