A Stochastic Model of Singing Voice F_0 Contours for Characterizing Expressive Dynamic Components

Yasunori Ohishi[†], *Hirokazu Kameoka*[†], *Daichi Mochihashi*[‡], *Kunio Kashino*[†]

[†]NTT Communication Science Laboratories, NTT Corporation,
[‡]The Institute of Statistical Mathematics
{ohishi.yasunori,kameoka.hirokazu,kashino.kunio}@lab.ntt.co.jp, daichi@ism.ac.jp

Abstract

We present a novel stochastic model of singing voice fundamental frequency (F_0) contours for characterizing expressive dynamic components, such as *vibrato* and *portamento*. Although dynamic components can be important features for any singing voice applications, modeling and extracting these components from a raw F_0 contour have yet to be accomplished. Therefore, we describe a process for generating dynamic components explicitly and represent the process as a stochastic model. Then we develop an algorithm for estimating the model parameters based on statistical techniques. Experimental results show that our method successfully extracts the expressive components from raw F_0 contours.

Index Terms: Singing voice, Fundamental frequency, Second-order linear system, Stochastic model

1. Introduction

The fundamental frequency (F_0) contour in a singing voice contains two main types of dynamic components. One is those generated by the physical constraints of the vocal folds, such as *overshoot*, *preparation* and *fine fluctuations* [1, 2]. The other type is dynamic components generated by singer's musical expressive intentions, such as *vibrato* and *portamento* [3]. Most previous papers have reported that these dynamic components strongly affect singing-voice perception, and that the former relates to the naturalness and individuality of a singing voice while the latter relates to singing styles and skills [4]. Accordingly, extracting these components from a raw F_0 contour automatically can be potentially very beneficial for any singing voice applications, such as singer identification, singing skill evaluation and the synthesis of more natural and varied singing voices [5, 6, 7].

Previous studies have represented the dynamic components generated by the physical constraints as the output of a second-order linear system [8, 9]. The input is a stepwise signal representing a musical-note sequence. The transfer function of the system is described as

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

where ζ and Ω denote the damping ratio and the natural angular frequency, respectively. The dynamic components were controlled by adjusting these parameters manually [9]. In [10], we proposed a method for solving the inverse problem of estimating the parameters from a raw F_0 contour. However, modeling and extracting the dynamic components generated by expressive intentions from the raw F_0 contour have not yet been accomplished. For example, although the rate and amplitude of *vibrato* are time-varying, *vibrato* has been conventionally modeled as a sinusoidal signal [9]. We consider that *vibrato* is controlled variedly by the singer's expressive intentions.

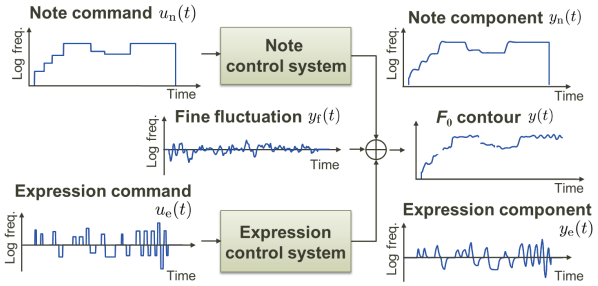

Figure 1: *Process of generating a singing voice F_0 contour*

In this paper we propose a new stochastic model of singing voice F_0 contours to describe the process of generating various dynamic components explicitly. This model is based on an analogy with the Fujisaki model [11], which describes the process of generating speech F_0 contours, and assumes that an F_0 contour on a logarithmic scale, $y(t)$, where t is time, is the superposition of three components (Fig. 1). The note and expression components are the outputs of second-order linear systems driven by the note and expression commands that correspond to the musical note sequence and the musical expressive intentions, respectively. The note component contains the note transition and *overshoot*, and the expression component contains *vibrato* and *portamento*. The fine fluctuation component consists of an irregular fluctuation higher than 10 Hz [9]. Then, we formulate a discrete-time stochastic version of this process and develop an algorithm for estimating the model parameters based on statistical techniques. Experimental results show that our method successfully extracts the expressive intentions for *vibrato* and *portamento* from a raw F_0 contour. Furthermore, we verify that the extracted expression components are perceived as *vibrato* and *portamento* through a psychoacoustic experiment.

