Statistical Modeling of F0 Dynamics in Singing Voices Based on Gaussian Processes with Multiple Oscillation Bases

Yasunori Ohishi, Hirokazu Kameoka, Daichi Mochihashi, Hidehisa Nagano, Kunio Kashino

NTT Communication Science Laboratories, NTT Corporation

{ohishi,kameoka,nagano,kunio}@cs.brl.ntt.co.jp, daichi@cslab.kecl.ntt.co.jp

Abstract

We present a novel statistical model for dynamics of various singing behaviors, such as *vibrato* and *overshoot*, in a fundamental frequency (F0) contour. These dynamics are the important cues for perceiving individuality of a singer, and can be a useful measure for various applications, such as singing skill evaluation and singing voice synthesis. While most previous studies have modeled the dynamics using a second-order linear system, the automatic and accurate estimation of model parameters has yet to be accomplished. In this paper, we first develop a complete stochastic representation of the second-order system with Gaussian processes from parametric discretization, and propose a complete, efficient scheme for parameter estimation using the Expectation-Maximization (EM) algorithm. Experimental results show that the proposed method can decompose an F0 contour into a musical component and a dynamics component. Finally, we discuss estimating singing styles from the model parameters for each singer.

Index Terms: Singing voices, Fundamental frequency (F0), Gaussian Processes, EM algorithm, Singing voice synthesis

1. Introduction

The goal of this work is to characterize both musical-note information and the dynamics of various singing behaviors, such as *vibrato* and *overshoot*, in a sung melodic contour, i.e., an F0 sequence. The importance of dynamics for perceiving individuality of a singer is reported in [1, 2] based on psycho-acoustic experiments, and this means that the dynamics can be a useful measure for identifying of singing styles [3], singing skill evaluation [4], and the synthesis of more natural and various singing voices [5, 6, 7, 8]. On the other hand, musical-note information is useful for various applications such as Query-By-Humming and the automatic clustering of songs [9, 10].

Most previous studies have used a deterministic, second-order linear system to represent the F0 dynamics of singing voices [1, 11]. The transfer function of the system is described as

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

where ζ and Ω denote the damping ratio and the damped natural frequency, respectively. In the F0 control model for singing voices, the dynamics are represented not only by *Critical damping* ($\zeta = 1$) used in the Fujisaki model [12] that controls the F0 contours of speaking voices, but also by *Over-damping* ($\zeta > 1$), *Under-damping* ($0<\zeta<1$), and *Steady oscillation* ($\zeta=0$). This is because the F0 fluctuations in singing voices are larger and more rapid than those of speaking voices. In [1], using the F0 contour generated by the convolution of a step-wise signal which corresponds to the musical-note sequence and the impulse response of Eq. (1), natural and expressive singing voices were synthesized. However, ζ and Ω were controlled manually based on psycho-acoustic experiments.

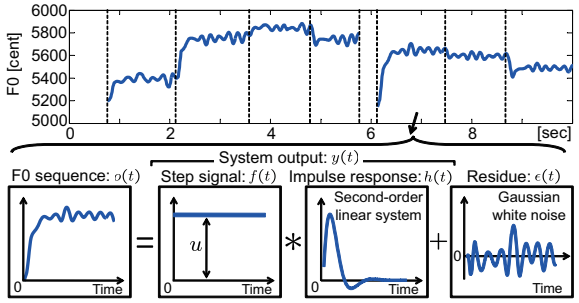

Figure 1: *Decomposition of F0 contour into three components based on second-order linear system*

On the other hand, we have previously proposed a stochastic framework that learns these parameters and the step-wise signal simultaneously from an observed F0 contour [13]. As seen in Fig. 1, we divided the F0 contour into segments and decomposed the F0 contour for each segment into three components, i.e., the step signal $f(t)$, the impulse response $h(t)$ of the second-order system, and the residue $\epsilon(t)$. $f(t)$ indicates the relative pitch that the singer attempts to sing. $h(t)$ represents the temporal attack of the musical note and *overshoot*. System output $y(t)$ denotes the convolution of $f(t)$ and $h(t)$. $\epsilon(t)$ is the residual component including *vibrato*. Therefore, we approximated the second-order differential equation given by the inverse Laplace transform of Eq. (1) by the difference equation:

$$(a\boldsymbol{A} + b\boldsymbol{B} + \boldsymbol{C})\boldsymbol{y} = \boldsymbol{f} \quad (a := 1/\Omega^2,\ b := 2\zeta/\Omega), \qquad (2)$$

where $\boldsymbol{f} = [f_1, f_2, \ldots, f_N]^{\mathrm{T}}$ and $\boldsymbol{y} = [y_1, y_2, \ldots, y_N]^{\mathrm{T}}$ denote discrete signals sampled with a sampling period Δ. N is the signal length, and $a\boldsymbol{A} + b\boldsymbol{B} + \boldsymbol{C}$ represents the inverse impulse response. We define $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ as follows:

$$\boldsymbol{A} := \begin{bmatrix} \frac{1}{\Delta} & & & \\ \frac{-2}{\Delta} & \frac{1}{\Delta} & & \boldsymbol{0} \\ \frac{1}{\Delta} & \frac{-2}{\Delta} & \frac{1}{\Delta} & \\ & \ddots & \ddots & \ddots \\ \boldsymbol{0} & & \frac{1}{\Delta} & \frac{-2}{\Delta} & \frac{1}{\Delta} \end{bmatrix}, \quad \boldsymbol{B} := \begin{bmatrix} \frac{1}{2\Delta} & & & \\ 0 & \frac{1}{2\Delta} & & \boldsymbol{0} \\ -\frac{1}{2\Delta} & 0 & \frac{1}{2\Delta} & \\ & \ddots & \ddots & \ddots \\ \boldsymbol{0} & & -\frac{1}{2\Delta} & 0 & \frac{1}{2\Delta} \end{bmatrix},$$

$$\boldsymbol{C} := \begin{bmatrix} 0 & & & \\ 1 & 0 & & \boldsymbol{0} \\ 0 & 1 & 0 & \\ & \ddots & \ddots & \ddots \\ \boldsymbol{0} & & 0 & 1 & 0 \end{bmatrix}. \qquad (3)$$

Then, we estimated the model parameters iteratively based on the maximum-likelihood approach, so that the residual error between both sides of Eq. (2) were minimized. However, the estimation accuracy was poor in parts of the transition between musical notes. Since various dynamics, such as the attack and *vibrato*, are mixed in each segment, we consider that over-fitting occurs with the conventional method which estimates the model parameters of the second-order linear system directly.

In this paper, we propose a complete stochastic representation of the second-order linear system with Gaussian processes

(GPs) [14]. Therefore, we parametrically approximate the impulse response of the system using multiple oscillation bases and explicitly introduce the residual component $\boldsymbol{\epsilon}$ as a random variable. Then, we develop a efficient scheme for parameter estimation using the EM algorithm. Our experimental results show that the proposed method can appropriately decompose the F0 contour into the target musical note and the dynamics. Furthermore, we also discuss estimating singing styles from the model parameters for each singer.

2. Impulse response approximation with multiple oscillation bases

We introduce a discrete-time representation of Eq. (1) to replace the difference approximation Eq. (2). The impulse response of Eq. (1) has four cases:

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

Discretizing these impulse responses at Δ, the input-output relation is described by $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{f}$. For example, when $\zeta = 1$, $\boldsymbol{\Phi}$ is described as a lower triangular matrix

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

However, since $h(t)$ has multiple cases, we compose matrix $\boldsymbol{\Phi}$ as follows:

$$\boldsymbol{\Phi}^{-1} \simeq w_1\boldsymbol{\Upsilon}^{(1)} + w_2\boldsymbol{\Upsilon}^{(2)} + \ldots + w_I\boldsymbol{\Upsilon}^{(I)}, \qquad (4)$$

where we determine the values of ζ and Ω manually and preliminarily calculate I oscillation bases $\{\boldsymbol{\Phi}^{(1)}, \boldsymbol{\Phi}^{(2)}, \ldots, \boldsymbol{\Phi}^{(I)}\}$. For convenience, we define the impulse response $\boldsymbol{\Upsilon}^{(i)} := (\boldsymbol{\Phi}^{(i)})^{-1}$ of the inverse filter and approximate $\boldsymbol{\Phi}^{-1}$ by combining these matrices linearly, where we assume that weight parameters $\boldsymbol{w} := \{w_1, w_2, \ldots, w_I\}$ are sparse. This means that $\boldsymbol{\Phi}^{-1}$ are expressed by only a few oscillation bases. Therefore, the input-output relation of the system is obtained by

$$(w_1\boldsymbol{\Upsilon}^{(1)} + w_2\boldsymbol{\Upsilon}^{(2)} + \ldots + w_I\boldsymbol{\Upsilon}^{(I)})\boldsymbol{y} = \boldsymbol{f}. \qquad (5)$$