2. Discretization of proposed model

We apply a backward difference s-to-z transform, $s \simeq (1-z^{-1})/t_0$, to the note and expression control mechanisms described as continuous-time linear systems to obtain a discrete-time version, where t_0 is the sampling period for the discretization. The transfer function of the inverse system $\mathcal{H}_{\rm n}^{-1}(s)$ for the note control mechanism can be written in the z-domain as

$$\mathcal{H}_{\rm n}^{-1}(z) = a_2 z^{-2} + a_1 z^{-1} + a_0, \qquad (2)$$

where $a_2 = \varphi^2,\ a_1 = -2\varphi(\psi+\varphi),\ a_0 = 1+2\varphi\psi+\varphi^2,\ \varphi = 1/(\Omega t_0)$, and $\psi = \zeta$. Here we use $u_{\rm n}[k]$ to denote the discrete-time version of the note command where k indicates the discrete-time index. The discrete-time version of the note component, $y_{\rm n}[k]$, can thus be regarded as the output of a constrained all-pole system whose characteristics are governed by parameters $\varphi,\ \psi$, such that

$$u_{\rm n}[k] = a_0 y_{\rm n}[k] + a_1 y_{\rm n}[k-1] + a_2 y_{\rm n}[k-2]. \qquad (3)$$

In the same way, the relationship between the expression command $u_{\rm e}[k]$ and the expression component $y_{\rm e}[k]$ is described as

$$u_{\rm e}[k] = b_0 y_{\rm e}[k] + b_1 y_{\rm e}[k-1] + b_2 y_{\rm e}[k-2], \qquad (4)$$

where $b_2 = \xi^2,\ b_1 = -2\xi(1+\xi),\ b_0 = 1+2\xi+\xi^2$, and $\xi = 1/(\Omega t_0)$. For the fine fluctuation component, we use $y_{\rm f}[k]$ to denote the discrete-time version of $y_{\rm f}(t)$. Altogether, the discrete-time version of our model can be expressed as the superposition of the three components: $y[k] = y_{\rm n}[k] + y_{\rm e}[k] + y_{\rm f}[k]$.

3. Statistical formulation

We assume that note and expression commands represent the musical-note sequence and the musical expressive intentions of a singer, respectively. As a convenient way of incorporating these assumptions into the command functions, we propose modeling the $u_{\rm n}[k]$ and $u_{\rm e}[k]$ pair using a hidden Markov model (HMM). Let us arrange $u_{\rm n}[k]$ and $u_{\rm e}[k]$ into a vector $\boldsymbol{o}[k]$. We assume that $\boldsymbol{o}[k]$ is a random vector driven by an additive white Gaussian noise $(\epsilon_{\rm n}[k], \epsilon_{\rm e}[k])^{\rm T}$. Let $\epsilon_{\rm n}[k] \sim \mathcal{N}(0, \sigma_{\rm n}^2)$ and $\epsilon_{\rm e}[k] \sim \mathcal{N}(0, \sigma_{\rm e}^2)$ be mutually independent, then

$$\boldsymbol{o}[k] \sim \mathcal{N}(\boldsymbol{\nu}[k], \boldsymbol{\Upsilon}),\ \boldsymbol{\nu}[k] := \begin{bmatrix}\mu_{\rm n}[k]\\ \mu_{\rm e}[k]\end{bmatrix},\ \boldsymbol{\Upsilon} := \begin{bmatrix}\sigma_{\rm n}^2 & 0\\ 0 & \sigma_{\rm e}^2\end{bmatrix}.$$

This can be viewed as an HMM in which the output distribution of each state is a Gaussian distribution. The mean vector $\boldsymbol{\nu}[k]$ is thus considered to evolve in time as the result of a state transition. This way of thinking allows us to incorporate the assumptions into $\mu_{\rm n}[k]$ and $\mu_{\rm o}[k]$ by simply constraining the path of the state transitions (Fig. 2).

The present HMM consists of $I \times J$ distinct states. In each of these states, $\mu_{\rm n}[k]$ can take a non-zero value $A_{\rm n}^{(i)} + d_i$, which is assumed to be constant in time. $A_{\rm n}^{(i)}$ is the pitch value of the i-th melodic note in a given musical score, and d_i is a pitch shift parameter. In

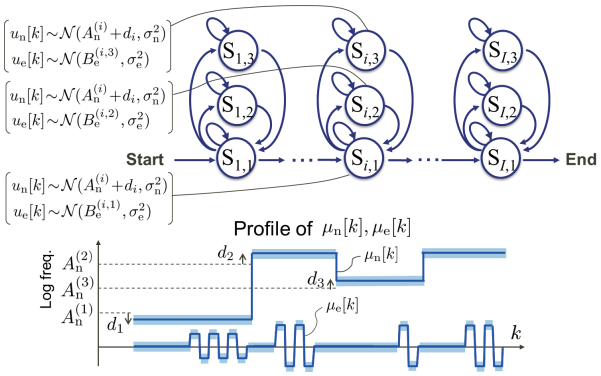

Figure 2: Command modeling with HMM

the i-th note, direct state transitions from state $\mathrm{S}_{i,j}$ to state $\mathrm{S}_{i,j'}$ $(j \neq j',\ 2 \leq j \leq J,\ 2 \leq j' \leq J)$ without passing through state $\mathrm{S}_{i,1}$ are not allowed. This constraint restricts $\mu_{\rm e}[k]$ to consisting of rectangular pulses. To sum up, the proposed HMM is defined as follows:

$$\begin{aligned}
\text{Output sequence: } & \{\boldsymbol{o}[k]\}_{k=1}^{K}\\
\text{Set of states: } & \mathcal{S} := \{\mathrm{S}_{i,j}\}_{i=1,j=1}^{I,J}\\
\text{State sequence: } & \{s_k\}_{k=1}^{K}\\
\text{Output distribution: } & P(\boldsymbol{o}[k]|s_k = n) = \mathcal{N}(\boldsymbol{c}_n[k], \boldsymbol{\Upsilon}),\\
& \boldsymbol{c}_n[k] = (A_{\rm n}^{(i)} + d_i, B_{\rm e}^{(i,j)})^{\rm T},\ n = \mathrm{S}_{i,j}\\
\text{Transition probability: } & \phi_{n',n} := \log P(s_k = n | s_{k-1} = n')
\end{aligned}$$

For simplicity, we treat the transition probabilities $\phi_{n',n}$ as constant parameters in this paper, so that the free parameters to be determined in our command model consist of the state sequence, $\{s_k\}_{k=1}^{K}$, the pitch shift parameters, $\{d_i\}_{i=1}^{I}$, the magnitude of the expression command, $\{B_{\rm e}^{(i,j)}\}_{i=1,j=1}^{I,J}$, and the variance of the output distribution, $\sigma_{\rm n}^2,\ \sigma_{\rm e}^2$. Hereafter, we use $\theta_{\rm u}$ to denote all these parameters. Once the state sequence is specified, the mean sequences are determined simultaneously by $(\mu_{\rm n}[k],\ \mu_{\rm e}[k])^{\rm T} \leftarrow \boldsymbol{c}_{s_k}[k]$.

We derive the probability density function of the F_0 contour, $y[1], \ldots, y[K]$, based on the statistical modeling of the commands,

$$\boldsymbol{u}_{\rm n}|\theta_u \sim \mathcal{N}(\boldsymbol{\mu}_{\rm n}, \sigma_{\rm n}^2 \boldsymbol{I}_K),\ \boldsymbol{u}_{\rm e}|\theta_u \sim \mathcal{N}(\boldsymbol{\mu}_{\rm e}, \sigma_{\rm e}^2 \boldsymbol{I}_K), \qquad (5)$$

where $\boldsymbol{u}_{\rm n} := (u_{\rm n}[1], \ldots, u_{\rm n}[K])^{\rm T}$, $\boldsymbol{u}_{\rm e} := (u_{\rm e}[1], \ldots, u_{\rm e}[K])^{\rm T}$, $\boldsymbol{\mu}_{\rm n} := (\mu_{\rm n}[1], \ldots, \mu_{\rm n}[K])^{\rm T}$, $\boldsymbol{\mu}_{\rm e} := (\mu_{\rm e}[1], \ldots, \mu_{\rm e}[K])^{\rm T}$, and $\boldsymbol{I}_K$ denotes the $K \times K$ identity matrix. By using the linear equation given in Section 2, the note component $\boldsymbol{y}_{\rm n} := (y_{\rm n}[1], \ldots, y_{\rm n}[K])^{\rm T}$ and the expression component $\boldsymbol{y}_{\rm e} := (y_{\rm e}[1], \ldots, y_{\rm e}[K])^{\rm T}$ can be written in terms of $\boldsymbol{u}_{\rm n}$ and $\boldsymbol{u}_{\rm e}$, respectively, such that $\boldsymbol{u}_{\rm n} = \boldsymbol{A}\boldsymbol{y}_{\rm n}$ and $\boldsymbol{u}_{\rm e} = \boldsymbol{B}\boldsymbol{y}_{\rm e}$, where