Note that the transfer function Eq. (1) is converted into the form $\boldsymbol{\Psi}\boldsymbol{y} = \boldsymbol{f}$ as seen from Eqs. (2) and (5).

3. Statistical modeling of second-order linear system based on Gaussian Processes

We statistically model the input-output relations of the system represented by Eqs. (2) and (5) based on GPs.

3.1. Modeling of input step signal

We assume input $\boldsymbol{f}$ is a step signal (see Fig. 1). We statistically model $\boldsymbol{f}$ as a random variable generated from the multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{u},\ \alpha\boldsymbol{I}_N)$, where we define $\boldsymbol{u} = [u_1, \ldots, u_N]^{\mathrm{T}} = u[1, 1, \ldots, 1]^{\mathrm{T}} = u\boldsymbol{1}$. Scalars u and α are the relative pitch and a hyperparameter representing the variance of the distribution, respectively. α is set experimentally. $\boldsymbol{I}_N$ denotes the $N \times N$ identity matrix. Since $\boldsymbol{y}$ is a linear combination of Gaussian distributed variables given by the elements of $\boldsymbol{f}$, $\boldsymbol{y}$ is itself Gaussian $\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{\Psi}^{-1}\boldsymbol{u},\ \alpha\boldsymbol{\Psi}^{-1}(\boldsymbol{\Psi}^{-1})^{\mathrm{T}}\right)$.

It should be noted that this model provides us with a particular example of GPs. A key point about GPs is that the joint distribution over N elements of $\boldsymbol{y}$ is specified completely by the second-order statistics, namely the mean and covariance. In vanilla GPs, since the mean is mainly zero, the covariance matrix, which is usually denoted by the Gram matrix $\boldsymbol{K}$, is represented by combining multiple kernel matrices linearly, referred to as the Multiple Kernel Learning (MKL) [15]. This technique can exploit the temporal correlations between observations instead of treating the observations as i.i.d. The GPs with the MKL have received increased attention in the machine-learning community. However, in our case, since the mean includes the variable $\boldsymbol{\Psi}^{-1}$, we represent $\boldsymbol{\Psi}$ by the linear combination of multiple oscillation bases. In this respect, we refer to this representation as GPs with the multiple basis learning.

3.2. Likelihood function and prior probability

We introduce the residual component $\boldsymbol{\epsilon}$ (Gaussian white noise, $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \beta\boldsymbol{I}_N)$) and assume that the observed F0 sequence $\boldsymbol{o} = [o_1, o_2, \ldots, o_N]^{\mathrm{T}}$ is given by output $\boldsymbol{y}$ with $\boldsymbol{\epsilon}$ so that $\boldsymbol{o} = \boldsymbol{y} + \boldsymbol{\epsilon}$, where β is a hyperparameter representing the variance of the noise. Since we suppose that $\boldsymbol{y}$ and $\boldsymbol{\epsilon}$ are mutually independent, from the definition of GPs, the likelihood function of $\Theta := \{\boldsymbol{\Psi}, u, \beta\}$ is given as follows:

$$P(\boldsymbol{o}|\Theta) = \frac{1}{(2\pi)^{N/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\Big\{ -\frac{1}{2}(\boldsymbol{o}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{o}-\boldsymbol{\mu})\Big\},$$
$$\Big(\boldsymbol{\mu} = \boldsymbol{\Psi}^{-1}\boldsymbol{u}, \quad \boldsymbol{\Sigma} = \alpha\boldsymbol{\Psi}^{-1}(\boldsymbol{\Psi}^{-1})^{\mathrm{T}} + \beta\boldsymbol{I}_N\Big). \qquad (6)$$