$$\boldsymbol{A} := \begin{bmatrix} a_0^{(1)} & & & & O\\ \ddots & \ddots & \ddots & & \\ & a_2^{(i)} & a_1^{(i)} & a_0^{(i)} & \\ & & \ddots & \ddots & \ddots\\ O & & & a_2^{(I)} & a_1^{(I)} & a_0^{(I)} \end{bmatrix},\ \boldsymbol{B} := \begin{bmatrix} b_0 & & & O\\ b_1 & b_0 & & \\ b_2 & b_1 & b_0 & \\ & \ddots & \ddots & \ddots\\ O & & b_2 & b_1 & b_0 \end{bmatrix}.$$

On the basis of a previous study [9], we assume that the note control parameters, $\varphi,\ \psi$, vary for each note and

describe the parameters as $\{\varphi^{(i)}, \psi^{(i)}\}_{i=1}^{I}$. Hence, it follows from Eq. (5) that

$$\boldsymbol{y}_{\mathrm{n}} \sim \mathcal{N}(\boldsymbol{A}^{-1}\boldsymbol{\mu}_{\mathrm{n}}, \sigma_{\mathrm{n}}^2 \boldsymbol{A}^{-1}(\boldsymbol{A}^{-1})^{\mathrm{T}}), \tag{6}$$

$$\boldsymbol{y}_{\mathrm{e}} \sim \mathcal{N}(\boldsymbol{B}^{-1}\boldsymbol{\mu}_{\mathrm{e}}, \sigma_{\mathrm{e}}^2 \boldsymbol{B}^{-1}(\boldsymbol{B}^{-1})^{\mathrm{T}}). \tag{7}$$

As for the fine fluctuation component $\boldsymbol{y}_{\mathrm{f}} := (y_{\mathrm{f}}[1], \ldots, y_{\mathrm{f}}[K])^{\mathrm{T}}$, we assume that it is white Gaussian noise such that $\boldsymbol{y}_{\mathrm{f}} \sim \mathcal{N}(\boldsymbol{0}, \sigma_{\mathrm{f}}^2 \boldsymbol{I}_K)$.

Overall, the likelihood function of the model parameters Θ given $\boldsymbol{y}$ can be written as

$$P(\boldsymbol{y}|\Theta) = \frac{|\boldsymbol{\Sigma}^{-1}|^{1/2}}{(2\pi)^{K/2}} \exp\Big\{ -\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\Big\},$$

$$\boldsymbol{\mu} = \boldsymbol{A}^{-1}\boldsymbol{\mu}_{\mathrm{n}} + \boldsymbol{B}^{-1}\boldsymbol{\mu}_{\mathrm{e}}, \tag{8}$$

$$\boldsymbol{\Sigma} = \sigma_{\mathrm{n}}^2 \boldsymbol{A}^{-1}(\boldsymbol{A}^{-1})^{\mathrm{T}} + \sigma_{\mathrm{e}}^2 \boldsymbol{B}^{-1}(\boldsymbol{B}^{-1})^{\mathrm{T}} + \sigma_{\mathrm{f}}^2 \boldsymbol{I}_K,$$

where $\Theta := \{\theta_{\mathrm{u}}, \{\varphi^{(i)}, \psi^{(i)}\}_{i=1}^{I}, \xi, \sigma_{\mathrm{f}}^2\}$. As for the prior probability of Θ, we assume that the parameters are independent of each other and distributed according to the following probability density functions: $P(s_1, \ldots, s_K) = P(s_1)\prod_{k=2}^{K} P(s_k|s_{k-1})$, $B_{\mathrm{e}}^{(i,j)} \sim \mathcal{N}(\mu_{B^{(i,j)}}, {\sigma_B}^2)$, $\varphi^{(i)} \sim \mathcal{N}(\mu_\varphi, \sigma_\varphi^2)$, $\psi^{(i)} \sim \mathcal{N}(\mu_\psi, \sigma_\psi^2)$, $\xi \sim \mathcal{N}(\mu_\xi, \sigma_\xi^2)$, $d_i \sim \mathcal{N}(0, \sigma_{\mathrm{d}}^2)$. σ_{n}^2, σ_{e}^2 and σ_{f}^2 are uniformly distributed.