We assume that the prior distributions for $\boldsymbol{\Psi}$, u, and β are independent, which yields $P(\Theta) = P(\boldsymbol{\Psi})P(u)P(\beta)$, and that $P(u)$ and $P(\beta)$ are uniform distributions. For $P(\boldsymbol{\Psi})$, we also assume the independence of parameters, which yields

$$P(\boldsymbol{\Psi}) = \begin{cases} P(a)P(b), & (\text{Difference approximation}) \\ P(\boldsymbol{w}), & (\text{Impulse response approximation}) \end{cases},$$

and that $P(a)$ and $P(b)$ are uniform distributions. $P(\boldsymbol{w})$ corresponds to the sparsity cost described in Section 2, for which a natural choice is a generalized Gaussian prior

$$P(\boldsymbol{w}) = \prod_{i=1}^{I} \frac{\lambda p}{2\Gamma(1/p)} \exp^{-\lambda^p|w_i|^p} \qquad (7)$$

where p and λ are the parameters that determine the shape of the distribution. When $0 < p < 2$, $P(\boldsymbol{w})$ becomes super-Gaussian and promotes sparsity if the norm of $\boldsymbol{w}$ is bounded.

4. Parameter estimation algorithm based on Expectation-Maximization algorithm

Given observed F0 sequence $\boldsymbol{o}$, we want to determine the estimate of Θ that maximizes the posterior density $P(\Theta|\boldsymbol{o}) \propto P(\boldsymbol{o}|\Theta)P(\Theta)$. However, it is difficult to obtain an optimum solution for the maximum a posterior (MAP) estimate of Θ analytically. This is because (1) $\boldsymbol{o}$ is given by $\boldsymbol{y}$ with $\boldsymbol{\epsilon}$, (2) the objective function is nonlinear with respect to $\boldsymbol{w}$. To cope with these problems, (1) we partition $\boldsymbol{o}$ into $\boldsymbol{y}$ and $\boldsymbol{\epsilon}$ using the EM algorithm, (2) Applying the auxiliary function method [16] to the M-step, we design an auxiliary function of the Q-function.

4.1. Definition of complete data

When applying the EM algorithm to the current MAP estimation problem, the first step is to define the "complete data". We

denote the complete data as $\boldsymbol{x} := [\boldsymbol{y}^{\mathrm{T}}, \boldsymbol{\epsilon}^{\mathrm{T}}]^{\mathrm{T}}$. Taking the conditional expectation of the log-likelihood given $\boldsymbol{x}$ and $\Theta = \Theta'$ and then adding $\log P(\Theta)$, we obtain the Q-function as follows:

$$Q(\Theta, \Theta') \stackrel{\mathrm{c}}{=} \frac{1}{2}\Big[\log|\boldsymbol{\Lambda}^{-1}| - \mathrm{tr}\left(\boldsymbol{\Lambda}^{-1}\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}|\boldsymbol{o};\Theta']\right) + 2\boldsymbol{m}^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\mathbb{E}[\boldsymbol{x}|\boldsymbol{o};\Theta'] - \boldsymbol{m}^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\boldsymbol{m}\Big] + \log P(\Theta), \quad (8)$$

$$\left(\boldsymbol{x} := \begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{\epsilon}\end{bmatrix},\ \boldsymbol{m} := \begin{bmatrix}\boldsymbol{\Psi}^{-1}\boldsymbol{u}\\ \boldsymbol{0}\end{bmatrix},\ \boldsymbol{\Lambda}^{-1} := \begin{bmatrix}\frac{1}{\alpha}\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi} & \boldsymbol{0}\\ \boldsymbol{0} & \frac{1}{\beta}\boldsymbol{I}_N\end{bmatrix}\right),$$

where the many-to-one relationship between the complete data $\boldsymbol{x}$ and the incomplete data $\boldsymbol{o}$ is described as $\boldsymbol{o} = \boldsymbol{H}\boldsymbol{x}$ with $\boldsymbol{H} := [\boldsymbol{I}_N\ \boldsymbol{I}_N]$. $\mathbb{E}[\boldsymbol{x}|\boldsymbol{o};\Theta']$ and $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}|\boldsymbol{o};\Theta']$ are described by

$$\mathbb{E}[\boldsymbol{x}|\boldsymbol{o};\Theta'] = \boldsymbol{m} + \boldsymbol{\Lambda}\boldsymbol{H}^{\mathrm{T}}(\boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^{\mathrm{T}})^{-1}(\boldsymbol{o} - \boldsymbol{H}\boldsymbol{m}), \quad (9)$$

$$\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}|\boldsymbol{o};\Theta'] = \boldsymbol{\Lambda} - \boldsymbol{\Lambda}\boldsymbol{H}^{\mathrm{T}}(\boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^{\mathrm{T}})^{-1}\boldsymbol{H}\boldsymbol{\Lambda} + \mathbb{E}[\boldsymbol{x}|\boldsymbol{o};\Theta']\mathbb{E}[\boldsymbol{x}|\boldsymbol{o};\Theta']^{\mathrm{T}}. \quad (10)$$

For convenience, we segment $\mathbb{E}[\boldsymbol{x}|\boldsymbol{o};\Theta']$ and $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}|\boldsymbol{o};\Theta']$ into small sections as follows:

$$\mathbb{E}[\boldsymbol{x}|\boldsymbol{o};\Theta'] = \begin{bmatrix}\bar{\boldsymbol{x}}_y\\ \bar{\boldsymbol{x}}_\epsilon\end{bmatrix}, \quad \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}|\boldsymbol{o};\Theta'] = \begin{bmatrix}\boldsymbol{R}_y & *\\ * & \boldsymbol{R}_\epsilon\end{bmatrix}. \quad (11)$$

4.2. M-step update formulae

We can derive the M-step update formulae for all the model parameters by employing the definitions of the complete data and the prior. In this subsection we derive the M-step update formulae for the impulse response approximation method described in Section 2 owing to space limitations, where $\boldsymbol{\Psi}$ corresponds to $\boldsymbol{\Phi}^{-1}$ of Eq. (4). Collecting terms that depend on Θ from Eq. (8), we obtain the objective function as follows:

$$f(\boldsymbol{w}, u, \beta) := -\frac{N}{2}\log\alpha\beta + \sum_{n=1}^{N}\log\Big(\sum_{i=1}^{I} w_i \Upsilon_{n,n}^{(i)}\Big) + \frac{1}{\alpha}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Psi}\bar{\boldsymbol{x}}_y - \frac{1}{2\alpha}\mathrm{tr}(\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi}\boldsymbol{R}_y) - \frac{1}{2\beta}\mathrm{tr}(\boldsymbol{R}_\epsilon) - \frac{1}{2\alpha}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u} - \lambda^p\sum_{i=1}^{I}|w_i|^p, \quad (12)$$

where $\Upsilon_{n,n}^{(i)}$ is the n^{th} diagonal component of the matrix $\boldsymbol{\Upsilon}^{(i)}$. We utilize the auxiliary function method [16] to maximize $f(\boldsymbol{w}, u, \beta)$. To define an auxiliary function of Eq. (12), we use two inequalities:

$$\sum_{n=1}^{N}\log\Big(\sum_{i=1}^{I} w_i\Upsilon_{n,n}^{(i)}\Big) \geq \sum_{n=1}^{N}\sum_{i=1}^{I}\gamma_{i,n}\log\frac{w_i\Upsilon_{n,n}^{(i)}}{\gamma_{i,n}}, \quad (13)$$

$$|w_i|^p \leq p|\bar{w}_i|^{p-1}w_i + |\bar{w}_i|^p - p|\bar{w}_i|^p \ \ (0 < p \leq 1), \quad (14)$$

with auxiliary variables $\bar{\boldsymbol{w}} := \{\bar{w}_1, \bar{w}_2, \ldots, \bar{w}_I\}$ and $\boldsymbol{\gamma} := \{\gamma_{1,1}, \ldots, \gamma_{I,N}\}$. Hence, we define the auxiliary function

$$f^+(\boldsymbol{w}, u, \beta, \bar{\boldsymbol{w}}, \boldsymbol{\gamma}) := -\frac{N}{2}\log\alpha\beta + \sum_{n=1}^{N}\sum_{i=1}^{I}\gamma_{i,n}\log\frac{w_i\Upsilon_{n,n}^{(i)}}{\gamma_{i,n}} + \frac{1}{\alpha}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Psi}\bar{\boldsymbol{x}}_y - \frac{1}{2\alpha}\mathrm{tr}(\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi}\boldsymbol{R}_y) - \frac{1}{2\beta}\mathrm{tr}(\boldsymbol{R}_\epsilon) - \frac{1}{2\alpha}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u} - \lambda^p\sum_{i=1}^{I}\Big(p|\bar{w}_i|^{p-1}w_i + |\bar{w}_i|^p - p|\bar{w}_i|^p\Big), \quad (15)$$