4. Parameter estimation algorithm

We describe an iterative algorithm, which locally maximizes the posterior density $P(\Theta|\boldsymbol{y}) \propto P(\boldsymbol{y}|\Theta)P(\Theta)$. By regarding a set consisting of the note, expression and fine fluctuation components, $\boldsymbol{x} := (\boldsymbol{y}_{\mathrm{n}}^{\mathrm{T}}, \boldsymbol{y}_{\mathrm{e}}^{\mathrm{T}}, \boldsymbol{y}_{\mathrm{f}}^{\mathrm{T}})^{\mathrm{T}}$, as the complete data, this problem can be viewed as an incomplete data problem, which can be dealt with using the EM algorithm. Taking the conditional expectation of the log-likelihood with respect to $\boldsymbol{x}$ given $\boldsymbol{y}$ and $\Theta = \Theta'$, and then adding $\log P(\Theta)$, we obtain the Q function

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

$$\boldsymbol{m} := \begin{bmatrix} \boldsymbol{A}^{-1}\boldsymbol{\mu}_{\mathrm{n}} \\ \boldsymbol{B}^{-1}\boldsymbol{\mu}_{\mathrm{e}} \\ \boldsymbol{0} \end{bmatrix}, \boldsymbol{\Lambda}^{-1} := \begin{bmatrix} \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}/\sigma_{\mathrm{n}}^2 & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}/\sigma_{\mathrm{e}}^2 & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{I}_K/\sigma_{\mathrm{f}}^2 \end{bmatrix}.$$

Because the relationship between the incomplete data and the complete data can be written as $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x}$ where $\boldsymbol{H} := [\boldsymbol{I}_K, \boldsymbol{I}_K, \boldsymbol{I}_K]$, $\mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\Theta']$ and $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}|\boldsymbol{y};\Theta']$ are given explicitly as

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

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

The E-step computes $\mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\Theta']$ and $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}|\boldsymbol{y};\Theta']$. In the M-step, we maximize $Q(\Theta, \Theta')$ with respect to each parameter of Θ in which we treat $\mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\Theta']$ and $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}|\boldsymbol{y};\Theta']$ as constants. In particular, the state sequence $\{s_k\}_{k=1}^{K}$ can be solved efficiently using the Viterbi algorithm [12]. Owing to space limitations, we omit all the mathematical details.

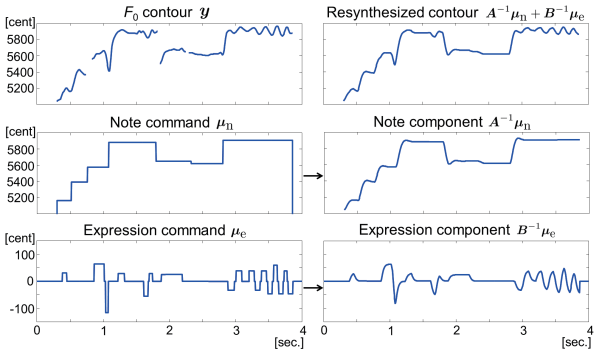

Figure 3: *Estimation result from a raw F_0 contour*

Looking back at Eq. (8), we have thus far implicitly assumed that we are given a set of F_0 observations on the whole sample period. However, the F_0 data in unvoiced regions are missing. This can simply be viewed as a missing data imputation problem, which can be effectively dealt with using EM algorithm [12].

5. Experimental evaluations

We tested our method in two experiments. Experiment A evaluated the expression commands estimated from raw F_0 contours, experiment B evaluated singing voices synthesized using resynthesized contours subjectively.

In these experiments, we used an F_0 contour annotated manually in the AIST annotation [13]. The title of the song is "PROLOGUE" (RWC-MDB-P-2001 [14], Song No.07) . Although the F_0 contour should essentially be estimated from the acoustic signal, we used these data to evaluate the upper limit of the performance of our method. The F_0 values were annotated every 5 ms ($t_0 = 5$ ms) and were represented in cents so that one equal-tempered semitone corresponded to 100 cents. We used the MIDI data to determine I and $\{A_{\mathrm{n}}^{(i)}\}_{i=1}^{I}$. The initial conditions are shown in the footnote[1].