where $f(\boldsymbol{w}, u, \beta) \geq f^+(\boldsymbol{w}, u, \beta, \bar{\boldsymbol{w}}, \boldsymbol{\gamma})$ is satisfied. We have equality if $\bar{w}_i = w_i$ and $\gamma_{i,n} = \bar{w}_i\Upsilon_{n,n}^{(i)}/\sum_{i'}\bar{w}_{i'}\Upsilon_{n,n}^{(i')}$, hence Eq. (15) satisfies the definition of the auxiliary function method. By iteratively updating $\mathbb{E}[\boldsymbol{x}|\boldsymbol{o};\Theta']$, $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}|\boldsymbol{o};\Theta']$, Θ, $\bar{\boldsymbol{w}}$, and $\boldsymbol{\gamma}$, $f(\boldsymbol{w}, u, \beta)$ will converge to a stationary point. As the update rules for $\bar{\boldsymbol{w}}$ and $\boldsymbol{\gamma}$ are shown above, we only need to derive the update rule for Θ. Differentiating Eq. (15) partially w.r.t. $w_{i'}$ and setting to zero, we obtain

$$\frac{1}{\alpha}\sum_{i=1}^{I}\mathrm{tr}\Big(\boldsymbol{R}_y{}^{\mathrm{T}}\boldsymbol{\Upsilon}^{(i)\,\mathrm{T}}\boldsymbol{\Upsilon}^{(i')}\Big)w_i - \frac{1}{\alpha}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Upsilon}^{(i')}\bar{\boldsymbol{x}}_y + \lambda^p p|\bar{w}_{i'}|^{p-2}w_{i'} - \sum_{n=1}^{N}\frac{\gamma_{i',n}}{w_{i'}} = 0. \quad (16)$$

Solving this nonlinear simultaneous equation for $i' = 1, \ldots, I$ based on the coordinate descent method, we obtain $\boldsymbol{w}$. Update rules for u and β are the followings:

$$u = \frac{1}{N}\boldsymbol{1}^{\mathrm{T}}\boldsymbol{\Psi}\bar{\boldsymbol{x}}_y, \quad \beta = \frac{1}{N}\mathrm{tr}\big(\boldsymbol{R}_\epsilon\big). \quad (17)$$

The proposed algorithm is summarized as follows:

Initial step: Initialize Θ ($\boldsymbol{w}$, u, and β).

E-step: Evaluate $\mathbb{E}[\boldsymbol{x}|\boldsymbol{o};\Theta']$ and $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}|\boldsymbol{o};\Theta']$ using the current parameter values Θ' and update $\bar{\boldsymbol{w}}$ and $\boldsymbol{\gamma}$.

M-step: Update $\boldsymbol{w}$, u, and β according to Eqs. (16) and (17)

If the convergence criterion of Eq. (12) is not satisfied, then let $\Theta' \leftarrow \Theta$ and return to the E-step. We can similarly derive the M-step for the difference approximation method Eq. (2) based on the auxiliary function method.

5. Experiment

The effectiveness of using the proposed method to decompose F0 contours into musical-note components and dynamics components is evaluated experimentally. We compare the parameter estimation accuracy of three methods, i.e., the conventional method based on the maximum-likelihood approach [13] (Conventional), the proposed method using the difference approximation (Method 1), and the proposed method using the impulse response approximation (Method 2). a and b were initially set at $a = 100$ and $b = 20$ calculated by $\zeta = 1.0$ and $\Omega = 0.1$. Using ζ values that varied from 0.01 to 2 at 0.02 intervals and Ω values that varied from 0.05 to 3 at 0.005 intervals, we calculated $\boldsymbol{\Upsilon}$ ($I = 5100$). α, β, and w_i were initially set at $\alpha = 2$, $\beta = 100$, and $w_i = 1/I$. u was initially set at the median value of the observed signal $\boldsymbol{o}$. λ and p were set at $\lambda = 10000$ and $p = 0.8$. We determined these initial parameters empirically.

For the first experiment we evaluated the step signal and the impulse response decomposed from the F0 contour to determine whether the proposed methods were affected by local minimum problem. First, we synthesized one hundred F0 contours artificially using randomly-determined ζ, Ω, and u values, where N, α, and β were set at 300, 2, and 100, respectively. Then, we estimated model parameters for each F0 contour. Finally, we calculated the root mean square error (RMSE) between the step signals, which we obtained by using the original u and the estimated parameter u, respectively. We also calculated the RMSE between the impulse responses, which we obtained by using the original ζ and Ω values and the model parameters, respectively. Tab. 1 shows the average values of the RMSEs between these signals over all segments. The best value was obtained with Method 2. We consider that Method 2 solves the local minimum problem more effectively.

Table 1: *Root mean square error (RMSE) between signals*

	Step signals	Impulse responses	$\boldsymbol{o}$ and $\boldsymbol{\mu}$ [cent]
Conventional	21.7	5.29×10^{-2}	44.7
Method 1	1.94	65.2×10^{-4}	42.9
Method 2	0.67	2.98×10^{-4}	34.7

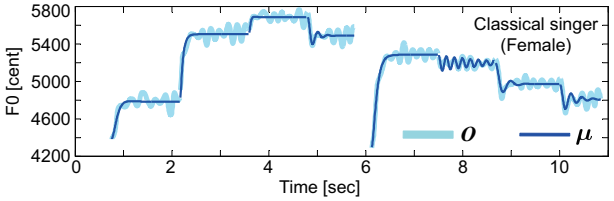

Figure 2: *Comparing observed signal $\boldsymbol{o}$ with signal $\boldsymbol{\mu}$ generated by model parameters of Method 2*

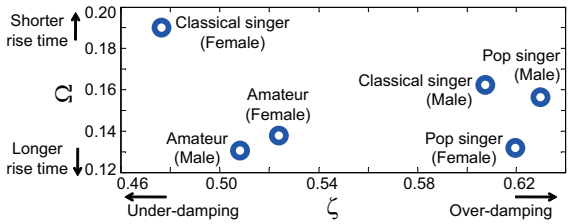

Figure 3: *Estimated ζ and Ω for each singer*

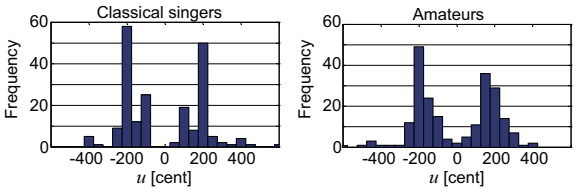

Figure 4: *Histograms of u for classical and amateur singers*

For the second experiment we evaluated the convergence performance of the proposed methods using real F0 contours. We used singing signals provided by six singers. We recorded signals for each gender in the professional classical, professional pop, and amateur categories. Each subject sang songs with Japanese lyrics (two patterns) and hummed, in both cases without musical accompaniment. The songs were "Twinkle, Twinkle, Little Star" and "Ode to Joy". In all, 36 singing signals were recorded. The F0 contour was estimated every 5 ms ($\Delta = 5$ms) using YIN [17] and represented in cents, so that one-tempered semitone corresponded to 100 cents. Using the linear interpolation, we smoothed F0 in voiceless consonants. We then divided the F0 contours into 1323 segments manually as seen in Fig. 1 and subtracted the starting F0 value over each segment so that the starting F0 value of the segment was zero. Finally, we estimated the model parameters for each segment.

The average values of the RMSE between an observed signal $\boldsymbol{o}$ and the signal $\boldsymbol{\mu}$ generated by model parameters are shown in Tab. 1. We obtained the smallest RMSE for Method 2. The temporal attack of the musical note was successfully represented by the model parameters (see Fig. 2). Hence, the RMSE corresponds to a residual component such as *vibrato*. Since Method 2 represents the dynamics by a linear combination of various oscillation bases and regularizes weighting parameters $\boldsymbol{w}$, we consider that over-fitting for the model is controlled and the RMSE is the smallest among three methods. Furthermore, the periodic components such as *vibrato* are included in the residue $\boldsymbol{\epsilon}$. Instead of modeling $\boldsymbol{\epsilon}$ by Gaussian white noise, we plan to model $\boldsymbol{\epsilon}$ using the periodic kernel similarly with GPs.