Experiment A: Fig. 3 shows note command $\boldsymbol{\mu}_{\mathrm{n}}$, expression command $\boldsymbol{\mu}_{\mathrm{e}}$, note component $\boldsymbol{A}^{-1}\boldsymbol{\mu}_{\mathrm{n}}$ and expression component $\boldsymbol{B}^{-1}\boldsymbol{\mu}_{\mathrm{e}}$ estimated from a raw F_0 contour. Resynthesized contour $\boldsymbol{A}^{-1}\boldsymbol{\mu}_{\mathrm{n}} + \boldsymbol{B}^{-1}\boldsymbol{\mu}_{\mathrm{e}}$ is represented as the sum of two components, and is similar to the F_0 contour. Since σ_{f} is 13.6 cent, we confirm that $\boldsymbol{y}_{\mathrm{f}}$ can be estimated as a fine fluctuation component.

In the F_0 contour of Fig. 4(a), *portamento*, which is a vocal slide between two notes, is observed. The result shows that two rectangular pulses for changing pitch gradually are estimated in expression command $\boldsymbol{\mu}_{\mathrm{e}}$. On

[1]The iteration was run for 1000 iterations. J was set at 5. The state transition probabilities were set at $\phi_{\mathrm{S}_{i,1},\mathrm{S}_{i,1}} = \log(0.9999 \times (J-1)/J)$, $\phi_{\mathrm{S}_{i,1},\mathrm{S}_{i,j}} = \log(0.9999/J)$, $\phi_{\mathrm{S}_{i,1},\mathrm{S}_{i+1,1}} = \log(0.0001)$, $\phi_{\mathrm{S}_{i,j},\mathrm{S}_{i,j}} = \log(0.9999)$, $\phi_{\mathrm{S}_{i,j},\mathrm{S}_{i,1}} = \log(0.0001)$, with $1 \le i \le I$ and $2 \le j \le J$. The parameters of the prior distributions were set at $\mu_\varphi = 6$, $\sigma_\varphi^2 = 0.1$, $\mu_\psi = 0.6$, $\sigma_\psi^2 = 0.02$, $\mu_\xi = 3$, $\sigma_\xi^2 = 0.1$, $\sigma_{\mathrm{d}}^2 = 2500$, $\sigma_B^2 = 100$, $\mu_{B^{(i,1)}} = 0$, $\mu_{B^{(i,2)}} = 30$, $\mu_{B^{(i,3)}} = -30$, $\mu_{B^{(i,4)}} = 60$, $\mu_{B^{(i,5)}} = -60$, with $1 \le i \le I$.

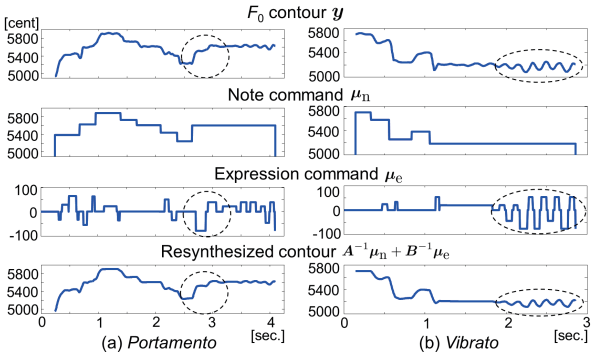

Figure 4: *Estimation of commands for portamento and vibrato*

the other hand, in the F_0 contour in Fig. 4(b), *vibrato* is observed in the latter part of the note. The result shows that the expression command consisted of quasi-periodic rectangular pulses. We found that our method extracted these commands from the raw F_0 contours and obtained resynthesized contours similar to the F_0 contours.

Experiment B: We conducted a psychoacoustic experiment to evaluate the estimated expression components subjectively. First, we prepared the following contours: (1) Note command $\boldsymbol{\mu}_{\mathrm{n}}$, (2) Note component $\boldsymbol{A}^{-1}\boldsymbol{\mu}_{\mathrm{n}}$, (3) Resynthesized contour $\boldsymbol{A}^{-1}\boldsymbol{\mu}_{\mathrm{n}} + \boldsymbol{B}^{-1}\boldsymbol{\mu}_{\mathrm{e}}$, (4) F_0 contour $\boldsymbol{y}$. Then, we synthesized four kinds of singing voices using singing synthesis software based on Yamaha's Vocaloid3 technology [15]. Specifically, notes and lyrics were manually entered using the estimated note command (all the syllables of the lyrics were /na/), *pitch bend (PIT)* and *pitch bend sensitivity (PBS)* were adjusted using the difference between the note command and each contour. Finally, we divided synthesized singing voice signals into short signals consisting of four bars and used these signals as stimuli. The total number of stimuli was 96 signals (24 signals for each contour).