Fig. 3 shows ζ and Ω for each singer. Estimating $\boldsymbol{w}$ for each segment using Method 2, we plotted the average values of ζ and Ω corresponding to the maximum value of the elements of $\boldsymbol{w}$ over all segments. If ζ and Ω are small, this means that the oscillation phenomenon is *Under-damping* such as *overshoot* and the rise time of the musical note is longer. Since amateurs have poor singing skills, ζ and Ω are small. Fig. 4 shows histograms of parameter u for classical and amateur singers. We confirmed that the histogram has peaks at the integral multiples of the semitone for classical and pop singers. However, the histogram peaks for amateurs are blurred. This means that it is difficult for amateurs to sing at the correct pitch. In the future, we plan to analyze singing styles based on estimated ζ, Ω, and u values using a large singing database.

6. Concluding remarks

This paper proposed a statistical representation of F0 dynamics in singing voices based on GPs with multiple oscillation bases, and derived a model parameter estimation algorithm based on the EM method. Our experiments showed that the proposed method can solve the local minimum problem and decompose an F0 contour into a musical note component and a dynamics component. Furthermore, we examined the differences between singing styles based on the estimated parameters. In the future, we plan to divide the F0 contour into segments using Hidden Markov Model (HMM), and evaluate its ability to detect particular singing behaviors automatically such as *vibrato* and *overshoot* and singing voice synthesis. We also want to employ the proposed method for automatic speech recognition and gesture recognition.

7. References

[1] T. Saitou *et al.*, "Speech-To-Singing Synthesis: Converting speaking voices to singing voices by controlling acoustic features unique to singing voices," in *WASSPA 2007*, pp. 215–218.

[2] ——, "Acoustic and perceptual effects of vocal training in amateur male singing," in *EUROSPEECH 2009*, pp. 832–835.

[3] T. Kako *et al.*, "Automatic identification for singing style based on sung melodic contour characterized in phase plane," in *Proc. ISMIR 2009*, pp. 393–397.

[4] T. Nakano *et al.*, "An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features," in *Proc. ICSLP 2006*, pp. 1706–1709.

[5] J. Sundberg, "The KTH synthesis of singing," *Advances in Cognitive Psychology. Special issue on Music Performance*, vol. 2, pp. 131–143, 2006.

[6] J. Bonada and A. Loscos, "Sample-based singing voice synthesizer by spectral concatenation," in *Proc. SMAC 2003*.

[7] T. Nakano *et al.*, "VocaListener: A Singing-to-Singing synthesis system based on iterative parameter estimation," in *Proc. SMC 2009*, pp. 343–348.

[8] S. Fukayama *et al.*, "Orpheus: Automatic composition system considering prosody of Japanese lyrics," in *Proc. ICEC 2009*, pp. 309–310.

[9] R. B. Dannenberg, W. P. Birmingham, *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.

[10] P. Proutskova and M. Casey, "You call *That* singing? ensemble classification for multi-cultural collections of music recordings," in *Proc. ISMIR 2009*, pp. 759–764.

[11] N. Minematsu *et al.*, "Prosodic modeling of Nagauta singing and its evaluation," in *Proc. SpeechProsody 2004*, pp. 487–490.

[12] H. Fujisaki, "A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour," in *Vocal Physiology: Voice Production, Mechanisms and Functions, (O. Fujimura, ed.)*. Raven Press, 1988, pp. 347–355.

[13] Y. Ohishi *et al.*, "Parameter estimation method of F0 control model for singing voices," in *Proc. ICSLP 2008*, pp. 139–142.

[14] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, Mass, USA, 2006.

[15] F. Bach *et al.*, "Multiple kernel learning, conic duality, and the smo algorithm," in *Proc. ICML 2004*, pp. 6–13.

[16] H. Kameoka *et al.*, "Complex NMF: A new sparse representation for acoustic signals," in *Proc. ICASSP 2009*, pp. 3437–3440.

[17] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *JASA*, vol. 111, no. 4, pp. 1917–1930, 2002.