Scheffe's method of paired comparison was used to evaluate the expressiveness of the stimuli. As shown at the top of Fig. 5, the subjects decided which stimulus was a more expressive singing voice according to an eleven-grade evaluation measure. The expressiveness of a singing voice has a multidimensional meaning. However, in this experiment, we defined an expressive singing voice as a singing voice with vocal ornamentations such as *vibrato* and *portamento*. The pair-wise stimuli were presented through binaural headphones at a comfortable sound pressure level. Each paired stimulus was randomly presented to each subject. The subjects were five males.

Fig. 5 shows the results. The evaluation values are normalized for each subject and are averaged over all subjects. Subjects perceived singing voices based on (3) and (4) contours to be more expressive than those based on (1) and (2) contours. Additionally, it was difficult for subjects to determine the difference between singing voices based on (3) and (4) contours. Therefore, we found that the singer's expressiveness is represented as the expression components estimated from the F_0 contour. Future work is to compare the expression components with *vibrato* modeled by a sinusoidal signal subjectively.

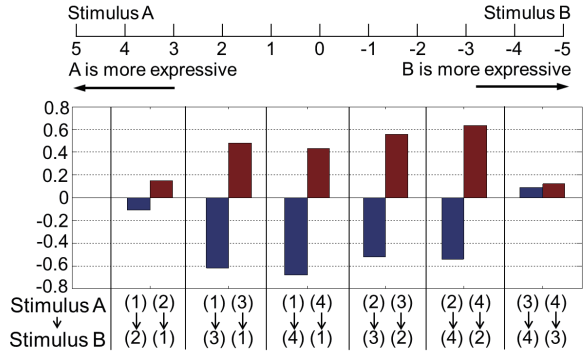

Figure 5: *Subjective evaluation results: (1) Note command, (2) Note component, (3) Resynthesized contour, and (4) F_0 contour*

6. Conclusions

We proposed a stochastic model of singing voice F_0 contours for characterizing and extracting various dynamic components and develop a model parameter estimation algorithm based on the EM approach. Experimental results show that our method decomposes a raw F_0 contour into the note and expression components explicitly and that the singer's expressiveness is contained in the expression component. In the future, we plan to evaluate model validity using the large data sets and learn the singing styles using the expression commands.

7. References

[1] J. Sundberg, *The Science of the Singing*, Northern Illinois University Press, 1987.

[2] G de Krom and G Bloothooft, "Timing and accuracy of fundamental frequency changes in singing," in *Proc. ICPhS95*.

[3] C. E. Seashore, "A musical ornament, the vibrato," in *Psychology of Music*, 1938, pp. 33–52.

[4] T. Saitou et al., "Acoustic and perceptual effects of vocal training in amateur male singing," in *Proc. EUROSPEECH 2009*.

[5] W. Tsai et al., "An automated singing evaluation method for karaoke systems," in *Proc. ICASSP 2011*.

[6] K. Oura et al., "Pitch adaptive training for HMM-based singing voice synthesis," in *Proc. ICASSP 2012*.

[7] L. Regnier et al., "Singer verification: singer model .vs. song model," in *Proc. ICASSP 2012*.

[8] N. Minematsu et al., "Prosodic modeling of Nagauta singing and its evaluation," in *Proc. SpeechProsody 2004*.

[9] T. Saitou et al., "Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis," *Speech Communication*, vol. 46, pp. 405–417, 2005.

[10] Y. Ohishi et al., "Statistical modeling of F0 dynamics in singing voices based on Gaussian processes with multiple oscillation bases," in *Proc. INTERSPEECH 2010*.

[11] 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*, pp. 347–355. Raven Press, 1988.

[12] H. Kameoka et al., "A statistical model of speech F0 contours," in *Proc. SAPA 2010*.

[13] M. Goto, "AIST annotation for the RWC music database," in *Proc. ISMIR 2006*.

[14] M. Goto et al., "RWC music database: Popular, classical, and jazz music databases," in *Proc. ISMIR 2002*.

[15] H. Kenmochi et al., "Singing synthesis as a new musical instrument," in *Proc. ICASSP 2012*.